IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 9, 2021, accepted July 29, 2021, date of publication August 9, 2021, date of current version August 17, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3103697*

# A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis

**FLOR MIRIAM PLAZA-DEL-ARCO, M. DOLORES MOLINA-GONZÁLEZ, L. ALFONSO UREÑA-LÓPEZ, AND MARÍA TERESA MARTÍN-VALDIVIA**

Department of Computer Science, Advanced Studies Center in Information and Communication Technologies (CEATIC), Universidad de Jaén, 23071 Jaén, Spain

Corresponding author: Flor Miriam Plaza-del-Arco (fmplaza@ujaen.es)

This work was supported in part by the Grant from European Regional Development Fund (ERDF), LIVING-LANG Project under Grant RTI2018-094653-B-C21, and in part by the Ministry of Science, Innovation and Universities Scholarship through the Spanish Government under Grant FPI-PRE2019-089310.

**ABSTRACT** The rise of social media platforms has significantly changed the way our world communicates, and part of those changes includes a rise in inappropriate behaviors, such as the use of aggressive and hateful language online. Detecting such content is crucial to filtering or blocking inappropriate content on the Web. However, due to the huge amount of data posted every day, automatic methods are essential for identifying this type of content. Seeking to address this issue, the Natural Language Processing community is increasingly involved in testing a wide range of techniques for hate speech detection. While achieving promising results, these techniques consider hate speech detection as the sole optimization objective, without involving other related tasks such as polarity and emotion classification that are strongly linked to offensive behavior. In this paper, we propose the first Multi-task approach that leverages the shared affective knowledge to detect hate speech in Spanish tweets, using a well-known Transformer-based model. Our results show that the combination of both polarity and emotional knowledge helps to detect hate speech more accurately across datasets.

**INDEX TERMS** Natural language processing, sentiment analysis, multi-task learning, Spanish hate speech, offensive language.

## I. INTRODUCTION

In recent decades, people have been getting hooked on social media platforms as a way of relating and connecting to other people. More and more users are expressing and sharing their opinions, inner thoughts, or emotions through social networks like Twitter and Facebook. However, sometimes these posts express prejudice and harmful content targeted to a specific individual or group. A complete and comprehensive definition for Hate Speech (HS) is given by General Policy Recommendation no. 15 of the European Commission[1] and defines it as ''the advocacy, promotion or incitement, in any form, of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of race, color, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status''.

Today, the vast and uncontrolled content posted every day on the Web makes difficult or even impossible to manually track the content of comments. One strategy employed to tackle this problem is through legislation. In order to prevent and counter the spread of HS online, in May 2016, the European Commission reached an agreement with Facebook, Microsoft, Twitter, and YouTube a ''Code of conduct on countering illegal HS online''.[2] During the course of 2018, Instagram, Snapchat, and Dailymotion joined the Code of Conduct. Jeuxvideo.com joined in January 2019, and TikTok announced their participation in the Code in September 2020. However, the EU Code of Conduct is hard to fulfill by these online platforms. Natural Language Processing (NLP) plays an important role as a powerful tool to try to tackle this problem. In recent years, the NLP community has trained

The associate editor coordinating the review of this manuscript and approving it for publication was Cheng Chin.

[1] http://hudoc.ecri.coe.int/eng?i=REC-15-2016-015-ENG

[2] https://bit.ly/2KI14cO

a variety of systems for detecting HS online, using a great range of techniques. The most common techniques used are traditional machine learning and deep learning approaches, which have yielded promising results.

On the other hand, in the context of Sentiment Analysis (SA), polarity and emotion classification from the text are two closely related research topics that have been investigated for a long time now, (especially polarity classification). Polarity classification focuses on determining the polarity of a document, sentence, or feature (*positive or negative*) and measuring the degree of the polarity expressed in the document [1]. Emotion classification aims to fine-grained automatic classification of texts on the basis of emotional categories. Most emotion analysis studies deal with six basic human emotions (*anger, fear, sadness, joy, surprise, and disgust*) as defined by Ekman [2]. According to a psychological study [3], negative sentiment messages are often indicators of emotions, such as *anger, disgust, fear or sadness*, and positive texts are related to *joy*. In the same way, abusive language and behavior are also inextricably linked to the emotional and psychological state of the speaker [4]. For instance, negative sentiments and emotions like *anger, disgust, fear, and sadness* are presented in HS messages as a number of studies have already revealed in recent years [5], [6]. Given the influence of emotions in HS messages, in this study, we investigated new ways of unearthing HS by modeling polarity and emotion analysis along with the HS detection task.

Although there are some studies on this topic for other languages – e.g. Portuguese [7], Italian [8] and Greek [9] –, and multilingual approaches [10]–[12], most of the resources and studies available are for English [13]–[15]. However, there is a huge demand for research into languages other than English [16]. Therefore, in this paper, we focus on Spanish a language whose presence on the Internet is growing every day.

In this study, we present a novel technique in classifying HS on social media. We evaluate our model on two datasets that contain tweets written in Spanish and annotated for HS. Our results show that models that learn multiple tasks at the same time (Multi-Task Learning (MTL)) achieve promising results. The main contributions of this paper can be summarized as follows:

- We perform HS detection for Spanish texts. HS detection is a worldwide phenomenon that involves other languages than English, including Spanish, which is the third most used language on the Web.
- We propose a MTL model using the monolingual Transformer-based model BETO that integrates polarity and emotion knowledge for HS detection, named $\text{MTL}_{sent+emo}$.
- We compare the MTL model performance with a monolingual Transformer-based model (our baseline) with the latest state-of-the-art results in HS detection for Spanish.
- We show the effectiveness of our proposed approach over the baseline model through detailed empirical evaluation results on two benchmark datasets and by analyzing the knowledge transfer from SA in the MTL proposed approach for HS detection.
- We perform an error analysis in order to gain deeper insight into the proposed MTL model performance. This analysis also allows us to identify some peculiarities of Spanish-speaking users while expressing HS.

The remaining structure of this paper is as follows. In Section II an overview of the related background literature is introduced. The data we used to evaluate our experiments are described in Section III. The proposed HS detection system is shown in Section IV. The experimental methodology and evaluation results are presented in Section V. Error analysis containing qualitative and quantitative analysis of the results are conducted in Section VI. Finally, the conclusion and directions for future research are presented in Section VII.

NOTE: This paper contains examples of potentially explicit or offensive content. They do not represent the views of the authors.

## II. RELATED WORK

In recent years, while HS continues to spread on Internet, the importance of addressing HS detection in textual information is becoming more and more significant in the NLP field, with a number of studies applying different machine learning approaches. Most of these studies focus on the detection of HS in social media, most notably on Twitter.

On the one hand, early studies experimented with traditional machine learning algorithms including Support Vector Machines, Random Forest, Decision Tree and Logistic Regression along with the combination of different types of syntactic, semantic, lexical, sentiment, and lexicon-based features [16]–[18].

On the other hand, given the promising results achieved by deep neural networks in several NLP tasks, extensive studies have recently explored the performance of these models in the task of HS detection with the Recurrent Neural Networks (RNNs) and Convolutional Neural Network (CNNs) being the most popular architectures for this task. For instance, in order to break the barrier of language dependency in the word embedding approach, [19] conducted an ensemble of RNN classifiers, incorporating various features associated with user-related information. [20] experimented with a robust system based on compositional RNNs able to handle even substantially noisy inputs, and reached competitive results for HS detection in English texts. In [21] authors developed a system for Twitter HS text identification based on two CNNs and feature embeddings including one-hot encoded character n-gram vectors and word embeddings, and they reported that the use of character n-gram does not help in the detection. More recently, Transformer-based models have made significant progress in most of the NLP tasks including text classification [22] and are currently applied to the HS identification. For instance, [23] proposed a multi-channel BERT model that integrates the hidden features of separate BERT models trained on different languages. The model

can capture a different semantic representation of different languages. The authors tested the model on three datasets from different competitions in different languages and it performed better or as well as previous state-of-the-art models. In [24] authors compared different Transformer-based models to address the task of HS identification in Spanish. They achieved state-of-the-art results with a monolingual pre-trained language model based on the Transformer mechanism and trained specifically on Spanish texts, BETO [25].

