

UACH-INAOE at HASOC 2019: Detecting Aggressive Tweets by Incorporating Authors' Traits as Descriptors

Marco Casavantes¹, Roberto López¹
, Luis Carlos González-Gurrola¹, and Manuel Montes-y-Gómez²

¹ Facultad de Ingeniería, Universidad Autónoma de Chihuahua (UACH), Mexico
{p271673,jrlopez,lcgonzalez}@uach.mx

² Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica, Óptica
y Electrónica (INAOE), Mexico
mmontesg@inaoep.mx

Abstract. In this paper, we describe our participation for the Aggressiveness Detection Track in English texts for HASOC 2019. We evaluate different strategies for text classification, including classifiers such as Logistic Regression and Support Vector Machines trained on n-grams (words and characters) and word embeddings for clustering techniques. We also study the incorporation of contextual characteristics to explore whether people verbally attack differently depending on their traits and environment.

Keywords: English text classification · Aggressiveness Detection · Twitter.

1 Introduction

As people increasingly communicate online through social media, they may deal with negative experiences such as being targets of cyberbullying or expose themselves to hateful and vulgar content. These problems have become more relevant in the past few years, as they pose several challenges to preserve the freedom of speech and sharing of ideas over these communication channels. The growth in the volume of the messages that are posted on social media on a daily basis demands more efficient means to detect and moderate the spread of offensive content and hate speech. Furthermore, administrators of social media platforms could prevent abusive behavior and harmful experiences. It is crucial to address the importance of early identification of users that promote hate speech, as this could enable important outreach programs, to prevent an escalation from speech to action [11]. Moreover, considering the high levels of aggressiveness and hostile behaviour of certain users towards particular groups or individuals, more serious

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). FIRE 2019, 12-15 December 2019, Kolkata, India.

real-life issues, like self-harm or suicide, could actually be prevented.

In the last years, several shared tasks have been organized with the purpose of attracting attention to these problems [14, 12, 7, 13]. Take for instance the second edition of MEX-A3T [4]. In that event, our participation focused on detecting aggressive tweets in a Mexican Spanish dataset, by incorporating traits of authors (e.g., occupation, location). Therefore, by participating in HASOC [9] (Sub-task A for English), we aimed to test our approach on a different collection of tweets, tweaking our system to face this new challenge.

In this study, we evaluate common strategies such as lexical feature engineering through *term frequency* representations (e.g., bag of words through *tfidf*), along with different approaches with the aim to enhance features by adding context to each document. Furthermore, we also advanced our research by including the authors' traits, and using the outcome of unsupervised methods as potential useful features.

The hypothesis behind our approach is that offensive messages could be better recognized by analyzing not only the message but the user profile. The rest of this document is organized as follows: in section 2 we describe our approach; in section 3, the results attained are detailed and analyzed; finally, in section 4 we state our conclusions and delineate some future work.

2 Proposed Method

Similar to our participation in MEX-A3T 2019 [5], we aim to enrich the classification of aggressive tweets by including a possible theme to which each tweet belongs, being this the main experiment that attempt to support our hypothesis. This section gives a complete description of the changes and adaptation of features that we propose in our approach.

2.1 Data Pre-processing

Once the text files were loaded using UTF-8 encoding, we conducted our experiments in a custom version of the dataset where:

- All words are made lowercase.
- Emojis are converted into their text representation.
(e.g., “:face_with_tears_of_joy:”)
- Tweets are stripped from non-alphanumeric characters excluding some relevant symbols (#, @ and _).
- Every URL (occurrence of the sequence “http”) was replaced with “weblink” to evenly represent references to external sources.

2.2 Features

We conducted our research using the following features:

Lexical: We use both word n-grams (n=1, 2) and char n-grams (n=2, 3, 4),

however this collection of terms was only weighted with its term frequency.

Document Embeddings: Using only the text available in both the train and test set, we employed a representation of the tweets through Word Embeddings [8] to feed different clustering strategies.

Grouping tweets by theme: We use different clustering methods (an implementation of Self Organizing Maps [1], K-Means and Affinity Propagation) to generate new features based on thematic terms in each tweet.

- The SOM allowed us to locate each tweet on a two-dimensional plane, taking the coordinates as new features.
- Using K-Means and Affinity Propagation we calculate, for every sample, the distance between itself and the rest of clusters.

Flesch Reading Ease and Flesch Kincaid Grade scores: Based on [6], we wanted to capture the quality of each tweet by getting the Flesch Reading Ease and Kincaid Grade scores using textstat [3]. In our experiments the number of sentences is also fixed at one.

Named Entity Recognition (NER) counters: Upon manual inspection of frequent tokens (Table 4), we observed that a big part of the dataset included references to people like Donald Trump (current president of USA), Boris Johnson (current Prime Minister of the United Kingdom), Mahendra Singh Dhoni (indian international cricketer) and organizations like ICC (International Cricket Council). Based on this information we decided to incorporate counters of how many persons, organizations and locations were mentioned in each text using polyglot [2].

