Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA
Unidad Cuajimalpa

9 de noviembre de 2020.
Dictamen C.I. 19/2020

DICTAMEN
QUE PRESENTA LA COMISIÓN DE INVESTIGACIÓN DE LA DIVISIÓN DE CIENCIAS DE LA COMUNICACIÓN Y DISEÑO

ANTECEDENTES

I. El Consejo Divisional de Ciencias de la Comunicación y Diseño, en la sesión 10.19, celebrada el 16 de julio de 2019, integró esta Comisión en los términos señalados en el artículo 55 del Reglamento Interno de los Órganos Colegiados Académicos.

II. El Consejo Divisional designó para esta Comisión a los siguientes integrantes:

a) Órganos personales:
- ✓ Dr. Jesús Octavio Elizondo Martínez, Jefe del Departamento de Ciencias de la Comunicación;
- ✓ Dra. Cecilia Castañeda Arredondo, Jefa del Departamento de Teoría y Procesos del Diseño;
- ✓ Dr. Carlos Joel Rivero Moreno, Jefe del Departamento de Tecnologías de la Información.

b) Representantes propietarios:
- Personal académico:
  - ✓ Dr. André Moise Dorcé Ramos, Departamento de Ciencias de la Comunicación;
  - ✓ Dra. Deyanira Bedolla Pereda, Departamento de Teoría y Procesos del Diseño.
  - ✓ Dr. Tiburcio Moreno Olivos, Departamento de Tecnologías de la Información.

CONSIDERACIONES

I. La Comisión recibió, para análisis y discusión, el informe de actividades académicas desarrolladas por el **Dr. Esaú Villatoro Tello**, durante el disfrute del año sabático comprendido del 9 de septiembre de 2019 al 8 de septiembre de 2020.

II. El año sabático fue aprobado en la Sesión 03.19 celebrada el 28 de mayo de 2019 mediante Acuerdo DCCD.CD.22.03.19 del Consejo Divisional de Ciencias de la Comunicación y Diseño.

III. En la sesión 11.19 celebrada el 30 de septiembre de 2019 el Consejo Divisional de Ciencias de la Comunicación y Diseño se aprobó una modificación al programa de actividades académicas correspondientes al año sabático del Dr. Villatoro. Las cuales contemplaron dar continuidad a compromisos contraídos previamente con alumnos y que se refieren básicamente a asesorías de forma virtual.

IV. La Comisión de Investigación sesionó vía remota el día 9 de noviembre de 2020, fecha en la que concluyó su trabajo de análisis y evaluación del informe, con el presente Dictamen.

V. La Comisión contó, para su análisis, con los siguientes elementos:

- Programa de actividades académicas por desarrollar durante el periodo sabático.
- Evaluación general.

VI. La Comisión evaluó el informe de actividades académicas, las constancias y documentos que demuestran las actividades realizadas por el **Dr. Esaú Villatoro Tello,** durante el disfrute del año sabático comprendido del 9 de septiembre de 2019 al 8 de septiembre de 2020.

El sabático se centraría en realizar trabajo de investigación alrededor de temas relacionados con Procesamiento de Lenguaje Natural y sus aplicaciones, el análisis y procesamiento de documentos orales, es decir, documentos que son generados a través de técnicas de transcripción automática.

Durante la estancia en el Centro de Investigación Idiap, el profesor trabajó y colaboró principalmente con el grupo de investigación "Voz y procesamiento de Audio" (Speech and Audio Processing Group).

Uno de los principales logros de esta colaboración fue el diseño e implementación de un algoritmo de categorización de documentos orales, el cual resultó ser altamente eficiente y que está inspirado en técnicas de representación semántica denominadas Representaciones de Segundo Orden. Este algoritmo permitió obtener una publicación en la revista "The Prague Bulletin of Methematical Linguistics", entre otras.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA
Unidad Cuajimalpa

DICTAMEN

ÚNICO:

Se recomienda al Consejo Divisional dar por recibido el informe de periodo sabático del **Dr. Esaú Villatoro Tello,** conforme al plazo establecido en el artículo 231 del Reglamento de Ingreso, Promoción y Permanencia del Personal Académico y del mismo se advierte que cumplió satisfactoriamente con el programa de actividades.

VOTOS:

Integrantes	Sentido de los votos
Dr. Jesús Octavio Elizondo Martínez	A favor
Dra. Cecilia Castañeda Arredondo	A favor
Dr. Carlos Joel Rivero Moreno	A favor
Dr. André Moise Dorcé Ramos	----
Dra. Deyanira Bedolla Pereda	----
Dr. Tiburcio Moreno Olivos	A favor
Total de los votos	**4 votos a favor**

Coordinadora

Dra. Gloria Angélica Martínez De la Peña
Secretaria del Consejo Divisional de
Ciencias de la Comunicación y Diseño

Reporte de Actividades del Periodo Sabático

ESAÚ VILLATORO-TELLO, PH.D.
Departamento de Tecnologías de la Información, UAM Cuajimalpa
Noviembre 2020

1 Introducción

El presente documento describe las actividades realizadas durante mi periodo sabático (Septiembre/2019 a Agosto/2020). Como se indicó en mi plan de trabajo entregado al Consejo Académico en mayo de 2019, mi sabático se centraría principalmente en realizar trabajo de investigación alrededor de temas relacionados con Procesamiento de Lenguaje Natural y sus aplicaciones el análisis y procesamiento de documentos orales, es decir, documentos que son generados a través de técnicas de transcripción automática.

Durante mi estancia sabática en el centro de investigación Idiap[1], trabajé y colaboré principalmente con el grupo de investigación "Voz y Procesamiento de Audio" (Speech and Audio Processing Group). Uno de los principales logros de esta colaboración fue el diseño e implementación de un algoritmo de categorización de documentos orales, el cual resultó ser altamente eficiente y que esta inspirado en técnicas de representación semántica denominadas Representaciones de Segundo Orden. Este algoritmo lo evaluamos en variados conjuntos de datos, y nos permitió obtener una publicación en la revista "The Prague Bulletin of Methematical Linguistics". Actualmente, tengo entendido que Idiap quiere llevar a producción dicho método, para ofertarlo como una solución eficiente y efectiva a la tarea de identificación automática de tópicos entre sus posibles clientes.

Además de esto, tuve la oportunidad de involucrarme en otros proyectos, relacionados más estrechamente a mis intereses de investigación. Por ejemplo, participamos en tareas de Análisis de Autoría, como son la identificación de mensajes agresivos en redes sociales (MEX-A3T), y la identificación de rasgos psicológicos de los autores (OMT Task). En las secciones siguientes doy una breve descripción de estos artículos, los cuales son adjuntados como documentos probatorios.

Finalmente debo mencionar que pude continuar mi colaboración con el DTI de la UAM Cuajimalpa. Continué con la dirección de proyectos terminales, y de proyectos de servicio social que estaban activos al momento de solicitar el sabático.

2 Trabajo de Investigación

En esta sección presento una breve descripción de los trabajos publicados durante mi sabático. Estos trabajos están ya publicados; se incluye como anexo la versión en extenso de todos éstos.

[1] `https://www.idiap.ch/en`

1. **Inferring Highly-dense Representations for Clustering Broadcast Media Content.**[2] Este trabajo describe el método desarrollado para la identificación automática de tópicos en documentos orales. EL trabajo fue realizado en colaboración con el *Dr. Shantipriya Parida, Dr. Petr Motlicek, y Ondrej Bojar*.

2. **Idiap Submission to Swiss-German Language Detection Shared Task.**[3] En este artículo describimos los experimentos realizados como parte de nuestra participación en la competencia nombrada Swiss-German Language Detection Shared Task, la cual se realizó en el marco del 5th SwissText & 16th KONVENS Joint Conference 2020. Este trabajo fue realizado en colaboración con *Dr. Shantipriya Parida, Dr. Petr Motlicek, Qingran Zhan y Sajit Kumar*.

3. **Idiap & UAM participation at GermEval 2020: Classification and Regression of Cognitive and Motivational Style from Text.**[4] Este artículo describe nuestra participación en la tarea de clasificación denominada Classification of the Operant Motive Test (OMT) subtask. Estrictamente hablando, esta tarea es un problema de Perfilado de Autor. Este evento se realizó como parte del 5th SwissText & 16th KONVENS Joint Conference 2020. El trabajo se realizó en colaboración con *Dr. Shantipriya Parida, Dr. Petr Motlicek, Qingran Zhan, y Sajit Kumar*. Nuestro sistema obtuvo el segundo mejor desempeño en la tarea.

4. **Idiap and UAM Participation at MEX-A3T Evaluation Campaign.**[5]. El artículo describe nuestra participación en la tarea de identificación de noticias falsas y detección de agresividad (Fake News and Aggressiveness Analysis shared tasks). Ambas tareas se evaluaron en datos en Español de México. El trabajo se realizó en colaboración con *Dr. Shantipriya Parida, Dr. Petr Motlicek, y Gabriela Ramírez de la Rosa*. Vale la pena mencionar que en esta competencia, nuestro sistema logró obtener el primer lugar en la tarea de detección de noticias falsas.

De los artículos mencionados hasta aquí, tres son el resultado de participar en distintas tareas compartidas (shared tasks). El objetivo principal de la participación fue evaluar el desempeño de una estrategia denominada "Supervised Autoencoders" (SAE). A pesar de que SAE es una tecnología que se propuso para hacer análisis de imágenes, nuestros artículos representan los primeros en evaluar este tipo de tecnología en tareas de Procesamiento de Lenguaje, lo cual se considera una contribución importante.

A continuación listo los artículos realizados en colaboración con mis alumnos (de licenciatura y posgrado) de la UAM-Cuajimalpa, como del INAOE-Puebla.

1. **Finding Evidence Of The Sexual Predators Behavior.** Este es un resumen extendido, el cual se envío para evaluación a la conferencia *LatinX in AI Research at NeurIPS 2019*. El trabajo fue aceptado para ser presentado de forma oral en Diciembre de 2019. Este trabajo es el resultado final del Proyecto terminal realizado por Ángeles López-Flores, alumna de la LTSI UAM-Cuajimalpa. La idea principal del trabajo fue desarrollar técnicas automáticas para la identificación precisa de evidencia de acoso en línea, específicamente, identificación de pedofilia.

[2] https://ufal.mff.cuni.cz/pbml/115/art-villatoro-tello-et-al.pdf

[3] http://ceur-ws.org/Vol-2624/germeval-task2-paper4.pdf

[4] https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/germeval-2020-cognitive-motive/ge20st1-paper-2.pdf

[5] http://ceur-ws.org/Vol-2664/mexa3t_paper3.pdf

El trabajo fue realizado en colaboración con Ángeles López-Flores, y Gabriela Ramírez-de-la-Rosa.

2. **Mental lexicon for personality identification in texts.** Al igual que el trabajo anterior, este es también un resumen extendido, el cual fue enviado y aceptado para ser presentado como póster en la conferencia *LatinX in AI Research at NeurIPS 2019* en Diciembre de 2019. La idea principal de este trabajo fue la de mostrar el impacto de utilizar el léxico mental para detectar automáticamente rasgos de personalidad. El trabajo es parte de la tesis de doctorado de Gabriela Ramírez-de-la-Rosa. En la colaboración participó también el Dr. Héctor Jiménez-Salazar.

3. **Predicting consumers engagement on Facebook based on what and how companies write.**[6] Este trabajo fue evaluado, y aceptado para publicación en la ocnferencia LKE 2019. [7] Como resultado de la calidad del trabajo, se nos invitó a enviar una versión para su publicación en la revista "Journal of Intelligent Fuzzy Systems" (Indexed in Journal Citation Reports-JCR, THOMSON REUTERS, Impact Factor 2017: 1.261). Este trabajo es el resultado final del Proyecto Terminal de Érika Rosas-Quezada, alumna de la LTSI de la UAM-Cuajimalpa. Trabajo realizado en colaboración con Érika Rosas-Quezada y la Maestra Gabriela Ramírez-de-la-Rosa.
   Abstract: Engaged customers are a very import part of current social media marketing. Public figures and brands have to be very careful about what they post online. That is why the need for accurate strategies for anticipating the impact of a post written for an online audience is critical to any public brand. Therefore, in this paper, we propose a method to predict the impact of a given post by accounting for the content, style, and behavioral attributes as well as metadata information. For validating our method we collected Facebook posts from 10 public pages, we performed experiments with almost 14000 posts and found that the content and the behavioral attributes from posts provide relevant information to our prediction model.

4. **Author Profiling in Social Media with Multimodal Information.**[8] Este trabajo resume la tesis de doctorado de mi (ex) alumno *Miguel Á. Álvarez-Carmona*, graduado del INAOE, Puebla. El artículo fue aceptado para su publicación en la revista Computación y Sistemas (CyS)[9].
   Abstract: In this paper, we propose a multi-modal approach for extracting information from written messages and images shared by users. Previous work has shown the existence of useful information within these modalities for the task of Author Profiling; however, our proposal goes further demonstrating the complementary of these modalities when merging these two sources of information. To do this, we propose to map images in a text, and with that, to have the same framework of representation through which to achieve the fusion of information. Our work explores different methods for extracting information either from the text or from the images. To represent the textual information, different distributional term representations approach were explored in order to identify the topics

[6] `https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs179897`

[7] `https://lkesymposium.tudublin.ie/index.html`

[8] https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3488

[9] CyS is a peer reviewed open access scientific journal of Computer Science and Engineering (`https://www.cys.cic.ipn.mx/ojs/index.php/CyS/index`)

addressed by the user. For this purpose, an evaluation framework was proposed in order to identify the most appropriate method for this task. The results show that the textual descriptions of the images contain information for the author profiling task, and the fusion of textual information with information extracted from the images increases the accuracy of this task.

En total, durante mi estancia sabática, se tuvieron tres artículos de revista, tres artículos en conferencias, y dos resúmenes extendidos aceptados para presentación. Al mismo tiempo continué en la asesoría de dos proyectos terminales, los cuales se concluyeron en Noviembre del 2019. Todos estos productos son anexados a este documento como elementos probatorios de dichos productos.

3 Probatorios

En el archivo que acompaña este documento, se proporcionan los elementos probatorios de los siguientes productos:

1. Artículo revista: Inferring Highly-dense Representations for Clustering Broadcast Media Content
2. Artículo revista: Predicting consumers engagement on Facebook based on what and how companies write
3. Artículo revista: Author Profiling in Social Media with Multimodal Information
4. Artículo conferencia: Idiap Submission to Swiss-German Language Detection Shared Task
5. Artículo conferencia: Idiap & UAM participation at GermEval 2020: Classification and Regression of Cognitive and Motivational Style from Text.
6. Artículo conferencia: Idiap and UAM Participation at MEX-A3T Evaluation Campaign
7. Resumen extendido: Finding Evidence Of The Sexual Predators Behavior
8. Resumen extendido: Mental lexicon for personality identification in texts
9. Asesoría de Proyecto Terminal: Prediciendo el impacto de una publicación en Facebook
10. Asesoría de Proyecto Terminal: Sistema web de apoyo para la identificación automática de evidencia textual en casos de pedofilia

The Prague Bulletin of Mathematical Linguistics
NUMBER 115 OCTOBER 2020

EDITORIAL BOARD

Published twice a year by Charles University (Prague, Czech Republic)

Editorial office and subscription inquiries:
ÚFAL MFF UK, Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
E-mail: pbml@ufal.mff.cuni.cz

ISSN 0032-6585

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 115 OCTOBER 2020

CONTENTS

Articles

Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources 5
Lukáš Kyjánek, Zdeněk Žabokrtský, Magda Ševčíková, Jonáš Vidra

Inferring Highly-dense Representations for Clustering Broadcast Media Content 31
Esaú Villatoro-Tello, Shantipriya Parida, Petr Motlicek, Ondřej Bojar

Every Layer Counts: Multi-Layer Multi-Head Attention for Neural Machine Translation 51
Isaac Kojo Essel Ampomah, Sally McClean, Lin Zhiwei, Glenn Hawe

The Design of Croderiv 2.0 83
Matea Filko, Krešimir Šojat, Vanja Štefanec

Morphological Networks for Persian and Turkish: What Can Be Induced from Morpheme Segmentation? 105
Hamid Haghdoost, Ebrahim Ansari, Zdeněk Žabokrtský, Mahshid Nikravesh, Mohammad Mahmoudi

Extending Ptakopět for Machine Translation User Interaction Experiments 129
Vilém Zouhar, Michal Novák

Are Multilingual Neural Machine Translation Models Better at Capturing Linguistic Features? 143
David Mareček, Hande Celikkanat, Miikka Silfverberg, Vinit Ravishankar, Jörg Tiedemann

Derivations and Connections: Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin 163
Eleonora Litta, Marco Passarotti, Francesco Mambrini

Focalizers and Discourse Relations 187
Eva Hajičová, Jiří Mírovský, Barbora Štěpánková

Instructions for Authors 198

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 115 OCTOBER 2020 31–50

Inferring Highly-dense Representations for Clustering Broadcast Media Content

Esaú Villatoro-Tello,[ab] Shantipriya Parida,[b] Petr Motlicek,[b] Ondřej Bojar[c]

[a] Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Mexico City, Mexico.
[b] Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland.
[c] Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25 118 00 Praha 1, Czech Republic

Abstract

We propose to employ a low-resolution representation for accurately categorizing spoken documents. Our proposed approach guarantees document clusters using a highly dense representation. Performed experiments, using a dataset from a German TV channel, demonstrate that using low-resolution concepts for representing the broadcast media content allows obtaining a relative improvement of 70.4% in terms of the Silhouette coefficient compared to deep neural architectures.

1. Introduction

Current broadcast platforms utilize the Internet as a cross-promotion source, thus, their produced materials tend to be very short and thematically diverse. Besides, modern Web technologies allow the rapid distribution of these informative content through several platforms. As a result, the broadcast media content monitoring represents a challenging scenario for current Natural Language Understanding (NLU) approaches to efficiently exploit this type of data due to a lack of structuring and reliable information associated with these contents (Morchid and Linarès, 2013; Doulaty et al., 2016; Staykovski et al., 2019). Furthermore, if we consider that documents are very short and that they come from a very narrow domain, the task of clustering becomes harder.

Traditionally, the Bag-of-Words (BoW) has been the most widely used text representation technique for solving many text-related tasks, including document cluster-

 Corresponding author: shantipriya.parida@idiap.ch
Cite as: Esaú Villatoro-Tello, Shantipriya Parida, Petr Motlicek, Ondřej Bojar. Inferring Highly-dense Representations for Clustering Broadcast Media Content. The Prague Bulletin of Mathematical Linguistics No. 115, 2020, pp. 31–50. doi: 10.14712/00326585.004.

ing, due to its simplicity and efficiency (Ribeiro-Neto and Baeza-Yates, 1999). However, the BoW has two major drawbacks: *i*) document representation is generated in a very high-dimensional space, *ii*) it is not feasible to determine the semantic similarity between words. As widely known, previous problems increase when documents are short texts (Li et al., 2016). It becomes more difficult to statistically evaluate the relevance of words given that most of the words have low-frequency occurrences, the BoW representation from short-texts results in a higher sparse vector, and the distance between similar documents is not very different than the distance between more dissimilar documents.

To overcome some of the BoW deficiencies, semantic analysis (SA) techniques attempt to interpret the meaning of the words and text fragments by calculating their relationship with a set of predefined concepts or topics (Li et al., 2011). Examples of SA techniques are LDA (Blei et al., 2003), LSA (Deerwester et al., 1990), and word embeddings (Le and Mikolov, 2014; Bojanowski et al., 2017; Devlin et al., 2019). Accordingly, these strategies learn word or document representations based on the combination of the underlying semantics in a dataset. Similarly, more recent approaches, with the help of word embeddings, learn text representations using deep neural network architectures for document classification (De Boom et al., 2016; Adhikari et al., 2019; Ostendorff et al., 2019; Sheri et al., 2019). However, most of these approaches focus either on solving supervised classification tasks or clustering formal-written short documents.

In this paper, we propose an efficient technological solution for the unsupervised categorization of broadcast media content, i.e., spoken documents. Our proposed approach generates document clusters using a highly dense representation, referred to as low-resolution concepts. We first identify the fundamental semantic elements (i.e., concepts) in the document collection, then, these are used to build the low-resolution representation, which is later used in an unsupervised categorization process. One major advantage of our proposed approach is it's easy to interpret, explicit, and profound representation, allowing the end-users understanding of document vectors and their differences.

The main contributions of this paper are summarized as follows: *i*) To the best of our knowledge, this is the first attempt to explore the feasibility and effectiveness of the low-resolution bag-of-concepts in solving one particular unsupervised task, broadcast media content categorization; *ii*) We conducted our experiments on a real-life dataset of German spoken documents, achieving good performance in terms of three internal evaluation metrics, allowing our method to be considered for practical deployment; *iii*) We evaluate the performance of our proposed method in three well-known datasets (formal written documents).

The remainder of the paper is organized as follows: a brief description of the related work is given in Section 2, in Section 3 we describe the proposed methodology, Section 4 we provide some details regarding the employed dataset. Experimental re-

sults and analyses are presented in Sections 5 and 6. Finally, in Section 7 we draw our main conclusions and future work directions.

2. Related Work

Our work is mainly related to topic modeling or topic discovery. As known, topic discovery aims to use statistical information of word occurrences to obtain the underlying semantics contained in a document set. The most popular textual topic modelling are based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Bayesian methods represented by latent semantic analysis (LSA) (Deerwester et al., 1990), Hierarchical Dirichlet Process (HDP) (Teh et al., 2005).

During recent years, models based on deep neural networks have emerged as a viable alternative for topic discovery. For example, the replicated softmax model (RSM), based on Restricted Boltzmann Machines (Hinton and Salakhutdinov, 2009), which is capable to estimate the probability of observing a new word in a document given previously observed words, thus RSM can learn efficient document representations. More recently, Variational Autoencoders (VAEs) have been successfully adapted for text topic modeling. The Neural Variational Document Model (NVDM) (Miao et al., 2016) for text modeling is an extension of a standard VAE, with an encoder that learns Gaussian distribution and a softmax decoder capable of reconstructing documents in a semantic word embedding space. In (Silveira et al., 2018) authors propose a VAE-based on Gumbel-Softmax (GSDTM) and Logistic-normal Mixture (LMDTM) for text topic modelling. In (Wang et al., 2020) authors propose a neural topic modeling approach, called Bidirectional Adversarial Topic (BAT) model, which builds a two-way projection between the document-topic distribution and the document-word distribution. Although these recent approaches have demonstrated great improvement in text clustering tasks using the topic information, they all have one major disadvantage, they require great amounts of data to infer accurate semantic representations, plus the lack of interpretability.

Despite the extensive exploration of this research field, scarce work has been done to evaluate the impact of these technologies in speech-documents, i.e., textual transcriptions obtained from speech. Contrary to formal documents, textual transcript represents a more challenging scenario as they represent very short documents, containing several speech phenomena such as hesitation, fillers, repetition, etc. Accordingly, in this paper, we evaluate the impact of several clustering strategies for broadcast media categorization. Our proposed approach generates document clusters using highly dense representation, which are easy to interpret by a human judge. The recent relevant work to ours is proposed by (Kim et al., 2017), which proposes a bag-of-concepts approach to generate alternative document representations to overcome the lack of interpretability of word2vec and doc2vec methodologies. However, contrary to this particular work, our method is particularly suited for very short spoken documents (transcripts), and we use highly dense representations, i.e., a very small

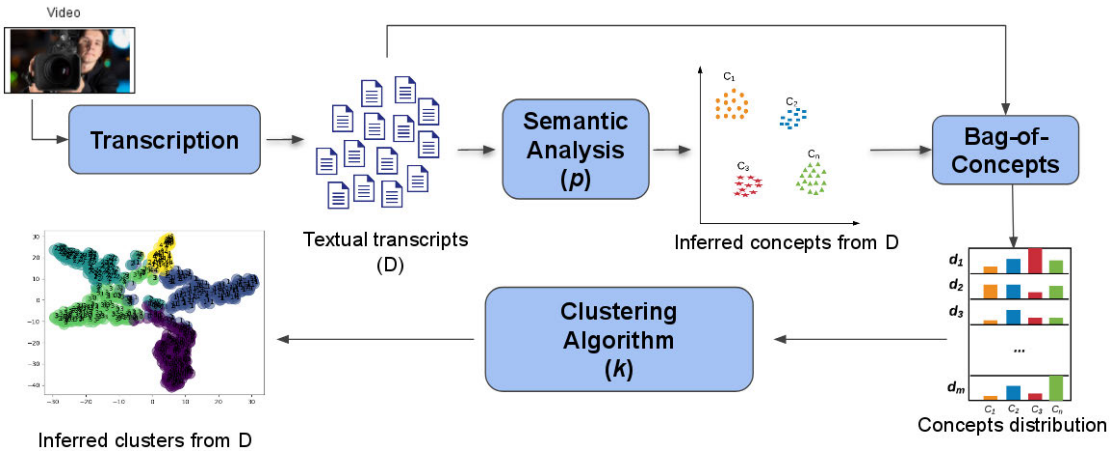


Figure 1: General framework to categorize spoken-documents using low-resolution concepts.

set of features is used to represent the concepts contained in the dataset. We evaluate our proposed method in a real-life dataset extracted to form a German tv channel and we also evaluate our method's performance in three benchmark corpora.

3. Proposed Method

Inspired by the work of (Kim et al., 2017; López-Monroy et al., 2018), we propose using a highly dense representation, denominated low-resolution concepts, for solving the task of clustering short transcript-texts, i.e., broadcast media documents. The intuition behind this approach is that highly abstract semantic elements (concepts) are good discriminators for clustering very short transcript texts that come from a narrow domain. The proposed methodology is depicted in Figure 1. Generally speaking, we first identify the underlying concepts contained in the dataset. For this, we can employ any semantic analysis (SA) approach for learning words representation; thus, learned representation allows us to generate sets of semantically associated words. After obtaining the main concepts, documents are represented by a condensed vector, which counts for the occurrences of the concepts, i.e., a concept distribution vector. Finally, the build texts representation serves as the input to a clustering process, in this case, the K-means algorithm.

More formally, let $\mathcal{D} \quad \{d_1, d_2, \ldots, d_n\}$ denote the set of short transcript texts, and let $\mathcal{V} \quad \{w_1, w_2, \ldots, w_m\}$ represent the vocabulary of the document collection $\mathcal{D}$. As first step, we aim at inferring the underlying set of concepts $\mathcal{C} \quad \{c_1, c_2, \ldots, c_p\}$ contained in $\mathcal{D}$, where every $c_l \in \mathcal{C}$ is a set formed by semantically related words. Notice that in order to obtain the concepts $\mathcal{C}$ we can apply any SA technique for learning the vector representation $\boldsymbol{v}_i$ of each word $w_i \in \mathcal{V}$, for example LDA, LSA, or word

embeddings. Next, for obtaining the document $\mathbf{d}_j$ representation, we account for the occurrence of each c_l within d_i, in other words, the document vector $\mathbf{d}_j$ is a vector that contains concepts distribution. Finally, the generated document-concepts matrix $\boldsymbol{M}_{\mathcal{D}\times\mathcal{C}}$ serves as the input to a clustering process aiming at finding the more suitable documents groups according to the concept-based representation. Henceforth, we will refer to the document-concepts matrix as the Bag-of-Concepts (BoC) representation.

The proposed method has two main parameters, the resolution parameter (p) and the group parameter (k). The former, p, represents the number of concepts that will be generated from the SA step. The lower the number of concepts, the more abstract the resolution. The second parameter, k, indicates the number of categories to be generated from the clustering process. Given the nature of the dataset, i.e., very short texts from a narrow domain, we hypothesize that the clustering algorithm will be able to find groups of documents that share the same amount of information about the same sub-set of concepts, resulting in a more coherent categorization of the documents. Thus, using low-resolution concepts will generate groups of documents referring to the same general topics, while using higher resolution values will result in a more fine-grained topic categorization of the documents.

4. Dataset Description

The dataset used in our paper is from n-tv[1], a German free-to-air television news channel. There are mainly two different sets of files in the proprietary data. One part of the dataset is represented by the speech segments (audio data) with an average duration of 1.5 minutes where each recording has multiple speakers recorded in a relatively noisy environment. The other part of the dataset is the textual transcripts (German) associated with the speech segments. Each of the transcript files represents an article (short text documents), which usually are spread across different topics. See for example a small fragment of an article shown in Table 1. This example, when given to experts, is categorized as 'politics' and as an 'economy' article, which is somehow correct given that both topics are present in the article. This occurs repeatedly across articles due to the interviewed people often mix topics when spontaneously speaking, making the categorization task even more challenging.

For our experiments, the employed dataset comprises a total of 697 articles. Table 2 shows some statistics from the employed dataset; before applying any pre-processing operation and after pre-processing. As pre-processing operations, we removed stopwords, numbers, special symbols, all the words are converted to lower-case. [2]We compute the average number of tokens, vocabulary, and lexical richness (LR) in the dataset. A couple of main observations can be done at this point. On the one hand,

[1] https://www.n-tv.de/

[2] We did not make any special processing for German compounds words.

Original German fragment
Arbeitsminister Hubertus Heil **kämpft** für **befristete Teilzeit**. Also dafür dass man nicht nur von Voll-zur Teilzeit sondern eben auch wieder zurück wechseln kann ...der **Arbeitgeber** darf den Antrag auf Teilzeit auch nicht einfach so ausschlagen außer es gibt betriebliche Gründe... bei **Unternehmen** mit mehr als 200 **Mitarbeitern** habe alle ein Recht auf befristete Teilzeit...zudem kann der Arbeitgeber den Antrag auf befristete Arbeitszeit ablehnen wenn diese ein Jahr unter- oder fünf Jahre überschreitet.
Closest English translation
Minister of Labor Hubertus Heil is **fighting** for **part-time work**. So that you can not only switch from full-time to part-time but also back again ... the **employer** may not simply refuse the application for part-time unless there are operational reasons ... in **companies** with more than 200 **employees**, everyone has a right to temporary part-time work ... the employer can also reject the application for limited working time if it exceeds one year less than or five years.

Table 1: Extracted fragment from the n-tv dataset. Letters in **bold** represent keywords associated with *politic* and *economic* topics.

we notice that individual texts are very short, on average 63.02 tokens with an average vocabulary of 47.86 words, resulting in a very high LR (0.785). This suggests that very few words are repeated within one article, very few redundancies, which represents a challenge for frequency-based methods. On the other hand, globally speaking, the complete dataset has an LR=0.272, which indicates, to some extent, that the information across texts is highly overlapped (narrow domains).

4.1. Benchmark datasets

To validate our proposal, we also evaluate our method in the following three benchmark datasets:

- **AG's news corpus.** We used the as employed in (Zhang et al., 2015). It contains categorized news articles (4 classes) from more than 2000 news sources. In total, this dataset contains 120000 documents in the train partition and 7600 in the test partition.
- **Reuters.** These documents appeared on the Reuters newswire in 1987 and were manually classified by personnel from Reuters Ltd. Particularly, we used for our experiments the R8 partition as provided in (Cardoso-Cachopo, 2007), i.e., 5845 documents for training, and 2189 for testing divided into eight categories.
- **10KGNAD.** This dataset, based on the One Million Posts Corpus (Schabus et al., 2017), is composed of 10273 German news articles collected from an Australian online newspaper. News is categorized into 9 different topics. The train partition contains 9245 documents, while the test partition contains 1028 documents.

	W/O Pre-processing Average (σ)	Total
Tokens	234.68 ($\pm$ 124.45)	163,572
Vocabulary	161.79 ($\pm$ 51.92)	22078
LR	0.717 ($\pm$ 0.073)	0.134
	W/ Pre-processing Average (σ)	Total
Tokens	63.02 ($\pm$ 31.52)	43,928
Vocabulary	47.86 ($\pm$ 16.30)	11,948
LR	0.785 ($\pm$ 0.092)	0.272

Table 2: Statistics of the n-tv dataset.

5. Experimental framework

This section describes the experimental setup. First, we describe the employed methods for learning word representations. Then, we briefly explain the evaluation metrics; and finally, we describe the approaches used for comparison purposes (baselines). For all the performed experiments we ran the k-means algorithm[3] for a range of k 2...15.

5.1. Obtaining word vectors

One crucial step of our approach is learning word representations, i.e., the semantic analysis process shown in Figure 1. For this, an important parameter is the resolution value (p), which indicates the number of concepts that will be employed for building the document-concepts matrix (BoC). Accordingly, we evaluate four different methods for inferring the set $\mathcal{C}$ ($|\mathcal{C}|$ p):

- **FastText:** Concepts are inferred from applying a clustering process over $\mathcal{V}$, using as word representation pre-trained word embeddings. We used word embeddings trained with FastText[4] (Bojanowski et al., 2017) on 2 million German Wikipedia articles. This configuration is referred as: **BoC(FstTxt)**.
- **BERT:** Similar to the previous configuration but, here we use BERT (Devlin et al., 2019), a very recent approach for getting contextualized textual representations. Thus, we feed every word in $\mathcal{V}$ to BERT and preserve the encode produced by

[3] As implemented in the scikit-learn library: https://scikit-learn.org/stable/modules/clustering.html

[4] https://www.spinningbytes.com/resources/wordembeddings/

the last hidden layer (768 units) as the word vector. Performed experiments were done using the pre-trained `bert-base-german-cased` model[5]. We refer to this configuration as **BoC(BERT)**.

- **LDA:** Latent Dirichlet Allocation (Blei et al., 2003) assumes that documents are probability distributions over latent concepts, and concepts are probability distributions over words. Thus, LDA backtracks from the document level to identify concepts that are likely to have generated the dataset. We used the Mallet's LDA implementation from Gensim[6]. After obtaining the concepts, we compute the document-concepts distribution over each d_j for generating the $\mathbf{d}_j$ representation. We refer to this experiment as **BoC(LDA)**.
- **LSA:** Latent Semantic Analysis (Deerwester et al., 1990) is a purely statistical technique that applies singular value decomposition (SVD) to the term-document matrix to identify the 'latent semantic concepts'. We employed the SVD (singular value decomposition) algorithm as implemented in sklearn[7]. Then, document-concepts representation $\mathbf{d}_j$ is obtained similarly to the LDA approach. We refer to this approach as **BoC(LSA)**.

5.2. Comparisons

We compare the proposed methodology against four different approaches:

- **BoW(*tf-idf*):** Short texts are represented using a traditional Bag-of-Words (BoW) considering a *tf-idf* weighing scheme. The top 10,000 most frequent terms are employed for generating the BoW representation. Thus, once we have the document's representation, we applied the traditional k-means algorithm.
  Avg-Emb: Every short text is represented using the average of the word embeddings which are respectively weighted with their *tf-idf* score. This strategy has been considered in previous research as a common baseline (Huang et al., 2012; Lai et al., 2015; Xu et al., 2015). We used the FastText embeddings for this experiment. Similarly to the BoW baseline, once the representation is generated, we applied the k-means algorithm to perform the clustering process.
  BERT: For this, every text is feed through BERT. As the $\mathbf{d}_j$ representation we use the values of the last hidden layer (768 units). We limit the input length to 510 tokens. After generating the BERT encoding of every document, we applied the k-means algorithm.
  CNNs: Contrary to the previous baselines, this is a specific convolutional neural network designed for clustering short texts[8]. The main idea of this method is to

[5] `https://huggingface.co/transformers/pretrained_models.html`

[6] `https://radimrehurek.com/gensim/models/wrappers/ldamallet.html`

[7] `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html`

[8] As implemented in `https://github.com/zqhZY/short_text_cnn_cluster`

learn deep features representations without using any external knowledge (Xu et al., 2015).

5.3. Evaluation metrics

For validating the clustering performance we employed three internal methods (Rendón et al., 2011), namely Silhouette (s) score (Rousseeuw, 1987), Calinski-Harabasz (CH) (Caliński and Harabasz, 1974), and Davies-Bouldin (DB) (Davies and Bouldin, 1979) index. Generally speaking, these metrics propose different strategies for combining the concepts of cohesion and separation for each point in the formed clusters. The cohesion value measures how closely the points in a cluster are related among them, and the separation value indicates how well a cluster is distinguished from other clusters.

Silhouette (s) score (Rousseeuw, 1987): this metric combines the concepts of cohesion and separation for each point in the formed clusters. The cohesion value measures how closely the points in a cluster are related among them, and the separation value indicates how well a cluster is distinguished from other clusters. Thus, the s score for a point i is computed as shown in expression 1.

$$s(i) \quad \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

where $a(i)$ is the cohesion score between point i and the rest of the points belonging to the same cluster; and $b(i)$ is the separation score, which represents the minimum average distance between point i and all the other points in any other cluster, of which i is not a member. At the end, the silhouette score of the clustering process is given by the mean $s(i)$ over all points. For this particular metric possible values range between -1 and 1, where a positive result indicates a better quality in the clustering.

Calinski-Harabasz (CH) index (Caliński and Harabasz, 1974): given a dataset $\mathcal{D}$ of size n, divided into k clusters, the CH index is defined as the ratio of the between-clusters dispersion mean and the within-cluster dispersion. The CH index is computed as shown in expression 2.

$$\text{CH} \quad \frac{SS_B}{SS_W} \times \frac{n-1}{n-k} \tag{2}$$

where SS_W is the overall within-cluster variance, and SS_B is the overall between-cluster variance. The SS_W term represents the sum of the within the sum of squares distances of each point in the cluster from that cluster's centroid, and it will decrease as the number of clusters goes up. On the other hand, the SS_B measures the variance of all the cluster centroids from the dataset's centroid. Hence, a big SS_B value means

that all centroids from all clusters are spread out, and consequently not too close to each other. Therefore, the biggest the CH index, the better the clustering output.

Davies-Bouldin (DB) index (Davies and Bouldin, 1979): this index aims to identify sets of clusters that are compact and well separated. The DB index is defined in expression 3.

$$\text{DB} \quad \frac{1}{k}\sum_{i,j=1}^{k}\max_{i\neq j}\left(\frac{d(i,c_i)+d(j,c_j)}{d(c_i,c_j)}\right) \tag{3}$$

where k denotes the number of formed clusters, i and j are cluster labels, then $d(i,c_i)$ is the average distance between each point of cluster i and the centroid of that cluster c_i, this is also know as cluster diameter. Likewise, $d(c_i,c_j)$ is the distance between centroids of cluster i and j respectively. Thus, the smaller the value of the DB index, the better the clustering solution.

Finally, it is worth mentioning that for the experiments performed in the AG's news, Reuters, and 10KGNAD datasets, we evaluate all the possible configurations and baselines on the test partition. Given that these datasets are labeled, we report the obtained results in terms of accuracy (ACC).

6. Results

First, we determine the impact of the resolution parameter (p) in the clustering task. Then, we compare the proposed method using the best value of p against methods described in section 5.2.

6.1. Impact of the resolution

In Figure 2 and Figure 3 we visually show the performance of the considered concepts-inferring approaches in the clustering task, i.e., BoC(FstTxt), BoC(BERT), Boc(LDA), and BoC(LSA). Each map depicts the performance of the different methods under several resolution values $p \quad 5, 10, 20, 50, 100, 500, 1000$ (y-axis), and several required clusters $k \quad 2, \ldots, 15$ (x-axis). In all cases, the darker the red color in the heat-map the better the performance, conversely, the darker the blue color the worst the performance, and if the cells tend to be white, it means an average performance. Each row in Figure 2 and Figure 3 represents the obtained performance under a different evaluation metric, s score, CH and DB index respectively. As mentioned, the lower the value of the DB index, the better the output of the clustering process. Thus, to provide the generated maps under the third row the same interpretation, we subtract the maximum obtained value under the DB metric to each of the original results.