### A. SENTIMENT ANALYSIS ON HATE SPEECH

SA offers a valuable tool that helps to enhance the performance of machine learning classification systems, as shown in [26] and [27]. A few recent studies have investigated the benefit of using SA features for HS detection. For instance, in [28] the authors follow the idea that the concept of HS can be split into two main components *hate* and *speech* [29], and based on this they proposed a new definition of HS in the scope of emotional analysis: ''any emotional expression imparting opinions or ideas - bringing a subjective opinion or idea to an external audience - with discriminatory purposes''. To predict HS in texts, they employed an emotional approach using a combination of lexicon-based and machine learning approaches and concluded that the emotional knowledge contained in the text helps to enhance the accuracy of HS detection. [30] introduced a new attention mechanism that embeds emotional knowledge from texts to find the most important words for the task of offensive language identification. They incorporated this module into a hybrid bidirectional LSTM and CNN neural architecture able to capture both local and sequential information from text. This supports the hypothesis that the affective knowledge behind the text plays a significant role to offensive language identification, as it boosts the performance of the system. [6] proposed a framework to identify Facebook pages that potentially promote HS. In order to obtain the most negative posts and comments, they applied polarity and emotion analysis, based on the idea that hateful texts contain negative emotions and sentiments. Finally, [31] developed a method for automatic data augmentation and deployed affective bidirectional Transformers models on offensive language detection and HS identification for Arabic. They demonstrated that fine-tuning such affective models is useful, especially in the case of offensive language detection.

Related to Spanish, [32], [33] incorporated sentiment features into a supervised classifier. [33] proposed a system based on linguistic features, semantic similarity with a domain-oriented lexicon, sentiments (using the sentiment vocabulary weighted by the TF-IDF measure), word embeddings, topic modeling (both LDA and hashtags), and TF-IDF n-grams of words and characters. These features were filtered and the 3000 best were selected. The machine learning algorithm selected for classification was linear SVM. [32] proposed a linear kernel SVM trained on a text representation composed of a bag of words, a bag of characters, and an embedding of tweets computed from fastText sentiment-oriented word vectors.

The studies mentioned were all carried out using a single-task learning approach. Single Task Learning (STL) is a paradigm that updates the weight of neural networks using the input sequence of a single classification task involving a dataset.

### B. MULTI-TASK LEARNING FOR HATE SPEECH DETECTION

The literature points out that MTL [34] has been successfully tested on a multitude of machine learning tasks [35], including NLP problems such as machine translation [36], SA classification [37] and biomedical entity recognition [38]. However, only a very few recent studies have employed the MTL paradigm to address the problem of HS identification. For instance, [39] proposed a deep MTL framework based on a stacked CNN and GRU architecture to leverage useful information from multiple datasets related to HS including racism, sexism, and offensive language identification. Its MTL model achieved better performance compared to its single-tasking framework. Another two studies experimented with a MTL approach using polarity and emotion information to detect HS. Firstly, [40] tested a MTL system exploring the effect of adding polarity information to perform the task of offensive language identification in Arabic tweets. They based their research on the fact that HS and offensive content always bear negative polarity. Their results showed that polarity information is correlated with HS and offensive language identification. Secondly, [41] were the first to take into account emotional features in order to gain auxiliary knowledge through a MTL framework to detect abuse in English tweets. They propose different MTL models, and the best result was achieved by a Gated Double Encoder model based on BiLSTM encoders. Their experiments showed that emotion detection is beneficial to abuse detection tasks in the Twitter domain.

As SA has been shown to be beneficial for HS detection systems and since most studies have used SA within a STL model, our proposed approach is focused on the MTL paradigm and differs from the previous studies because a) we use a Transformer-based model in the proposed MTL approach; b) we explore the combination of both polarity and emotion knowledge in the MTL approach for Spanish HS detection.

## III. DATASETS

We carried out experiments with four different tweets corpora, one related to polarity (InterTASS), another related to emotion (EmoEvent), and the last related to HS (HatEval) and aggressiveness (MEX-A3T). The datasets are described below:

International TASS Corpus (InterTASS) was released in 2017 [42] with Spanish tweets and updated in 2018 with texts written in three different variants of Spanish from Spain, Costa Rica and Peru [43] and in 2019, with new texts written in two new Spanish variants: Uruguayan and Mexican [44].

**TABLE 1.** Number of tweets in InterTASS dataset.

| | Training | | | | | Test | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Class** | ES | CR | PE | UR | MX | ES | CR | PE | UR | MX |
| P | 510 | 341 | 321 | 443 | 472 | 594 | 366 | 435 | 469 | 525 |
| N | 741 | 453 | 335 | 559 | 757 | 663 | 549 | 485 | 587 | 745 |
| NEU | 444 | 373 | 885 | 427 | 271 | 449 | 371 | 544 | 372 | 230 |
| Total | 1,695 | 1,167 | 1,541 | 1,429 | 1,500 | 1,706 | 1,196 | 1,464 | 1,428 | 1,500 |

The corpus released in 2019 is the one used in this paper. Each tweet was labeled with its level of polarity, which could be labeled as Positive (P), Negative (N), Neutral (NEU), and none (NONE). Each tweet was annotated by at least three annotators. Table 1 shows the number of tweets in the training and test sets for each different label or the levels of opinion intensity (P, N, NEU), where NEU will include NONE and for each corpus of tweets written in Spanish variants from Spain (ES), Costa Rica (CR), Peru (PE), Uruguay (UR) and Mexico (MX).

EmoEvent [45] is a multilingual emotion dataset based on events that took place in April 2019. It focuses on tweets in the areas of entertainment, catastrophes, politics, global commemoration, and global strikes. The authors collected Spanish and English tweets from the Twitter platform. Then, each tweet was labeled with one of seven emotions, six Ekman's basic emotions plus the ''neutral or other emotions'' label. The labeling was done by three Amazon Mechanical Turkers. Focusing on the Spanish language, a total of 8,409 were labeled. Table 2 shows the number of tweets in the training and test set for each label (emotion).

**TABLE 2.** Number of tweets by emotion and event in the EmoEvent dataset.

| **Label** | Training | Test |
|---|---|---|
| joy | 1,455 | 360 |
| sadness | 809 | 200 |
| anger | 687 | 170 |
| surprise | 276 | 68 |
| disgust | 129 | 32 |
| fear | 77 | 19 |
| others | 3,310 | 817 |
| Total | 6,743 | 1,666 |

The first HS dataset used in this paper was provided by the organizers in SemEval 2019 Task 5: HatEval [46]. The task consisted of detecting hateful content in social media texts, specifically in Twitter posts, against two targets: women and immigrants. For the creation of the corpus known as, the HatEval dataset, the data was collected using a different time frame. The majority of tweets against women were derived from an earlier collection made in the context of two earlier challenges on misogynistic speech identification, whose collection phase began in July 2017 and ended in November 2017 [47], [48]. The remaining tweets were collected from July to September 2018. The dataset contains tweets composed of an identifier (id), the text of the tweet (text), and the mark of HS, which is 0 if the text is not hateful and 1 if the text is hate speech against women or immigrants. In the task, this dataset was divided into three small sets: train, dev, and test. For our experiments, the union of train and dev builds the training set which contains 2,921 not hateful tweets and 2,079 hateful tweets. The test set contained 940 non-hateful tweets and 660 hateful tweets. Table 3 shows the number of tweets in the training set and test set targeted at women and immigrants.

