Automated Detection of Hate Speech towards Woman on Twitter

Havvanur Şahi
Computer Science
Ankara Yıldırım Beyazıt University
Ankara, Turkey
hwnrshn@gmail.com

Yasemin Kılıç
Computer Science
Ankara Yıldırım Beyazıt University
Ankara, Turkey
ysmnkilic93@gmail.com

Rahime Belen Sağlam
Computer Science
Ankara Yıldırım Beyazıt University
Ankara, Turkey
rbsaglam@ybu.edu.tr

Abstract—**Given the steadily growing body of social media content, hate speech towards women is increasing. Such kind of contents have the potential to cause harm and suffering on an individual basis, and they may lead to social tension and disorder beyond cyber space.**

To support the automatic detection of cyber hate online, specifically on Twitter, we build a supervised learning model which is developed to classify cyber hate towards woman on Twitter. Turkish tweets, with a hashtag specific to choice of clothing for women, have been collected and five machine learning based classification algorithms were applied including Support Vector Machines (using polynomial and RBF Kernel), J48, Naive Bayes, Random Forest and Random Tree. Preliminary results showed that hateful contents can be detected with high precision however more sophisticated approaches are necessary to improve recall.

Keywords—Hate speech recognition, machine learning, classification, tf-idf

I. INTRODUCTION

Due to lack of policy and consistent enforcement, both Facebook and Twitter are used for attacks on people based on characteristics like race, ethnicity, gender, and sexual orientation. Even though there is no formal definition of hate speech, there is a consensus that it targets disadvantaged social groups in a manner that is potentially harmful to them [1].

Oksanen et al. [2] investigated the extent of exposure to and victimization by online hate material among young social media users and reported that 67 percent of young social media users, between15 to 18 year olds, had been exposed to cyber hate on Facebook and YouTube where 21 percent of them had become the victims of such material.

In many countries, including United Kingdom, Canada, and France, there are laws prohibiting hate speech [3]. People convicted of using hate speech can often face large fines and even imprisonment [3]. These laws extend to the internet and social media, leading many sites to create their own provisions against hate speech [4].

The importance of the detection relies on the strong connection between hate speech and actual hate crimes [5]. Social media sites face the problem of identifying and censoring problematic posts [6]. However it is an infeasible task due to its scale and social media sites are not able to prevent or censor the all hate speech users.

In 2006,the Association for Progressive Communications (APC), which is an international network of organizations that provide communication infrastructure to groups and individuals who work for human rights launched an international online campaign "Take Back the Tech" [7]. Take Back The Tech is a global campaign that connects the issue of violence against women and information and communications technology (ICT). It aims to raise awareness on the way violence against women is occurring on ICT platforms such as the Internet and mobile phones, and to call for people to use ICT in activism to end violence against women. In 2014, after an analysis of user policies and redress framework of Twitter, Facebook and YouTube, Take Back the Tech launched the campaign "What are you doing about violence against women?" targeting these three intermediaries. In March 2015, they published a report on corporate policies to end technology-related violence against women, which includes a checklist for addressing violence [8]. In 2014, Twitter ran a pilot project for users to report sexist harassment and abuse with the aim of collecting data for analyses to improve responses to harassment on the social platform.

In Turkey, hate speech and hate crime towards woman have attracted more attention and have become part of public discussion after the death of a female university student (Özgecan Aslan) who was raped and killed by a bus driver. Some social media users came up with a charge such as she deserved that rape due to her dressing style and it has been discussed by several users. Today, hate speech towards woman in Turkey is considered a serious problem, and receives attention from politicians, journalists, and academics from all over the country. However to the best of our knowledge there is no study focusing on hate speech towards women in Turkey.

Due to the massive rise of user-generated web content automated techniques are needed that programmatically classify cyber hate. This task is non-trivial since there are several target groups including race, religion, disability, sexual orientation and transgender status. Each target group are exposed to specific hate related terms complicating the task of automated classification.

In the literature different studies have been conducted for different contents and different target groups. In 2004, Greevy [9] focused on racist content in Web pages and proposed a supervised machine learning approach using bag-of-words (BOW) as features. Dinakar et al. [10] also focused on the identification of cyberbullying using a BOW approach. However feature set was enriched with profane words, parts-of-speech and words with negative connotations.