3 Experiments and Results

The datasets were provided by the HASOC-2019 organization team. Table 1 shows the distribution of training and test partitions for English tweets.

Table 1. Data distribution for English tweets corpus used in HASOC-2019.

Class	Training	Test
Non Hate-Offensive (NOT)	3591	N/A
Hate and Offensive (HOF)	2261	N/A
Total	5852	1153

We started our research by recreating our baselines used in MEX-A3T 2019, this time focusing on the word unigrams and bigrams baseline, as it holds the best performance in this task in comparison to the character n-grams baseline. In order to generalize our results for the test set, we evaluated our experiments using two different configurations, a single stratified train-validation split and a 5-Fold Cross Validation.

We trained Linear Support Vector Machines and a Logistic Regression classifier for this task, and we decided to use both of them to submit our predictions:

- **Run 1** consists of a LinearSVM trained with the best 800 features from a Bag of Words of range=(1,2) considering the term frequency of all the tokens. The feature selection was done by a chi-squared statistics test on a 70-30% train-validation split.
- **Run 2** is the same as Run 1, but in this case the top 1250 features were selected from a stratified 5-fold cross validation on the train set, specifying a 20% split for the validation set.
- **Run 3** is the result of creating an ensemble of two Logistic Regression classifiers, one trained with a Bag of Words and the other one with a Bag of Character n-grams. The predictions were assigned by choosing the model with the highest probability for each tweet.

Table 2 shows the macro and weighted F1-score that we obtained over the two classes. We performed all modeling regarding the creation of term frequency feature matrices, classifiers, cross validation and Kmeans/Affinity Propagation clustering using scikit-learn[10].

3.1 Results for HASOC 2019

As stated before, a Linear Support Vector Machine was chosen as our system’s classifier adding Named Entity Recognition counters for runs 1 and 2, and a Logistic Regression classifier ensemble was used to submit run 3. Table 3 lists the results of our three submissions for the English Hate Speech and Offensive Content Identification Sub-task A for HASOC 2019, more information of all results of the contest is available at [9].

3.2 Analysis

We analyzed our participation in HASOC’19 in two ways. The first analysis focuses on observing what are the 10 most frequent n-grams (excluding stopwords) at word level (separated by length) in the Hate-Offensive class, these are shown in Table 4. We also exhibit in Table 5 the best word n-grams per class according to the Logistic Regression classifier (LRC) trained with the whole training set. In our final configuration, it was easier for an offensive tweet to be missclassified as non-aggressive, and despite running several experiments, most of our attempts to improve classification in this task by adding new features trying to give context to the tweets unfortunately affected the results negatively. After inspection, we observed that this could have happened because:

- The clustering techniques that we used didn’t add anything new since the tweets were kind of grouped from the beginning, as some main topics can be spotted (e.g., Trump, Dhoni/ICC and ”DoctorsFightBack” protest related tweets).

Table 2. Detailed classification with F1-scores in the validation stage.

Run	Features	Setup	Macro F1-score	Weighted F1-score
	Complete BoW	LinearSVM on single split	0.6315	0.6571
	Top 800 features from BoW	“	0.6452	0.6717
Run 1	NER + top features	“	0.6468	0.6731
	Complete BoW	LinearSVM on 5-FoldCV	0.6242	0.6527
	Top 1,250 features from BoW	“	0.6323	0.6645
Run 2	NER + top features	“	0.6352	0.6671
	Flesch Reading Ease Score	“	0.6181	0.6468
	Flesch Kincaid Grade Score	“	0.6205	0.6489
	NER counters	“	0.6250	0.6518
	K-Means Clustering	“	0.6237	0.6521
	Affinity Propagation	“	0.6213	0.6501
	SOM Coordinates	“	0.6185	0.6475
Run 3	Bag of Words and Chars.	Ensemble on single split	0.6174	0.6515
Run 3	Bag of Words and Chars.	Ensemble on 5-FoldCV	0.6227	0.6547

Table 3. Final scores of the 2019 Hate Speech and Offensive Content Identification Sub-task A in English.

Rank	Team	Macro F1-score	Weighted F1-score
1/79	YNU_wb	0.7882	0.8395
20/79	UACH-INA OE_english.1.run.3	0.7075	0.7828
29/79	UACH-INA OE_english.1.run.2	0.6765	0.7490
30/79	UACH-INA OE_english.1.run.1	0.6753	0.7491

- Since there were multiple cases of similar quality scores assigned to both not offensive and offensive messages, the classifiers could not pick a relevant pattern.

