ISSN 2007-9737

Do Linguistic Features Help Deep Learning? The Case of Aggressiveness in Mexican Tweets

Simona Frenda[1,2], Somnath Banerjee[3], Paolo Rosso[2], Viviana Patti[1]

[1] Università degli Studi di Torino,
Dipartimento di Informatica,
Italy

[2] Universitat Politècnica de València,
PRHLT Research Center, València,
Spain

[3] Jadavpur University,
Salt Lake,
India

sfrenda@unito.it, s.banerjee@ndsu.edu, prosso@dsic.upv.es, patti@di.unito.it

Abstract. In the last years, the control of online user generated content is becoming a priority, because of the increase of online aggressiveness and hate speech legal cases. Considering the complexity and the importance of this issue, this paper presents an approach that combines the deep learning framework with linguistic features for the recognition of aggressiveness in Mexican tweets. This approach has been evaluated relying on a collection of tweets released by the organizers of the shared task about aggressiveness detection in the context of the Ibereval 2018 evaluation campaign. The use of a benchmark corpus allows to compare the results with those obtained by Ibereval 2018 participant systems. However, looking at the achieved results, linguistic features seem not to help the deep learning classification for this task.

Keywords. Deep learning, aggressiveness automatic detection, Mexican Spanish language, twitter, linguistic analysis.

1 Introduction

The opinions expressed online by users are usually uncontrolled, and this lack of control facilitates and supports negative online behaviors such as cyberbullying, racism or sexism. Indeed, inciting the hate towards an individual or a group of people for characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion is the principal aim of any kind of hate speech on social media.

On these platforms, in fact, the anonymity and the interactivity facilitate the diffusion and the permanence of these offensive comments. Actually, this uncontrolled flow of thoughts that gives rise to negative online behaviors, can morally harm or incite physical violence. Indeed, it is not uncommon to find user-generate online contents impregnated of hate, especially towards women, which sometime can also be translated into actions of violence.

For instance, the social study by Fulper [15] demonstrates the existence of a correlation between the number of rapes and the amount of misogynistic tweets per state on USA in 2012, suggesting the fact that social media could be used as a social sensor of sexual violence. Moreover, the persistence and diffusion of misogynistic or offensive content hurt and distress above all psychologically the victims, causing sometime their suicide, like the case of the teenager Amanda

Todd in 2012[1]. Hate speech online is a real problem of modern society and, for this reason, in the last few years, governments, social media platforms, Internet companies and communities of citizens have been spending a growing amount of efforts to monitor and contrast such forms of online aggressive behaviors and attitudes.

An example of governmental dedication about this subject is the campaign *No Hate Speech Movement* of the Council of Europe[2] for human rights online. On the academic side, the research interest about this issue is increasing and the approach is naturally interdisciplinary, involving: computer science, psychology, sociology, law, gender studies, communication, media studies and natural language processing (NLP).

Especially in the NLP field, the attention is supported by national and international workshops, like ALW1[3] at international venues, that allow to share information and results exploring the different kinds of online hate speech, such as: racism, misogyny and cyberbulling. In addition, also campaigns of evaluation foster the research in this field in various languages, such as EVALITA 2018[4], SemEval 2019[5] or the competition proposed in the framework of IberEval 2018[6] by the organizers of MEX-A3T[7] [1] on the aggressiveness analysis in Twitter in Mexican language.

This shared task proposes to detect the aggressiveness on Mexican Spanish tweets providing tweets containing offensive messages that disparage or humiliate individuals or groups of persons. Especially online comments, like tweets, are difficult to manage considering the variety of informal language elements used by users, such as: slang words, typical expressions of the speech, abbreviations, alphanumeric words, acronyms, hashtags, emoticons or emoji.

This work proposes an innovative approach that incorporates linguistic features into a deep learning architecture, specifically Convolutional Neural Network (CNN). At this scope, a set of features specific for the aggressiveness detection in Mexican texts has been created.

Moreover, one of the principal aims of this paper is to analyze the contribution of linguistic features into deep learning architecture for aggressiveness detection, making a comparison with a system based on a simple deep learning approach.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 briefly delineates the dataset. Section 4 presents the proposed approach, followed by the description of the experiments and the results in Section 5. Finally, Sections 6 and 7 discuss the results and present some conclusions proposing the future works.

2 Related Work

The literature about hate speech detection on the web covers phenomena having different specific focuses, such as cyberbullying, misogyny, sexual predators, abusiveness, nastiness and aggressiveness. In order to address automatically the detection of these online misbehaviors, commercial and simple methods rely currently on the use of blacklists, essentially composed of slurs and swear words.