A framework was proposed by Burnap [11]with the aim of detecting online growing tension about a particular event . The study was conducted by collecting tweets published during the event with a specific hashtag. They aimed to classify level of tension (some tension, tension, high tension)

978-1-5386-7893-0/18/$31.00 ©2018 IEEE

where high tension covers hate speech and applied 3 different approaches; tension analysis engine, machine learning algorithms and sentiment analysis. Results were reported to be promising.

In another study, Burnap et al. [3] leveraged identification of 'othering' language characterized by an us-them dichotomy in racist communication for classifying cyber hate based on religious beliefs specifically for identifying anti-muslim sentiment. Examples of othering language were covered in the study that are about distancing particular social groups geographically (e.g. 'send them home'), justifying an expectation of malicious behavior from the group (e.g. 'told you so'), or openly derogatory (e.g. 'Muslim savages'). It has been reported that identifying othering terms and using it as a feature in a machine classifier for cyber hate targeted at specific religious groups yields better results.

In another study, Davidson et. al. focused on separation of hate speech from offensive language which is a challenging task [4]. They used a crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords and applied multi-class classifier to distinguish between these different categories.

In our study, we have collected the tweets with a specific hashtag (#kiyafetimekarisma) and obtained tweets that cover discussions about woman rights and their freedom on dressing styles. Although there are similar studies working on English tweets, to the best of our knowledge this is the first study that works on Turkish texts.

II. Data

Using the Twitter API we searched for tweets containing hashtag #kıyafetimekarışma (hands off my outfit), resulting in 1288 tweets from several users. A hashtag can be considered as the ground-truth cluster label for tweets which are popular words, beginning with a "#" character. They allow users to efficiently search tweets belonging to specific topics of interest at a certain time. The hashtag used in this study, #kıyafetimekarışma, was used for the discussions about the demonstration organized by women's rights activists after reports of women being attacked or harassed for their clothing. Data consisting of Turkish tweets were collected for a two-week window following the start of the demonstration. Four human annotators were employed to label each tweet as hate speech and non-hate speech. We retained tweets for which at least 3 human annotators (75%) agreed on the class it belonged to. Annotators were provided with each tweet and the question: 'is this text offensive or antagonistic towards woman?'. They were presented with three classes - yes, no, undecided. We removed all tweets with less than 75% agreement.

The results of the annotation presented a dataset of 1,288 tweets, with 159 instances of hate speech (12.34% of the annotated sample), and 1,129 instances of non-hateful content (87.66%). In our experiments, we covered 318 tweets 159 of which were labeled as hate speech and 159 of which were labeled as non-hate speech to have a balanced dataset. We split them into a 75-25 training and test sets.

During the preprocessing phase, we lowercased each tweet, stemmed them using the stemmer algorithm of Zemberek [12] and created bigram, unigram, and trigram which were weighted by their tf-idf scores. The tf-idf is the acronym of "term frequency - inverse document frequency" which is defined as the product of the term frequency $tf_{w,d}$, which is defined as the count of word w in a document d, and the inverse document frequency of word w. Here is the formula:

$$tf_{idf(w,d)} = tf_{w,d} * log \frac{N}{1+N_w} \quad (1)$$

where N is the total number of documents (2, collection of hate speech and non-hate speech tweets in our case), and N_w is the number of documents whose accounts mention the word w at least once.

III. Evaluation

We evaluate the influence of different subsets of features on classification using the library of Weka.

A. Feature

In addition to bigram, unigram, and trigram features, all of which are weighted by their tf-idf values, we use Flesch-Kincaid Grade Level and Flesch Reading Ease scores to capture the quality of the tweets. We used those scores due to the observation that non-hate speech contents tend to be more sophisticated texts whereas hate speech contents are short texts that are not much complex. We also include the features for the number of characters, words, sentences, and syllables in each tweet.

In order to evaluate the performance of the features, we have conducted different experiments with different feature subsets. We also conducted experiments using stemming in order to see the effect of stemming on the classification task.

B. Model

We applied a variety of models that have been used for the classification task in the literature including support vector machines, J48 which is a supervised technique for building a decision tree, Naive Bayes and another decision tree algorithm Random Tree. Random tree is a method for constructing a tree which considers K randomly chosen attributes at each node in order to test the influence of various features on prediction performance. We tested each model using 4-fold cross validation.

We find that the Naive Bayes and Linear SVM tended to perform significantly better than the other models.

IV. Result

We have run several experiments to evaluate the performance of the features and obtained promising results. In the first experiment, we have covered 50 content based features which were top scoring 50 words that have the highest tf*idf values. We have obtained the best results with SVM which reached the desired precision (0.97) however recall value was observed to be low.

