Vista.ue at SemEval-2019 Task 5: Single Multilingual Hate Speech Detection Model

Kashyap Raiyani, Teresa Gonçalves, Paulo Quaresma and Vitor Beires Nogueira
Computer Science Department, University of Évora, Portugal
(kshyp,tcg,pq,vbn)@uevora.pt

Abstract

This paper shares insight from participating in SemEval-2019 Task 5. The main propose of this system-description paper is to facilitate the reader with replicability and to provide insightful analysis of the developed system. Here in Vista.ue, we proposed a single multilingual hate speech detection model. This model was ranked 46/70 for English Task A and 31/43 for English Task B. Vista.ue was able to rank 38/41 for Spanish Task A and 22/25 for Spanish Task B.

1 Introduction

According to the article (Bosco et al., 2017), nearly a quarter of a billion people, throughout the world, currently live in a country other than their place of birth. This is an increase of 41% from 2000 to 2015. This figure includes more than 21 million refugees often vulnerable and dissatisfied. Since 2015, Europe is facing an unprecedented refugee crisis, the by-effect of the Syrian civil war and the terrible living conditions in equatorial Africa. 1,300,000 people have generated this increased migration flow to Europe which can only but increase, putting European stable societies, so far, under pressure.

Therefore, the implications for the European society and the way we behave towards immigration, immigrant integration and social inclusion for newcomers and their children, are becoming more decisive and must be addressed either at a local or global level, considering a political and social perspective. While this phenomenon stimulates the generation and diffusion of hate speech and hate crimes, at the same time several initiatives are promoted, but they should be further improved to increase the awareness and empathy of receiving populations while avoiding polarization against immigrants.

Hate speech analysis and hate maps allow both a greater understanding of social phenomena linked to the integration of migrants, that more targeted actions to improve it. The integration of migrants is strongly linked to the new cultural context where they try to rebuild their lives. The process of acculturation depends on personal and social variables of the migrant, in large part in turn dependent on the cultural context of his/her origin, on the characteristics of the context of resettlement and on events occurring during this life period. The different migrants strategies firstly affect the different outcomes achieved. In particular, he can decide whether or not to maintain the cultural identity of origin and whether or not to establish and maintain new relationships within the new contest. This gives rise to four possible different outcomes: integration, assimilation, separation/segregation, marginalization (Berry, 1997).

1.1 Motivation

Data released by European Community about population change (Union, 2015) show that from the 1990s onwards natural population change had a diminishing role in EU demographic developments, while the role of net migration became increasingly important. In the period 2011 to 2013, net migration contributed more than 80% to total population growth, drawing an overall pattern of growth of EUs populations driven increasingly by changes in migratory flows, which hides a range of demographic situations among the EU Member States. Between 2004 and 2013, indeed the population of 11 EU Member States decreased, with the biggest reductions recorded in Germany and Romania, but a high overall increase in population numbers was recorded in the other countries like UK (a gain of 4.51 million inhabitants), Spain (3.96 million), France (3.54 million) and Italy (3.29 million). Among these countries, character-

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 524–528
Minneapolis, Minnesota, USA, June 6–7, 2019.

ized by a negative natural population change, also compounded by negative net migration, Italy is affected by a negative natural change, that was completely offset by net migration which accounted for 108% of the total population change.

As a part of the motivation, we participated in the shared task named "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter". Section 2 outlines the existing approaches in a systematic manner and the description of the task mentioned in Section 3. Paper also provides a short, comprehensive and structured overview of automatic hate speech detection in Section 4 followed my result comparison and conclusion in Section 5 and 6 respectively.

2 Related Work

For any text classification task, the most obvious information to utilize is surface-level features, such as a bag of words. Indeed, unigrams and larger n-grams are included in the feature sets by a majority of authors (Chen et al., 2012; Sood et al., 2012; Xu et al., 2012; Warner and Hirschberg, 2012; Van Hee et al., 2015). These features are often reported to be highly predictive. Still, in many works, n-gram features are combined with a large selection of other features. For example, in their recent work, (Nobata et al., 2016) report that while token and character n-gram features are the most predictive single features in their experiments, combining them with all additional features further improves performance.