However, filtering the messages in this way does not provide a sufficient remedy because it falls short when user-generated content is more subtle. For instance, one of the first systems of recognition of flames is Smokey (implemented by Spertus [27]) that exploits specific rules designing syntactic and semantic structures in addition to lexicons of insults and profanities.

Today, the research challenges in this field are oriented at investigating deeply all dimensions of language and also the communication on the web, to envision deeper and more sophisticated solutions experimenting classic and deep learning methods.

[1] Here, some information about the case: https://www.theguardian.com/commentisfree/2012/oct/26/amanda-todd-suicide-social-media-sexualisation

[2] https://www.coe.int/en/web/no-hate-campaign

[3] The 1st Workshop on Abusive Language Online held at ACL 2017.

[4] https://amievalita2018.wordpress.com/; http://www.di.unito.it/~tutreeb/haspeede-evalita18/index.html#

[5] https://competitions.codalab.org/competitions/19935; https://competitions.codalab.org/competitions/20011

[6] https://sites.google.com/view/ibereval-2018

[7] https://mexa3t.wixsite.com/home/aggressive-detection-track

The former ones rely on manual feature engineering using rule-based systems [19] or machine learning approaches [5, 6, 12, 21, 18, 24], exploiting: simple surface characteristics [8, 20], linguistic features that take into account POS-tags or dependencies relationships [29, 6, 31, 19], semantic knowledge using word embedding techniques [21, 25] and conceptual and polarity information [18].

Recently, the authors of [12] experiment profile-based representation in the perspective to catch the sexual predators into chats online and to recognize aggressive messages. In [11] the authors use an ontology-based approach in order to predict the anti-LGBT hate speech. Other scholars take advantage of the connection between sentiment analysis and hate speech, benefiting from sentiment lexicons [28] or using a multi-step approach that combines sentiment or subjectivity classifiers with systems of hate speech detection [17]. This relation is due to the fact that hate speech expressions mostly exhibit a negative polarity, although the polarity intensity depends above all on cultural factors.

Some authors focus also on the extraction of meta-information from social platforms about users (like gender) and on their social activity (like history or geolocalization of posts) as predictive features [9]; while others concentrate on roles of persons involved in hate speech episodes [29].

More recently, some works investigate the potentiality of a deep learning approach [30] in aggressiveness detection. The authors of [20] compare the performance between the combination of Support Vector Machine with Naive Bayes and Recurrent Neural Network system, demonstrating the superiority of a character-based approach on the classical approach token-based. For the recognition of different classes of hate speech (sexism, racism, hate and not hate), the authors of [16] show the best performance of the CNN, compared to Logistic Regression with character n-grams. As well as, for Italian language, the authors of [10] display the optimal performance of bidirectional Long Short Term Memory architecture with word embedding and polarity compared to SVM with stylistic traits, linguistic features and lexicons.

To our knowledge, the MEX-A3T shared task is the first task about aggressive detection in a variant of the Spanish language. It can be also seen as one of the first efforts devoted to extend the investigation about hate speech in other languages, differently from English. In this framework, our contribution is an experimental work where, differently from the previous researches, feature engineering and CNN are combined. The implemented architecture, indeed, aims to catch: the characteristics of the aggressive language such as emotions that arouse this behavior; insults or profanities; the typical traits of informal texts (like tweets) such as emoticons, abbreviations or the use of capital letter; and the typical expressions of Mexican language.

3 Dataset

The proposed approach in this paper is tested using the benchmark corpus released by the organizers [1] of the MEX-A3T shared task about aggressiveness detection organized at IberEval 2018 in order to compare the performances with the participants in the task. In this shared task, the organizers propose to participants a classification task with the aim to distinguish aggressive tweets from the non-aggressive ones in Mexican language.

The collection is composed of tweets written by users with different backgrounds. The texts are very short and often contain orthographic or grammar mistakes. Dealing with such spontaneous texts allows researchers to face the multitude of speaking devices used in the digital writing, and to cope with the complexity of the figurative devices used by the users to express their opinions.

The dataset provided by the organizers has been collected between August and November 2017, using controversial hashtags (about politics, homophobia, and discrimination in general) and a fixed vocabulary about Mexican vulgarities and insults. This corpus has been collected considering the geolocalization in Mexico city and has been annotated manually by two annotators relying on an annotation scheme including the labels *aggressive* and *non-aggressive*.

