IMT Mines Ales at HASOC 2019: Automatic Hate Speech Detection

Jean-Christophe Mensonides, Pierre-Antoine Jean, Andon Tchechmedjiev, Sébastien Harispe

IMT Mines Ales at HASOC 2019: Automatic Hate Speech Detection

Jean-Christophe Mensonides, Pierre-Antoine Jean, Andon Tchechmedjiev, and Sébastien Harispe

LGI2P, IMT Mines Ales, Univ Montpellier, Ales, France
firstname.lastname@mines-ales.fr

Abstract. This paper presents the contribution of the LGI2P (Laboratoire de Génie Informatique et d'Ingénierie de Production) team from IMT Mines Alès to the Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) 2019 shared task. This challenge aims at automatically identifying hate speech content in social media through three sub-tasks, each available in three different languages (English, German and Hindi). We are interested in sub-tasks A and B, requiring to (A) classify tweets as offensive or as non offensive, and (B) to further classify offensive tweets from sub-task A as hate speech, offensive speech or profane. We trained a fastText model for each proposed language and obtained promising results on the Hindi dataset for both sub-tasks A and B.

Keywords: Hate Speech Identification and Offensive Detection[1] · Tweet Classification

1 Introduction

With the rise in popularity of social media in recent years, it has become easier than ever to convey a point of view and spread ideas across the world. While most ideas can be heard, some of them seem wrong from an ethical point of view, such as encouraging someone to commit a crime or harassing another human being because of his ethnicity. Protecting the youth from those kinds of unethical ideas is an important societal challenge to overcome (27% of children in UK had a social network profile in 2007 [4]). However, with the ever increasing number of tweets posted everyday, manual monitoring is not a practical solution. The Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) 2019 Shared Task [6] proposes to tackle the lack of scalability of human monitoring by automatically identifying hate speech content.

Specifically, three sub-tasks are proposed in this challenge. For each task, three languages are proposed (English, German and Hindi). In this paper, we are interested exclusively in sub-task A and sub-task B, for all three proposed languages.

Sub-task A requires to classify tweets into two classes:

- Hate and Offensive (HOF) - Tweets containing any form of non-acceptable language (hate speech, aggression, profanity)
- Non Hate-Offensive (NOT) - Tweets not containing any offensive content.

Sub-task B is a fine-grained classification of offensive tweets from task A. HOF tweets must be classified into three classes:

- Hate speech (HATE) - Tweets containing hateful comments toward groups because of their race, political opinion, sexual orientation, health condition or similar.
- Offensive (OFFN) - Tweets degrading, dehumanizing or insulting an individual.
- Profane (PRFN) - Tweets containing unacceptable language in the absence of insults or abuse. It mainly concerns the usage of swearwords.

Recents advances in natural language processing have been applied to tweet classification. Duppada et al. [3] used deep learning techniques with ensemble learning methods to perform sentiment analysis in tweets. Wu et al. [8] used densely connected Long Short-Term Memory recurrent neural networks to detect irony in tweets. Çöltekin and Rama [1] predicted emoji in tweets using Support Vector Machines. Pérez and Luque [7] detected hate speech against women and immigrants in spanish using Support Vecotr Machines.

2 Corpus description

2.1 Class imbalance analysis

One dataset was given for each language. The English, German and Hindi datasets respectively contain 5852, 3819 and 4665 labeled tweets. The class distribution for each pair (language, sub-task) is shown in Table 1.

Regarding sub-task A, the datasets are roughly balanced between OFF and NOT classes both in English and in Hindi, while being highly imbalanced in German (with only 11 % of OFF tweets). Regarding sub-task B, the datasets are slightly imbalanced for each language, where one class represents 50% of the tweets. It is important to note that due to the low number of OFF tweets in German for sub-task A, each class in sub-task B has a really low number of tweets, dropping to only 86 PRFN tweets. This makes this classification task closer to a one-shot learning problem than a traditional deep learning problem.

