IMDB SENTIMENT ANALYSIS

LUO, ROBIN (260851506), ROUSSEAU, MARC-ANDRE (260089646), AND XU BIDE, (260711367)

Abstract. We used ML techniques including logistic regression, naive Baye's and support vector machines to process the data from a large dataset of reviews and train our software to be able to distinguish a favourable review from a negatve one. In addition to the three methods listed, we used TF-IDF, where the frequency of occurrence of the words within the document and the dataset are both taken into account to ascertain word importance. Our model was submitted to a kaggle competition and we obtained a score of 0.91786 for the model using a linear SVM, tf-idf. This is a significant improvement on the Naive Bayes classification method which has an accuracy of approximately 0.83

1. Introduction

(5+sentences) The ubiquity of social networks is no longer a budding phenomenon, it is the reality of the world in which we live. Many popular sites have included ways for users to share their opinions on a variety of topics and therefore the ability to mine through these posts and determine how users feel about the things being discussed is extremely useful for businesses. Knowing that a user or group of users desire something or find it appealing creates a market of opportunity for companies in search of low risk opportunities to expand their business operations. In addition, the analyses performed, once properly summarized and visualized effectively have tremendous value in themselves. For our project, we were given 12500 positive and 12500 negative reviews to train our algorithm and another 25000 to test our code and submit our best guess as to the correct labeling of the test reviews as either positive or negative. Our best model which used a linear SVM (0.918 accuracy), was much better than Naive Bayes (0.83).

2. Related Work

(4+sentences) Machine learning and sentiment analysis are hot topics of research with many machine learning conferences having several talks on the topic. For example, Twitter has been releasing datasets to be mined for things like whether a piece of task is positive, negative or neutral. Recently, a group of researchers extended this problem to five classification categories and added arabic language content (Rosenthal, 2017). Many teams submitted ML proposals to classify the twitter posts and the top performing groups used deep neural networks (DNNs) (Rosenthal, 2017). In addition, out of the top 10 submissions, the second most successful approach to DNNs involved the use of SVMs which is consistent with our best performing model for this project. A more directly related work involved taking into account sentence negations (Das, 2018). In this paper, the authors have decided to use a shortcut for negation by negating the word immediately following a negation. One example of this method would be to take the sentence "I am not happy" which gets converted to "I am not_happy". The benefit of this is that by changing only one word, they are able to change the meaning of the entire sentence (Das, 2018).

3. Dataset and Setup

(3+sentences) Our training dataset consisted of a list of 25000 reviews which we began by tokenizing followed by some cleaning where unicode characters which were not relevant to analysis were mapped to the empty string.... (what else?)

4. Proposed Approach

(7+sentences) Here we need a description of the full model (should be in the image that Peter is planning to do).

- Discussion of algorithm selection - why did we end up using SVM? need plots to justify.
- Splitting of Data into validation/training, etc...
- Regularization strategies (did we use any? if so, what effect did it have)
- Did we use any optimization tricks?
- We need plots to justify hyperparameter selection.
- Background/motivation for each model (i can do this, but I need to know everything that was done)

Date: February 19, 2019.
1991 *Mathematics Subject Classification.* Comp 551.

5. Results

(7+sentences) This section needs to be filled with some of the plots needed to demonstrate the performance of the various models we used and should also incorporate plots for the hyperparameter fitting.

6. Discussion and Conclusion

(3+ Sentences) This part can be done last

7. Division of Work

- **Robin Luo**: Model fitting
- **Marc-Andre Rousseau**: Literature research, TeXing, some minor coding.
- **Peter Xu**: Coding, feature selection and hyperparameter fitting.

8. References

1 Rosenthal, Sara, et al. SemEval-2017 task 4: Sentiment analysis in Twitter Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pp. 502518.

2 Das, Bijoyan. Chakraborty, Sarit. "An improved text sentiment classification model using tf-idf and next word negation" June, 2017, eprint arXiv:1806.06407