Author Profiling in Informal and Formal Language Scenarios Via Transfer Learning

Perfilamiento de autor en escenarios lingüísticos informales y formales mediante aprendizaje por transferencia

Daniel Escobar-Grisales[1];
Juan Camilo Vásquez-Correa[2];
Juan Rafael Orozco-Arroyave[3]

[1] Universidad de Antioquia, Medellín-Colombia, daniel.esobar@udea.edu.co
[2] Universidad de Antioquia, Medellín-Colombia; Friedrich Alexander Universität, Erlangen Nürnberg-Germany; Pratech Group, Medellín-Colombia, jcvasquez@pratechgroup.com
[3] Universidad de Antioquia, Medellín-Colombia; Friedrich Alexander Universität, Erlangen Nürnberg-Germany, rafael.orozco@udea.edu.co

ISSN-p: 0123-7799
ISSN-e: 2256-5337

Vol. 24, **nro.** 52, e2166, 2021

Recibido: 28 junio 2021
Aceptado: 18 noviembre 2021
Disponible: 17 diciembre 2021

Cómo citar / How to cite

D. Escobar-Grisales; J. C. Vásquez-Correa; J. R. Orozco-Arroyave, "Author Profiling in Informal and Formal Language Scenarios Via Transfer Learning", *TecnoLógicas*, vol. 24, nro. 52, e2166, 2021. https://doi.org/10.22430/22565337.2166

Abstract

The interest in author profiling tasks has increased in the research community because computer applications have shown success in different sectors such as security, marketing, healthcare, and others. Recognition and identification of traits such as gender, age or location based on text data can help to improve different marketing strategies. This type of technology has been widely discussed regarding documents taken from social media. However, its methods have been poorly studied using data with a more formal structure, where there is no access to emoticons, mentions, and other linguistic phenomena that are only present in social media. This paper proposes the use of recurrent and convolutional neural networks and a transfer learning strategy to recognize two demographic traits, i.e., gender and language variety, in documents written in informal and formal language. The models were tested in two different databases consisting of tweets (informal) and call-center conversations (formal). Accuracies of up to 75 % and 68 % were achieved in the recognition of gender in documents with informal and formal language, respectively. Moreover, regarding language variety recognition, accuracies of 92 % and 72 % were obtained in informal and formal text scenarios, respectively. The results indicate that, in relation to the traits considered in this paper, it is possible to transfer the knowledge from a system trained on a specific type of expressions to another one where the structure is completely different and data are scarcer.

Keywords

Author profiling, Gender Recognition, Language variety recognition, Transfer learning, Natural language processing.

Resumen

El interés en tareas de perfilamiento de autor ha aumentado en la comunidad científica porque las aplicaciones han mostrado éxito en diferentes sectores como la seguridad, el mercadeo, la salud, entre otros. El reconocimiento e identificación de rasgos como el género, la edad, el dialecto o la personalidad a partir de datos de texto puede ayudar a mejorar diferentes estrategias de mercadeo. Este tipo de tecnología ha sido ampliamente discutida considerando documentos de redes sociales. Sin embargo, los métodos han sido pobremente estudiados en datos con una estructura más formal, donde no se tiene acceso a emoticones, menciones, y otros fenómenos lingüísticos que solo están presentes en redes sociales. Este trabajo propone el uso de redes neuronales recurrentes y convolucionales, y una estrategia de transferencia de aprendizaje para reconocer dos rasgos demográficos: el género y la variedad lingüística en documentos que están escritos en lenguajes informales y formales. Los modelos se prueban en dos bases de datos diferentes que consisten en Tuits (informal) y conversaciones de centros de llamadas (formal). Se obtienen precisiones del 75 % y del 68 % para el reconocimiento de género en documentos con una estructura informal y formal, respectivamente. Además, para el reconocimiento de variedad lingüística se obtuvieron precisiones del 92 % y del 72 % en documentos con una estructura informal y formal, respectivamente. Los resultados indican que, para los rasgos considerados, es posible transferir el conocimiento de un sistema entrenado en un tipo específico de expresiones a otro, donde la cantidad de datos es más escasa y su estructura es completamente diferente.

Palabras clave

Perfilamiento de autor; Reconocimiento de género y variedad lingüística; Transferencia de aprendizaje; Procesamiento del lenguaje natural.

1. INTRODUCTION

Author Profiling (AP) consists of recognizing demographic traits of a human being such as age, gender, personality, emotions, and others. Typically, its main aim is to create a user profile based on unstructured data. It has different applications in forensics, security, sales, marketing, healthcare, and many other sectors [1]. In e-commerce scenarios, this type of information gives companies advantages in competitive environments because it allows them to segment customers in order to offer personalized products and services, which strengthens their marketing strategies [2], [3]. Moreover, in chatbot systems, this type of technology is used to segment end users in order to provide them with personalized answers. Although most demographic factors are explicitly collected through a registration process, this approach could be limited given that most potential customers in online stores are anonymous. The automatic recognition of demographic variables such as gender or Language Variety (LV) according to geographic location can help to overcome these limitations [4].

