Sentiment classification over financial text

Name:	**Ritik**
Registration No./Roll No.:	19245
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	September 02, 2021
Date of Submission:	September 29, 2022

1 Introduction

The growth of financial texts in the wake of big data has challenged most organizations and brought escalating demands for analysis tools. In general, text streams are more challenging to handle than numeric data streams. Text streams are unstructured by nature, but they represent collective expressions that are of value in any financial decision. Sentiment analysis has turned out to be a complicated and strongly domain dependent task. For this task there are so many statistical[1] and Pre-trained NLP models exist. In this paper we experimented with such existed and our proposed approach in statistical modelling.

1.1 Dataset Details:

The dataset is part of financial news corpus consisting of a collection of 2264 sentences with manually annotated class labels. The dataset is very unbalanced contains three classes positive, negative and neutral whose count are 456, 242 and 1113 respectively, which are shown with color bar graph in figure1.

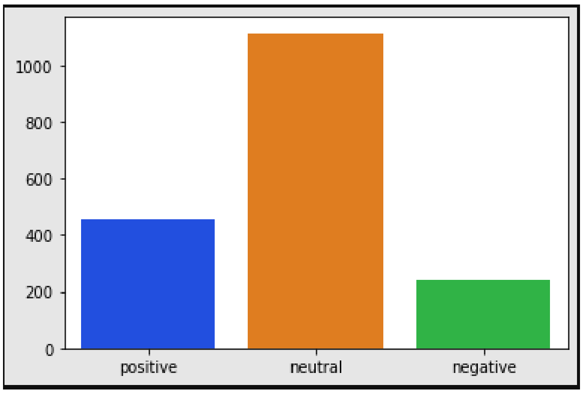

Figure 1: Overview of Corpus

2 Methods

In this project, we evaluate different modules and tricks that are adopted in our proposed model to show their effect on the Financial dataset. We rely on stastical approach for Data Preprocessing and model building in this task. Methods section can be further divided into following subsections:

2.1 Data Pre-Processing:

Steps for the pre-processing are mentioned below:
1. Data cleaning: Data Cleaning process done on all those words/character that don't add much information to the sentence, these words/ characters consists of Stopwords, Punctuation, Numbers, Emojis, Contractions and Emoticons. Lowecasing the data and Removing extra space are also done on dataset.
In addtion to this we performed the Stemming and Lemmatization techniques with the help of python nltk library to get the meaningful context of the words.

2.2 Exploration Analysis:

In this step, we inferred useful information and analyzed the financial text on basis of different text parameters with the help of some useful plots and curve using python matplotlib and seaborn libraries. Some the analysed plots with their parameters are described below:
1. Bar-Plot for each class label to analyze distribution of the classes in dataset.
2. Bar-plot for Doc-length vs class label.
3. Bar Blot for text length of all class labels.
4. Word clouds for all class labels to get insight trend of words to their respective labels.
5. Bar plots for Doc freq vs class labels to get frequency of words in figure2
The analysis will also help identify useful features for building predictive models.

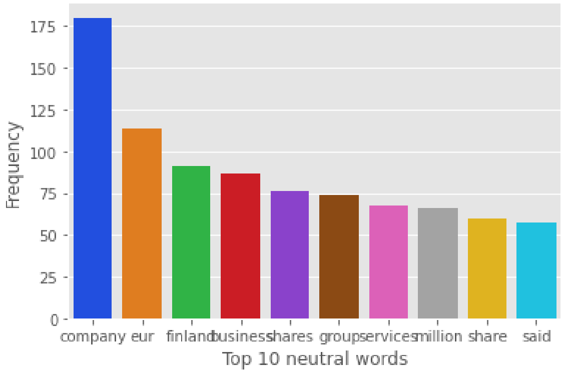

Figure 2: Top 10 Neutral words

2.3 Predictive Analysis:

After these data pre-processing and visualization steps we experimented with different Feature Selection techniques followed by both State of the arts and our Proposed model to get desired output.
We began our experiment with observing the performance of State of the art machine learning models(Logistic Regression, Support Vector Classifier, LinearSVC, Multinomial Naive Bayes and Descision Tree) with hyper-parameter tuning using grid-search technique on our unprocessed data. Out of these models Logistic Regression model outperformed others with F1 score of 0.878.
We continue with Logistic Regression with best hyper-parameters for further analysis on pre-processed data. We did two types of pre-processing, first one is by applying Data Cleaning steps including stopword removal, stemming and Lemmatization. We converted the cleaned data into vector form using tf-idf vectorizer followed by training the Logistic Regression model and we get F1 score of 0.83.
In the second way, we used only the stopwords removal step to get cleaned bag of words. After that we get the Pos tags[2] of the words by using python nltk library. We used these Pos tags as feature by combining them with the bag of words, followed by the tf-idf and Logistic Regression model and we get F1 score of 0.743.

