How can combination classifiers help in producing a more accurate model?

Anonymous

1. Introduction

As the world is becoming more reliant on technology and automation, there became more problems to be solved with an emphasis on accuracy. Simple problems can be solved with satisfying accuracy via training base classifiers due to a lack of features or dependencies. Complex problems, however, have an abundant number of features and dependencies, making base classifiers unable to suffice. To tackle this problem, classifiers are combined to achieve a robust model. But how can they help with increasing accuracy? To answer this, a dataset was prepared consisting of tweets being annotated with positive, negative, and neutral sentiments. Therefore, a model is to be trained to predict tweet sentiments.

2. Hypothesis

The effect of limitations in each model is being reduced due to the other models not having the same limitations.

3. Method

As a baseline, the zero-r baseline was utilised, which achieved an accuracy of 0.58 for the training dataset. As no features are considered for this baseline, it is not fit to be used as a classifier.

2.1 Vectorisation

At the current moment, the data has no features to help in training a model but as the data is language based, each tweet can be broken down into vectors, where each word can be a feature. To achieve this, both Bag of Words (**BoW**) and TFIDF will be implemented depending on the model being seed.

Theoretically, TFIDF will be perform better as the importance of each word is calculated based on its appearance in each document, or tweet in this case. In contrast to this, BoW emphasises importance based on the frequency in which a word appears. In other words, TFIDF in theory would be better at identifying words that have a strong association with certain sentiments.

2.2.1 Stop words

TFIDF was also implemented with stop words, so that grammatical words will not have an impact on the results. With all the vectors, there is a total of around 40000 features in the dataset.

2.2 Feature Selection

After the features have been obtained, the next step is to train the models, however, using all 40000 features would be too time consuming. Therefore, feature selection will be implemented to reduce complexity and to make the models more generalised. This would also further help with identifying important words that have strong associations with certain sentiments.

2.2.1 Chi-squared or Mutual information

The feature selection methods being used will be chi-squared or mutual information with k varying between 1000 and 10000. Both generally give the same results, but mutual information should be used as it is more robust in large datasets. Since important words could appear rarely, in the thousands of tweets, the chi-squared value would be low, causing a low dependency between the feature and the class and thus cause inaccurate predictions. Mutual information would not have that problem, as it has a bias towards rare information.

2.2.2 Implementation

Mutual information can only used on discrete values, which is why it can only be implemented on the BoW vectors. Chi-squared, however can be implemented with both BoW and TFIDF vectors. As TFIDF theoretically performs better, Chi-squared will be utilised on it.

3.3 Choosing the models

After feature selection, the next step is to choose the 2 base models to train and fit the data. The 2 models that will be trained are Naïve Bayes and Logistic Regression (**LR**). Multinomial Naïve Bayes (**MNB**) was chosen due to its high scalability, ease of implementation, and due to how it is being applied to sentimental analysis in the real world. LR was chosen due to how it is a discriminative model, meaning it finds the most important features and learns to distinguish between classes. This is different to Naïve bayes, which learns about every feature and determines class through that.

3.4 Choosing the combination classifier

For comparison, a combined classifier will be implemented to see how much of a difference it can make in the accuracy. Stacking will be used as it generally gives better results then the base classifiers at their best. As the predictions from each base classifier is used as an input for a meta-classifier to form a new prediction, the different trends in each base model are captured. As a result, the new prediction will capture less biases, thus leading to more accurate results. The base classifiers being used will be the same 2 models from above.

2.5 Evaluation metric

To check for consistency, cross-validation will suffice due to its simple implementation. The accuracy scores will be compared for this analysis.

Results

After implementing the method explained above, it is found that implementing mutual information for feature selection was too time consuming, so as a result, the chi-squared feature selection was applied on both BoW and TFIDF. Fitting the base models to the resultant selected features and testing the models led to the following graphs.

3.1 Base classifiers

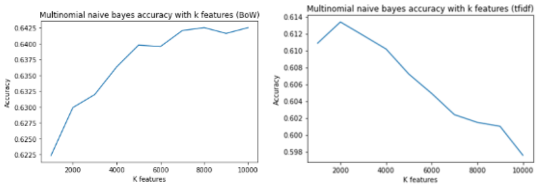

Figure 1- The left image is a graph showing the accuracy of Multinomial Naïve Bayes with k features using Bag of Words. The right image uses TFIDF.

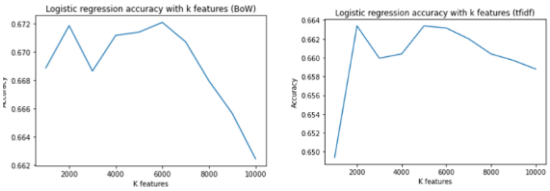

Figure 2- The left image is a graph showing the accuracy of Logical regression with k features using Bag of Words.The right image uses TFIDF.

From these 4 graphs, it can be said that LR generally has higher accuracy than MNB. For both BoW and TFIDF, LR follows a similar decreasing trend for an increasing k value. However, in the case of MNB, there is an increasing trend for BoW, but a decreasing trend for TFIDF. Also, although TFIDF should theoretically performs better, the results show that BoW has higher accuracy.

3.2 Stacking

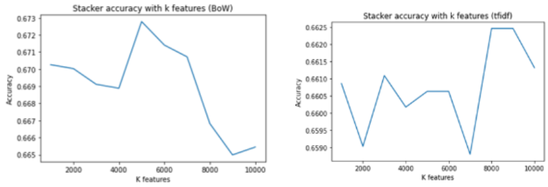

Figure 3- The left image is a graph showing the accuracy of Stacking with k features using Bag of Words.The right image uses TFIDF.