Character level n-gram features might provide a way to attenuate the spelling variation problem often faced when working with user-generated comment text. For instance, the phrase "ki11 yrslef a$$hole", which is regarded as an example of hate speech, will most likely pose problems to token based approaches since the unusual spelling variations will result in very rare or even unknown tokens in the training data. While using Character level approaches, on the other hand, are more likely to capture the similarity to the canonical spelling of these tokens. Author (Mehdad and Tetreault, 2016) systematically compare character n-gram features with token n-grams for hate speech detection and found that character n-grams prove to be more predictive than token n-grams.

Apart from word and character based features, hate speech detection can also benefit from other surface features (Chen et al., 2012; Nobata et al., 2016), such as information on the frequency of URL mentions and punctuation, comment and token lengths, capitalization, words that cannot be found in English dictionaries, and the number of non-alpha numeric characters present in tokens.

Hate speech and sentiment analysis are closely related, and it is safe to assume that usually, negative sentiment pertains to a hate speech message. Because of this, several approaches acknowledge the relatedness of hate speech and sentiment analysis by incorporating the latter as an auxiliary classification. Author (Dinakar et al., 2012; Sood et al., 2012; Njagi et al., 2015) followed a multistep approach in which a classifier dedicated to detect negative polarity is applied prior to the classifier specifically checking for evidence of hate speech. Further, (Njagi et al., 2015) run an additional classifier that weeds out non-subjective sentences prior to the aforementioned polarity classification.

3 Task Description and Dataset

The main task (Basile et al., 2019) was to detect Hate Speech in Twitter toward two different targets, immigrants and women. The data were available in a multilingual perspective, English, and Spanish.

3.1 Task Description

The task was partition into two groups: Task A and Task B. Making a total of four subtasks (English/Spanish task A/B).

TASK A - Hate Speech(HS) Detection against Immigrants and Women: a two-class (or binary) classification where systems have to predict whether a tweet in English or in Spanish with a given target (women or immigrants) is hateful or not hateful.

TASK B - Aggressive behavior(AG) and Target Classification(TR): where systems are asked first to classify hateful tweets for English and Spanish (e.g., tweets, where Hate Speech against women or immigrants has been identified,) as aggressive or not aggressive, and second to identify the target harassed as individual or generic (i.e. single human or group).

A binary value (1/0) indicating if HS is occurring against one of the given targets (women or immigrants). If HS occurs (i.e. the value for the feature at point HS is 1), a binary value indicat-

ing if the target is a generic group of people (0) or a specific individual (1) denoted as TR. And if HS occurs (i.e. the value for the feature at point HS is 1), a binary value indicating if the tweeter is aggressive (1) or not (0) denoted as AG. Thus, making 3 columns (named HS, TR and, AG) for each tweet.

3.2 Dataset

As per detail provided by the organizing committee, all data for the competition were collected from Twitter and manually annotated mainly via the "Figur8 crowdsourcing platform". The Table 1 describes the distribution of the dataset.

Language	Task	Train	Dev	Test
English	A	9000	1000	3000
English	B	9000	1000	3000
Spanish	A	5000	500	1600
Spanish	B	5000	500	1600

Table 1: Task Dataset Distribution.

The Table 2 and 3 describes the Hate Speech Tweet data distribution/property over training and development dataset.

Task	Non-HS	HS
EN-A	5790	4210
EN-B	5790	1463 (TR=AG=0)
ES-A	2921	2579
ES-B	2921	315 (TR=AG=0)

Table 2: Task A/B Data Property of Non-HS/HS.

Lan	TR(AG=0)	AG(TR=0)	AG=TR=1
EN	984	1187	576
ES	86	498	1180

Table 3: Task B Data Property (HS=1).

