CIC at SemEval-2019 Task 5: Simple Yet Very Efficient Approach to Hate Speech Detection, Aggressive Behavior Detection, and Target Classification in Twitter

Iqra Ameer, Muhammad Hammad Fahim Siddiqui, Grigori Sidorov, and Alexander Gelbukh
Instituto Politécnico Nacional (IPN),
Center for Computing Research (CIC),
Av. Juan Dios Batiz, s/n, Zacatenco, 07738, Mexico City,
Mexico
{iqraameer133,hammad.fahim57}@gmail.com, {sidorov,gelbukh}@cic.ipn.mx

Abstract

In recent years, the use of social media has increased incredibly. Social media permits Inter-net users a friendly platform to express their views and opinions. Along with these nice and distinct communication chances, it also allows bad things like usage of hate speech. Online automatic hate speech detection in various aspects is a significant scientific problem. This paper presents the Instituto Politécnico Nacional (Mexico) approach for the Semeval 2019 Task-5 [Hateval 2019] (Basile et al., 2019) competition for Multilingual Detection of Hate Speech on Twitter. The goal of this paper is to detect (A) Hate speech against immigrants and women, (B) Aggressive behavior and target classification, both for English and Spanish. In the proposed approach, we used a bag of words model with preprocessing (stemming and stop words removal). We submitted two different systems with names: (i) CIC-1 and (ii) CIC-2 for Hateval 2019 shared task. We used TF values in the first system and TF-IDF for the second system. The first system, CIC-1 got 2nd rank in subtask B for both English and Spanish languages with EMR score of 0.568 for English and 0.675 for Spanish. The second system, CIC-2 was ranked 4th in subtask A and 1st in subtask B for Spanish language with a macro-F1 score of 0.727 and EMR score of 0.705 respectively.

1 Introduction

The social media applications enable users to discover, create and share contents handily, without specific expertise. This remarkably boosted the amount of data generated by the users, within a process that some people call "democratization" of the web (Silva et al., 2016). Still, this liberty also permits for the publication of data, which is insulting and hurtful both regarding the ethics of democracy and the privileges of some categories of people – hate speech (HS). The Hate Speech (HS) term is defined in the literature as an expression *"that is abusive, insulting, intimidating, harassing, and incites to violence, hatred, or discrimination. It is directed against people by their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth."* (Erjavec and Kovacic, 2012). HS has turned into the main issue for each sort of online website, where user-produced content comes into sight: from the comments on any post to live chatting in online games. Such material can isolate users and inflame violence (Allan, 2013). Website operators as Facebook, Twitter, and gaming companies like Runic Games recognize that hateful data are creating both practical and ethical issues and have attempted to demoralize them, causing changes in their platforms or strategies.

As stated by Pew[1], women experienced more sexualized forms of abuse than men. Platforms as Twitter are flopping in acting immediately against real-time misogyny and taking a lot of time to delete the hateful data[2]. The researchers began to concentrate on this problem and are building techniques to detect misogyny in real time (Fersini et al., 2018; Hewitt et al., 2016; Poland, 2016). Real-time HS about groups of people like asylum searchers and visitors is common all over the world, but it is rarely investigated.

[1] http://www.pewinternet.org/2017/07/11/online-harassment-2017/ Last visited: 01/02/2019

[2] https://www.telegraph.co.uk/news/2017/08/21/twitter-failing-women-taking-long-remove-misogynistic-abuse/ Last visited: 01/02/2019

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 382–386
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

In this article, we worked on the detection of (A) Hate speech against immigrants and women, (B) Aggressive behavior and target classification, both for English and Spanish languages at Hateval 2019. For this task, we submitted two systems with names: (i) CIC-1 and (ii) CIC-2. We used the bag-of-words model (plus stemming) with TF and TF-IDF as feature values and then we classified these vectors using various machine learning classifiers. We submitted two approaches (systems). Subtask A is ranked by macro-F1 score, whereas subtask B is ranked by EMR score. Our system CIC-1 got 2nd rank in subtask B for the both English (2nd out of 42 teams) and Spanish (2nd out of 25 teams) languages with EMR score of 0.568 for English; 0.675 for Spanish (accuracy score of 0.766 for English; 0.787 for Spanish). The second system, CIC-2 was ranked 4th (out of 39 teams) in subtask A and 1st (out of 23 teams) in subtask B for Spanish language with a macro-F1 score of 0.727 and EMR score of 0.705 respectively (accuracy score of 0.727 in subtask A; 0.791 for subtask B).