2.2 Hashtag analysis

Intuitively one could think that it's easy to classify a tweet according to its hashtags, e.g tweets containing #FuckTrump, #DickHead or #DoucheBag hashtags

	Sub-task A			Sub-task B			
	OFF	NOT	Total	HATE	OFFN	PRFN	Total
English	2261	3591	5852	1143	451	667	2261
	0.39	0.61	1.0	0.5	0.2	0.3	1.0
German	407	3412	3819	111	210	86	407
	0.11	0.89	1.0	0.27	0.52	0.21	1.0
Hindi	2469	2196	4665	556	676	1237	2469
	0.53	0.48	1.0	0.23	0.27	0.5	0.1

Table 1. Class distribution analysis for each pair (language, sub-task)

should be offensive. Actually this is mostly not true. Figure 1 displays if an English tweet is labeled as NOT or HOF according to its hashtags. As we can see, most of those hashtags cannot solely be used to accurately choose between NOT and HOF, appart from a few specific ones such as #DoctorsFightBack and #DoctorsProtest.

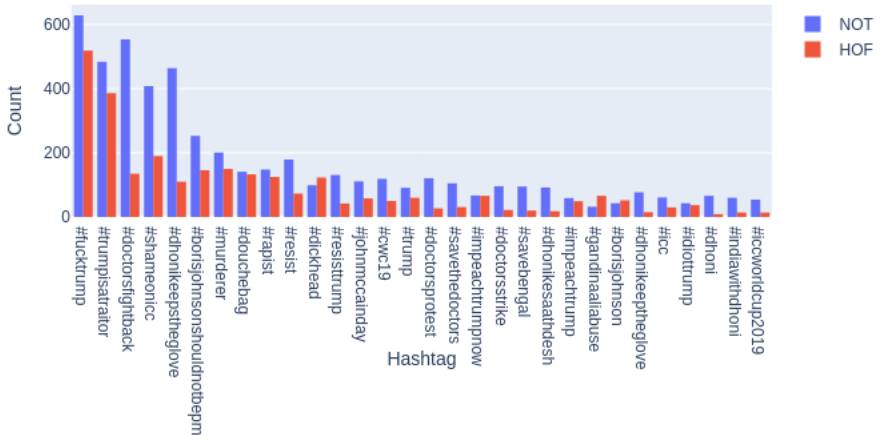

Fig. 1. OFF or NOT by hashtag according to the 30 most used hashtags in the English dataset

2.3 Emoji analysis

Emoji are little images used to convey feelings in electronic messages. As for hashtags, one could think there is a strong correlation between the use of emoji, usually carrying strong emotional information, and the HOF or NOT labels. Actually the presence of some emoji (folded hands, backhand index pointing

down, middle finger) in a tweet is a strong indicator of its offensiveness, as shown in figure 2.

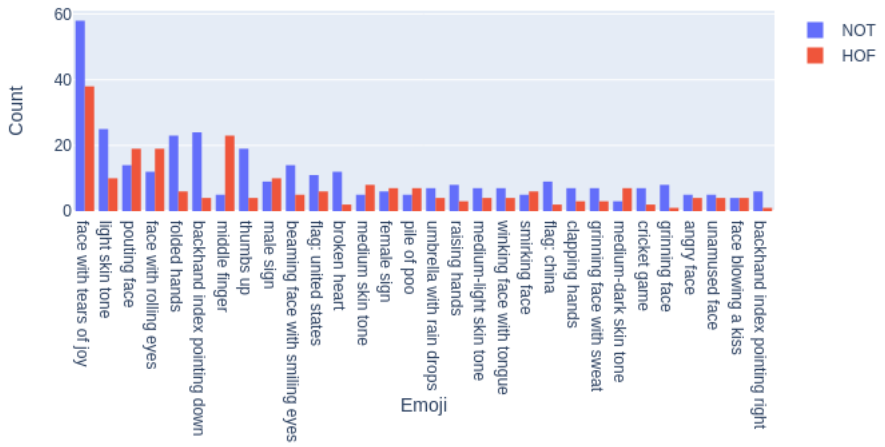

Fig. 2. OFF or NOT by emoji according to the 30 most used emoji in the English dataset