Table 1. MEX-A3T dataset

Training set		Test set
Aggressive	*Non-Aggressive*	
2,727	4,973	3,156

A message is considered offensive if it targets individuals or groups with the purpose to insult them. Therefore, it could contain jokes or humorous expressions, derogatory adjectives, profanities and also nicknames that underline, for instance, physical defects of the target. The entire data set has been split into training (70%) and test set (30%) as showed in Table 1.

4 Proposed Approach

This section describes the proposed approach for the detection of aggressive tweets in Spanish Mexican language.

One of the main difficulties related to treatment of the data in Mexican language is the lack of adequate linguistic resources. In fact, Mexican language is a variation of Spanish and the vocabulary often is not the same revealing other meanings[8]. Moreover, these linguistic differences are more evident in the informal register, daily used also in social platforms. Therefore, considering these complexities, the proposed approach incorporates linguistic features into a deep learning architecture, attempting to combine the features engineering with the deeper analysis of the deep learning technique.

Usually, a system based on deep learning technique derives the features from the data without feature engineering. In this work, the linguistic features are incorporated into deep learning architecture in order to guide and help the system to choose the correct class for each tweet. In particular, CNN is employed, and to extract the designed features, the dataset was preprocessed deleting symbols and urls, that can hinder the process of extraction of features, and pos-tagged the texts using FreeLing [7]. The next subsections describe the deep learning architecture and the used linguistic features.

[8]About that, the *Academia Mexicana de la Lengua* proposes a dictionary of Mexican language: `https://www.academia.org.mx/obras/obras-de-consulta-en-linea/diccionario-de-mexicanismos`

4.1 Deep Learning Architecture

This section describes the deep learning (DL) framework, inspired by the deep architecture proposed by the authors of [3]. The approach successfully enhanced the classification accuracy for the code-mixed question classification task [4] with a few training data. In order to understand the impact of the feature engineering in the DL framework, two models have been implemented, both based on CNN: i) feature engineering combined with deep learning framework (DL+FE), and ii) simple DL framework (DL). Figure 1 shows the schema of the proposed architecture that is described below.

Embedding layer: For the embedding, any pre-trained word embedding has not been used. Instead, a vocabulary table is prepared by compiling the training data. Therefore, the embedding layer acts as a lookup table which maps vocabulary word indices into low-dimensional vector representations. The length of all the aggressive tweets is not same. Therefore, the zero-padding (i.e., the missing part replaced by zeros) has been employed to maintain the input vector to a fixed size L.

Features: As mentioned earlier, one of the models combines the feature engineering with deep learning framework. Therefore, the said model employs the described features in Subsection 4.2. The other model does not combine the features with deep learning framework.

Convolutional layer: Let $W_j \in \mathbb{R}^p$ be the p-dimensional vector corresponding to the j-th word in the tweet. Here, a tweet is represented as $W_{1:m} = W_1 \oplus W_2 \oplus ... \oplus W_m$, where, $W_1, W_2, \ldots, W_m$ are the words in the tweet and $\oplus$ is the concatenation operator.

Also, let $F_{1:n} = F_1 \oplus F_2 \oplus ... \oplus F_n$ be the feature set for the tweet $W_{1:m}$. The resulting vector is $r1 : n + m = F_{1:n} \oplus W_{1:m}$ after combining the feature set $F_{1:n}$ with the vector representation of