2 Related work

A wide range of work has been devoted to HS detection. Xu et al. (2012) applied sentiment analysis to classify bullying in tweets with the usage of Latent Dirichlet Allocation (LDA) topic models (Blei et al., 2003) to recognize related topics in these scripts.

HS detection has been improved by a diverse range of features such as n-grams (Nobata et al., 2016), character n-grams (Mehdad and Tetreault, 2016), paragraph embeddings (Nobata et al., 2016; Djuric et al., 2015) and average word embeddings. (Silva et al., 2016) proposed to detect target groups regarding their class and background on Twitter by looking for sentence structures like "I <intensity> hate <targeted group>".

Currently, interest is increasing in the identification of HS against women on the web (Ging et al., 2018). Initially, Hewitt (2016) worked on the identification of HS against women in social media. Fox (2015) observed that the reaction on hated contents posted against women by unknown and know accounts is different. In (Fox et al., 2015), the authors study the roles of anonymity and interactivity in response to sexist content posted on social media. They inferred that content from unknown account advances more prominent threatening sexism than the known ones.

3 Corpora and task description

Multilingual detection of hate speech on Twitter shared task at Hateval 2019 had two datasets for the English and Spanish language. We participated in both subtasks for both languages.

3.1 Corpora

Corpora for the training of the model consist of 9,000 labeled tweets, and the development dataset includes 1,000 unlabeled tweets. The English data set statistics of different labels is given in Table 1 and the Spanish statistics in Table 2. The corpora is manually labeled by different annotators according to three types:

- Hate speech (present vs not-present),
- Target range (whole group vs individual),
- Aggressiveness (present vs not-present).

We describe these types in the following section.

Type	Labels	Train	Dev
Hate Speech	Present (0)	2631	278
	Absent (1)	1838	222
Target Range	Whole group (0)	3352	363
	Individual (1)	1117	137
Aggressiveness	Present (0)	2984	324
	Absent (1)	1485	176

Table 1: Spanish dataset statistics.

Type	Labels	Train	Dev
Hate Speech	Present (0)	3783	427
	Absent (1)	5217	573
Target Range	Whole group (0)	7659	781
	Individual (1)	1341	219
Aggressiveness	Present (0)	1559	204
	Absent (1)	7441	796

Table 2: English dataset statistics.

3.2 Description of the subtasks

Subtask A: Hate speech detection against immigrants and women: it is a binary classification problem, where it is asked to predict if a specific piece of text (tweet) with a given target (women or immigrants) expresses hatred or not. The systems are evaluated using standard evaluation metrics, containing accuracy, precision, recall, and macro-F1 score. The submissions are ranked by macro-F1 score.

Subtask B: aggressive behavior and target classification: it is required to identify hatred text (tweet) (e.g., tweets, where there is HS against

Team	Task	Classifier	English	Rank$_{eng.}$	Spanish	Rank$_{spa.}$
CIC-1	A$_{F1}$	Logistic Regression	0.462	29 of 69	0.703	18 of 39
	B$_{EMR}$	Majority Voting	0.568	**2 of 41**	0.675	**2 of 23**
CIC-2	A$_{F1}$	MultinomialNB	0.494	14 of 69	0.727	**4 of 39**
	B$_{EMR}$	Classifiers Chain	0.314	19 of 41	0.705	**1 of 23**

Table 3: Our results of Hateval 2019 shared task with ranking for both subtasks A and B.

women or immigrants were marked before) as aggressive or not, and on the second place to recognize a harassing target, either the text (tweet) is against an individual or a group. The evaluation of subtask B was carried out using a partial match and exact match (Basile et al., 2019). The submissions are ranked by EMR score. A tweet must be identified exclusively in one of the following types:

1. **Hateful:** an expression with feelings of dislike, very unpleasant or filled with hatred.
2. **Target Range:** the tweet contains offensive messages intentionally sent to a particular individual or to a group.
3. **Aggressiveness:** it is based on the person's purpose to be aggressive, damaging, or even to provoke.

