Sentiment Analysis of Tweets

November 26, 2021

1 Introduction

Sentiment classification is the process of predicting a user's attitude from a piece of social media text, classifying the opinion as positive, negative or neutral. This paper utilises a large data-set of 180,000 tweets from users, acquired from two papers. The first running a scraper that queries the Twitter API from April to June 2009, and the second running a streaming crawler on the Twitter website from July to December 2016 (Go, Bhayani, & Huang, 2009; Vadicamo et al., 2017). Further application of various feature engineering techniques are performed to build machine learning models aimed at predicting current attitude of users. Sentiment classification has many implementations, such as location based identification of user opinions to certain events, customer satisfaction with products, etc. The main purpose of this paper will be the implementation of various machine learning models on tweets for sentiment classification.

2 Literature review

There have been substantial research in the literature of sentiment analysis, driven by an absence of a perfect solution to the problem associated with the nuances of language. This is a field that extended the foundations of statistical natural language processing (Manning & Schutze, 1999).

In Go et al. (2009), use of Naive Bayes, maximum entropy and SVM models on Twitter messages for sentiment analysis are investigated, with feature engineering that created unigrams and bigrams of the data-set.

In Mohsen et al. (2021), Arabic quarantine tweets are investigated for sentiment analysis, where Synthetic Minority Oversampling TEchnique (SMOTE), combined with Edited Nearest Neighbor (ENN) provided the greatest performance for the majority of machine learning models, such as stochastic gradient descent, multi layer and support vector machines.

In Naresh and Venkata Krishna (2021), sentiment analysis on Twitter data using a sequential minimal optimization, combined with a decision trees model provided the greatest accuracy of 89.47%.

3 Method

Three feature engineering techniques of the raw data-set is performed to normalise the text and present the data in a meaningful way. These methods are used for unigrams of the raw tweets, which considers each instance of a single word for each tweet.

1. **Term frequency**: The raw text is converted into a count, where each unique word is assigned an identifier. To prevent common or rare words that provide little insight for the models, only 5000 of the meaningful words are retained, where the others are discarded.

2. **Term frequency-Inverse Document Frequency**: In TF-IDF, the top 5000 words are found and given values based on a weight of the frequency of terms, inversely proportionate to how often they occur different tweets.

3. **GloVe**: Each tweet will be represented by a 100 length vector which is a sum of the the associated GloVe vector for each term in the tweet. In Global Vectors for Word Representation, each word has a corresponding 100 length vector which attempts to encode the semantics of words through nearest neighbours and linear substructures (Pennington, Socher, & Manning, 2014).

For these representation of the data-set, a holdout strategy is used with a training set of 160,000 instances, and the remaining 20,000 instances for a development and testing set used to cross-validate the models, ensuring they are not over-fitted or under-fitted. This is the chosen evaluation instead of using K-fold cross validation strategy as the data-set is very large, therefore, no overlap between data-sets is necessary.

Before the use of machine learning models, a zero-R baseline classifier is applied, where the most frequent label in the training data-set is used to classify all test instances. This will be used as a baseline for the subsequent machine learning models as a means to determine the efficacy of various models in contrast to this simple baseline. To determine the efficacy of the models, accuracy and a macro-average F1 score is calculated with the validation set. Macro-averaging is used due to the multi-class nature of the classification (pos, neg, neu). This measures calculates the F1 score for each classification label and equally weighs them to calculate a macro-averaged F1 score, ensuring that there will not be a single classification label with particularly bad performance. The following machine learning models are applied:

- **Multinomial Naive Bayes (NB):** A generative model that calculates the prior class probabilities for the three sentiment classifications (neg, pos, neu). After this, the posterior class probabilities for each word $p(x|y)$ will be calculated, where $p(x|y)$ is determined using a multinomial distribution. A multinomial distribution is chosen and applied onto the term frequency data-set to account for the non-binary classification labels and the multiple occurrences of words that accounts for the total number of tweets (Kibriya, Frank, Pfahringer, & Holmes, 2004). Finally, the the $p(y|x)$ determined with $p(y)\prod_{i=0}^{n} p(x_i|y)$ for n words in the tweet corpus, and the class label the NB model predicts is found using the maximum probability log transformation to avoid underflow with floating-point number representation:

$$\underset{y\in Y}{\operatorname{argmax}}[\log(p(y)) + \sum_{i=0}^{n} \log(p(x_i|y))]$$

- **Multinomial multi layer (LR):** A discriminative model that maximises the posterior probability directly by minimising the negative conditional log likelihood through iterative gradient descent. The model will use a softmax function to classify the instance since there more than 2 class labels. This model will be applied to all 3 feature engineered data-sets after a standard scaler is performed to the data-set to normalise the data to a mean of 0 and a standard deviation of 1.

