# Sentiment Analysis of financial news using NLP

| | |
|---|---|
| Name: | **Ankit Anil Lade** |
| Registration No./Roll No.: | 18038 |
| Institute/University Name: | IISER Bhopal |
| Program/Stream: | EECS |
| Problem Release date: | September 02, 2021 |
| Date of Submission: | September 26, 2021 |

## 1 Introduction

In this assignment we classify the financial text data into three classes; 'Neutral', 'Negative' and 'Positive'. The training dataset has 1811 text sentences which are labelled. Of these 1811 labelled dataset there are '242' negative labelled, '456' positive labelled and '1113' neutral labels. We apply wordnet lemmatisation technique and bag-of-words model to predict the classes. We also use 5 classifiers during this process namely, Decision tree classifier, Multinomial Naive Bayes, Logistic regression, Support vector machine, and random forest classifier. These help us in determining the classes of the given text data. The data-set was provided by the instructor which had training data, training data labels and test data for which we predict the classes 'positive', 'negative', and 'neutral'. We also use n-grams during this process.

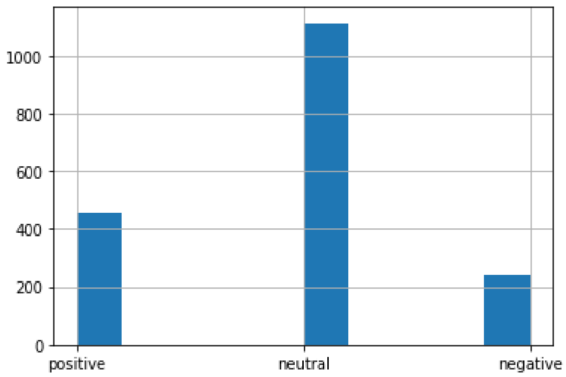


Figure 1: Overview of Corpus

## 2 Methods

First we pre-process the data using the bag-of-words model and remove stopwords from the given data-set in the data cleaning process. This helps in removing repetitive words from the text dataset. Now we apply the tf-idf for a Feature extraction.

**Multinomial Naive Bayes Classifier** The multinomial Naive Bayes classifier is suitable for classification with discrete features. It is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. Bayes Theorem is given as

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{1}$$

**Logistic regression :** Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. [1]
**Random forest classifier:** Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees[2].
**Support vector machine:** In machine learning, support-vector machines (SVMs, also support-vector networks)[3] are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.
**Wordnet:** WordNet is a lexical database of semantic relations between words in more than 200 languages [4]. WordNet links words into semantic relations including synonyms, hyponyms, and meronyms. The synonyms are grouped into synsets with short definitions and usage examples[5].
**Lemmatization** Lemmatisation (or lemmatization) in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form [6].
**Stopwords** Stopwords are meaningless, non-significant terms that frequently occur in a document.They usually have no real purpose in describing document contents and they can- not discriminate between relevant and non-relevant items[7].
**Decision Tree Classifier** A decision tree (it may be also called Classification Tree) is a predictive model that can be used to represent the classification model. Decision trees are usually represented graphically as a hierarchical structure that makes them easier to be interpreted than other techniques. This structure mainly contains a starting node (called root) and group of branches (conditions) that lead to other nodes until we reach leaf node that contain final decision of this route[8].

## 3 Evaluation Criteria

We use Confusion Matrix, accuracy, precision, recall, f1 score(both micro and macro) for the evaluation of the complete process. Accuracy, Recall and Precision can be calculated from the values in Confusion Matrix.
**Confusion Matrix:** It is a clean method of representing prediction results. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class[9].

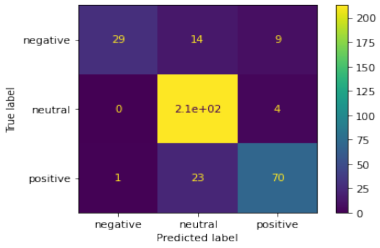


Figure 2: 3 x 3 Confusion Matrix

**Precision:** Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances. From the confusion Matrix we can calculate the Precision(p) as

$$p = \frac{tp}{tp + fp} \tag{2}$$

**Recall:** Recall (also known as sensitivity) is the fraction of relevant instances that were

Table 1: TP, TN, FP and FN for the three classes

| | Negative | Neutral | Positive |
|---|---|---|---|
| TP | 29 | 213 | 70 |
| FP | 14+9 | 0+4 | 1+23 |
| TN | 213+4+23+70 | 29+9+70+1 | 29+213+0+14 |
| FN | 0+1 | 14+23 | 9+4 |

retrieved. From the confusion Matrix we can calculate the Recall(r) as

$$r = \frac{tp}{tp + fn} \tag{3}$$

**F1 Score:** It is the harmonic mean of precision and recall. The F1 score conveys the balance between the precision and the recall. Higher F1 Score is evaluated as better classification of data.