From these 2 graphs, it can be said that the stacker does give overall better accuracy than both the base models, however there are instances where LR does have better results. There are instances where accuracy decreases, then increases and decreases again, indicating that there is no trend in the stacker. The stacker in BoW has achieved the highest accuracy in all the graphs with a value of 0.67278 and k=5000.

3.3 Evaluation results

The cross validation has shown a low variance in results with an average accuracy of 0.6665 for BoW and 0.6636 for TFIDF with a k value of 3965. All models were able to achieve higher accuracy than the zero-r baseline, indicating that they have less bias than the baseline.

Analysis

There have been positive results but there are parts of the methods that need improvements

4.1 Logistic Regression trend

There is a decreasing trend in the LR models due to their tendency to overfit data with high dimensionality or a high number of features. Usually, to avoid overfitting with LR, regularisation techniques are applied which result in more complex model. As the method utilised a simple implementation of the model, no regularisation techniques were applied. LR also requires only the important features to be used to build the model, otherwise the probabilistic predictions made by the model may be incorrect. Therefore, as the number of features had increased, there would be more less informative features in the dataset being included.

4.2 Unimportant features

The inclusion of unimportant features is also related to BoW and TFIDF being simple vectorisation methods and so are not well fit for the task of vectorizing tweets. Although TFIDF attempts to fix the weakness of BoW, they both have the problem that they are unable account for the contextual meaning of a sentence. For example, both methods would give the same vector for the sentences “Text processing is easy but tedious.” and “Text processing is tedious but easy.” even though they have different meanings. The difference in meaning can also result in different sentiments so it is evident that this played a partial negative role in the accuracy.

4.3 Multinomial Naïve Bayes performance

MNB is shown to perform worse than Logical Regression and this is likely due to how the prediction accuracy of the algorithm it uses is lower than other probability algorithms. It is also due to Naïve Bayes assumes independence between features, while LR does not. Words can be dependent on each other. For example, the word ‘heaven’ can be correlated with the word ‘god’ and these 2 features can affect how the model perceives their sentiment.

The BoW and TFIDF versions of MNB have opposite trends. This is due to how other than stop words, there are other common English words or topics that are being counted in the documents. Therefore, if a lot of the common topics such as ‘car’ were being vectorized, it would affect the accuracy of the model. For smaller k values, less of the common words would be considered, which explains the decreasing trend. BoW would not have this problem as the value of a word is solely based on its frequency compared to TFIDF where the value is dependent on a word’s appearance in other documents. This also contributes to TFIDF having worse results than BoW.

4.4 Stacker performance

As the stacker is based off its base models, its accuracy would be somewhat reflective of the base model accuracies. As BoW had more accurate results, the stacker equivalent would also have more accuracy than the TFIDF counterpart. This dependency is also the reason why Logical Regression sometimes performed better than the stacker. When the difference between Logical Regression and MNB became too large, the stacker would be unable to perform well. Therefore, it is important that the base models have similar results.

As a matrix of predictions is made by the base models, the meta-classifier will be trained using these predictions. If a model predicts a sentiment correctly and the other predicts it incorrectly, the meta-classifier will be trained to make less mistakes and predict the sentiment more accurately. In other words, the effects of the limitations of each base model are being reduced through the stacking process. While Logical Regression tends to overfit, MNB tends to underfit, and while MNB uses a less accurate algorithm, Logical Regression uses a more accurate one.

4.5 Error Analysis

Cross validation was implemented for evaluating the variance of the stacker model, which resulted in a low variance, but as the accuracy is still quite low, the bias is still quite high. To reduce the bias or improve the model, nested cross validation should have been utilised to train the model more. More base classifiers should have also been added to improve the reliability of the model. With this, there should be a significant increase accuracy.

Conclusion

This problem has demonstrated how just using a base classifier is not enough to reliably solve a complex problem. An accuracy of 0.66 is not enough to approve a method to be implemented. Each model has their limitations and that is what makes them incapable of solving real world problems. Combined classifiers have demonstrated that more reliable results can be achieved making those limitations or bias play less of a part in the classification. Although the results shown give the impression that they do not seem to have huge impact, the implementation was simple and could have been further improved.

References

Mukherjee, A., Venkataraman, V., Liu, B. & Glance, N. What Yelp fake review filter might be doing? 7th International AAAI Conference on Weblogs and Social Media, 2013.

Rayana, S. & Akoglu, L. Collective opinion spam detection: Bridging review networks and metadata. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015. 985-994.

(What is a meta-classifier? – Terasolartisans.com, n.d.)

Terasolartisans.com. n.d. What is a meta-classifier? – Terasolartisans.com. [online] Available at: <https://www.terasolartisans.com/john/notes-of-a-writers/what-is-a-meta-classifier/> [Accessed 12 May 2022].

(Naive Bayes vs Logistic Regression | Top 5 Differences You Should Know, n.d.)

EDUCBA. n.d. Naive Bayes vs Logistic Regression | Top 5 Differences You Should Know. [online] Available at: <https://www.educba.com/naive-bayes-vs-logistic-regression/> [Accessed 12 May 2022].

(Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2022 | upGrad blog, 2021)

upGrad blog. 2021. Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2022 | upGrad blog. [online] Available at: <https://www.upgrad.com/blog/multinomial-naive-bayes-explained/> [Accessed 12 May 2022].

(Grover, n.d.)

Grover, K., n.d. Advantages and Disadvantages of Logistic Regression. [online] OpenGenus IQ: Computing Expertise & Legacy. Available at: <https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/> [Accessed 12 May 2022].

(An Introduction to TF-IDF using Python, 2019)

Medium. 2019. An Introduction to TF-IDF using Python. [online] Available at: <https://medium.com/analytics-vidhya/an-intr

oduction-to-tf-idf-using-python-5f9d1a343f77
> [Accessed 12 May 2022].