Text data from customers can be obtained via transcripts of voice recordings, chats, surveys, and social media. These text resources can be processed to automatically recognize the gender or LV of the users. Different studies have applied Natural Language Processing (NLP) techniques to recognize Demographic Traits (DTs) of the author based on text data, mainly from social media posts [5]-[7]. Term Frequency-Inverse Document Frequency (TF-IDF) is a classical method to extract features from text data, and it is widely used to resolve different NLP tasks including AP [8], [9]. This feature represents each document based on the frequency of occurrence of the words in the document, weighted by their occurrence in all the documents in the corpus. In [10] and [11], the authors used TF-IDF to extract features from tweets in the PAN17 corpus [12], which has labels for gender and LVs of different Spanish speaking countries such as Argentina, Colombia, Venezuela, and others. By using a Support Vector Machine (SVM), they reported accuracies for gender classification and LV recognition around 81 % and 94 %, respectively. A similar approach was presented in [13], where the frequency of female and male emojis was used to recognize gender and LV. The authors reported an accuracy of 83.2 % in the PAN17 [12] corpus for gender recognition and 96.2 % for LV classification. In spite of the high accuracy reported in [13] and [14], this type of methodology would not be accurate to model text data written in more formal scenarios such as customer reviews, product surveys, opinion posts, and customer service chats, which have a different structure compared to text data available in social media. Moreover, these language features highly depend on the corpus, reducing generalization to other domains. For instance, some studies, such as [15] and [16], have concluded that females use emoticons more often than males, while another study [17] concluded the opposite.

Recently, text representations based on word embeddings have been successful in different applications including AP. In [18], the authors proposed a system to classify the gender of people who wrote 100,000 posts taken from Weibo (a Chinese social network like Tweeter) based on Word2Vec. The system achieved an accuracy of 62.9 %, which is nearly 3 % better than that of the human judgments reported in this corpus. This fact proves that the problem of recognizing gender in written texts is very hard, even for human readers. Regarding LV recognition tasks, Word2Vec has been successfully used. In [19], the authors represented words based on a Word2Vec model. The average word embedding computed along the words that form the post was used to represent each document in the HispaBlogs database, which has posts from five different countries: Argentina, Chile, Mexico, Peru, and Spain. They reported an accuracy of 73.6 % in LV recognition.

Deep Neural Networks (DNNs), such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been widely explored for various NLP tasks due to their high performance without the need for engineered features [20]-[22]. CNNs have shown to be efficient in author profiling tasks such as personality and author identification [23] and [24]. In [25], the author used a methodology based on word and sentence level embeddings with CNNs for gender and geographic identification. Word2Vec and FastText, which is a variation of Word2Vec at character-level, were employed. The model was evaluated in the VarDial corpus, which is composed of news articles in different Spanish dialects (Argentine, Venezuelan, Guatemalan, Spanish, and others). The proposed methodology was compared with a machine learning approach based on traditional features and SVM classifiers.

Accuracies up to 73 % and 92 % were achieved in the CNN approach for gender and geographic location identification, respectively. The results indicated that CNN models outperformed traditional machine learning algorithms for this type of AP application. RNNs have also been used to identify the author's demographic variables. In [26], the authors proposed a methodology based on bidirectional Gated Recurrent Units (GRUs) and an attention mechanism for gender and LV recognition in the PAN17 corpus. They used a Word2Vec model as input for their deep learning architecture and reported accuracies of up to 72.2 % and 91.4 % for gender and LV recognition, respectively. Other studies have considered the use of embedding layers within neural networks to automatically learn text representations of the documents [25], [26]. The advantage of these models is that extracted embeddings are specifically designed for each corpus. However, these types of approaches must deal with out-of-vocabulary problems [27]. Moreover, the number of parameters in a deep architecture increases considerably, therefore a large amount of data is required to correctly fit the model.

According to the reviewed literature, AP has been mainly explored in social media scenarios, where the language is informal, and the documents do not follow a formal structure [28]. There is a gap between models trained on formal and informal written language because a model trained with formal language data for a specific purpose will not achieve comparable results in an informal language scenario, or vice versa [29]. Due to this reason, it is important to validate trained models to estimate demographic variables in both types of language: formal and informal. In addition, AP problems have been under-explored in documents about e-commerce or customer service interactions because, compared to data from social media, it is difficult to collect a large amount of such labeled data.

This paper proposes a methodology based on RNNs and CNNs for gender and LV recognition in informal and formal language scenarios. First, the models were trained and tested within the PAN17 corpus, which is a traditional dataset for AP in tweets [12]. The models originally trained using the PAN17 corpus were re-trained using a transfer learning strategy with data from call-center conversations, which are structured in a more formal language. The aim of this study is twofold: 1) to recognize gender and nationality in the Tweeter corpus and 2) to recognize gender and dialect (*Antioqueño* vs. *Bogotano*) in the customer service corpus. Accuracies of up to 75 % and 92 % were obtained for gender and nationality recognition, respectively, in the first corpus, and of up to 68 % and 72 % for gender and dialect recognition, respectively, in the customer service corpus. The results in the customer service corpus outperformed the accuracies obtained with a baseline model widely used in AP. Finally, the best models were used to compare inter- and intra-country LV recognition. In this study, *inter-country* refers to the recognition of the Colombian dialect among Spanish-speaking countries, and *intra-country*, to the recognition of LVs from different regions in Colombia. The results indicate that the proposed methodology is accurate for DT recognition in documents written either in formal or in informal language. Moreover, fine-tuned models using transfer learning showed that, despite the noise and lack of structure in documents written in informal language, they can be used to improve the accuracy of DT recognition in documents written in formal language.