From these experiments we observe the following: (*1*) Using low-resolution values ($p \quad 5, 10$) allows us to obtain better performance, showing a consistent behavior

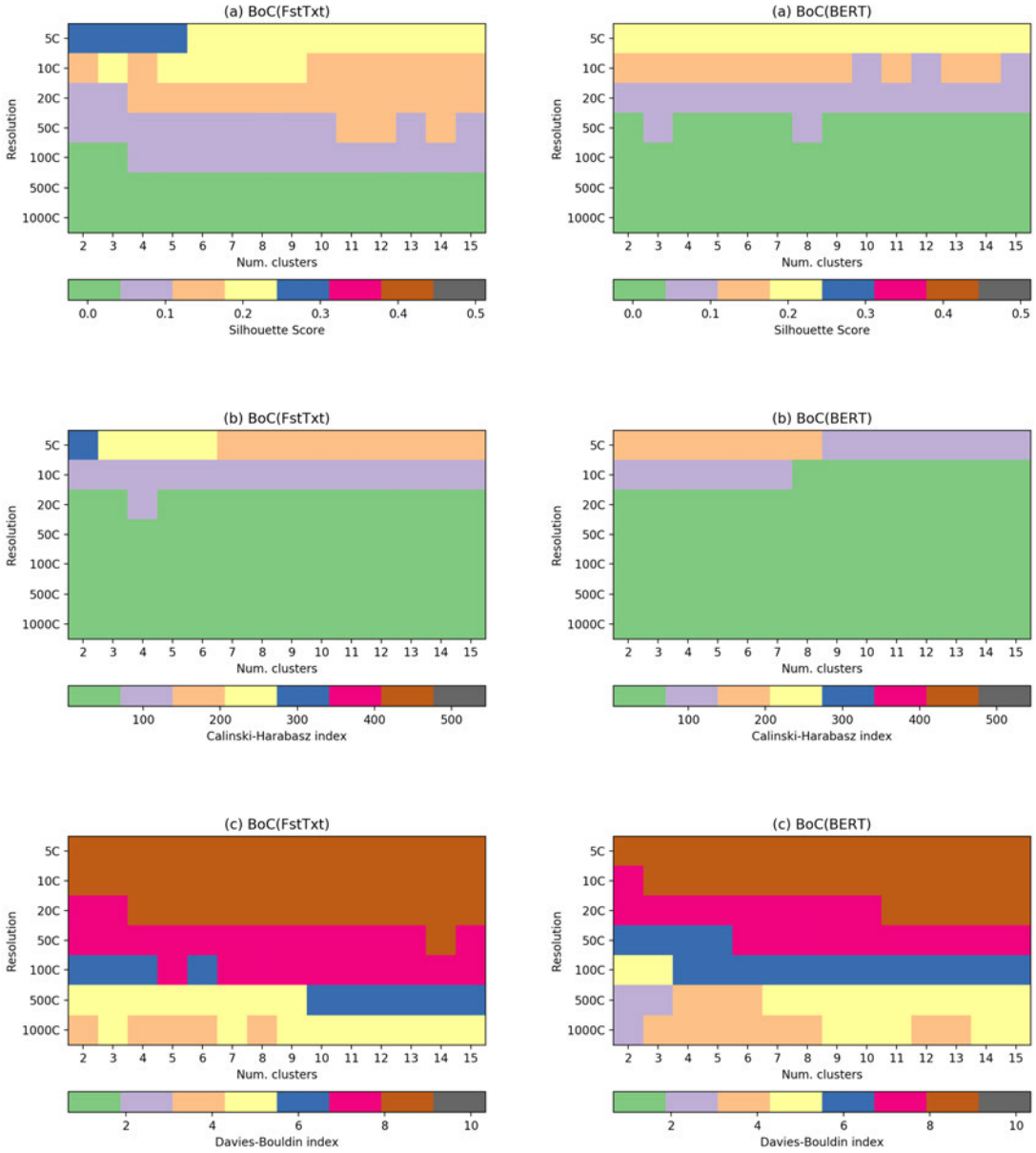


Figure 2: Heatmaps showing the impact of the resolution parameter (p) in the clustering task. First row depicts results in terms of the s score, second row shows the CH index, and third row represents the DB index. Graphs in the same column were generated using the same approach for inferring word representations, specifically, here we are comparing FastText and BERT approaches.

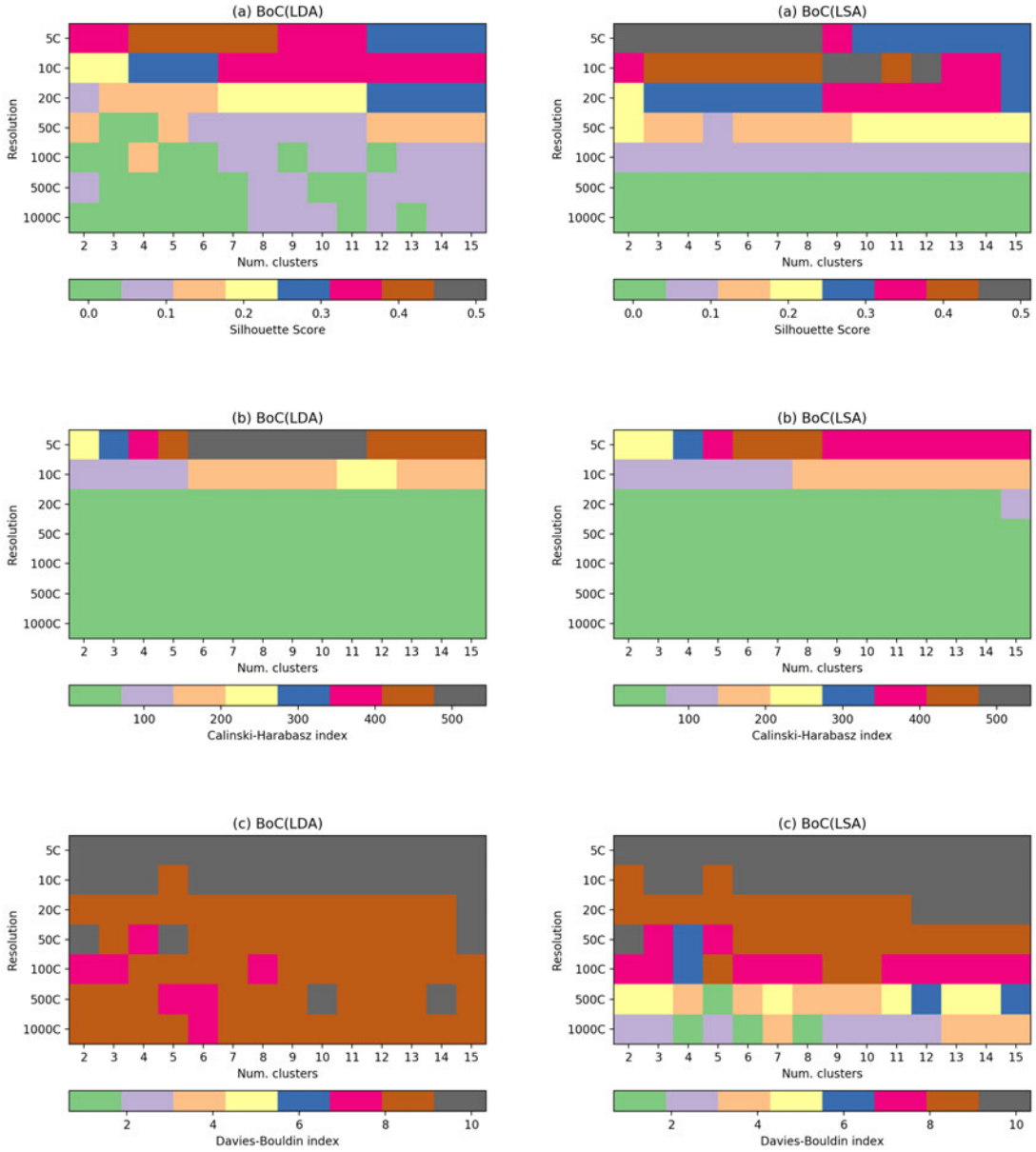


Figure 3: Heatmaps showing the impact of the resolution parameter (p) in the clustering task. First row depicts results in terms of the s score, second row shows the CH index, and third row represents the DB index. Graphs in the same column were generated using the same approach for inferring word representations, specifically, here we are comparing LDA and LSA approaches.

across the three evaluation metrics, although is more clear for the s and CH indexes; (2) inferring word representations with LDA and LSA (Figure 3) allows us to obtain better performance across different values of k. In general, these experiments indicate that low-resolution values (5C, 10C) are preferable for obtaining the best clustering performance in the n-tv dataset.

Additionally, we evaluated our proposed method in three benchmark datasets, namely: Reuters 8 (Cardoso-Cachopo, 2007), AG's News (Zhang et al., 2015), and 10KGNAD (Schabus et al., 2017). Table 3 shows the obtained results in terms of the s score (SH), and clustering accuracy (ACA) values. It is important to mention that although these three datasets are labeled, we cannot compute the traditional Accuracy as in a supervised classification task because the k-means will assign an arbitrary label to every cluster it forms. However, what we can do is to compute the Average Clustering Accuracy (ACA) measure, which gives the accuracy of the clustering no matter what the actual labeling of any cluster is, as long as the members of one cluster are together. Traditionally, for obtaining the ACA value it is necessary to figure out what is the best setting that would yield me the maximum clustering accuracy. For our performed experiments, we used the sklearn linear_assignmen function, which uses the Hungarian algorithm to solve this problem.

As can be observed in Table 3, Boc(LDA) experiments were performed only for 5 and 10 concepts. We do not report results with a higher number of concepts because the LDA approach was not able to obtain more than 10 topics with high probability distributions, in other words, for greater values than 10 the employed LDA implementation generated empty topics for all the three datasets.

The first four rows represent the considered baselines. As can be noticed, the CNN approach performs well in the AGs News and 10KGNAD dataset, while for the R8 dataset, the traditional BoW obtains a competitive performance. In general, we can conclude that using the LDA approach for inferring the underlying semantics represents the best approach for inferring efficient highly-dense concepts. The BoC(LDA-5C) and BoC(LDA-5C) configurations obtain good results in terms of SH and ACA metrics in the R8 and AGs News datasets respectively.

6.2. Overall performance

From the previous analysis, we choose p 5 as the best resolution value, since in two out of the three considered metrics, when the number of concepts is equal 5 we obtain better performances. Therefore, the next set of experiments was done using this as the number of concepts[9] and we compare our proposed approach against baselines described in section 5.2. Figure 4 shows the obtained results across the three considered evaluation metrics. Contrary to the previous section, here we kept the

[9] Represented as the '-5C' suffix in the experiments.

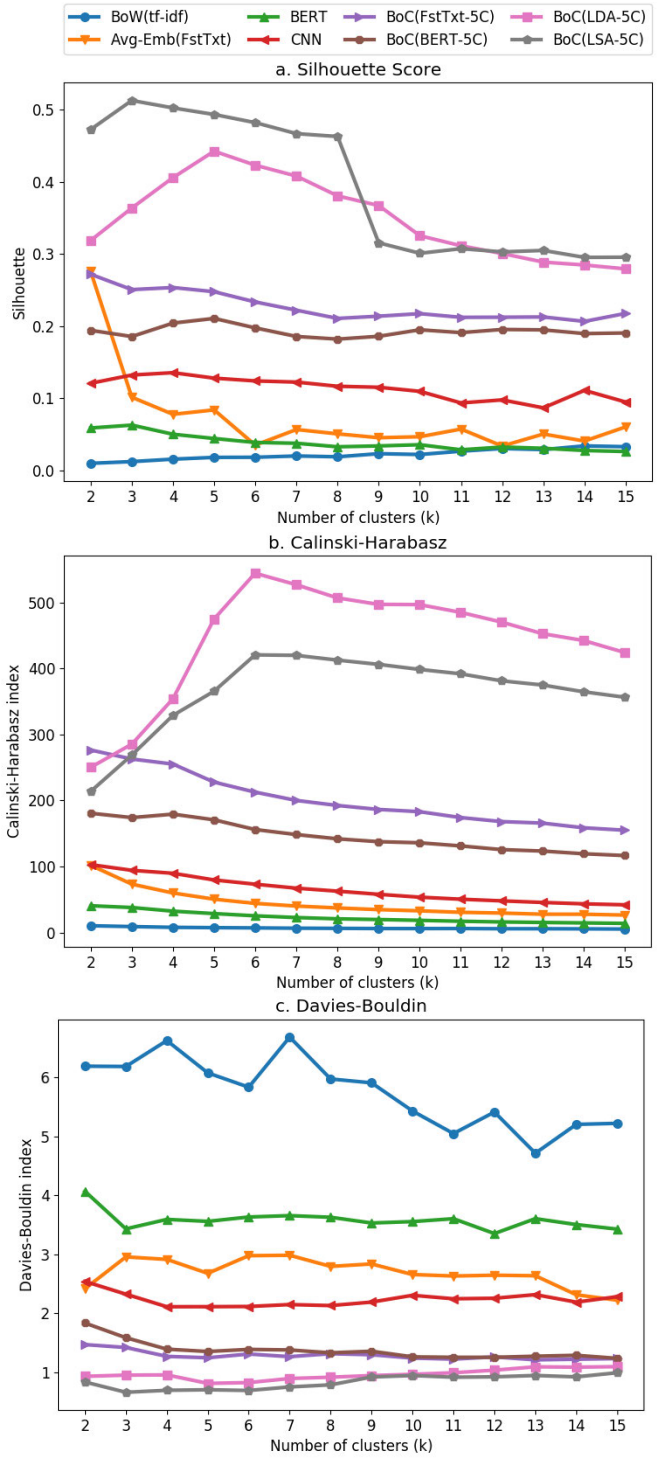


Figure 4: Clustering performance across several values of k: (a) s score, (b) CH index, and (c) DB index

Model	R8		AGs News		10KGNAD	
	SH	ACA	SH	ACA	SH	ACA
BOW	0.055	0.641	0.012	0.271	0.020	0.424
Avg-Emb(FstTxt)	0.054	0.474	0.042	0.409	0.223	0.225
BERT	0.077	0.378	0.041	0.599	0.039	0.368
CNN	0.079	0.407	0.057	0.623	0.158	**0.618**
BoC(FstTxt-5C)	0.279	0.312	0.300	0.361	0.221	0.327
BoC(FstTxt-10C)	0.199	0.325	0.203	0.344	0.231	0.513
BoC(FstTxt-20C)	0.131	0.322	0.158	0.403	0.185	0.503
BoC(FstTxt-50C)	0.098	0.319	0.122	0.579	0.116	0.495
BoC(FstTxt-100C)	0.088	0.364	0.086	0.539	0.073	0.485
BoC(FstTxt-500C)	0.057	0.392	0.043	0.610	0.034	0.527
BoC(FstTxt-1000C)	0.066	0.446	0.030	0.597	0.025	0.499
BoC(LSA-5C)	0.162	0.453	0.236	0.457	0.225	0.476
BoC(LSA-10C)	0.304	0.583	0.183	0.484	0.236	0.471
BoC(LSA-20C)	0.262	0.595	0.095	0.454	0.179	0.421
BoC(LSA-50C)	0.149	0.619	0.158	0.292	0.127	0.427
BoC(LSA-100C)	0.133	0.633	0.057	0.292	0.073	0.448
BoC(LSA-500C)	0.056	0.701	0.050	0.396	0.005	0.418
BoC(LSA-1000C)	0.085	0.592	-0.010	0.459	0.027	0.453
BoC(LDA-5C)	0.349	0.504	**0.424**	**0.793**	0.341	0.455
BoC(LDA-10C)	**0.388**	**0.721**	0.237	0.617	**0.384**	0.495

Table 3: Additional experiments on three benchmark datasets. Results are reported in terms of Silhouette score (SH), and average clustering accuracy (ACA).

original configuration of the DB index, i.e., the lower the obtained score, the better the performance of the clustering approach.

Notice that traditional BoW(tf-idf) and Avg-Emb(FstTxt) techniques obtain the worst performance overall. Similarly, the BERT approach, which represents each document using the produced encoded by the last hidden layer of the pre-trained model of BERT, obtains comparable results to those from the Avg-Emb(FstTxt) technique. Although the CNNs method (Xu et al., 2015) improves the performance of the three previous baselines, its obtained results are far from reaching those obtained with the different configurations of our proposed approach.

From these experiments, it becomes clearer that the proposed approach performs better when concepts are inferred using either LDA or LSA techniques. If we concentrate on the s score only, the best performance is obtained when using BoC(LSA-5C)

at k 3 (s 0.51), which represents a relative improvement of 73% against the best baseline, i.e., the CNN approach. Similarly, if we observe the CH index, the best result is obtained with BoC(LDA-5C) at k 6 (CH 544.19), which represents a relative improvement of 81.1% against the best result of the CNN approach. And finally, in terms of the DB index, the best performance is obtained with BoC(LSA-5C) at k 3 (DB 0.66), which represents a relative improvement of 68% in comparison to the CNN approach. Hence, the main observations from this analysis are: (*1*) proposed approach consistently improves, across three different metrics, traditional clustering techniques as well as some more recent approaches based on deep NN; (*2*) LDA and LSA techniques allow inferring better word representations, improving clustering results in comparison to SOTA methods such as BERT encodings.

6.3. Manual evaluation

To judge the quality of the generated groups, we have taken a subset of 30 articles and performed a small manual annotation experiment using 6 human experts.

For this exercise, we randomly select 30 articles from the n-tv dataset. Every annotator was instructed to identify 5 different clusters, i.e., they had to organize the information into five semantically related groups. The only restriction given is that each group should have at least one document and the same document can not be assigned to more than one cluster. We choose 5 as the number of clusters to identify, as from the previous experiments (see Figure 4) we observed that with k 5 as a middle point, it is possible to obtain good performance on all the considered metrics. We evaluated the annotator's agreement using the Kappa metric (Cohen, 1968). Resulting in a Kappa score of **0.49** which indicates a moderate agreement.

We performed a detailed analysis of the identified groups, and it was clear from the exercise that spotted topics were: 'technology', 'economy', 'politics', 'car industry', and 'financial education'. We observed that annotators tend to disagree on the class of the document when the categories might be related to 'economy', 'politics', and 'financial education', similarly when a document might belong to 'technology' and 'car industry'. However, using a majority vote scheme, we decided on the final class of each document, and we used these 30 documents as a test set. We evaluate our method using the BoC(LDA-5C) configuration, and we were able to obtain a **70%** accuracy in the classification process. In Figure 5 we show the clusters' visualization under this configuration.

7. Conclusions

In this paper, we proposed using highly dense representations, denominated low-resolution concepts, for clustering German broadcast media contents. The proposed approach infers the fundamental semantic elements contained in the input dataset, which are used for suggesting optimal clusters configuration. Performed experiments

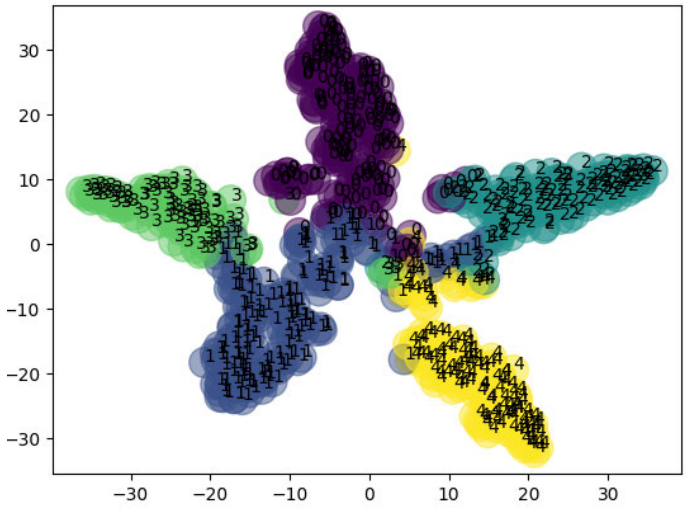


Figure 5: Formed clusters using the BoC(LDA-5C) configuration with k 5. Found topics with the LDA approach are: *i*) chef (boss), autos (cars), deutschland (germany), zukunft (future), diesel (diesel); *ii*) euro (euro), prozent (percent), geld (money), experten (experts), deutschland (germany); *iii*) unternehmen (company), usa (USA), milliarden (billions), trump (Trump), eu (EU); *iv*) kunden (customers), google (Google), mitarbeiter (employees), online (online), facebook (facebook); and *v*) startup (sartup), deutschland (germany), daten (data), idee (idea), welt (world).

demonstrate that using small resolution values provides a better clustering performance, which is consistent across three different internal evaluation metrics, and in four different datasets. Particularly, the proposed framework is not dependent of any particular concise semantic analysis method for inferring concepts; however, when concepts are detected using the LDA and LSA approaches, the clustering performance tends to improve, obtaining relative improvements of 73%, 81%, and 68% under Silhouette, Calinski-Harabasz, and Davies-Bouldin indexes respectively. Finally, we would like to highlight one major advantage of our proposed approach, which is interpretability. As a result of the representation process, produced vectors are easy to interpret, facilitating end users understanding the found semantics and the decisions made by the system.

As future work, we plan to evaluate our proposed approach in similar datasets, i.e., very short texts, from a very narrow domain, and as the result of automatic transcription process from spontaneous speech. Is it possible to imagine, the latter represents a more challenging scenario since automatic transcription systems have many errors that might affect the performance of text-based methods.

Acknowledgments

We are very grateful to the annotators who help us manually validating the topics in the n-tv dataset: Alicia Illi, Noémi Stalder, Luca Schöb, Jasmin Staubli, Chaira Tremml, Angela Mürner. We also like to thank Daniele Zuccheri, and Jan Aeberli for providing access to the n-tv transcripts in the first place.

The work was supported by an innovation project (under an InnoSuisse grant) oriented to improve the automatic speech recognition and natural language understanding technologies for German. Title: "SM2: Extracting Semantic Meaning from Spoken Material" funding application no. 29814.1 IP-ICT and EU H2020 project "Real-time network, text, and speaker analytics for combating organized crime" (ROXANNE), grant agreement: 833635. The first author, Esaú Villatoro-Tello is supported partially by Idiap, SNI-CONACyT, CONACyT project grant CB-2015-01-258588, and UAM-C Mexico during the elaboration of this work.

Bibliography

Adhikari, Ashutosh, Achyudh Ram, Raphael Tang, and Jimmy Lin. DocBERT: BERT for Document Classification. *CoRR*, abs/1904.08398, 2019.

Blei, David M, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051.

Caliński, Tadeusz and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101.

Cardoso-Cachopo, Ana. Improving Methods for Single-label Text Categorization. PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa, 2007.

Cohen, Jacob. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968. doi: 10.1037/h0026256.

Davies, David L and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.

De Boom, Cedric, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156, 2016. doi: 10.1016/j.patrec.2016.06.012.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Doulaty, M, O Saz, RWM Ng, and T Hain. Automatic Genre and Show Identification of Broadcast Media. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2016. doi: 10.21437/Interspeech.2016-472.

Hinton, Geoffrey E and Russ R Salakhutdinov. Replicated softmax: an undirected topic model. In *Advances in neural information processing systems*, pages 1607–1614, 2009.

Huang, Eric H, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proc. ACL*, pages 873–882, 2012.

Kim, Han Kyul, Hyunjoong Kim, and Sungzoon Cho. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266:336–352, 2017. doi: 10.1016/j.neucom.2017.05.046.

Lai, Siwei, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

Le, Quoc and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

Li, Chenliang, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174, 2016. doi: 10.1145/2911451.2911499.

Li, Zhixing, Zhongyang Xiong, Yufang Zhang, Chunyong Liu, and Kuan Li. Fast text categorization using concise semantic analysis. *Pattern Recognition Letters*, 32(3):441–448, 2011. doi: 10.1016/j.patrec.2010.11.001.

López-Monroy, Adrian Pastor, Fabio A González, Manuel Montes, Hugo Jair Escalante, and Thamar Solorio. Early text classification using multi-resolution concept representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1216–1225, 2018. doi: 10.18653/v1/N18-1110.

Miao, Yishu, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736, 2016.

Morchid, Mohamed and Georges Linarès. A LDA-based method for automatic tagging of Youtube videos. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4. IEEE, 2013. doi: 10.1109/WIAMIS.2013.6616126.

Ostendorff, Malte, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, and Bela Gipp. Enriching BERT with Knowledge Graph Embeddings for Document Classification, 2019.

Rendón, Eréndira, Itzel Abundez, Alejandra Arizmendi, and Elvia M Quiroz. Internal versus external cluster validation indexes. *International Journal of computers and communications*, 5 (1):27–34, 2011.

Ribeiro-Neto, Berthier and Ricardo Baeza-Yates. Modern information retrieval. *Addison-Wesley*, 4:107–109, 1999.

Rousseeuw, Peter J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7. URL http://www.sciencedirect.com/science/article/pii/0377042787901257.

Schabus, Dietmar, Marcin Skowron, and Martin Trapp. One Million Posts: A Data Set of German Online Discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1241–1244, Tokyo, Japan, August 2017. doi: 10.1145/3077136.3080711.

Sheri, Ahmad Muqeem, Muhammad Aasim Rafique, Malik Tahir Hassan, Khurum Nazir Junejo, and Moongu Jeon. Boosting Discrimination Information Based Document Clustering Using Consensus and Classification. *IEEE Access*, 7:78954–78962, 2019. doi: 10.1109/ACCESS.2019.2923462.

Silveira, Denys, Andr'e Carvalho, Marco Cristo, and Marie-Francine Moens. Topic modeling using variational auto-encoders with Gumbel-softmax and logistic-normal mixture distributions. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018. doi: 10.1109/IJCNN.2018.8489778.

Staykovski, Todor, Alberto Barrón-Cedeño, Giovanni Da San Martino, and Preslav Nakov. Dense vs. Sparse Representations for News Stream Clustering. In *Text2Story@ ECIR*, pages 47–52, 2019.

Teh, Yee W, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392, 2005.

Wang, Rui, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. Neural Topic Modeling with Bidirectional Adversarial Training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 340–350, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.32. URL https://www.aclweb.org/anthology/2020.acl-main.32.

Xu, Jiaming, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. Short Text Clustering via Convolutional Neural Networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69, 2015. doi: 10.3115/v1/W15-1509.

Zhang, Xiang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

Address for correspondence:
Shantipriya Parida
shantipriya.parida@idiap.ch
Idiap Research Institute
Rue Marconi 19, 1920 Martigny
Switzerland.

Predicting consumers engagement on Facebook based on *what* and *how* companies write

Issue title: Special section: Selected papers of LKE 2019

Guest editors: David Pinto, Vivek Singh and Fernando Perez

Article type: Research Article

Authors: Rosas-Quezada, Érika S.[a] | Ramírez-de-la-Rosa, Gabriela[a] | Villatoro-Tello, Esaú[a; b; *]

Affiliations: [a] Language and Reasoning Research Group, Information Technologies Department, Universidad Autónoma Metropolitana, Cuajimalpa, Mexico | [b] Idiap Research Institute, Rue Marconi 19, Martigny, Switzerland

Correspondence: [*] Corresponding author. Esaú Villatoro-Tello. E-mails: esau.villatoro@idiap.ch and evillatoro@correo.cua.uam.mx.

Abstract: Engaged customers are a very import part of current social media marketing. Public figures and brands have to be very careful about what they post online. That is why the need for accurate strategies for anticipating the impact of a post written for an online audience is critical to any public brand. Therefore, in this paper, we propose a method to predict the impact of a given post by accounting for the content, style, and behavioral attributes as well as metadata information. For validating our method we collected Facebook posts from 10 public pages, we performed experiments with almost 14000 posts and found that the content and the behavioral attributes from posts provide relevant information to our prediction model.

Keywords: Social media branding, impact analysis, data mining, features engineering, natural language processing

DOI: 10.3233/JIFS-179897

Journal: Journal of Intelligent & Fuzzy Systems, vol. 39, no. 2, pp. 2365-2377, 2020

Published: 31 August 2020

Journal of Intelligent & Fuzzy Systems xx (20xx) x–xx
DOI:10.3233/JIFS-179897
IOS Press

Predicting consumers engagement on Facebook based on *what* and *how* companies write

Érika S. Rosas-Quezada[a], Gabriela Ramírez-de-la-Rosa[a] and Esaú Villatoro-Tello[a,b,*]
[a]*Language and Reasoning Research Group, Information Technologies Department Universidad Autónoma Metropolitana, Cuajimalpa, Mexico*
[b]*Idiap Research Institute, Rue Marconi 19, Martigny, Switzerland*

Abstract. Engaged customers are a very import part of current social m dia marketing. Public figures and brands have to be very careful about what they post online. That is why the need for accura e strategies for anticipating the impact of a post written for an online audience is critical to any public brand Ther fore, in this paper, we propose a method to predict the impact of a given post by accounting for the content, style and behavioral attributes as well as metadata information. For validating our method we collected Facebook posts f om 10 public p ges, we performed experiments with almost 14000 posts and found that the content and the behavioral attributes from post provide relevant information to our prediction model.

Keywords: Social media branding, impact analysis, data mining, features engineering, natural language processing

1. Introduction

Nowadays, people world ide are largely engaged and attached to diff rent ypes of Internet technologies and social media platforms. All these technologi s combined have provided new ways for exchangi g feedback on products and services. As stated in [9], this type of circumstances has boosted customer empowerment. Accordingly, customers have the potential of becoming influential with their opinions, recommendations or complaints.

This situation requires the constant incorporation of novel strategies for effectively managing brand's aims and marketing plans, especially aspects related to customers' involvement, relationship, and communication [1]. Thus, measuring the impact of produced advertising is an important issue that needs to be included by brands as part of their social media management strategies [10]. According to previous research, the impact of a published post is measured through several available metrics, mainly related to the consumer's visualizations, reactions, comments, and interactions. Hence, increasing the impact of the published posts will lead to stronger relationships among brand and consumers, allowing customers to create valuable content through social media [14].

Recently, the community of electronic commerce and business research has started to pay attention to how effectively exploit the mechanisms to interact with their customers. Researches have focused on studying phenomena such as the role of social media on advertising, the electronic word of mouth, customer's relationships management, brand's performance, among others [1, 2, 11, 13]. Although many works have proposed techniques for finding

*Corresponding author. Esaú Villatoro-Tello. E-mails: esau.villatoro@idiap.ch and evillatoro@correo.cua.uam.mx.

the relationships between online posts on social media and the impact of such publications measured by users interactions, the vast majority of these research do it as a posteriori analysis [1–4, 9, 14]. This means they focus on finding those characteristics that allowed a post to be appealing for their customers, obtaining valuable insights that enable designing powerful marketing strategies. However, in spite of all the knowledge that these methodologies can provide to specific firms, it is not enough for predicting the impact a post will have prior to its publication. Therefore, a system able to anticipate the impact of individual posts can provide an enormous advantage when deciding to communicate something to the costumers through social media platforms.

In this paper, we propose a novel framework for predicting the impact of publishing posts on a social media network, namely Facebook. Contrastingly to traditional approaches in the field, our method incorporates features that are able to capture content, style, and behavioral features when representing posts. The proposed approach is based on a supervised machine learning strategy, which allows anticipating post's impact, i.e., either high- or low-impact. For validating the proposed method, we took on the task of collecting a dataset from ten renowned brands on Facebo k Mexico. Our performed experiments, over more than 13,000 posts, for six different classification problems, indicate that the combination of the pr pos d features with some metadata-based attr but s, allows an automatic system to obtain acceptab e performance results.

We foresee this work wil repres nt an important contribution to the dev l pment of novel methodologies in the field of el ctronic commerce and business research, as well as m tivate urther research from the intelligent systems and text mining research communities.

The main c ntributions of this paper are as follows:

1. We collected and labeled more than 13,000 posts from ten renowned brands on Facebook Mexico. This dataset represents a valuable resource for future research work on the field of electronic commerce and business, as well as for the intelligent systems community.
2. We provide evidence on the importance of content-based, stylistic, and behavioral features in combination with metadata-based attributes for solving the task of impact prediction on Facebook posts.
3. We proposed a novel framework, based on a supervised machine learning approach, for solving the problem of anticipating the impact of publishing a post on Facebook.

The rest of the paper is organized as follows. The next section provides a review of related work on the problem of social media and customer relationships management. Section 3 describes the followed methodology for collecting the employed dataset, how it was labeled, and provides some tatistics regarding its composition. Section 4 explains he proposed framework based on an supervi ed approach for predicting the impact of publishin posts n Facebook. Section 5 depicts the experimen l setup, and the obtained results for all he per ormed experiments. Finally, in Section 6 we d aw some conclusions and future work dir ctions

2. Related w rk

Consumer engagement is measured by the number of performed activities by users within the social media platform. Normally, these activities vary from platform to platform[1], however, on Facebook, a typi al set of metrics that help to evaluate the level of engagement are: generated reactions (positive, negative, and neutral reactions), number of comments, and the number of times a post is shared [14]. Thus, posts having elevated or low numbers under these metrics, are considered examples of high or low impact posts respectively; meaning a healthy/unhealthy customer engagement relationships. An additionally employed metric is the ROI (return-on-investment) indicator, which is defined as the profit of an investment divided by the cost of the investment [7]. The ROI indicator is one of the most important engagement metrics employed by many companies [8], however, the core of our research is not related to the ROI analytics' field since we are not interested in the direct sales reported by companies. Instead, we aim at developing automatic models that can anticipate the impact of a publication in terms of popularity, i.e., how reached customers will interact with the publication of some post.

Accordingly, literature establishes that the more capable are the organizations building and sustaining emotional and social ties between their customers and their brands, i.e., a healthy level of customer

[1]The research undertaken by [5], describes some of the most relevant metrics over 350 social media marketers.

engagement; the more the benefits that can be obtained. Therefore, many research groups have tackled the problem of how to contribute to both customers experience and customer relationships using social media platforms [1, 2].

On the one hand, the vast majority of the previous work has faced the problem as a knowledge extraction technique for designing powerful marketing strategies. In other words, this type of research proposes analyzing the relationships between several variables and the level of engagement of customers. Thus, it is possible to find what are the main characteristics that provoke customers manifestations (reactions, comments, and sharing). However, a major drawback of these approaches is that they do not consider using this knowledge as part of an automatic method for anticipating the impact of a post. Recent examples of this type of methodologies can be found in [1–4, 9, 14].

On the other hand, a few research works have proposed and evaluated distinct methodologies for implementing predictive systems [10, 12, 15, 18]. In [10] authors proposed using seven features for representing the information contained in a post, namely: category of the post (action, product, or inspirational), the total likes of the brand's page, the type of content (photo, video, or link), time of the publication, month, weekday and hour of the post, and a feature that indicates if the post was paid for advertising. These features were employed fo predicting 12 distinct Facebook metrics. For their experiments, authors employed a SVM regresor, and evaluated their method in 790 posts from a cosmetic company's page. A similar appr ach is described in [15] but for estimating the cess of eBay smartphone sellers. For representing th data authors proposed near 20 me adata-bas d fea ures extracted from the eBay plat orm, such as reachability and engagement (followers) custome feedback (number of positive and negative eviews) and seller information (name, country, etc.). In the work of [12], 164 posts were analyzed from five distinct tourism brands in Spain (dataset is in Spanish). Authors trained a regression model for predicting the number of likes and the number of comments a post will generate. For this, authors proposed as features the post richness (defined as the number of videos, pictures, links are included in the post), time frame (weekday and time of the publication), plus a couple of features associated to the size of the post (in characters) and the number of followers of the brand's page. Similar to the above-described research, a few studies analyze the importance of the so-called contextual features (URLs, mentions, hashtags) to infer the number of replies a tweet may provoke [6, 16]. Finally, in the work described in [18], authors model the relationship between the text of a political blog post and the number of comments that such post will receive. Authors approached the problem both as a regression problem and as a classification task. An interesting aspect of this work is that as features, authors employed a topic based representation (LDA) instead of me adata-based features. Given the nature of their data, they hypothesize that the nature of the topic contai d in the post will influence the number of genera ed comments.

A common characteristic in previous research is the exclusion of text-based feat res (except for [18]). Thus, contrary to previ us res arch, our proposed framework incorp rates three feature categories: stylistic, conten based and behavioral. Our main hypothes s establishes that the content of a post (what it say , as well as the style in how is writte (how it say it), in combination with how the post s designed for interacting with the community (behavioral aspects) are important elements for accura ely predicting the impact of a post. We validate our proposal on a dataset with near 14,000 posts from ten different brands on Facebook Mexico, and compare our results against traditional metadata-based features.

3. Dataset

Given the lack of a standard corpus for evaluating impact prediction systems, we took on the task of collecting and standardizing a large dataset[2] of Facebook posts from different brands that have an important presence in Mexico[3]. Collected corpus represents a valuable resource, in a non-English language, that can be used for training and evaluating automatic systems that aim at predicting several customer's engagement metrics, specifically Facebook's reactions (i.e., Like, Love, Haha, Wow, Sad and Angry), sharing amount, and the number of comments generated by a post. Table 1 summarizes the composition of the dataset.

Under the columns **Num. of Posts**, we report the original (OG) number of collected posts and the resultant number of posts after filtering (FL) the data and

[2]The dataset is available for download in: https://github.com/lyr-uam/CorpusReaccion

[3]Compilation of the data was done from November 2018 to January 2019.

Table 1
Table shows the absolute number of reactions ($|R|$), comments ($|C|$) and shares ($|S|$) in the data set. Additionally, average ($\bar{x}$) and standard deviation (σ) values of these characteristics are shown

Brand's Name	Reactions (R)			Comments (C)			Shares (S)			Num. of Posts	
	$\|R\|$	$\bar{x}$	σ	$\|C\|$	$\bar{x}$	σ	$\|S\|$	$\bar{x}$	σ	(OG)	(FL)
Clash Royale ES	3,464,687	6209.12	11457.63	561,540	1006.34	2341.34	258,904	463.99	1495.29	561	558
Canon Mexicana	2,316,406	2079.36	6626.56	112,280	100.79	265.89	426,502	382.88	1030.55	1157	1114
Muy Interesante México	14,900,267	6872.82	11314.93	258,214	119.10	296.41	4,801,967	2214.93	8436.72	2175	2168
Cinépolis	28,074,276	14404.45	28790.81	2,784,917	1428.90	4179.64	8,165,961	4189.82	1865 .82	1985	1949
Discovery Channel	2,422,143	1446.06	2164.16	44,258	26.42	63.43	392,068	234.07	474 24	1712	1675
National Geographic	2,700,761	1548.60	3680.85	107,884	61.86	247.85	1,034,515	593.19	4643 73	2076	1744
Fisher-Price	3,550,516	4216.76	6053.21	155,811	185.05	400.89	249,981	296.89	709.63	848	842
Xbox México	4,697,179	2839.89	6360.91	479,033	289.62	749.36	513,323	310 35	60 45	1737	1654
Nikon	829,560	673.34	1306.00	36,308	29.47	81.29	145,370	18.00	262.84	1357	1232
Lacoste	1,057,118	1478.49	2559.50	10,005	13.99	36.79	37,037	51.80	266.68	914	715
Total:	*64,012,913*	-	-	*4,550,250*	-	-	*16,025,6 8*		-	*14522*	*13651*

Table 2
This table shows the total number of tokens, vocabulary, and lexical richness of each ra d's posts. Additionally, we show the average number of tokens, and characters for each post; between parenthesis the tandard deviation is indicated

Brand's Name	Total number of:			Average number per post:	
	tokens	*vocabulary*	*le cal ichness*	*tokens* (σ)	*characters* (σ)
Clash Royale ES (CR)	14,531	4,046	0.27	26.04 ($\pm$27.53)	163.21 ($\pm$165.07)
Canon Mexicana (CM)	21,885	5,006	0.22	19.65 ($\pm$12.63)	128.02 ($\pm$81.06)
Muy Interesante México (MI)	42,321	8,916	0.21	19.52 ($\pm$14.74)	117.70 ($\pm$88.79)
Cinépolis (CI)	44,071	7 536	0.17	22.61 ($\pm$10.78)	133.95 ($\pm$64.36)
Discovery Channel (DC)	44,659	10,8 2	0.24	26.66 ($\pm$12.94)	158.09 ($\pm$75.28)
National Geographic (NG)	46,039	6 988	0.15	26.40 ($\pm$13.33)	153.16 ($\pm$73.32)
Fisher-Price (FP)	19,863	4,788	0.24	23.59 ($\pm$78.70)	149.89 ($\pm$517.25)
Xbox México (XM)	27,639	5,283	0.19	16.71 ($\pm$7.21)	112.27 ($\pm$52.98)
Nikon (NK)	28,251	6,044	0.21	22.93 ($\pm$42.99)	147.28 ($\pm$278.2)
Lacoste (LC)	13,455	3,570	0.26	18.82 ($\pm$24.75)	128.06 ($\pm$155.12)

eliminating those posts tha were identified as useless. Particularly, we re oved ll the posts that fulfill any of the following cond tions: *i)* does not contain any tex , *ii)* does not have any reaction, and *iii)* the only reaction contained is 'like'. At the end, a total of 8 1 posts w re removed after applying the previous condition . The first columns report some statistics regarding the number of **Reactions** (R), **Comments** (C), **Shares** (S), for each brand. Values below columns $|R|$, $|C|$, and $|S|$, represent the total number of reactions, comments, and shares, respectively. Values under the columns $\bar{x}$ and σ indicates the average and standard deviation of reactions, comments, and shares for each post. It is worth mentioning that these statistics were computed with the filtered (FL) version of the dataset. In addition, keep in mind that these numbers may vary if the corpus is re-downloaded; since the date of compilation, posts could have generated more manifestations in any of the considered metrics, or perhaps some posts are no longer available.

Observe in Table 1 that the brand with the highest number of reactions, comments, and shares is *Cinépolis*. This brand is a very well known firm in Mexico, devoted to the movie theater business. The second place in the number of reactions and shares is held by *Muy Interesante México*. This is a firm mainly dedicated to science and technology diffusion. It is interesting to notice that even though *Cinépolis* provokes a high number of manifestations from users, is not the brand that produces the most number of posts, which is the case of *Muy Interesante México* with the highest number of posts.