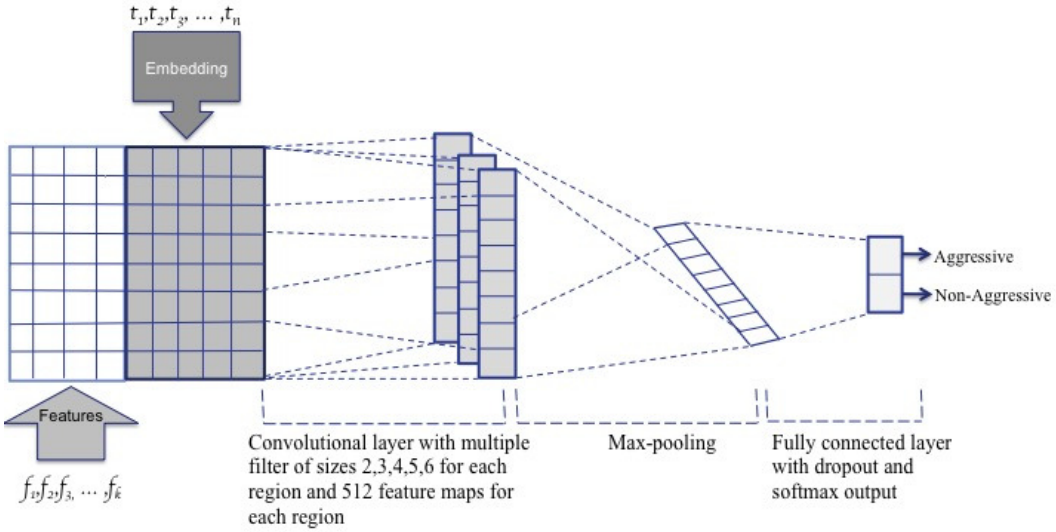

Fig. 1. Deep learning architecture

the tweet $W_{1:m}$. Therefore, $r_{1:n+m} = r_1 \oplus r_2 \oplus ... \oplus r_{n+m}$, where either $r_i \in F_{1:n}$ or $r_i \in W_{1:m}$.

Let $r_{j:j+i}$ refer to the concatenation of $r_j, r_{j+1}, \ldots, r_{j+i}$. In the convolution operation, the filter $t \in \mathbb{R}^{hp}$ is applied to a window of h words to produce new features such as feature A_j is generated from a window of words $r_{j:j+h-1}$ by $a_i = f(t.r_{j:j+h-1} + b)$, where, $b \in \mathbb{R}$ is a bias term and f is a non-linear function. A feature map $O = [O_1, O_2, \ldots, O_{m-h+1}]$ (where, $O \in \mathbb{R}^{m-h+1}$) is produced by applying the aforesaid filter to each possible window of h words (i.e., $\{r_{1:h}, r_{2:h+1}, \ldots, r_{m-h+1:m}\}$) in the tweet. After applying the max-pooling operation to the feature map O, the maximum value $O' = \max\{O\}$ is obtained for the particular filter.

The prime goal of the max pooling operation is to capture the most important feature with the highest value for each feature map. In this work, the proposed framework uses multiple filters with varying window sizes to obtain multiple features.

After extracting the features, these features are provided as input to the fully-connected layer.

Fully-connected layer: In the fully-connected layer (sometimes called as the dense layer), the best features which are selected by the max-pooling operation from the convolutional kernel are combined. The output of this layer is passed to the next layer, i.e., the output layer.

Output layer: This layer is the final layer of this proposed architecture. This layer is made of 2 neurons. Each neuron is for a target class, i.e., one neuron for the aggressive class and another for the non aggressive class. The 'softmax' is used as the nonlinear activation function.

4.2 Linguistic Features

As said before, in one of the models feature engineering is combined with DL architecture. The set of selected features comes from an accurate analysis of the dataset, and involves stylistic, emotional and semantic aspects of Mexican tweets annotated like aggressive.

Stylistic features: Aspects like affect and personality could be captured by stylistic information [2]. For this reason, specific traits of author writing have been taken into account, like: the presence of Mexican abbreviations more used in tweet context, such as hdp, alv, the use of punctuation elements (question *¿?*, exclamation marks *¡!* and sequences of dots ...) and uppercase characters.

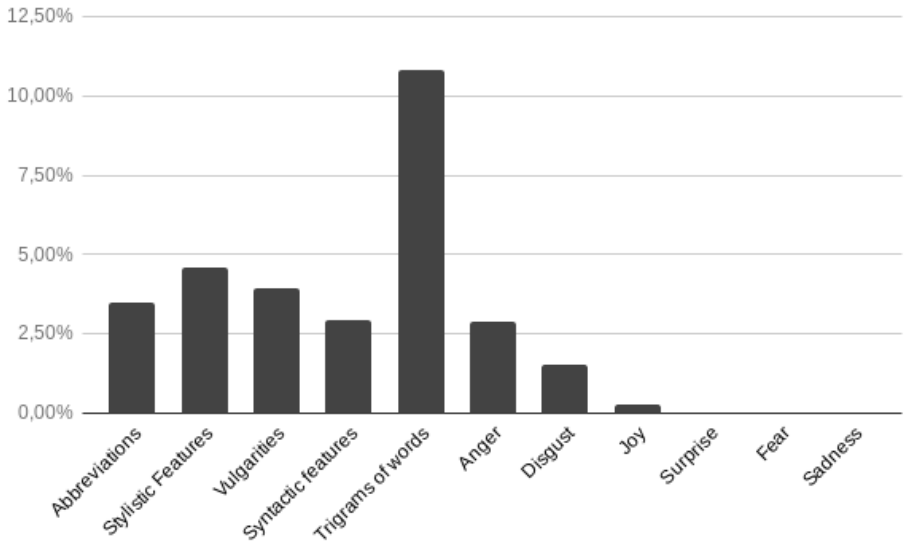

Fig. 2. Relevance of linguistic features by InfoGain measure

In particular, the system inspects if the user writes all the word in uppercase, or only the first letter, or uses capital letters inside the word. Moreover, quantitative features, such as the number of characters and words per sentence and the average word length, are considered.