2. DATA

2.1 PAN17

We used the Spanish data in the PAN17 corpus [12]. In this database, there are variants of Spanish from seven countries: Argentina, Chile, Colombia, Mexico, Peru, Spain, and Venezuela. The training set was composed of texts by 600 subjects from each country (300 female). Since each subject had 100 tweets, there was a total of 4,200 subjects and 420,000 tweets in the dataset. The test set comprised data from 400 subjects from each country (200 female), for a total of 2,800 subjects and 280,000 tweets. For the sake of comparison with previous studies, we kept the original training and test sets as in [30]. The training set was randomly divided into 80 % for training and 20 % to optimize the hyper-parameters of the models (development set). All the data distribution was performed subject-independent to avoid subject-specific bias and to guarantee a better generalization capability of the models.

2.2 Call-Center Conversations

Conversations between customers and call center agents of a pension administration company were collected. Transcripts of the conversations were manually generated by linguistics experts. Labels for gender and perceived dialect were assigned in each conversation. Formal language is typically used by customers when asking for a service, making a request, asking about certificates, and other questions about the service provided by the company. Nevertheless. This corpus was highly unbalanced. There were Colombian dialects with very few samples, and some conversations did not have a specific Colombian dialect. For these reasons, two sub-databases were built, depending on the DT. First, the gender corpus was composed of 220 samples (110 female). Second, the dialect corpus contained two classes that represent two Colombian dialects: "Antioqueño" (from Antioquia) and "Bogotano" (from Bogotá). In the second corpus, each class had 80 samples. These two sub-databases shared 130 common customers, 76 female and 72 from Bogotá. A summary of the metadata of these corpora is provided in Table 1.

Table 1. Description of the sub-databases created based on the call-center conversation corpus for each label
Source: Created by the authors.

Label	Class	Samples	Words per conversation
Gender	Female	110	607.5 ± 460.9
	Male	110	592.4 ± 497.9
LV	*Bogotano*	80	556.8 ± 316.9
	Antioqueño	80	581.3 ± 471.1

3. METHODS

We implemented two Deep Learning (DL) architectures in this case: an RNN with bidirectional Long Short-Term Memory (LSTM) cells, and a CNN with multiple temporal resolutions. These networks were trained with data from the PAN17 corpus. Then, a transfer learning strategy was applied to recognize the trait (gender or dialect) from the call-center conversation data.

3.1 Bidirectional Long Short-Term Memory

The main idea of RNNs is to model a sequence of feature vectors based on the assumption that the output depends on the input features at the present time-step and on the output at the previous time-step. Conventional RNNs have a causal structure, i.e., the output at the present time-step only contains information from the past. However, many applications require information from the future [31]. Bidirectional RNNs are created to address such requirement by combining a layer that processes the input sequence forward through time with an additional layer that moves backwards the input sequence. Traditional RNNs also exhibit a vanishing gradient problem, which appears when long temporal sequences are modeled. LSTM layers have been proposed to solve this vanishing gradient problem by including a long-term memory to produce paths where the gradient flows for long duration sequences, such as sentences of a tweet or the ones that appear in a conversation with a call-center agent [32]. We propose the use of a Bidirectional LSTM (Bi-LSTM) network for our application. These architectures are widely used for different NLP tasks such as sentiment analysis in social media and product reviews [33], [34], and [35]. Figure 1 presents a scheme of the implemented architecture. Words from the data are represented using a word-embedding layer. The input to the Bi-LSTM layer consists of k d-dimensional word-embedding vectors, where k is the length of the sequence. The final decision about the DTs of the subject is made at the output layer by using the Softmax activation function.

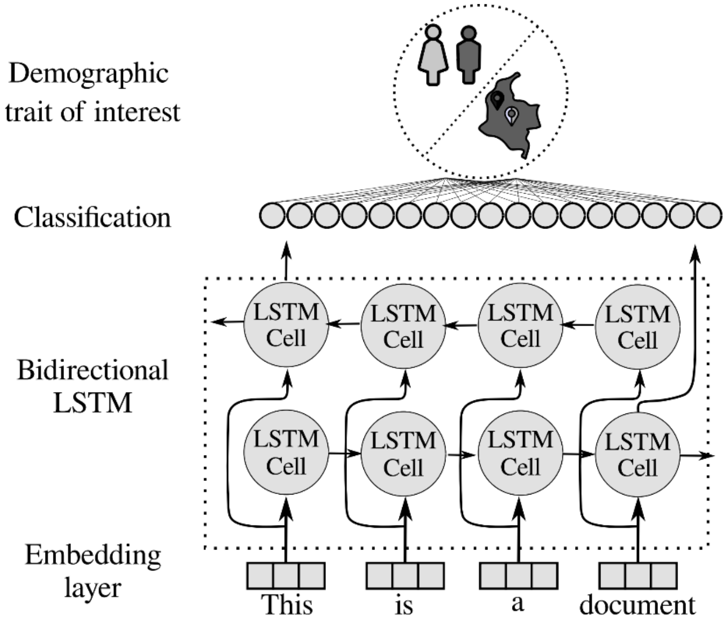

Figure 1. Bi-LSTM architecture for gender classification of a tweet. Source: Created by the authors.

3.2 Convolutional Neural Network

CNN-based architectures are designed to extract sentence representations by a composition of convolutional layers and a max-pooling operation over all the word embeddings that compose the embedding matrix. We propose the use of a parallel CNN architecture with different filter orders to exploit different temporal resolutions at the same time. Details of the architecture can be found in Figure 2. The output of the word-embedding layer is convolved with filters of n different orders, where n denotes the number of elements in an n-gram. The proposed CNN computes the convolution only in the temporal dimension. After convolution, a max-pooling operation is applied to reduce redundant information. Finally, a fully connected layer is employed for classification using a Softmax activation function.