Table 4. Most frequent n-grams at word level in training set

Length	N-gram	Freq. in HOF class	Freq. in NOT class
Unigram	'fucktrump'	515	628
	'trumpisatraitor'	386	484
	'realdonaldtrump'	383	347
	'trump'	280	323
	'icc'	270	543
Bigram	'fucktrump weblink'	130	236
	'world cup'	65	127
	'trumpisatraitor weblink'	58	80
	'borisjohnsonshouldnotbepm weblink'	50	73
	'resisttrump fucktrump'	36	129

Table 5. Best word n-grams per class in training set

Class	N-gram	LRC Weight	- Class	N-gram	LRC Weight
HOF	'fuck'	-4.51	NOT	'dhonikeepstheglove'	3.67
	'fucking'	-2.90		'doctorsfightback'	3.16
	'dickhead'	-1.88		'dhoni'	1.62
	'youre'	-1.85		'shameonicc'	1.50
	'gandinaaliabuse'	-1.83		'doctors'	1.32
	'traitor'	-1.78		'borisjohnsonshouldnotbepm'	1.13
	'shit'	-1.60		'new'	1.08
	'you'	-1.55		'happy'	1.06
	'president'	-1.52		'happy johnmccainday'	1.05
'hes'	-1.51	'real'	0.98		

The second analysis addresses the performance of our proposal, regarding F1-score and contrasted against the rest of the competitors. Fig. 1 presents two box plots for the complete distribution of competitors in terms of Macro F1 and Weighted F1. This analysis suggests that the outcome achieved by our proposal is competitive, practically been located within the first quartile for all participants.

4 Conclusions and Future Work

In this paper, we describe our strategy to classify offensive and non-offensive tweets in a relatively new English collection of tweets. Regarding our experi-

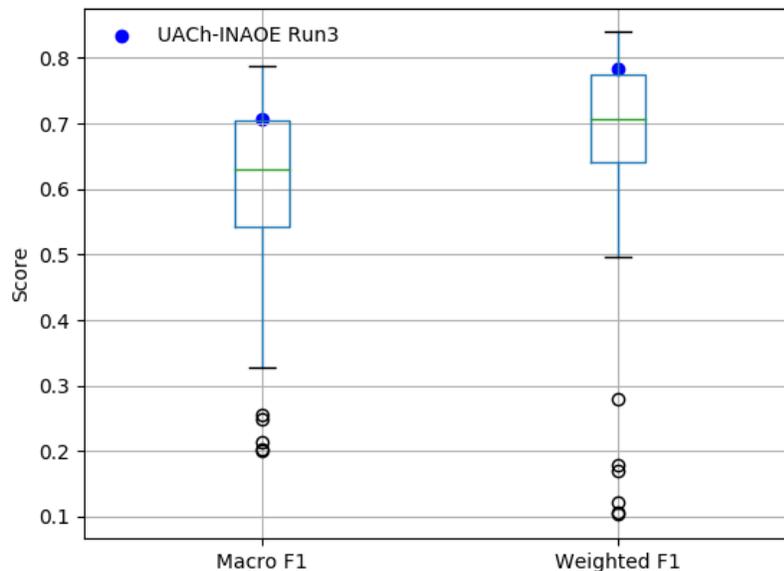


Fig. 1. Box plots of the results for Sub-task A for English.

ments for this task we can conclude that, in our best performing system, term frequency matrices of words and character n-grams complement each other in an ensemble of Logistic Regression classifiers. After seeing that the NER counters were basically the only useful features in the validation stage and the fact that we could not improve our classification scores with our current approach on providing context to tweets motivates the idea of future work focusing on finding new features to help us in our goal to see if it's possible to differentiate an offensive text from a non offensive one based on the message's underlying properties and the author's attributes.

References

- [1] Github - justglowing/minisom: Minisom is a minimalistic implementation of the self organizing maps. <https://github.com/JustGlowing/minisom>. (Accessed on 06/03/2019).

- [2] polyglot· PyPI. <https://pypi.org/project/polyglot/>. (Accessed on 09/09/2019).
- [3] textstat· PyPI. <https://pypi.org/project/textstat/>. (Accessed on 09/09/2019).
- [4] Aragón, M. E., Álvarez-Carmona, M. Á., Montes-y Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., and Moctezuma, D. (2019). Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets. In *Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain, September*.
- [5] Casavantes, M., López, R., and González, L. C. (2019). UCh at MEX-A3T 2019 : Preliminary Results on Detecting Aggressive Tweets by Adding Author Information Via an Unsupervised Strategy.
- [6] Davidson, T., Warmesley, D., Macy, M. W., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.
- [7] Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of TRAC*.
- [8] Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14*, pages II–1188–II–1196. JMLR.org.
- [9] Modha, S., Mandl, T., Majumder, P., and Patel, D. (2019). Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*.
- [10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- [11] Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. pages 88–93.
- [12] Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the semeval 2018 shared task on the identification of offensive language.
- [13] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019a). Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.
- [14] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019b). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.