TAR System Description Paper Template

Stjepanović Mateo, Žabčić Mislav, Tolić Filip

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
mateo.stjepanovic@fer.hr, {mislav.zabcic,filip.tolic}@fer.hr

Abstract

This document provides the instructions on formatting the TAR system description paper in LaTeX. This is where you write the abstract (i.e., summary) of the work you carried out within the project. The abstract is a paragraph of text ranging between 70 and 150 words.

1. Introduction

This section is the introduction to your paper. Introduction should not be too elaborate, as that is what other sections are for (the Introduction should definitely not spill over to the second page).

This is the second paragraph of the introduction. In LaTeX, paragraphs are separated by inserting an empty line in between them. Avoid very large paragraphs (larger than half of the page height), but also avoid tiny paragraphs (e.g., one-sentence paragraphs).

2. Related work

Hate speech is great problem of today. Not many works has been done on this subject yet. From that reason we don't have dataset of set of features for hate speech. It is important to separate hate speech from offensive language, even so because of new laws against hate speech(Davidson et al.2017). Papers so far shows that best approach to given problem is using Support Vector Machine (SVM) and Bag-of-Words(BoW).(Davidson et al.2017).
Researchers tried to create topology for abusive language, which could help to specify features for identifying abusive language.(Waseem et al.2017) stated that there is four types of abusive language that one should take into account.

- *Directed Implicit*
- *Generalized Implicit*
- *Directed Explicit*
- *Generalized Explicit*

In (Davidson et al.2017) they used several features to capture information about syntactic and semantic structures. Porter stemmer is used to create unigram, bigram and trigram features and then used TF-IDF to put weights to them. Using NLTK they constructed Part of Speech tags as features. They showed that their model is working great, and got pretty high official metrics (precision, recall and F1 score).

(Chen and Lin2006) shows that one can implement feature selection methods into SVM it self. This could possibly make great future work in trying to achieve end-to-end solution for hate speech identification problem, as we encountered problem of high number of features, and high CPU and RAM requirement as result of that.
Giving the problem of difficulty to get relevant data and to manually annotate enough of them, active learning is developed. (Luo et al.2005) proposed new active learning model based on multi-class support vector machines. They showed that there is a way to make SVM work with probabilities and give us proper active learning model. Even though we didn't implement this model, plan is to implement it in the future and test it on state-of-the-art models. Their algorithm works as follows:

- 1. Start with an initial training set and an unclassified set.
- 2. A multi-class support vector machine is built using the current training set.
- 3. Compute the probabilistic outputs of the classification results for each data on the unclassified set. Suppose the class with highest probability is a and the class with second highest probability is b. Record the value of P(a) and P(b) for each unclassified data.
- 4. Remove the data from the unclassified set that have the smallest difference in probabilities between them (P(a) P(b)) for the two highest probability classes, obtain the correct label from human experts and add the labeled data to the current training set.
- 5. Go to 2

In (Yang et al.2009) authors introduced their own active learning model based on multi-class SVMs. It is little different from one before on manner that they calculate probability on each data in unlabeled pool. Given that model is computational expensive if unlabeled pool in large size. Their experiment also show that they outperform state-of-the-art models, which can only be a sign to continue further on out work to get end-to-end solution for hate speech identification.

References

Yi-Wei Chen and Chih-Jen Lin, 2006. *Combining SVMs with Various Feature Selection Strategies*, pages 315–324. Springer Berlin Heidelberg.

Thomas J Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Tong Luo, Kurt Kramer, Dmitry B. Goldgof, Lawrence O. Hall, Scott Samson, Andrew Remsen, and Thomas Hopkins. 2005. Active learning to recognize multiple types of plankton. *Journal of Machine Learning Research*, pages 589–613, April.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84. Association for Computational Linguistics.

Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. 2009. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 917–926. ACM.