4 System Description

This section will talk about the preprocessing of the data, the experimental setup, and the multilingual system architecture.

4.1 Tweet Preprocessing

Here, for EN/ES tweets, we are only removing "url" from each tweet. This is done with the help of regular expression.

```
r"http\S+", "url", tweet
```

4.2 Experimental Setup

Here, a common architecture is used for all the four subtasks. The only difference is the hyper-parameter. The Table 4 shows the experimental parameter values.

Task	Paramter	Value
EN/ES - A/B	batch_size	1
EN/ES - A/B	epochs	2
EN/ES - A/B	optimizer	Adam
EN/ES - A/B	validation_split	0.20

Table 4: Experimental Parameter.

4.3 Single Multilingual System Architecture

Author (Raiyani et al., 2018) have used simple feedforward dense architecture and able to achieve beyond the average result for finding aggression over social media (Facebook and Twitter). In particular, their model was able to stand the best performing model for English Tweets. Using a similar architectural concept, here, we are using a character-based dictionary. First of all, all the unique characters from the dataset are stored in the form of a dictionary. Then, using this dictionary, each character in the dataset are replaced by its key value. Thus, this transforms the dataset into an integer from the text. Finally, this integer data is further transformed into a binary array and fed to the Dense architecture. The Figure 1 shows the flowchart of system process. Where as the Figure 2 shows the Dense Architecture.

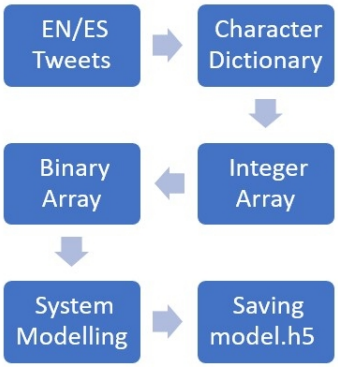

Figure 1: System Flowchart.

To store the intermediate character into the dictionary, pika library[1] was used. The number of unique characters found for English and Spanish

[1] https://pika.readthedocs.io/en/stable/

is respectively 169 and 172 (this also includes all the special character and emojis).

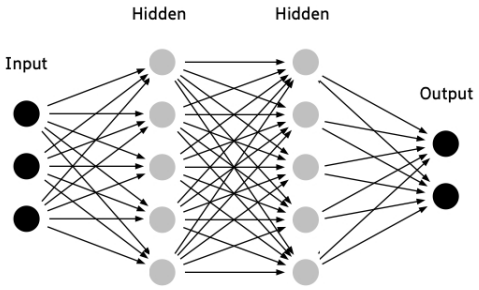

Figure 2: Feedforward Dense Architecture.

The architectural hyper parameter were selected based on trail and run. The same can be found in the Table 5 . The code of the entire task could be found in the online GitHub repository (Raiyani, 2019).

Task	Dense	Value	Activation
EN-A/B	layer 1	100	Relu
EN-A	layer 2	200	Sigmoid
EN-B	layer 2	200	Relu
ES-A/B	layer 1/2	50	Relu
EN/ES-A/B	layer 3	2	Softmax

Table 5: Architecture Parameter.

In the next section we will talk about the system performance and its global standing in the task.

5 Result Comparison and Discussion

The Table 6 shows the English task A average precision, recall, and F1 measure in reference to the baseline (SVC and MFC). The same for Spanish task A is found in the Table 7. The Table 8 shows the F1 measure over all the three parameter (namely, Hate Speech (HS), Target Classification(TR), and Aggressive(AG)). The ranking of task B is done using the value of Exact Match Ratio(EMR) (the evaluation formula could be found here[2]). The Table 9 shows the EMR value in reference to the baseline SVC and MFC.

The provided final ranking among all the subtasks are shown in Table 10.