- **Multi-Layer Perceptron (MLP):** Fully connected sets of neurons in a series of layers that alters the weighting of each neuron through back-propagation after each epoch. Gradient descent will be used to minimise the error and to acquire a high training data accuracy. A maximum epochs of 20 with a single hidden layer of 100 neurons is chosen. This classification can handle non linearly-separable data and uses a softmax activation function for the output layer to address the multi-class nature of the data. This model will be applied to all 3 feature engineered data-sets.

After applying the models on the different feature-engineered data-sets, the data-set producing the best accuracy and F1 score will be used to obtain a learning curve of the model to determine the model bias and variance for each model. Models will then be further tuned based on the underfitting or overfitting issues identified. The final tuned accuracy and F1 score will be compared against the Zero-R baseline and each other to determine the model with the greatest performance.

4 Results

For the zero-R baseline, the accuracy is 0.4067 and the F1 score is 0.1927. This is from the distribution of the training set labels, with 64872 negative labels, 62447 positive labels and 31934 neutral labels. This means that predicting every instance in the training set as the majority class negative will provide the obtained accuracy. For the F1 score, each label will calculate the F1 score and a evenly weighted average is taken across the three classes, which means the F1 score will be very low for the positive and neutral labels, causing a very low F1 score using the zero-R baseline.

Multinomial Naive Bayes:

- **TF-IDF:** Accuracy = 0.7256 F1 = 0.7430
- **TF:** Accuracy = 0.7308 F1 = 0.7467

Using NB, the term frequency data-set produced the highest accuracy and F1 score as expected, this is because the multinomial distribution already accounts for the document frequency, meaning that the TF-IDF data-set will produce an sub-optimal multinomial distribution (Kibriya et al., 2004). Using the term frequency data-set, the learning curve of the model can be examined. The accuracy for both the training and validation data-set is high for a large training data size, with a gradual increase in validation accuracy and a gradual decrease in the training accuracy with a larger training data size. This leads to a small difference in the training and validation accuracy for a large data-set, which indicates this model does not underfit or overfit the training data.

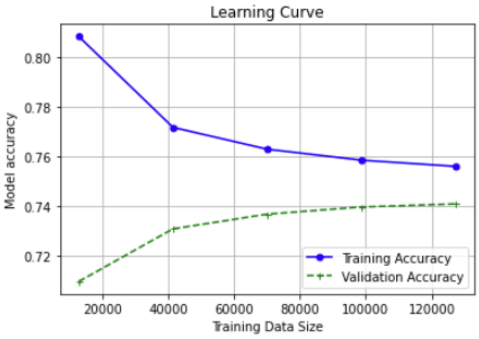


Figure 1: Learning curve of multinomial Naive Bayes for term frequency

Multinomial Logistic Regression:

- **TF-IDF:** Accuracy = 0.7373 F1 = 0.7528
- **TF:** Accuracy = 0.7365 F1 = 0.7517
- **GloVe:** Accuracy = 0.6777 F1 = 0.694

Using LR, the TF-IDF data-set produced the highest accuracy and F1 score. This learning curve shows a similar curve response to the NB, with a high accuracy for both the training and validation accuracy for a large data-set as the gap between the responses becomes small. This shows that the model does not underfit or overfit the data. However, there are apparent differences that can be observed from the NB learning curve. The impact of increasing the training data size has a greater impact on the LR validation accuracy, with approximately 0.63 model accuracy with a data-set of 20,000 compared to 0.71 with the multinomial Naive Bayes model

