SINAI at SemEval-2019 Task 6: Incorporating lexicon knowledge into SVM learning to identify and categorize offensive language in social media

Flor Miriam Plaza-del-Arco, M. Dolores Molina-González,
M. Teresa Martín-Valdivia, L. Alfonso Ureña-López

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{fmplaza, mdmolina, maite, laurena}@ujaen.es

Abstract

Offensive language has an impact across society. The use of social media has aggravated this issue among online users, causing suicides in the worst cases. For this reason, it is important to develop systems capable of identifying and detecting offensive language in text automatically. In this paper, we developed a system to classify offensive tweets as part of our participation in SemEval-2019 Task 6: OffensEval. Our main contribution is the integration of lexical features in the classification using the SVM algorithm.

1 Introduction

In recent years, with the emergence of social media, the user-generated content on the Web has grown exponentially. This content has the potential to be transmitted quickly, reaching anywhere in the world in a matter of seconds. Due to the exchange of ideas between users, we find not only positive comments, but also a wide diffusion of aggressive and potentially harmful content. Consequently, this type of remarks affects millions of online users. In fact, it has been reported that these incidents have not only created mental and psychological agony to the online users, but have forced people to deactivate their accounts and, in severe cases like cyberbullying, to commit suicides (Hinduja and Patchin, 2018). One of the strategies used to deal with aggressive behavior in social media is to monitor or report this type of content. However, this strategy is not entirely feasible due to the huge amount of data that is generated daily by users. Therefore, it is necessary to develop systems capable of identifying this type of content on the Web.

In order to tackle this problem, firstly it is important to define the toxic language. The toxic language can be broadly divided into two categories: hate speech and offensive language (Cheng, 2007; Davidson et al., 2017; Gaydhani et al., 2018). According to Cambridge Dictionary, hate speech is defined as "public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation". Offensive language is defined as the text which uses hurtful, derogatory or obscene terms made by one person to another person.

In this paper, we present the system we developed as part of our participation in SemEval-2019 Task 6 OffensEval: Identifying and Categorizing Offensive Language in Social Media) (Zampieri et al., 2019b). In particular, we participated in subtask A: Offensive language identification. It is a binary classification task and consists of identifying if a post contains or not offense or profanity language.

The rest of the paper is structured as follows. In Section 2, we explain the data used in our methods. Section 3 introduces the lexical resources used for this work. Section 4 presents the details of the proposed systems. In Section 5, we discuss the analysis and evaluation results for our system. We conclude in Section 6 with remarks on future work.

2 Data

To run our experiments, we used the English dataset provided by the organizers in SemEval19 Task 6 OffensEval: Identifying and Categorizing Offensive Language in Social Media (Zampieri et al., 2019a).

The datasets contain tweets with five fields. Each tweet comprises an identifier (id), the tweet text (tweet), field for subtask A (subtask_a), field for subtask B (subtask_b) and field for subtask C (subtask_c). Since we have only participated in sub-task A, we are interested in the fields id, tweet

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 735–738
Minneapolis, Minnesota, USA, June 6–7, 2019.

and subtask_a.

In sub-task A, we are interested in the identification of offensive tweets and tweets containing any form of (untargeted) profanity. In this sub-task, there are 2 categories in which the tweet could be classified:

(NOT) Not Offensive - This post does not contain offense or profanity.

(OFF) Offensive - This post contains offensive language or a targeted (veiled or direct) offense. In the annotation, this category includes insults, threats, and posts containing profane language and swear words.

During pre-evaluation period, we trained our models on the train set, and evaluated our different approaches on the trial set. During evaluation period, we trained our models on the train and trial sets, and tested the model on the test set. Table 1 shows the number of tweets per class for English language used in our experiments.

Dataset	NOT	OFF	Total
Train	8840	4400	13,240
Trial	243	77	320
Test	-	-	860

Table 1: Number of tweets per class in OffensEval dataset.

3 Resources

For this subtask A, we used different lexicons that we explain in detail below.

VADER (Valence Aware Dictionary and sEntiment Reasoner) (Gilbert, 2014). The VADER sentiment lexicon is a rule-based sentiment analysis tool. This is sensitive both to the polarity and the intensity of sentiments expressed in social media contexts, and is also generally applicable to sentiment analysis in other domains. VADER has been validated by multiple independent human judges. The tool return four values: positive, negative, neutral and compound. The first three scores represent the proportion of text that falls in these categories. The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized between -1 (most extreme negative) and +1 (most extreme positive).

Offensive/Profane Word List (von Ahn, 2009). A list of 1,384 English terms (unigrams and bigrams) that could be found offensive. The list contains some words that many people won't find offensive, but it's a good start for anybody wanting to detect offensive or profane terms.

4 System Description

In this section, we describe the systems developed for the subtask A in OffensEval task. During our experiments, scikit-learn machine learning in Python library (Pedregosa et al., 2011) was used for benchmarking. A general scheme of the system can be seen in Figure 1.