In Table 2 we show some basic statistics regarding the size of the corpus. The first three columns indicate the size of the collected data for each brand in terms of the number of tokens, the size of the vocabulary, and the lexical richness of the posts. Next two columns

Table 3
Number of *high-* and *low-* impact instances for each problem

	$\|R\|$		$\|R+\|$		$\|R-\|$		$\|R\odot\|$		$\|C\|$		$\|S\|$	
	high	*low*	*high*	*low*	*high*	*low*	*high*	*low*	*high*	*low*	*high*	*low*
CR	189	369	165	393	209	349	217	341	264	294	39	519
CM	100	1014	94	1020	43	1071	90	1024	78	1036	96	1018
MI	775	1393	790	1378	136	2032	374	1794	153	2015	824	1344
CI	991	958	966	983	353	1596	729	1190	883	1067	745	1204
DC	109	1566	112	1563	110	1565	85	1590	9	1666	60	1615
NG	124	1620	132	1612	110	1634	53	1691	50	1694	115	1629
FP	248	594	266	576	15	827	31	811	119	723	43	799
XM	230	1424	230	1424	168	1486	155	1499	299	1355	92	1562
NK	24	1208	33	1199	16	1216	6	1226	12	1220	10	222
LC	46	669	57	658	0	715	2	713	2	713	4	7 1

show the average number of tokens, and characters contained in every post of every brand. For the latter two, the standard deviation of these metrics is shown between parenthesis.

From Table 2 we can remark that the brands with the largest number of tokens are *National Geographic* and *Discovery Channel*, both dedicated to promoting a great variety of programs related to ecology, wildlife, science, among others. Having a great number of tokens indicates that, in general, published posts from these brands are larger in terms of words per post. This phenomeno can be observed in the fifth column of Table 2 where it is possible to see the average number of tokens in the published posts. Lexical richness (LR) s a val e that indicates how the terms from the v abul ry are used within a text. Is defined as the r tio between the vocabulary size and the number f tok ns from a text ($LR = |V|/|T|$). Thus, a val e c ose to 1 indicates a higher LR, which means v cabulary terms are used only once, while value ear 0 represent a higher number of tokens u ed mo e frequently (i.e., more repetitive). From our dataset, observe that the brands with the l west LR values are *National Geographic* and *Cinépolis*, which means their produced posts employ a similar ocabulary. We hypothesize that this could be a marketing strategy since, for the case of *Cinépolis*, allows them to reach a high number of consumers manifestations in their posts in spite of being reiterative.

3.1. Labeling methodology

As we mentioned, our goal was to collect a dataset for evaluating the performance of automatic methods for determining the impact of publishing a post on Facebook, in other words, anticipate the consumers' engagement. For this purpose, traditional engagement metrics we e considered [5]: reactions, comments, and sharing.

Therefore and i spired on the work of [17, 18], we define he task of predicting consumer's engagement as he pr cess of classifying whether a post will have hi her (or lower) impact volume than the av rage seen i training data. Even though more fine- rained predictions are possible as well (e.g., predicti g the absolute number of distinct reactions, th number of provoked comments, and the number of times is shared), our goal in this paper was not or ented to propose a methodology based on regression algorithms. Consequently, we define six binary classification problems, namely: *i)* comments ($|C|$), *ii)* sharing ($|S|$), *iii)* total reactions ($|R|$), *iv)* positive reactions ($|R+|$), *v)* negative reactions ($|R-|$) and, *vi)* neutral reactions ($|R\odot|$). Each classification problem has the categories *high-impact* and *low-impact*.

The followed methodology for assigning each post's category, i.e., either *high-* or *low-* impact, consists in the following steps: for each classification problem (i.e., the considered metrics), we compute the average value of metric k among all the posts from the ten brands, this is referred as $\bar{x}_k$. Once we know the value $\bar{x}_k$, for each post contained in brand i, we review the value of metric k in post p_i, thus, if $p_{i,k} > \bar{x}_k$ the category of the post is assigned to *high-impact*, or *low-impact* otherwise. This process represents a very straightforward approach for the problems of *total reactions, comments*, and *sharing*. However, for labeling *positive*, *negative* and *neutral* reactions we acted as follows: we grouped as positive reactions the Like and Love responses, as negative reactions the Sad and Angry responses, and as neutral reactions the Wow and Haha responses. Table 3 shows the number of instances on each category after the labeling process. As expected, the dataset has a

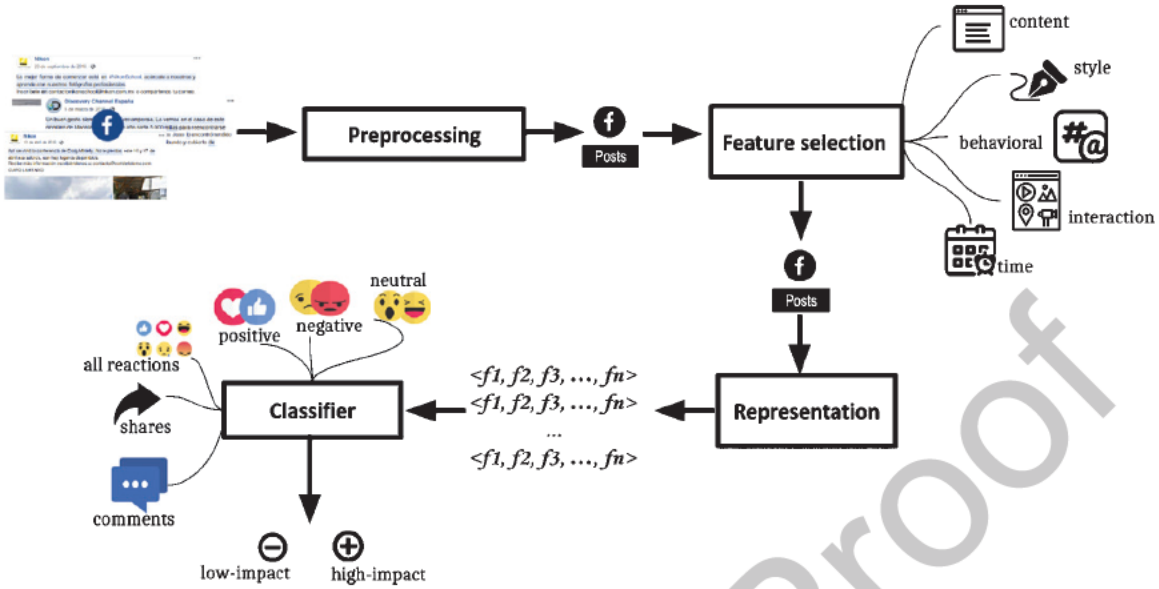


Fig. 1. General framework of our propose meth d.

much greater rate of low-impact volume posts for all the classification problems, potentially making the prediction task a much harder problem.

4. Proposed framework

Our general framework relies on the traditional pipeline of an automatic classification sys em. The classification problem is a learning problem whe e the function $F(x) = y$ needs to be learn d given a series of pairs $< x, y >$ where x is an example of an instance and y is the class of such e ample. Usually, $y \in Y$ and $|Y|$ is the to al number of predefined classes for a given classifica ion problem.

Particularly, our goal is o lea n six functions, one for each metric that re rele ant to know the overall impact of a Facebook's post. Thus, our six classification problems re: impact of total number of comments, impact of number of shares, impact of total reactions, s well as, impact of positive reactions, impact of negative reactions, and impact of neutral reactions. And the predefined classes are *high-impact* and *low-impact* for each of the previously mentioned problems.

The methodology used in all classification problems is showed in Figure 1. Each process in the figure is explained below:

4.1. Preprocessing

First, for each post (p) a preprocessing is performed. The general idea to this step is to standardize th posts conte t to avoid having textual attributes with irrelevant semantic information. For instance, we do n t care about all different URLs included in th posts, we only need to know that a post has an URL.

In this regard we replace all different url, hashtag, emojis and users' mentions to unique tags such as <*url*>, <*hashtag*>, <*emoji*>, and <*mentions*>, respectively.

4.2. Feature selection

The next process in Figure 1 is a feature selection process. As we established, we want to include information about the *what* is been posted as well as *how* these posts are written. Consequently, in this process of our methodology we extracted the following types of features: Content-based that capture *the what*, Style-based and Behavioral to capture *the how*, and we include two more matadata-based features (as these are usually included in the previous works): Interaction and Time.

For the **content type**, we only considered single words as attributes. This feature can give us a general idea of *what* is being saying in a post. The number of features of this type can vary accordingly to the vocabulary of the dataset. The **style type** features account for differences in the writing style of an author of that post. Ideally, a brand's post needs to have a similar writing style that aligns with such brand. In this regard, we included five features: the

post's length (measure in words), the total number of upper-cased and lower-cased words used, the total number of numerals and the total number of symbols (including punctuation marks and other non-alphanumeric symbols). To take into account the level of engagement in the posts, we incorporated the **behavioral type**. For this type of features we considered: total number of emojis, total number of hashtags, total numbers of users' mentions and total number of links.

Additionally, we include two types of metadata attributes: type of links included in the posts (we called this type **interaction**) and the time in which the post is written (**time**). Particularly, for the Interaction type we included five features: number of links to images, number of links to albums, number of links to videos and number of other links. For the Time feature we include the percentages of posts written at the same time if a given post, these percentages are compute independently for: hour, day, month and year.

4.3. Representation

Once we had selected the corresponding feature type, we represent each post in a multidimensional vector, where the number of dimensions corresp nd to the total number of features of a given repr enta tion.

The vectors are normalized to values be ween 0 and 1 to reduce the impact of differences between ranges of different type of featur s.

4.4. Classification model

The fourth phase of the general framework (see Figure 1) i to apply learning algorithm for each classificat on problem. For this stage, we apply four of the wid ly algorithms used for text classification. At the same ime we selected one algorithm of 4 different families: Probabilistic (Naïve Bayes), Decisions Trees (DT), with kernel functions (SVM), and Instance-based (k-NN).

As we have mentioned, to provide an overview of the general impact of a post in the consumer, we generate six different prediction algorithms. At the end, the content manager of a given brand can determined the average impact given the predicted impact of comments, shares, total reactions, as well as, positive reactions, negative reactions and neutral reactions.

In the next section we describe the experiments performed as well as the obtained results.

5. Experiments and results

To test our proposed method, we used the filtered dataset (FL in Table 2) with a total of 13651 Facebook's posts. To evaluate the classification performance we used the F-score metric, and for all experiments we employ a stratified 10 fold cross validation technique to compute the performance. Note that we do not make any distinctions among the posts of particular brands, we aim at building a general classifier instead of having a specific model for ach brand.

One of our research question stabl shes if the combination of the *what* plus the *how* n the process of post's representation can be be ter at predicting the impact of our six metrics than u ing only features that answered the *how*. With this in mind we performed two sets of experim nts. Fir tly, we used as features only singl types of at ibutes for representing post's informati n. Th s type of configuration aims at validating the p tinence of these type of features as has been proposed in many of the previous work. Second, we incorporate the content features to determine the imp ct of considering textual information on the p sed task.

Figure 2 shows obtained results for the first set of experiments. It is important to mention that the size of the representation vector for each of these experiments is very small (between 4 and 5 features). One detail to notice in the Figure 2 is that the best classification algorithm for all problems is Decision Trees, which makes sense given the small number of attributes used as post's representation. Also, we can observe that the style attribute alone, is the second best predictor for each problem. However, the best performance outcome happens when a combination of the four type of attributes is used (b+s+i+t). Among the less useful set of features are the metadata-based ones: interactions and time, where all instances were classify as the majority class (i.e., low-impact); Therefore, from hereafter , these two types of features are not used in the second set of experiments.

So far, Figure 2 shows very consistent results for all problems; but nevertheless, minor aspects are worth mentioning. For instance, the most difficult classification problem is predicting the impact of negative reactions. However, one of the best performances is in predicting the impact of positive reactions. This results can be due to the fact that the positive reactions problem is trained with a slightly less unbalanced dataset in contrast to the negative reactions problem (see Table 3).

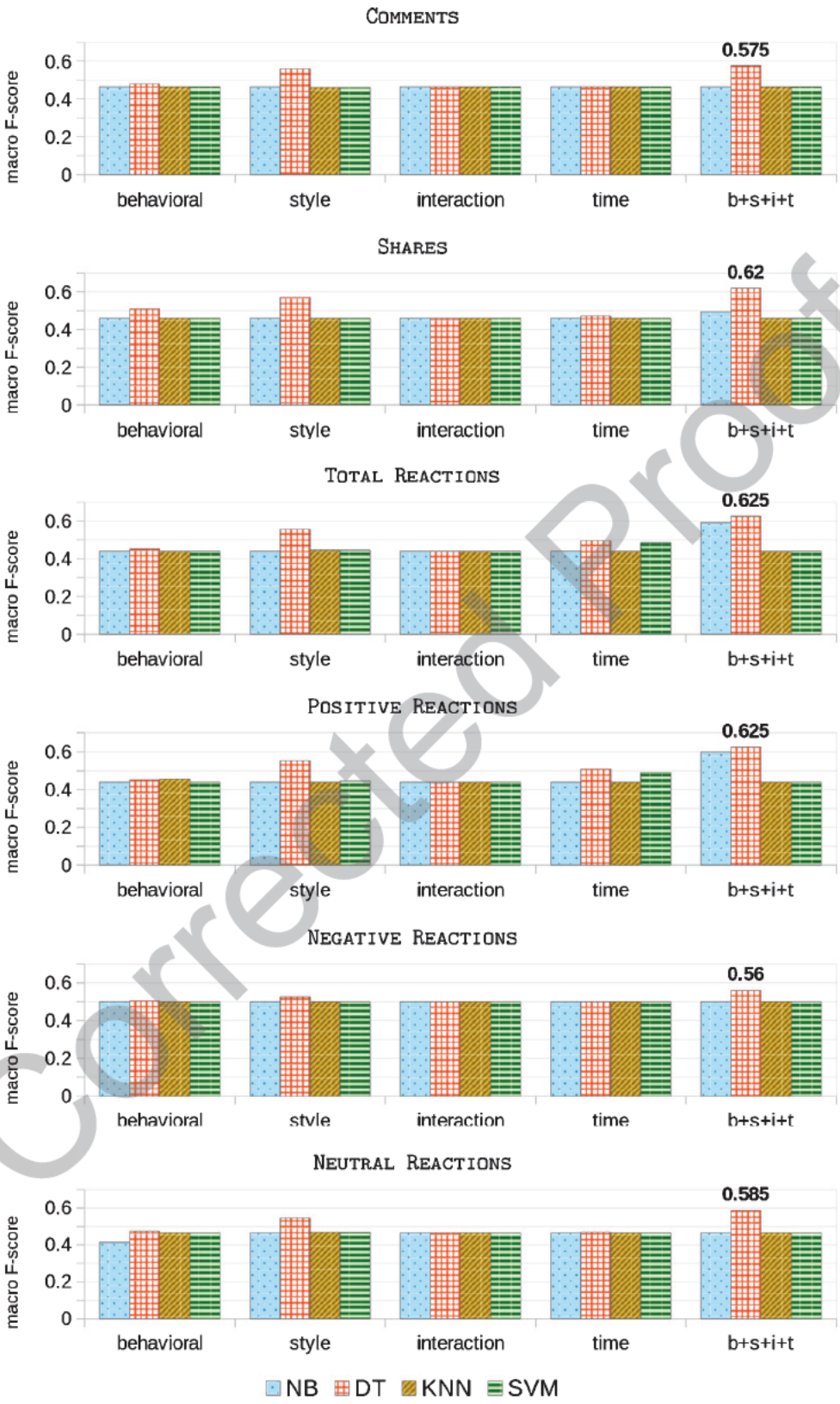


Fig. 2. Performance of our method for predicting the impact of brand's posts using our proposed features types independently. *b+s+i+t* stands for the combination of behavioral, style, interactions and time features in the vector representation.

Figure 3 shows the results of the second set of experiments. For including the text, we used a traditional bag-of-word approach to represent each post. We used only the 10000 tokens more frequent in each problem. The black solid line in each graph indicates the best performance of the previous set of experiments (i.e., from Figure 2). In general, we notice that for all problems using the content feature (alone or in combination with other type of feature) outperformed the best results using only single types of attributes; we compare the best performance from the first set of experiments (black solid line) against the best result obtained in the second set of experiment (c+b+s+i+t in Figure 3) and we found that in five out of six problems, the differences are statistically significant with a $p < 0.0001$; for the sixth problem, *Positive Reactions*, the difference is also statistically significance but for $p = 0.01$ (for this test we use two-tailed t-test). Another aspect to note is that, on one hand, the best learning algorithm for four out of six problems is the probabilistic one. Support Vector Machines, on the other hand, is also the best algorithm predicting the impact of negative and neutral reactions.

As shown in Figure 2, the poorest performance was in the prediction of impact of negative and neutral reactions. That is, those two problems are very difficult to solve. On the contrary, the best overall performances were obtained for predicting the impact of total reactions and positive reacti ns, foll w by predicting the impact of shares and omments.

In all the evaluated classification problems, the best performance was obtained u ing the combination of all our proposed features: th *what* nd the *how* plus the metadata informati On o e hand, the small different in the performance of using only the content feature (the what) wi h the best results, particularly for predic ing the impact of Comments and Shares, gives us some clue o the importance of the content in predicting our s x variables. On the other hand, for predicting reactions (total or positive reactions, particularly) there is a clear improvement in the performance when combining content with behavioral features. This means that using the number of occurrences of hashtag, emojis, users' mentions and links is important to predicting the impact of a post. This type of features were included to give some information regarding the social media lingo used when communicating some information. According to our results, including this type of attributes helps to reach the consumers and induce them to express their feelings towards the brand.

5.1. Qualitative results

Aside from the prediction tasks such as described above, the proposed approach itself can be informative for people in charge of designing marketing strategies. As stated so far, our proposed framework is able to determine the impact of publishing a post on Facebook. Given that part of our goals was to design a generic impact prediction method, i.e., not brand dependent, our approach allows us to envisage characteristics from high and low imp ct posts.

In order to exemplify the typ of information that can be obtained with o r prop sed method, we retrieve eight examples (fo r high-impact, and four low-impact posts) and analyze its characteristics. In Figure 4 we show two high-impact posts (*a* and *b*), and two ow-impact posts (*e* and *f*) from *Cannon Mexicana*. Similarly, we retrieved two high-impact p sts (*c* and *d*) and two low-impact posts (*g* and *h*) from *Nikon's* Facebook page.

Given the n ture of these two brands, we found inte sting to analyze its publications. As it is known, these two firms compete in the field of photography, th y both promote photography courses, professional photography equipment, etc. If we observe Table 1, notice that Nikon publishes a bit more posts than Cannon Mexicana (1,357 *vs.* 1,157). However, Cannon Mexicana has a significantly greater number of reactions, comments, and shares than Nikon; for example, Cannon has more than 2 million reactions while Nikon has barely 829,560. After examining their most representative posts (Figure 4), we notice the following: *i)* High-impact Cannon's posts have a more juvenile way for interacting with their customers, they employ emojis, drawings, as well as less-formal language; *ii)* contrastingly, Nikon uses a more formal style of writing, and their posts refer (mainly) to photography courses, while for Cannon, their posts refer to photographers activities or situations.

With respect to the low-impact posts for both firms, it is interesting that for Cannon, their less popular posts talk about technicalities of the cameras, such as focus points and lens' characteristics. A similar phenomenon occurs for the case of Nikon, where their less popular posts talk about the results of a workshop. Thus, as a preliminary result from this analysis, we could conclude that Nikon needs to produce less formal posts in order to reach a higher level of consumer engagement activities, in other words, change the way they use emojis, hashtags, mentions or links in their published posts.

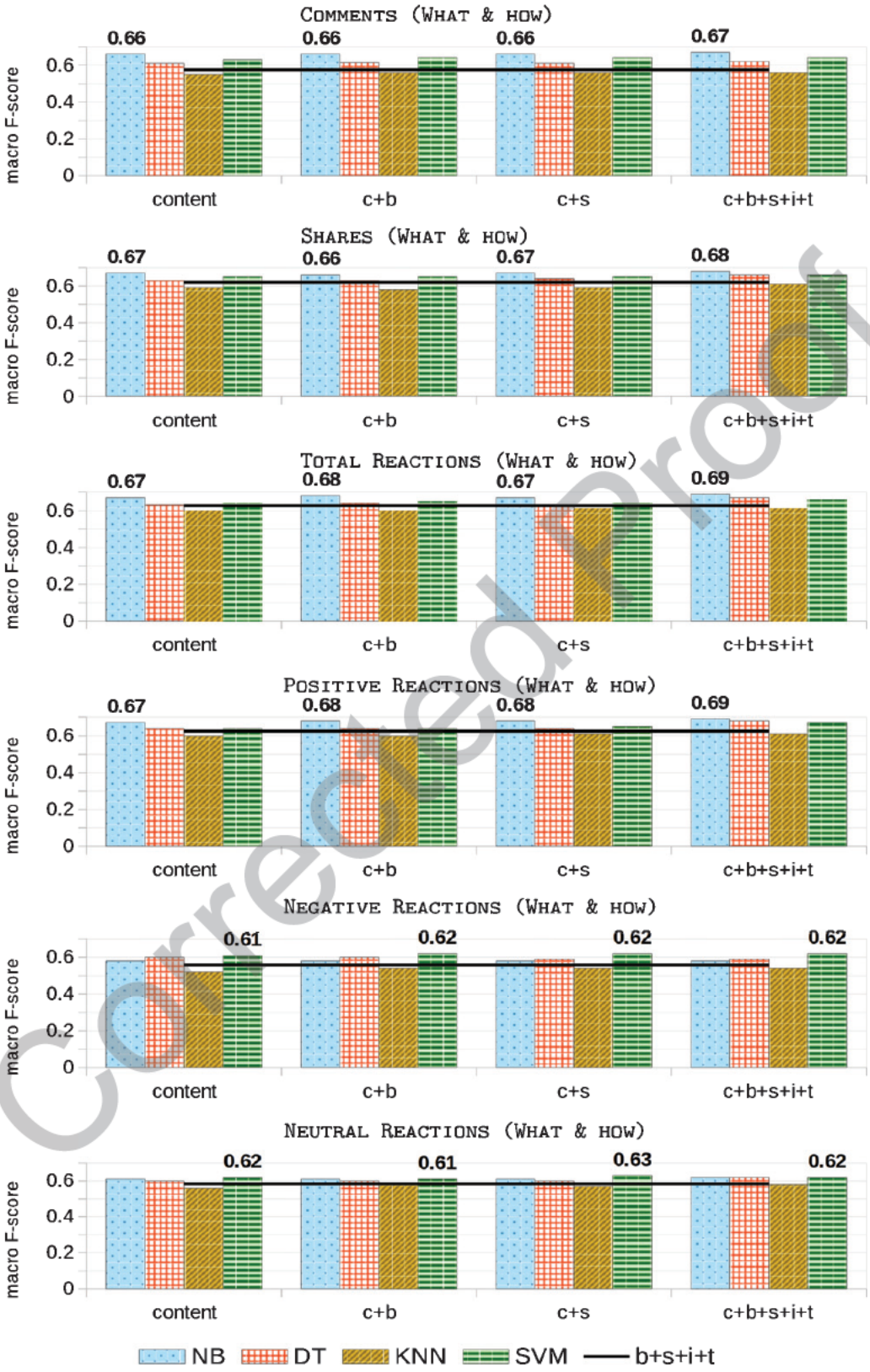
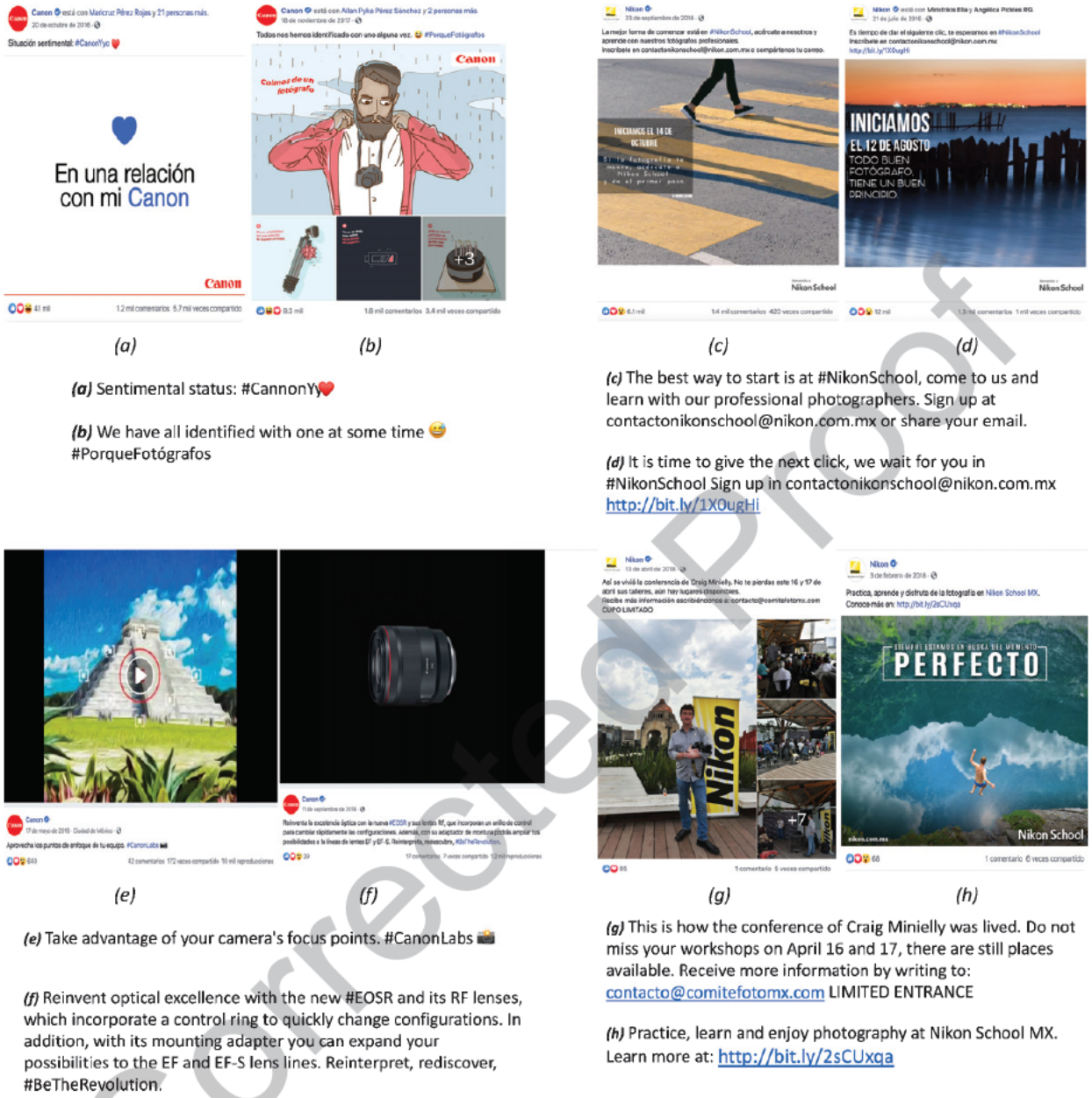


Fig. 3. Performance of our method for predicting impact of brand's posts using our proposed features types in combination with the text (content feature). c, b, s, i and t stands for content feature, behavioral feature, style feature, interaction and time features, respectively.



(a) (b) (c) (d)

(a) Sentimental status: #CannonYy❤️

(b) We have all identified with one at some time 😅 #PorqueFotógrafos

(c) The best way to start is at #NikonSchool, come to us and learn with our professional photographers. Sign up at contactonikonschool@nikon.com.mx or share your email.

(d) It is time to give the next click, we wait for you in #NikonSchool Sign up in contactonikonschool@nikon.com.mx http://bit.ly/1X0ugHi

(e) (f) (g) (h)

(e) Take advantage of your camera's focus points. #CanonLabs 📸

(f) Reinvent optical excellence with the new #EOSR and its RF lenses, which incorporate a control ring to quickly change configurations. In addition, with its mounting adapter you can expand your possibilities to the EF and EF-S lens lines. Reinterpret, rediscover, #BeTheRevolution.

(g) This is how the conference of Craig Minielly was lived. Do not miss your workshops on April 16 and 17, there are still places available. Receive more information by writing to: contacto@comitefotomx.com LIMITED ENTRANCE

(h) Practice, learn and enjoy photography at Nikon School MX. Learn more at: http://bit.ly/2sCUxqa

Fig. 4. Examples of high- mpact (*a, b, c*, and *d*) and low-impact (*e, f, g*, and *h*) published posts extracted from the *Cannon Mexicana* and *Nikon*.

We performed a similar analysis between the high- and low-impact posts from *Discovery Channel* and *National Geographic*. Due to the lack of space, we don't show the obtained most relevant posts. However, for this particular case, we found that for both brands, their less popular publications always refer to TV programming of their respective channels. Regarding their most popular posts, we found that every time these brands publish something related to science and technology diffusion, customers engage positively.

6. Conclusions and future work

This paper focused on proposing a novel framework for anticipating the impact of publishing a post on a company's Facebook page. Our main hypothesis establishes that if an automatic classification algorithm is able to accurately model the what and the how a post should be written, then it will be possible to predict its impact, i.e., its consumer engagement level. Thus, our proposed approach incorporates features that are able to capture

content, style, and behavioral characteristics from posts.

In order to validate our hypothesis, and given the lack of a standard corpus for evaluating this type of approaches, we took in the task of collecting and standardizing a large dataset of Facebook posts from different brands in Mexico. The collected corpus represents a major contribution of this work, and aims at providing resources for future research work in non-English languages. Accordingly, we evaluated our proposed approach in predicting traditional engagement metrics, such as reactions (total reactions, positive, negative, and neutral), comments, and sharing. We performed experiments on our collected dataset, which contains more than 13,000 posts from ten different brands on Facebook Mexico, and compare our results against traditional metadata-based features. Obtained results indicate that *what* and *how* the companies write, in combination with some traditional metadata-based features, allows to obtain the best performance. A qualitative analysis allowed us to observe what are the aspects our proposed model is learning. For instance, we could notice that for some particular brands, competing in the same market, their behavior (i.e., the use of emojis, hashtags, or mentions), in combination with the topics of the pos are very important for improving costumers engagement.

Some relevant advantages of the proposed method are: is a language-independent approach, is not biased towards a specific brand or product type, and ll ws to obtain relevant insights that could be beneficial for community managers providing them some interesting knowledge.

Several ideas arise from this ini ial research for future work. First, th proposed model could be enriched with other s ylistic nd content features. For example, character n- rams are known for providing valuable s ylistic information. Regarding content, we plan to inc rporate some topic-based features, such a LDA or seco d or er representations. Finally, there exist some evidence on the relevance of detecting the post's sentiment as a feature, we plan to evaluate how beneficial could be to incorporate this type of features in our framework.

Acknowledgments

First Author was partially supported by the CONACyT Thematic Networks program (RedTTL Language Technologies Network) with project numbers: 281795 and 295022; and by UAM Cuajimalpa. The third author was partially supported by ADOBE project, from Idiap Research Institute, Switzerland, and by the Information Technologies Department from UAM Cuajimalpa, Mexico. Authors also thank the facilities provided by the Information Technologies Department from UAM Cuajimalpa, Mexico to develop this research. Finally, we thank Orlando Hernández Hernández who was responsible for developing part of the tools that helped in the recollection of the compiled corpus.

References

[1] A.A. Alalwan, N.P Rana Y.K. Dwivedi and R. Algharabat, Social m in m rketing: A review and analysis of the existing literat re, *Te matics and Informatics* **34**(7) (2017), 117 1190.

[2] A. Am do P. Cortez, P. Rita and S. Moro, Research trends on big dat in marketing: A text mining and topic modeling based literat e analysis, *European Research on Management and Business Economics* **24**(1) (2018), 1–7.

3] E. Bonsón, S. Royo and M. Ratkai, Citizens' engagement on l al governments' facebook sites, an empirical analysis: Th impact of different media and content types in western europe, *Government Information Quarterly* **32**(1) (2015), 52–62.

[4] Q. Gao and C. Feng, Branding with social media: User gratifications, usage patterns, and brand message content strategies, *Computers in Human Behavior* **63** (2016), 868–890.

[5] Simply Measured Inc. The state of social marketing 2016 annual report, *Technical Report*, available at: https://www.michigan.org/lib/content/industry/Social_Media_Learning_Library/2016%20State%20of%20Social%20Marketing.pdf, 2016. Accessed October 11, (2019).

[6] M. Jenders, G. Kasneci and F. Naumann, Analyzing and predicting viral tweets. In *Proceedings of the 22Nd International Conference on World Wide Web, WWW'13 Companion*, pages 657–664, New York, NY, USA, (2013). ACM.

[7] M.B.C. Menezes, S. Kim and R. Huang, Return-on-investment (roi) criteria for network design, *European Journal of Operational Research* **245**(1) (2015), 100–108.

[8] E. Michopoulou and D.G. Moisa, Hotel social media metrics: The roi dilemma, *International Journal of Hospitality Management* **76** (2019), 308–315.

[9] S. Moro and P. Rita, Brand strategies in social media in hospitality and tourism, *International Journal of Contemporary Hospitality Management* **30**(1) (2018), 343–364.

[10] S. Moro, P. Rita and B. Vala, Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach, *Journal of Business Research* **69**(9) (2016), 3341–3351.

[11] G. Ramírez-de-la Rosa, E. Villatoro-Tello, H. Jiménez-Salazar and C. Sánchez-Sánchez, Towards automatic detection of user influence in twitter by means of stylistic and behavioral features, In *Human-Inspired Computing and its Applications*, pp. 245–256. Springer International Publishing, (2014).

[12] F. Sabate, J. Berbegal-Mirabent, A. Cañabate and P.R. Lebherz, Factors influencing popularity of branded content in facebook fan pages, *European Management Journal* **32**(6) (2014), 1001–1011.

[13] M. Schreiner, T. Fischer and R. Riedl, Impact of content characteristics and emotion on behavioral engagement in social media: literature review and research agenda, *Electronic Commerce Research* (2019), 1–17.

[14] C.D. Schultz, Proposing to your fans: Which brand post characteristics drive consumer engagement activities on social media brand pages? *Electronic Commerce Research and Applications* **26**(2017), 23–34.

[15] A.T. Silva, S. Moro, P. Rita and P. Cortez, Unveiling the features of successful ebay smartphone sellers, *Journal of Retailing and Consumer Services* **43** (2018), 311–324.

[16] B. Suh, L. Hong, P. Pirolli and E.H. Chi, Want to be retweeted? large scale analytics on factors impacting retweet in twitter network, In *2010 IEEE Second International Conference on Social Computing*, pages 177–184, (2010).

[17] T. Yano, W.W. Cohen and N.A. Smith, Predicting response to political blog posts with topic models, In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 477–485. Association for Computational Linguistics, (2009).

[18] T. Yano and N.A Smith, What's worthy of comment? content and comment volume in political bl gs, In *Fourth International AAAI Conference on Weblogs nd Social Media*, (2010).

Computación y Sistemas

AN INTERNATIONAL JOURNAL OF COMPUTING SCIENCE AND APPLICATIONS

ISSN 1405-5546 (print)
ISSN 2007-9737 (electronic)

Apartado Postal 75-546
C.P. 07738 México, D.F.
Tel (+52)-55-5729-6000
Ext. 56518, 56643
Fax Ext. 56607
computacion-y-sistemas@cic.ipn.mx

http://cys.cic.ipn.mx

Mexico City, June 15, 2020

Dear Miguel A. Álvarez-Carmona, Esaú Villatoro-Tello, Manuel Montes-y-Gómez, and Luis Vilaseñor-Pineda

We are pleased to inform you that after a thorough reviewing process your paper

"Author Profiling in Social Media with Multimodal Information"

has been **accepted for publication** in the journal *Computación y Sistemas*. Results of the reviewing process were sent to you by email.

It is scheduled for publication in Vol. 24, No. 3, 2020.

Sincerely,

Prof. Dr. Grigori Sidorov
Editor-in-Chief

ISSN 2007-9737

Author Profiling in Social Media with Multimodal Information

Miguel Á. Álvarez Carmona[1,2], Esaú Villatoro Tello[4],
Manuel Montes y Gómez[3], Luis Vilaseñor Pineda[3]

[1] Consejo Nacional de Ciencia y Tecnología,
Mexico

[2] Centro de Investigación Científica y de Educación Superior de Ensenada,
Unidad de Transferencia Tecnológica,
Mexico

[3] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE),
Laboratorio de Tecnologías del Lenguaje,
Mexico

[4] Universidad Autónoma Metropolitana Unidad Cuajimalpa,
Departamento de Tecnologías de la Información,
Mexico

malvarez@cicese.mx

Abstract. This paper summarizes the thesis: "Author Profiling in Social Media with Multimodal Information." Our solution uses a multimodal approach to extracting information from written messages and images shared by users. Previous work has shown the existence of useful information for this task in these modalities; however, our proposal goes further, demonstrating the complementarity of the modalities when merging these two sources of information. To do this, we propose to transform images to texts, and with them, to have the same framework of representation for both kinds of information, which allow to achieve their fusion. Our work explores different methods for extracting information either from the text and the images. To represent the extracted information, different distributional term representations approaches were explored in order to identify the topics addressed by the user. For this purpose, an evaluation framework was proposed in order to identify the most appropriate method for this task. The results show that the textual descriptions of the images contain useful information for the author profiling task, and that the fusion of textual information with information extracted from the images increases the accuracy of this task.

Keywords. Author profiling, multimodal information, natural language processing, text classification.

1 Introduction

The author profiling task (AP) is to extract *demographic aspects* of a person from their texts. For example gender, age, location, occupation, socio-economic level or native language [21] [41]. Efforts have also been made to determine other aspects such as the level of well-being [42], personality traits such as extraversion or neuroticism [40] [39] as well as political ideology [19], an affinity for some products [7], among others [13].

In the AP context, it can be seen that most of the recent works, in the field of social networks, have focused mainly on the definition of thematic attributes and style-metrics appropriate for this task.

However, there is a sign of progress towards the description of multimodal representations that, for example, integrate different types of information. Due to the nature of social networks, the images shared by users or their social environment are also incorporated. This thesis work particularly considered the author profiling in social networks with multimodal information [32].

1.1 Problem

Most of the works that have tried to solve the task of AP are based solely on the textual information that users share on social networks. Utilize only text generates that much of the information available by the nature of social networks is not exploited. Most approaches do not take advantage of images, videos, contact lists, activity schedules, or other information. For this reason, it is not known which of these different information modalities is more valuable for the AP task. This is why it is essential to analyze how multimodal information impacts the AP task.

Another aspect to highlight is that the works in AP have given evidence of the importance of the content of the texts. Nevertheless, the most common approach that has been used is the Bag of Words (BoW). The problem with this approach when working on social networks is the lack of information because regular short texts are analyzed. Besides that, the texts are not formal, which causes that there are words out-of-the-dictionary and spelling mistakes.

A set of approaches that have not been deepened enough to represent the content of the texts and, that can be useful for the AP task are the **distributional term representations** (**DTRs**). The basic intuition behind the DTR's is called the distributional hypothesis [44], which states that terms with similar distributional patterns tend to have the same meaning [23, 25]. This distributional hypothesis could capture the content of the users' text in a better way than the traditional content approaches used for AP. In this thesis work, we compared these representations experimentally to know their impact on the AP task.

On the other hand, few works have taken advantage of the information extracted from the images shared by users, this even though various works in psychology have concluded that the photos that are shared on social networks can tell a lot of the people [15, 10, 50, 17]. Some works have applied the color histogram of the images to determine the gender of the users, but no studies have been done for other traits of the authors. Other works have converted the images to texts with automatic labelers of images, through **automatic images annotation** techniques, that assign a list of labels from a previously established set, and from there, infer the user's profile.

These approaches are commonly supervised and with a **closed vocabulary**. This means that the labelers select from a limit list of labels the elements in each image. The problem is that a limited vocabulary could be insufficient to represent the interest of the profiles in a collection. In this thesis work, we proposed to apply an approach based on **open vocabulary** to the AP task, under the idea that it describes in a better way the social media profiles. The automatic annotation of images based on open vocabulary approaches does not select the labels set from a limit list, but they select the vocabulary from an extensive collection, usually extracted from Internet pages. With this idea, we could represent each image in the collection as a text, and we were able to apply text processing approaches to classify the profile of each user.

1.2 Research Questions

Throughout this thesis, we intend to answer the following research questions:

1. What kind of information could be captured by distributional-based methods, and how effective are they for representing user's information when facing the problem of author profiling?
2. How to extract information from the images shared by users through an open vocabulary approach, and how to use them to determine their profile?
3. How to jointly take advantage of the information obtained from texts and images for solving the author profiling task?

1.3 Contributions

The main contributions derived from this work are:

1. A novel corpus including information about Mexican twitter accounts with text and image information. Also, the extension of the well knows PAN@14 corpus. This collection had only text information. For this study, we include the image information for this collection.
2. A comparison among different distributional based methods for the AP task. For this study, we apply DOR, TCOR, word2vec, and SSR.
3. A multimodal method for the AP task taking advantage of textual and image information.
4. The evidence that it is possible to classify profiles from different countries and language trough the different images shared on the networks.

In the following sections, we describe each of the main contributions of this work.

2 Corpora

We presented two new corpora that have been designed for the Author Profiling task evaluation with text and image information.