Low recall values could be related to the slurs frequently used in the hate speech contents. The models classified the tweets covering those slurs as hate speech. However agnostic tweets without explicit slurs have been observed to be misclassified.

TABLE I. PERFORMANCE OF THE CLASSIFIERS WITH 50 CONTENT BASED FEATURES

	Precision	Recall	Accuracy	F Measure
Naive Bayes	0.56	0.72	% 57.81	0.62
SVM using PolyKernel	0.97	0.32	% 65.31	0.48
Random Tree	0.56	0.52	% 55.31	0.54
J48	0.53	0.46	% 51.56	0.49
Random Forest	1.00	0.40	% 69.69	0.57
SVM using RBF Kernel	0.75	0.52	% 50.63	0.61

With the aim of getting better results we applied stemming using Zemberek and obtained the features accordingly. The results are represented in TABLE II. Stemming significantly increased the precision for three models however it did not yield significantly better recall values.

TABLE II. PERFORMANCE OF THE CLASSIFIERS WITH 50 CONTENT BASED FEATURES OBTAINED USING STEMMING

	Precision	Recall	Accuracy	F Measure
Naive Bayes	1.00	0.50	% 74.69	0.67
SVM using PolyKernel	1.00	0.47	% 73.44	0.64
Random Tree	1.00	0.49	% 74.38	0.66
J48	0.50	1.00	% 50.00	0.67
Random Forest	1.00	0.51	% 75.31	0.67
SVM using RBF Kernel	1.00	0.24	% 61.88	0.39

In the next two experiments, we have expanded the content based features and covered 100 words that have the highest tf*idf values. We have applied stemming in one of the experiments and skipped it in the other.

TABLE III. PERFORMANCE OF THE CLASSIFIERS WITH 100 CONTENT BASED FEATURES

	Precision	Recall	Accuracy	F Measure
Naive Bayes	0.55	0.73	% 57.50	0.63
SVM using PolyKernel	1.00	0.33	% 66.00	0.50
Random Tree	0.60	0.47	% 43.44	0.53
J48	0.63	0.42	%51.88	0.50
Random Forest	1.00	0.41	% 70.63	0.58
SVM using RBF Kernel	0.75	0.58	% 53.75	0.65

TABLE IV. PERFORMANCE OF THE CLASSIFIERS WITH 100 CONTENT BASED FEATURES OBTAINED USING STEMMING

	Precision	Recall	Accuracy	F Measure
Naive Bayes	1.00	0.53	% 75.00	0.69
SVM using PolyKernel	1.00	0.50	% 74.69	0.67
Random Tree	1.00	0.56	% 77.81	0.72
J48	0.50	1.00	% 50.00	0.67
Random Forest	1.00	0.52	% 75.94	0.68
SVM using RBF Kernel	0.63	0.80	% 52.50	0.70

As observed in the tables above, stemming increased the precision values however we could not get the desired recall values in any of the experiments.

Observing the limitation of the content based features on the prediction, we have enriched the dataset with the structural features and rerun the experiments. We applied stemming to obtain content based features since it yielded better precision in the previous experiments. However, the gap between the precision and recall values remained almost the same.

TABLE V. PERFORMANCE OF THE CLASSIFIERS WITH STRUCTURAL FEATURES AND 50 CONTENT BASED FEATURES OBTAINED USING STEMMING

	Precision	Recall	Accuracy	F Measure
Naive Bayes	0.58	0.72	% 59.38	0.64
SVM using PolyKernel	1.00	0.42	% 70.94	0.59
Random Tree	0.59	0.49	% 56.89	0.54
J48	0.48	0.35	% 50.63	0.40
Random Forest	0.69	0.49	% 63.44	0.58
SVM using RBF Kernel	0.58	0.80	%59.69	0.66

TABLE VI. PERFORMANCE OF THE CLASSIFIERS WITH STRUCTURAL FEATURES AND 100 CONTENT BASED FEATURES OBTAINED USING STEMMING

	Precision	Recall	Accuracy	F Measure
Naive Bayes	0.57	0.72	% 58.75	0.64
SVM using PolyKernel	1.00	0.45	% 72.5	0.62
Random Tree	0.65	0.49	% 61.25	0.59
J48	0.51	0.45	% 49.69	0.48
Random Forest	0.70	0.43	% 61.88	0.53
SVM using RBF Kernel	0.67	0.69	% 58.44	0.68

To sum up, tweets with the highest prediction probabilities of being hate speech tend to contain multiple sexual or homophobic slurs which can be learned easily in training phase. In testing phase, tweets that cover those slurs were classified correctly we obtained high precision values in most of the experiments. However, given the variety of words used in agnostic contents, it is a non-trivial task to get high recall values. The success of the study is limited to the offensive words covered in the training set.

