PressAcademia Procedia

2nd World Conference on Technology, Innovation and Entrepreneurship
May 12- 14, 2017, Istanbul, Turkey. Edited by Sefer Şener

AUTOMATIC HATE SPEECH DETECTION IN ONLINE CONTENTS USING LATENT SEMANTIC ANALYSIS

DOI: 10.17261/Pressacademia.2017.612
PAP-WCTIE-V.5-2017(50)-p.368-371

Xhemal Zenuni[1], Jaumin Ajdari[2], Florije Ismaili[3], Bujar Raufi[4]
[1]South East European University. xh.zenuni@seeu.edu.mk
[2]South East European University. j.ajdari@seeu.edu.mk
[3]South East European University. f.ismaili@seeu.edu.mk
[4]South East European University. b.raufi@seeu.edu.mk

ABSTRACT

Internet in general and social media in particular have greatly facilitated the communication, interaction and collaboration among people and different entities. As generally there is no censorship, these media sometimes are used to proliferate discourses that contain hateful messages targeting ethnic origin, religious or sexual groups, which potentially may degenerate to violent acts against individuals of such groups. Therefore, we explore the idea of building of automatic classifier that can be used for detection of hate speech in public Albanian language pages. A hate speech corpus for Albanian language is created, and then based on Support Vector Machine (SVM) approach, an automatic hate speech detection system is proposed. Such system can be used to detect and analyze hate speech in online contents over time and to enhance our knowledge on how they affect opinion creation in society.

Keywords: Hate speech detection, text classification, support vector machines, NLP, Albanian language

1. INTRODUCTION

The continous growth of social media and other Internet services, such as Facebook, Twitter, microblogging or Web services among others has greately facilitated the information exchange, interaction and collaboration among people and different entities. However, the widespread adoption of social media and other online services offers new opportunities to dissiminate hateful messages. Up to date, there is very little research and evidence how the diffusion of hate speech in online contents could trigger hate crimes, yet this potential is recently recognised. For example, Facebook and Twitter pledge to remove hate speech contents within 24 hours after they are reported (Kottasova, 2016). On the other side, EU despite its security and political situations, launched a “code of conduct” to establish public commitments for the biggest Internet companies that the valid hate speech contents will be removed and yet the right to freedom of expression will be preserved (Commision, 2016).

In this context, automatic detection of abusive and hate speech in online contents becomes important topic and task. An automatic detection method could scan large amount of text, analyze and categorize it as hateful or not. The trends of hateful messages could not only be reported to relevant authorities, but it could provide a solid ground to researchers to understand how hateful messages in online contents affect the social processes.

But as noted in (Thomas Davidson, 2017), effective automatic hate speech detection is challenging and very difficult task. The difficulties mostly come from the complexity of natural language processing. The ambiguity and language variability represents a real challenge to be solved. On the other hand, when building more complex and effective automatic machine learning text classifier, the training data becomes crucial.

In this paper, we aim to develope a method to detect hate speech in public online contents in Albanian language, while also addressing the above mentioned challenges. We have collected data from public Facebook pages in Albanian language, and

labeled them as hate speech or not. Than a classifier based on SVM (support vector machines) is trained to differentiate between these categories. To our best knowldge, the contribution in this paper is two-fold:

- It represents the first attempt to create a hate speech corpus in Albanian language
- We make the first attempts to create a hate speech text classifier for Albanian language based on supervised machine learning approach

2. LITERATURE REVIEW

Bag-of-word approaches like in (Kwok & Wang, 2013) are simpler to implement, especially if the classifier is targeting racial hate of speech, but such approaches are insufficieent for accurate classification as it leads to high rates of false positives.

Syntactic features have been explored in (Gitari, Zuping, Damien, & Long, 2015). The experimental results has shown improvements both on precision and recall when used semantic, hate and theme-based features. Chen (Chen, Zhu, Zhou, & Xu, 2012) utilize the profanties, obscenities and pejorative terms as features, weighted accordingly and produced a set of rules to model offensive content, which improved the precision on standard machine learning approaches.

Leveraging morpho-syntactical features, sentiment polarity and word embedding lexicons, Vigna (Vigna, Cimino, Dell'Orlleta, Petrocchi, & Tesconi, 2017) proposed two hate speech classifiers for Italian language based on Support Vector Machines (SVM) and on Reccurent Neural Network named Long Short Term Memory.

Other suppervised approaches to hate speech classification have been proposed as well. Neural language models have potential (Djuric, Zhou, & Morris, 2015), but in all cases the training set data is important. Moreover, the accuracy of hate speech classsifiers could be improved by non-linguistic features, like the gender, ethnicity or age of the author, but this information is often unreliabale or unavailable (Waseem & Hovy, 2016).

3. HATE SPEECH CORPUS

To our best knowledge, there is no previous work on building a hate speech corpus for Albanian language. Therefore, during a period of time, we collected data from Facebook pages in Albanian language and prepared a hate speech corpus that could be used by a classifier. This section reports on data collection, annotatin phase, preprocessing and feature selection in data.