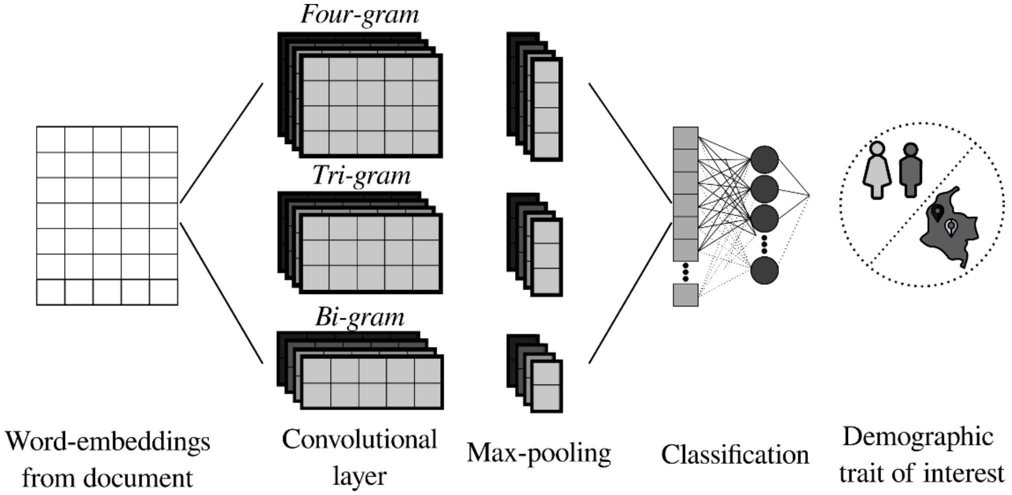

Figure. 2. CNN architecture for demographic trait recognition. Source: Created by the authors.

3.3 Training

The networks described in this article were implemented in Tensorflow 2.0 and trained with a sparse categorical cross-entropy loss function using an Adam optimizer. An early stopping criterion was used to stop training when the validation loss was not improved after 10 epochs. The embedding dimension d was set to 100. The vocabulary size for the tokenizer was set to 5,000 for experiments with the PAN17 corpus and to 1,500 for experiments with call-center conversations. This number was computed as the number of words with a frequency higher than 5 % of the number of documents in the training set of each corpus. The hyper-parameters were optimized based on the validation accuracy and the simplest model.

3.4 Transfer learning

We tested two approaches for the call-center conversation data: (1) training the network only using the data from the corresponding corpus and (2) training the model via transfer learning by using a pre-trained model generated with the PAN17 corpus. Regarding the transfer learning experiment, the most accurate model for the PAN17 data was fine-tuned but freezing the embedding layer in order to keep the tokenizer and a larger vocabulary.

Experiments without freezing the embedding layer were also performed, but the results were not satisfactory. The motivation for using transfer learning here is to test whether the knowledge acquired by a model trained with text data in informal language is useful to improve AP systems based on texts with formal language.

4. RESULTS AND DISCUSSION

Two experiments were performed to recognize each DT. The first one consisted of evaluating short sequences of texts; thus, the architectures were trained and the DT (gender or LV) of the subject was computed based on the average classification scores of all short texts by the same subject. In this experiment, in the PAN17 corpus, each tweet was a short text; in turn, each call-center transcript was divided into 60-word chunks, like in [26]. The second experiment consisted of evaluating long texts. In this case, the complete text data of the subjects was fed to the network at the same time. In the PAN17 corpus all tweets by the same subject were concatenated, and in the call-center corpus, the complete transliteration of each conversation was included. This strategy was evaluated using only the CNN-based approach because longer segments produced vanishing gradient problems in the Bi-LSTM network.

The experiments with the call-center conversation data were compared with a baseline model, where the documents were represented using TF-IDF and an SVM with a Gaussian kernel used to classify the DTs. The vocabulary used in the baseline model was the same as that used in the embedding layer for the proposed models. This baseline model was only tested in the long text approach following the methodology used in different studies. This type of baseline has been used successfully as a benchmark in several databases for AP, including the PAN17 corpus [10] and [11].

Additionally, in this paper we present a cluster analysis based on k-means in order to implement a customer segmentation strategy using the prediction scores of our neural networks. We focused especially on Colombian DT recognition for inter-country assessment using the PAN17 corpus, and intra-country recognition using the call-center conversation corpus. The number of clusters was defined using the elbow method and the Kneedle algorithm [36].

4.1 AP in informal structured language (PAN17 corpus)

The results obtained for the PAN17 corpus considering only Spanish data are shown in Table 2.

Table 2. Accuracies of gender and Language Variety (LV) classification in the PAN17 database. All values are given in %. Source: Created by the authors.

	Short texts		Long texts
	Bi-LSTM	CNN	CNN
		Gender	
Acc per tweet	60.5	61.1	-
Acc per subject	71.3	71.4	75.9
Precision	69.6	81.1	75.6
Recall	72.0	68.0	76.1
F1-score	70.8	73.9	75.8
		LV	
Acc per tweet	44.1	48.5	-
Acc per subject	83.3	89.8	92.3
F1-score	83.4	89.8	92.3
Kappa score	80.5	88.0	91.0

The analysis based on long texts to classify gender showed an improvement of 4 % compared to that based on short texts. The improvement when classifying LV was about 3 %.

4.2 AP in formal structured language (call-center conversation corpus)

All the experiments performed with this corpus were validated following a 10-fold cross-validation strategy due to the small size of the dataset. Table 3 shows the results of the baseline model, where accuracies of up to 57 % and 63 % were obtained for gender and LV classification, respectively.

Table 3. Results of gender and dialect classification in the call-center conversation data using TF-IDF and SVM (baseline model). Source: Created by the authors.