$$f1 - score = \frac{2 * p * r}{p + r} \tag{4}$$

**Accuracy:** It is the ratio of correctly predicted documents to the total number of documents. Higher Accuracy value is good. Also,

$$Accuracy = (tp + tn)/(tp + fp + fn + tn)$$

## 4 Analysis of Results

In this assignment, I analyzed and visualized the given dataset of sentiment analysis for the financial news. All the observations on the training labelled dataset. There were 1802 unique words in the document. The maximum number of unique words were observed for the 'neutral' labels (= 1110). It was observed that the number of words in a sentence were more in case of 'neutral' labelled sentence than 'negative' and 'positive' labels. The punctuations were used more often in 'neutral' labelled sentences than 'negative' and 'positive' labels. The stopwords were more prominent in the case of the 'negative' labelled sentence. On average, we find that the number of words is more in 'neutral' labelled sentences. It may be due to more number of neutral labelled sentences in the dataset.
It was also observed that there were 40601 words in total in the text file, with an average of 22 per sentence before stopword removal and applying wordnet. After using stopword and wordnet lemmatization, the total number of words in the document was only 23898, with an average of 13 per sentence. This shows us there were too many stopwords and punctuations in the given dataset.

Table 2: Analysis of text data on the basis of their classes

| Sentiment | Text Count | Unique | Probability | No. of words | No. of stopwords | Punctuations |
|---|---|---|---|---|---|---|
| Negative | 242 | 242 | 0.614578 | 54 | 14 | 9 |
| Neutral | 1113 | 1110 | 0.251795 | 81 | 13 | 36 |
| Positive | 456 | 456 | 0.133628 | 57 | 13 | 11 |

I applied the bag of words model and calculated all the results with and without using wordnet lemmatization. The accuracy score obtained when using wordnet that compared to not using it were different. It was observed that when we used wordnet, the accuracy score fell but not by a huge margin. We observed that the accuracy score for all the classifiers was higher when we did not use wordnet. It may be due to wordnet lemmatization. The following Table 5 explains the trend.

Table 3: Performance Of Different Classifiers Using All Terms

| Using wordnet | | | | |
|---|---|---|---|---|
| Classifier | Precision | Recall | Macro F1-score | Micro F1-score |
| LR | 0.85 | 0.85 | 0.78 | 0.85 |
| MNB | 0.83 | 0.83 | 0.77 | 0.83 |
| RF | 0.83 | 0.86 | 0.80 | 0.86 |
| SVM | 0.83 | 0.86 | 0.83 | 0.86 |
| DTC | 0.80 | 0.80 | 0.71 | 0.80 |
| Without using wordnet | | | | |
| Classifier | Precision | Recall | Macro F1-score | Micro F1-score |
| LR | 0.79 | 0.79 | 0.66 | 0.79 |
| MNB | 0.83 | 0.83 | 0.77 | 0.83 |
| RF | 0.85 | 0.85 | 0.78 | 0.85 |
| SVM | 0.78 | 0.83 | 0.76 | 0.83 |
| DTC | 0.74 | 0.74 | 0.63 | 0.74 |

