COMP90049 Report: Feature Selection in Tweet Sentiment Analysis

Anonymous

1. Introduction

Recently, natural language process (NLP) has gained much attention in data science research and sentiment analysis is one of the most common application in this field (Manning & Schütze, 1999). For processing sentiments, a task with thousands of text documents can be processed by a computer in seconds, which, however, may take a person hours to handle these documents manually. This project aims to gain some knowledge from the sentiment analysis of short texts. Sentiment analysis refers to the process of identifying and categorising writers' attitudes through machine learning methods (Vinodhini & Chandrasekaran, 2012). Nowadays, people usually write 'tweets' on social media to express their feelings or mood. The hypothesis of this paper is that the accuracy of sentiment classification can be improved by filtering sentiment-unrelated features such as id, pronouns, and conjunctions.

2. Related Work

Sentiment analysis is an important analysis type because it can help people in decision making through understanding sentiments (Zhou, Tao, et al., 2013). To deal with the problem that tweets are usually informal and in creative language, Efthymios, Theresa, et al. (2011) capture those new features as a complement to the existing lexical resources in their supervised approach. Similarly, Neethu, et al. (2013) also develop a new feature vector to classify tweets sentiment in the specific domain of tweets about electronic products, the results from which may be used to create a new marketing strategy. All of those research shows the significance of the feature selection which can absolutely affect the machine learning methods performance.

3. Dataset

The dataset applied in the experiments is from the social media platform Twitter contributors including a train tweets text file along with a train labels text file and a corresponding evaluation tweets text file along with its labels (Rosenthal, Sara, et al., 2017). Based on those raw tweets and labels, preprocessed features are stored in two CSV (comma-separated values) files for training and evaluating respectively.

4. Evaluation Metrics

The evaluation of utilized machine learning classifiers can be based on the following metrics:

- Accuracy: For each classification model evaluation, the accuracy of the classification model is the proportion of correct predictions in all attempts.
- Precision: For classification model which predicts more than one prediction, the precision measures the proportion of correct predictions among all candidates.
- Recall: Similarly, the recall measures the proportion of the predictions that are correct.
- F1 score: By considering both precision and recall, the F1 score of a classification model is a measure of the accuracy.

The common knowledge is that all those metrics are calculated based on the TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative) of the classification results. In this report, the evaluation metric is mainly selected as the accuracy.

5. Methodology

In this report, three machine learning classifiers will be discussed, which are Naïve Bayes, Decision Tree, and Support-Vector Machines (SVM) and the baseline performance is measured by ZeroR classifier. All those classifiers are machine leaning methods that can classify tweets into different classifications.

6. Initial Data Cleaning

Before the experiment, it is important to pre-process the text data for training and evaluating as it can make the raw data ready for mining and easy for extracting information and applying machine learning (Joshi, 2018).

Despite that the pre-processed CSV files are given, an experiment for initial data cleaning of text files is conducted to illustrate the relationship between features and sentiments. The results for five most frequent terms for each sentiment are shown in Figure 1.

From Figure 1, we can see that most terms are noun, which may be a person or an organisation. From this experiment, we can know that through feature extraction, some information can be

gained from the raw tweets such as people tend to be positive or negative towards a specific person or an organisation. In terms of the feature "ISIS", for example, we can see that more counts of features are found in neutral tweets than negative, which may tell us that most people hold neutral attitude to the "ISIS".

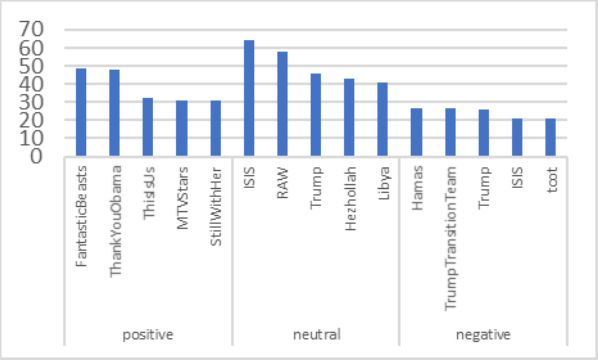

Figure 1. Most Frequent Terms in Tweets

7. Classification

In the CSV file, there are in total 46 features extracted in the pre-processing step. However, not all those features are good for classifiers to predict in better performance. The experiment aims to find out the impacts of some sentiment-unrelated features such as id, pronouns, and conjunctions on the classification.

