CE807 – Assignment 1 - Interim Practical Text Analytics and Report

Anonymous ACL submission

Abstract

This document puts a bit of light on different methods avail be for text classification and hate speech recognition and differentiates between them. The aim of this report is to have a better understanding of these methods and find out there advantages and disadvantages. This report also introduces the OLID data-set general view. The main purpose of this report is strictly study based and the opinions are not from an expert.

1 Generic text Classification(Task 1)

Text classification is a necessity in today's research due to high number of electronic documents that need to be categorized and organized throughout the internet. There have been many different methods used for test classification. Every approach has its own advantages and disadvantages as well. They are divided into two major sections, Statistical and Machine learning and in ML into sub two more sections, Supervised and unsupervised.

There has been different research done on the effectiveness of stemming and stop words removal. These both methods have improved the efficiency of the system Uysal and Gunal (2014). Other well known techniques used are tokenization, lower case conversion. One of the other approaches that was really interesting was the hierarchical classification using top down classification algorithm Sun and Lim (2001). Generally naïve bayes classifier is not the best algorithm for this problem as it is not able to cope with text classification due to parameter estimation process, but attribute weighing methods and document normalization has ben proven to improve the effectiveness of the model Kim et al. (2006). The proposed system called Poisson naïve bayes classifier has performed subsequently better if compared to original model.

Statistical approach has also been implemented on some scenarios; it is quite hard approach with comparison to other methods as it is highly based on the mathematical calculations. According to research done on Arabic language, it's been proven that the semi-automatic or ML based algorithms work better than the normal Statistical approach Fodil et al. (2014). Another interesting method used for test classification is the use of the capsules, usually used for image classification but also very effective for text classification [5]Kim et al. (2020). Capsules work like a group to learn patterns and encode properties in form of vectors.

Rocchio Classification is another algorithm used for text classification. It finds centroid from the training set and it classifies it to the nearest classified text documentKim et al. (2020). There is another improved version if Rocchio Classification called Hierarchical Rocchio Vijayan et al. (2017) which is not only able to generate new classes much quicker as compared to original model but can also get relationship hierarchy between different classes. SVM(Support Vector Machine) are used for high dimensional Classification problems. They are generally good with text classification and there are many implementations of SVM on this particular problem Zeng and Huang (2012). The SVM needs to be trained using predefined labeled data.

1.1 Critical Discussion (Task 1)

Although Hierarchical classification has proven to be good in terms of accuracy and also uses misclassified samples for classification, the drawback is the amount of training samples needed for model to be trained and resources. The Poisson naïve bayes classifier has good accuracy but fails to outperform the traditional SVM. On the other hand, the major problem with statistical methods is that they are mathematical based calculations which makes them harder to implement for more complex problem solving.

The use of capsules for text classification is somewhat of a new idea even though it is quite effective, there needs to be lots of work and research on this method as it's relatively new. In the case of Rocchio Classification, everything is dependent on the centroid of the class and the algorithm fails when centroid is not able to represent the class. SVMs are the most frequently used for text classification but need labeling of data samples in the training stage.

2 Offensive Language Detection Methods (Task 2)

With the ease of access of social media, it has become much more challenging to control hate speech crimes as anyone can say whatever they want without thinking of any cause and effectWilliam et al. (2022). This is somewhat of a sensitive issue, and we can implement ML methods to overcome the problem. SVM is the most widely used algorithm giving around 79 percent of the overall accuracy Pawar et al. (2022). The results were much easier to comprehend as compared to implication of other methods which tend to be more complicated Biere et al. (2018). The annotation of hate speech dataset needs to be done manually.

NLP has many implementations and it is of much more importance in this case as we are looking for specific sensitive words in the text on order to look for hate speech. Another good way is using a method with TF-IDF which has proven to give good results around 92 percent which is remarkable Sachdeva et al. (2021). Other way is using the Ensemble based learning method like Random Forest and Logistic Regression which have proven to do well on there own but CNN models work well on pretrained embedded models Martins et al. (2018).

The last and personally a totally different and unique approach was using Emotion models for detecting the emotions in the actual text itself in order to accurately predict what other person means. It gives more control over things as compare to looking for certain words Zampieri et al. (2019). It could increase the accuracy of the actual hate speech recognition by double. Discrete emotional models group words and make decision in terms of note individual but whole group evaluation, in order to differentiate between actual context or hate speech.