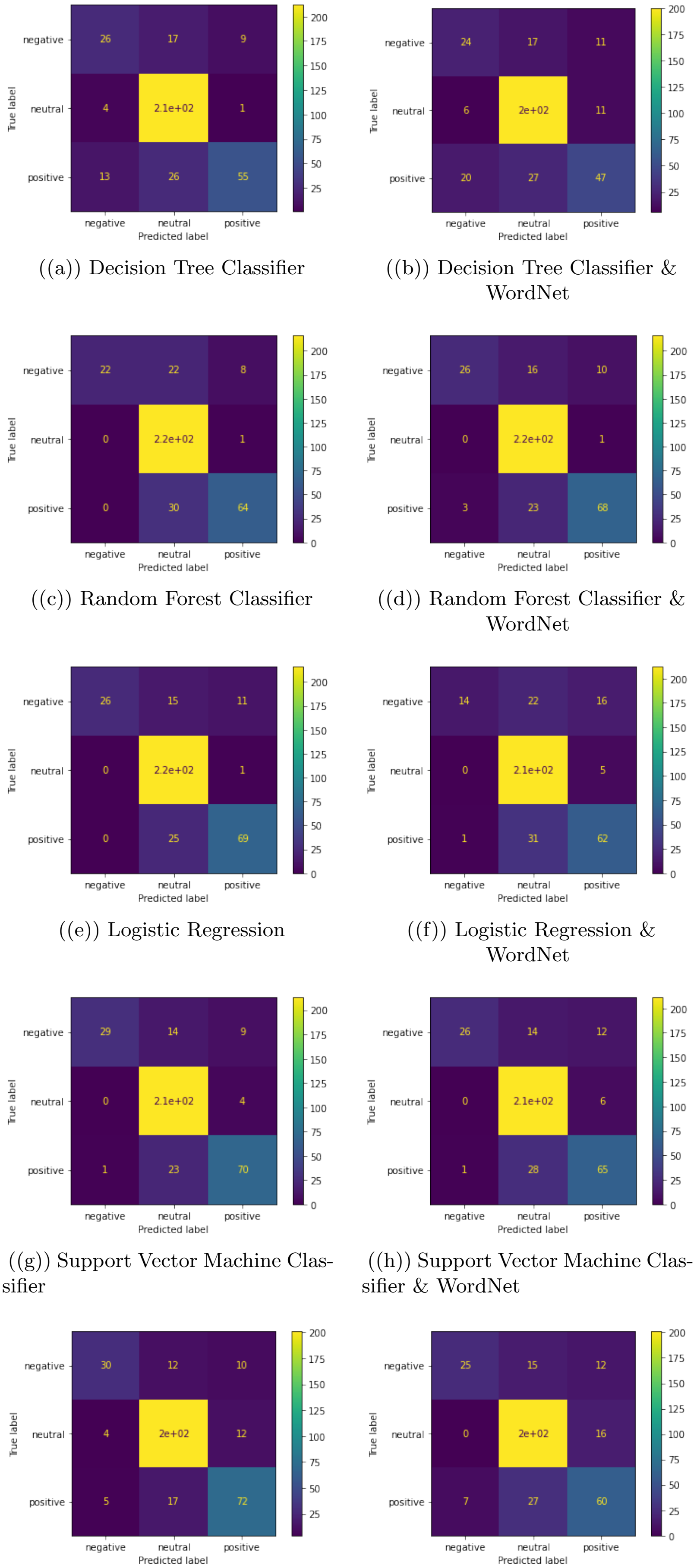
It is also observed that the recall values are generally very low for the ’negative’ labelled dataset. This means that predicted ’negative’ labels are more comparatively false than they are true. The recall value is significantly less due to the small dataset of ’negative’ labels and, hence, less training. The precision, recall and f1-score of the ’neutral’ label is very high due to it’s large training dataset. Recall value of ’neutral’ label is very high for every classifier. This is due to large training dataset in this label.

We also observe that the confusion matrix for the classifiers used give the following results. We use classifiers to predict the classes of the text data. In this confusion matrices we observe that number of false positives are in general more than false negatives for all the cases.

## 5 Discussions and Conclusion

In this assignment we apply different classifier once by cleaning the test dataset by the wordnet and one without it. The result obtained by training the dataset is obtained and we find that in both the cases Random forest classifier performs great. This is due to the fact that the random forest classifier performs great for the multi-class problems. It is optimised to work for multiclass classification.

The study we have explored is the performance of the classifiers using bag of words model and wordnet. In future we also aim to use Word2vec model in this project as the Python3 currently doesn’t support the Word2Vec.



((a)) Decision Tree Classifier

((b)) Decision Tree Classifier & WordNet

((c)) Random Forest Classifier

((d)) Random Forest Classifier & WordNet

((e)) Logistic Regression

((f)) Logistic Regression & WordNet

((g)) Support Vector Machine Classifier

((h)) Support Vector Machine Classifier & WordNet

((i)) Naive Bayes Classifier

((j)) Naive Bayes & WordNet

Figure 3: 4*2

# References

[1] Juliana Tolles and William J Meurer. Logistic regression: relating patient characteristics to outcomes. *Jama*, 316(5):533–534, 2016.

[2] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[3] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[4] GA Miller, R Beckwith, CD Fellbaum, D Gross, and K Miller. Wordnet: An online lexical database. 1990. *Int. J. Lexicograph*, 3(4).

[5]

[6] Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, 2015.

[7] Giuliano Armano, Francesca Fanni, and Alessandro Giuliani. Stopwords identification by means of characteristic and discriminant analysis. In *ICAART (2)*, pages 353–360, 2015.

[8] Abdul Fattah Mashat, Mohammed M Fouad, S Yu Philip, and Tarek F Gharib. A decision tree classification model for university admission system. *Editorial Preface*, 3(10), 2012.

[9] Claude Sammut and Geoffrey I Webb. *Encyclopedia of machine learning and data mining.* Springer Publishing Company, Incorporated, 2017.