First, we presented an extension of the well known PAN 14 Twitter corpus [38], aiming to use a well-known corpus enriching it with image information.

Also, the thesis presented a Mexican Twitter corpus for the AP task. The specific application of this corpus is in the analysis of several traits of Mexican Twitter users by text and image information. The data contains for each account the activity schedule on Twitter, its tweets, and its images. This corpus is labeled for gender, the place where he/she lives, and occupation. The annotation of the data was been accomplished manually.

The rest of this chapter is organized as follows. Section 2.1 describes the PAN 14 corpus for the text experiments. Section 2.2 shows the description of the images extension for the PAN 14 Twitter corpus.

Table 1. Distribution of the gender and age classes across the different social media domains

Classes	Genres			
	Blogs	*Reviews*	*Social-media*	*Twitter*
Female	73	2080	3873	153
Male	74	2080	3873	153
Total:	147	4160	7746	306
18-24	6	360	1550	20
25-34	60	1000	2098	88
35-49	54	1000	2246	130
50-64	23	1000	1838	60
65+	4	800	14	8
Total:	147	4160	7746	306

Finally, Section 2.3 describe the new Mexican Twitter corpus for the author profiling task.

2.1 Pan 14 Corpus

For our experiments, we employed the English dataset from the PAN 14 AP track. This corpus was specially built for studying AP in social media. It is labeled by gender (i.e., female and male), and five non-overlapping age categories (18-24, 25-34, 35-49, 50-64, 65+). Although all documents are from social media domains, four distinct genres were provided: blogs, social media, hotel reviews, and Twitter posts. A more detailed description of how these datasets were collected can be found in [38]. Table 1 provides some basic statistics regarding the distribution of profiles across the different domains (i.e., genres). It can be noticed that gender classes are balanced, whereas, for the age classification task, the classes are highly unbalanced. Notably, there are very few instances for the 65+ category.

2.2 Extended PAN 14 Corpus

Images shared by social media users tend to be strongly correlated with their thematic interests as well as to their style preferences. Motivated by these facts, we tackled the task of assembling a corpus considering text and images from Twitter users. Mainly, we extended the PAN-2014 [38] dataset by obtaining images from the already existing Twitter users.

Table 2. Statistics of images shared by each age category

Ages	Profiles	Average images (α)	Average tweets (α)
18-24	17	246.45 ($\pm$80.34)	706.18($\pm$361.76)
25-34	78	286.42 ($\pm$202.65)	796.01($\pm$291.18)
35-49	123	301.74 ($\pm$253.83)	640.41($\pm$362.28)
50-64	54	334.19 ($\pm$238.24)	527.68($\pm$354.24)
65+	7	441.65 ($\pm$102.52)	651.85($\pm$432.28)

Table 3. Statistics of images shared by each gender category

Ages	Profiles	Average images (α)	Average tweets (α)
Female	140	162.21 ($\pm$294.13)	543.53($\pm$395.93)
Male	139	141.76 ($\pm$274.98)	784.88($\pm$265.86)

The PAN-2014 dataset includes tweets (only textual information) from English users. Based on this dataset, we obtained more than 42,000 images, corresponding to a subset of 279 profiles in English[1]. The images associated with all of the users were downloaded to existing user profiles, resulting in a new multimodal Twitter corpus for the AP task. Each profile has an average of 304 images.

Tables 2 and 3 present additional statistics on the values that both variables, gender and age can take, respectively. On the one hand, Table 2 divides profiles by age ranges, i.e., 18-24, 25-34, 35-49, 50-64 and 65+. It shows a great level of imbalance, being the 35-49 class, the one having the greatest number of users.

Nonetheless, the users from the 65+ range are the ones with the greatest number of posted images as well as the lower standard deviation values. It is also important to notice that the users belonging to the 50-64 range share in average a lot of images, but show a large standard deviation, indicating the presence of some users with too many and very few images.

On the other hand, Table 3 reports some statistics for each gender profile. It is observed a balanced number of male and females users in both corpora as well as a similar number of shared images.

[1]Note that the PAN-2014 corpus includes more profiles; however, for some Twitter users, it was impossible to download their associated images.

Fig. 1. Regional division for Mexico. Source: `http://www.conafor.gob.mx/`

2.3 Mex-A3T-500 Corpus

To study the characteristics of the different Mexican Twitter profiles, we built a Mexican corpus for author profiling named Mex-A3T-500[2]. Each of the Twitter users was labeled with gender, occupation, and place of residence information. For the occupation label, we considered the following eight classes: arts, student, social, sciences, sports, administrative, health, and others. For the place of residence trait, we considered the following six classes: north (norte), northwest (noroeste), northeast (noreste), center (centro), west (occidente), and southeast (sureste). Figure 1 shows the division in Mexico's map.

2.3.1 Construction of the Corpus

Two human annotators, working three months each, were needed for building this corpus. They applied the following methodology: (i) to find a set of Twitter accounts corresponding to famous persons and/or organizations from each region of interest. These accounts usually were from local civil authorities, known restaurants, and universities; (ii) to search for followers of the initial accounts, assuming that most of them belong to the same region with the initial accounts; (iii) to select only those followers that explicitly mention, in Twitter or another social network (as

[2]This is a subset of the corpus used for the MEX-A3T forum for the 2018 and 2019 editions [2, 6]. https://sites.google.com/view/mex-a3t/.

Table 4. Example of tweets mentioning information related to the place of residence and/or occupation of users

Trait detected	Original text	Translation
Residence	La pura carnita asada en **Monterrey**	Roast beef in **Monterrey**
Residence	Nunca me canso de pasear en el zócalo de **Puebla**	I never get tired of walking in the **Puebla** Zocalo
Occupation	Porque los **arquitectos** nunca descansamos	Because we, the **architects** never rest
Occupation	**Programando** en el trabajo ando	**Programming** at work

Table 5. Mexican author profiling corpus: distribution of the gender trait

Class	Profiles	Average images (α)	Average tweets (α)
Female	250	715.46 ($\pm$722.89)	1225.00($\pm$868.17)
Male	250	480.90($\pm$459.36)	1500.01($\pm$946.66)

Facebook and Instagram) their place of residence and occupation. Table 4 shows some examples of tweets where users reveal information from their place of residence and occupation.

2.3.2 Statistics

The corpus consists of 500 profiles from Mexican Twitter users. Each profile is labeled with information about the gender, occupation, and place of residence of the user. Tables 5, 6 and 7 present additional statistics on the distribution of user accounts on gender, occupation and location.

Table 6 divides profiles into the different Mexican regions on the corpus, i.e., north, northeast, northwest, center, west, and southeast. Also, it shows an important level of imbalance, being the center class, the one having the greatest number of users, while the north is the class with the lowest.

On the other hand, Table 7 divides profiles on the eight different occupations on the corpus. It is possible to see that the majority class is the central region, whereas the classes with the least instances are the others and sports.

Table 6. Mexican author profiling corpus: distribution of the place of residence trait

Class	Profiles	Average images (α)	Average tweets (α)
North	13	625.23($\pm$442.49)	1594.23($\pm$855.17)
Northwest	80	385.92($\pm$345.95)	1162.17($\pm$866.14)
Northeast	123	460.54($\pm$482.02)	1071.60($\pm$800.66)
Center	191	755.58($\pm$732.74)	1597.83($\pm$922.49)
West	46	611.91($\pm$488.10)	1525.80($\pm$990.62)
Southeast	47	659.12($\pm$732.35)	1284.51($\pm$916.36)

Table 7. Mexican author profiling corpus: distribution of the occupation trait

Class	Profiles	Average images (α)	Average tweets(α)
Arts	38	826.21($\pm$754.71)	1828.23($\pm$834.09)
Student	253	336.57($\pm$259.81)	1184.66($\pm$838.81)
Social	64	1158.15($\pm$867.03)	1362.62($\pm$921.89)
Sciences	25	474.28($\pm$461.97)	1549.64($\pm$947.44)
Sports	12	682.41($\pm$652.27)	1113.00($\pm$892.95)
Administrative	82	894.59($\pm$651.72)	1597.52($\pm$965.65)
Health	15	248.20($\pm$275.05)	1410.20($\pm$1127.04)
Others	11	1026.90($\pm$747.28)	1873.27($\pm$965.63)

3 Analysis of Distributional Term Representations

This section describes a general framework for Author Profiling using distributional term representations (DTRs). Our goal is to overcome, to some extent, the issues naturally inherited by the BoW representation and build instead of a more semantically related representation. Intuitively, DTRs can capture the semantics of a term t_i by exploiting the distributional hypothesis: "words with similar meanings appear in similar contexts". Thus, different DTRs can capture the semantics through the context in different ways and at different levels.

Traditionally, the Author Profiling task has been approached as a single-labeled classification problem, where the different categories (e.g., *male* vs. *female*, or *teenager* vs. *young* vs.

old) stand for the target classes. The common pipeline is as follows: *i*) extracting textual features from the documents; *ii*) building the documents' representation using the extracted features, and *iii*) learning a classification model from the built representations [5].

As it is possible to imagine, extracting the relevant features is a key aspect for learning the textual patterns of the different profiles. Accordingly, previous research has evaluated the importance of thematic (content-based) features [20, 37] and stylistic characteristics [8].

More recently, some works have also considered learning such representations utilizing Convolutional and Recurrent Neural Networks [43, 18, 45].

Although many textual features have been used and proposed, a common conclusion among previous research is that content-based features are the most relevant for this task. The latter can be confirmed by reviewing the results from the PAN[3] competitions [39], where the best-performing systems employed content-based features for representing documents regardless of their genre. This result is somehow intuitive since AP is not focused on distinguishing a particular author through modeling his/her writing style, but on characterizing a group of authors.

The idea is to enrich representations that help to overcome the small-length and high-sparsity issues of social media documents by considering contextual information computed from document occurrence and term co-occurrence statistics. Mainly, we proposed a family of distributional representations based on second-order attributes that allow capturing the relationships between terms and profiles and sub-profiles [29].

These representations obtained the best results in the AP tasks at PAN 2013 and PAN 2014 [28]. Also, we evaluated topic-based representations such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) in the AP task [3], obtaining the best performance at the PAN 2015 as well as showing its superiority against a representation based on manually defined topics utilizing LIWC [4].

[3] A set of shared tasks on digital text forensics: http://pan.webis.de/

In this section, we present a thorough analysis of the pertinence of *distributional term representations* (DTRs) for solving the problem of AP in social media. We aim to highlight the advantages and disadvantages of this type of representation in comparison with traditional topic-based representations such as LSA and LDA.

In summary, the main contributions of this section are:

— We introduce a framework for supervised author profiling in social media domains using DTRs. This framework encompasses the extraction of distributional representation terms as well as the construction of the authors' representation by aggregating the representations of the terms from their documents.

— We evaluate for the first time the document-occurrence representation (DOR) and the term co-occurrence representation (TCOR) in the AP task. These are two simple and well-known term representations from distributional semantics [24].

— We present a comparative analysis of several distributional representations, namely DOR, TCOR, SSR, and word2vec, using the proposed framework for AP. Additionally, we compare their performance against the results from classic bag-of-words and topic-based representations.

3.1 Distributional Term Representations

Let us consider words in the vocabulary as the base terms for building the DTR. More formally, let $\mathcal{D} = \{(d_1, y_1), \ldots, (d_n, y_n)\}$ be a training set of $n-$pairs of documents (d_j) and labels/categories $y_i \in \mathcal{C} = \{C_1, \ldots, C_q\}$. Also let $\mathcal{V} = \{t_1, \ldots, t_m\}$ be the collection vocabulary. In this context, DTRs associates each term $t_i \in \mathcal{V}$ with a term vector $\vec{w_i} \in R^r$, i.e., $\vec{w_i} = \langle w_{i,1}, \ldots, w_{i,r} \rangle$. In this notation $w_{i,j}$ indicates the contribution of distributional feature j to the representation of term t_i. This contribution is particular of each DTR and can be computed in a number of ways.

In the following sections we describe in detail each of the DTRs that we selected for this study. The second step consists in building the document representations by using the term vectors. Formally, the representation of document a d_j, the vector $\vec{d_j}$, is obtained by using the expression 1, where the scalar α_i weighs the relevance of term t_i in the document d_j. Although there are several ways to define this weighting, the most widely used approach is the average of the distribution (i.e., α_i is proportional to the number of terms in the document):

$$\vec{d_j} = \sum_{t_i \in d_j} \alpha_i \cdot \mathbf{w_i}. \quad (1)$$

Different ways to define vectors w_i are briefly explained below. For more details of the formal implementation, consult [23, 47, 1, 28]

3.1.1 Document Occurrence Representation

The document occurrence representation (DOR) can be considered the dual of the TF-IDF representation widely used in the Information Retrieval field [23]. DOR is based on the hypothesis that the semantics of a term can be revealed by its distribution of occurrence-statistics over the documents in the corpus. A term t_i that belongs to the vocabulary $\mathcal{V}$ is represented by a vector of weights associated to documents $\vec{w_i} = \langle w_{i,1}, \cdots, w_{i,N} \rangle$ where N is the number of documents in the collection and $0 \leq w_{i,j} \leq 1$ represents the contribution of document d_j.

3.1.2 Term Co-Occurrence Representation

Term Co-Occurrence Representation (TCOR) is based on co-occurrence statistics [23]. The underlying idea is that the semantics of a term t_i can be revealed by the terms that co-occur with it across the documents collection. Here, each term $t_i \in \mathcal{V}$ is represented by a vector of weights $\vec{w_i} = \langle w_{i,1}, \cdots, w_{i,|\mathcal{V}|} \rangle$ where $0 \leq w_{i,j} \leq 1$ represents the contribution of term t_j to the semantic description of t_i.

3.1.3 Word Embeddings: Word2vec

Recently, a prevalent group of related models for producing word embeddings is word2vec [35]. These models are shallow, two-layer neural networks trained to reconstruct the linguistic contexts of words.

Word2vec takes as its input a large corpus of texts and produces a vector space, typically of a few hundreds of dimensions, where each term in the corpus is assigned to a corresponding vector $\vec{w_i}$ in the space. Thus, once the word vectors have been computed and positioned in the vector space, words that share common contexts in the corpus are located close to each other in the space [34].

In our experiments, we built the word embeddings (i.e., vectors $\vec{w_i}$) using the skip-gram model.

3.1.4 Subprofile Specific Representation

The intuitive idea of the second order attributes consists in representing the terms by their relation with each target class [26, 29]. This can be done by exploiting occurrence-statistics over the set of documents in each one of the target classes.

In this way, we represent each term $t_i \in \mathcal{V}$ with a vector $\vec{w_i} = \langle w_{i,1}, \cdots, w_{i,q} \rangle$, where the scalar $w_{i,k}$ is the degree of association between word t_i and class C_k. Under this DTR, the weight $w_{i,k}$ is directly related to the number of occurrences of term t_i in documents that are labeled with class C_k.

In [29], second order attributes were modeled at sub-profile level; mainly, it was proposed to cluster the instances from each target in order to generate several subclasses. The idea was to consider the high heterogeneity of social media users.

Utilizing this process, the set of target classes C will now correspond to the set of all subgroups from the original target classes. This new representation is called *Subprofile-based Representation* (SSR), and is considered one of the state-of-the-art representations for AP.

3.2 Experiments and Results

This section explains the experiments that were carried out using the proposed framework. As we have previously mentioned, we aim at determining the pertinence of distributional term representations (DTRs) to the AP task in distinct social media domains. Accordingly, this section is organized as follows: first, Subsection 3.2.1 explains the experimental settings for all the experiments, then, Subsection 3.3 describes the results obtained by each DTR in the four different social media domains.

3.2.1 Experimental Setup

Preprocessing: For computing the DTRs of each social media domain we considered the 10,000 most frequent terms. We did not remove any term, i.e., we preserved all content words, stop words, emoticons, punctuations marks, etc. In one previous work [29] demonstrated that preserving only the 10,000 most frequent words is enough for achieving a good representation of the documents.

Text representation: The different DTRs were computed as described in Section 3.1.

Classification: Following the same configuration as in previous works (please refer to 4), in all the experiments we used the linear Support Vector Machine (SVM) from the LIBLINEAR library with default parameters [11].

Baseline: As baseline we employed the traditional bag-of-words (BoW) representation. We also compared the results from the different DTRs to those obtained by topic modeling representations such as LSA and LDA as well as to those from the top systems from the PAN@2014 AP track.

Evaluation: We performed a stratified 10 cross-fold validation (10-CFV) strategy. For comparison purposes, and following the PAN guidelines, we employed the accuracy as the main evaluation measure. Finally, we evaluated the statistical significance of the obtained results using a 0.05 significance level utilizing the Wilcoxon Signed-Ranks test since is recommended for these cases by [9].

Table 8. F-measure results obtained by the DTRs for the *age* classification problem

Approach	Text genres			
	Blogs	Reviews	Social Media	Twitter
DOR	**0.38**	**0.30**	**0.29**	**0.35**
TCOR	0.22	0.21	0.23	0.31
w2v-wiki	0.21	0.21	0.23	0.30
w2v-sm	0.20	0.20	0.24	0.28
SSR	0.36	0.27	0.26	0.33
Baseline	0.21	0.19	0.23	0.21

Table 9. F-measure results obtained by the employed DTRs for the *gender* classification task

App.	Text genres			
	Blogs	Reviews	Social Media	Twitter
DOR	**0.78***	**0.69***	0.52	0.70
TCOR	0.56	0.62	0.41	0.54
w2v-wiki	0.75*	0.64	0.52	0.69
w2v-sm	0.74	0.64	0.54	0.66
SSR	**0.78***	**0.69***	**0.55***	**0.71**
Baseline	0.72	0.62	0.52	0.70

3.3 Results

This section is organized as follows: first, we show the results from different DTRs for the age and gender classification tasks; then, we compare them against some topic-based representations and the best approaches from PAN 2014.

3.3.1 Age and Gender Identification Using DTRs

Table 8 shows the F-measures results for *age*. Also, Table 9 shows the obtained results for the *gender* classification problems respectively. Each row represents one of the described DTRs, i.e., DOR, TCOR, word2vec, and SSR, while the last row represents the baseline results. Every column refers to a distinct social media genre. In these tables, the best results are highlighted using boldface, and the star symbol ($\star$) indicates the differences that are statistically significant concerning the baseline results (in accordance to the used test; for details refer to Section 3.2.1).

Obtained results indicate that all DTRs, except for TCOR, outperformed the baseline method.

Table 10. Comparison of the best DTRs against topic-based methods in the *age* classification task

Approach	Text genres			
	Blogs	Reviews	Social Media	Twitter
DOR	**0.49**†	0.36†	**0.38**†‡	0.47‡
SSR	0.48†	0.34†	0.37‡	**0.48**†‡
LDA	0.44	0.27	0.37	0.47
LSA	**0.49**	**0.37**	0.36	0.45
[33]	0.38	0.33	0.36	0.44
[48]	0.39	0.31	0.35	0.41
[49]	0.45	0.37	0.42	0.52

Table 11. Comparison of best DTRs against topic-based methods in the *gender* classification task

Approach	Text genres			
	Blogs	Reviews	Social Media	Twitter
DOR	**0.78**†	**0.69**†	0.52	0.70†
SSR	**0.78**†	**0.69**†	**0.55**†‡	
LDA	0.61	0.55	0.52	0.64
LSA	**0.78**	**0.69**	0.53	0.70
[33]	0.57	0.66	0.53	0.66
[48]	0.64	0.68	0.54	0.51
[49]	0.82	0.71	0.57	0.78

In particular, DOR and SSR show statistically significant differences. These two methods obtained comparable results, being DOR slightly better than SSR in 5 out of 8 classification problems, which is an interesting result since SSR was among the winning approaches at PAN 2014. On the other hand, we attribute the low accuracy results showed by TCOR to the strong expansion that it imposes to the document representations.

Considering direct term co-occurrences causes the inclusion of many unrelated and unimportant terms in the document vectors, and, therefore, it complexities the extraction of profiling patterns.

Finally, another essential aspect to notice is the fact that both *w2v-wiki* and *w2v-sm* obtained similar results in each of the classifications problems, although the former learned the embeddings from a corpus that is not thematically and neither stylistically similar to the social media content. We presume these results could be explained by the relatively small size of the social media training collections, and, at the same time, by the large size and broad coverage of the used Wikipedia dataset, which has a vocabulary of 1,033,013 words.

Tables 10 and 11 compare the results from DOR and SSR, the best DTRs according to the previous results, against the results from two well-known topic-based representations, namely LDA and LSA.

Regarding the LSA results, it is possible to observe, on the one hand, that for *age* classification (refer to Table 10), its average performance is similar to the one from DOR, i.e., 42%. However, the only domain in which LSA outperforms DOR is in the reviews dataset. Nonetheless, there is no significant difference between these results. On the other hand, for *gender* classification (Table 11), LSA was not able to improve any result from DOR and SSR. It is important to mention that, although their results are comparable, LSA is a parametric method, and, therefore, tunning is required.

Finally, the works [33], [48] and [49] are the best results for the PAN@2014 forum.

4 Image Author Profiling Approach

4.1 Open-Vocabulary Method

The adopted UAIA method for labeling images with an open vocabulary approach was proposed in [36]. The general idea of this method relies on the use of a multimodal indexing $\mathcal{M}$ composes of visual prototypes that are used for labeling new images.

Given a reference collection of documents $\mathcal{D}$ that include texts $\mathcal{T}$ and images $\mathcal{I}$. First, each image in $\mathcal{V}$ is represented by a visual feature $\mathbf{v}_i$. In our case, we use the VGG-16 pre-trained model proposed in [46] for visual extraction. Then, each extracted word, i.e. $\mathbf{t}_i$, from $\mathcal{T}$ is represented by visual vector resulting from combining images that co-occur with the word, i.e., visual features of images included in documents where the word appears.

Mathematically, multimodal indexing could be done as follows:

$$\mathcal{M} = \mathcal{T}^T \cdot \mathcal{V}, \tag{2}$$

ISSN 2007-9737

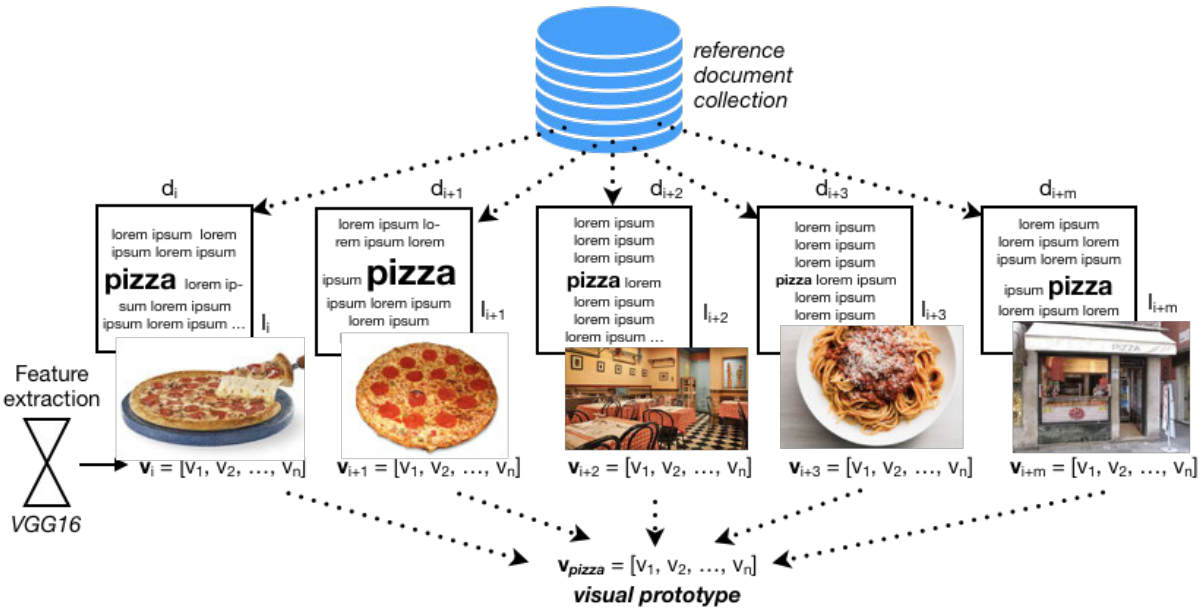


Fig. 2. Illustrate of how the annotator builds visual prototypes. In this example the method uses all images where appears the word 'pizza' and as result obtains its visual prototype

where $\mathcal{M}$ is the multimodal indexing obtained by the product of textual $\mathcal{T}$ and visual features $\mathcal{V}$ of documents. The advantage of the UAIA method is the capability to build visual prototypes from free and large vocabularies extracted from reference collections. The whole process is illustrated in Figure 2, for the case, note that the word 'pizza' is bigger when it appears with more frequency in the document. This mechanism prevents to add images with no relevance to the word.

For annotating a new image, first, it is described in a common representation to the visual prototypes, then it is compared for estimating a similarity score based on cosine distance:

$$cosine(\mathbf{q}, \mathcal{M}) = \frac{\mathbf{q} \cdot \mathbf{m}_i}{|\mathbf{q}| \times |\mathbf{m}_i|}, \quad (3)$$

where $\mathbf{q}$ is the visual representation of the query image, and $\mathbf{m}_i$ is the $i-th$ visual prototype in $\mathcal{M}$. The query image is compared with each visual prototype in $\mathcal{M}$, and n of the most similar visual prototypes, that is, the n of the most similar words are used for annotating the image.

LSA captures the topics in a corpus applying a mathematical technique called singular value decomposition (SVD) while preserving the similarity structure among the texts. The underlying idea is that the aggregate of all the word contexts, in which a given word does and does not appear, it provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. Unlike BoW, LSA represents each document D_j into a k-dimensional vector where k represents the number of discovered topics, thus, $D_j = \{x_1, x_2, ..., x_i, ..., x_k\}$.

Each dimension i in the vector represents the weight of a topic i in the document j [16]. For the performed experiments, when we apply LSA on the posts' text we express it as LSA_T, and when we use LSA on the labels extracted from the image annotation methods, we express it as LSA_I.

4.2 Results

Table 12 shows the results of the proposal compared with the individual text results as BoW and LSA_T for the Mexican collections for the 3 traits. Also, we compared the proposal results with the AlexNet [22] and RCNN [14] models. These models are based on deep learning and represent

Table 12. Obtained performance for the gender, occupation, and location tasks on the MEX-A3T corpus

Approach		*Accuracy*	*F1*
Gender Identification Task			
Textual baselines	BoW	0.80	0.80
	LSA_T	0.79	0.79
Visual baselines	AlexNet [22]	0.65	0.65
	RCNN [14]	0.64	0.64
Proposed	BoL	0.74*	0.74
	LSA_I	**0.79***	**0.79**
Multi modal	DOR+LSA_I	**0.79***	**0.79**
Occupation Identification Task			
Textual baselines	BoWk	0.64	0.34
	LSA_T	0.65	0.25
Visual baselines	AlexNet [22]	0.52	0.23
	RCNN [14]	0.54	0.24
Proposed	BoL	0.63*	**0.34**
	LSA_I	**0.65***	**0.34**
Multi modal	DOR+LSA_I	**0.68***	**0.39**
Location Identification Task			
Textual baselines	BoW	0.52	0.37
	LSA_T	0.71	0.57
Visual baselines	AlexNet [22]	0.35	0.24
	RCNN [14]	0.35	0.23
Proposed	BoL	0.44*	0.28
	LSA_I(k=100)	**0.50***	**0.31**
Multi modal	DOR+LSA_I	**0.68***	**0.58**

each image as a vector of 1000 semantic features. The proposal is represented as BoL for the bag of labels and LSA_I for the implementation of LSA for the labels of BoL.

For all traits, the textual representations obtain better or very similar results than the image representations. This indicates that the textual information is more valuable than the image information. Also, for the three traits, the proposal results overcome the deep learning based methods. Particularly, LSI_I overcomes the BoL implementation, it seems that to group the labels by their contest provides a better representation.

Finally, we implement a fusion strategy for taking advantage of both modalities. We use the late fusion [30] concatenated both spaces, DOR (the best DTR result) and LSA_I (the best image representation). As we can see, the most noticeable difference occurs for the location trait

however, for occupation, the results are better with the fusion too. But, for gender trait, the text result is still the best result. This could occurs because the occupation and location are traits unbalanced and the gender trait is balanced, Thus the occupation and location traits are harder tasks than the gender classification and it is necessary more information to help the classification [31].

5 Cross-Language Gender Prediction

We appraise the robustness of our proposed method under a cross-lingual scenario [12, 27]. For this, we performed several experiments training and evaluating using distinct *source* and *target* languages, and compared against the best results obtained in a monolingual situation.

The hypothesis behind this idea establishes that users with distinct native languages, having a similar profile, will share analogous images. To the best of our knowledge, this is the first attempt in proposing a cross-language gender prediction method based on merely visual information.

In order to prove this hypothesis, we performed a series of experiments for the gender prediction task[4]. Similar to the previous experiments, we compare the performance of the closed vocabulary approaches (AlexNet and RCNN) against the performance of the open vocabulary approach (LSA$_I$) under a cross-lingual scenario.

Table 13. Cross language results using AlexNet as image annotation method

Source language	*Target language*	*Acc.*	F_1	F_1 Male	F_1 Female
EN	EN	0.58	0.58	0.56	0.60
SP	EN	0.59	0.59	0.59	0.59
SP+EN	EN	**0.61**	**0.61***	0.60	0.62
SP	SP	0.65	0.65	0.65	0.65
EN	SP	0.64	0.64	0.64	0.64
SP+EN	SP	**0.66**	**0.66***	0.66	0.66

[4]The *gender* trait is the only common trait among both datasets, i.e., PAN 2014 and MEX-A3T

Table 14. Cross language results using RCNN as image annotation method

Source language	*Target language*	*Acc.*	F_1	F_1 Male	F_1 Female
EN	EN	0.56	0.56	0.56	0.56
SP	EN	0.60	0.55	0.70	0.40
SP+EN	EN	**0.61**	**0.61***	0.61	0.61
SP	SP	0.64	0.64	0.64	0.64
EN	SP	0.64	0.59	0.46	0.72
SP+EN	SP	**0.65**	**0.65***	0.64	0.66

Table 15. Cross language results using the proposed method under the LSA$_I$. The number between parenthesis indicates the value of the k parameter for the LSA method

Source language	*Target language*	*Acc.*	F_1	F_1 Male	F_1 Female
EN	EN(100)	0.72	0.72	0.71	0.72
SP	EN(100)	0.60	0.55	0.70	0.40
SP+EN	EN(50)	**0.84**	**0.84***	0.84	0.84
SP	SP(100)	0.79	0.79	0.79	0.79
EN	SP(100)	0.64	0.59	0.46	0.72
SP+EN	SP(50)	**0.80**	**0.80***	0.80	0.80

5.1 Results

Table 13, 14 and 15 show the obtained results using ALexNet, RCNN and LSA$_I$ methods for labeling the visual information. It is interesting to observe that when only one language is used for training (EN$\rightarrow$ EN, SP$\rightarrow$EN, SP$\rightarrow$SP, EN$\rightarrow$SP), achieved performance is very similar for all AIA methods. However, a significant improvement is obtained when the combination of the two languages (SP+EN) is employed to train the classification model Particularly, observe the LSA$_I$ method (Table 15) outperforms both AlexNet (Table 13) and RCNN (Table 14) configurations.

These results evidentiate that similar users share in fact similar images, allowing an automatic classifier to distinguish among users, regardless of their native language. In order to exemplify this affirmation, we took on the task of retrieving the most important images from the top 5 topics

(a) Topic 1: People **(b)** Topic 2: Sports

(c) Topic 3: Women **(d)** Topi 4: Technology

(e) Topic 5: Diagrams

Fig. 3. Representative images for each topic extracted with LSA

identified by the LSA_I approach. Figure 3 illustrates the retrieved images.

For each topic six images are shown, where the three from the left correspond to Spanish speaking users, and the three on the right to English speaking users. After observing the retrieved images, is possible to conclude that shared images by users not sharing language, at least between males and females, contain similar characteristics. This language and culture independent phenomenon indicates that is possible to configure cross-lingual AP methods.

6 Conclusions

As a result of this work, the following conclusions were obtained.

DTR's have advantages in the author profiling task compared with other approaches to capture the content of the texts. In particular, DOR presents the best behavior, besides that DOR is not a parameterized approach, which causes it to be a simpler and more efficient approach to this task. Also, a significant advantage of DOR is its robustness across different social media genres, contrary to others approaches.

Automatic image annotation based on open vocabulary approaches is better to represent the images than the closed vocabulary approaches for the Author profiling task.

There is complementarity among the textual and image modalities since it is possible to overcome the individual results with fusion schemes.

Also, it is possible to use image information from another corpus, even if the corpus is in another language. This seems reasonable, taking into account that images are language independent. It seems that the open vocabulary approach with LSA_I represents better the images from different native languages users.

Acknowledgment

Álvarez-Carmona thanks for doctoral scholarship CONACyT-Mexico 401887.

References

1. **Álvarez-Carmona, M. Á. (2019).** *Author profiling in social media with multimodal information*. Ph.D. thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica.
2. **Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018).** Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September*.
3. **Álvarez-Carmona, M. A., López-Monroy, A. P., Montes-y Gómez, M., Villaseñor-Pineda, L., & Escalante, H. J. (2015).** INAOE's participation at PAN'15: Author profiling task. *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, Vol. 1391.
4. **Álvarez-Carmona, M. A., López-Monroy, A. P., Montes-y Gómez, M., Villaseñor-Pineda, L., & Meza, I. (2016).** Evaluating topic-based representations for author profiling in social media. *Ibero-American Conference on Artificial Intelligence*, Springer, pp. 151–162.
5. **Alvarez-Carmona, M. A., Villatoro-Tello, E., Villasenor-Pineda, L., et al. (2019).** A comparative analysis of distributional term representations for author profiling in social media. *Journal of Intelligent & Fuzzy Systems*, Vol. 36, No. 5, pp. 4857–4868.
6. **Aragón, M. E., Álvarez-Carmona, M. Á., Montes-y Gómez, M., Escalante, H. J., Villasenor-Pineda, L., & Moctezuma, D. (2019).** Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. *Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain*.
7. **Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. W. (2005).** Lexical predictors of personality type. *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*.
8. **Bergsma, S., Post, M., & Yarowsky, D. (2012).** Stylometric analysis of scientific articles. *Proceedings of the 2012 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, pp. 327–337.

9. **Demšar, J. (2006).** Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, Vol. 7, No. Jan, pp. 1–30.
10. **Eftekhar, A., Fullwood, C., & Morris, N. (2014).** Capturing personality from Facebook photos and photo-related activities: How much exposure do you need? *Computers in Human Behavior*, Vol. 37, pp. 162–170.
11. **Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008).** Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874.
12. **Feliciano-Avelino, I., Álvarez-Carmona, M. Á., Escalante, H. J., Montes-y Gómez, M., & Villaseñor-Pineda, L. (2019).** Cross-cultural image-based author profiling in twitter. *Mexican International Conference on Artificial Intelligence*, Springer, pp. 353–363.
13. **Gelbukh, A. (2019).** Computational linguistics: Introduction to the thematic issue. *Computación y Sistemas*, Vol. 23, No. 3.
14. **Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014).** Rich feature hierarchies for accurate object detection and semantic segmentation. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587.
15. **Grimshaw, M. (2013).** *The Oxford handbook of virtuality*. Oxford University Press.
16. **Hofmann, T. (1999).** Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 50–57.
17. **Kharroub, T. & Bas, O. (2015).** Social media and protests: An examination of Twitter images of the 2011 Egyptian revolution. *New Media & Society*, pp. 1461444815571914.
18. **Kodiyan, D., Hardegger, F., Neuhaus, S., & Cieliebak, M. (2017).** Author profiling with bidirectional rnns using attention with grus, pp. 1–10.
19. **Koppel, M., Akiva, N., Alshech, E., & Bar, K. (2009).** Automatically classifying documents by ideological and organizational affiliation. *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on*, IEEE, pp. 176–178.
20. **Koppel, M., Argamon, S., & Shimoni, A. R. (2002).** Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, Vol. 17, No. 4, pp. 401–412.
21. **Koppel, M., Schler, J., & Zigdon, K. (2005).** Determining an author's native language by mining a text for errors. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, ACM, pp. 624–628.
22. **Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012).** Imagenet classification with deep convolutional neural networks. In **Pereira, F., Burges, C. J. C., Bottou, L., & Weinberger, K. Q.**, editors, *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., pp. 1097–1105.
23. **Lavelli, A., Sebastiani, F., & Zanoli, R. (2004).** Distributional term representations: an experimental comparison. *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, ACM, pp. 615–624.
24. **Lavelli, A., Sebastiani, F., & Zanoli, R. (2004).** Distributional term representations: An experimental comparison. *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, ACM, New York, NY, USA, pp. 615–624.
25. **Levy, O., Goldberg, Y., & Dagan, I. (2015).** Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, Vol. 3, pp. 211–225.
26. **Li, Z., Xiong, Z., Zhang, Y., Liu, C., & Li, K. (2011).** Fast text categorization using concise semantic analysis. *Pattern Recognition Letters*, Vol. 32, No. 3, pp. 441–448.
27. **López, R., Peñaloza, D., Beingolea, F., Tenorio, J., & Sobrevilla Cabezudo, M. (2019).** An exploratory study of the use of senses, syntax and cross-linguistic information for subjectivity detection in spanish. *Computación y Sistemas*, Vol. 23, No. 3.
28. **López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., & Villaseñor-Pineda, L. (2014).** Using intra-profile information for author profiling. *CLEF 2014 Working Notes*, pp. 1116–1120.
29. **López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., & Stamatatos, E. (2015).** Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems*, Vol. 89, pp. 134–147.

30. **Loyola-González, O., López-Cuevas, A., Medina-Pérez, M. A., Camiña, B., Ramírez-Márquez, J. E., & Monroy, R. (2019).** Fusing pattern discovery and visual analytics approaches in tweet propagation. *Information Fusion*, Vol. 46, pp. 91–101.
31. **Loyola-González, O., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & García-Borroto, M. (2016).** Effect of class imbalance on quality measures for contrast patterns: An experimental study. *Information Sciences*, Vol. 374, pp. 179–192.
32. **Loyola-González, O., Medina-Pérez, M. A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., Monroy, R., & García-Borroto, M. (2017).** Pbc4cip: A new contrast pattern-based classifier for class imbalance problems. *Knowledge-Based Systems*, Vol. 115, pp. 100–109.
33. **Maharjan, S., Shrestha, P., & Solorio, T. (2014).** A simple approach to author profiling in mapreduce. *CLEF (Working Notes)*, pp. 1121–1128.
34. **Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013).** Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
35. **Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013).** Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pp. 3111–3119.
36. **Pellegrin, L., Escalante, H. J., Montes-y Gómez, M., & González, F. A. (2016).** Local and global approaches for unsupervised image annotation. *Multimedia Tools and Applications*, Vol. 76, No. 15, pp. 16389–16414.
37. **Poulston, A., Waseem, Z., & Stevenson, M. (2017).** Using tf-idf n-gram and word embedding cluster ensembles for author profiling, pp. 1–6.
38. **Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., & Daelemans, W. (2014).** Overview of the 2nd author profiling task at pan 2014. *Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes)*, pp. 1–30.
39. **Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., & Stein, B. (2018).** Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. *Working Notes Papers of the CLEF*, pp. 1–38.
40. **Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015).** Overview of the 3rd author profiling task at pan 2015. *CLEF*, sn, pp. 2015.
41. **Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006).** Effects of age and gender on blogging. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pp. 199–205.
42. **Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E., et al. (2013).** Characterizing geographic variation in well-being using tweets. *ICWSM*, pp. 583–591.
43. **Sierra, S. & González, F. A. (2018).** Combining textual and visual representations for multimodal author profiling. *Working Notes Papers of the CLEF*, Vol. 2125, pp. 219–228.
44. **Skalmowski, W. (2016).** Review of harris, zellig (1968) mathematical structures of language. *ITL-International Journal of Applied Linguistics*, Vol. 4, No. 1, pp. 56–61.
45. **Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., & Ohkuma, T. (2018).** Text and image synergy with feature cross technique for gender identification. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, volume 2125, pp. 10–22.
46. **Tindall, L., Luong, C., & Saad, A. (2015).** Plankton classification using vgg16 network.
47. **Villegas, M. P., Garciarena Ucelay, M. J., Fernández, J. P., Álvarez Carmona, M. A., Errecalde, M. L., & Cagnina, L. (2016).** Vector-based word representations for sentiment analysis: a comparative study. *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016).*
48. **Villena Román, J. & González Cristóbal, J. C. (2014).** Daedalus at pan 2014: Guessing tweet author's gender and age, pp. 1157–1163.
49. **Weren, E. R., Kauer, A. U., Mizusaki, L., Moreira, V. P., de Oliveira, J. P. M., & Wives, L. K. (2014).** Examining multiple features for author profiling. *Journal of Information and Data Management*, Vol. 5, No. 3, pp. 266.
50. **Wu, Y.-C. J., Chang, W.-H., & Yuan, C.-H. (2014).** Do Facebook profile pictures reflect user's personality? *Computers in Human Behavior*, Vol. 51, pp. 880–889.