**TABLE 3.** Number of tweets in Spanish HatEval dataset.

| | Training | | | Test | | |
|---|---|---|---|---|---|---|
| **Class** | **Women** | **Immigrants** | **Total** | **Women** | **Immigrants** | **Total** |
| 0 (Non-HS) | 1,881 | 1,040 | 2,921 | 464 | 476 | 940 |
| 1 (HS) | 1,328 | 751 | 2,079 | 336 | 324 | 660 |
| Total | 3,209 | 1,791 | 5,000 | 800 | 800 | 1,600 |

The last dataset is MEX-A3T [49]. It was provided by the organizers in IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets [50]. They built a corpus of tweets to detect aggressiveness from Mexican accounts collected from August to November of 2017. They selected a set of terms that served as seeds for extracting the tweets. They used both words classified as vulgar and non-colloquial in the Dictionary of Mexicanisms of the Mexican Academy of the Language and hashtags classified as controversial according to the National Institute of Women. The hashtags were related to politics, sexism, homophobia, and discrimination. Tweets were collected taking into account their geo-location. They used Mexico City as the center and extracted all tweets that were within a radius of 500 km. Finally, the collected tweets were labeled by two people. The dataset contains tweets composed of an identifier (id), the text of the tweet (text), and the mark of aggressiveness, being 0 if the tweet is not aggressive and 1 if the tweet is aggressive. For our experiments, the corpus is divided into two parts, training and test sets. The non-aggressive class is the majority class in both partitions. The test set with the gold labels is not freely available but the organizers provide us with the evaluation results based on our predicted labels. Table 4 shows the number of tweets in MEX-A3T dataset per class.

## IV. SYSTEM DESCRIPTION

Following the notation of [51] with the binary classification of documents as a running example, transfer learning involves

**TABLE 4.** Number of tweets in MEX-A3T dataset per class.

| Class | Training tweets | Test tweets |
|---|---|---|
| 0 (Non-Aggressive) | 5,222 | 2,238 |
| 1 (Aggressive) | 2,110 | 905 |
| Total | 7,332 | 3,143 |

the concepts of a domain and a task. Given a source domain $D_S$ and learning task $T_S$, a target domain $D_T$ and learning task $T_T$, transfer learning aims to help improve the learning of the target predictive function $f_T$ (.) in $D_T$ using the knowledge in $D_S$ and $T_S$, where $D_S = D_T$ or $T_S = T_T$.

Transfer learning is widely used to develop models for solving tasks where the availability of large data to train is limited, as is the case with Spanish.

In this study, we follow the transfer learning taxonomy defined by [52] to introduce the MTL scenario explored in this study. In the scope of transfer learning, when there are different tasks and labeled data in a target domain, the taxonomy refers to inductive transfer learning. Within inductive transfer learning, there are two possible scenarios: sequential learning (if the tasks are learned sequentially), and MTL (if the tasks are learned simultaneously).

### A. SINGLE-TASK LEARNING

STL is a paradigm that updates the weight of neural networks using the input sequence of a single classification task involving a dataset. In order to establish a baseline in our study and compare the results with the MTL scenario, we use a STL model that uses the HS task as the sole optimization objective. To do so, we focus on pre-trained language models based on Transformer. The Transformer is an attention mechanism that learns contextual relations between words (or sub-words) in a text and includes two separate mechanisms: an encoder that reads the text input and a decoder that produces a prediction for the task [53]. In contrast to directional models, where the text input is read sequentially, the Transformer encoder reads the whole word sequence at once, which allows the model to learn the context of a word on the basis of its entire surroundings.

In particular, we experiment with a well-known model, the Bidirectional Encoder Representations from Transformers (BERT) [54]. As far as we know, there are two variants of BERT trained on Spanish texts: the Multilingual BERT (mBERT) and BETO [25]. mBERT was pre-trained on the concatenation of monolingual Wikipedia corpora from 104 languages, including Spanish but it does not provide a language detection mechanism therefore the word piece tokenizer could confuse languages. Moreover, it does not have any explicit procedures to encourage translation equivalent pairs to have similar representations. For this reason, we use the BETO model as our baseline since it was trained on Spanish data. We refer to this baseline as $\text{STL}_{BETO}$ in the paper. Specifically, we use the BETO cased checkpoint.[3]

[3]https://github.com/dccuchile/beto

We address two $\text{STL}_{BETO}$ scenarios to detect HS, one for the HatEval task and the other one for the MEX-A3T task:

- **HatEval**. The task consisted of detecting hateful content in Spanish tweets against two targets: women and immigrants.
- **MEX-A3T**. It focused on the detection of aggressive tweets in Mexican Spanish on Twitter. According to the authors in [55], aggressive language seeks to harm or hurt a group or individual by referring to or inciting violence. Therefore, aggression could involve HS.

### B. MULTI-TASK LEARNING

In the MTL scenario, the goal is to use the process of learning multiple tasks in order to improve the performance on each task [34]. These tasks are usually related and share some commonalities, though they may have different data or features. When the model learns these tasks, some clues from one task can be used to improve the other by sharing features. In this study, we use related tasks (HS detection, polarity classification, and emotion classification) sharing the same source of data: Twitter. For the polarity classification task, we use the InterTASS dataset and for emotion classification, we used the EmoEvent corpus. The aim is to check whether the use of polarity and emotion classification tasks assists in the identification of HS by applying a MTL scenario. The reason for incorporating polarity and emotion information to detect HS is that HS is usually emotional and expresses a negative emotion and polarity towards the recipient.

In the field of Deep Learning, MTL is typically implemented with either *hard* or *soft parameter sharing* of hidden layers. In the following, these two methods will be introduced.

- *Soft parameter sharing*. In this setting, each task has its model with its parameters. The distance between the parameters of the model is regularized in order to encourage similarity between the parameters.
- *Hard parameter sharing*. This technique is the most widely used approach to MTL in neural networks and was introduced by [34]. It consists of a single encoder that is shared and updated between all tasks while keeping several task-specific output layers [52].
  In this study, we use the Transformer-based model BETO which uses BERT encoder. Figure 1 shows our proposed MTL model. In this architecture, we use the data labeled for the n-tasks. As we have two different datasets related to HS (HatEval and MEX-A3T), we first trained the model on HatEval and n-1 tasks obtaining the evaluation for HatEval. Secondly, for evaluating our proposal, we trained the model on MEX-A3T and n-1 tasks obtaining the performance for MEX-A3T. As mentioned at the beginning of this Section, in our MTL model, the tasks used to improve the detection of HS are polarity and emotion classification. In the first step, all the inputs are converted to WordPieces [56], an extension of Byte Pair Encodings (BPE) [57]. The sequence has at most two segments, the first token

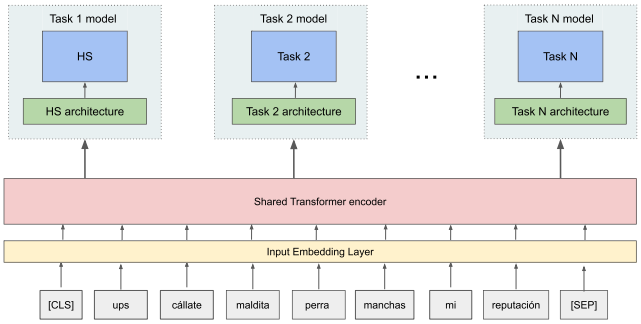


**FIGURE 1.** Proposed MTL system on the n-tasks. English translation: ups shut up fucking bitch stains my reputation.

[CLS] contains the special classification embedding and another special token [SEP] is used to separate segments. We take the first token [CLS] in the final hidden layers as the representation of the whole sequence. As we can see in Figure 1 the BERT encoder is shared by all three tasks simultaneously, in this way each task benefits from the other as they share features with each other and therefore the model can learn what these features are. Then, the task-specific output heads are created for each task, and task heads are attached to a common sentence encoder. The shared Transformer encoder ensures that during training, all upgrades will update the same encoder weights. Then, the layers are fine-tuned according to the given set of downstream tasks.