Among the stylistic features, the emoticons play an important role in the digital writing. Thus, the use of emoticons[9] annotated with polarity (positive / negative / neutral) is taken into account. Actually, emoticons are used as representations of facial expressions giving contextual information to readers.

Vulgarities and profanities: Aggressive texts aim to offend and attack psychologically the victims addressing their dignity with insults, humiliating adjectives or vulgar expressions. Therefore, two collections of profanities and derogatory adjectives have been created in order to help the system to detect aggressive texts like:

> "*Te vas a chingar a tu madre pinche estupido pendejo!!!*[10]"

Syntactic features: Considering the fact that the purpose of an aggressive tweet is to insult and offend someone, identifying the presence of a target is important for this task. Thus, one of the method employed is to locate the mention of the target by means of specific syntactic patterns: the mention of the proper name or *@usuario* followed or preceded by words or expressions from the lists of impolite adjectives or vulgar expressions.

Trigrams of words: Considered the variety of combination of Mexican insulting expressions and their length, the 100 most relevant sequences of trigrams of words have been chosen and weighted with term frequency-inverse document frequency (TF-IDF). In order to understand the importance of this feature for this task, Table 2 reports some of the most frequent trigrams in the analyzed dataset in comparison with unigrams and bigrams of words.

Actually, the different possible n-grams have been tested and the trigrams obtained the best performance. Indeed, as Table 2 shows, the trigrams of words are the most significant respect to unigrams and bigrams, because they capture the typical multi-word-expressions used in Mexican language as vulgar or semantically altered expressions.

[9] The annotated list of used emoticons for this work is provided by the Unicode Consortium: `http://www.unicode.org/`

[10] *Fuck you, stupid asshole!!!*

Table 2. The most frequent unigrams, bigrams and trigrams of words

Unigrams	Bigrams	Trigrams
('verga')	('la', 'verga')	('a', 'la', 'verga')
('madre')	('a', 'la')	('hasta', 'la', 'madre')
('putas')	('de', 'la')	('tu', 'puta', 'madre')
('putos')	('que', 'me')	('me', 'vale', 'verga')
('loca')	('que', 'no')	('a', 'chingar', 'a')
('pinche')	('los', 'putos')	('su', 'puta', 'madre')
('puta')	('la', 'madre')	('chingar', 'a', 'su')
('todo')	('en', 'la')	('sus', 'putas', 'madres')
('joto')	('en', 'el')	('todos', 'los', 'putos')
('ser')	('que', 'se')	('a', 'la', 'chingada')
('q')	('su', 'madre')	('chingas', 'a', 'tu')
('vida')	('las', 'putas')	('hijo', 'de', 'tu')
('vale')	('lo', 'que')	('mandar', 'a', 'la')
('marica')	('a', 'su')	('hijos', 'de', 'su')
('ver')	('y', 'no')	('la', 'puta', 'madre')
('luchona')	('que', 'te')	('me', 'vale', 'madre')
('mierda')	('puta', 'madre')	('chinga', 'tu', 'madre')
('solo')	('voy', 'a')	('estoy', 'hasta', 'la')

Moreover, the data analysis reveals that the majority of aggressive expressions in Mexican language are long combinations of insults, like the tweet:

> "*@USUARIO Adios, hijo te toda tu perra celestial puta madre.*[11]"

Therefore, the trigrams seem to represent a good choice for this specific task in Mexican language.

Affective information: In this study, one of the purposes is to examine the emotion related to aggressive language. Therefore, the system tends to capture the emotions that are expressed in the aggressive texts by the use of the Spanish Emotion Lexicon (SEL) provided by the authors of [26, 23]. Each word in this lexicon is associated with the six principal Ekman emotions (Joy, Anger, Surprise, Disgust, Sadness and Fear) in accordance with the *Probability Factor of Affective use* in Spanish. In this work, the words with a higher probability factor for each emotion have been considered. Moreover, the lexicon is extended by synonyms and slang forms usually used in social networks [22].

Finally, by means of the Information Gain analysis, the impact of emotions and the relevance of the delineated features in the recognition of aggressive tweets has been analyzed. The Figure 2 show the results.