3.1. Data Collection

We explored the Graph API (https://developers.facebook.com/docs/graph-api) provided from Facebook to reterieve and build a corpus of comments from two public pages that publish posts on variety of topics on different political and social events, and which we suspected to find a lot of comments contating hateful speeches. On the other side, we also looked forward posts that contained a significant number of comments. Table 1 summarizes the pages that where crawled and the number annotated posts and comments

Table 1: Dataset Description and Annotations

Facebook pages	Annotated Post	# of Commnets
jetaoshqef	108	4737
tvklan	19	149

3.2. Data Annotation

Two annotators were asked to analyze the content of the crawled comments and to categorize them as *hate* or *no hate*. Overall, 4886 comments received two annotations, and as *hate* were considered only the comments that were categorized as such from both annotators. In total, 2764 comments were categorized as containing hateful content and on other comments either both annotators agreed that the message is not hateful or no concensus was reached.

3.3 Data Preprocesing

In order to prepare the data for the supervides learning algorithm, several pre – processing steps were undertaken. First, the collected text was transformed to lowercase with the objective to improve syntactic matching. Then, extra white spaces, puncuation marks, digits and emoji were removed from the text as they were not considered important in the classifiction process. Finally, we removed from the text the words which we find redundant for text classification (such as conjuctions) and consequently reduced the size of document-term matrix.

4. TEXT CLASSIFICATION MODEL

We tested the Support Vector Machines (SVM) as supervised learning technique used for text classification. As algorithm it captures sparse and discrete features in text classification, which makes it good candidate in our case. On the other side, as noted in (Joachims, 1998) there are theoretical evidence that SVM is an extremely strong performer when having high dimensional input space, few irrelevant features and especilly when most of text classification problems are linerely separable.

We implemented the approach in R System (http://www.rsystems.com/) based on RTextTool. RtextTool is an easy to use tool that can be used for end-to-end implementation by interfacing with existing pre-processing routines and machine learning algorithms. The supporting features include the process from document-term matrix creation, data pre – porcessing, training, classification, up to analytical reports which help users to understand the classification of the employed model.

The speech corpus was divided in two parts. 4000 records were used as training set, and the rest of 886 records were used as testing set. And than based on this dataset, RTextTool functons were used to implement the text classification workflow.

5. FINDINGS AND DISCUSSIONS

While there are many techniques to evaluate the performance of the algorithm, precission, recall and F-score are considered standard evaluation metrics in classification tasks. Accuracy measures In the context of the hate speech system, the accuracy tells what propotion of hate speech comments, are actually hate speech content. Recall tells what percentage of hate speech comments did the algorithm correctly classify, and F-score produces a weighted average of precision and recall.

Table 2 reports the results for the conducted experiement. And the numbers were generated through *create_analytics( )* function contained in RTextTool.

Table 2: Evaluation of Classification Model

Classifier	Precission	Recall	F-Score
SVM	.61	0.57	0.58

6. CONCLUSION

This paper presents the first efforts in bulding an automated hate speech classifier for Albanian language texts. The first experiments show that binary classificaion based on Suport Vector Machines are a promising approach toward building an automated hate speech detection system for online text contents for Albanian language. We are encouraged by initial results, however for the hate classfier of Albanian language to achieve results comparable with similar approaches, it needs richer hate speech corpus and and to explore other language processing features of Albanian language, which for the time being are lacking. However, we believe this work represents the basis toward a building an automated system that could be used to track and monitor online content.

As future work, we intend to extend the annotaed hate speech corpus from different Facebook sites and crawl more comments. This will make richer the current training set, which we believe it will consequently increase the evaluation metrics employed in standard classification tasks. Another important aspect will be to see how other similar supervised learning models will work under the same speech training set.

REFERENCES

Chen, Y., Zhu, S., Zhou, Y., & Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. *Proceedings of the Fourth ASE/IEEE International Conference on Social Computing.* Amsterdam.

Commision, E. (2016). *CODE OF CONDUCT ON COUNTERING ILLEGAL HATE SPEECH ONLINE.*

Djuric, N., Zhou, J., & Morris, R. (2015). Hate Speech Detection with Comment Embeddings. *Proceedings of the 24th International Conference on World Wide Web*, (s. 29-30).

Gitari, N., Zuping, Z., Damien, H., & Long, J. (2015). A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, 2015-230.

(tarih yok). *http://www.rsystems.com/.*

https://developers.facebook.com/docs/graph-api. (tarih yok).

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *European Conference on Machine Learning*, (s. 137-142).

Kottasova, I. (2016). *Facebook and Twitter pledge to remove hate speech within 24 hours.* http://money.cnn.com/2016/05/31/technology/hate-speech-facebook-twitter-eu/.

Kwok, I., & Wang, Y. (2013). Locate the Hate: Detecting Tweets against Blacks. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, (s. 1621-1622).

Thomas Davidson, D. W. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *In the Proceedings of ICWSM 2017.*

Vigna, D. V., Cimino, A., Dell'Orlleta, F., Petrocchi, M., & Tesconi, M. (2017). Hate Me, Hate Me Not: Hate Speech Detection on Facebook. ITASEC.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (s. 88-93).