## V. EXPERIMENTS AND RESULTS

### A. DATA PRE-PROCESSING

The text pre-processing technique is a fundamental step in NLP systems since it helps improve the performance of the classifier and speed up the classification process. Since we are dealing with Twitter data, we perform a Twitter-specific data cleaning before including the tweets in the models. Tweets have specific peculiarities because they are written in colloquial language. Therefore, there are numerous challenges in tokenizing tweets such as word length, duplication, use of informal terms, noise (misspellings and slang words), user mentions, hashtags, and emojis. To tackle these challenges and remove the noise we apply the following practices to prepare the text for machine learning experiments using the ekphrasis library [58]:

- All tweets are converted to lowercase so words such as "HATE", "Hate", and "hate" will have the same syntax when converting text into features.
- URLs, emails, users' mentions, percentages, monetary amounts, time and date expressions, and phone numbers are normalized.
- Hashtags are unpacked and split into their constituent words.
- Elongated words and repeated characters in words are annotated and reduced.

| | |
|---|---|
| **Tweet** | @user tu sonrisa es tu superrr poder!!! #FelizDía |
| **Tokenization** | ['<user>', 'tu', 'sonrisa', 'es', 'tu', 'super', '<elongated>', 'poder', '!', '<repeated>', '<hashtag>', 'feliz', 'día', '</hashtag>'] |

**FIGURE 2.** Example of a tokenized tweet. English translation: Your smile is your superpower! happy day.

- Emojis are converted to its alias using the emoji module[4] of Python.
- The maximum sequence length is set to 80 words. Post padding is done if any sentence is less than 80.

Figure 2 shows an example tweet and the result of the pre-processing.

### B. EXPERIMENTAL SETUP

All the models has been implemented using PyTorch, a high-performance deep learning library [59] based on the Torch library. We trained the models on the training set and then we tested it on the test set for both datasets. The experiments were run on a single Tesla-V100 32 GB GPU with 192 GB of RAM.

We evaluated both HatEval and MEX-A3T datasets conducting four different experiments for each HS dataset:

1) In order to obtain the baseline, we evaluate the corresponding HS dataset with the Transformer-based BETO model, namely $\text{STL}_{BETO}$ method.
2) We perform the MTL approach proposed in this study and explained in Section IV-B. Specifically, for the MTL approach we experimented with three different configurations to detect HS:
   a) We train and evaluate the model on the corresponding HS dataset and InterTASS ($\text{MTL}_{sent}$).
   b) We train and evaluate the model on the corresponding HS dataset and EmoEvent dataset ($\text{MTL}_{emo}$).
   c) We train and evaluate the model on the corresponding HS dataset. InterTASS and EmoEvent ($\text{MTL}_{sent+emo}$).

We use grid search to tune the hyperparameters of the models on the development sets of the tasks (HatEval and MEX-A3T). The optimal values for each model and task are shown in this section. Across the two $\text{STL}_{BETO}$ experiments, MEX-A3T was fine-tuned BETO for three epochs, the learning rate was set to 4e-05, and the batch size to 16. For HatEval we used the model proposed by [24] with the same hyperparameters (epoch: 3, batch size: 16, learning rate: 2e-05). For the proposed MTL settings, in the case of HatEval we trained the model for two epochs, the learning rate was set to 4e-05 and the batch size was set to 32. For MEX-A3T, the model was trained for three epochs, the learning rate was set to 3e-05 and the batch size was set to 16. In order to optimize both approaches $\text{STL}_{BETO}$ and MTL in both datasets we use Adam optimizer and the epsilon was set to 1e-8.

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

**TABLE 5.** $STL_{BETO}$ and MTL settings results on the Spanish HS datasets. Class 0: Non-HS or Non-Aggressiveness, class 1: HS or Aggressiveness.

| | | Class 0 | | | Class 1 | | | Macro-Avg | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Model** | **P (%)** | **R (%)** | **F1 (%)** | **P (%)** | **R (%)** | **F1 (%)** | **P (%)** | **R (%)** | **F1 (%)** |
| **HatEval** | $STL_{BETO}$ [24] | 86.16 | 74.15 | 79.70 | 69.28 | 83.03 | 75.53 | 77.72 | 78.59 | 77.62 |
| | $MTL_{sent}$ | 87.34 | 73.40 | 79.77 | 69.14 | 84.85 | 76.19 | 78.24 | 79.13 | 77.98 |
| | $MTL_{emo}$ | 87.53 | **74.68** | **80.60** | **70.18** | 84.85 | 76.82 | 78.85 | 79.76 | **78.71** |
| | $MTL_{sent+emo}$ | **88.92** | 72.55 | 79.91 | 69.03 | **87.12** | **77.03** | **78.97** | **79.84** | 78.47 |
| **MEX-A3T** | $STL_{BETO}$ | 91.97 | 91.15 | 91.56 | 78.59 | 80.33 | 79.45 | 85.28 | 85.74 | 85.51 |
| | $MTL_{sent}$ | 93.39 | 90.93 | **92.14** | 78.94 | 84.09 | **81.43** | **86.17** | 87.51 | **86.79** |
| | $MTL_{emo}$ | 92.36 | **91.33** | 91.84 | **79.14** | 81.33 | 80.22 | 85.75 | 86.33 | 86.03 |
| | $MTL_{sent+emo}$ | **93.69** | 90.21 | 91.92 | 77.83 | **84.97** | 81.25 | 85.76 | **87.59** | 86.58 |

### C. RESULTS ANALYSIS

In this section, we report the performance of our methodology along with the comparison with the latest state-of-the-art studies. In order to accomplish this, we have employed the usual metrics in NLP tasks, including Precision (P), Recall (R), F1-score (F1), and macro scores. The equations of the different metrics are described below:

$$P(c) = \frac{TP}{TP + FP} \tag{1}$$

$$R(c) = \frac{TP}{TP + FN} \tag{2}$$

where c is equal to the class, TP = True Positive, FP = False Positive and FN = False Negative.

$$F1 = \frac{2 * P * R}{P + R} \tag{3}$$

#### 1) SINGLE-TASK LEARNING VS. MULTI-TASK LEARNING

We compare the performance of our baseline ($STL_{BETO}$) with the proposed MTL configurations on HatEval and MEX-A3T datasets. The results are reported in Table 5 which shows the prediction performances of each model, and each dataset. For the baseline experiments, in the case of HatEval we use the results obtained by the authors in [24] who applied the Transformer-based model BETO on the HatEval dataset, and in the case of MEX-A3T, we report the results we have obtained with $STL_{BETO}$. The performance of the baseline experiments in both datasets is very promising and shows that the $STL_{BETO}$ model works very well when fine-tuned on a small Spanish dataset. Specifically, in the MEX-A3T task, $STL_{BETO}$ achieved high results with a macro F1-score of 85.51%, compared to the result obtained in the HatEval dataset, 77.62%. In both datasets, we can see that the most challenging class to identify correctly by $STL_{BETO}$ is class 1 (HS and Aggressiveness). This behavior has been observed during the participants' results in workshops related to the HS detection task.