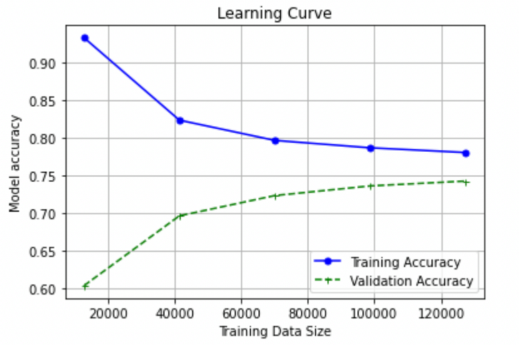
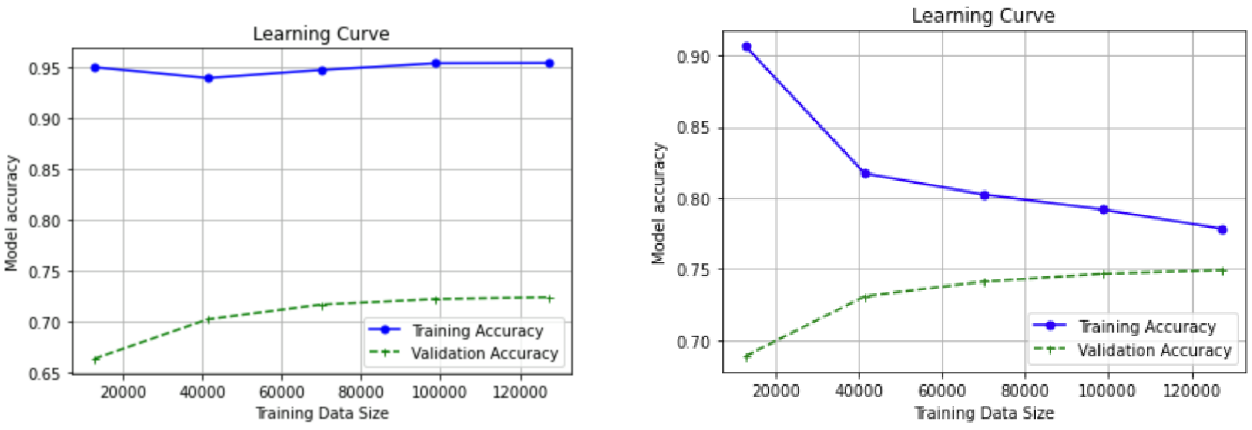


Figure 2: Learning curve of multinomial Naive Bayes for term frequency

Multi Layer Perceptron:

- **TF-IDF:** Accuracy = 0.7136 F1 = 0.7316
- **TF:** Accuracy = 0.7094 F1 = 0.7275
- **GloVe:** Accuracy = 0.7066 F1 = 0.7277

Using MLP, the TF-IDF data-set produced the highest accuracy and F1 score. The learning curve indicates that the MLP is overfitting to the training data, with a high training accuracy but a much lower validation accuracy. Given that overfitting occurs with a maximum epoch of 20, a L2 regularisation is applied instead to reduce the high variance from the complexity of the MLP model. This will result in smaller weights, resulting in a decision boundary plot with less curvatures to allow for greater generalisation of the model. Using a regularisation term of 0.08, the MLP model has improved performance, with an increased accuracy of 0.7436 and an F1 score of 0.7597. Additionally, this removed the overfitting problem, with the L2 regularisation term reducing the gap accuracy difference between the training and validation data-set.



(a) Max Epoch = 20

(b) L2 regularisation = 0.08

Figure 3: MLP Learning curves for TF-IDF

5 Discussion

In terms of the data-sets used, for the NB model, the term frequency data-set produced the best results due to the multinomial distribution innately using the counts of words in modelling the probabilities (Kibriya et al., 2004). For LR, the TF-IDF data-set produced marginally better results compared to the term frequency data-set and GloVe. The improvement over term frequency is negligable, however, GloVe heavily under performs the other data-sets because information is loss when when compressing every word into a 100 length vector representation. For the MLP, all three data-sets performs similarly, with TF-IDF producing slightly better results compared to GloVe and TF. This can be attributed to neural network's ability to learn features themselves from the provided input data. This means that the feature engineering performed onto the data-sets only cause minor changes to the accuracy.