	Gender	Dialect
Acc per subject	57.50 +/- 3.50	63.400 +/- 2.800
Precision	58.00 +/- 3.90	63.500 +/- 2.900
Recall	54.60 +/- 5.10	63.500 +/- 4.700
F1-Score	56.20 +/- 4.20	63.400 +/- 3.000
C; γ	10.00; 0.01	10.000; 0.001

Table 4 reports the AP results of the proposed models for the call-center conversations obtained with and without applying transfer learning. The highest accuracy was obtained with long texts, in the same way as in the PAN17 corpus. In addition, note that, for both DTs, the accuracy improved by up to 13 % when the transfer learning strategy was applied. In gender recognition, the base model was not very accurate; thus, the knowledge transferred to the target model did not cause a significant improvement, considering the complexity of the architecture and the number of samples available in the target model.

Table 4. Results of gender and dialect classification in the call-center conversation data. TL: transfer learning. Source: Created by the authors.

	Short texts				Long texts	
	Bi-LSTM		CNN		CNN	
			Gender			
	no TL	TL	no TL	TL	no TL	TL
Acc per text	52.7 +/- 6.43	51.6 +/- 5.07	57.9 +/- 9.2	58.3 +/- 6.48	-	-
Acc per	54.2 +/- 10.1	56.4 +/- 12.1	65.9 +/- 12.7	62.9 +/- 14.9	56.9 +/- 9.5	68.7+/- 10.3
Precision	65.3 +/- 29.2	55.0 +/- 13.8	52.0 +/- 22.2	61.1 +/- 17.9	48.0 +/- 27.0	70.8 +/- 13.6
Recall	53.3 +/- 22.7	56.7 +/- 13.2	70.4 +/- 17.2	64.6 +/- 15.2	54.9 +/- 37.8	65.8 +/- 13.2
F1-Score	55.2 +/- 19.8	55.2 +/- 11.9	57.8 +/- 19.9	61.1 +/- 15.7	47.1 +/- 28.1	67.1 +/- 10.6
			Dialect			
Acc per text	51.8 +/- 12.1	57.3 +/- 5.74	60.4 +/- 7.9	59.8 +/- 6.5	-	-
Acc per	55.8 +/- 16.6	63.0 +/- 11.1	66.2 +/- 12.1	67.8 +/- 12.0	59.4 +/- 14.5	72.8 +/- 11.7
Precision	52.8 +/- 27.5	62.9 +/- 17.8	62.8 +/- 19.7	68.2 +/- 17.0	54.4 +/- 30.9	70.5 +/- 15.5
Recall	71.5 +/- 36.8	68.2 +/- 22.8	75.4 +/- 20.5	73.3 +/- 22.2	51.7 +/- 35.3	75.9 +/- 19.1
F1-Score	55.6 +/- 25.9	62.3 +/- 16.2	66.4 +/- 16.7	67.7 +/- 15.6	48.1 +/- 28.4	72.1 +/- 15.2

Additionally, the models that used transfer learning outperformed the results obtained with the baseline model for both DTs. However, the results obtained with the baseline model show lower standard deviations, which likely indicates that baseline methods baseline models.

4.3 Analysis of recognized Colombian DTs for user segmentation (inter- vs. intra-country)

Figure 3 shows the results of a cluster analysis using all the samples of the test set from the PAN17 corpus. We used the best resulting model from the previous experiments in order to employ user segmentation strategies. We plotted the gender score in the horizontal axis vs. the probability of being classified as Colombian in the vertical axis. These data were obtained from the output of the CNNs after the Softmax activation function. The results indicate the presence of three clusters, where 95.2 % of subjects in Cluster 1 are Colombian, while 97 % of the subjects in Clusters 2 and 3 are non-Colombian. Regarding gender, Cluster 2 is mainly formed by female subjects (75.5 %), while Cluster 3 is formed by 75.2 % male subjects. Cluster 1 does not have a dominant gender. In addition, note that the Colombian dialect recognition based on text is more accurate than gender recognition; however, in relation to non-Colombian samples, each cluster is composed of at least 75 % subjects with a specific gender.