Regarding the results obtained by the different settings of the MTL model proposed, it is worth noting that, for both HatEval and MEX-A3T tasks, all the MTL configurations ($MTL_{sent}$, $MTL_{emo}$, and $MTL_{sent+emo}$) succeeded in surpassing our baseline $STL_{BETO}$ in terms of macro-P, macro-R, and macro-F1. In particular, for HatEval the best configuration is $MTL_{emo}$ while for MEX-A3T the $MTL_{sent}$ model achieves the best macro-F1 score. It could be because MEX-A3T is a dataset of tweets written in Mexican, and the dataset used for the sentiment task (InterTASS) contains texts written in different variants of Spanish including Mexican, therefore, a deeper knowledge of this linguistic variant is obtained. Concerning $MTL_{sent+emo}$ model, it behaves in the same way in both datasets, as shown in Table 5. Observing the performance of the model for class 1, it is worth mentioning that our proposal $MTL_{sent+emo}$ outperforms the precision of $STL_{BETO}$, and achieves a significantly higher recall by increasing 4.09 points in the case of HatEval and 4.64 points in MEX-A3T. This observation is remarkable since the $MTL_{sent+emo}$ succeeds at enhancing particularly the most challenging class (1), by detecting the HS and Aggressiveness tweets that $STL_{BETO}$ was not able to identify.

#### 2) COMPARISON TO THE STATE-OF-THE-ART SYSTEMS

Table 6 summarizes the comparative results of previous studies for Spanish HS detection. In particular, we have selected the state-of-the-art systems which have evaluated both HatEval and MEX-A3T datasets.

Regarding HatEval, in Table 6 we show the top three teams' [32], [60], [61] that achieved the best results in SemEval-2019 Task 5 as well as other systems that outperformed the results of the competition. The best model in SemEval-2019 Task 5 was presented by [32] where authors obtained a macro-F1 score of 73.0% using a linear kernel SVM trained on a text representation composed of a bag of words, a bag of characters, and an embedding of tweets computed from fastText sentiment-oriented word vectors. The system proposed by [60] was based on a linear kernel SVM. The study focused on a combinatorial framework used to search for the best feature configuration among a combination of linguistic pattern features, a lexicon of aggressive words and, different types of n-grams (characters, words, POS tags, aggressive words, word breaks, function words, and punctuation symbols). They obtained a macro-F1 score of 73.0%. [61] achieved a macro-F1 score of 72.9%, presenting a pre-trained BERT model on Twitter data and using a corpus of tweets collected over the same period of time from the HatEval training dataset. Another study to consider in SemEval-2019 Task 5 was the system presented by [33] which incorporated

**TABLE 6.** Comparative results for the HS detection task in Spanish. Results on class 0 and 1 indicate the F1-score.

| Dataset | System | Class 0 | Class 1 | Macro-F1 |
|---|---|---|---|---|
| **HatEval** | SVM with sentiment features [33] | 75.3 | 69.8 | 72.5 |
| | BERT [61] | 73.0 | 72.7 | 72.9 |
| | SVM with features [60] | 76.1 | 69.9 | 73.0 |
| | SVM with fastText sentiment embedding [32] | 74.9 | 71.1 | 73.0 |
| | Ensemble voting classifier [17] | **80.0** | 68.8 | 74.2 |
| | multi-channel BERT [23] | - | - | 76.6 |
| | BETO [24] | 79.7 | 75.5 | 77.6 |
| | **MTL$_{sent+emo}$ (Proposed approach)** | 79.91 | **77.03** | **78.47** |
| **MEX-A3T** | EvoMSA7 [62] | 89.33 | 74.68 | 82.00 |
| | Ensemble BETO models and adversarial data augmentation [63] | **91.95** | 79.98 | 85.96 |
| | BETO [64] | 91.07 | 79.69 | 85.38 |
| | **MTL$_{sent+emo}$ (Proposed approach)** | 91.92 | **81.25** | **86.58** |

sentiment features. This system achieved a macro-F1 score of 72.5%, presenting a linear SVM model based on linguistic features, semantic similarity with a domain-oriented lexicon, sentiments, word embeddings, topic modeling and TF-IDF n-grams of words and characters. On the other hand, we found the following studies [17], [23], [24] on the HatEval dataset out of the competition. [17] proposed an ensemble of traditional machine learning classifiers (Naive Bayes and Logistic Regression) using the Term Frequency scheme (TF) with a combination of unigrams and bigrams for the feature representation. The system is particularly successful in identifying the Non-HS class as it achieved a macro-F1 score of 80.0%. With regard to Transformer-based systems, [23], [24] used BERT model and succeeded in surpassing previous results. The first study used the multilingual BERT model trained for several languages and the second one which uses BETO obtained state-of-the-art results. After analyzing these studies, as can be seen in Table 6 it is worth mentioning that our proposed model MTL$_{sent+emo}$ significantly outperforms the best result of the SemEval-2019 Task 5 by 5.47 points in terms of the macro-F1 score and also slightly outperforms the results of the latest study on the state-of-the-art [24] using this dataset. Moreover, it should be noted that our model successfully detected the HS class obtaining an F1 score of 77.03%. Related to the MEX-A3T dataset, [62] proposed a text classifier that combines two models called B4MSA and EvoDAG. B4MSA is a minimalistic classifier independent from domain and language and EvoDAG is a classifier based on Genetic Programming. [63], [64] used the BETO model trained specifically for Spanish and similar to the studies of HatEval, it improves also the previous system. Our proposed model MTL$_{sent+emo}$ achieved the best results by obtaining a macro-F1 score of 86.58%. Similar to the previous dataset, our model successfully identified the HS class obtaining an F1 score of 81.25%.

#### 3) KNOWLEDGE TRANSFER FROM SENTIMENT ANALYSIS

Our results show that SA improves HS detection on both HatEval and MEX-A3T datasets. Table 7 introduces examples of improvements in HatEval achieved by the MTL$_{sent+emo}$ system, over the STL$_{BETO}$ model. As the SA tasks lead the MTL model to learn how to predict the polarity and emotion labels for the instances, the representations computed by the encoder embed the affective knowledge. This allows the & MTL$_{sent+emo}$ model to classify HS more accurately by leveraging the affective nature of the instance. Looking at the examples in Table 7 it is important to point out that people often use some expressions that contain offensive words, however, the expression is not necessarily offensive since it conveys a positive polarity and emotion. For instance, tweet number 4 contains the expression *puta ama* (fucking boss) which is positive although the presence of the offensive word *puta* (whore) is used. In this case, the STL$_{BETO}$ model mislabeled the tweet as HS, whereas the MTL classified it as Non-HS since the polarity and emotion predicted were *positive* and *joy*, respectively. Similarly, tweets 1, 2, and 3 with *positive* polarity and *joy* emotion were correctly classified by our proposed model as Non-HS but not by the STL$_{BETO}$ model which prediction was HS. These examples, as well as the expressions, also contain offensive words associated with misogyny and xenophobia, but the emotion they evoke is *positive*. The rest of the examples with *negative* polarity and conveying *anger* and *sad* emotions were misclassified by STL$_{BETO}$ but not by the MTL$_{sent+emo}$ model. For instance, in tweet 5 with the negative words *redadas* (raids) and *hieren* (hurt), tweet 6 with *parasitos subordinados* (subordinate parasites) and tweet 8 with the expression *el que tenga huevos* (whoever has balls), it is again shown that MTL benefits from the affective knowledge learned from SA tasks.

## VI. ERROR ANALYSIS

In order to gain deeper insight about the proposed MTL model performance, we conducted an error analysis from both quantitative and qualitative levels. We mainly analyzed the instances in the test set that were wrongly labeled by the STL$_{BETO}$ and the MTL$_{sent+emo}$ models in HatEval dataset. Since the gold labels of the MEX-A3T test set are not publicly available, we have not performed the analysis for this dataset.