The anger and disgust are the principal emotions involved in hate speech respect to joy, surprise, fear and sadness. Indeed, the anger, as well as the disgust, is the main emotion that drives the hate message toward someone. With respect to target, the syntactic patterns seem that they can help the system to classify correctly the aggressive tweets addressed to individuals. Moreover, the features most relevant for this task seem to be the trigrams of words, and this factor confirms the fact that insults in Mexican language are, in most of cases, longer composition of slurs than simple derogatory adjectives.

The next section describes the experiments carried out and the obtained results.

5 Experiments and Evaluation

In order to understand the impact of the features, two experiments have been carried out: firstly, the approach is employed splitting the training data and creating an appropriated development set; secondly, the system is evaluated by using the test set provided by the organizers to compare its performance with those of the participants.

In the first phase, taking into account the unbalanced distribution of aggressive and non-aggressive tweets (see Table 1), the set of 7700 tweets is split into 7000 samples for training and 700 as development set, perfectly separated in 350 positive and 350 negative samples.

For the evaluation phase, the same measure used in the competition is employed. Considering the unbalanced corpus, the organizers used as evaluation measure the F-measure for the positive class (i.e., aggressive tweets).

Table 3 shows the results obtained using the development set, and those obtained with the test set, with the ranks also that this approach could have obtained in the framework of the competition: 3th for DL and 10th for DL+FE.

These results are compared with the ones obtained by the best performing system and with the baselines provided by the organizers of the shared task.

[11] *@USUARIO Bye, son of fucking bitch.*

Table 3. Results obtained with F-score for positive class on the MEXA3T aggressiveness corpus

	Development set	**Test set**			
	F-score	*Precision*	*Recall*	*F-score*	*Rank*
Best performing system				*0.49*	*1*
DL	0.82	0.37	0.53	**0.44**	3
3-grams char baseline				*0.43*	-
DL+FE	**0.83**	0.38	0.42	0.40	10
BoW baseline				*0.37*	-

Using the development set, the linguistic features seem to help the classification task. However, in the evaluation with the test set, the values for DL+FE decrease compared to DL. Moreover, comparing the obtained results with the baselines, it is evident that the results by DL+FE overcomes only the bag of word (BoW) baseline and not the 3-grams char baseline. The 3-grams char baseline is obtained by the organizers, training the data on 3-grams of characters using SVM with linear kernel and C=1 [1]. A possible justification of the high result obtained by the 3-grams of characters is that with n-grams of characters the system could detect also the typographical mistakes or variations often found in informal texts like tweets.

6 Discussion

Considering the low results obtained including the linguistic features in the deep learning architecture, an error analysis has been carried out. In particular, the special cases will be discussed.

Jokes: Analyzing the data, several humorous cases are located. Actually, the users usually disguise aggressive comments as humorous, involving, principally, sarcasm or irony in their utterances. However, these are experienced by the target like harassment. For instance, misogynistic comments are perceived like sexual harassment [13] and, also, the continue exposition to sexist jokes can also modify the perception of sexism like norm and not like negative behavior [14]. These types of linguistic devices, like irony and sarcasm, could hinder the correct classification. Below some examples:

"*@USUARIO @USUARIO El señor tiene el superpoder de hablar mierda , cagar la madre y cambiar su color de piel a color naranja*[12]";

"*@USUARIO #LOS40MeetAndGreet 9 . Por q es una mamá luchona que cuida a su bendiciòn*[13]";

"*Gracias Facebook , pero no son personas que "quizá conozca" , son personas que conozco pero que me valen verga y no las quiero agregar.*[14]";

"*Aunque me cagues. . . . brilla , pero . . Algún día construiré mi súper misil para mandarte a la Merga #namaste #mamaste*[15]".

Laughters: The presence of certain elements such as the laughters generally imply that the text is not aggressive. However, in some cases the laughter elements seem to emphasize the offensive mockery expressed in the tweet. For instance the following tweet are not classified as aggressive probably because of these misleading elements.

[12] *@User @User This man has the superpower of producing shit with his mouth, fucking and changing the colour of his skin to orange.*

[13] *@User #LOS40MeetAndGreet 9 . Because she is a fighter mother who takes care of her kid.*

[14] *Thanks Facebook, but they are not persons who "maybe" I know, they are persons that I know but I don't care about them and I don't want to join them.*

[15] *Even I don't like you...shine, but.. one day I will build my super missile for sending you to Marshit #namaste #yousuck*

"*TRUMP , ESTADO UNIDOS Y SY PUTO MURO SE FUERO A CHINGAR A SU MEDRE SE QUEDARON SI MUNDIAL POR PUTOS JEJEJE*[16]";

"*LOS PUTOS SIEMPRE QUIEREN TODO DE A GRATIS jajaja no mamen :D*[17]".