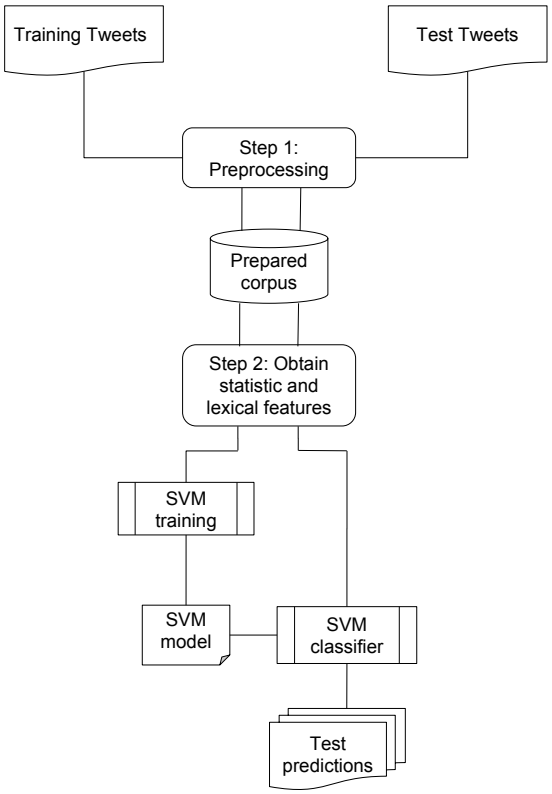

Figure 1: Systems architecture.

4.1 Data Preprocessing

In first place, we preprocessed the corpus of tweets provided by the organizers. We applied the following preprocessing steps: the documents were tokenized using NLTK, the URLs and mentions users are removed and all letters were converted to lower-case.

4.2 Feature Extractor

Converting sentences into feature vectors is a focal task of supervised learning based sentiment analysis. Therefore, the features we chose in our system can be divided into two parts: statistic features and lexical features.

- **Statistic features**. We employed the features that usually perform well in text classification: Term Frequency (TF) taking into account unigrams.

- **Lexical features**. As we explained in Section 3, we used two lexicons to obtain our features in the following way:

  1. **VaderSentiment**. We use the sentiment.vader module[1] provided by the Natural Language Toolkit (NLTK). With this module, we analyze each sentence and we obtained a vector of four scores: negative sentiment, positive sentiment, neutral sentiment and compound polarity.
  2. **Offensive/Profane Word List**. We checked the presence of each word of offensive/profane word list in the tweet and if it exists we assigned 1 as Confidence Value (CV). Then, we summed the CV of all the words finding in the tweet and this value is divided for the total number of words of tweet. As a result, we obtained a parameter that will be used as a feature applied for the classifier.

4.3 Classifier

The concatenation of the features described before are applied for the classification using the SVM algorithm. We selected the Linear SVM formulation, known as C-SVC and the value of the C parameter was 1.0.

5 Analysis of results

During the pre-evaluation phase we carried out several experiments and the best experiment were taken into account for the evaluation phase. The system has been evaluated using the official competition metric, the macro-averaged F1-score. The metric has been computed as follows:

$$\text{Macro-F1} = \frac{2 * \text{Macro-Prec} * \text{Macro-Rec}}{\text{Macro-Prec} + \text{Macro-Rec}} \quad (1)$$

The results of our participation in the subtask A of OffensEval task during the evaluation phase can be seen in Table 2.

[1] https://www.nltk.org/_modules/nltk/sentiment/vader.html

Class	precision	recall	f1-score
NOT	0.81	0.95	0.88
OFF	0.77	0.44	0.56
avg / total	0.8	0.81	**0.79**

Table 2: System test results per class in subtask A of OffensEval task.

User name (ranking)	Macro-F1
pliu19 (1)	0.83
DA-LD-Hildesheim (22)	0.78
fmplaza (68)	**0.72**
gretelliz92 (80)	0.67
AyushS (102)	0.42

Table 3: System Results per participating team in subtask A of OffensEval task.

In relation to our results, it should be noted that we achieve better score in case of the class NOT offensive (F1: 0.88). However, our system is not able to classify well the OFF class (F1: 0.56). This issue may be due to overtraining for the NOT class since as we can see in the Table 1 of Section 2, around 67% of the total tweets belong to that class in the training set in comparison to 33% of OFF class.

With respect to other users, we were ranked in the 68th position as can be seen in Table 3.

6 Conclusions and Future Work

In this paper, we present the system we have developed as part of our participation in SemEval-2019 Task 6: OffensEval: Identifying and Categorizing Offensive Language in Social Media. Specifically, we have participated in subtask A. To solve this task, we have developed a classifier system based on SVM incorporating lexical features from a polarity lexicon and a offensive/profane word list.

Our next study will focus on exploring more features from lexicons because in SemEval-2018 Task 1 (Mohammad et al., 2018), most of the top-performing teams relied on features derived from existing affective lexicons. Also, we will continue working on classifying offensive tweets because today it is a very important task due to the large amount of offensive data generated by users on the Web and we need to prevent the serious consequences it can have on other users.

7 Acknowledgments

This work has been partially supported by Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

Luis von Ahn. 2009. Offensive/profane word list. *Retrieved June*, 24:2018.

J Cheng. 2007. Report: 80 percent of blogs contain offensive content. *Ars Technica*, 2011.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. 2018. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*.

Sameer Hinduja and Justin W Patchin. 2018. Connecting adolescent suicide to the severity of bullying and cyberbullying. *Journal of School Violence*, pages 1–14.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval)*.