Article received on 14/06/2020; accepted on 21/07/2020.
Corresponding author is Manuel Montes y Gómez.

CEUR Workshop Proceedings
http://ceur-ws.org
ISSN 1613-0073

Vol-2624
urn:nbn:de:0074-2624-4

SWISSTEXT & KONVENS 2020
Swiss Text Analytics Conference & Conference on Natural Language Processing 2020

Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)

Zurich, Switzerland, June 23-25, 2020 (held online due to COVID19 pandemic).

Edited by

Sarah Ebling *
Don Tuggener **
Manuela Hürlimann **
Mark Cieliebak **
Martin Volk *

* University of Zurich, Department of Computational Linguistics, 8050 Zurich, Switzerland
** ZHAW Zurich University of Applied Sciences, Institute of Applied Information Technology (InIT), 8401 Winterthur, Switzerland

Table of Contents

- Preface

Scientific Track

- Investigating the Influence of Selected Linguistic Features on Authorship Attribution using German News Articles
  Manuel Sage, *Pietro Cruciata*, *Raed Abdo*, *Jackie Chi Kit Cheung*, *Yaoyao Fiona Zhao*
- On the Comparability of Pre-trained Language Models
  Matthias Aßenmacher, *Christian Heumann*
- Supervised Pun Detection and Location with Feature Engineering and Logistic Regression
  Jingyuan Feng, *Özge Sevgili*, *Steffen Remus*, *Eugen Ruppert*, *Chris Biemann*
- Compiling a Large Swiss German Dialect Corpus
  Manuela Weibel, *Muriel Peter*
- To BERT or not to BERT – Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of four Types of Speech, Thought and Writing Representation
  Annelen Brunner, *Ngoc Duyen Tanja Tu*, *Lukas Weimer*, *Fotis Jannidis*
- Cultural Differences in Bias? Origin and Gender Bias in Pre-Trained German and French Word Embeddings
  Mascha Kurpicz-Briki
- Evaluating German Transformer Language Models with Syntactic Agreement Tests
  Karolina Zaczynska, *Nils Feldhus*, *Robert Schwarzenberg*, *Aleksandra Gabryszak*, *Sebastian Möller*
- Predicting the Concreteness of German Words
  Jean Charbonnier, *Christian Wartena*
- X -stance: A Multilingual Multi-Target Dataset for Stance Detection
  Jannis Vamvas, *Rico Sennrich*
- Harmonization Sometimes Harms
  Manfred Klenner, *Anne Göhring*, *Michael Amsler*
- Ranking Georeferences for Efficient Crowdsourcing of Toponym Annotations in a Historical Corpus of Alpine Texts
  Janis Goldzycher, *Isabel Meraner*, *Martin Volk*, *Simon Clematide*
- HarryMotions – Classifying Relationships in Harry Potter based on Emotion Analysis
  Albin Zehe, *Julia Arns*, *Lena Hettinger*, *Andreas Hotho*
- Cross-lingual Transfer-learning Approach to Negation Scope Resolution
  Anastassia Shaitarova, *Lenz Furrer*, *Fabio Rinaldi*

- Psychological Distance in German and English Brand Language of Eight International Brands
  Simone Griesser

GermEval 2020 Task 2: Swiss German Language Identification

- Overview of the GermEval 2020 Shared Task on Swiss German Language Identification
  Pius von Däniken, *Manuela Hürlimann*, *Mark Cieliebak*
- Spoken Dialect Identification in Twitter using a Multi-filter Architecture
  Mohammadreza Banaei, *Rémi Lebret*, *Karl Aberer*
- Detecting Noisy Swiss German Web Text Using RNN- and Rule-Based Techniques
  Janis Goldzycher, *Jonathan Schaber*
- Idiap Submission to Swiss-German Language Detection Shared Task
  Shantipriya Parida, *Esaú Villatoro-Tello*, *Sajit Kumar*, *Petr Motlicek*, *Qingran Zhan*

GermEval 2020 Task 3: 2nd German Text Summarization Challenge

- 2nd German Text Summarization Challenge
  Dominik Frefel, *Manfred Vogel*, *Fabian Märki*
- UPB at GermEval-2020 Task 3: Assessing Summaries for German Texts using BERTScore and Sentence-BERT
  Andrei Paraschiv, *Dumitru-Clementin Cercel*
- Hybrid Ensemble Predictor as Quality Metric for German Text Summarization: Fraunhofer IAIS at GermEval 2020 Task 3
  David Biesner, *Eduardo Brito*, *Lars Patrick Hillebrand*, *Rafet Sifa*

GermEval 2020 Task 4: Low-Resource Speech-to-Text

- GermEval 2020 Task 4: Low-Resource Speech-to-Text
  Michel Plüss, *Lukas Neukom*, *Manfred Vogel*
- LTL-UDE at Low-Resource Speech-to-Text Shared Task: Investigating Mozilla DeepSpeech in a low-resource setting
  Aashish Agarwal, *Torsten Zesch*
- ZHAW-InIT at GermEval 2020 Task 4: Low-Resource Speech-to-Text
  Matthias Büchi, *Malgorzata Anna Ulasik*, *Manuela Hürlimann*, *Fernando Benites*, *Pius von Däniken*, *Mark Cieliebak*
- UZH TILT: A Kaldi recipe for Swiss German Speech to Standard German Text
  Tannon Kew, *Iuliia Nigmatulina*, *Lorenz Nagele*, *Tanja Samardžić*

Additional Material

- Abstracts of the Applied Track

2020-06-11: submitted by Don Tuggener, metadata incl. bibliographic data published under Creative Commons CC0
2020-06-18: published on CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073) |valid HTML5|

Idiap Submission to Swiss-German Language Detection Shared Task

Shantipriya Parida[1], **Esaú Villatoro-Tello**[2,1], **Sajit Kumar**[3], **Petr Motlicek**[1] **and Qingran Zhan**[1]

[1]Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland.
firstname.lastname@idiap.ch
[2]Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Mexico City, Mexico.
evillatoro@correo.cua.uam.mx
[3]Centre of Excellence in AI, Indian Institute of Technology, Kharagpur, West Bengal, India.
kumar.sajit.sk@gmail.com

Abstract

Language detection is a key part of the NLP pipeline for text processing. The task of automatically detecting languages belonging to disjoint groups is relatively easy. It is considerably challenging to detect languages that have similar origins or dialects. This paper describes Idiap's submission to the 2020 Germeval evaluation campaign[1] on Swiss-German language detection. In this work, we have given high dimensional features generated from the text data as input to a supervised autoencoder for detecting languages with dialect variances. Bayesian optimizer was used to fine-tune the hyper-parameters of the supervised autoencoder. To the best of our knowledge, we are first to apply supervised autoencoder for the language detection task.

1 Introduction

The increased usage of smartphones, social media, and the internet has led to rapid growth in the generation of short linguistic texts. Thus, identification of language is a key component in building various NLP resources (Kocmi and Bojar, 2017). Language detection is the task of determining the language for the given text. Although it has progressed substantially, still few challenges exist: (1) distinguishing among similar languages, (2) detection of languages when multiple language contents exist within a single document, and (3) language identification in very short texts (Balazevic et al., 2016; Lui et al., 2014; Williams and Dagli, 2017).

It is a difficult task to discriminate between very close languages or dialects (for example, German dialect identification, Indo-Aryan language identification (Jauhiainen et al., 2019a)). Although dialect identification is commonly based on the distributions of letters or letter n-grams, it may not be possible to distinguish related dialects with very similar phoneme and grapheme inventories for some languages (Scherrer and Rambow, 2010).

Many authors proposed traditional machine learning approaches for language detection like Naive Bayes, SVM, word and character n-grams, graph-based n-grams, prediction partial matching (PPM), linear interpolation with post-independent weight optimization and majority voting for combining multiple classifiers, etc. (Jauhiainen et al., 2019b).

More recently, deep learning techniques have shown substantial performance in many NLP tasks including language detection (Oro et al., 2018). In the context of deep learning techniques, many papers have demonstrated the capability of semi-supervised autoencoders solving different tasks, indicating that the use of autoencoders allows learning a representation when trained with unlabeled data. (Ranzato and Szummer, 2008; Rasmus et al., 2015). However, as per our literature survey, none of the recent research has applied autoencoder for the language detection task. In this paper, we propose a supervised configuration of the autoencoders, which utilizes labels for learning the representation. To the best of our knowledge, this is the first time this technology is evaluated in the context of the language detection task.

1.1 Supervised Autoencoder

An autoencoder (AE) is a neural network that learns a representation (encoding) of input data and then learns to reconstruct the original input from the learned representation. The autoencoder is mainly

[1]https://sites.google.com/view/gswid2020

used for dimensionality reduction or feature extraction (Zhu and Zhang, 2019). Normally, it is used in an unsupervised learning fashion, meaning that we leverage the neural network for the task of representation learning. By learning to reconstruct the input, the AE extracts underlying abstract attributes that facilitate accurate prediction of the input.

Thus, a supervised autoencoder (SAE) is an autoencoder with the addition of a supervised loss on the representation layer. For the case of a single hidden layer, a supervised loss is added to the output layer and for a deeper autoencoder, the innermost (smallest) layer would have a supervised loss added to the bottleneck layer that is usually transferred to the supervised layer after training the autoencoder.

In supervised learning, the goal is to learn a function for a vector of inputs $\mathbf{x} \in \mathbb{R}^d$ to predict a vector of targets $\mathbf{y} \in \mathbb{R}^m$. Consider SAE with a single hidden layer of size k, and the weights for the first layer are $\mathbf{F} \in \mathbb{R}^{k \times d}$. The function is trained on a finite batch of independent and identically distributed (i.i.d.) data, $(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_t, \mathbf{y}_t)$, with the goal of a more accurate prediction on new samples generated from the same distribution. The weight for the output layer consists of weights $\mathbf{W}_p \in \mathbb{R}^{m \times k}$ to predict $\mathbf{y}$ and $\mathbf{W}_r \in \mathbb{R}^{d \times k}$ to reconstruct $\mathbf{x}$. Let L_p be the supervised loss and L_r be the loss for the reconstruction error. In the case of regression, both losses might be represented by a squared error, resulting in the objective:

$$\frac{1}{t}\sum_{i=1}^{t}\Big[L_p(\mathbf{W}_p\mathbf{F}\mathbf{x}_i, \mathbf{y}_i) + L_r(\mathbf{W}_r\mathbf{F}\mathbf{x}_i, \mathbf{x}_i)\Big] = \frac{1}{2t}\sum_{i=1}^{t}\Big[||\mathbf{W}_p\mathbf{F}\mathbf{x}_i - \mathbf{y}_i||_2^2 + ||\mathbf{W}_r\mathbf{F}\mathbf{x}_i - \mathbf{x}_i||_2^2\Big] \tag{1}$$

The addition of supervised loss to the autoencoder loss function acts as regularizer and results (as shown in equation 1) in the learning of the better representation for the desired task (Le et al., 2018).

1.2 Bayesian Optimizer

In the case of SAE, there are many hyperparameters related to (a) Model construction and (b) Optimization. Hence, SAE training without any hyperparameter tuning usually results in poor performance due to the dependencies that may result in simultaneous over/under-fitting.

Global optimization is considered to be a challenging problem of finding the globally best solution of (possibly nonlinear) models, in the (possible or known) presence of multiple local optima. Bayesian optimization (BO) is shown to outperform other state-of-the-art global optimization algorithms on several challenging optimization benchmark functions (Snoek et al., 2012; Bergstra and Bengio, 2012). BO provides a principled technique based on Bayes theorem to direct a search for a global optimization problem that is efficient and effective. It works by building a probabilistic model of the objective function, called the surrogate function, that is then searched efficiently with an acquisition function before candidate samples are chosen for evaluation on the real objective function. It tries to solve the minimization problem:

$$X^* = \arg\min_{x \in \chi} f(x), \tag{2}$$

where we consider χ to be a compact subset of $\mathbb{R}^k$ (Snoek et al., 2015).

Thus, we employed BO for hyperparameter optimization where the objective is to find the hyperparameters of a given machine learning algorithm, for this, we preserved the best performance as measured on a validation set.

2 Proposed Method

The architecture of the proposed model is shown in Figure 1. We used character n-grams as features from the input text. In comparison to word n-grams, which only capture the identity of a word and its possible neighbors, character n-grams are additionally capable of providing an excellent trade-off between sparseness and word's identity, while at the same time they combine different types of information: punctuation, morphological makeup of a word, lexicon and even context (Wei et al., 2009; Kulmizev et al., 2017; Sánchez-Vega et al., 2019). The extracted n-gram features are input to the deep SAE as shown in the Figure 1. The deep SAE contains multiple hidden layers. We used the BO for selecting the optimal parameters.

3 Experimental Setup and Datasets

The training dataset was provided by the organizers of the shared task. The training[2] dataset consists of 2,000 tweets in the Swiss-German language. The

[2]Although 2K Twitter ids were provided, we were not able to retrieve them all, resulting in 1976 training instances.

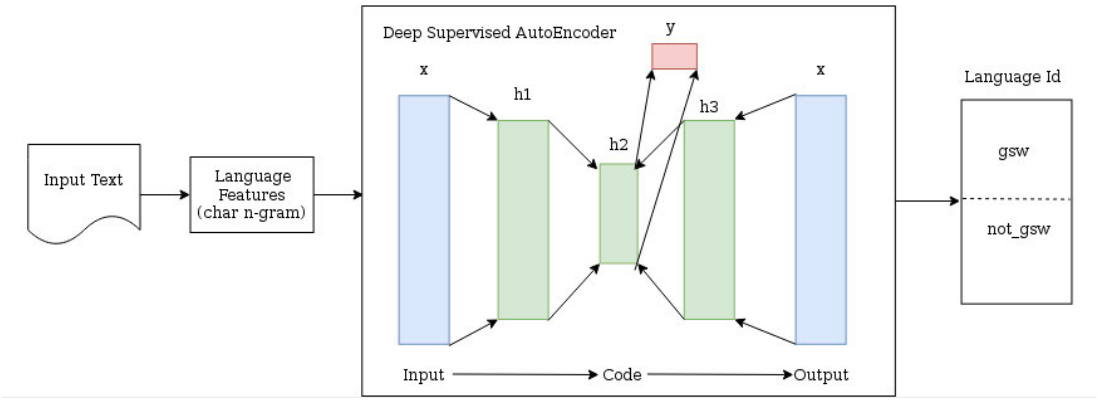


Figure 1: Proposed model architecture. The extracted features of the text are fed to the supervised autoencoder. The targets "y" are included. The classification output are the language ids for the classified languages.

participants were allowed to use any additional resources as training datasets. As part of the additional resources recommended by the organizers, the following Swiss-German datasets were suggested: NOAH [3] (Hollenstein and Aepli, 2015), and SwissCrawl [4](Linder et al., 2019); which we used in our experiments.

The test data released by the organizers consists of 5,374 Tweets (mix of different languages) to be classified as Swiss-German versus not Swiss-German.

The training dataset provided by the organizer did not have any non-Swiss-German text. In addition to the recommended Swiss-German datasets, we have used other non-Swiss-German datasets (DSL [5] (Tan et al., 2014a), and Ling10 [6]) for training our models.

- *DSL Dataset:* The data obtained from the "Discriminating between Similar Language (DSL) Shared Task 2015" contains 13 different languages as shown in Table 1. The DSL corpus collection have different versions based on different language group which provides datasets for researchers to test their systems (Tan et al., 2014a). We selected DSLCC version 2.0 [7] in our experiments (Tan et al., 2014b).
- *Ling10 Dataset :* The Ling10 dataset contains 190,000 sentences categorized into 10 languages (English, French, Portuguese, Chinese Mandarin, Russian, Hebrew, Polish, Japanese, Italian, Dutch) mainly used for language detection and benchmarking NLP algorithms. We considered "Ling10-trainlarge" (one of the three variants of Ling10 dataset) in our experiment.

Group Name	Language	Id
South Eastern Slavic	Bulgarian Macedonian	bg mk
South Western Slavic	Bosnian Croatian Serbian	bs hr sr
West-Slavic	Czech Slovak	cz sk
Ibero-Romance(Spanish)	Peninsular Spain Argentinian Spanish	es-ES es-AR
Ibero-Romance(Portuguese)	Brazilian Portuguese European Portuguese	pt-BR pt-PT
Astronesian	Indonesian Malay	id my

Table 1: DSL Language Group. Similar languages with their language code.

As the task is a binary classification of Swiss-German versus not Swiss-German, we have split all our collection of datasets including the training set provided by the organizers into two categories as follows:

- Swiss-German (NOAH, SwissCrawl, Swiss-German Training Tweets).
- not Swiss-German (DSL, Ling10).

Accordingly, we labeled the target class of all the Swiss-German text as "gsw" (Swiss-German) and labeled the target class of all other language

[3] https://noe-eva.github.io/NOAH-Corpus/

[4] https://icosys.ch/swisscrawl

[5] http://ttg.uni-saarland.de/resources/DSLCC/

[6] https://github.com/johnolafenwa/Ling10

[7] https://github.com/Simdiva/DSL-Task/tree/master/data/DSLCC-v2.0

text as "not_gsw").

We prepared three settings (S1, S2, and S3) combining the above datasets in different proportions of Swiss-German versus not Swiss-German languages for training the model. The statistics of the datasets for the settings are shown in Table 2.

We mixed the datasets of Swiss-German and other languages and split them into different ratios for training and development as per the settings. In each setting, the training and development set is different based on the selection of the number of sentences from each dataset. We used the test set provided by the shared task organizers. As the test set includes twitter text during preprocessing, we removed emojis and other unnecessary symbols.

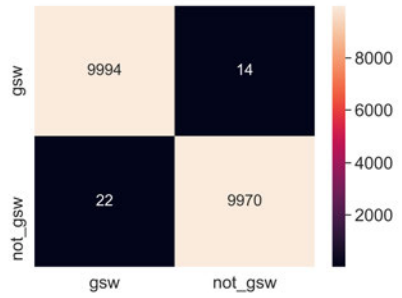
The range of values for the hyperparameters search space is shown in Table 3. During training, BO chooses the best hyperparameters from this range. The overall configuration of the SAE model is shown in Table 4.

4 Results and Discussion

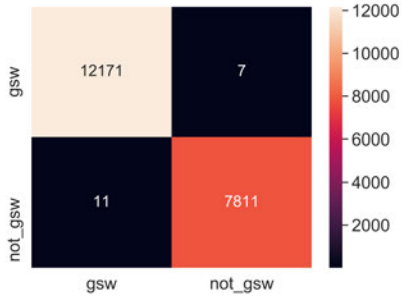
We evaluated the development set performance and the test set evaluation performed by the shared task organizers. The development set performance is given in section Section 4.1 and the test set performance in Section 4.2.

Our evaluation includes calculating classification accuracy based on the predicted label compared with the actual label. The organizers calculated *precision*, average *precision*, *recall*, and *F1* score for each of the submissions. As known, *precision* is the ratio of correctly predicted positive observations to the total predicted positive observations; *recall* (or sensitivity) is the ratio of correctly predicted positive observations to all observations in actual positive class, and the *F1* score is the weighted average of *precision* and *recall*.

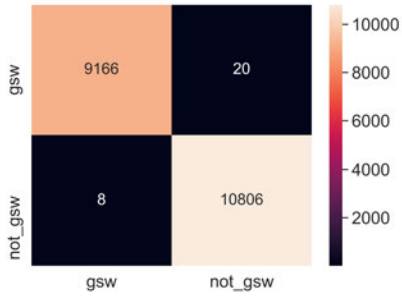
Organizers also generated the Receiver Operating Characteristic curve (ROC), Area Under the ROC Curve (AUC), and Precision-Recall (PR) curves. The AUC - ROC curve is a performance measurement at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It indicates how much a trained model is capable of distinguishing between classes, thus, the higher the AUC, the better the model performance. Finally, PR curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds; hence, a good model is represented by a curve that bows towards (1,1).



Confusion matrix for setting S1 on dev set.



Confusion matrix for setting S2 on dev set.



Confusion matrix for setting S3 on dev set.

Figure 2: Confusion matrix on the development (dev) set for the setting S1, S2, and S3. The confusion matrix shows the correct and incorrect predictions with count values broken down by each class i.e. "gsw" (Swiss-German) or "not_gsw" (not Swiss-German).

4.1 Development Set

The SAE model performance for the three settings (S1, S2, and S3) on the development set is shown in Table 5. The confusion matrix for all the settings on the development set is shown in Figure 2. The confusion matrix shows the correct and incorrect predictions with count values broken down by each class i.e. "gsw" (Swiss-German) or "not_gsw" (not

Setting	Datasets and Language	Distribution	Distribution (Overall)	Training	Dev	Test
S1	NOAH (*Swiss-German*) SwissCrawl (*Swiss-German*) SwissTextTrain (*Swiss-German*) DSL (*not Swiss-German*) Ling10 (*not Swiss-German*)	7,327 (8%) 40,697 (40%) 1,976 (2 %) 25,000 (25 %) 25,000 (25 %)	50% Swiss-German 50% not Swiss-German	80,000	20,000	5,374
S2	NOAH (*Swiss-German*) SwissCrawl (*Swiss-German*) SwissTextTrain (*Swiss-German*) DSL (*not Swiss-German*) Ling10 (*not Swiss-German*)	7,327 (5%) 81,841 (55 %) 1,976 (1 %) 25,000 (17 %) 33,856 (22 %)	61% Swiss-German 39% not Swiss German	130,000	20,000	5,374
S3	NOAH (*Swiss-German*) SwissCrawl (*Swiss-German*) SwissTextTrain (*Swiss-German*) DSL (*not Swiss-German*) Ling10 (*not Swiss-German*)	7,327 (4 %) 81,841 (41 %) 1,976 (1 %) 50,000 (25 %) 58,856 (29 %)	46% Swiss-German 54% not Swiss-German	180,000	20,000	5,374

Table 2: Dataset Statistics. The training-development-test set distribution for each of setting (S1, S2 and S3). The distribution is based on the number of sentences selected from the datasets.

Hyper Parameter	Range
number of layer	1-5
learning rate	$10^{-5} - 10^{-2}$
weight decay	$10^{-6} - 10^{-3}$
activation functions	'relu', 'sigma'

Table 3: Search space hyper parameter range.

Parameter	Value
char `n_gram` range	1-3
number of target	2
embedding dimension	300
supervision	'clf' (classification)
converge threshold	0.00001
number of epochs	500

Table 4: SAE model configuration used for training.

Swiss-German).

		Accuracy (%)
Model	**Setting**	**Development Set**
SAE (char-3gram)	S1	100
SAE (char-3gram)	S2	100
SAE (char-3gram)	S3	100

Table 5: Swiss-German language detection performance (classification accuracy) of the proposed model on the development set based on the setting S1, S2, and S3.

4.2 Test Set

The overall result announced by the organizers on test set is shown in the Table 6 and in the Figure 3. Our submission labeled as *"IDIAP"*, obtained the results *0.777*, *0.998*, and *0.872* for precision (prec), recall (rec), and F1 score respectively for the setting S3 as shown in Table 6. The detailed performance of each of our setting is shown in Table 7.

	Precision	Recall	F1
IDIAP	0.775	0.998	0.872
jj-cl-uzh	0.945	0.993	0.968
Mohammadreza Banaei	0.984	0.979	**0.982**

Table 6: Shared task result announced by the organizers displaying participant team and their model performance (Precision, Recall, and F1).

Setting	Prec (gsw)	Rec (gsw)	F1 (gsw)	Avg. Prec	AUROC
S1	0.649	0.997	0.786	0.871	0.924
S2	0.673	0.997	0.804	0.911	0.946
S3	**0.775**	**0.998**	**0.872**	**0.965**	**0.975**

Table 7: Performance of setting S1, S2, and S3.

Based on our initial analysis, we presume that the low performance of the SAE on the test set is due to the very few samples of twitter data available in the training data.

5 Conclusion

In this paper, we have shown the pertinence of SAE with Bayesian optimizer for the language detection task. Obtained results are encouraging, and SAE was found effective for discriminate between very close languages or dialects. The proposed model can be extended by creating a host of features such as character n-gram, word n-gram, word counts, etc and then passing it through autoencoder to choose the best features. In future work, we plan to (i) verify our model (SAE with BO) with other language detection datasets, and (ii) include more short texts, particularly Twitter data, in the training set and

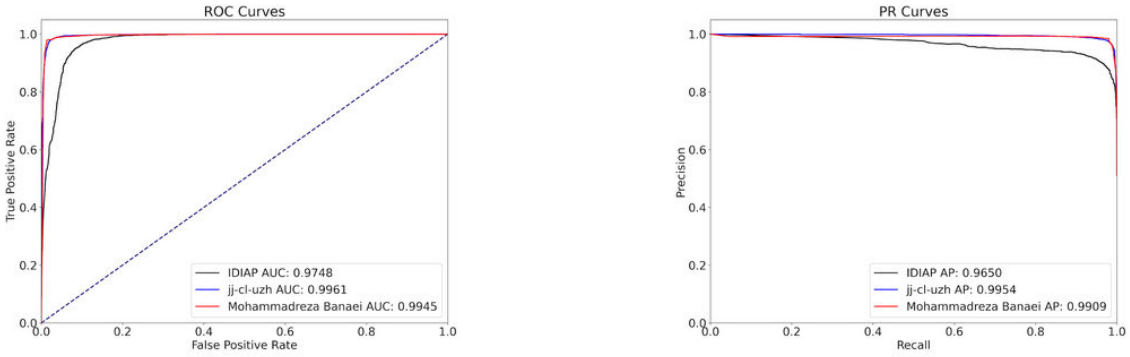


Figure 3: Official results announced by the organizers displaying team's performance (ROC, PR curves).

verify the performance of our model under a more balanced data type scenario.

Acknowledgments

The work was supported by an innovation project (under an InnoSuisse grant) oriented to improve the automatic speech recognition and natural language understanding technologies for German. Title: "SM2: Extracting Semantic Meaning from Spoken Material" funding application no. 29814.1 IP-ICT and EU H2020 project "Real-time network, text, and speaker analytics for combating organized crime" (ROXANNE), grant agreement: 833635. The second author, Esaú Villatoro-Tello is supported partially by Idiap, UAM-C Mexico, and SNI-CONACyT Mexico during the elaboration of this work.

References

Ivana Balazevic, Mikio Braun, and Klaus-Robert Müller. 2016. Language detection for short text messages in social media. *arXiv preprint arXiv:1608.08515*.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb):281–305.

Nora Hollenstein and Noëmi Aepli. 2015. A resource for natural language processing of swiss german dialects.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2019a. Language model adaptation for language and dialect identification of text. *Natural Language Engineering*, 25(5):561–583.

Tommi Sakari Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019b. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Tom Kocmi and Ondřej Bojar. 2017. Lanidenn: Multilingual language identification on character window. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 927–936.

Artur Kulmizev, Bo Blankers, Johannes Bjerva, Malvina Nissim, Gertjan van Noord, Barbara Plank, and Martijn Wieling. 2017. The power of character n-grams in native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 382–389.

Lei Le, Andrew Patterson, and Martha White. 2018. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In *Advances in Neural Information Processing Systems*, pages 107–117.

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Musat, and Andreas Fischer. 2019. Automatic creation of text corpora for low-resource languages from the internet: The case of swiss german.

Marco Lui, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40.

Ermelinda Oro, Massimo Ruffolo, and Mostafa Sheikhalishahi. 2018. Language identification of similar languages using recurrent neural networks. In *ICAART*.

Marc'Aurelio Ranzato and Martin Szummer. 2008. Semi-supervised learning of compact document representations with deep networks. In *Proceedings of the 25th international conference on Machine learning*, pages 792–799.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. In *Advances in neural information processing systems*, pages 3546–3554.

Fernando Sánchez-Vega, Esaú Villatoro-Tello, Manuel Montes-y Gómez, Paolo Rosso, Efstathios Stamatatos, and Luis Villaseñor-Pineda. 2019. Paraphrase plagiarism identification with character-level features. *Pattern Analysis and Applications*, 22(2):669–681.

Yves Scherrer and Owen Rambow. 2010. Natural language processing for the swiss german dialect area. In *Semantic Approaches in Natural Language Processing-Proceedings of the Conference on Natural Language Processing 2010 (KONVENS)*, pages 93–102. Universaar.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.

Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. 2015. Scalable bayesian optimization using deep neural networks. In *International conference on machine learning*, pages 2171–2180.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014a. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014b. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

Zhihua Wei, Duoqian Miao, Jean-Hugues Chauchat, Rui Zhao, and Wen Li. 2009. N-grams based feature selection and text representation for chinese text classification. *International Journal of Computational Intelligence Systems*, 2(4):365–374.

Jennifer Williams and Charlie Dagli. 2017. Twitter language identification of similar languages and dialects without ground truth. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 73–83.

Qiuyu Zhu and Ruixin Zhang. 2019. A classification supervised auto-encoder based on predefined evenly-distributed class centroids. *arXiv preprint arXiv:1902.00220*.

GERMEVAL 2020 SHARED TASK ON THE CLASSIFICATION AND REGRESSION OF COGNITIVE AND MOTIVATIONAL STYLE FROM TEXT

Important: We with all institutions involved acknowledge that the boundary between science and pseudo-science is contested and highly politicized when the science of IQ is involved.
Please notice the section on the aptitude test and criticism or visit the provided link on that topic here under 'Aptitude test and criticism'.

Show content

Announcements

General Information

The validity of high school grades as a predictor of academic development is controversial. Researchers have found indications that linguistic features such as function words used in a prospective student's writing perform better in predicting academic development (Pennebaker et al., 2014) than other methods such as GPA values.

During an aptitude test, participants are asked to write freely associated texts to provided questions, regarding shown images. Psychologists can identify so-called implicit motives from those expressions. Implicit motives are unconscious motives, which are measurable by operant methods. Psychometrics are metrics, which can be utilized for assessing psychological phenomena. One flawed but well-known example are the infamous ink dots, which ought to be described. Operant methods, in turn, are psychometrics, which is collected by having participants write free texts (Johannßen et. al, 2019). Those motives are said to be predictors of behavior and long-term development from those expressions (McClelland, 1988, Scheffer 2004, Schultheiss, 2008).

From a small sample of an aptitude test collected at a college in Germany, the classification and regression of cognitive and motivational styles from a text can be investigated. Such an approach would extend sole text classification and could reveal insightful psychological traits.

Operant motives are unconscious intrinsic desires that can be measured by implicit or operant methods, such as the Operant Motive Test (OMT) or the Motive Index (MIX). Psychologists label these textual answers with one of five motives (M - power, A - affiliation, L - achievement, F - freedom, 0 - zero) and corresponding levels (0 to 5), which roughly describe the emotional mood from positive to negative. The identified motives allow psychologists to predict behavior and longterm development. For our task, we provide extensive amounts of textual data from both, the OMT and MIX, paired with IQ and high school grades and labels.

With this task, we aim to foster novel research within the context of NLP and the psychology of personality and emotion. This task is focusing on utilizing German psychological text data for researching the connection of text to cognitive and motivational style. For this, contestants are asked to build systems to restore an artificial 'rank' as well as performing classification on an image description that psychologists can investigate on implicit motives.

The shared task is organized by Dirk Johannßen, Chris Biemann, Steffen Remus, and Timo Baumann from the Language Technology group of the University of Hamburg (Germany), as well as David Scheffer from the NORDAKADEMIE Elmshorn (Germany).

Proceedings

GermEval 2020 Task 1: Classification and Regression of Cognitive and Motivational Style from Text
Dirk Johannßen, Chris Biemann, Steffen Remus, Timo Baumann, and David Scheffer
Idiap & UAM participation at GermEval 2020: Classification and Regression of Cognitive and Motivational Style from Text
Esáu Villatoro-Tello, Shantipriya Parida, Sajit Kumar, Petr Motlicek, and Qingran Zhan
Predicting Cognitive and Motivational Style from German Text using Multilingual Transformer Architecture
Henning Schäfer, Ahmad Idrissi-Yaghir, Andreas Schimanowski, Michael Raphael Bujotzek, Hendrik Damm, Jannis Nagel, and Christoph M. Friedrich
Predicting Educational Achievement Using Linear Models
Çağrı Çöltekin
Ethical considerations of the GermEval20 Task 1. IQ assessment with natural language processing: Forbidden research or gain of knowledge?
Dirk Johannßen, Chris Biemann, and David Scheffer

Please visit the official SWISSTEXT YouTube channel with the SWISSTEXT + KONVENS playlist, which includes the GermEval20 Task 1 video presentations:
https://www.youtube.com/channel/UCGsc1P6JWvWBxHwWPvnQ43A/playlists

The Aptitude Test, college and criticism

Since 2011, the private university of applied sciences NORDAKADEMIE performs an aptitude college application test.

Zimmerhofer and Trost (2008, p. 32ff.) describe the developments of the German Higher Education Act. A so-called Numerus Clausus (NC) Act from 1976 and 1977 ruled that colleges in Germany with a significant amount of applications have to employ a form of selection mechanism. For most colleges, the NC was the threshold for many applicants. Even though this value is more complex, it roughly can be understood as a GPA threshold. Since this second Higher Education Act, colleges are also free to employ alternate selection forms, as long as they are scientifically sound, transparent and commonly accepted in Germany (BVerfGE 43, 291 - numerus clausus II).

Even though Hell, Trapmann und Schuler (2008, p. 46) found the correlation coefficient of high school grades of r = 0.517 to be the most applicable measure for academic suitability, criticism emerged as well. The authors criticized that the measure of grades by just one single institution (i.e. a high school) does not reflect upon the complexity of such a widely questioned concept of intellectual ability. Schleithoff (2015, p. 6) researched the high school grade development of different German federal states on the issue of grade inflation in Germany and found evidence, that supports this claim. Furthermore, in most parts of Germany, the participation grade makes up 60% of the overall given grade and thus is highly subjective.

Since operant motives are said to be less prone to subjectivity, the NORDAKADEMIE decided to employ an assessment center (AC) for research purposes and a closely related aptitude test for the application procedure (NORDAKADEMIE, 2018). Rather than filtering the best applicants, the NORDAKADEMIE aims with the test for finding and protecting applicants that they suspect to not match the necessary skills required at the college (Sommer, 2012). Thus, every part of the aptitude test is skill-oriented.

During an hour-long aptitude test, participants are asked to write freely associated texts to provided questions and images. Those motives are said to be predictors of behavior and long-term development from those expressions. This test contains multiple parts, e.g. a math- and an English test, Kahnemann scores, intelligence quotient (IQ) scores, a visual questionnaire, knowledge questions to the applied major or the implicit motives, called the Motive Index (MIX).

The MIX measures implicit or operant motives by having participants answer questions to those images like the one displayed on the Tasks tab such as "who is the main person and what is important for that person?" and "what is that person feeling". Furthermore, those participants answer the question of what motivated them to apply for the NORDAKADEMIE.

Even though parts of this test are questionable and are currently under discussion, no single part of this test leads to an application being rejected. Only when a significant amount of those test parts

Furthermore, every applicant has the option to decline the data to be utilized for research purposes and still can apply to study at the Nordakademie. All anonymized data instances emerged from college applicants that consented for the data to be utilized in this type of research setting and have the opportunity to see any stored data or to have their personal data deleted at any given moment (e.g. sex, age, the field of study).

Any research performed on this aptitude test or the annually conducted assessment center (AC) at the NORDAKADEMIE is under the premise of researching methods of supporting personnel decision-makers, but never to create fully automated, stand-alone filters (NORDAKADEMIE, 2019). First of all, since models might always be flawed and could inherit biases, it would be highly unethical. Secondly, the German law prohibits the use of any - technical or non-technical - decision or filter system, which can not be fully and transparently be explained. Aptitude diagnostics in Germany are highly legally regulated.

The most debated upon the part of the aptitude test is the intelligence quotient (IQ). Intelligence in psychology is understood as results measured by an intelligence test (and thus not the intelligence of individuals itself). Furthermore, intelligence is always a product of both, genes and the environment. Even though there are hints that the IQ does not measure intellectual ability but rather cognitive and motivational style (DeYoung, 2011), it is defined and broadly understood as such.

Mainly companies in Europe employ IQ tests for selecting capable applicants. In the United Kingdom, roughly 69 percent of all companies utilize the IQ. In Germany, the estimate is 13 percent (Nachtwei & Schermuly, 2009).

Since IQ tests only measure the performance in certain tasks that rather ask for skill in certain areas (logics, language, problem-solving) than cognitive performance, such intelligence tests should rather be called comprehension tests. Minorities can be discriminated by a biased due to unequal environmental circumstances and measurements in non-representative groups (Rushton, & Jensen, 2005). One result of research on the connection between implicit motives and intelligence testing could help to improve early development and guided support.

It is this bias, which leads to unequal opportunities especially in countries where there is a rich diversity among the population. Intelligence testing has had a dark history. Eugenics during the great wars e.g. in the US by sterilizing citizens (Buck v. Bell) or in Germany during the Third Reich are some of the most gruesome parts of history.

But even in modern days, the IQ is misused. Recently, IQ scores have been used in the US to determine which death row inmate shall be executed and which might be spared. Since IQ scores show a too large variance, the Supreme Court has ruled against this definite threshold of 70 (Hall v. Florida). However, Sanger (2015) has researched an even more present practice of 'racial adjustment', adjusting the IQ of minorities upwards to take countermeasures on the racial bias in IQ testing, resulting in death row inmates, which originally were below the 70 points threshold, to be executed.

There is an ethical necessity to carefully view, understand and research the way intelligence testing is conducted and how those scores are - if at all - correlated with what we understand as 'intelligence', as they might be mere cognitive and motivational styles. Further valuable research can be conducted to investigate connections between other personality tests such as implicit motives with intelligence or comprehension tests. Racial biases are measurable, variances are great and many critics state that IQ scores reflect upon skill or cognitive and motivational style rather than real intelligence as it is broadly understood.

Regarding commercial interests: While of course there is interest from the people that provide this data, we find it remarkable that the data is made available freely. We aim to share the data with the international scientific community, to better understand and learn from the data and discuss interesting findings publicly, for the benefit of everyone. Note that this is the entire data that currently exists, not a sub-sample, so it likewise supports the commercial interests of competitors. Furthermore, professors at Universities for Applied Sciences in Germany (especially private colleges) are supposed to work in the private industry on their specific research field (Wikipedia, 2019). Thus, an alleged conflict of interest is a result of the educational system in Germany. The interests of the task organizers are strictly scientific. There is no funding for this task, neither from the public nor from commercial sources.

FAQs

Is this task about building automated filters?

How does the aptitude test lead to being rejected?

Do the organizers have a commercial interest?

Some organizers work in the private sector. Do we work for them when we participate in the task?

Why would you ever want to build a system for classifying or regressing psychological traits? What is the purpose of this task?

Intelligene structure tests (i.e. IQ) are said to have a racial bias. Why not simply use high school grades alone?

Can resulting systems be used for predicting school grades or IQ scores on any text?

Evaluation

System submissions are done in teams. There is no restriction on the number of people in a team. However, keep into consideration that a participant is allowed to be in multiple teams, so splitting up into teams with overlapping members is a possibility. Every participating team is allowed to submit 3 different systems to the competition. For submission in the final evaluation phase, every team must name their submission (.zip and the actual submission .txt file) in the form "[Teamname]__[Systemname]" (note the two underscores!). Important: Please do not include more than one double underscore in the naming scheme. E.g. your submission could look like

```
Funtastic4__SVM_ensemble1.zip
|
+-- Funtastic4__SVR_TF_IDF_ensemble1_task1.txt
```

Or

```
Funtastic4__SVM_ensemble1.zip
|
+-- Funtastic4__SVC_TF_IDF_ensemble1_task2.txt
```

We also ask you to put exactly this name into the description before submitting your system. This identification method is needed to correctly associate each submitted system with its description paper. Thus, please make sure to write the name exactly as it will appear in your description paper (i.e. case sensitive). If your submission does not follow these rules it might not be evaluated. The evaluation script has been adopted for a formality check.

Only the person who submits is required to register for the competition. All team members need to be stated in the description paper of the submitted system. The last submission of a system will be used for the final evaluation. Participants will see whether the submission succeeds, however, there will be no feedback regarding the score. The leaderboard will thus be disabled during the test phase.

The evaluation script is provided with the data so that participants can still evaluate their data splits. The zip located at the Data section contains the evluate.py program among other files. If you use the standalone functionality of this file, you need to call it as:

```
python evaluate.py <input_dir> <output_dir>
```

The submission files have to comply with the tab-separated format as follows for Subtask 1, reproducing the target rank (as averaged z-standardized scores of a participant) relative to all participants in a collection (i.e. test / dev / train):

```
student_ID rank
```

and for Task 2:

```
UUID motive level
```

On task1, the script computes multiple correlation coefficients. The Pearson rank correlation coefficient will be the main evaluation metric. On task2, the script computes for each class precision, recall and F1 score. As a summarizing score, the tool computes accuracy and macro-average precision, recall and F1 score.