Emphasized negative emotions: The users use also linguistic devices like superlative adjectives to emphasize the anger or the disgust toward someone. Analyzing the wrong predicted tweets, some tweets with sequences of insults that contain also superlative adjectives are not correctly classified. Below some examples:

"*Mándame una de 1000 por que te voy a mandar a chingar a su reputisima madre por putos y ratas*[18]";

"*@USUARIO Otra rata más ! ! , por igual que lo consiguen éstos imbéciles corruptos HDP no tienen madre ! !*[19]".

Abbreviation: In the majority of cases the users tend to abbreviate the words, especially the functional words such as pronouns or relative connectors. This kind of abbreviations, that in this work are not taken into account in the preprocessing of the data, seems to hinder the the correct classification. For instance, the following tweets are misclassified:

"*@USUARIO @USUARIO @USUARIO A ti t da pena mostrar tu foto , por tu cara de estupido y Maricon que tienes ve con el america a chingar a su madre*[20]";

"*@USUARIO HDP ! Citen 5 cosas q pasan en Venezuela y q temen los fanáticos y empinados a EPNdejo ! D las q digan , mencion...*[21]".

Particular exclamations: These cases are the most common misclassified tweets, but there are also some cases where the vulgar expression are not used to insult someone, such as:

"*No hay mejor sensación que darte cuenta que algo te vale chingos de verga*[22]".

Indeed, these kind of tweets aim to emphasize a subjective opinion and they are not addressed to someone. Like those, the vulgar expressions are also used as exclamations, for instance:

"Puta madre quiero dormirrrrrer[23]".

These reported examples reveal the difficulty to detect automatically aggressiveness especially in a context such as social platform. Indeed, the typical misspellings or grammar mistakes with our approach are difficulty treated. Moreover, the informal language is complex and overflowing with semantic exceptions that mislead the decision of the system.

7 Conclusions and Future Work

The main aim of this paper is to investigate the impact of linguistic features into a deep learning architecture for the detection of aggressive tweets. Although the deep learning architecture achieved a result comparable to the best performing systems (3rd ranking), feeding the DL-based approach with linguistic features did not help in obtaining better results on the test corpus, while it did on the training corpus. This seems to be due to the unbalanced distribution of topics into the two datasets (training and test).

[16] *Trump, USA and their wall go to hell, they are out of the World Cup because they are motherfucker jejeje.*

[17] *The motherfuckers always want all free jajaja fuck off*

[18] *Send me one of the 1000 because of you're bitch, fuck you.*

[19] *@User Other rat more!!, however they reach it, these fucking corrupt son of bitch don't have no shame*

[20] *@User @User @User You have shame to show your photo because of your face of jerk and gay, go to hell with the american.*

[21] *@User Son_of_bitch! Cite 5 events in Venezuela and of whom the fanatic people and raised to jerk are afraid! Of those I mention...*

[22] *There is nothing like that realize that you don't care about something.*

[23] *Fuck I want to sleeeeeeep.*

It will be important to investigate this issue further in the future, possibly on bigger datasets, in order to allow the DL-based approach to generalize from big data.

Another significant point is the highlighted presence of sarcastic and ironic utterances in this dataset. Indeed, as noticed in the error analysis, these linguistic devices in some aggressive tweets made their comprehension particularly difficult. Therefore, as future work, in order to improve our system's performance in this task it will be interesting to address the issues related to the specific use of figurative language devices such as irony and sarcasm in the hate speech detection.

Finally, in order to fully understand the impact of linguistic features on the deep learning approach, it will be helpful to experiment such technique with other corpora, also in other languages, in the framework of related tasks such as hate speech detection, misogyny and racism identification.

Acknowledgment

The work of Simona Frenda and Paolo Rosso was partially funded by the Spanish MINECO under the research project SomEMBED (TIN2015-71147-C2-1-P).