3.3 Baselines

The Hateval 2019 has set up two following baselines:

- **SVC baseline:** the SVC baseline is a linear Support Vector Machine (SVM) based on TF-IDF representation.
- **MFC baseline:** The MFC baseline is a trivial model that assigns the most frequent label (estimated on the training set) to all the instances in the test set.

4 Description of our approach

In this section, we describe the two submitted approaches (systems) considering the features and machines learning models used for this shared task.

4.1 Pre-processing

We performed pre-processing on raw tweets before feature extraction. Pre-processing helps in these kind of tasks (Markov et al., 2017). For both approaches we used stemming and stop words removal. In CIC-2 we additionally made the following steps:

- we removed HTML tags,
- punctuation marks are removed,
- special characters are removed, like “&”, “$”, “_”, “, ”, etc.

4.2 Features

The pre-processed text was used to generate the features for the machine learning (ML) algorithms. We used a well-known bag of words model, for example, (Sidorov, 2013; Sidorov, 2019). For the first system, we used TF and for the second system TF-IDF values.

4.3 Machine learning algorithms

In our two systems, we used four different classifiers for both subtasks A and B. In CIC-1: Subtask A: Logistic regression, subtask B: Majority voting. In CIC-2: Subtask A: Multinomial Naive Bayes, subtask B: Classifier chains. For all classifiers, we used available implementation in scikit-learn[3].

5 Results and analysis

Results of our both systems CIC-1 and CIC-2 are presented in Table 3, for both shared subtasks, i.e., A and B with our rank in Hateval 2019 competition. Table 3, subtask A ranked by macro-F1 and B by EMR, we used the following conventions. In the first column, “Team” refers to both different systems (CIC-1 and CIC-2) submitted for the shared task. “Task” represents two different subtasks A and B (A$_{F1}$ means that scores of the subtask A are ranked by macro-F1 and B$_{EMR}$ means that scores of the subtask B are ranked by EMR), see section 3.2. “Classifier” states different classifiers, which we used in this competition. “English” and “Spanish” indicate scores for English and Spanish respectively. “Rank$_{eng.}$” and “Rank$_{spa.}$” mean our team's rank in the competition in both subtasks.

The system CIC-1 got 2nd rank in subtask B for the both English and Spanish languages with EMR score of 0.568 for English; 0.675 for Spanish. We used majority voting classifier for both languages.

[3] https://scikit-learn.org/sFigure/ Last visited: 01/04/2019

User name	Macro-F1	Acc.	Rank
saradhix	0.651	0.653	1
Panaetius	0.571	0.572	2
YunxiaDing	0.546	0.560	3
alonzorz	0.535	0.558	4
amontejo	0.519	0.535	5
:	:	:	:
:	:	:	:
hammad.fahim57	**0.494**	0.523	**14**
:	:	:	:
:	:	:	:
Iqraameer133	**0.462**	0.505	**29...**
SVC baseline	0.451	0.492	-
MFC baseline	0.367	0.579	-

Table 4: Results for Subtask A - English.

User name	EMR	Acc.	Rank
MFC baseline	0.580	0.802	-
ninab	0.570	0.802	1
iqraameer133	**0.568**	0.766	**2**
scmhl5	0.483	0.770	3
garain	0.482	0.763	4
gertner	0.399	0.631	5
:	:	:	:
:	:	:	:
hammad.fahim57	**0.314**	0.711	**19...**
SVC baseline	0.308	0.692	-

Table 5: Results for Subtask B - English.

User name	Macro-F1	Acc.	Rank
francolq2	0.730	0.731	1
luiso.vega	0.730	0.734	2
gertner	0.729	0.729	3
hammad.fahim57	**0.727**	0.758	**4**
dibesa	0.725	0.728	5
:	:	:	:
:	:	:	:
Iqraameer133	**0.703**	0.708	**18...**
SVC baseline	0.701	0.705	-
MFC baseline	0.370	0.588	-

Table 6: Top 5 teams for Subtask A - Spanish.