Although the evaluation tool outputs several evaluation measures, the official ranking of the systems will be based on the macro-average F1 score for task2 or the Pearson correlation coefficient for task1 only. Please remember this when tuning your classifiers. A classifier that is optimized for the accuracy or the Spearman correlation coefficient may not necessarily produce optimal results in terms of the macro-average F1 score.
The evaluation tool on Codalab and the download versions is the same and accepts both tasks simultaneously.

System submissions are done in teams. There is no restriction on the number of people in a team. However, keep into consideration that a participant is allowed to be in multiple teams, so splitting up into teams with overlapping members is a possibility. Every participating team is allowed to submit as many different systems to the competition as they wish.

Subtasks

The shared task on classification and regression of cognitive and motivational style of text consists of two subtasks, described below. You can participate in any of them, may use external data and/or utilize the other data respectively for training, as well as perform e.g. multi-task or transfer learning. Both tasks are closely related to the main research objective: implicit motives. Those motives are said to describe the intrinsic desires of students and allow for psychologists to, after identifying those motives, make statements on long-term behavior and development. For this first task, the so-called Motive Index (MIX) texts are the basis for classifying cognitive and motivational style. For the second task, so-called Operant Motives (OMT), which are implicit motives as well, can be classified into main motives and so-called levels, describing the emotional exertion expressed.

We encourage every participant to also include ethical positions and discussions in their system descriptions that can be the basis for an insightful and reflected podium discussion during the workshop session at GermEval 2020.

Subtask 1: Regression of artificially ranked cognitive and motivational style

This task has yet never been researched and is open: It is neither certain, whether this task can be achieved, nor how well this might be possible due to the novelty and sparsity of research.

The task is to predict measures of cognitive and motivational style solemnly based on text. For this, z-standardized high school grades and intelligence quotient (IQ) scores of college applicants are summed and globally 'ranked'. This rank is utterly artificial, as no applicant in a real-world-setting is ordered in such fashion but rather there is a certain threshold over the whole of the hour-long aptitude test with multiple different test parts, that may not be undergone by applicants. Only about 10% of initial applicants get declined and may not proceed to a second step, the application at a private company. The resulting system would be of no real-world use as those motive texts still ought to be collected and strict European data protection laws prohibit any use of unexplainable, intransparent aptitude systems (Sommer, 2012 and NORDAKADEMIE b, 2019).

The goal of this subtask is to reproduce this 'ranking', systems are evaluated by the Pearson correlation coefficient between system and gold ranking. An exemplary illustration can be found in the Data area. We are especially interested in the analysis of possible connections between text and cognitive and motivational style, which would enhance later submission beyond the mere score reproduction abilities of a submitted system.

One z-standardized example instance looks as follows (including spelling errors made by the participant) with the unique ID (consisting of studentID_imageNo_questionNo), a student ID, an image number, an answer number, the German grade points, the English grade points, the math grade points, the language IQ score, the math IQ score and the average IQ score (all z-standardized)

The data is delivered in two files, one containing participant data, the other containing sample data, each being connected by a student ID. The rank in the sample data reflects the averaged performance relative to all instances within the collection (i.e. within train / test / dev), which is to be reproduced for the task.

```
student_ID german_grade english_grade math_grade lang_iq logic_iq
1034-875791 -0.08651999119820285 0.3747985587188588 0.5115559707967757
-0.010173719700624676 -0.13686707618782515

student_ID rank
1034-875791 15
```

The training data set contains 80% of all available data, which is 62,280 expressions and the development and test sets contain roughly 10% each, which are 7,800 expressions for the dev set and 7,770 expressions for the test set (this split has been chosen in order to preserve the order and completeness of the 30 answers per participant).

For the final results, participants of this shared task will be provided with a MIX_text only and are asked to reproduce the ranking of each student relative to all students in a collection (i.e. within the test set).

The success will be measured with the pearson rank correlation coefficient.

Subtask 2: Classification of the Operant Motive Test (OMT).

Operant motives are unconscious intrinsic desires that can be measured by implicit or operant methods, such as the Operant Motive Test (OMT)(Kuhl and Scheffer, 1999). During the OMT, participants are asked to write freely associated texts to provided questions and images. An exemplary illustration can be found under the Data tab. Psychologists label these textual answers with one of five motives. The identified motives allow psychologists to predict behavior and long-term development.

For this task, we provide the participants with a large dataset of labeled textual data, which emerged from an operant motive test. The training data set contains 80% of all available data (167,200 instances) and the development and test sets contain 10% each (20,900 instances)

```
UUID OMT_text
6221323283933528M10 Sie wird ausgeschimpft, will jedoch das Gesicht
bewahren.Beleidigt.Weil sie sich schämt, ausgeschimpft zu werden.
Die blaue Person ist verletzt und hört nicht auf die Worte der weißen Person.

UUID motive level
6221323283933528M10 F 5
```

The success will be measured with the macro-averaged F1-score.

Data

Development data sets, example systems and first evaluation script

NORDAKADEMIE Aptitude Data Set

Since 2011, the private university of applied sciences NORDAKADEMIE performs an aptitude college application test, where participants state their high school performance, perform an IQ test and a psychometrical test called the Motive Index (MIX). The MIX measures so-called implicit or operant motives by having participants answer questions to those images like the one displayed below such as "who is the main person and what is important for that person?" and "what is that person feeling". Furthermore, those participants answer the question of what motivated them to apply for the NORDAKADEMIE.

The data consists of a unique ID per entry, one ID per participant, of the applicants' major and high school grades as well as IQ scores with one textual expression attached to each entry. high school grades and IQ scores are z-standardized for privacy protection.

In total there are 2,595 participants, who produced 77,850 unique MIX answers. The shortest textual answers consist of 3 words, the longest of 42 and on average there are roughly 15 words per textual answer with a standard deviation of 8 words. The (not z-standardized) average grades and IQ scores are as follows:

German: 9.4 points
English: 9.5 points
math: 10.1 points

IQ language: 118 points
IQ logic: 72.6 points
IQ averaged: 77 points

The IQ language measures the use of language and intuition such as the comprehension of sayings. IQ logic tests the relations of objects and an intuitive understanding of mainly verbalized truth systems. The averaged IQ includes IQ language and logic as well as further IQ tests (i.e. language, logic, calculus, technology and memorization). To enhance data protection, all provided high school grades and IQ scores are z-standardized.

To enhance data protection, all provided high school grades and IQ scores are z-standardized.

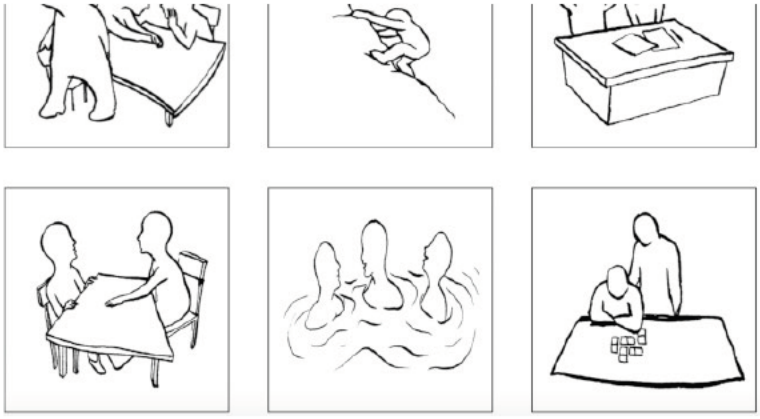
```
student_ID image_no answer_no UUID MIX_text
1034-875791 2 2 1034-875791_2_2 Die Person fühl sich eingebunden in
die Unterhatung.

student_ID german_grade english_grade math_grade lang_iq logic_iq
1034-875791 -0.08651999119820285 0.3747985587188588 0.5115559707967757
-0.010173719700624676 -0.13686707618782515

student_ID rank
1034-875791 15
```

Operant Motive Test (OMT)

The available data set has been collected and hand-labeled by researchers of the University of Trier. More than 14,600 volunteers participated in answering questions to 15 provided images such as displayed in the figure below.

The pairwise annotator intraclass correlation was r = .85 on the Winter scale (Winter, 1994).

The length of the answers ranges from 4 to 79 words with a mean length of 22 words and a standard deviation of roughly 12 words.



Some example answers to the very first image above are as follows (with A being the so-called affiliation motive and M being the power motive, two out of the five motives besides L for achievement, F for freedom and 0 for the zero / unassigned motive):

```
A sie nimmt am Gespräch nicht teil und wendet sich ab. gelangweilt. es interessiert sie nicht, worüber die anderen beiden reden. schlecht.

M weicht ängstlich zuruück. unterlegen. wird zurechtgewiesen. Gelegenheit den Fehler zu korrigieren
```

(Translation: A she does not take part in the conversation and turns away. bored. She does not care what the other two are talking about. Bad. M withdraws anxiously. Inferior. is rebuked. Opportunity to correct the mistake.)

The number of motives in the available data is unbalanced with power (M) being by far the most frequent with 54.5%, achievement (L) constituting 19% of the data, affiliation (A) 17%, freedom (F) 5.6% and zero 5%.

For each instance there is a unique ID, the expressed textually answers with a label for the main motive and a level. The data structure of the whole OMT data set looks as follows and is tab-separated:

```
UUID OMT_text
6221323283933528M10 Sie wird ausgeschimpft, will jedoch das Gesicht bewahren.Beleidigt.Weil sie sich schämt, ausgeschimpft zu werden. Die blaue Person ist verletzt und hört nicht auf die Worte der weißen Person.

UUID motive level
6221323283933528M10 F 5
```

System Description Paper Author Guidelines

Please use the LaTeX template provided by the main conference under https://swisstext-and-konvens-2020.org/call-for-papers/
Language: English
All submissions must be in PDF format and must conform to the official style guidelines, which are contained in the template files that are available above.
The decision on paper acceptance will be based on the feedback from the reviewers.
The review process will be single-blind, i.e. authors are allowed to enter information that might reveal their identity.
Accepted system description papers will appear in an online workshop proceeding.
Manuscripts must describe original work that has neither been published before nor is currently under review elsewhere.
Submission will be made through EasyChair: https://easychair.org/conferences/?conf=gest201
A draft of the description paper can be found here.

The shared task is organized by Dirk Johannßen, Chris Biemann, Steffen Remus and Timo Baumann from the Language Technology group of the University of Hamburg, as well as David Scheffer from the NORDAKADEMIE Elmshorn, Nicola Baumann from the Universität Trier and the Gudula Ritz from the Impart GmbH (Germany).

Important Dates

~~01 Dec 2019: Release of trial data~~
~~01 Jan 2020: Release of training data (train + validation)~~
~~08 May 2020: Release of test data~~
~~01 Jun 2020~~ ~~05 Jun 2020~~: Final submission of test results
~~03 Jun 2020~~ ~~09 Jun 2020.~~ Submission of description paper
~~04 11 Jun 2020~~ ~~14 Jun 2020~~: Peer reviewing: participants are expected to review other participant's system descriptions
~~12 Jun 2020~~ ~~15 Jun 2020~~: Notification of acceptance and reviewer feedback
~~18 Jun 2020~~ ~~20 Jun 2020~~: Camera-ready deadline for system description papers
~~23 Jun 2020:~~ Workshop in Zurich, Switzerland at the KONVENS 2020 and SwissText joint conference

People

Dirk Johannßen
Chris Biemann
Steffen Remus
Timo Baumann
David Scheffer

Terms and Conditions

The copyright to the provided data belongs to the NORDAKADEMIE and for the OMT related tasks to the University of Trier and Impart GmbH, its licensors, vendors and/or its content providers. The scores and instances serve promotional/public purposes and permission has been granted by the NORDAKADEMIE and the University of Trier, which both share this dataset. This dataset is redistributed under the creative commons license CC BY-NC-SA 4.0.

By participating at this competition, you consent the public release of your anonymized scores at the GermEval-2020 workshop and in respective proceedings, at the task organizers' discretion.

References

BVerfGE. (1977). BVerfGE. 43, 291 - numerus clausus II. URL http://www.servat.unibe.ch/dfr/bv043291.html [12/08/2019]
DeYoung, C. G. (2011). Intelligence and personality. In R. J. Sternberg & S. B. Kaufman (Eds.), Cambridge handbooks in psychology. The Cambridge handbook of intelligence (p. 711-737). Cambridge University Press. https://doi.org/10.1017/CBO9780511977244.036 [12/08/2019]
Hall v. Florida. Docket number 12-10882. SCOTUSblog. 27 May 2014. Retrieved 29 May 2014.
Hell, Benedikt, Sabrina Trapmann und Heinz Schuler (2007). Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum. In: Jahrgang 21 (Heft 3), S. 251-270.
Johannßen, D., Biemann, C. and Scheffer, D. (2019): Reviving a psychometric measure: Classification and prediction of the Operant Motive Test. Proceedings of CLPsych 2019, Minneapolis, MN, USA.
Kuhl, Julius and Scheffer, David. 1999. Der operante Multi-Motiv-Test (OMT): Manual [The operant multi-motive-test (OMT): Manual]. Impart, Osnabrück, Germany: University of Osnabrück.
McClelland, David C. 1988. Human Motivation. Cambridge University Press. Nachtwei, Jens & Schermuly, Carsten. (2009). Acht Mythen über Eignungstests. Harvard Business Manager. 09. 6-10.
NORDAKADEMIE. (2018). Campus Forum Nr. 66/Juni 2018. P. 8. Assessment Center an der NORDAKADEMIE. [online] https://www.nordakademie.de/sites/default/files/2019-08/CF_66_final.pdf URL. [12/08/2019].
NORDAKADEMIE b. (2019). Datenschutzbestimmungen. [online] Available at: https://auswahltest.nordakademie.de/datenschutz URL [12/08/2019]
NORDAKADEMIE. (2019). Digitale Unterstützung für Personaler - Mitarbeitende finden mithilfe von Künstlicher Intelligenz. [online] Available at: https://www.nordakademie.de/news/digitale-unterstuetzung-fuer-personaler-mitarbeitende-finden-mithilfe-von-kuenstlicher [12/08/2019].
Pennebaker, James W. , Chung, C. K., Frazee, J., Lavergne, G. M., and Beaver, D. I., When small words foretell academic success: The case of college admissions essays, PLOS ONE, vol. 9, no. 12, e115844,

Psychology, Public Policy, and Law, 11(2), 235-294. https://doi.org/10.1037/1076-8971.11.2.235
Sanger, Robert M., IQ, Intelligence Tests, 'Ethnic Adjustments' and Atkins (November 21, 2015). American University Law Review, Vol. 65, No. 1, 2015. Available at SSRN: https://ssrn.com/abstract=2706800
Scheffer, David. 2004. Implizite Motive: Entwicklung, Struktur und Messung [Implicit Motives: Development, Structure and Measurement]. Hogrefe Verlag, Go Î^ttingen, Germany, 1st edition.
Schleithoff, Fabian (2015). Noteninflation im deutschen Schulsystem - Macht das Abitur hochschulreif. In: ORDO - Jahrbuch für die Ordnung von Wirtschaft und Gesellschaft. Bd. 66. De Gruyter Oldenbourg, S. 3-26. ISBN: 978-3-8282-0621-2.
Schultheiss, O. C. (2008). Implicit motives. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), Handbook of personality: Theory and research (p. 603-633). The Guilford Press.
Sommer, Kristina. (2012). Erst testen, dann bewerben. [online] Available at: https://idw-online.de/de/news492748 URL [12/08/2019].
Wikipedia. 2019. Fachhochschule. [online] Available at: https://en.wikipedia.org/wiki/Fachhochschule URL [12/08/2019].
Winter, David. 1994. Manual for scoring motive imagery in running text. Dept. of Psychology, University of Michigan (unpublished).
Zimmerhofer, Alexander und Günter Trost (2008). Auswahl- und Feststellungsverfahren in Deutschland - Vergangeheit, Gegenwart und Zukunft. In: Studierendenauswahl und Studienentscheidung. 1., Aufl. Hogrefe Verlag, S. 32-42.

NEWS

UHH | 23 OCTOBER 2020

NLP Master project entitled „BigBlueButton video conference transcription and keyword lookup extension" from LT group won EXPO 2020

Congratulations to Fabian Rausch and Felix Welter

Fabian Rausch and Felix...

UHH | 1 OCTOBER 2020

Two papers accepted at COLING 2020

The 28th International Conference on Computational Linguistics (COLING 2020)...

UHH | 2 SEPTEMBER 2020

Paper accepted at ISWC 2020

The 19th International Semantic Web Conference (ISWC 2020) accepted the...

VERWALTUNG | 7 AUGUST 2020

A paper accepted at INTERSPEECH 2020 main conference

The 21th Conference of the International Speech Communication...

UHH | 4 APRIL 2020

Paper accepted at ACL 2020

ACL 2020, the 58th Annual Meeting of the Association for Computational...

VERWALTUNG | 25 MARCH 2020

LT Group Teams rank 1st and 2nd at SemEval-2020 Task "Multilingual Offensive Language Identification in Social Media"

In this year's SemEval Shared Task 12 (OffensEval 2020) on "Multilingual...

Our application for a collaboration project betwen literary science and...

UHH | 30 NOVEMBER 2019

2nd Phase of Transregio-SFB "Crossmodal Learning"

We happily announce that the second phase of the 2nd phase of Transregio-SFB...

UHH | 14 OCTOBER 2019

Rami Aly was honored with the GSCL best thesis award

Every two years the GSCL (German Society of Computational Linguists) awards...

UHH | 1 OCTOBER 2019

New publication in LREV journal

A new publication from LT group member Gregor Wiedemann has appeared in...

> ARCHIVE OF CURRENT NEWS

Idiap & UAM participation at GermEval 2020: Classification and Regression of Cognitive and Motivational Style from Text

Esaú Villatoro-Tello[1,2], **Shantipriya Parida**[2], **Sajit Kumar**[3], **Petr Motlicek**[2] **and Qingran Zhan**[2]

[1]Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Mexico City, Mexico.
evillatoro@correo.cua.uam.mx
[2]Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland.
firstname.lastname@idiap.ch
[3]Centre of Excellence in AI, Indian Institute of Technology, Kharagpur, West Bengal, India.
kumar.sajit.sk@gmail.com

Abstract

In this paper, we describe the participation of the Idiap Research Institute at GermEval 2020 shared task on the Classification and Regression of Cognitive and Motivational style from Text, specifically on subtask 2, Classification of the Operant Motive Test (OMT). Generally speaking, GermEval 2020 aims at encouraging the Natural Language Understanding (NLU) research community in proposing novel methodologies for assessing the connection between freely written texts and its cognitive and motivational styles. For evaluating this task, organizers provided a large dataset containing textual descriptions, in German language, generated by more than 14,000 participants. Our participation aims at evaluating the impact of advanced language representation, e.g., Bert, XLM, and DistilBERT in combination with some traditional machine learning algorithms. Our best configuration was able to obtain an F1 macro of 69.8% on the test partition, which represents a relative improvement of 7.4% in comparison to the proposed baseline.

1 Introduction

The idea that language use reveals information about personality has long circulated in the social and medical sciences. The ways people use words convey a great deal of information about themselves (Pennebaker et al., 2003). Psycholinguistics theory has shown the presence of linguistic indicators that could be important for determining aptitudes and academic development in subjects (Pennebaker et al., 2014), however, many of these research has focused on the analysis of self-reports or essays.

In contrast, implicit motives, indicators used by psychologies during aptitude diagnosis, are not readily accessible features to the conscious mind and, therefore, not assessable using self-reports of personal needs (Gawronski and De Houwer, 2014). Instead, implicit motives are primarily assessed using indirect measures that rely on projective techniques that instruct individuals to produce imaginative stories based on ambiguous picture stimuli that depict people in different situations. Such stimuli influence the content of the individual's fantasy and are projected onto the characters of the stories which the individual writes about from these pictures (Johannßen and Biemann, 2018, 2019; Johannßen et al., 2019). Consequently, this motivational response emerges through the contents of the written imaginative material and can be coded for its motive imagery using standardized and validated content coding systems.

The most frequently used measures of implicit motives are Picture Story Exercise (PSE) (McClelland et al., 1989), Thematic Apperception Test (TAT) (Murray, 1943), Multi-Motive Grid (MMG) (Sokolowski et al., 2000), and Operant Motive Test (OMT) (Kuhl and Scheffer, 1999; Denzinger and Brandstätter, 2018). Generally speaking, these tests are based on the operant methods, i.e., participants are asked ambiguous questions or are shown simple images, which they have to describe. Specifically, during the OMT test, subjects are shown sketched scenarios with multiple persons in non-specified situations, which required to use introspection and assess their psychological attributes unconsciously. Psychologists label these textual answers with one of five motives, namely M-power, A-affiliation, L-achievement, F-freedom,

and 0-zero. And, each motive is associated with its corresponding level (from 0 to 5).

Accordingly, the "GermEval 2020 Task on the Classification and Regression of Cognitive and Emotional Style from Text",[1] shared subtask 2, proposes an exploratory task on the Classification of the Operant Motive Test (OMT). The challenge consists of automatically processing pieces of text, generated by undergraduate students during an OMT test, and to correctly detect subjects corresponding motive/level combination.

To address the OMT task, we evaluate the impact of deep learning architectures such as Transformers (Wolf et al., 2019), namely Bert (Devlin et al., 2019), XLM (Conneau and Lample, 2019), DistilBert (Sanh et al., 2019). We compare its performance against traditional classification methods, e.g., fully connected neural networks. We compared the efficiency of these recent methodologies and compare them under different configuration parameters. Our results indicate that performing a fine-tuning of Bert is possible to obtain a 7.4% relative improvement in comparison to the proposed baseline, and the 2nd place overall during GermEval 2020 edition.

The rest of the paper is organized as follows. Section 2 describes the dataset and provides some statistics. The details of our methodology are provided in Section 3. Performed experiments and obtained results are shown in Section 3.2. Finally, we share the conclusion of our work in Section 5.

2 Dataset

To perform our experiments, we employed the dataset available in the GermEval 2020 shared task on the "Classification and Regression of Cognitive and Motivational style from the text", described in Johannßen et al. (2020). The provided data, in German language, has been collected from more 14,600 subjects that participated in the OMT test. Each answer was manually labeled with the motives (0, A, L, M, F) and the levels (from 0 to 5), resulting in a 30 class classification problem. This annotation was performed by an expert psychologist, trained by the OMT manual as described in (Kuhl and Scheffer, 1999). The distribution of the dataset is: 167,200 for training (*train*), 20,900 for development (*dev*), and 20,900 for testing (*test*).[2]

	Training	
	Average (σ)	*Total*
Tokens	20.27 ($\pm$12.08)	3,389,945
Vocabulary	18.07 ($\pm$9.82)	267,620
LR	0.92 ($\pm$ 0.08)	0.08
	Development	
	Average (σ)	*Total*
Tokens	20.38 ($\pm$12.17)	425,880
Vocabulary	18.17 ($\pm$9.94)	55,606
LR	0.92 ($\pm$ 0.08)	0.13
	Test	
	Average (σ)	*Total*
Tokens	20.24 ($\pm$12.01)	423,018
Vocabulary	18.05 ($\pm$9.76)	55,592
LR	0.92 ($\pm$0.08)	0.13

Table 1: Statistics of the OMT dataset in terms of number of tokens, vocabulary size and lexical richness. The minimum length of the texts is 1 token, while the maximum length is 99, 90, and 96 tokens for *train*, *dev*, and *test* partitions respectively. In all partitions, the 75% of the data has a length of 27 tokens.

Table 1 shows some statistics of the GermEval 2020 dataset, for *train*, *dev*, and *test* partitions. We compute the average number of tokens, vocabulary, and lexical richness of each text in the dataset. Lexical richness (LR), also known as "type/token ratio" is a value that indicates how the terms from the vocabulary are used within a text. LR is defined as the ratio between the vocabulary size and the number of tokens from a text ($LR = |V|/|T|$). Thus, a value close to 1 indicates a higher LR, which means vocabulary terms are used only once, while values near to 0 represent a higher number of tokens used more frequently (i.e., more repetitive).

Two main observations can be done at this point. On the one hand, notice that for the three partitions (i.e., *train, dev, and test*), textual descriptions are very short, on average 20 tokens with a vocabulary of 18 words, resulting in a very high LR (0.92). The high LR value means that very few words are repeated within each textual description, i.e., very few redundancies. On the other hand, globally speaking, the complete dataset has a low LR (0.08 for *train* and 0.13 for *dev* and *test*). Although these values are not directly comparable due to the size of each partition, they indicate, to some extent, that information across texts is very repetitive, i.e., simi-

[1] https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/germeval-2020-cognitive-motive.html

[2] During our experimentation a total of 13 instances were removed from the training partition due to its lack of label, leaving 167,187 instances.

lar types of words are being used by tested subjects for describing different images, even though they belong to different classes (motives and levels). We are aware of the necessity from a deeper analysis of the data in order to reach concrete conclusions about the nature of the texts; however, this initial analysis helped us to envision the complexity and nature of the data.

3 Methodology

We aim to automate the annotation of participant responses for the OMT task by training a machine learning model. Machine learning (ML) models as such cannot use raw text as input. Therefore it is necessary to transform the input to a feature representation understandable by the model. Accordingly, we evaluate two ML approaches for solving the OMT task: fine-tuning of transformers based architectures (Section 3.1), and a traditional fully-connected neural network (Section 3.2).

It is important to mention that instead of facing the OMT task as a 30 class classification problem, we split the problem into two separate classification tasks: motives (5 classes), and levels detection (6 classes). For each of classification problem, we applied the exact same methodology as described in the following sections. Finally, in order to produce the required output by the organizers, we merge the predicted motive and the predicted level for every instance.

3.1 Simple Transformer

The transformer model (Vaswani et al., 2017) introduces an architecture that is solely based on attention mechanism and does not use any recurrent networks but yet produces results superior in quality to Seq2Seq (Sutskever et al., 2014) models, incorporating the advantage of addressing the long term dependency problem found in Seq2Seq model.

For our experiments using Simple Transformers (ST) architectures, we setup three different configurations:

1. Bert (Devlin et al., 2019): we use a pre-trained model referred as `bert-base-german-cased`, with 12-layer, 768-hidden, 12-heads, 110M parameters.[3] The model is pre-trained on German Wikipedia dump (6GB of raw text files), the OpenLegalData dump (2.4 GB), and news articles (3.6 GB). We refer to this configuration as ST-Bert in our experiments.
2. XLM (Conneau and Lample, 2019): for this configuration we use a model with 6-layer, 1024-hidden, 8-heads, which is an English-German model trained on the concatenation of English and German Wikipedia documents (`bert-base-german-cased`). We refer to this configuration as ST-XLM in our experiments.
3. DistilBert (Sanh et al., 2019): fir this model we used a model with 6-layer, 768-hidden, 12-heads, 66M parameters (`distilbert-base-german-cased`). We refer to this configuration as ST-DistilBert in our experiments.

For all the previous configurations, in order to perform the fine-tuning of the ST architecture, we added an untrained layer of neurons on the end, and re-train the model for the OMT classification task. To perform these experiments, we used the Simple Transformers library which allows us to easily implement the proposed idea.[4] For all the experiments done using simple transformers architecture we set the `max_length` parameter to 90, and we re-trained the models up to 2 epochs. Further details of employed models can be found at huggingface web page.[5]

Hyper Parameter	Range
number of layers	3
number of hidden layers	1
nodes in hidden layer	16
activation function	ReLU

Table 2: Fully connected neural network configuration parameters.

3.2 Fully Connected Neural Network

As an additional classification method, we configured a fully connected neural network (FC). This type of artificial neural network is configured such that all the nodes, or neurons, in one layer, are connected to all neurons in the next layer. The network and configuration parameters are mentioned in Table 2.

For our performed experiments using FCs, we passed as input features to the FC the sentence rep-

[3] https://deepset.ai

[4] https://pypi.org/project/simpletransformers

[5] https://huggingface.co/transformers/pretrained_models.html

Method	*Configuration type*	*Configuration sub-type*	**F1-macro** *(dev)*	**F1-macro** *(test)*
ST	Bert	bert-base-german-cased	**0.694**	**0.698**
ST	XLM	xlm-mlm-ende-1024	0.688	0.686
ST	DistilBert	distilbert-base-german-cased	0.692	0.688
FC	Bert (pre-trained)	LHL	0.589	0.589
FC	Bert (pre-trained)	Concat4LHL	0.616	0.579
FC	Bert (fine-tuned)	LHL	0.673	0.671
FC	Bert (fine-tuned)	Concat4LHL	0.675	0.230
Baseline	SVM	*tf-idf*	0.639	0.644
1st place	–	–	–	0.704

Table 3: Obtained results on the *dev* and *test* partitions of the OMT classification task. Results are reported in terms of the F1 macro measure. Baseline and 1st place results were extracted from the companion paper (Johannßen et al., 2020).

resentation generated using Bert encoding. Thus, to generate the representation of the sentence, we evaluate several configurations, namely: last hidden layer (LHL), concatenation of the 4 last hidden layers (Concat4LHL), min, max and mean pool of the last hidden layers. However, we only report the best performances obtained during the validation stage, i.e., LHL and Concat4LHL configurations. On the one hand, for generating the Concat4LHL representation we concatenate the last four layers values from the token CLS. As known, the CLS token at the beginning of the sentence is treated as the sentence representation. On the other hand, for the LHL configuration, we preserve as the sentence representation the values of the last hidden layer from the token CLS.

For the reported experiments under the FC method, two configurations of Bert were tested for generating the LHL and Concat4LHL representation: i) pre-trained German encodings of Bert (`distilbert-base-german-cased`), referred as Bert(pre-trained); and ii) resultant fine-tuned Bert encodings from the re-training we explained in Section 3.1, referred as Bert(fine-tuned).

4 Experiments and Results

The results of each considered method are shown in Table 3. The proposed baseline by the GermEval 2020 organizers, is a linear Support Vector Classifier (SVC) using as a form of representation of the documents a traditional *tf-idf* strategy, specifically a 30 (combined motive/level labels) binary SVCs (one-vs-all) classifiers. Results are reported in terms of F1-macro, for both *dev* and *test* partitions. As can be observed in table 3, the proposed baseline obtains an F1=64.4%, representing a strong base method. During the competition, the best reported performance was an F1 macro of 70.4% (last row of Table 3).

During the validation stage, the best result using the FC method was obtained under the Concat4LHL configuration, i.e., when texts are represented using as features the concatenation of the four last hidden layers from 'Bert (fine-tuned)' model. However, notice that the same configuration obtained the worst performance during the test stage (23%). We think that some errors occurred during the setup of the output file, or at worst, maybe some error occurred during final training, provoking some overfitting situation. In spite of this result, the 'Bert fine-tuned' consistently improves the performance of the experiments using a fully connected neural network. Particularly, during the development stage, both experiments using the fine-tuned version of Bert outperformed the same configuration that uses the pre-trained version of Bert. Except for the FC(Bert pre-trained-Concat4LHL), a similar situation occurred during the test phase, i.e., adjusting the attention of Bert to the OMT task, helped the FC method for obtaining a more relevant results.

Finally, the best performance was obtained by the simple transformers architectures. As expected, the best performance is obtained when the Bert model is employed, followed by DistilBert and XML models. Generally speaking, ST-BERT configuration obtains a relative improvement of 7.4% over the competition baseline. Overall, the ob-

tained performance by the three considered configurations exhibits marginal differences, thus, the performance obtained by the DistilBert could be considered a very good alternative given that represents a significantly smaller, faster, cheaper and lighter transformer model.

5 Conclusion

This paper describes Idiap's participation at the GermEval 2020 shared task on the Classification and Regression of Cognitive and Motivational Style from the text. Our participation aimed at analyzing the performance of recent NLP technologies for solving the OMT classification task. To this end, we performed a comparative analysis among Simple Transformers based architectures, e.g., Bert, XLM, and DistilBert, and traditional machine learning techniques. Notably, transformers based methods exhibit the best empirical results, obtaining a relative improvement of 7.7% over the baseline suggested as part of the GermEval 2020 challenge. Overall, our system obtained the second-best place in terms of the F1 macro among participant teams during the GermEval 2020 edition.

As future work, we plan to evaluate the impact of hyperparameter tuning through optimization methods, such as Bayes optimizer (Snoek et al., 2012), and to perform further analysis on how the attention-mechanism from the transformers architecture is working in the OMT task.

Acknowledgments

The work was supported by an innovation project (under an InnoSuisse grant) oriented to improve the automatic speech recognition and natural language understanding technologies for German. Title: "SM2: Extracting Semantic Meaning from Spoken Material" funding application no. 29814.1 IP-ICT and EU H2020 project "Real-time network, text, and speaker analytics for combating organized crime" (ROXANNE), grant agreement: 833635. The first author, Esaú Villatoro-Tello is supported partially by Idiap, SNI-CONACyT, CONACyT project grant CB-2015-01-258588, and UAM-C Mexico during the elaboration of this work.

References

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

Ferdinand Denzinger and Veronika Brandstätter. 2018. Stability of and changes in implicit motives. a narrative review of empirical studies. *Frontiers in psychology*, 9:777.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Bertram Gawronski and Jan De Houwer. 2014. Implicit measures in social and personality psychology. *Handbook of research methods in social and personality psychology*, 2:283–310.

Dirk Johannßen and Chris Biemann. 2018. Between the Lines: Machine Learning for Prediction of Psychological Traits - A Survey. LNCS-11015:192–211. Part 2: MAKE-Text.

Dirk Johannßen and Chris Biemann. 2019. Neural classification with attention assessment of the implicit-association test omt and prediction of subsequent academic success. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 68–78, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Dirk Johannßen, Chris Biemann, Steffen Remus, Timo Baumann, and David Sheffer. 2020. Germeval 2020 task 1 on the classification and regression of cognitive and emotional style from text: Companion paper. In *Proceedings of the 5th SwissText & 16th KONVENS Joint Conference 2020*, pages 1–9.

Dirk Johannßen, Chris Biemann, and David Scheffer. 2019. Reviving a psychometric measure: Classification and prediction of the operant motive test. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 121–125.

Julius Kuhl and David Scheffer. 1999. Der operante multi-motiv-test (omt): Manual [the operant multi-motive-test (omt): Manual]. *Germany: University of Osnabrück*.

David C McClelland, Richard Koestner, and Joel Weinberger. 1989. How do self-attributed and implicit motives differ? *Psychological review*, 96(4):690.

H.A. Murray. 1943. *Thematic Apperception Test*. Harvard University Press.

James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12):e115844.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.

Kurt Sokolowski, Heinz-Dieter Schmalt, Thomas A Langens, and Rosa M Puca. 2000. Assessing achievement, affiliation, and power motives all at once: The multi-motive grid (mmg). *Journal of personality assessment*, 74(1):126–145.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

CEUR Workshop Proceedings
http://ceur-ws.org
ISSN 1613-0073

Vol-2664
urn:nbn:de:0074-2664-7

IberLEF 2020
Iberian Languages Evaluation Forum 2020

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020)

Málaga, Spain, September 23th, 2020.

Edited by

Miguel Ángel García Cumbreras *
Julio Gonzalo **
Eugenio Martínez Cámara ***
Raquel Martínez Unanue **
Paolo Rosso ****
Salud Jiménez Zafra *
Jenny A. Ortiz-Zambrano *****
Antonio Miranda-Escalada ******
Jordi Porta-Zamorano, *******
Yoan Guitiérrez ********
Aiala Rosá *********
Manuel Montes-y-Gómez **********
Manuel García-Vega *

* Universidad de Jaén, Spain
** Universidad Nacional de Educación a Distancia, Spain
*** Universidad de Granada, Spain

**** PRHLT Research Center, Universitat Politècnica de València, Spain
***** Universidad de Guayaquil, Ecuador
****** Barcelona Supercomputing Center, Spain
******* Centro de Estudios de la Real Academia Española, Madrid, Spain
******** Department of Software and Computing systems,University of Alicante, Spain
********* Universidad de la República, Uruguay
********** Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico

Table of Contents

- Preface
- Organization

Track 1: Lexical Analysis at SEPLN (ALexS)

- Overview of ALexS 2020: First Workshop on Lexical Analysis at SEPLN 1-6
  Jenny Ortiz-Zambrano, *Arturo Montejo-Ráez*
- Vicomtech at ALexS 2020: Unsupervised Complex Word Identification Based on Domain Frequency 7-14
  Elena Zotova, *Montse Cuadros*, *Naiara Perez*, *Aitor García-Pablos*
- General Lexicon-Based Complex Word Identification Extended with Stem N-grams and Morphological Engines 15-23
  Antonio Rico-Sulayes
- Hulat - ALexS CWI Task - CWI for Language and Learning Disabilities Applied to University Educational Texts 24-30
  Rodrigo Alarcón, *Lourdes Moreno*, *Paloma Martínez*

Track 2: Named Entity Recognition and Classification, and Universal Dependency Parsing of Spanish News (CAPITEL)

- Overview of CAPITEL Shared Tasks at IberLEF 2020:Named Entity Recognition and Universal Dependencies Parsing 31-38
  Jordi Porta-Zamorano, *Luis Espinosa-Anke*
- Combining Different Parsers and Datasets for CAPITEL UD Parsing 39-44
  Fernando Sánchez-León
- Projecting Heterogeneous Annotations for Named Entity Recognition 45-51
  Rodrigo Agerri, *German Rigau*
- Vicomtech at CAPITEL 2020: Facing Entity Recognition and Universal Dependency Parsing of Spanish News Articles with BERT Models 52-59
  Aitor García-Pablos, *Montse Cuadros*, *Elena Zotova*
- System Report of HW-TSC on the CAPITEL NER Evaluation 60-63
  Lizhi Lei, *Minghan Wang*, *Hao Yang*, *Shiliang Sun*, *Ying Qin*, *Daimeng Wei*
- Two Models for Named Entity Recognition in Spanish. Submission to the CAPITEL Shared Task at IberLEF 2020 64-70
  Elena Álvarez Mellado

Track 3: eHealth Knowledge Discovery (eHealth-KD)

- Overview eHealth-KD 2020 71-84

Alejandro Piad-Morffis, *Yoan Gutiérrez*, *Hian Cañizares-Diaz*, *Suilan Estevez-Velarde*, *Rafael Muñoz*, *Andrés Montoyo*, *Yudivián Almeida-Cruz*

- TALP at eHealth-KD Challenge 2020 85-93
  Salvador Medina, *Jordi Turmo*
- ExSim at eHealth-KD Challenge 2020 94-101
  Zainab H Almugbel
- Vicomtech at eHealth-KD Challenge 2020 102-111
  Aitor García-Pablos, *Naiara Perez*, *Montse Cuadros*, *Elena Zotova*
- UH-MatCom at eHealth-KD Challenge 2020 112-124
  Juan Pablo Consuegra-Ayala, *Manuel Palomar*
- UH-MAJA-KD at eHealth-KD Challenge 2020 125-135
  Alejandro Rodríguez-Pérez, *Ernesto Quevedo-Caballero*, *Jorge Mederos-Alvarado*, *Rocío Cruz-Linares*, *Juan Pablo Consuegra-Ayala*
- HapLap at eHealth-KD Challenge 2020 136-140
  Sergio Santana, *Alicia Pérez*, *Arantza Casillas*
- SINAI at eHealth-KD Challenge 2020 141-150
  Pilar López-Úbeda, *José Manuel Perea-Ortega*, *Manuel Carlos Díaz-Galiano*, *María Teresa Martín-Valdivia*, *Luis Alfonso Ureña-López*
- IXA-NER-RE at eHealth-KD Challenge 2020 151-162
  Edgar Andrés, *Óscar Sainz*, *Aitziber Atutxa*, *Oier Lopez de Lacalle*

Track 4: Sentiment Analysis in Spanish (TASS)

- Overview of TASS 2020: Introducing Emotion Detection 163-170
  Manuel García-Vega, *Manuel Carlos Díaz-Galiano*, *Miguel Ángel García-Cumbreras*, *Flor Miriam Plaza del Arco*, *Arturo Montejo-Ráez*, *Salud María Jiménez-Zafra*, *Eugenio Martínez Cámara*, *César Antonio Aguilar*, *Marco Antonio Sobrevilla Cabezudo*, *Luis Chiruzzo*, *Daniela Moctezuma*
- Palomino-Ochoa at TASS 2020: Transformer-based Data Augmentation for Overcoming Few-Shot Learning 171-178
  Daniel Palomino, *José Ochoa-Luna*
- ELiRF-UPV at TASS 2020: TWilBERT for Sentiment Analysis and Emotion Detection in Spanish Tweets 179-186
  José-Ángel González, *José Arias Moncho*, *Lluís-Felip Hurtado*, *Ferran Pla*
- UMUTeam at TASS 2020: Combining Linguistic Features and Machine-learning Models for Sentiment Classification 187-196
  José Antonio García-Díaz, *Ángela Almela*, *Rafael Valencia-García*

Track 5: Factuality Analysis and Classification Task (FACT)

- Overview of FACT at IberLEF 2020: Events Detection and Classification 197-205
  Aiala Rosá, *Laura Alonso*, *Irene Castellón*, *Luis Chiruzzo*, *Hortensia Curell*, *Ana Fernandez Montraveta*, *Santiago Góngora*, *Marisa*