Based on quantitative analysis, we analyzed the confusion matrices of STL$_{BETO}$ and MTL$_{sent+emo}$ models and compare them in Table 8. The MTL system in HatEval mislabeled only 85 HS instances to Non-HS compared to 104 instances

**TABLE 7.** $STL_{BETO}$ vs. $MTL_{sent+emo}$ samples from HatEval dataset, showing improved MTL performance. English translation of Spanish tweets is provided.

| | Tweet | Gold label | $STL_{BETO}$ | $MTL_{sent+emo}$ | | |
|---|---|---|---|---|---|---|
| | | | | HS | polarity | emotion |
| 1 | *Enseñando a ser puta yo las amo* (Teaching to be a whore I love them) | 0 | 1 | 0 | P | joy |
| 2 | *puta la madre, tu eres una mujer muy guapa @user* (fucking hell, you are a very beautiful woman @user) | 0 | 1 | 0 | P | joy |
| 3 | *@user Que ganen la sudaca, amén* (@user May they win the sudaca, amen) | 0 | 1 | 0 | P | joy |
| 4 | *ANA, tu eres una GUERRERA, no te vamos a dejar abandonar este barco PUTA AMA, te quiero mi niña* (ANA, you are a WARRIOR, we will not let you leave this ship FUCKING BOSS, I love you my girl) | 0 | 1 | 0 | P | joy |
| 5 | *Redadas y devoluciones en caliente y frío! Estamos hasta los CO-JO-NES. Más de 100 inmigrantes hieren a siete guardias civiles con ácido y cal viva para saltar la valla de Ceuta* (Hot and cold raids and returns! We are fucking done. More than 100 immigrants hurt seven civil guards with acid and quicklime to jump over the fence in Ceuta) | 1 | 0 | 1 | N | anger |
| 6 | *La versión sudaka del Isis. Me parece que cambiemos tendría que abrir un poco más los ojos y tomar más en serio las acciones estos parásitos subordinados* (The sudaka version of Isis. It seems to me that we should change and open our eyes a little more and take the actions of these subordinate parasites more seriously) | 1 | 0 | 1 | N | anger |
| 7 | *Te vengo a enseñar y a educar que vos puta a mi no me vas a ganar, también rima?* (I come to teach you and to educate you that you whore to me you are not going to win, also it rhymes?) | 1 | 0 | 1 | N | anger |
| 8 | *400 voltios y que quiten las concertinas, y el que tenga huevos que salte* (400 volts and that they remove the razor wire fences, and whoever has balls should jump) | 1 | 0 | 1 | N | anger |
| 9 | *Por desgracia, no queda otra, aportan poco y nos cuestan mucho, incluido nuestra seguridad* (Unfortunately, there is no other way, they contribute little and cost us a lot, including our security) | 1 | 0 | 1 | N | sadness |

**TABLE 8.** Confusion matrix of HatEval.

| Class | $STL_{BETO}$ | | $MTL_{sent+emo}$ | |
|---|---|---|---|---|
| | Non-HS | HS | Non-HS | HS |
| Non-HS | 685 | 255 | 682 | 258 |
| HS | 104 | 556 | 85 | 575 |

misclassified with the $STL_{BETO}$ model. As we highlight in the analysis of the results, it shows that the MTL model's ability to distinguish the hateful text by reducing the number of false positives in HS class. On the other hand, the $MTL_{sent+emo}$ system mislabeled 258 Non-HS instances to HS compared to 255 instances misclassified with the $STL_{BETO}$ model. Therefore, we consider that the knowledge provided by the external datasets (InterTASS and EmoEvent), although very slightly detrimental to the Non-HS class, not just improves the prediction in general, but the performance on the HS class is particularly enhanced. This is important in a detection task, more than in a classification task.

Concerning a qualitative analysis, we analyzed some tweets misclassified by the $MTL_{sent+emo}$ system to identify the possible challenges that the system faces with the Spanish and the HS detection. In addition, we also looked at the predictions related to emotion and polarity classification in order to analyze how they contribute to the detection of HS. We select some mislabeled tweets predicted by the $MTL_{sent+emo}$ system: two false positive and two false-negative tweets which can be seen in Table 9. In the first false positive, there is the presence of the xenophobic word *negrata* (nigga) but at the same time, the user includes the positive word *top* (top). Consequently, due to the lack of context and the short length of the tweet, the system is labeled as HS. Moreover, it should be noted that this tweet is classified as *neutral* polarity and *others* emotion which shows that is not clear if the tweet is positive or negative and therefore it demonstrates the difficulty of detecting the HS by the $MTL_{sent+emo}$ model. In the second false positive, the user employs the misogynist word *puta* (whore), but without offending anyone because she is referring to herself and is annoyed with the judgments that other people make about women. It is possible that the $MTL_{sent+emo}$ model mislabeled it because there is a misogynist word and also the emotions and sentiments expressed in the tweet are negative, two factors that, together, are usually associated with the presence of HS. In the case of the false-negative instances,

**TABLE 9.** Tweets mislabeled by the $\text{MTL}_{sent+emo}$ model. Two false positives and two false negatives, respectively. English translation of Spanish tweets is provided.

| Tweet | Gold label | $\text{MTL}_{sent+emo}$ | | |
|---|---|---|---|---|
| | | HS | polarity | emotion |
| *Para mi con el negrata la delantera es top*<br>(For me with the nigga in the front is top) | 0 | 1 | NEU | others |
| *O sea, ¿se supone que no puedo tener opiniones? ¿debo estar siempre callada porque soy una mujer? me dices puta porque expreso lo que pienso, supongo que prefieres que me limite a sentarme y sonreír. Cuando una mujer se harta y contraataca de repente el macho no sabe como actuar*<br>(I mean, am I not supposed to have opinions? should I always be quiet because I am a woman? you call me a whore because I express what I think, I guess you prefer me to just sit and smile. When a woman gets fed up suddenly man doesn't know how to act) | 0 | 1 | N | anger |
| *Devolución exprés ahora y siempre. Y más concertinas y lo que haga falta para que no entren*<br>(Express return now and always. And more razor wire fences and whatever else is needed to keep them out) | 1 | 0 | N | others |
| *Quitar las concertinas y poner ametralladoras*<br>(Remove the razor wire fences and set up the machine guns) | 1 | 0 | N | others |

the system has may predict it wrongly because although the user is expressing xenophobia, there is no explicit mention of immigrants in the tweet, which is a challenge for NLP systems due to the implicit information. In addition, the polarity is classified correctly but the emotion predicted is other which shows the difficulty of detecting HS in this tweet.

Finally, it is worth noting that we have detected some mislabeled tweets in the datasets, which complicates the learning process of the models.

## VII. CONCLUSION

The spread of HS has increased in recent years becoming a major challenge for online platforms and national governments, which need to rely on automated systems to identify and remove this type of content. Spanish is becoming increasingly globalized since it is the official language of 21 countries and is spoken by over 420,000,000 native speakers. Therefore, our paper focuses on HS detection in Spanish corpora, proposing a MTL model to take advantage of related tasks including polarity and emotion classification. Experiments conducted on two benchmark corpora show the efficacy of our proposed approach in achieving convincing performance over an $\text{STL}_{BETO}$ model and state-of-the-art results. The performance achieved by our proposed model and a detailed knowledge transfer analysis from SA shows that polarity and emotion classification tasks help the MTL model to classify HS more accurately by leveraging on the affective knowledge. The correlated effects of affective knowledge and HS provide the opportunity to investigate new ways of improving NLP systems in other domains as well, where polarity and emotion could play an important role. The limitations found in the model lie in the fact that multitasking leverages other corpora for classification, so the computational cost is higher. In addition, the quality of corpora is important in a multitask learning environment, so finding such resources is not always possible, especially in low-resource languages. Finally, as future work, we plan to develop a complex model that incorporates other related tasks, such as irony or sarcasm detection, that could be beneficial for HS detection.

