IRlab@IIT-BHU at SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media using SVM

Anita Saroj, Supriya Chanda, Sukomal Pal
Department of Computer Science and Engineering
IIT(BHU), Varanasi, INDIA
{anitas.rs.cse16, supriyachanda.rs.cse18, spal.cse}@itbhu.ac.in

Abstract

This paper describes the IRlab@IIT-BHU system for the OffensEval 2020. We take the SVM with TF-IDF features to identify and categorize hate speech and offensive language in social media for two languages. In subtask A, we used a linear SVM classifier to detect abusive content in tweets, achieving a macro F_1 score of 0.779 and 0.718 for Arabic and Greek, respectively.

1 Introduction

Nowadays, the internet is accessed by half of the world population[1]. People from different cultures and educational backgrounds interact on social media. They express their opinions, dissent and even hatred against any individual, or a group using some hate and offensive language. But use of hate and offensive language not only lowers the morale of individuals, but may cause mental agony and trauma. So automatic identification of offensive languages is need of the hour. Almost all social media have policies against use of abusive language, but identification is a challenge. It is not possible to monitor manually or using a static set of rules. Hate speech and offensive language belongs to natural language and therefore NLP techniques can be tried to search for offensive content from textual data. The enthusiasm to explore these sub-fields is to possibly limit the offensive language and hate speech on user-generated content, particularly on social media. Two popular social media platforms for researchers to study are Twitter and Facebook, a social network website where people "tweet" and "message" short posts.

We mainly focus on detecting whether a tweet or message contains offensive content or not (Subtask A). In this paper, a machine learning technique is used to identify the offensive language. Using the Arabic dataset annotated by Mubarak et al.,(2020) and Greek dataset annotated by Pitenis et al.,(2020), we train our Support Vector Machine (SVM) classifier using term frequency-inverse document frequency(TF-IDF) as features. The main aim of this paper is to establish a baseline for this two languages (Arabic and Greek) with traditional machine learning approach.

The rest of this paper is organized as follows. In Section 2, we briefly outline some previous work on offensive language identification. The dataset description, features extraction, and model description are presented in Section 3. Results are presented in section 4, followed by conclusion and future work in Section 5.

1.1 SemEval Sub-Task

According to Zampieri et al.,(2020) the aim of SemEval2020 is to build a system able to detect offensive content. For the first time in SemEval2019, three subtasks were proposed by Zampieri et al.,(2019). We only participate in Sub-task A for Arabic and Greek languages. Some tweet examples from training dataset are listed in Table 2 and Table 3 respectively.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

[1] https://news.itu.int/itu-statistics-leaving-no-one-offline/

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 2012–2016
Barcelona, Spain (Online), December 12, 2020.

- **Sub-task A : Offensive language Identification**
  Sub-task A is coarse-grained binary classification system which classify tweets into two different classes
  - **Non Offensive (NOT)** - This tweet does not contain any offense speech.
  - **Offensive (OFF)** - This tweet contains offense speech.
- **Sub-task B : Categorization of Offensive post**
  Sub-task B is only for English dataset. It is a binary classification task. The posts labeled as OFF in Sub-task A, are further categorised into the following two:
  - **Targeted Insult (TIN):** - These posts contain an insult or threat to an individual or a group.
  - **Untargeted (UNT):** - These posts contain swear words, expression of frustration but not targeted to any person or group.
- **Sub-task C : Offense target identification**
  Sub-task C is a multi-class classification task. After classifying a post as TIN in Sub-task B, it is further classified into following three types:
  - **Individual (IND):** - The post contains offense contents like dehumanizing, insulting an individual or threatening.
  - **Group (GRP):** - This post targeting a group or a member of a group knowing that he or she belongs to that group. Any hateful comments because of political opinion, sexual orientation, gender, social status, health condition or similar.
  - **Other (OTH):** - If any post does not belong to any of the previous two categories then it will be of OTH class.

2 Related Work

Many organizations from academia and industry regularly host workshops and seminars such as TA-COS[2], Abusive Language Online(ALW)[3], TRAC[4] by Kumar et al.(2018a). Recently two surveys by Schmitdt and Wiegand (2017) and Fortuna and Nunes (2018) show how to tag an offensive tweet from features like sentiment, linguistic features, different lexical resources. Task 6 of SemEval 2019 by Zampieri et al.,(2019) was one such effort which helps to solve the three subtasks mentioned in Sec 1.1 for twitter data (only English language).

Most of the previous works are based on English, but some work also reported in other languages like Arabic by Mubarak et al.,(2017), German by Ross et al.,(2017), Slovene by Fiser et al.,(2017)) and Chinese by Su et al.,(2017). Mubarak et al.,(2017) created a set of obscene words as a seeding list called SeedWords (SW) and computed Log Odds Ratio (LOR) for each word unigram and bigram. Like SemEval, GermEval was organized by Wiegand et al., (2018). Not only monolingual, but some attempted a more challenging case of code mixed language like Hinglish by Mathur et al.,(2018). HatEval[5] also organised to detect hate speech in Twitter. In this competition, they fixed two specific different targets, immigrants and women in both languages, Spanish and English. HASOC[6] is a shared task on Hate Speech and Offensive Content Identification in Indo-European Languages(English, Hindi, and German).