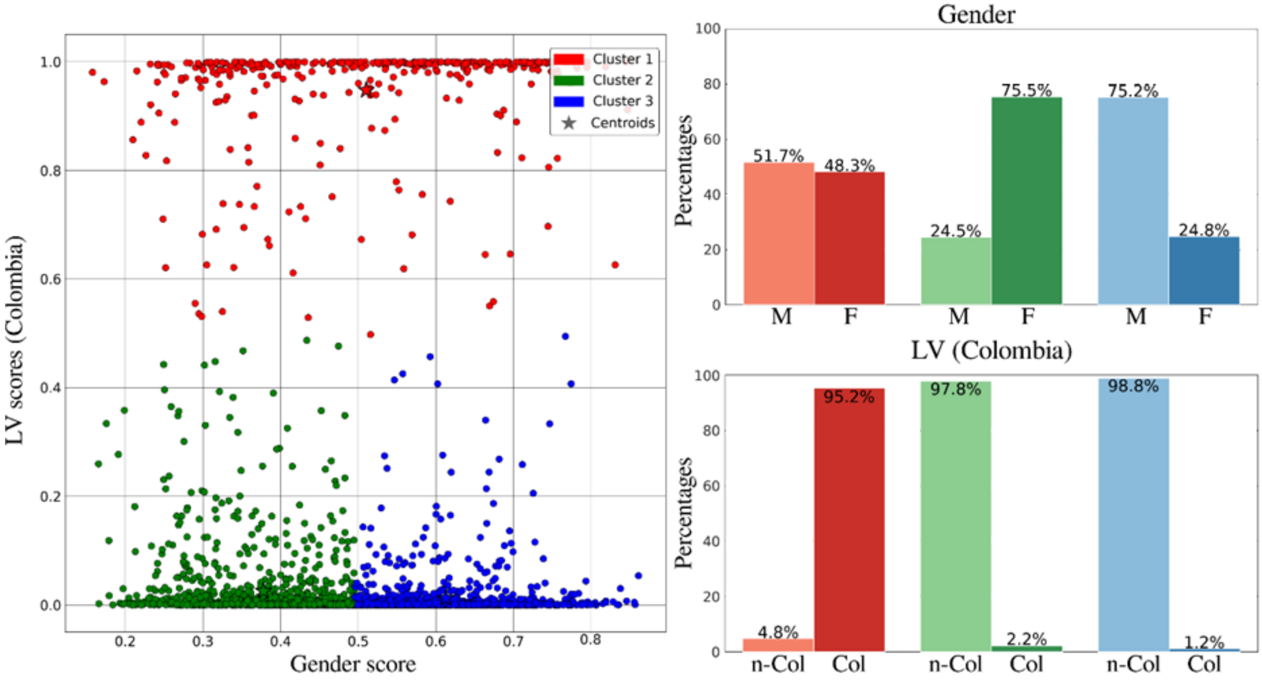

Figure 3. Results of k-means experiments using the scores of the best models for Colombian DT inter-country recognition. M: Male, F: Female, LV: Language Variety, n-Col: non-Colombian, Col: Colombian
Source: Created by the authors.

Figure 4 shows the results of the intra-country analysis of the samples in the call-center conversation corpus. There was a total of 130 subjects, distributed as 72 *Bogotanos* and 58 *Antioqueños*, and 77 female and 53 male users. According to Figure 4, Cluster 1 is composed mainly of subjects from Bogotá, Cluster 2 of subjects from Antioquia, and Cluster 3 is slightly balanced in terms of dialect with a larger number of subjects from Bogotá. Regarding gender, only Cluster 3 has a larger percentage of female subjects. The other two clusters are not gender specific.

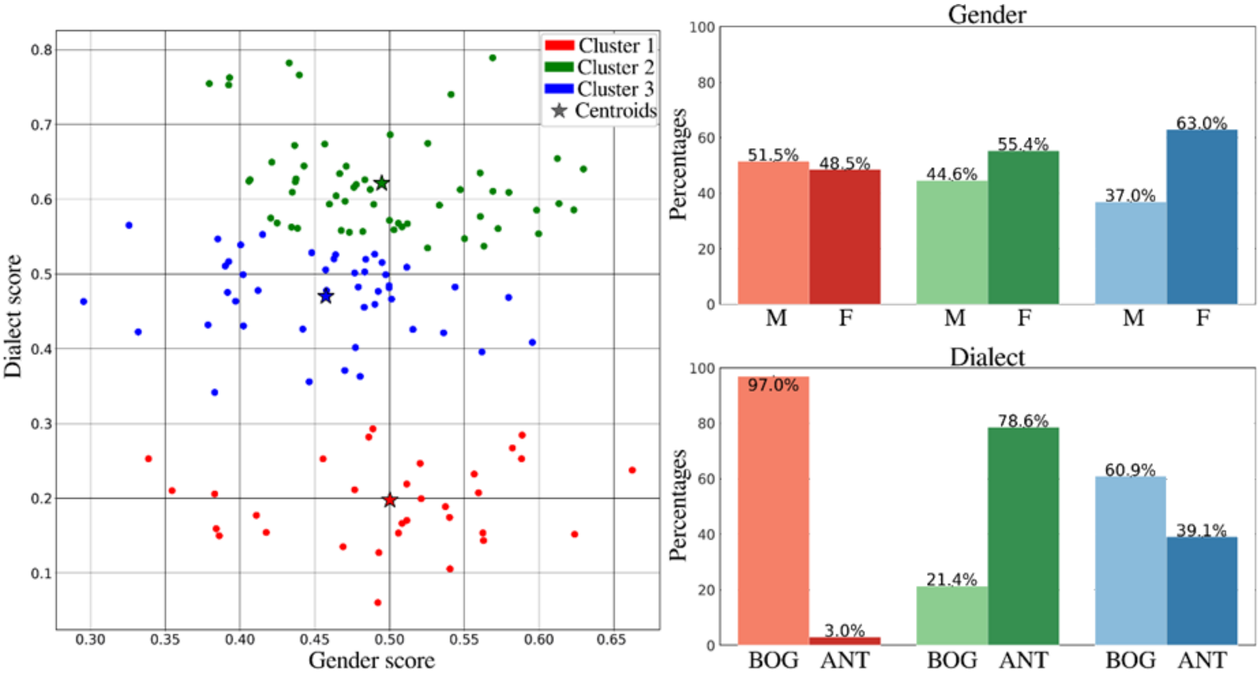

Figure. 4. Results of k-means experiments using the scores of the best models for intra-country DT recognition. M: Male, F: Female, BOG: *Bogotano*, ANT: *Antioqueño*. Source: Created by the authors.

In both approaches, the subjects tend to be grouped according to their LV. This is better observed in the inter-country analysis, but it also occurs in its intra-country counterpart. This can be explained by the fact that the differences in dialects in the same country are more subtle than those observed among different countries that share the same native language.