## REFERENCES


[1] B. Pang, "Foundations and trends in information retrieval," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2008.

[2] P. Ekman, "An argument for basic emotions," *Cognit. Emotion*, vol. 6, nos. 3–4, pp. 169–200, 1992, doi: 10.1080/02699939208411068.

[3] J. B. Nezlek and P. Kuppens, "Regulating positive and negative emotions in daily life," *J. Personality*, vol. 76, no. 3, pp. 561–580, Jun. 2008.

[4] G. T. Patrick, "The psychology of profanity," *Psychol. Rev.*, vol. 8, no. 2, p. 113, 1901.

[5] W. Alorainy, P. Burnap, H. Liu, A. Javed, and M. L. Williams, "Suspended accounts: A source of tweets with disgust and anger emotions for augmenting hate speech data sample," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, vol. 2, Jul. 2018, pp. 581–586.

[6] A. Rodriguez, C. Argueta, and Y.-L. Chen, "Automatic detection of hate speech on Facebook using sentiment and emotion analysis," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Feb. 2019, pp. 169–174.

[7] A. Silva and N. Roman, "Hate speech detection in Portuguese with Naïve Bayes, SVM, MLP and logistic regression," in *Proc. Anais do XVII Encontro Nacional de Inteligência Artif. e Computacional*, 2020, pp. 1–12.

[8] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and M. Stranisci, "An Italian Twitter corpus of hate speech against immigrants," in *Proc. 11th 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 1–8.

[9] Z. Pitenis, M. Zampieri, and T. Ranasinghe, "Offensive language identification in Greek," 2020, *arXiv:2003.07459*. [Online]. Available: https://arxiv.org/abs/2003.07459

[10] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata, "A multilingual evaluation for online hate speech detection," *ACM Trans. Internet Technol.*, vol. 20, no. 2, pp. 1–22, 2020.

[11] E. W. Pamungkas, V. Basile, and V. Patti, "A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection," *Inf. Process. Manage.*, vol. 58, no. 4, Jul. 2021, Art. no. 102544.

[12] N. Vashistha and A. Zubiaga, "Online multilingual hate speech detection: Experimenting with Hindi and English social media," *Information*, vol. 12, no. 1, p. 5, Dec. 2020.

[13] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," in *Proc. 10th ACM Conf. Web Sci. (WebSci)*, 2019, pp. 105–114.

[14] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion)*, 2017, pp. 759–760.

[15] P. Kapil and A. Ekbal, "Leveraging multi-domain, heterogeneous data using deep multitask learning for hate speech detection," 2021, *arXiv:2103.12412*. [Online]. Available: https://arxiv.org/abs/2103.12412

[16] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–30, 2018.
[17] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "Detecting misogyny and xenophobia in Spanish tweets using language technologies," *ACM Trans. Internet Technol.*, vol. 20, no. 2, pp. 1–19, 2020.
[18] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 11, no. 1, 2017, pp. 512–515.
[19] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks," *Appl. Intell.*, vol. 48, no. 12, pp. 4730–4742, 2018.
[20] G. H. Paetzold, M. Zampieri, and S. Malmasi, "UTFPR at SemEval-2019 task 5: Hate speech identification with recurrent neural networks," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 519–523, doi: 10.18653/v1/S19-2093.
[21] B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in *Proc. 1st Workshop Abusive Lang. Online*, 2017, pp. 85–90.
[22] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in *Proc. China Nat. Conf. Chin. Comput. Linguistics*, 2019, pp. 194–206.
[23] H. Sohn and H. Lee, "MC-BERT4HATE: Hate speech detection using multi-channel BERT for different languages and translations," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2019, pp. 551–559.
[24] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "Comparing pre-trained language models for Spanish hate speech detection," *Expert Syst. Appl.*, vol. 166, Mar. 2020, Art. no. 114120.
[25] J. Canete, G. Chaperon, R. Fuentes, and J. Pérez, "Spanish pre-trained bert model and evaluation data," in *Proc. PML4DC ICLR*, 2020, pp. 1–0.
[26] R. Martins, J. Almeida, P. Henriques, and P. Novais, "Increasing authorship identification through emotional analysis," in *Proc. World Conf. Inf. Syst. Technol.* Springer, 2018, pp. 763–772.
[27] F. M. Plaza-del-Arco, M. D. Molina-González, M. Martin, and L. A. Ureña-López, "SINAI at SemEval-2019 task 6: Incorporating lexicon knowledge into SVM learning to identify and categorize offensive language in social media," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 735–738.
[28] R. Martins, M. Gomes, J. J. Almeida, P. Novais, and P. Henriques, "Hate speech classification in social media using emotional analysis," in *Proc. 7th Brazilian Conf. Intell. Syst. (BRACIS)*, Oct. 2018, pp. 61–66.
[29] A. Brown, "What is hate speech? Part 1: The myth of hate," *Law Philosophy*, vol. 36, no. 4, pp. 419–468, Aug. 2017.
[30] N. S. Samghabadi, A. Hatami, M. Shafaei, S. Kar, and T. Solorio, "Attending the emotions to detect online abusive language," 2019, *arXiv:1909.03100*. [Online]. Available: https://arxiv.org/abs/1909.03100
[31] A. Elmadany, C. Zhang, M. Abdul-Mageed, and A. Hashemi, "Leveraging affective bidirectional transformers for offensive language detection," in *Proc. 4th Workshop Open-Source Arabic Corpora Process. Tools, Shared Task Offensive Lang. Detection*, Marseille, France, May 2020, pp. 102–108.
[32] J. M. Pérez and F. M. Luque, "Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification," in *Proc. 13th Int. Workshop Semantic Eval.*, Minneapolis, MN, USA, 2019, pp. 64–69, doi: 10.18653/v1/S19-2008.
[33] D. Benito, O. Araque, and C. A. Iglesias, "GSI-UPM at SemEval-2019 task 5: Semantic similarity and word embeddings for multilingual detection of hate speech against immigrants and women on Twitter," in *Proc. 13th Int. Workshop Semantic Eval.*, Minneapolis, MN, USA, 2019, pp. 396–403, doi: 10.18653/v1/S19-2070.
[34] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
[35] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 31, 2021, 10.1109/TKDE.2021.3070203.
[36] A. Tebbifakhr, M. Negri, and M. Turchi, "Automatic translation for multiple NLP tasks: A multi-task approach to machine-oriented NMT adaptation," in *Proc. 22nd Annu. Conf. Eur. Assoc. Mach. Transl.*, Lisboa, Portugal, Nov. 2020, pp. 235–244.
[37] S. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, and S. Kurohashi, "All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework," *IEEE Trans. Affect. Comput.*, early access, Jul. 5, 2019, doi: 10.1109/TAFFC.2019.2926724.
[38] G. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen, "A neural network multi-task learning approach to biomedical named entity recognition," *BMC Bioinf.*, vol. 18, no. 1, p. 368, Dec. 2017.
[39] P. Kapil and A. Ekbal, "A deep neural network based multi-task learning approach to hate speech detection," *Knowl.-Based Syst.*, vol. 210, Dec. 2020, Art. no. 106458.
[40] I. A. Farha and W. Magdy, "Multitask learning for Arabic offensive language and hate-speech detection," in *Proc. 4th Workshop Open-Source Arabic Corpora Process. Tools, Shared Task Offensive Lang. Detection*, 2020, pp. 86–90.
[41] S. Rajamanickam, P. Mishra, H. Yannakoudakis, and E. Shutova, "Joint modelling of emotion and abusive language detection," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4270–4279, doi: 10.18653/v1/2020.acl-main.394.
[42] E. Martínez-Cámara, M. C. Díaz-Galiano, M. A. García-Cumbreras, M. García-Vega, and J. Villena-Román, "Overview of TASS 2017," in *Proc. TASS*, 2017, pp. 13–21.
[43] E. Martínez-Cámara, Y. Almeida-Cruz, M. C. D. Galiano, S. Estévez-Velarde, M. Á. G. Cumbreras, M. G. Vega, and Y. Gutiérrez, "Overview of TASS 2018: Opinions, health and emotions," in *Proc. TASS 2018: Workshop Semantic Anal. at SEPLN, TASS@SEPLN 2018*, vol. 2172, 2018, pp. 13–27.
[44] M. C. Díaz-Galiano, M. G. Vega, and E. Casasola, "Overview of TASS 2019: One more further for the global Spanish sentiment analysis corpus," in *Proc. IberLEF@ SEPLN*, 2019, pp. 550–560.
[45] F. M. Plaza-del-Arco, C. Strapparava, L. A. Ureña-López, and M. T. Martín-Valdivia, "EmoEvent: A multilingual emotion corpus based on different events," in *Proc. 12th Lang. Resour. Eval. Conf.*, Marseille, France, May 2020, pp. 1492–1498.
[46] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proc. 13th Int. Workshop Semantic Eval.*, Minneapolis, MN, USA, 2019, pp. 54–63, doi: 10.18653/v1/S19-2007.
[47] E. Fersini, P. Rosso, and M. Anzovino, "Overview of the task on automatic misogyny identification at IberEval 2018," in *Proc. 3rd Workshop Eval. Hum. Lang. Technol. Iberian Lang. (IberEval)*, 2150, 2018, pp. 214–228.
[48] E. Fersini, D. Nozza, and P. Rosso, "Overview of the evalita 2018 task on automatic misogyny identification (AMI)," *Proc. EVALITA Eval. NLP Speech Tools Italian*, vol. 12, 2018, p. 59.
[49] M. Ezra-Aragón and H. J. Jarquín-Vásquez, "Overview of MEX-A3T at IberLEF 2020: Fake news and aggressiveness analysis in Mexican Spanish," in *Proc. Iberian Lang. Eval. Forum (IberLEF) 36th Conf. Spanish Soc. Natural Lang. Process. (SEPLN)*, Málaga, Spain, vol. 2664, Sep. 2020, pp. 222–235.
[50] M. A. Á. Carmona, E. Guzmán-Falcón, M. M.-Y. Gómez, H. J. Escalante, L. Villasenor-Pineda, V. Reyes-Meza, and A. Rico-Sulayes, "Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets," in *Proc. 3rd Workshop IberEval, (SEPLN)*, vol. 2150, Sep. 2018, pp. 74–96.
[51] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[52] S. Ruder, "Neural transfer learning for natural language processing," M.S. thesis, NUI Galway, 2019. [Online]. Available: http://hdl.handle.net/10379/15463
[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[55] M. J. Díaz-Torres, P. A. Morán-Méndez, L. Villaseñor-Pineda, M. Montes, J. Aguilera, and L. Meneses-Lerín, "Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset," in *Proc. 2nd Workshop Trolling, Aggression Cyberbullying*, 2020, pp. 132–136.
[56] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*. [Online]. Available: http://arxiv.org/abs/1609.08144
[57] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting (ACL)*, vol. 1, Berlin, Germany, Aug. 2016, pp. 1715–1725, doi: 10.18653/v1/P16-1162.