Malcuori, *Glòria Vàzquez*, *Dina Wonsever*

- FACT2020: Factuality Identification in Spanish Text 206-213
  Arturo Collazo, *Agustín Rieppi*, *Tiziana Romani*, *Guillermo Trinidad*
- Factuality Classification Using BERT Embeddings and Support Vector Machines 214-221
  Biswarup Ray, *Avishek Garain*

Track 6: MEX-A3T: Authorship and aggressiveness analysis in Twitter: case study in Mexican Spanish

- Overview of MEX-A3T at IberLEF 2020: Fake News and Aggressiveness Analysis in Mexican Spanish 222-235
  Mario Ezra Aragón, *Horacio Jarquín-Vásquez*, *Manuel Montes-y-Gómez*, *Hugo Jair Escalante*, *Luis Villaseñor-Pineda*, *Helena Gómez-Adorno*, *Juan-Pablo Posadas-Durán*, *Gemma Bel-Enguix*
- Detecting Aggressiveness in Mexican Spanish Social Media Content by Fine-Tuning Transformer-Based Models 236-245
  Mircea-Adrian Tanase, *George-Eduard Zaharia*, *Dumitru-Clementin Cercel*, *Mihai Dascalu*
- GRU with Author Profiling Information to Detect Aggressiveness 246-251
  María Guadalupe Garrido-Espinosa, *Alejandro Rosales-Pérez*, *Adrián Pastor López-Monroy*
- Idiap and UAM Participation at MEX-A3T Evaluation Campaign 252-257
  Esaú Villatoro-Tello, *Gabriela Ramírez-de-la-Rosa*, *Sajit Kumar*, *Shantipriya Parida*, *Petr Motlicek*
- ITCG's Participation at MEX-A3T 2020: Aggressive Identification and Fake News Detection Based on Textual Features for Mexican Spanish 258-264
  Diego Zaizar-Gutiérrez, *Daniel Fajardo-Delgado*, *Miguel Ángel Álvarez-Carmona*
- TecNM at MEX-A3T 2020: Fake News and Aggressiveness Analysis in Mexican Spanish 265-272
  Samuel Arce-Cardenas, *Daniel Fajardo-Delgado*, *Miguel Ángel Álvarez-Carmona*
- UACh at MEX-A3T 2020: Detecting Aggressive Tweets by Incorporating Author and Message Context 273-279
  Marco Casavantes, *Roberto López*, *Luis Carlos González*
- Detecting Aggressiveness in Mexican Spanish Tweets with LSTM + GRU and LSTM + CNN Architectures 280-286
  Victor Peñaloza
- UMUTeam at MEX-A3T'2020: Detecting Aggressiveness with Linguistic Features and Word Embeddings 287-292
  José Antonio García-Díaz, *Rafael Valencia-García*
- Transformers and Data Augmentation for Aggressiveness Detection in Mexican Spanish 293-302

Mario Guzman-Silverio, *Ángel Balderas-Paredes*, *Adrián Pastor López-Monroy*

Track 7: CANcer TExt Mining Shared Task (Cantemist)

- Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results 303-323
  Antonio Miranda-Escalada, *Eulàlia Farré-Maduell*, *Martin Krallinger*
- Extracting Neoplasms Morphology Mentions in Spanish Clinical Cases through Word Embeddings 324-334
  Pilar López-Úbeda, *Manuel Carlos Díaz-Galiano*, *María Teresa Martín-Valdivia*, *Luis Alfonso Ureña-López*
- NLNDE at CANTEMIST: Neural Sequence Labeling and Parsing Approaches for Clinical Concept Extraction 335-346
  Lukas Lange, *Xiang Dai*, *Heike Adel*, *Jannik Strötgen*
- NCU-IISR: Pre-trained Language Model for CANTEMIST Named Entity Recognition 347-351
  Jen-Chieh Han, *Richard Tzong-Han Tsai*
- Recognai's Working Notes for CANTEMIST-NER Track 352-357
  David Carreto Fidalgo, *Daniel Vila-Suero*, *Francisco Aranda Montes*
- End-to-End Neural Coder for Tumor Named Entity Recognition 358-367
  Mohammed Jabreel
- Using Embeddings and Bi-LSTM+CRF Model to Detect Tumor Morphology Entities in Spanish Clinical Cases 368-375
  Sergio Santamaria Carrasco, *Paloma Martínez*
- Tumor Entity Recognition and Coding for Spanish Electronic Health Records
  Fadi Hassan, *David Sánchez*, *Josep Domingo-Ferrer* 376-384
- Deep Neural Model with Contextualized-word Embeddings for Named Entity Recognition in Spanish Clinical Text 385-395
  Renzo Rivera-Zavala, *Paloma Martinez*
- Exploring Deep Learning for Named Entity Recognition of Tumor Morphology Mentions 396-411
  Gema de Vargas Romero, *Isabel Segura-Bedmar*
- Tumor Morphology Mentions Identification Using Deep Learning and Conditional Random Fields 412-421
  Utpal Kumar Sikdar, *Björn Gambäck*, *M Krishna Kumar*
- LasigeBioTM at CANTEMIST: Named Entity Recognition and Normalization of Tumour Morphology Entities and Clinical Coding of Spanish Health-related Documents 422-437
  Pedro Ruas, Andre Neves, Vitor D.T. Andrade and Francisco M. Couto, *Mario Ezra Aragón*
- A Parallel-Attention Model for Tumor Named Entity Recognition in Spanish
  Tong Wang, *Yuanyu Zhang*, *Yongbin Li* 438-446

- A Tumor Named Entity Recognition Model Based on Pre-trained Language Model and Attention Mechanism 447-457
  Xin Taou, *Renyuan Liu*, *Xiaobing Zhou*
- Identification of Cancer Entities in Clinical Text Combining Transformers with Dictionary Features 458-467
  John D Osborne, *Tobias O'Leary*, *James Del Monte*, *Kuleen Sasse*
- ICB-UMA at CANTEMIST 2020: Automatic ICD-O Coding in Spanish with BERT 468-476
  Guillermo López-García, *José Manuel Jerez*, *Nuria Ribelles*, *Emilio Alba*, *Francisco Javier Veredas*
- Automatic ICD Code Classification with Label Description Attention Mechanism 477-488
  Kathryn Chapman, *Günter Neumann*
- Vicomtech at CANTEMIST 2020 489-498
  Aitor García-Pablos, *Naiara Perez*, *Montse Cuadros*
- A Joint Model for Medical Named Entity Recognition and Normalization 499-504
  Ying Xiong, *Yuanhang Huang*, *Qingcai Chen*, *Xiaolong Wang*, *Yuan Nic*, *Buzhou Tang*
- Clinical NER using Spanish BERT Embeddings 505-511
  Ramya Vunikili, *Supriya H N*, *Vasile George Marica*, *Oladimeji Farri*

2020-09-21: submitted by Miguel Ángel García Cumbreras, metadata incl. bibliographic data published under Creative Commons CC0

2020-09-27: published on CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073) |valid HTML5|

Idiap and UAM Participation at MEX-A3T Evaluation Campaign

Esaú Villatoro-Tello[a,b], Gabriela Ramírez-de-la-Rosa[b], Sajit Kumar[c], Shantipriya Parida[b] and Petr Motlicek[b]

[a]Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Mexico City, Mexico
[b]Idiap Research Institute, Rue Marconi 19, 1920, Martigny, Switzerland
[c]Centre of Excellence in AI, Indian Institute of Technology Kharagpur, West Bengal, India

Abstract

This paper describes our participation in the shared evaluation campaign of MexA3T 2020. Our main goal was to evaluate a Supervised Autoencoder (SAE) learning algorithm in text classification tasks. For our experiments, we used three different sets of features as inputs, namely classic word n-grams, char n-grams, and Spanish BERT encodings. Our results indicate that SAE is adequate for longer and more formal written texts. Accordingly, our approach obtained the best performance (F = 85.66%) in the fake-news classification task.

Keywords

Supervised Autoencoders, Text Representation, Deep Learning, Natural Language Processing

1. Introduction

In this era where social media and instant messaging is widely used for communication, the reach and volume of these text messages are enormous. The use of aggressive language or dissemination of false news is widespread across these communication channels. It is impossible to verify the text messages manually. We need automated systems that help users of these communication channels to determine if they are reading real or fake news or to try to flag when someone has been targeted with aggressive messages.

Besides the fact that most of the previous works done in these two tasks, namely aggressiveness detection and fake-news detection, are for English, little research has been done for Spanish using the most recent NLP techniques such as deep learning approaches. On the one hand, for aggressiveness detection, in past editions of the MEX-A3T[1] challenge [1], only three out of nine approaches used some deep learning classifier, particularly for CNN, LSTM, and GRU, with no good performances [2]. On the other hand, most of the current research on fake-news detection has been done for the English language, using graph CNNs [3], and more recently attention mechanism-based transformer models [4].

Our participation at MEX-A3T 2020 aimed at exploring the use of Supervised Autoencoder (SAE) [5] in two different text classification tasks: *i)* aggressiveness detection in Spanish tweets, where documents are very short and informal texts; and, *ii)* fake-news detection from Spanish newspapers,

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)
EMAIL: evillatoro@correo.cua.uam.mx, esau.villatoro@idiap.ch (E. Villatoro-Tello); gramirez@correo.cua.uam.mx (G. Ramírez-de-la-Rosa); kumar.sajit.sk@gmail.com (S. Kumar); shantipriya.parida@idiap.ch (S. Parida); petr.motlicek@idiap.ch (P. Motlicek)
ORCID: 0000-0002-1322-0358 (E. Villatoro-Tello); 0000-0003-3387-6300 (S. Parida)

[1]https://sites.google.com/view/mex-a3t/home

Table 1
Features as inputs for the Supervised Autoencoder Method.

Features type	Sub-type	Identifier
Word n-grams	n=(1,2) and n=(1,3)	W
Char n-grams	n=(1,2) and n=(1,3)	C
BETO	*min*, *max*, and *mean* pooling	B
Word n-grams and Char n-grams		W+C
BETO and Word n-grams		B+W
BETO and Char n-grams		B+C
BETO, Word n-grams and Char n-grams		B+W+C

where documents are larger and contain a more formal written style. We found that SAE can generalize well for both tasks, particularly, for the aggression detection our approach obtains an F1 macro of 80.7%, while for the fake-news detection we reached the best score with an F1 macro of 85.6%.

2. Methodology

For both tasks, we aimed at evaluating the impact of recent generalization techniques, namely SAE [5] with a varied set of features as input vectors. Although SAE has been extensively evaluated in image classification tasks [6], very few works exist evaluating the impact of SAE in text classification tasks, e.g. language detection [7]. Next, we briefly describe the SAE theory, and we provide some details on how the document representation was generated for all the explored features.

2.1. Supervised Autoencoder

An autoencoder (AE) is a neural network that learns a representation (encoding) of input data and then learns to reconstruct the original input from the learned representation. The autoencoder is mainly used for dimensionality reduction or feature extraction [5]. Normally, it is used in an unsupervised learning fashion, meaning that we leverage the neural network for the task of representation learning. By learning to reconstruct the input, the AE extracts underlying abstract attributes that facilitate accurate prediction of the input.

Thus, an SAE is an autoencoder with the addition of a supervised loss on the representation layer. The addition of supervised loss to the autoencoder loss function acts as a regularizer and results in the learning of the better representation for the desired task [6]. For the case of a single hidden layer, a supervised loss is added to the output layer and for a deep supervised autoencoder, the innermost (smallest) layer would have a supervised loss added to the bottleneck layer that is usually transferred to the supervised layer after training the autoencoder.

For all our performed experiments, the overall configuration of the SAE model was done using nonlinear activation function (ReLU) with 3 hidden layers, the number of nodes in the representation layer was set to 300, and we trained to a maximum of 100 epochs.

2.2. Input Features

The SAE receives as input the representation of the document build using Spanish pre-trained BERT encodings (BETO [8]), traditional text representation techniques such as word and char n-grams (ranges 1-2 and 1-3), and, combinations of BETO encodings plus traditional words/char n-grams vectors.

We choose to evaluate the impact of word and char n-grams since as previous research has shown [9, 10, 11], word n-grams are capable of capturing the identity of a word and its contextual usage, while character n-grams are additionally capable of providing an excellent trade-off between sparseness and word's identity, while at the same time they combine different types of information: punctuation, morphological makeup of a word, lexicon and even context. For generating this type of features we used the `CountVectorizer` and `TfidfTransformer` libraries from the scikitlearn[2] toolkit. For the case of the fake-news detection task, we empirically chose the best values for the `min-df` and `max-df` parameters, which are reported on Table 3. For the aggressiveness task, these values were fixed (for all the experiments) to `min-df=` 0.001 and `max-df=` 0.3.

Additionally, we evaluate the impact of transformer-based models [12] as a language representation strategy. For our experiments we tested BETO[3], a BERT model trained on a large dataset of Spanish documents [8]. As known, the [CLS] token acts an "aggregate representation" of the input tokens, and can be considered as a sentence representation for many classification tasks [13]. Accordingly, we apply the following approaches for generating the representation of the document: *i)* for the aggressiveness task, each tweet is directly passed to the BETO model, and is represented using the encoding of the last hidden layer from the [CLS] token; ii) for the fake-news detection task, we split the news document into smaller chunks, obtain the [CLS] encoding of each chunk, and then we apply either a *min*, *max*, *mean* pooling for generating the final document representation. Table 1 depicts the type and variations of features tested during the training phase.

Finally, it is worth mentioning that we did not apply any preprocessing steps in any of the tasks. To validate our experiments, we performed a stratified 10 cross-fold validation strategy.

3. Aggressiveness Identification

The offensive language in Mexican Spanish corpus used for this task has 10,475 Spanish tweets. The training partition contains 7332 tweets with two possible classes (aggressive or non-aggressive). More details of this corpus can be found in [14]. Table 2 shows the results obtained in both, the validation phase and our two runs submitted for the final evaluation of this task over 3143 unseen tweets. The difference between the two submitted outputs, i.e., run id 1 and 2 (†), is the classifier, submission 2 was trained using a Multi-Layer Perceptron (MLP).

4. Fake-News Identification

The fake-news Spanish corpus used in this task has 971 news from 9 different topics. The training partition provided for the development stage has 676 news with a binary class (fake or true). Each news is compose by the headline, body, and the URL from where the news was published (the complete description of this corpus can be found in [15]). For our experiments, we used only the headline and the body of the news as a single document. Table 3 shows the results obtained in the development stage of the challenge, and the two runs submitted for the final evaluation of the tasks over 295 unseen news.

[2] https://scikit-learn.org/stable/index.html
[3] https://github.com/dccuchile/beto

Table 2
Results in validation and test phases reported in F-score for aggressive (F+), non-aggressive (F-), and macro average of the F-score (Fm).

	Validation phase			*Test phase*			
Input features	Fm	F+	F-	ID	Fm	F+	F-
W (1,2)	0.783	0.698	0.868	-	-	-	-
W (1,3)	0.777	0.690	0.864	-	-	-	-
C (1,2)	0.726	0.601	0.850	-	-	-	-
C (1, 3)	0.778	0.689	0.866	-	-	-	-
B (LHL)	0.742	0.628	0.856	-	-	-	-
C (1, 3) + W (1,2)	0.780	0.702	0.857	-	-	-	-
B + W (1,2)	0.787	0.694	0.879	-	-	-	-
B + C (1,3)	0.780	0.684	0.876	-	-	-	-
B + W (1,2) + C (1,3)	**0.803**	0.716	0.889	1	0.807	0.725	0.888
B + W (1,2) + C (1,3)†	**0.798**	0.702	0.894	2	0.801	0.706	0.895
Bi-GRU (baseline-given by track organizers)					0.798	0.712	0.884
BOW-SVM (baseline-given by track organizers)					0.777	0.676	0.878
Best system (in the task [1])					0.859	0.799	0.919

Table 3
Results in validation and test phases reported in F-score for fake-news (F+), real-news (F-), and macro average of F-score (Fm).

		Validation phase			*Test phase*			
Input features	`min-df,max-df`	Fm	F+	F-	ID	Fm	F+	F-
W(1,2)	0.01, 0.5	0.775	0.793	0.758	-	-	-	-
W(1,3)	0.01, 0.5	0.778	0.798	0.758	-	-	-	-
C(1,2)	0.01, 0.5	0.697	0.719	0.674	-	-	-	-
C(1, 3)	0.01, 0.5	0.757	0.768	0.745	-	-	-	-
B(min-pooling)		**0.843**	0.842	0.845	2	**0.856**	**0.844**	**0.868**
B(max-pooling)		0.830	0.830	0.830	-	-	-	-
B(mean-pooling)		0.833	0.831	0.835	-	-	-	-
C(1, 3)+W(1,2)	0.01, 0.5	0.805	0.807	0.802	-	-	-	-
B+W(1,2)	0.01, 0.3	**0.845**	0.846	0.844	1	0.850	0.840	0.859
B+C(1,3)	0.01, 0.3	0.834	0.834	0.835	-	-	-	-
B+W(1,2)+C(1,3)	0.01, 0.3	0.833	0.831	0.835	-	-	-	-
B+W(1,2)+C(1,3)	0.01, 0.5	**0.848**	0.846	0.850	-	-	-	-
Third best system (in the track)						0.817	0.819	0.817
BOW-RF (baseline-given by track organizers)						0.786	0.785	0.787

5. Conclusions

This paper describes Idiap & UAM participation at the MEX-A3T 2020 shared task on the Classification of Fake-News and Aggressiveness analysis. Our participation aimed at analyzing the performance of recent generalization techniques, namely deep supervised autoencoders. To this end, we performed a comparative analysis among simple transformers based language representation strategies and traditional text representations such as word and character n-grams. Notably, the SAE method benefits the most when it is feed with input features generated from the combination of BERT encodings and word/char n-grams. Particularly, for the aggression detection task, our proposed approach can obtain

a relative improvement of 1.1% over the stronger baseline, while for the fake-news detection task the improvement over the baseline is 8.1%.

As future work, we plan to perform an analysis of what are the dataset characteristics that allow the SAE approach to provide good performances. Also, we want to evaluate the impact of SAE's hyperparameter tuning through optimization methods, such as Bayes Optimizer[16], and evaluate our proposed approach on other similar classification tasks.

Acknowledgments

The work was supported by an innovation project (under an InnoSuisse grant) oriented to improve the automatic speech recognition and natural language understanding technologies for German. Title: "SM2: Extracting Semantic Meaning from Spoken Material" funding application no. 29814.1 IP-ICT and EU H2020 project "Real-time network, text, and speaker analytics for combating organized crime" (ROXANNE), grant agreement: 833635. The first author, Esaú Villatoro-Tello is supported partially by Idiap, SNI-CONACyT, CONACyT project grant CB-2015-01-258588, and UAM-C Mexico during the elaboration of this work.

References

[1] M. E. Aragón, H. Jarquín, M. Montes-y Gómez, H. J. Escalante, L. Villaseñor-Pineda, H. Gómez-Adorno, G. Bel-Enguix, J.-P. Posadas-Durán, Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish, in: Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain, September, 2020.

[2] M. E. Aragón, M. Á. Álvarez-Carmona, M. Montes-y Gómez, H. J. Escalante, L. Villasenor-Pineda, D. Moctezuma, Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets, in: Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain, 2019.

[3] F. Monti, F. Frasca, D. Eynard, D. Mannion, M. M. Bronstein, Fake news detection on social media using geometric deep learning, arXiv preprint arXiv:1902.06673 (2019).

[4] M. Qazi, M. U. S. Khan, M. Ali, Detection of fake news using transformer model, in: 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2020, pp. 1–6.

[5] Q. Zhu, R. Zhang, A classification supervised auto-encoder based on predefined evenly-distributed class centroids, arXiv preprint arXiv:1902.00220 (2019).

[6] L. Le, A. Patterson, M. White, Supervised autoencoders: Improving generalization performance with unsupervised regularizers, in: Advances in Neural Information Processing Systems, 2018, pp. 107–117.

[7] S. Parida, E. Villatoro-Tello, S. Kumar, P. Motlicek, Q. Zhan, Idiap submission to swiss-german language detection shared task, in: Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS), 2020.

[8] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, in: to appear in PML4DC at ICLR 2020, 2020.

[9] Z. Wei, D. Miao, J.-H. Chauchat, R. Zhao, W. Li, N-grams based feature selection and text representation for chinese text classification, International Journal of Computational Intelligence Systems 2 (2009) 365–374.

[10] A. Kulmizev, B. Blankers, J. Bjerva, M. Nissim, G. van Noord, B. Plank, M. Wieling, The power of character n-grams in native language identification, in: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, 2017, pp. 382–389.

[11] F. Sánchez-Vega, E. Villatoro-Tello, M. Montes-y Gómez, P. Rosso, E. Stamatatos, L. Villaseñor-Pineda, Paraphrase plagiarism identification with character-level features, Pattern Analysis and Applications 22 (2019) 669–681.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[14] M. J. Díaz-Torres, P. A. Morán-Méndez, L. Villasenor-Pineda, M. Montes-y Gómez, J. Aguilera, L. Meneses-Lerín, Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 132–136. URL: https://www.aclweb.org/anthology/2020.trac-1.21.

[15] J.-P. Posadas-Durán, H. Gomez Adorno, G. Sidorov, J. Moreno, Detection of fake news in a new corpus for the spanish language, Journal of Intelligent Fuzzy Systems 36 (2019) 4869–4876. doi:10.3233/JIFS-179034.

[16] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, in: Advances in neural information processing systems, 2012, pp. 2951–2959.

Gmail

Esaú Villatoro <villatoroe@gmail.com>

[LatinX in AI Research Workshop at NeurIPS 2019] Submission 44 - Accepted

2 messages

Microsoft CMT <email@msr-cmt.org> Wed, Oct 2, 2019 at 4:23 AM
Reply-To: Pablo Fonseca <palefo@gmail.com>
To: Esaú Villatoro-Tello <villatoroe@gmail.com>

Dear Esaú Villatoro-Tello Villatoro-Tello,

We are very excited to let you know that your submission "Finding Evidence Of The Sexual Predators Behavior " was accepted to the LatinX in AI Research Workshop at NeurIPS 2019 to be held in Vancouver, Canada.

Also, we are excited to share that we almost doubled our submissions for the LXAI workshop at Neural Information Processing Systems this year!! With ~180 entries we are super grateful for the community outreach and of course, all our Program Committee who assisted in reviewing all the amazing submissions.

We hope the feedback from the reviewers becomes a useful asset for your research.

We are looking forward to see you in Vancouver!

Best,

The Chairs of the LatinX in AI Research Workshop@NeurIPS 2019

Esaú Villatoro <villatoroe@gmail.com> Wed, Oct 2, 2019 at 12:16 PM
To: Angeles Lopez <angeleslopezf10@gmail.com>, Gabriela Ramírez de la Rosa <a.gaby.rr@gmail.com>

[redacted]

View Reviews

Paper ID
44
Paper Title
Finding Evidence Of The Sexual Predators Behavior

Reviewer #1

Questions

1. What's your overall score for the submission?
Met expectations

4. Please provide feedback for the authors. Eg. What are the strong points of the submission? What are improvement points? What are interesting directions for this research? Please give feedback to authors even if double-blind or format style was not respected in the submission. The main factors considered by the LXAI Research workshop

chairs in qualifying abstracts for acceptance are: -The quality of the research, scholarship, or creative work -The content of the abstract -Adherence to submission criteria referenced below
It's a specific kind of analysis that you don't often see.
I'll be careful about the example displayed to represent the different labels. Would be more reasonable to chose some of them to be less "strong in language"
In the point of the benchmark I think that usually a logistic regression on top of BoW is usually a stronger benchmark than what was used and would be nice to add

5. The submission respects the double-blind criteria?
Yes

Reviewer #2

Questions

1. What's your overall score for the submission?
Failed to meet expectations

4. Please provide feedback for the authors. Eg. What are the strong points of the submission? What are improvement points? What are interesting directions for this research? Please give feedback to authors even if double-blind or format style was not respected in the submission. The main factors considered by the LXAI Research workshop chairs in qualifying abstracts for acceptance are: -The quality of the research, scholarship, or creative work -The content of the abstract -Adherence to submission criteria referenced below
Overall domain is of social importance and the methods are technically sound.

I have concerns with regards to the ways in which the authors handle the material:

The abuse of children is a troubling subject, but the authors address the subject in a way that could benefit from additional sensitivity. For example, the use of sentence samples use language that are highly sexualized is not considered appropriate for a general audience. Although this language may be a typical example, how this language is presented in an academic context should be considered.

Secondly, given the importance of the data source to this task, additional effort could be made to explain the contents of the data and its curation. The definition of the label "incriminatory" is not defined. Is the assumption that this model would be used only with data from minors? How would one handle that ethically? How does one confirm that one party is a minor? How does one ensure that consenting adults are not affected by these methods? These questions remain unclear.

5. The submission respects the double-blind criteria?
Yes

Reviewer #3

Questions

1. What's your overall score for the submission?
Met expectations

4. Please provide feedback for the authors. Eg. What are the strong points of the submission? What are improvement points? What are interesting directions for this research? Please give feedback to authors even if double-blind or format style was not respected in the submission. The main factors considered by the LXAI Research workshop chairs in qualifying abstracts for acceptance are: -The quality of the research, scholarship, or creative work -The content of the abstract -Adherence to submission criteria referenced below
This work is very interesting; However, the results are not clearly explained. What does P, R and F mean in Table 1? I suggest to the authors that they generate a hierarchical classification strategy and improve the classifiers (fine tuning)

5. The submission respects the double-blind criteria?

Yes

Esaú
[Quoted text hidden]
--

--

Esaú Villatoro Tello Ph.D
Titular Professor C
Information Technologies Dept.
Universidad Autónoma Metropolitana (UAM)
Unidad Cuajimalpa
Phone: +52 (55) 5814-6500 Ext. 3540
email: evillatoro@correo.cua.uam.mx
url: http://ccd.cua.uam.mx/~evillatoro

Finding Evidence Of The Sexual Predators Behavior

Ángeles López-Flores
Universidad Autónoma Metropolitana
Mexico City
angeleslopezf10@gmail.com

Esaú Villatoro-Tello
Idiap Researh Institute
Rue Marconi 19, 1920 Martigny, Switzerland.
esau.villatoro@idiap.ch
Universidad Autónoma Metropolitana
Mexico City.
evillatoro@correo.cua.uam.mx

Gabriela Ramírez-de-la-Rosa
Universidad Autónoma Metropolitana
Mexico City
gramirez@correo.cua.uam.mx

1 Introduction

Sexual predator identification is a critical problem given that the majority of cases of sexually assaulted children have agreed voluntarily to meet with their abuser [10]. Traditionally, a term that is used to describe malicious actions with a potential aim of sexual exploitation or emotional connection with a child is referred to as "*Child Grooming*" or "*Grooming Attack*" [6]. This attack is defined by [4] as "*a communication process by which a perpetrator applies affinity seeking strategies, while simultaneously engaging in sexual desensitization and information acquisition about targeted victims in order to develop relationships that result in need fulfillment*" (*e.g.* physical sexual molestation). Clearly, the detection of a malicious predatory behavior against a child could reduce the number of abused children.

Given the difficulties involved in having access to useful data, *i.e.,* where real pedophiles are involved, nowadays the problem of sexual predator identification through pattern recognition techniques is still a challenging research area. The usual approach to catch sexual predators is by means of police officers or volunteers who behave as fake children in chat rooms and provoke sexual offenders to approach them[1]. Unfortunately, online sexual predators always outnumber the law enforcement officers and volunteers. Therefore, tools that can automatically detect and to evidence sexual predators in chat conversations (or at least serve as a support tool for officers) are highly needed. Recently, different research groups have proposed distinct approaches for anticipating the presence of a predator in a chat, i.e., deciding whether or not a conversation is suspicious, and if so, to point the predator [1, 2, 3, 7, 9]. However, an important aspect of the problem has been left behind, i.e., once the predator is identified, officers need to collect all the necessary evidence for sentencing a pedophile. The later is known as the identification of predatory behavior and implies to detect those lines (interventions within a conversation) that are distinctive of the predatory activities.

Accordingly, in this work we focus on the problem of detecting the predatory behavior. Our main proposal is focused on the representation of the chat interventions, thus we incorporate features that capture content, style, and contextual information. For performing our experiments, we used the only publicly available data set for sexual predator detection [5]. This data set was released in the context

[1]The American foundation, called Perverted Justice (PJ) (http://www.perverted-justice.com/), follows the above mentioned approach.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Table 1: Results obtained using three distinct families of features: *content*, *style*, and *behavioral*.

Representation		Classifiers performance								
		NB			SVM			RF		
		P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
BoW	1-gram	0.54	0.47	0.50	**0.75**	**0.48**	**0.59**	0.68	0.37	0.48
	2-gram	0.51	0.39	0.44	0.70	0.33	0.45	0.51	0.37	0.43
	3-gram	0.52	0.17	0.26	0.66	0.16	0.26	0.49	0.21	0.30
POS	1-gram	0.29	0.33	0.31	**0.50**	0.02	0.04	0.31	0.14	0.19
	2-gram	0.31	**0.41**	**0.36**	**0.50**	0.01	0.03	0.38	0.18	0.25
	3-gram	0.33	0.37	0.35	0.46	0.11	0.18	0.35	0.18	0.24
LIWC	—	0.30	**0.58**	0.39	**0.69**	0.09	0.16	0.62	0.37	**0.46**

of the sexual predator identification task (SPI) at PAN-CLEF'12[2] and comprises a large number of chat conversations that include real sexual predators.

2 Proposed framework and initial experiments

For our performed experiments, we followed a traditional supervised machine learning framework. However, as we previously mentioned, we are mainly focus on proposing a suitable representation for the posed task, namely: content, stylistic, and behavioral features. Thus, for our initial set of experiments we used as content features a traditional *Bag-of-Words* with the 10K most frequent features. As for the stylistic features, we considered as features the 36 POS tags contained in the TreeTagger[3] part-of-speech tagger. Finally, as contextual features we account the 68 LIWC [8] psychologically meaningful categories. The LIWC representation provides richer information regarding the words contained in a text, therefore gives context. For example, the word 'cried' matches with four word categories: sadness, negative emotion, overall affect, and a past tense verb.

For training our evidence detection model we used the test partition of the corpus described in [5][4]. In the test partition, a total of 3,737 conversations contain at least one sexual predator[5], and within these conversations, predators interventions are labeled as *incriminatory* or *not-incriminatory*. In order to perform our training, we firstly filtered the 3,737 conversations as done in [9], resulting in a total of 1,466 conversations containing full conversations between victims and a predators. Then, from the filtered version of the corpus we preserve the predator's interventions, giving a total of 59,410 interventions, where 6,395 (11%) are *incriminatory*, and 53,015 (89%) are *not-incriminatory*. As can be noticed, a highly unbalanced problem. Thus, to evaluate the classification performance (using three well know learning algorithms: Naive Bayes, Support Vector Machines and Random Forest) we used precision, recall and the F-score metric of the positive class (i.e., *incriminatory*), and for all experiments we employ a stratified 10 fold cross validation technique to compute the performance.

We observe from Table 1, the best performance ($F = 0.59$) is obtained by the SVM classifier when BoW (*content*) features are used, with $n = 1$ for the *n-gram* size. With respect to the *style* features, the best result was obtained when POS 2-grams are used as features with the NB classifier. As for the *contextual* features, we notice that is not possible to obtain a good performance in terms of F; however, the NB classifier obtains a very high recall level ($R = 0.58$). According to [5], having lot of relevant incriminatory lines, augments the possibility of finding good evidences towards a suspect. Thus, during SPI task at CLEF'12, organizers proposed using the F measure with the β factor equal to 3, hence emphasizing recall. Consequently, our best configuration so far is the one generated by the BoW (1-gram) representation with the SVM classifier, which obtains an $F_{(\beta=3)} = 0.4979$; outperforming the best result reported during CLEF'12 $F_{(\beta=3)} = 0.4762$. Table 2 shows a few examples of the type of evidence we are able to obtain with our proposed method.

As future work, we plan to evaluate fusion methods in order to exploit the best from every family of features. Additionally, we are interested in evaluating the performance of representing the information using word embedding strategies.

[2] `https://pan.webis.de/clef12/pan12-web/`

[3] `https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

[4] The training partition is not labeled with the incriminatory lines.

[5] The total number of conversation on the test partition is near 155K.

Table 2: Examples of incriminatory and not incriminatory evidence found by our proposed method.

Incriminatory	Not-incriminatory
» i'd be so excited with u i'd probably cum just touchin u » you like that I'd do nasty things to your young little body » i will wear condom for you	» do i have anything to be jealous about? » i cant beelieve that i am nervous abt tonmorrow » If u were here we would not be worrying about internet either baby

References

[1] S. G. Burdisso, M. Errecalde, and M. Montes-y Gómez. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197, 2019.

[2] C. Cardei and T. Rebedea. Detecting sexual predators in chats using behavioral features and imbalanced learning. *Natural Language Engineering*, 23(4):589–616, 2017.

[3] H. J. Escalante, E. Villatoro-Tello, S. E. Garza, A. P. López-Monroy, M. Montes-y Gómez, and L. Villaseñor-Pineda. Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, 89:99–111, 2017.

[4] C. Harms. Grooming: An operational definition and coding scheme. *Sex Offender Law Report*, 8(1):1–6, 2007.

[5] G. Inches and F. Crestani. Overview of the international sexual predator identification competition at pan-2012. In *CLEF (Online working notes/labs/workshop)*, volume 30, 2012.

[6] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, 44(4):1448–1466, 2008.

[7] A. P. López-Monroy, F. A. González, and T. Solorio. Early author profiling on twitter using profile features with multi-resolution. *Expert Systems with Applications*, page 112909, 2019.

[8] J. W. Pennebaker. The secret life of pronouns. *New Scientist*, 211(2828):42–45, 2011.

[9] E. Villatoro-Tello, A. Juárez-González, H. J. Escalante, M. Montes-y Gómez, and L. V. Pineda. A two-step approach for effective detection of misbehaving users in chats. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 1178, 2012.

[10] J. Wolak, K. J. Mitchell, and D. Finkelhor. Online victimization of youth: Five years later. 2006.

Speakers

Angeles Lopez Flores
BSc Student at Universidad Autonoma Metropolitana

Posters

- Abraham sanchez "Deep learning models for diabetic retinophaty screening program" (Gobierno de Jalisco)
- Adilson L Khouri "An ontology and frequency-based approach, with machine learning, to recommend activities in scientific workflows" (USP)
- Agostina Larrazabal "Anatomical Priors for Image Segmentation via Post-Processing with Denoising Autoencoders" (Universidad Nacional del Litoral)
- Alfredo A De La Fuente "Expressiveness of Neural Processes" (Schlumberger software Technology innovation Center)
- Arturo oncevay "Revisiting Syllable-aware Language Modelling" (University of edinburgh)
- Bruna Silva "Understanding Algorithmic Fairness in Health Care: A Proposed Case Study with Three Datasets" (Universidade Federal de Minas Gerais)
- Daniel Alcides Saromo Mori "Auto-Rotating Perceptrons" (Pontificia Universidad Católica del Peru)
- Daniel Alfredo Palomino Paucar "Advanced Transfer Learning Approach for Improving Sentiment Analysis on Different Dialects of Spanish" (National University of Engineering)
- Daniel Buades Marcos "Using a self-supervised encoder for anticipating failures in industrial equipment" (Polytecnique Montreal)
- Daniel Ruiz-Perez "Application of Bayesian Techniques to Multiomic Longitudinal Data" (Florida International University)
- Daniel Ruiz-Perez "Role of gut microbiota and their temporal interactions in kidney transplant recipients" (Florida International University)
- Darwin D Saire Pilco "Semantic Segmentation on Image Using Multi-task Hourglass Networks" (University of Campinas)
- David Brenes "Multi-Task Deep Learning Model for Improved Histopathology Prediction from In-Vivo Microscopy Images" (Rice University)
- Dennis H Núñez Fernández "Development of a hand pose recognition

system on an embedded computer using CNNs" (Universidad Nacional de ingenieria)

- Dennis H Núñez Fernández "Portable system for the prediction of anemia based on the ocular conjunctiva using Artificial Intelligence" (Universidad Peruana Cayetano Heredia)
- Dennis H Núñez Fernández "Prediction of gaze direction using Convolutional Neural Networks for Autism diagnosis" (Universidad Peruana Cayetano Heredia)
- Diana M Diaz Herrera "Incorporating Climate Change in Spatiotemporal Species Distribution Models for cattle tick Rhipicephalus (Boophilus) microplus" (Wayne State University)
- Dustin Javier Carrion "Biometric system based on electroencephalogram analysis" (Yachay Tech University)
- Edson Luque "A novel stochastic model based on echo state networks for hydrological time series forecasting"
- Edward Jorge Yuri Cayllahua cahuina "A study of observation scales based on the FH dissimilarity measure" (San Pablo Catholic University)
- Erick D Tornero "Reinforcement Learning Approach to Fly Quadcopters with a Faulted Rotor" (UCSP)
- Errol Wilderd Wilderd Mamani Condori "Aggressive Language Identification in Social Media using Deep Learning" (UCSP)
- Esaú Villatoro-Tello "Finding Evidence Of The Sexual Predators Behavior" (Universidad Autonoma Metropolitana)
- Felipe A. Moreno-Vera "Understanding Safety Based on Urban Perception" (Universidad Catolica San Pablo)
- Felipe Leno da Silva "Autonomously Reusing Knowledge in Multiagent Reinforcement Learning" (university of Sao Paulo)
- Fernando J Yanez "Building Bridges: Implementing Redundancy Analysis by means of a Neural Network" (Universidad Metropolitana)
- Gabriel Jimenez "Skin Cancer Analysis using Deep Learning" (PaPaMED)
- Gabriela Ramirez-de-la-Rosa "Mental lexicon for personality identification in texts" (Universidad Autónoma Metropolitana)
- Gean T Pereira "Transfer Learning for Algorithm Recommendation" (University of São Paulo)
- Gerson Waldyr Vizcarra Aguilar "Paraphrase Generation via Adversarial Penalizations" (San Pablo Catholic University)
- Gilberto Ochoa-Ruiz "Road Damage Acquisition System based on RetinaNet for Physical Asset Management" (Tec de Monterrey)
- Giri Narasiman "EXP4-DFDC: A Non-Stochastic Multi-Armed Bandit for Cache Replacement" (Floridad International University)

- Hansenclever F Bassani "Learning to Play Soccer by Reinforcement and Applying Sim-to-Real to Compete in the Real World" (Universidade Federal de Pernambuco)
- Isabela Albuquerque "Adversarial target-invariant representation learning" (Institut National de la Recherche Scientifique)
- Israel Nazareth Chaparro Cruz "Generative Adversarial Networks for Image Synthesis and Semantic Segmentation in Brain Stroke Images" (Universidad Católica San Pablo)
- Ivan D Arraut Guerrero "Mapping the loss of information of Bosonic (Physical) systems into neural networks with applications in Machine learning" (The Open University of Hong Kong)
- Ivan Vladimir Meza Ruiz "Towards Identifying for Evidence of Drain Brain from Web Search Results using Reinforcement Learning" (universidad Nacional Autonoma de México)
- Jefferson J Quispe Pinares "Automatically Personalized Pain Intensity Estimation from Facial Expressions using CNN-RNN and HCRF in videos". (Universidad CatólicaSanPablo)
- Jessica Soares dos Santos "Investigating Transfer Learning Approaches for Mining Opinions in the Electoral Domain" (Universidade Federal Fluminense)
- Jesús Castillo Cabello "Solving the generalized non-linear Schrödinger equations with genetic algorithms" (Tec de Monterrey)
- Jesús García-Ramírez "Which Kernels to Transfer in Deep Q-Networks?" (INAOE)
- Joao B Monteiro "An end-to-end approach for the verification problem through learned metric-like spaces" (Institut National de la Recherche Scientifique)
- JOHAN SAMIR OBANDO CERON "Exploiting the potential of deep reinforcement learning for classification tasks in high-dimensional and unstructured data" (Universidad Autonoma de Occidente)
- Johnny Torres "Seq2Seq Neural Architecture for Recommending Short Text Conversations" (ESPOLUniversity)
- Jose Paniagua "Generation of time response of linear and nonlinear dynamic systems using autoencoders" (Universidad Autonoma de Occidente)
- Joseph A Gallego "Robust Estimation in Reproducing Kernel Hilbert "(National University Of Colombia)
- José Chávez Ambient "Lighting Generation for Flash Images with Conditional Adversarial Networks" (UCSP)
- Juan A Carvajal "Augmented Curiosity: Depth and Optical Flow Prediction

for Efficient Exploration" (Purdue University)