3 Data and Methodology

The subtasks in the shared task are different from each other. In Sub-task A, the aim is to identify offensive tweets. Sub-task A requires sensitivity to meticulous changes in the meaning of the word, while

[2] http://ta-cos.org/
[3] https://sites.google.com/view/alw3/
[4] https://sites.google.com/view/trac1/home
[5] https://competitions.codalab.org/competitions/19935
[6] https://hasocfire.github.io/hasoc/2019/index.html

Sub-task B is more pronounced. However, both suffer from data sparsity. The earlier work on hate speech and offensive content identification depended on feature engineering. The handmade feature extraction requires knowledge about language, which is a difficult task for a person who is new to the language. Our classification method depends upon TF-IDF features. Figure 1 shows the methodology of the paper.

Figure 1: Block Diagram of Hate Speech and Offensive Classifier

3.1 Dataset

We have used a training dataset which contains 8000 tweets for Arabic language (2020) and 8743 tweets for Greek language (2020) provided by SemEval 2020 Task 12. The statistics of training and test data corpus collection and class distribution is shown in Table 1.

Data	Language	NOT	OFF	TOTAL
Train	Arabic	6416	1584	8000
	Greek	6257	2486	8743
Test	Arabic	1598	402	2000
	Greek	1302	242	1544

Table 1: Training and Testing Data set Collection and Class Distribution

Sample tweet from the class	SubTask A
السَلامُ تذاكر يا هانم ..... تذاكر يا بيه ■ بالكلمات ■ باداء خيالي ل عبير الشاذلي اوغى فا	NOT
غة االوصيغ: كل حركة ،كل همسة ،كل فكرة،كل نية ،كل كلمة ،كل لحظة حتتسجل ف كتاب و محسوبة يا علينا يا لنا .. لازم نستوعب انه ورانا الكثي	OFF

Table 2: Example tweets from the Arabic training dataset for both classes

3.2 Data Preprocessing

We perform the following pre-processing operations on the dataset.

- All Twitter data are cleaned using tweet pre-processing library[7].
- All Retweet Symbols (RT), Hashtag, URL's, Twitter Mentions, Smileys and Emoji's are removed from the tweets.
- All stopwords are excluded using NLTK[8] library and we apply tokenization.

3.3 Feature Selection

Feature extraction plays an essential role in any classification task. The TF-IDF is used to extract the features from the pre-processed data. It is a weighting scheme based on the count of terms that are present in every post with the terms present in the entire corpus. We have chosen the feature-weight because it extracts the most descriptive terms from the post.

$$tfidf(d,t) = tf(t) * idf(d,t)$$

[7] https://pypi.org/project/tweet-preprocessor/
[8] https://www.nltk.org

Sample tweet from the class	SubTask A
Οι γυναίκες χρειάζονται περισσότερο ύπνο γιατί έχουν πιο σύνθετο εγκέφαλο	NOT
συγνωμη ρε μαλακες ειναι τα ροδακινα στην κοζανη αμαζωχτα κι ακομα μαλακιζεστε εδω μεσα	OFF

Table 3: Example tweets from the Greek training dataset for both classes

3.4 Model

Sub-task A requires the classifier to identify offensive language or not. As training dataset is imbalanced, we decided to use linear SVM classifier for binary classification. As mentioned by Fountana et al.(2018) and Kumar et al.(2018b) SVM classifier performs well when features are well selected. As noted by Dinakar et al.,(2011), SVM are a class of powerful methods for classification tasks, involving the construction of hyper-planes that at the largest distance to the nearest training points. Several papers cite support-vector machines as the state of the art methods for textual classification. For SVM classifier we used `Scikit-learn` implementation and data provided by shared task.

4 Results

We trained a linear SVM model on the training set and tested them on the validation set, using `Scikit-learn`[9]. The experimental results were based on tf-idf features. The result of the classifier on validation data is shown in Table 4. The highest F_1 scores are reported by team `hwijeen`(0.852), `ALAMIHamza`(0.9017) for Greek and Arabic language respectively.

Language	System	F_1 score	Ranking
Greek	SVM$_{TF-IDF}$ classifier	0.718	43/53
Arabic	SVM$_{TF-IDF}$ classifier	0.779	32/37

Table 4: The results of Sub-task A

5 Conclusion and future work

In this study, we report our systems for OffensEval Sub-task A. We trained a linear SVM classifier with a TF-IDF feature model to detect offensive tweets and messages. The evaluation result indicates that our system is capable of detecting offensive language and provides a good potential of identifying targets. However, there is room for improvement. In the future, to capture the subtle meaning and to overcome data sparsity, we plan to investigate a combination of selected surface features and pre-trained word embedding.

References

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online*, pages 46–51, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), July.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

[9] `https://scikit-learn.org/stable/`

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018b. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia, July. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *CoRR*, abs/1701.08118.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April. Association for Computational Linguistics.

Hui-Po Su, Zhen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing profanity in Chinese text. In *Proceedings of the First Workshop on Abusive Language Online*, pages 18–24, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018, pages 1 – 10. Austrian Academy of Sciences, Vienna, Austria.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. 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.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.