For model performance, all three models performed much better than the zero-R baseline, which had and accuracy of 0.4067 and an F1 score of 0.1927. The best performing model is the MLP model, fitted with the TF-IDF data-set, a tuned L2 regularisation of 0.08 and a maximum epoch of 20. The MLP model used contained a single hidden layer with 100 neurons. This produced the highest accuracy of 0.7436 and F1 score of 0.7597 and the effects of the L2 regularisation vastly improved the model performance from an accuracy of 0.7136 and F1 score of 0.7316. This additional L2 regularisation addressed the overfitting problem by adding a weight-delta term during back-propogation to each weight, penalising large weight values to produce a decision boundary with less curvatures (Phaisangittisagul, 2016). Additionally, the ReLu activation function is used in the hidden layer for the computational advantage, which is an important factor given the large training data-set of 160,000 tweets.

In terms of the learning rate of each model, NB performs the best with a small data-set, obtaining a 0.715 accuracy with a training data size of 20,000 compared to 0.63 for LR and 0.7 for the tuned MLP. This is because in NB, the assumption of conditional independence between the features is used to arrive at the posterior probability, which allows the model to converge quicker. In comparison, LR requires a large amount of data to converge as it makes no conditional independence assumptions of the features. Overall, this assumption is incorrect in the semantics of language, however, it is difficult to model using uni-grams because the conditional dependence of words is highly subject to the order of the words. This leads to the trained LR model slightly outperforming the NB model with an LR accuracy of 0.7373 and F1 score of 0.7528 compared to the NB accuracy of 0.7308 and F1 score of 0.7467.

6 Conclusion

To conclude, the tuned MLP implementation with the normalised TF-IDF data-set provided the best results with an accuracy of 0.7436 and F1 score of 0.7597. Further improvements of a sentiment classification model could be made by using additional feature-engineering techniques. These include considering bi-grams or tri-grams of words to better contextualise the attitude of a tweet with phrases such as 'not bad' that could completely change the sentiment of the tweet. Additionally, emoticons can be included as they heavily contribute to the tone of attitudes in social media. For improvements based on the machine learning models, tuning parameters such as the number of hidden layers, number of neurons in each hidden layer and activation function used for the hidden layers could improve the model performance.

An important aspect of sentiment classification on social media is the ethical considerations. Depending on the use of the sentiment classifications, it could lead to profiling of individuals, monitoring their private behaviour on social media to classify how they behaviour in real life. This could be governments monitoring people with negative social media posts or insurance companies using variable pricing based on an individual's online behaviour. As stated by the prominent whistleblower Edward Snowden, "privacy is a precondition of free speech, free inquiry, association and dissent. In other words, pervasive surveillance is a threat to privacy, but also free speech" (Harting, 2019).

References

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, *1*(12), 2009.

Harting, C. (2019, Nov). *Edward snowden on surveillance and free speech.* Retrieved from `https://news.columbia.edu/news/snowden-mass-surveillance-knight-institute-jameel-jaffer-amy-davidson-sorkin`

Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004). Multinomial naive bayes for text categorization revisited. In *Australasian joint conference on artificial intelligence* (pp. 488–499).

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

Mohsen, A., Ali, Y., Al-Sorori, W., Maqtary, N. A., Al-Fuhaidi, B., & Altabeeb, A. M. (2021). A performance comparison of machine learning classifiers for covid-19 arabic quarantine tweets sentiment analysis. In *2021 1st international conference on emerging smart technologies and applications (esmarta)* (pp. 1–8).

Naresh, A., & Venkata Krishna, P. (2021). An efficient approach for sentiment analysis using machine learning algorithm. *Evolutionary Intelligence*, *14*, 725–731.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Phaisangittisagul, E. (2016). An analysis of the regularization between l2 and dropout in single hidden layer neural network. In *2016 7th international conference on intelligent systems, modelling and simulation (isms)* (p. 174-179). doi: 10.1109/ISMS.2016.14

Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., Dell'Orletta, F., Falchi, F., & Tesconi, M. (2017). Cross-media learning for image sentiment analysis in the wild. In *Proceedings of the ieee international conference on computer vision workshops* (pp. 308–317).