6 Conclusion and Future Work

In this system description paper, we presented a single multilingual model for hate speech detection among immigrant and women. Through the

[2] https://competitions.codalab.org/competitions/19935

System	P	R	F1
Heigh	0.690	0.679	0.651
SVC	0.595	0.549	0.451
Vista.ue	0.483	0.488	0.420
MFC	0.289	0.500	0.367

Table 6: English - Task A Result.

System	P	R	F1
Heigh	0.734	0.741	0.730
SVC	0.701	0.707	0.701
Vista.ue	0.596	0.593	0.594
MFC	0.294	0.500	0.370

Table 7: Spanish - Task B Result.

System F1	Low	High	Obtain
EN B - HS	0.348	0.602	0.463
EN B - TR	0.372	0.752	0.596
EN B - AG	0.214	0.621	0.530
ES B - HS	0.370	0.761	0.573
ES B - TR	0.424	0.824	0.640
ES B - AG	0.413	0.760	0.578

Table 8: English/Spanish Task B F1 Result.

EMR	MFC	SVC	High	Obtain
EN B	0.580	0.308	0.580	0.284
ES B	0.588	0.605	0.635	0.536

Table 9: English/Spanish Task B EMR Result.

Task	System	Rank
EN A	SVC	35
	Vista.ue	46
	MFC	68
EN B	MFC	1
	SVC	27
	Vista.ue	31
ES A	SVC	21
	Vista.ue	38
	MFC	41
ES B	SVC	13
	Vista.ue	23
	MFC	18

Table 10: System Ranking.

system ranking, we can see that for task A of both the languages, the system is performing better than MFC baseline where on task B results could be improved. Further, We consider that our system can be grown, mainly due to the following facts: (1) The system does not count any NLP feature into account (2) Due to this, many hate tweets are missed. (3) Especially, for task B, features like Part of Speech (POS) tagging and Entity Extraction (EE) can improve the result. Lastly, how to address these aspects and generate a more accurate, comprehensive and fine-grained hate speech detection remains our further work.

Acknowledgments

The authors would like to thank COMPETE 2020, PORTUGAL 2020 Programs, the European Union, and ALENTEJO 2020 for supporting this research as part of Agatha Project SI & IDT number 18022 (Intelligent analysis system of open of sources information for surveillance/crime control).

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*. Association for Computational Linguistics.

John W. Berry. 1997. Immigration, acculturation, and adaptation. *Applied Psychology*, 46(1):5–34.

Cristina Bosco, Viviana Patti, Marcello Bogetti, Michelangelo Conoscenti, Giancarlo Francesco Ruffo, Rossano Schifanella, and Marco Stranisci. 2017. Tools and resources for detecting hate and prejudice against immigrants in social media. In *SYMPOSIUM III. SOCIAL INTERACTIONS IN COMPLEX INTELLIGENT SYSTEMS (SICIS) at AISB 2017*, pages 79–84. AISB.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 71–80, Washington, DC, USA. IEEE Computer Society.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Trans. Interact. Intell. Syst.*, 2(3):18:1–18:30.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303. Association for Computational Linguistics.

Dennis Njagi, Z Zuping, Damien Hanyurwimfura, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10:215–230.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Kashyap Raiyani. 2019. Single multilingual hate speech detection model. https://github.com/kraiyani/Single-Multilingual-Hate-Speech-Detection-Model.

Kashyap Raiyani, Teresa Gonçalves, Paulo Quaresma, and Vitor Beires Nogueira. 2018. Fully connected neural network with advance preprocessor to identify aggression over facebook and twitter. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 28–41. Association for Computational Linguistics.

Sara Owsley Sood, Elizabeth F. Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *J. Am. Soc. Inf. Sci. Technol.*, 63(2):270–285.

European Union. 2015. *People in the EU: who are we and how do we live?*, 2015 edition edition. Luxembourg.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the 10th Recent Advances in Natural Language Processing (RANLP 2015)*, Hissar, Bulgaria.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 656–666, Stroudsburg, PA, USA. Association for Computational Linguistics.