On the other hand, this approach leads high rates of false positives since the presence of offensive words can lead to the misclassification of tweets as hate speech. This could be the main reason of low precision values returned by Naïve Bayes.

Another problem occurs due to the vague definition of hate speech. Even though there is no formal definition for hate speech, there is a consensus that it is speech that targets disadvantaged social groups in a manner that is potentially harmful to them [13]. Consequently, we have asked our annotators to label a tweet as hate speech if it covers hateful message towards woman. This approach yielded hateful tweets towards the accounts that post hateful posts towards woman have been labeled as non-hate speech by the annotators. Such kind of tweets are really challenging and hard-to-classify instances which requires detecting the target group in the sentences.

V. CONCLUSION

Given the legal and moral implications of hate speech, it is important to detect and monitor the hateful contents on the Web. In this study, we proposed a machine learning based classification model to detect hate speech towards women on Twitter. Even though the study is conducted on tweets, it is possible to apply it on any textual content. Our experiments revealed that it was possible to detect hate speech with high precision even if we could not increase the recall values as desired.

Consistent with the other works applied on English texts for similar purposes, we find that certain terms are particularly useful for distinguishing hate speech (*r*spu, k*hpe, namussuz etc.). There exist several publicly available lists that consist of general hate-related terms in English however we could not make use of such a list since it does not exist for Turkish. We call for the wider community to explore such a list for Turkish to be used in the future studies.

Many of the tweets labeled as hateful contain sexual and homophobic slurs. This results misclassification for the hateful tweets that do not contain any offensive words. In our future work, we are considering to focus on those hard-to-classify instances and try sentiment analysis to improve results.

REFERENCES

[1] "Hate Crimes: Criminal Law and Identity Politics - James B. Jacobs, Kimberly Potter - Google Books." [Online]. Available: https://books.google.com.tr/books?hl=en&lr=&id=G6V0Qb4GtNUC&oi=fnd&pg=PR9&dq=Jacobs+and+Potter+2000&ots=JVHG0RY2OJ&sig=v_fajEA9khdAvBVKnHEWmfPRHYw&redir_esc=y#v=onepage&q=Jacobs and Potter 2000&f=false. [Accessed: 30-May-2018].

[2] P. Oksanen, Atte and Hawdon, James and Holkeri, Emma and N{"a}si, Matti and R{"a}s{"a}nen, "Exposure to online hate among young social media users," *Soul Soc. a Focus lives Child. \& youth*, vol. 18, pp. 253--273, 2014.

[3] P. Burnap and M. L. Williams, "Us and them: identifying cyber hate on Twitter across multiple protected characteristics," *EPJ Data Sci.*, vol. 5, no. 1, p. 11, Dec. 2016.

[4] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," Mar. 2017.

[5] "Hate Crimes: Consequences of hate speech.," 2014. [Online]. Available: http://www.nohatespeechmovement.org/hate-speech-watch/focus/consequences-of-hate-speech. [Accessed: 30-May-2018].

[6] "No place for hate speech on facebook." .

[7] "Take Back the Tech." [Online]. Available: https://www.takebackthetech.net/research-resources. [Accessed: 25-May-2018].

[8] A. Rima, "From impunity to justice: Improving corporate policies to end technology-related violence against women." [Online]. Available: http://www.genderit.org/sites/default/upload/flow_corporate_policies_formatted_final.pdf.%0D. [Accessed: 15-May-2018].

[9] E. Greevy, A. S.-P. of the 27th annual International, and U. 2004, "Classifying racist texts using a support vector machine," *dl.acm.org*, 2004.

[10] K. Dinakar, B. Jones, C. Havasi, … H. L.-A. T. on, and undefined 2012, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *dl.acm.org*.

[11] P. Burnap *et al.*, "Detecting tension in online communities with computational Twitter analysis," *Technol. Forecast. Soc. Change*, vol. 95, pp. 96–108, Jun. 2015.

[12] "Zemberek," 2018. [Online]. Available: http://zemberekn1p.blogspot.com. [Accessed: 15-May-2018].

[13] P. K. Jacops James B, "No Title," in *Hate Crimes: Criminal Law and Identity Politics*, 2000.