In addition, gender-dependent clusters are created in the inter-country scenario. Conversely, in the intra-country analysis, the clusters are more gender-balanced, although in some clusters there is a slight tendency toward a specific gender.

5. CONCLUSIONS

In this study, we proposed a methodology for AP in which two DTs, i.e., gender and LV, are automatically recognized in informal texts from social media posts and formal texts of call-center conversations. Different deep learning models were evaluated, including CNNs and LSTMs. We implemented a transfer learning approach where base models are pre-trained with data collected from social networks and then fine-tuned with call-center conversation data, which have a more formal structure than the social media posts used for pre-training.

The results indicate that it is possible to classify the gender and LV of a subject based on his/her social media posts, with accuracies of up to 75 % and 92 %, respectively. Regarding formal scenarios, we obtained accuracies of up to 68 % and 72 % for gender and dialect recognition, respectively. These results outperformed those obtained with a baseline model using TF-IDF in combination with an SVM classifier. The use of a transfer learning strategy improved the accuracy in scenarios where it is more difficult to collect data, like in call-center conversations, which suggests that such strategy is suitable for companies or sectors where it is not possible to create large datasets from scratch. The models that use transfer learning are also more stable and generalize better than others where the neural networks are trained from scratch. Furthermore, the knowledge acquired by the models to recognize LVs of Spanish-speaking countries can be successfully used to fine-tune models that recognize more subtle LVs, such as those inside the same country.

We believe that these results are very positive because they show that AP can benefit from large amounts of text data available in other domains, such as social media. Even though the accuracy of the models does not seem to be very high, (especially for gender recognition in call-center conversations), it is relevant because other studies, such as [18], have reported accuracies under 61 % using human judgment for gender recognition based on text data. The proposed approaches can be extended to other applications related to AP such as age, personality, and educational attainment, which would allow for the building of more complete and specific subject/customer profiles.

6. ACKNOWLEDGMENTS

This study was supported by Pratech Group SAS and CODI-UdeA (grants # PI 2019-24110 and PRG2020-34068). The authors would like to thank the Natural Language Engineering Laboratory of the Universidad Politécnica de Valencia and Pratech Group SAS for giving us access to the databases used in this paper.

CONFLICTS OF INTEREST

The Authors declare no conflict of financial, professional, or personal interests that may inappropriately influence the results that were obtained or the interpretations that are proposed.

AUTHOR CONTRIBUTIONS

Daniel Escobar-Grisales: statistical analysis, data curation, figure design, and writing of the first draft.

Juan Camilo Vasquez-Correa: study conception, study design, statistical design, writing of the first draft, review, and critique.

Juan Rafael Orozco-Arroyave: funding acquisition, study design, writing, review, and critique.

All the authors take responsibility for the integrity of the data and the accuracy of the data analysis.

7. REFERENCES

[1] F. Chiu Hsieh; R. F. Sandroni Dias; I. Paraboni, “Author profiling from Facebook corpora,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* pp. 2566-2570, 2018.URL

[2] O. Dogan; B. Oztaysi, *“Gender prediction from classified indoor customer paths by fuzzy C-medoids clustering,”* in *Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making INFUS 2019. Advances in Intelligent Systems and Computing*, vol 1029. Springer, Cham., pp. 160-169. https://doi.org/10.1007/978-3-030-23756-1_21

[3] R. Hirt; N. Kühl; G. Satzger, “Cognitive computing for customer profiling: meta classification for gender prediction,” *Electron. Mark.,* vol. 39, no. 1, pp. 93-106, Feb. 2019. https://doi.org/10.1007/s12525-019-00336-z

[4] D. Fernandez-Lanvin; J. de Andres-Suarez; M. Gonzalez-Rodriguez; B. Pariente-Martinez, "The dimension of age and gender as user model demographic factors for automatic personalization in e-commerce sites," *Comput. Stand. Interfaces,* vol. 59, pp. 1-9, Aug. 2018. https://doi.org/10.1016/j.csi.2018.02.001

[5] M. Arroju; A. Hassan; G. Farnadi, "Age, gender and personality recognition using tweets in a multilingual setting Notebook for PAN at CLEF 2015". in *6th Conference and Labs of the Evaluation Forum (CLEF)*, 2015, pp. 23-31. URL

[6] A. Nemati, "Gender and Age Prediction Multilingual Author Profiles Based on Comments". in *FIRE (Working Notes)*, 2018.URL

[7] P. Mishra; M. Del Tredici; H. Yannakoudakis; E. Shutova, "Author profiling for abuse detection". in *Proceedings of the 27th international conference on computational linguistics*, 2018. URL

[8] B. G. Gebre; M. Zampieri; P. Wittenburg; T. Heskes, "Improving native language identification with TF-IDF weighting". in *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications,* 2013, pp. 216-223. URL

[9] K. M. Alomari; H. M. ElSherif; K. Shaalan, "Arabic tweets sentimental analysis using machine learning". in Advances in *Artificial Intelligence: From Theory to Practice. IEA/AIE 2017. Lecture Notes in Computer Science*, vol 10350. Springer, Cham. https://doi.org/10.1007/978-3-319-60042-0_66

[10] I. Markov; H. Gómez-Adorno; G. Sidorov, "Language-and subtask-dependent feature selection and classifier parameter tuning for author profiling Notebook for PAN at CLEF 2017," *CLEF (Working Notes)*, 2017. URL

[11] M. Martinc; I. Skrjanec; K. Zupan; S. Pollak, "PAN 2017: Author profiling-gender and language variety prediction," in *CLEF (Working Notes)*, 2017. URL