3 System description

3.1 Tweets preprocessing

We used tweettokenize[2] to replace tweeter specific tokens to *USERNAME*, *NUMBER* and *URL* special tokens. Emoji were replaced by their name[3], as suggested by Duppada et al. [3]. As an example, ☺ was replaced by *emoji_start Smiling Face emoji_end*, where *emoji_start* and *emoji_end* are special tokens.

3.2 System

Since the number of labeled tweets in German was on the low side, especially regarding sub-task B, we made the choice to not try models comprising of millions of parameters (e.g, BERT [2]) and aimed at smaller ones. We achieved our best F1-scores on our development set[4] using fastText [5]. We trained a fastText model for each pair (language, sub-task) independently.

[2] https://github.com/jaredks/tweetokenize

[3] Emoji names were scrapped from https://www.compart.com and https://emojipedia.org

[4] The development was obtained by randomly sampling 20% of the training set.

4 Results

We report our results for each pair (language, sub-task) in Table 2. The *F1-score* column represents the best F1-score we achieved on the test set, the *Best F1-score* column represents the best F1-score achieved by any team in this challenge and the *Rank* column represents our ranking (the denominator represents the number of different teams reported on the leaderboard).

	Sub-task A			**Sub-task B**		
	F1-score	Best F1-score	Rank	F1-score	Best F1-score	Rank
English	0.6649	0.7882	17/36	0.4188	0.5446	9/24
German	0.4641	0.6162	14/15	0.2348	0.3468	11/12
Hindi	0.8111	0.8149	2/18	0.5617	0.5812	3/15

Table 2. Results obtained for each pair (language, sub-task)

While we used the same method for every language, the results obtained are really heterogeneous. Indeed, our system performed really well in Hindi both on sub-tasks A and B, performed averagely in English on sub-task A and above average on sub-task B, while performing poorly in German on both sub-tasks.

We can see that the obtained results are scaling according to the balance of the datasets. Hence, spending time tackling the class imbalance issue should be a priority in our future works. The poor results on the German dataset can be explained since the model predicted the NOT labels for every tweet during inference, except once.

5 Conclusion

This paper presents the contribution of the LGI2P on sub-tasks A and B of the HASOC 2019 shared task. Those tasks required, for three different languages (English, German, Hindi), to (A) detect if tweets were containing any form of hate or offensive speech, and to (B) further classify offensive tweets from sub-task A as either hate speech, offensive speech or profane. We applied some simple prepossessing methods and trained the same model once for each language. We obtained promising results on the Hindi dataset both on sub-tasks A and B, ranking respectively as second and third.

References

1. Çöltekin, Ç., Rama, T.: Tübingen-oslo at SemEval-2018 task 2: SVMs perform better than RNNs in emoji prediction. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 34–38. Association for Computational Linguis-

tics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/S18-1004, https://www.aclweb.org/anthology/S18-1004
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Duppada, V., Jain, R., Hiray, S.: Seernet at semeval-2018 task 1: Domain adaptation for affect in tweets. arXiv preprint arXiv:1804.06137 (2018)
4. Goldsmith, A.: Disgracebook policing: social media and the rise of police indiscretion. Policing and society **25**(3), 249–267 (2015)
5. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics (April 2017)
6. Modha, S., Mandl, T., Majumder, P., Patel, D.: 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 (December 2019)
7. Pérez, J.M., Luque, F.M.: Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 64–69. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2008, https://www.aclweb.org/anthology/S19-2008
8. Wu, C., Wu, F., Wu, S., Liu, J., Yuan, Z., Huang, Y.: THU_NGN at SemEval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 51–56. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/S18-1006, https://www.aclweb.org/anthology/S18-1006