7.1. All Features

First, all features are used for classifiers to identify the sentiments of tweets and the classification result is shown in Figure 2.

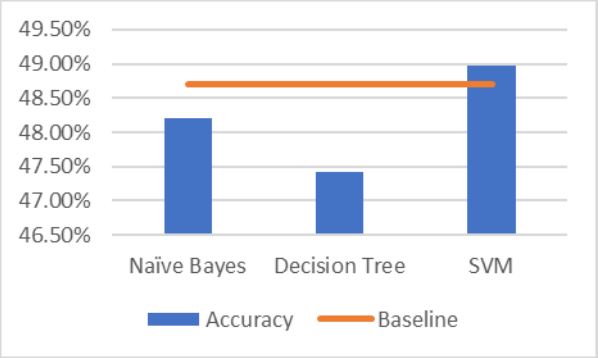

Figure 2. Classification with All Features

As we can see from the Figure 2, only SVM classifier meets the baseline performance, both the Naïve Bayes classifier and Decision Tree classifier are not in expected performance.

7.2. Features without id

One reason for that result may be the interference of the ids of tweets. In this case, features without id are applied in the second trial and the result is shown in Figure 3.

From Figure 3, we can see that even though the performance of both Naïve Bayes and Decision Tree classifiers are still under the baseline, the result shows that the feature 'id' is not an independent or unique attribute. In other words, there can be tweets with the same id in the text file.

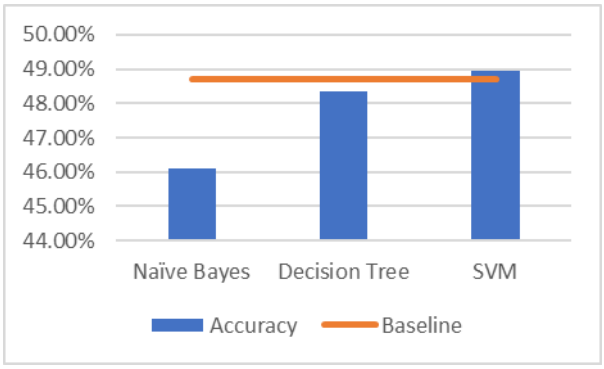

Figure 3. Classification without id

7.3. Features without Pronouns

Similarly, pronouns can be removed from features to experiment classifications without pronouns, which contain 'we', 'my', 'they', and 'their' in the 46 features. The result is shown in Figure 4. By comparing this result with the classification without id, we can see that they are quite similar with each other, with only marginal differences.

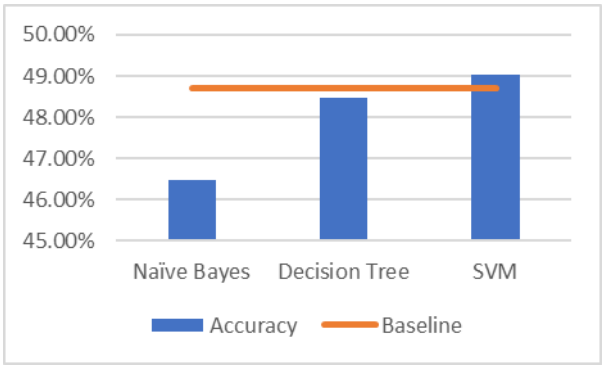

Figure 4. Classification without Pronouns

7.4. Features without Conjunctions

In terms of conjunctions, there are only two conjunctions in the features, 'and' and 'so'. The result for classification without conjunctions is shown in Figure 5.