[12] F. Rangel; P. Rosso; M. Potthast; B. Stein, "Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in twitter," in *Working notes papers of the CLEF*, pp. 1613-0073, 2017. URL

[13] A. Basile; G. Dwyer; M. Medvedeva; J. Rawee; H. Haagsma; M. Nissim, "N-gram: New Groningen author-profiling model," Jul. 2017. URL

[14] M. Potthast; T. Gollub; F. Rangel; P. Rosso; E. fstathios Stamatatos; B. Stein, "Improving the reproducibility of PAN's shared tasks," in *Information Access Evaluation. Multilinguality, Multimodality, and Interaction. CLEF 2014. Lecture Notes in Computer Science,* vol 8685. Springer, Cham, 2014, pp. 268-299. https://doi.org/10.1007/978-3-319-11382-1_22

[15] M. L. Newman; C. J. Groom; L. D. Handelman; J. W. Pennebaker, "Gender differences in language use: An analysis of 14,000 text samples," *Discourse Processes,* vol. 45, no. 3, pp. 211-236, Jun. 2008. https://doi.org/10.1080/01638530802073712

[16] D. Rao; D. Yarowsky; A. Shreevats; M. Gupta, "Classifying latent user attributes in twitter," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents - SMUC '10*, 2010, pp. 37-44. https://doi.org/10.1145/1871985.1871993

[17] H. A. Schwartz *et al.,* "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS one,* vol. 8, no. 9, e73791, Sep. 2013. https://doi.org/10.1371/journal.pone.0073791

[18] W. Li; M. Dickinson, "Gender prediction for Chinese social media data," in Proceedings of Recent Advances in Natural Language Processing, Varna, Bulgaria, 2017, pp. 438-445. https://doi.org/10.26615/978-954-452-049-6_058

[19] M. Franco-Salvador; G. Kondrak; P. Rosso, "Bridging the native language and language variety identification tasks", *Procedia Computer Science*, vol. 112, pp. 1554-1561, 2017. https://doi.org/10.1016/j.procs.2017.08.068

[20] M. E. Aragón; A. P. López-Monroy, "Author profiling and aggressiveness detection in Spanish tweets: Mex-a3t 2018," in *IberEval@SEPLN,* 2018, pp. 134-139.

[21] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Doha, 2014, pp. 1746-1751. https://doi.org/10.3115/v1/D14-1181

[22] N. Kalchbrenner; E. Grefenstette; P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, 2014, pp. 655-665. https://doi.org/10.3115/v1/P14-1062

[23] N. Majumder; S. Poria; A. Gelbukh; E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 74-79, Mar. 2017. https://doi.org/10.1109/mis.2017.23

[24] S. Ruder; P. Ghaffari; J. Breslin, "Character-level and multi-channel convolutional neural networks for large-scale authorship attribution," *ArXiv,* Sep. 2016. URL

[25] H. Gómez-Adorno *et al.*, "A convolutional neural network approach for gender and language variety identification," *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, pp. 4845-4855, May. 2019. https://doi.org/10.3233/JIFS-179032

[26] D. Kodiyanet, “Author profiling with bidirectional RNNs using attention with GRUs,” Notebook for PAN at CLEF 2017. URL

[27] J. V. Lochter; R. M. Silva; T. A. Almeida, “Deep learning models for representing out-of-vocabulary words”. in *Brazilian Conference on Intelligent Systems*. Springer, Cham, 2020, pp. 418-434. https://doi.org/10.1007/978-3-030-61377-8_29

[28] M. González Bermúdez, “An analysis of twitter corpora and the differences between formal and colloquial tweets,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3153-3159. URL

[29] J. Gu; Z. Yu, “Data annealing for informal language understanding tasks,” *arXiv,* Apr. 2020. URL

[30] M. Potthast, F. Rangel; M. Tschuggnall; E. Stamatatos; P. Rosso; B. Stein, “Overview of PAN’17”. in CLEF 2017: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer, Cham, pp 275-290. https://doi.org/10.1007/978-3-319-65813-1_25

[31] D. W. Otter; J. R. Medina; J. K. Kalita, “A survey of the usages of deep learning for natural language processing,” *IEEE Trans. Neural Networks Learn. Syst., vol.* 32, no. 2, pp. 604-624, Feb. 2021. https://doi.org/10.1109/TNNLS.2020.2979670

[32] A. Torfi; R. A. Shirvani; Y. Keneshloo; N. Tavvaf; E. A Fox, “Natural language processing advancements by deep learning: A survey.” ArXiv, Mar. 2020. URL

[33] L. Arras; G. Montavon; K. R. Müller; W. Samek, “Explaining recurrent neural network predictions in sentiment analysis,” *in proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Copenhagen, 2017. https://doi.org/10.18653/v1/W17-5221

[34] S. Minaee; E. Azimi; A. Abdolrashidi, “Deep-sentiment: Sentiment analysis using ensemble of CNN and bi-LSTM models,” ArXiv, Apr. 2019. URL

[35] J. Trofimovich, “Comparison of neural network architectures for sentiment analysis of Russian tweets,” in *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue*, Moscow. 2016, pp. 50-59. http://www.dialog-21.ru/media/3380/arkhipenkoetal.pdf

[36] V. Satopaa; J. Albrecht; D. Irwin; B. Raghavan, “Finding a" kneedle" in a haystack: Detecting knee points in system behavior,” in *31st International Conference on Distributed Computing Systems Workshops*, Minneapolis, 2011, pp. 166-171. https://doi.org/10.1109/ICDCSW.2011.20