- Juan M Banda "Weak supervision for electronic phenotyping using electronic health records" (Georgia State University)
- Karla C Otiniano-Rodríguez "Object Recognition using a Region Detector Based on Hierarchies of Partitions" (Esiee Paris (Paris-Est))
- Karol Baca-Lopez "A genetic algorithm implementation for spatio-temporal variogram modelling to determine air quality monitoring network representativeness" (Autonomous University of the State of Mexico)
- Lazo Quispe Cristian "Segmentation of skin lesions and their attributes using Generative Adversarial Networks" (Universidad Nacional de ingenieria)
- leandro ticlia de la cruz "A Machine Learning Approach For Blood Vessels Segmentation In Chorioallantoic Membrane Images" (IO-USP)
- Lourdes Martinez-Villaseñor "Overview of UP-Fall Detection Project" (Universidad Panamericana)
- Lourdes Ramírez Cerna "Emotion recognition using Texture Maps and Convolutional Neural Networks" (National University of Trujillo)
- Lucas Oliveira Souza "Dynamic Sparse Neural Networks" (Numenta)
- Lucy Alsina Choque Mansilla "Object Segmentation by Oriented Image Foresting Transform with Connectivity Constraints" (University of São Paulo)
- Luis A Avendaño Muñoz "Transfer Learning applied to Reinforcement Learning problem with continuous state space using Human-like recall/association" (Universidad de los Andes)
- Luis E Colchado "Interpolation and Prediction of PM2.5 based on Conditional Generative Adversarial Network and a forecasting model" (Universidad Católica San Pablo)
- Luis Fernando Cantu "Algorithmic Targeting of Social Policies: Accuracy & Fairness" (ITAM)
- Manasses A. Mauricio "Pain Intensity Estimation using Spatiotemporal Facial Features" (Universidad Católica San Pablo)
- Marcio Fonseca "Deep Predictive Coding for Multimodal Spatiotemporal Representation Learning" (Câmara dosDeputados)
- Marleny Hilasaca "User-Centered Feature Space Transformation" (University of Sao Paulo)
- Mateo Dulce "Crime prediction using self-exciting point processes and image features as covariates" (Quantil)
- Mikiyas Gulema Tefera "Music video classification using audio and visual features" (Bahir Dar Univerity)

- Mohammed Ali Mr. Adem "Energy Optimization of Wireless Sensor Network Using Neuro-Fuzzy Algorithms" (Bahirdar University)
- Musfiqur Sazal "Signed Causal Bayesian Networks for Microbiomes" (FIU)
- Nicolas Ignacio Fredes Y Nicolás Nieto "On the Impact of Gender Bias in Medical Imaging Classifiers for Computer-aided Diagnosis" (Research institute for signals, systems and computational inteligence)
- Nils Murrugarra-Llerena "Involving humans to learn attributes" (University of Pittsburgh)
- Omar DeGuchy "Relation Augmentation: A Gradient Boosting Approach for Detecting Genomic Anomalies" (University of california, merced)
- Omar U Florez "Memory Networks Encode Knowledge Bases to Generate More Fluent Dialogue Responses" (Capital One)
- Omar U Florez "On the Unintended Social Bias of Training Language Generation Models with Latin American Newspapers" (Capital One)
- Oralia Nolasco-Jauregui "A Machine Learning approach to Neural Information Decoding of Spike Train Distances in the Peripheral Nervous System" (Tecana American University)
- Oscar F Leong "Low Shot Learning with Untrained Neural Networks for Imaging Inverse Problems" (RiceUniversity)
- Pablo Rivas "DiPol-GAN: Generating Molecular Graphs Adversarially with Relational Differentiable Pooling" (Marist College)
- Paul Augusto Bustios Belizario "A Deep Learning Model for Motor Imagery Classification" (University of Sao Paulo)
- Paul Augusto Bustios Belizario "Learning Bandpass and Common Spatial Pattern Filters for Motor Imagery Classification" (University of Sao Paulo)
- Paula Rodriguez "Efficient allocation of law enforcement resources using predictive police patrolling" (Quantil)
- Paulo Mann "See and Read: Detecting Depression Symptoms in Higher Education Students Using Multimodal Social Media Data" (Institute ofComputing / Universidade Federal Fluminense)
- Pedro A Colon-Hernandez "Does a dog desire cake? - Expanding Knowledge Base Assertions Through Deep Relationship Discovery "(MIT Media Lab)
- Pedro H. M. Braga "Backpropagating the Unsupervised Error of Self-Organizing Maps to Deep Neural Networks" (Universidade Federal de Pernambuco)
- Rensso V. H. Mora Colque "Anomaly event detection based on people trajectories for surveillance videos" (UFMG)
- Ricardo Benitez-Jimenez "Meta-Webly Supervised Learning for object recognition" (Instituto Nacional de Astrofísica, Óptica y Electrónica

(INAOE))

- Ricardo Carrillo Mendoza "Model car architecture for education in Robotics and Deep Neural Networks" (FU Berlin)
- Robert A Aduviri "Feature Selection Algorithm Recommendation for Gene Expression data with Meta Learning" (Pontifical Catholic University of Peru)
- Rocio M Zorrilla "On The Selection of Predictive Models in Production" (Laboratorio Nacional de Computacao Cientifica)
- Rodrigo A Toro Icarte "Learning Reward Machines for Partially Observable Reinforcement Learning (Abridged Report) "(University of Toronto and Vector Institute)
- Rodrigo A. Vargas-Hernandez "Gaussian Processes for simulating complex quantum systems" (Chemical Physics Theory Group, Department of Chemistry, University of Toronto)
- Rodrigo C Bonini "Speeding up Reinforcement Learning for Inference and Control of Gene Regulatory Networks" (Universidade Federal do ABC)
- Santiago Miret "Neural Network Autoencoders for Compressed Neuroevolution" (Intel AI Lab)
- Santiago Toledo-Cortés "Large Scale Learning Techniques For Least Squares Support Vector Machines" (Universidad Nacional deColombia)
- Sara I Garcia "Meta-learning for skin cancer detection using Deep Learning techniques" (UniversityCoventry)
- Sidney Araujo Melo "Representation Learning in Game Provenance Graphs" (Institute of Computing / Universidade Federal Fluminense)
- Susana Benavidez "Improving Hate Speech Classification on Twitter" (Stanford University)
- Túlio Corrêa Loures "An Evaluation Benchmark for Online Discussion Representation Models" (Universidade Federal de Minas Gerais)
- Victoria Peterson "Optimizing the regularization parameters selection in sparse modeling" (Instituto de Matemática Aplicada del Litoral)
- Vítor Lourenço "Towards Learning Better Representations for Completion of Real-World Knowledge Bases"(Universidade Federal Fluminense)
- walter M Mayor "Divide and Conquer: an Accurate Machine Learning Algorithm to Process Split Videos on a Parallel Processing Infrastructure" (Univesity Autonoma de Occidente)
- Xochitl Watts "Global Model Explanation for Time Series" (Stanford University Alumni)
- Y "Fast Calorimeter Simulation with Wasserstein Generative Adversarial Networks "(University of Helsinki)

The LXAI Workshop at NeurIPS is only one of the many research and engineering programs our organization is hosting. Please visit our main page to learn more about Latinx in AI initiatives around the world.

Subscribe to our mailing list email address

Subscribe

Build with Jekyll and ❤ by LatinX in AI

Monday

Dec 9th

Schedule of Events

- **7:00am - 8:45am** Check-in
  Vancouver Convention Center Registration
- **8:45am - 9:00am** Opening Ceremony
  MC by David Ramirez
  Vancouver Convention Center East Building East Ballroom A
- **9:00am - 9:30am** Keynote: Carlos Guestrin, Apple AI
  Maximizing the ultimate impact of our ML-based applications
  Vancouver Convent on Center, East Bu d ng, East Ba room A
- **9:30am - 9:40am** Lourdes Ramírez Cerna, National University of Trujillo
  Emotion recognition using Texture Maps and Convolutional Neural Networks
  Vancouver Convent on Center, East Bu d ng, East Ba room A
- **9:40am - 9:50am** Sara I Garcia, University Coventry
  Meta-learning for skin cancer detection using Deep Learning techniques
  Vancouver Convention Center East Building East Ballroom A
- **9:50am - 10:00am** Erick D Tornero, UCSP
  Reinforcement Learning Approach to Fly Quadcopters with a Faulted Rotor
  Vancouver Convention Center East Building East Ballroom A

- **10:00am - 10:10am** Farzana Yusuf, Florida International University
  EXP4-DFDC A Non-Stochastic Multi-Armed Bandit for Cache Replacement
  Vancouver Convent on Center, East Bu d ng, East Ba room A

- **10:10am - 10:50am** Coffee Break
  Vancouver Convention Center East Building East Ballroom A

- **10:55am - 11:05am** João Monteiro, Institut National de la Recherche Scientifique
  An end-to-end approach for the verification problem through learned metric-like spaces
  Vancouver Convent on Center, East Bu d ng, East Ba room A

- **11:05am - 11:35am** Keynote: Barbara Poblete, Instituto Milenio Chile
  ML for Misinformation in Social Media
  Vancouver Convention Center East Building East Ballroom A

- **11:35am - 11:40am** Pablo Rivas, Marist College
  DiPol-GAN Generating Molecular Graphs Adversarially with Relational Differentiable Pooling
  Vancouver Convention Center East Building East Ballroom A

- **11:45am - 11:55pm** Oscar F Leong, Rice University
  Low Shot Learning with Untrained Neural Networks for Imaging Inverse Problems
  Vancouver Convention Center East Building East Ballroom A

- **11:55am - 12:05pm** Daniel Alcides Saromo Mori, PUCP
  Auto-Rotating Perceptrons
  Vancouver Convention Center East Building East Ballroom A

- **12:05pm - 2:00pm** Lunch & Sponsor Keynote: Sergio Guadarrama, Google Brain
  TF-Agents A reliable scalable and easy to use reinforcement learning library for TensorFlow
  Vancouver Convention Center East Building East Ballroom A

- **2:05pm - 2:15pm** Esaú Villatoro-Tello, Universidad Autónoma Metropolitana
  Finding Evidence Of The Sexual Predators Behavior
  Vancouver Convention Center East Building East Ballroom A

- **2:15pm - 2:45pm** Keynote: Alan Aspuru Guzik, University of Toronto
  The materials for tomorrow... today Machine Learning for Chemistry and

Materials

Vancouver Convent on Center, East Bu d ng, East Ba room A

- **2:45pm - 2:55pm** Paula Rodriguez, Quantil

  Efficient allocation of law enforcement resources using predictive police patrolling

  Vancouver Convention Center East Building East Ballroom A

- **2:55pm - 3:30pm** Coffee Break

  Vancouver Convention Center East Building East Ballroom A

- **3:30pm - 5:30pm** Roundtable & Research Mentoring Hour

  Vancouver Convent on Center, East Bu d ng, East Ba room A

- **5:30pm - 6:30pm** Coffee Break

  Vancouver Convention Center East Building East Ballroom A

- **6:30pm** Poster Session

  Vancouver Convention Center East Building

Tuesday

Dec 10th

- **8:00pm - 10:00pm** LXAI Networking Reception

  Sponsored by OpenAI

  Sa t ast ng Room, 45 B ood A ey Square, Vancouver, BC V6B 1C7, Canada

The LXAI Workshop at NeurIPS is only one of the many research and engineering programs our organization is hosting. Please visit

our main page to learn more about Latinx in AI initiatives around the world.

Subscribe to our mailing list email address

Subscribe

Build with Jekyll and ❤ by LatinX in AI

Gmail

Esaú Villatoro <villatoroe@gmail.com>

Fwd: [LatinX in AI Research Workshop at NeurIPS 2019] Submission 149 - Accepted

Gabriela Ramírez de la Rosa <a.gaby.rr@gmail.com> Wed, Oct 2, 2019 at 4:41 AM
To: Esaú Villatoro <villatoroe@gmail.com>, Héctor Jiménez Salazar <hgimenezs@gmail.com>

Buenas noches,
Perdón por la hora, pero acabo de recibir la notificación de aceptación del abstract que envié el mes pasado, anexo los comentarios que subieron al sistema pero no enviaron por correo:

Paper ID
149
Paper Title
Mental lexicon for personality identification in texts

Reviewer #1

Questions

1. What's your overall score for the submission?
Met expectations

4. Please provide feedback for the authors. Eg. What are the strong points of the submission? What are improvement points? What are interesting directions for this research? Please give feedback to authors even if double-blind or format style was not respected in the submission. The main factors considered by the LXAI Research workshop chairs in qualifying abstracts for acceptance are: -The quality of the research, scholarship, or creative work -The content of the abstract -Adherence to submission criteria referenced below
General comments:
The study asses personality identification as an author profiling task in NLP and cognitive linguistics. The motivation and background of the work are well defined, however, and given the limited extension of the abstract, it could be more concise, to free space for further discussion and analysis of the results.

-The quality of the research, scholarship, or creative work
The main contribution is the method to represent texts in the lexical availability context. It has been detailed extensively, although there are some factors that are not entirely clear (e.g. "-2.3" in the exponential of LA(tj)). Besides, the authors have not identified which is the learning algorithm that provides the results in Table 1. Furthermore, as the results are very close between the proposed method and baselines, it is not clear if there is a significant difference. There should be some hypothesis testing for comparing distributions.

-The content of the abstract
The first page of the abstract could have been more concise. More robust and detailed analysis is required when the proposed method cannot overcome the baselines for significant margins.

-Adherence to submission criteria referenced below
Double-blind and format style are respected in the submission.

5. The submission respects the double-blind criteria?
Yes

Reviewer #2

Questions

1. What's your overall score for the submission?

Met expectations

4. Please provide feedback for the authors. Eg. What are the strong points of the submission? What are improvement points? What are interesting directions for this research? Please give feedback to authors even if double-blind or format style was not respected in the submission. The main factors considered by the LXAI Research workshop chairs in qualifying abstracts for acceptance are: -The quality of the research, scholarship, or creative work -The content of the abstract -Adherence to submission criteria referenced below

This paper describes a system for detecting personality in texts. The approach is based on creating list of discriminative units and weights for such units, then using different classifiers to detect the personality of a text. The problem is interesting, and as mentioned by the authors, it is quite similar to style classification. The paper will benefit of testing approaches that are also standard for that problem, on providing examples of the texts, length, confusion matrices, size of the lists, examples of the extracted lists, etc. Definition of LIWC should be given. Abstract is not provided.

5. The submission respects the double-blind criteria?

Yes

Reviewer #3

Questions

1. What's your overall score for the submission?

Met expectations

4. Please provide feedback for the authors. Eg. What are the strong points of the submission? What are improvement points? What are interesting directions for this research? Please give feedback to authors even if double-blind or format style was not respected in the submission. The main factors considered by the LXAI Research workshop chairs in qualifying abstracts for acceptance are: -The quality of the research, scholarship, or creative work -The content of the abstract -Adherence to submission criteria referenced below

It seems a very interesting work!

strong points:
- the task is very well defined.
- the text is short but well written.

improvement:
-talk about the method employed.
-give details about the results.

I cannot give any further advice since the method isn't mentioned in the abstract.

5. The submission respects the double-blind criteria?

Yes

Gabriela.

---------- Forwarded message ---------
De: **Microsoft CMT** <email@msr-cmt.org>
Date: mar., 1 de oct. de 2019 a la(s) 21:23
Subject: [LatinX in AI Research Workshop at NeurIPS 2019] Submission 149 - Accepted
To: Gabriela Ramirez-de-la-Rosa <a.gaby.rr@gmail.com>

Dear Gabriela Ramirez-de-la-Rosa Ramirez-de-la-Rosa,

We are very excited to let you know that your submission "Mental lexicon for personality identification in texts" was accepted to the LatinX in AI Research Workshop at NeurIPS 2019 to be held in Vancouver, Canada.

Also, we are excited to share that we almost doubled our submissions for the LXAI workshop at Neural Information Processing Systems this year!! With ~180 entries we are super grateful for the community outreach

and of course, all our Program Committee who assisted in reviewing all the amazing submissions.

We hope the feedback from the reviewers becomes a useful asset for your research.

We are looking forward to see you in Vancouver!

Best,

The Chairs of the LatinX in AI Research Workshop@NeurIPS 2019

Mental lexicon for personality identification in texts

Gabriela Ramírez-de-la-Rosa
Universidad Autónoma Metropolitana
Mexico City
gramirez@correo.cua.uam.mx

Esaú Villatoro-Tello
Idiap Researh Institute
Rue Marconi 19, 1920 Martigny, Switzerland.
esau.villatoro@idiap.ch
Universidad Autónoma Metropolitana
Mexico City
evillatoro@correo.cua.uam.mx

Héctor Jiménez-Salazar
Universidad Autónoma Metropolitana
Mexico City
hjimenez@correo.cua.uam.mx

1 Introduction

Personality identification from texts is a relative new area of interest in the natural language processing (NLP) community. The benefits of helping to identify the personality of a subject solely on the text they write are manifolds. For one, it can help directly to the authors of such texts to understand their social interactions, and their behaviour in general [5, 12]. Beyond that, personality identification is useful for many other research areas. For instance, in human computer interactions (HCI), interactive systems may be able to adapt to user's personality, providing a better experience [2]. In education, building intelligent tutors compatible with the student's personality can improve, not only the experience of the student with the system, but also the system could provide more adequate material from a educative program in accordance to the particular student's preferences [10, 6].

From the NLP perspective, personality identification from texts can be treated as an author profiling problem. Author profiling consists on, given a text, determine some demographics characteristics of the author of such text. In this context, the representation of a given text such that the model can extract relevant information according to the specific demographic interest [7, 1] is of a relevant importance.

In the mental health context, the main interest is not only to build accurate systems, but to provide interpretable results that in turn, would serve as additional and reliable elements to a therapist. Accordingly, we focused on developing an automatic method for personality identification, able to provide valuable information regarding the language usage of subjects being analyzed.

Specifically, we use the linguistic theory behind lexical availability to first compute a set of relevant mental lexicon from groups of subjects (e.g. *introverts* vs *extroverts* for the Extroversion trait) and then we use this mental lexicon in a representation stage. For our experiments, we use two data sets: English essays and Spanish essays; these datasets use the Big Five Model of Personality [9].

2 Lexical Availability as language descriptor

Lexical availability methods were developed to provide useful vocabulary to immigrants in early 60's in France [13]; where word's frequencies do not necessary means importance of such a word in a given context. Traditionally, the lexical availability elicitation approach consists on ask to a group of subjects to write, in a small period of time (usually 2 to 5 minutes), a set of terms given a specific center of interest [4, 13].

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

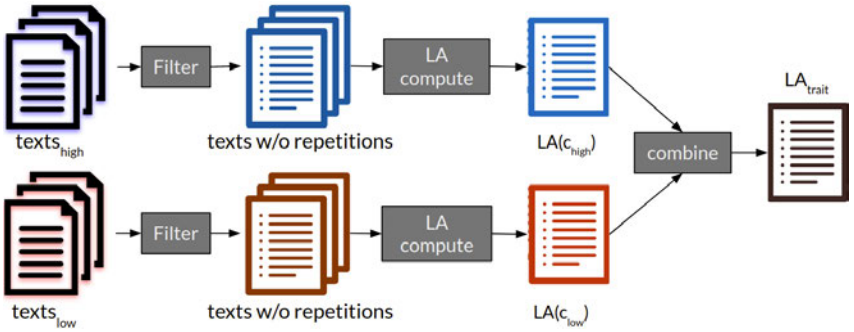


Figure 1: Schema to generate a mental lexicon given a set of written texts.

Table 1: Results with the best configuration from our proposed method and traditional baseline. In bold are mark results of our method when outperform the baseline.

	RxPI Spanish[11]		**English essays**[8]	
Trait	F-macro (Ours)	F-macro (Baseline)	F-macro (Ours)	F-macro (Baseline)
EXT	**0.6018**	0.5640	0.5753	0.5788
AGR	0.5697	0.5711	**0.5615**	0.5530
CON	**0.5857**	.5800	0.5795	0.5806
STA	**0.6026**	0.5828	**0.5918**	0.5785
OPE	0.5704	0.5722	**0.6414**	0.6237

We use a linguistically motivated approach aiming to identify those lexical markers that represent the words springing to mind in response to a specific topic. Lexical Availability score (LA) measures the ease with which word is generated in a given communicative situation [4], and allows to obtain the *mental lexicon* which represents the vocabulary flow usable of a group of people [3].

In general, the terms with greater LA score can be seen as the most important ones for a group of people with the same personality trait. Thus, we computed the mental lexicon for each pole in a trait, and then a general list (LA_{trait}) was generated to be use in a vectorial representation model with dimension equal to $|LA_{trait}|$.

3 Proposed framework and evaluation

We proposed the method in Figure 1 to use lexical availability for texts representation. Our method has three main processes: The *filter process* generates a list of terms without repetitions given any instance text. The *LA compute process* computes the lexical availability score of a list of terms as $LA(t_j) = \sum_{i=1}^{n} e^{(-2.3*\frac{i-1}{n-1})} * \frac{f_{ij}}{I}$, where t_j is the term j in a list; n is the lowest position of a term j in some list; i is the position of term j in a list; f_{ij} is the number of lists in where term j appears in position i, and I is the total number of lists. Finally, the *combine process*, takes as input the lists generated for each class and using set operations combine them into a single general list that we called LA_{trait}.

Once we have the mental lexicon of a trait (a.k.a. LA_{trait}), we use the scores and terms in this list to generate a vector representation of a given instance text. In order to weight each term in our vector, we use three approaches as follows. If w_k is the weight of a term k and $LA(w_k)$ is the score of lexical aviability of word k in the list LA then: 1) $w_k^{global} = LA_{trait}(w_k)$, 2) $w_k^{comb} = LA_{trait}(w_k) * LA_{instance}(w_k)$ where $LA_{instance}$ is the score of a term in the unseen instance, and 3) $w_k^{tfla} = tf * LA_{trait}(w_k)$, where tf is the frequency of the term (w_k) in the unseen instance.

To compare our performance in classification, we used three representation baselines: n-grams of words and characters, and a dictionary based representations such as LIWC. For each of these baselines we experimented with several configuration parameters (e.g. the number of n). To train a model we used traditional learning algorithms such as probabilistic, decision trees, support vector

Table 2: Results with the best configuration from our proposed method and traditional baseline. In bold are mark results of our method when outperform the baseline.

Trait	**RxPI Spanish** (Ramirez et al., 2018) F-macro (Ours)	F-macro (Baseline)	**English essays** (Mairesse et al.,2007) F-macro (Ours)	F-macro (Baseline)
EXT	**0.6018**	0.5640	0.5753	0.5788
AGR	0.5697	0.5711	**0.5615**	0.5530
CON	**0.5857**	.5800	0.5795	0.5806
STA	**0.6026**	0.5828	**0.5918**	0.5785
OPE	0.5704	0.5722	**0.6414**	0.6237

machine, and instance based. Table 2 shows the results with the best parameters for our method as well as for the baselines.

Our ongoing work in this project is to analyze the semantic categories in each lists that are relevant to the expert when identify the personality of a subject. At the same time we want to use more sophisticated methods that take advantage of our proposed representation to improve the classification performance.

References

[1] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni. Gender, genre, and writing style in formal written texts. *TEXT*, 23:321–346, 2003.

[2] T. W. Bickmore and R. W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.*, 12(2):293–327, June 2005.

[3] R. M. J. Catalán. *Lexical availability in English and Spanish as a second language*, volume 17. Springer, 2013.

[4] N. R. Dimitrijević. A comparative study of the lexical availability of monolingual and bilingual schoolchildren. 1981.

[5] K. A. Holder. M. D. Temperament and happiness in children. *Journal of Happiness Studies*, 2009.

[6] M. Komarraju and S. J. Karau. The relationship between the big five personality traits and academic motivation. *Personality and Individual Differences*, 39(3):557 – 567, 2005.

[7] M. Koppel, S. Argamon, and A. R. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[8] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR*, pages 457–500, 2007.

[9] R. R. McCrae and P. T. Costa Jr. Personality trait structure as a human universal. *American psychologist*, 52(5):509, 1997.

[10] M. Pavalache-Ilie and S. Cocorada. Interactions of students' personality in the online learning environment. *Procedia - Social and Behavioral Sciences*, 128:117 – 122, 2014.

[11] G. Ramírez-de-la Rosa, E. Villatoro-Tello, and H. Jiménez-Salazar. Txpi-u: A resource for personality identification of undergraduates. *Journal of Intelligent & Fuzzy Systems*, 34(5):2991–3001, May 2018.

[12] G. B. Svendsen, J.-A. K. Johnsen, L. Almås-Sørensen, and J. Vittersø. Personality and technology acceptance: the influence of personality factors on the core constructs of the technology acceptance model. *Behaviour & Information Technology*, 32(4):323–334, 2013.

[13] A. Ávila Muñoz and J. Sanchez Saez. Fuzzy sets and prototype theory: Representational model of cognitive community structures based on lexical availability trials. *Review of Cognitive Linguistics*, 12, 01 2014.

Speakers

Posters

- Abraham sanchez "Deep learning models for diabetic retinophaty screening program" (Gobierno de Jalisco)
- Adilson L Khouri "An ontology and frequency-based approach, with machine learning, to recommend activities in scientific workflows" (USP)
- Agostina Larrazabal "Anatomical Priors for Image Segmentation via Post-Processing with Denoising Autoencoders" (Universidad Nacional del Litoral)
- Alfredo A De La Fuente "Expressiveness of Neural Processes" (Schlumberger software Technology innovation Center)
- Arturo oncevay "Revisiting Syllable-aware Language Modelling" (University of edinburgh)
- Bruna Silva "Understanding Algorithmic Fairness in Health Care: A Proposed Case Study with Three Datasets" (Universidade Federal de Minas Gerais)
- Daniel Alcides Saromo Mori "Auto-Rotating Perceptrons" (Pontificia Universidad Católica del Peru)
- Daniel Alfredo Palomino Paucar "Advanced Transfer Learning Approach for Improving Sentiment Analysis on Different Dialects of Spanish" (National University of Engineering)
- Daniel Buades Marcos "Using a self-supervised encoder for anticipating failures in industrial equipment" (Polytecnique Montreal)
- Daniel Ruiz-Perez "Application of Bayesian Techniques to Multiomic Longitudinal Data" (Florida International University)
- Daniel Ruiz-Perez "Role of gut microbiota and their temporal interactions in kidney transplant recipients" (Florida International University)
- Darwin D Saire Pilco "Semantic Segmentation on Image Using Multi-task Hourglass Networks" (University of Campinas)
- David Brenes "Multi-Task Deep Learning Model for Improved Histopathology Prediction from In-Vivo Microscopy Images" (Rice University)
- Dennis H Núñez Fernández "Development of a hand pose recognition

system on an embedded computer using CNNs" (Universidad Nacional de ingenieria)
- Dennis H Núñez Fernández "Portable system for the prediction of anemia based on the ocular conjunctiva using Artificial Intelligence" (Universidad Peruana Cayetano Heredia)
- Dennis H Núñez Fernández "Prediction of gaze direction using Convolutional Neural Networks for Autism diagnosis" (Universidad Peruana Cayetano Heredia)
- Diana M Diaz Herrera "Incorporating Climate Change in Spatiotemporal Species Distribution Models for cattle tick Rhipicephalus (Boophilus) microplus" (Wayne State University)
- Dustin Javier Carrion "Biometric system based on electroencephalogram analysis" (Yachay Tech University)
- Edson Luque "A novel stochastic model based on echo state networks for hydrological time series forecasting"
- Edward Jorge Yuri Cayllahua cahuina "A study of observation scales based on the FH dissimilarity measure" (San Pablo Catholic University)
- Erick D Tornero "Reinforcement Learning Approach to Fly Quadcopters with a Faulted Rotor" (UCSP)
- Errol Wilderd Wilderd Mamani Condori "Aggressive Language Identification in Social Media using Deep Learning" (UCSP)
- Esaú Villatoro-Tello "Finding Evidence Of The Sexual Predators Behavior" (Universidad Autonoma Metropolitana)
- Felipe A. Moreno-Vera "Understanding Safety Based on Urban Perception" (Universidad Catolica San Pablo)
- Felipe Leno da Silva "Autonomously Reusing Knowledge in Multiagent Reinforcement Learning" (university of Sao Paulo)
- Fernando J Yanez "Building Bridges: Implementing Redundancy Analysis by means of a Neural Network" (Universidad Metropolitana)
- Gabriel Jimenez "Skin Cancer Analysis using Deep Learning" (PaPaMED)
- Gabriela Ramirez-de-la-Rosa "Mental lexicon for personality identification in texts" (Universidad Autónoma Metropolitana)
- Gean T Pereira "Transfer Learning for Algorithm Recommendation" (University of São Paulo)
- Gerson Waldyr Vizcarra Aguilar "Paraphrase Generation via Adversarial Penalizations" (San Pablo Catholic University)
- Gilberto Ochoa-Ruiz "Road Damage Acquisition System based on RetinaNet for Physical Asset Management" (Tec de Monterrey)
- Giri Narasiman "EXP4-DFDC: A Non-Stochastic Multi-Armed Bandit for Cache Replacement" (Floridad International University)

- Hansenclever F Bassani "Learning to Play Soccer by Reinforcement and Applying Sim-to-Real to Compete in the Real World" (Universidade Federal de Pernambuco)
- Isabela Albuquerque "Adversarial target-invariant representation learning" (Institut National de la Recherche Scientifique)
- Israel Nazareth Chaparro Cruz "Generative Adversarial Networks for Image Synthesis and Semantic Segmentation in Brain Stroke Images" (Universidad Católica San Pablo)
- Ivan D Arraut Guerrero "Mapping the loss of information of Bosonic (Physical) systems into neural networks with applications in Machine learning" (The Open University of Hong Kong)
- Ivan Vladimir Meza Ruiz "Towards Identifying for Evidence of Drain Brain from Web Search Results using Reinforcement Learning" (universidad Nacional Autonoma de México)
- Jefferson J Quispe Pinares "Automatically Personalized Pain Intensity Estimation from Facial Expressions using CNN-RNN and HCRF in videos". (Universidad CatólicaSanPablo)
- Jessica Soares dos Santos "Investigating Transfer Learning Approaches for Mining Opinions in the Electoral Domain" (Universidade Federal Fluminense)
- Jesús Castillo Cabello "Solving the generalized non-linear Schrödinger equations with genetic algorithms" (Tec de Monterrey)
- Jesús García-Ramírez "Which Kernels to Transfer in Deep Q-Networks?" (INAOE)
- Joao B Monteiro "An end-to-end approach for the verification problem through learned metric-like spaces" (Institut National de la Recherche Scientifique)
- JOHAN SAMIR OBANDO CERON "Exploiting the potential of deep reinforcement learning for classification tasks in high-dimensional and unstructured data" (Universidad Autonoma de Occidente)
- Johnny Torres "Seq2Seq Neural Architecture for Recommending Short Text Conversations" (ESPOLUniversity)
- Jose Paniagua "Generation of time response of linear and nonlinear dynamic systems using autoencoders" (Universidad Autonoma de Occidente)
- Joseph A Gallego "Robust Estimation in Reproducing Kernel Hilbert "(National University Of Colombia)
- José Chávez Ambient "Lighting Generation for Flash Images with Conditional Adversarial Networks" (UCSP)
- Juan A Carvajal "Augmented Curiosity: Depth and Optical Flow Prediction

for Efficient Exploration" (Purdue University)

- Juan M Banda "Weak supervision for electronic phenotyping using electronic health records" (Georgia State University)
- Karla C Otiniano-Rodríguez "Object Recognition using a Region Detector Based on Hierarchies of Partitions" (Esiee Paris (Paris-Est))
- Karol Baca-Lopez "A genetic algorithm implementation for spatio-temporal variogram modelling to determine air quality monitoring network representativeness" (Autonomous University of the State of Mexico)
- Lazo Quispe Cristian "Segmentation of skin lesions and their attributes using Generative Adversarial Networks" (Universidad Nacional de ingenieria)
- leandro ticlia de la cruz "A Machine Learning Approach For Blood Vessels Segmentation In Chorioallantoic Membrane Images" (IO-USP)
- Lourdes Martinez-Villaseñor "Overview of UP-Fall Detection Project" (Universidad Panamericana)
- Lourdes Ramírez Cerna "Emotion recognition using Texture Maps and Convolutional Neural Networks" (National University of Trujillo)
- Lucas Oliveira Souza "Dynamic Sparse Neural Networks" (Numenta)
- Lucy Alsina Choque Mansilla "Object Segmentation by Oriented Image Foresting Transform with Connectivity Constraints" (University of São Paulo)
- Luis A Avendaño Muñoz "Transfer Learning applied to Reinforcement Learning problem with continuous state space using Human-like recall/association" (Universidad de los Andes)
- Luis E Colchado "Interpolation and Prediction of PM2.5 based on Conditional Generative Adversarial Network and a forecasting model" (Universidad Católica San Pablo)
- Luis Fernando Cantu "Algorithmic Targeting of Social Policies: Accuracy & Fairness" (ITAM)
- Manasses A. Mauricio "Pain Intensity Estimation using Spatiotemporal Facial Features" (Universidad Católica San Pablo)
- Marcio Fonseca "Deep Predictive Coding for Multimodal Spatiotemporal Representation Learning" (Câmara dosDeputados)
- Marleny Hilasaca "User-Centered Feature Space Transformation" (University of Sao Paulo)
- Mateo Dulce "Crime prediction using self-exciting point processes and image features as covariates" (Quantil)
- Mikiyas Gulema Tefera "Music video classification using audio and visual features" (Bahir Dar Univerity)

- Mohammed Ali Mr. Adem "Energy Optimization of Wireless Sensor Network Using Neuro-Fuzzy Algorithms" (Bahirdar University)
- Musfiqur Sazal "Signed Causal Bayesian Networks for Microbiomes" (FIU)
- Nicolas Ignacio Fredes Y Nicolás Nieto "On the Impact of Gender Bias in Medical Imaging Classifiers for Computer-aided Diagnosis" (Research institute for signals, systems and computational inteligence)
- Nils Murrugarra-Llerena "Involving humans to learn attributes" (University of Pittsburgh)
- Omar DeGuchy "Relation Augmentation: A Gradient Boosting Approach for Detecting Genomic Anomalies" (University of california, merced)
- Omar U Florez "Memory Networks Encode Knowledge Bases to Generate More Fluent Dialogue Responses" (Capital One)
- Omar U Florez "On the Unintended Social Bias of Training Language Generation Models with Latin American Newspapers" (Capital One)
- Oralia Nolasco-Jauregui "A Machine Learning approach to Neural Information Decoding of Spike Train Distances in the Peripheral Nervous System" (Tecana American University)
- Oscar F Leong "Low Shot Learning with Untrained Neural Networks for Imaging Inverse Problems" (RiceUniversity)
- Pablo Rivas "DiPol-GAN: Generating Molecular Graphs Adversarially with Relational Differentiable Pooling" (Marist College)
- Paul Augusto Bustios Belizario "A Deep Learning Model for Motor Imagery Classification" (University of Sao Paulo)
- Paul Augusto Bustios Belizario "Learning Bandpass and Common Spatial Pattern Filters for Motor Imagery Classification" (University of Sao Paulo)
- Paula Rodriguez "Efficient allocation of law enforcement resources using predictive police patrolling" (Quantil)
- Paulo Mann "See and Read: Detecting Depression Symptoms in Higher Education Students Using Multimodal Social Media Data" (Institute ofComputing / Universidade Federal Fluminense)
- Pedro A Colon-Hernandez "Does a dog desire cake? - Expanding Knowledge Base Assertions Through Deep Relationship Discovery "(MIT Media Lab)
- Pedro H. M. Braga "Backpropagating the Unsupervised Error of Self-Organizing Maps to Deep Neural Networks" (Universidade Federal de Pernambuco)
- Rensso V. H. Mora Colque "Anomaly event detection based on people trajectories for surveillance videos" (UFMG)
- Ricardo Benitez-Jimenez "Meta-Webly Supervised Learning for object recognition" (Instituto Nacional de Astrofísica, Óptica y Electrónica

(INAOE))

- Ricardo Carrillo Mendoza "Model car architecture for education in Robotics and Deep Neural Networks" (FU Berlin)
- Robert A Aduviri "Feature Selection Algorithm Recommendation for Gene Expression data with Meta Learning" (Pontifical Catholic University of Peru)
- Rocio M Zorrilla "On The Selection of Predictive Models in Production" (Laboratorio Nacional de Computacao Cientifica)
- Rodrigo A Toro Icarte "Learning Reward Machines for Partially Observable Reinforcement Learning (Abridged Report) "(University of Toronto and Vector Institute)
- Rodrigo A. Vargas-Hernandez "Gaussian Processes for simulating complex quantum systems" (Chemical Physics Theory Group, Department of Chemistry, University of Toronto)
- Rodrigo C Bonini "Speeding up Reinforcement Learning for Inference and Control of Gene Regulatory Networks" (Universidade Federal do ABC)
- Santiago Miret "Neural Network Autoencoders for Compressed Neuroevolution" (Intel AI Lab)
- Santiago Toledo-Cortés "Large Scale Learning Techniques For Least Squares Support Vector Machines" (Universidad Nacional deColombia)
- Sara I Garcia "Meta-learning for skin cancer detection using Deep Learning techniques" (UniversityCoventry)
- Sidney Araujo Melo "Representation Learning in Game Provenance Graphs" (Institute of Computing / Universidade Federal Fluminense)
- Susana Benavidez "Improving Hate Speech Classification on Twitter" (Stanford University)
- Túlio Corrêa Loures "An Evaluation Benchmark for Online Discussion Representation Models" (Universidade Federal de Minas Gerais)
- Victoria Peterson "Optimizing the regularization parameters selection in sparse modeling" (Instituto de Matemática Aplicada del Litoral)
- Vítor Lourenço "Towards Learning Better Representations for Completion of Real-World Knowledge Bases"(Universidade Federal Fluminense)
- walter M Mayor "Divide and Conquer: an Accurate Machine Learning Algorithm to Process Split Videos on a Parallel Processing Infrastructure" (Univesity Autonoma de Occidente)
- Xochitl Watts "Global Model Explanation for Time Series" (Stanford University Alumni)
- Y "Fast Calorimeter Simulation with Wasserstein Generative Adversarial Networks "(University of Helsinki)

The LXAI Workshop at NeurIPS is only one of the many research and engineering programs our organization is hosting. Please visit our main page to learn more about Latinx in AI initiatives around the world.

Subscribe to our mailing list email address

Subscribe

Build with Jekyll and ❤ by LatinX in AI

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA
Unidad Cuajimalpa

Comunidad académica comprometida con el desarrollo humano de la sociedad.

Enero 17, 2020

A QUIEN CORRESPONDA

ASUNTO: Constancia Asesor de Proyecto Terminal

Por medio de la presente hago constar que los profesores investigadores **M.C. Adriana Gabriela Ramírez de la Rosa** y el **Dr. Esaú Villatoro Tello**, adscritos al Departamento de Tecnologías de la Información de la Universidad Autónoma Metropolitana Unidad Cuajimalpa, asesoraron a la alumna **Erika Saraí Rosas Quezada** (matrícula 2153067524) en el desarrollo de su Proyecto Terminal de carrera titulado "**Prediciendo el impacto de una publicación de Facebook**". Este proyecto se concluyó satisfactoriamente en el trimestre 19P.

Se extiende la presente para los fines que convengan a los interesados.

Atentamente,
Casa abierta al tiempo

Dr. Carlos Roberto Jaimez González
Coordinador de la Licenciatura en Tecnologías y Sistemas de la Información
Email: cjaimez@correo.cua.uam.mx

Universidad Autónoma Metropolitana
División de Ciencias de la Comunicación y Diseño
Departamento De Tecnologías de la Información

Licenciatura en Tecnologías y Sistemas de Información

Prediciendo el impacto de una publicación de Facebook

Por:

Erika Sarai Rosas Quezada

Asesores:

Mtra. Adriana Gabriela Ramírez de la Rosa

Dr. Esaú Villatoro Tello

Noviembre 2019

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA
Unidad Cuajimalpa

Comunidad académica comprometida con el desarrollo humano de la sociedad.

Enero 17, 2020

A QUIEN CORRESPONDA

ASUNTO: Constancia Asesor de Proyecto Terminal

Por medio de la presente hago constar que los profesores investigadores **Dr. Esaú Villatoro Tello** y la **M.C. Adriana Gabriela Ramírez de la Rosa**, adscritos al Departamento de Tecnologías de la Información de la Universidad Autónoma Metropolitana Unidad Cuajimalpa, asesoraron a la alumna **Ángeles López Flores** (matrícula 2153067542) en el desarrollo de su Proyecto Terminal de carrera titulado "**Sistema web de apoyo para la identificación automática de evidencia textual en casos de pedofilia**". Este proyecto se concluyó satisfactoriamente en el trimestre 19P.

Se extiende la presente para los fines que convengan a los interesados.

Atentamente,
Casa abierta al tiempo

Dr. Carlos Roberto Jaimez González
Coordinador de la Licenciatura en Tecnologías y Sistemas de la Información
Email: cjaimez@correo.cua.uam.mx

UNIVERSIDAD AUTÓNOMA METROPOLITANA
Casa abierta al tiempo
Unidad Cuajimalpa

División de Ciencias de la Comunicación y Diseño

Licenciatura en Tecnologías y Sistemas de Información

Proyecto Terminal

Sistema web de apoyo para la identificación automática de evidencia textual en casos de pedofilia

Por:

Angeles López Flores

Asesores:

Dr. Esaú Villatoro Tello

Mtra. Gabriela Ramírez de la Rosa

Mexico, D.F.

Noviembre 2019