From the Figure 5, we can see that the condition is also similar to the one without id. Both the Naïve Bayes classifier and Decision Tree

classifier are affected by the decline of features. The former is declined while the latter is increased. Possible reasons for that can be that the conjunctions still are independent with other features and those conjunctions may be related to the decision making in the Decision Tree method.

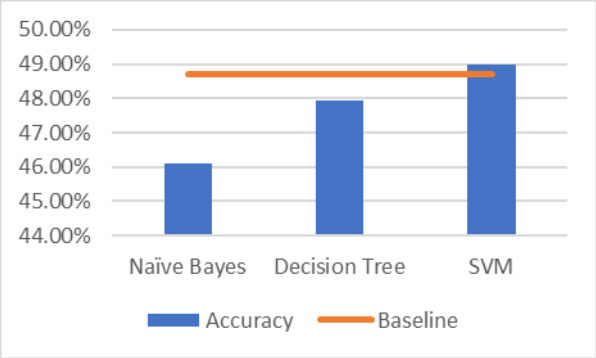

Figure 5. Classification without Conjunctions

7.5. Features without Combinations of id, Pronouns, and Conjunctions

As those features selected are separate in the previous experiments, to make it more convincing, we should consider the combination of those conditions. The classification results for four different combination situations are shown in Figure 6.

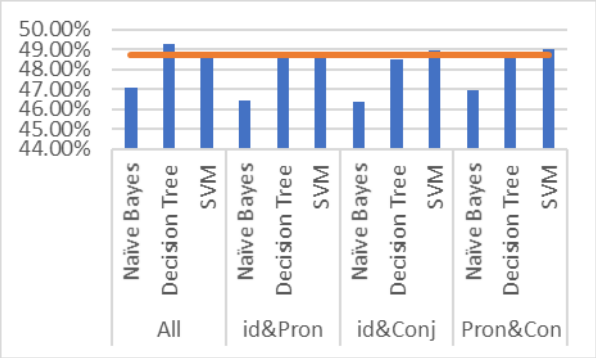

Figure 6. Combination Situation Classifications

From the results of those combination situations, by comparing with other results, we can easily identify that the maximum performance one is the Decision Tree classifier with filtering sentiment-unrelated features of id, pronouns, and conjunctions. In this case, the hypothesis can be determined as true. However, through all the experiments, we should be aware of that features are not absolutely independent, which is the reason why the performance of Naïve Bayes classifiers is not satisfied.

8. Conclusion

In conclusion, this report applies three machine learning classifiers which are Naïve Bayes, Decision Tree, and SVM to verify the correctness of the hypothesis. It is shown that SVM has stable performance among all experiments. And the best performance is from the Decision Tree classifier with filtering sentiment-unrelated features of id, pronouns, and conjunctions, which is the reason why the hypothesis is decided to be true. Overall, feature engineering is both required at the beginning and during the sentiment analysis process and different feature collections are appropriate for different types of machine learning classifiers.

9. References

Joshi, P. (2018). Comprehensive Hands on Guide to Twitter Sentiment Analysis with dataset & code. Retrieved from https://www.analyticsvidhya.com/blog/2018/07/hands-on-sentiment-analysis-dataset-python/

Kouloumpis, E., Wilson, T., & Moore, J. (2011, July). Twitter sentiment analysis: The good the bad and the omg!. In Fifth International AAAI conference on weblogs and social media.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.

Neethu, M. S., & Rajasree, R. (2013, July). Sentiment analysis in twitter using machine learning techniques. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

Rosenthal, Sara, Noura Farra, and Preslav Nakov (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval '17). Vancouver, Canada.

Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. International Journal, 2(6), 282-292.

Zhou, X., Tao, X., Yong, J., & Yang, Z. (2013, June). Sentiment analysis on tweets for social events. In Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 557-562). IEEE.