[58] C. Baziotis, N. Pelekis, and C. Doulkeridis, "DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis," in *Proc. 11th Int. Workshop Semantic Eval. (SemEval)*, 2017, pp. 747–754.
[59] A. Paszke, S. Gross, F. Massa, and A. Lerer, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[60] L. E. Argota Vega, J. C. Reyes-Magaña, H. Gómez-Adorno, and G. Bel-Enguix, "MineriaUNAM at SemEval-2019 task 5: Detecting hate speech in Twitter using multiple features in a combinatorial framework," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 447–452.
[61] A. Gertner, J. Henderson, E. Merkhofer, A. Marsh, B. Wellner, and G. Zarrella, "MITRE at SemEval-2019 task 5: Transfer learning for multilingual hate speech detection," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 453–459.
[62] M. Graff *et al.*, "INGEOTEC at MEX-A3T: Author profiling and aggressiveness analysis in Twitter using TC and EvoMSA," in *Proc. IberEval@SEPLN*, 2018, pp. 128–133.
[63] M. Guzman-Silverio and Á. Balderas-Paredes, "Transformers and data augmentation for aggressiveness detection in Mexican Spanish," in *IberEval@SEPLN*, Malaga, Spain, 2020, pp. 293–302.
[64] M.-A. Tanase, G. E. Zaharia, D. C. Cercel, and M. Dascalu, "Detecting aggressiveness in Mexican Spanish social media content by fine-tuning transformer-based models," in *Proc. Notebook Papers 2nd SEPLN Workshop Iberian Lang. Eval. Forum (IberLEF)*, Malaga, Spain, 2020, pp. 236–245.

**FLOR MIRIAM PLAZA-DEL-ARCO** received the B.S. and M.S. degrees in computer science from the University of Jaén, Spain, in 2016 and 2018, respectively, where she is currently pursuing the Ph.D. degree in computer science.

Her research interests include computational linguistics and natural language processing, especially deep learning, text categorization, sentiment and emotion analysis, offensive language detection, and low-resource generation. She is a member of the Spanish Society for Natural Language Processing (SEPLN). She has been a member of the Organizing Committee of the 36th Annual SEPLN Conference and the TASS workshop at IberLeF 2021.

**M. DOLORES MOLINA-GONZÁLEZ** received the B.S. degree in telecommunication engineering from the University of Valencia, Spain, in 2005, and the Ph.D. degree in computer sciences engineering from the University of Jaén, Spain, in 2014.

She is currently an Associate Professor with the Department of Engineering of Telecommunication, University of Jaén. She is the author or coauthor of more than 25 scientific publications. Her current research interests include natural language processing, machine learning, sentiment and emotion analysis, and offensive language detection. She is a technical reviewer for several journals. She is also a member of the Spanish Society for Natural Language Processing (SEPLN).

**L. ALFONSO UREÑA-LÓPEZ** received the B.S. and Ph.D. degrees in computer science from the University of Granada, in 1991 and 2000, respectively.

He is currently a Full Professor with the Department of Computer Science, University of Jaén, Spain, where he is the Director of the Research Institute in Information and Communication Technologies. He is the author of more than 200 publications on various topics of natural language processing (NLP). He is also the President of the Spanish Society for Natural Language Processing (SEPLN). He is the programme chair and a keynote speaker of several major international conferences. He is the Editor-in-Chief of *Procesamiento de Lenguaje Natural* journal.

**MARÍA TERESA MARTÍN-VALDIVIA** received the B.S. degree in computer science from the University of Granada, in 1992, and the Ph.D. degree in computer science engineering from Málaga University, in 2004.

She is currently a Full Professor with the Department of Computer Science, University of Jaén, Spain. She is the author or coauthor of more than 100 scientific publications. Her current research interests include natural language processing, machine learning algorithms, sentiment analysis, and text mining techniques. In addition, she is an Editor of the *Procesamiento de Lenguaje Natural* journal. She is a technical reviewer of several journals and a member of the program committee of several major conferences.