2.1 Critical Discussion (Task 2)

The most difficult part of hate speech recognizing is defining it. As there is no one practical definition of hate speech and it varies from person to person according to their ideologies and beliefs, which are also strongly affected by religious views as well. The only major difference between the generally used classification method and the TF-IDF is the time complexity.

Even though the use of emotional analysis is a really unique idea, it is restricted by the amount of training data and language barrier itself. Overall Hate speech recognition is much more complex problem to handle as compared to text classification. It is dependent on too many variables and needs constant updating. The models now available are generic and need continuous update of training in order to cope with the problem. Nevertheless, hate speech recognition is getting better but still needs a lot of research.

3 Data-set (Task 3)

The dataset used in this research paper is the OLID dataset (Zampieri et al., 2019)

- Who: OLID is created by Zampieri et al. on year 2019 and is a publicly available dataset

- What: It is a dataset which contains 14,200 tweets which are Annotated in English.

- Where: It is available on the OffensEval website.

- Why: The purpose of the dataset was to able to find offensive words in the language and it's been used in many competitions and is from reliable source.

- When: The first training set release was on 28th of November 2018 and first full dataset was available by 20th April 2019

4 Summary (Task 4)

For text classification undoubtedly the SVM models are state of the art and there are so many variations and combinations of them available for text classification problem. But at the end of the day not all models are generic.For example, Statistical Methods outperform others on certain problems but lack when there is a learning element to it. Otherwise they have better accuracy in average. On the other hand the ML methods have proven to be much more easy and longer term solution for these kind of problems.

For Hate speech recognition, it is subjective to just decide any one algorithm as it is too complicated. Every now and then something finds it's way through the system. There's lack of training data to specifically train the model to figure out exactly if particular text is a hate speech or not but the Discrete Emotional model is a good approach and a way step forward towards AI with critical thinking. The application of the method is a bit more complicated but at the same time it is much more effective in terms o recognizing the hate speech pattern.

For text classification using SVM the only problem that's faced is manually labeling of the data-set. But we can use a pre-trained language annotated model that is available for example by google which might make tings much more easier. The performance of SVM overall is really good. And in terms of Hate Speech recognition we still need alot of research and proper definition of Hate speech words incorporated in the Discreet Emotional Model which can subsequently increase the accuracy but still figuring out all hate speech patterns is going to be chlallenig but in terms of accuracy this model has huge potential.

References

Shanita Biere, Sandjai Bhulai, and Master Business Analytics. 2018. Hate speech detection using natural language processing techniques. *Master Business AnalyticsDepartment of Mathematics Faculty of Science*.

Leila Fodil, Halim Sayoud, and Siham Ouamour. 2014. Theme classification of arabic text: A statistical approach. In *Terminology and Knowledge Engineering 2014*, pages 10–p.

Jaeyoung Kim, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020. Text classification using capsules. *Neurocomputing*, 376:214–221.

Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng. 2006. Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering*, 18(11):1457–1466.

Ricardo Martins, Marco Gomes, Jose Joao Almeida, Paulo Novais, and Pedro Henriques. 2018. Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 61–66. IEEE.

AB Pawar, Pranav Gawali, Mangesh Gite, MA Jawale, and P William. 2022. Challenges for hate speech recognition system: Approach based on solution. In *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pages 699–704. IEEE.

Janak Sachdeva, Kushank Kumar Chaudhary, Harshit Madaan, and Priyanka Meel. 2021. Text based hate-speech analysis. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 661–668. IEEE.

Aixin Sun and Ee-Peng Lim. 2001. Hierarchical text classification and evaluation. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 521–528. IEEE.

Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information processing & management*, 50(1):104–112.

Vikas K Vijayan, KR Bindu, and Latha Parameswaran. 2017. A comprehensive study of text classification algorithms. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1109–1113. IEEE.

P William, Ritik Gade, Rup esh Chaudhari, AB Pawar, and MA Jawale. 2022. Machine learning based automatic hate speech recognition system. In *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pages 315–318. IEEE.

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.

Anping Zeng and Yongping Huang. 2012. A text classification algorithm based on rocchio and hierarchical clustering. In *Advanced Intelligent Computing: 7th International Conference, ICIC 2011, Zhengzhou, China, August 11-14, 2011. Revised Selected Papers 7*, pages 432–439. Springer.