References

1. **Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018).** Overview of MEX-3AT at IBEREVAL: Authorship and aggressiveness analysis in Mexican Spanish tweets. *Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September (2018)*.
2. **Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., & Levitan, S. (2007).** Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 6, pp. 802–822.
3. **Banerjee, S., Naskar, S., Rosso, P., & Bandyopadhyay, S. (2018).** Code mixed cross script factoid question classification-a deep learning approach. *Journal of Intelligent & Fuzzy Systems*, Vol. 34, No. 5, pp. 2959–2969.
4. **Banerjee, S., Naskar, S. K., Rosso, P., & Bandyopadhyay, S. (2016).** The first cross-script code-mixed question answering corpus. *MultiLingMine@ ECIR*, pp. 56–65.
5. **Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., Morgan, J., & Sloan, L. (2015).** Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change*, Vol. 95, pp. 96–108.
6. **Burnap, P. & Williams, M. L. (2014).** Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making. *Social Network Analysis and Mining*.
7. **Carreras, X., Chao, I., Padró, L., & Padró, M. (2004).** Freeling: An open-source suite of language analyzers. *LREC*, pp. 239–242.
8. **Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012).** Detecting offensive language in social media to protect adolescent online safety. *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, IEEE, pp. 71–80.
9. **Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013).** Improving cyberbullying detection with user context. *European Conference on Information Retrieval*, Springer, pp. 693–696.
10. **Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017).** Hate me, hate me not: Hate speech detection on facebook. *Proceedings of ITASEC17*.
11. **Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012).** Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, Vol. 2, No. 3, pp. 18.
12. **Escalante, H. J., Villatoro-Tello, E., Garza, S. E., López-Monroy, A. P., Montes-y Gómez, M., & Villaseñor-Pineda, L. (2017).** Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications*, Vol. 89, pp. 99–111.

13. **Ford, T. E. & Boxer, C. F. (2011).** Sexist humor in the workplace: A case of subtle harassment. In *Insidious Workplace Behavior*. Routledge, pp. 203–234.
14. **Ford, T. E., Wentzel, E. R., & Lorion, J. (2001).** Effects of exposure to sexist humor on perceptions of normative tolerance of sexism. *European Journal of Social Psychology*, Vol. 31, No. 6, pp. 677–691.
15. **Fulper, R., Ciampaglia, G. L., Ferrara, E., Ahn, Y., Flammini, A., Menczer, F., Lewis, B., & Rowe, K. (2014).** Misogynistic language on Twitter and sexual violence. *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*.
16. **Gambäck, B. & Sikdar, U. K. (2017).** Using convolutional neural networks to classify hate-speech. *Proceedings of the First Workshop on Abusive Language Online*, pp. 85–90.
17. **Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015).** A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 10, No. 4, pp. 215–230.
18. **Justo, R., Corcoran, T., Lukin, S. M., Walker, M., & Torres, M. I. (2014).** Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, Vol. 69, pp. 124–133.
19. **Mahmud, A., Ahmed, K. Z., & Khan, M. (2008).** Detecting flames and insults in text. *Proc. of 6th International Conference on Natural Language Processing (ICON' 08)*.
20. **Mehdad, Y. & Tetreault, J. (2016).** Do characters abuse more than words? *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 299–303.
21. **Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016).** Abusive language detection in online user content. *Proceedings of the 25th international conference on world wide web*, pp. 145–153.
22. **Posadas-Durán, J.-P., Markov, I., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Gelbukh, A., & Pichardo-Lagunas, O. (2015).** Syntactic n-grams as features for the author profiling task. *Working Notes Papers of the CLEF*.
23. **Rangel, I. D., Guerra, S. S., & Sidorov, G. (2014).** Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomázein: Revista de lingüística, filología y traducción de la Pontificia Universidad Católica de Chile*, , No. 29, pp. 31–46.
24. **Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017).** A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*.
25. **Samghabadi, N. S., Maharjan, S., Sprague, A., Diaz-Sprague, R., & Solorio, T. (2017).** Detecting nastiness in social media. *Proceedings of the First Workshop on Abusive Language Online*, pp. 63–72.
26. **Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Dıaz-Rangel, I., Suárez-Guerra, S., Trevino, A., & Gordon, J. (2012).** Empirical study of opinion mining in Spanish tweets. *LNCS*, Vol. 7629, pp. 1–14.
27. **Spertus, E. (1997).** Smokey: Automatic recognition of hostile messages. *AAAI/IAAI*, pp. 1058–1065.
28. **Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., & Hoste, V. (2015).** Detection and fine-grained classification of cyberbullying events. *International Conference on RANLP*, pp. 672–680.
29. **Xu, J.-M., Jun, K.-S., Zhu, X., & Bellmore, A. (2012).** Learning from bullying traces in social media. *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, Association for Computational Linguistics, pp. 656–666.
30. **Zhang, Z. & Luo, L. (2018).** Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *arXiv preprint arXiv:1803.03662*.
31. **Zhong, H., Li, H., Squicciarini, A. C., Rajtmajer, S. M., Griffin, C., Miller, D. J., & Caragea, C. (2016).** Content-driven detection of cyberbullying on the instagram social network. *IJCAI*, pp. 3952–3958.

Article received on 30/10/2019; accepted on 07/03/2020.
Corresponding author is Simona Frenda.