User name	EMR	Acc.	Rank
SVC baseline (as by organizers)	0.771	0.771	-
hammad.fahim57	**0.705**	**0.791**	**1**
MFC baseline	0.704	0.704	-
iqraameer133	**0.675**	0.787	**2**
gertner	0.671	0.758	3
francolq2	0.657	0.749	4
OscarGaribo	0.644	0.732	5...
SVC baseline(Our)		0.550	

Table 7: Top 5 teams for subtask B - Spanish

The system CIC-2 CIC-2 ranked 4th in subtask A and 1st in subtask B for Spanish language with a macro-F1 score of 0.727 and EMR score of 0.705 respectively by using MultinomialNB classifier. It is clear that our system was able to get good results in subtask B (to classify aggressive behavior and target), but was not able to perform well in subtask A (to detect hate speech against immigrants and women) for English language, although we obtained the 2nd position in subtask A for Spanish language. For Spanish subtask B, we tried to reproduce SVM baseline as by organizers but we failed, our SVM baseline gave us 0.550 accuracy.

We made experiments without stop words removal and stemming, and accuracy, in this case, goes down by 2-3%. We discovered that imbalanced data was the main reason for poor performance on English for subtasks A and B. We noticed that most of the submitted systems achieved poor results on the subtask A.

6 Conclusion and future work

In this article, we described our approach to detect (1) Hate Speech Detection against immigrants and women; (2) aggressive behavior and target on the Twitter corpus. We submitted two different systems namely: (i) CIC-1 and (i) CIC-2. We used a bag of words model with TF and TF-IDF values. The vectors are then used as features for classifiers like MultinomialNB, Majority Voting, Logistic Regression, and Classifier Chains. Our CIC-1 system ranked 2nd in task B for both English and Spanish languages. Our system CIC-2 ranked 1st in task B for Spanish and 4th for the same language in task A.

In future work, we can consider embeddings with TF-IDF weighting (Arroyo-Fernández et al., 2019) and learning of document embeddings like in (Gómez-Adorno et al,. 2018). We also plan to consider syntactic n-grams (n-grams obtained by following paths in syntactic dependency trees) (Sidorov 2013; 2019).

We have also made the winning model public[4] for other researchers to use.

Acknowledgement

The is work done with partial support of Mexican government (Conacyt, SNI, COFAA-IPN). We also acknowledge the support of Conacyt project 240844 and SIP-IPN projects 20171813, 20181849, and 20195719.

[4] https://github.com/iqraameer133/HateEval2019 Last visited 13/02/2019

References

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT),* Montreal, Canada, 2012, pp.656-666.

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*, pages 145–153, Geneva, Switzerland.

David M. Blei, Andrew Y. Ng and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3(Jan):993–1022.

Debbie Ging and Eugenia Siapera (eds.). 2018. Special issue on online misogyny. *Feminist Media Studies*, pages 515–524.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18),* Turin, Italy. CEUR.org.

Grigori Sidorov. *Syntactic N-grams in Computational Linguistics*. Springer, 2019, 125 p.

Grigori Sidorov. 2013. *Construcción no lineal de n-gramas en la lingüística computacional* [*Non-linear Construction of N-grams in Computational Linguistics*]. Sociedad Mexicana de Inteligencia Artificial, 166 p.

Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov and David Pinto. 2018. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, pages 1–16.

Ilia Markov, Efstathios Stamatatos and Grigori Sidorov. 2017. Improving Cross-Topic Authorship Attribution: The Role of Pre-Processing. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing* (*CICLing 2017*)*,* Budapest, Hungary. Springer.

Ignacio Arroyo-Fernández, Carlos-Francisco Méndez-Cruz, Gerardo Sierra, Juan-Manuel Torres-Moreno and Grigori Sidorov. 2019. Unsupervised Sentence Representations as Word Information Series: Revisiting TF–IDF. *Computer Speech & Language*, 10.1016/j.csl.2019.01.005.

Jesse Fox, Carlos Cruz, and Ji Young Lee. 2015. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in Human Behavior,* pages 436–442.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabricio Benevenuto and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media. In *Proceedings of the 10th International AAAI Conference on Web and Social Media* (*ICWSM'16*)*,* pages 687–690.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *Proceedings of the 24th International Conference on World Wide Web*.

Sarah Hewitt, Thanassis Tiropanis and Christian Bokhove. 2016. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science* (*WebSci'16*)*,* pages 333–335, ACM.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Rangel Francisco, 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,* Minneapolis, Minnesota, USA.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue,* pages 299–303, Los Angeles, CA, USA.