In this second way we also experimented with our proposed model **Model_1**. The basic intuition behind the model is if one word has any financial sense (positive, negative or neutral), the word having same similarity i.e., similar word should have the same financial sense. We used this idea to generate

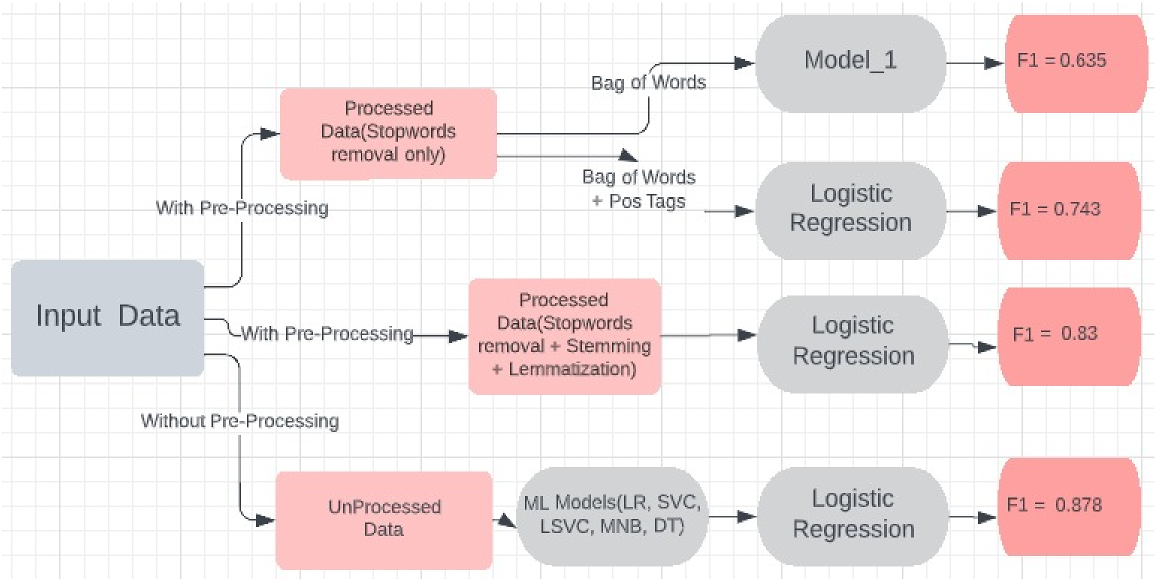

Figure 3: Flowchart of the System

Model	Precision	Recall	Micro F-1 score
Descision Tree	0.754	0.754	0.754
MultinomialNB	0.856	0.856	0.856
SVC	0.865	0.865	0.865
Logistic Regression	***0.878***	0.878	***0.878***
LinearSVC	0.876	0.876	0.876

Table 1: Performance measures of our models on unprocessed dataset

the vector for the entire corpus comparing wu-palmer similarity [3] of all the words in the corpus to top 50 (most frequently occurred) words from each class and assigning the word a tag based on the class of word it most similar is. Then by taking the linear sum of all the elements in a single vector (representing a sentence) we assign the sense to the whole sentence based on the addition score. In this way, we developed a basic model with the micro averaged f1 score of 0.635.
A brief demonstration of our system can be seen by the figure3.

3 Evaluation Criteria

The performance evaluation criteria used in this text classification problem is Confusion Matrix, which describes precision, recall, f-measure, micro averaged techniques. We are using micro averaged F1score because it gives equal weight to all the class and is not affected by the deviations in the distribution of the class log the Log loss it penalizes small deviations in the class.

4 Analysis of Results

For analysing the result we present our result in tables: table 1 and table 2.
Table 11 describes the performance measures of our machine learning models on unprocessed dataset in which the Logistic Regression model classified with best F1 score.
Table 2 describes the performance measures of Logistic Regression on both type of features(Bag of words and Bag of words + Pos tags) and Model_1 on processed data. Logistic Regression is the best model on unprocessed dataset with the hyper-paramters (class_1weight = 'balanced', solver = 'newton-cg')

Model	Dataset	Micro F-1 score
Logistic Regression	Data with Stopwords removal, Stemming and Lemmatization	0.830
Logistic Regression	Data with only Stopwords removal + Pos tags	0.743
Model_1	Data with only Stopwords removal	0.635

Table 2: Performance measures of our models on processed dataset

5 Discussions and Conclusion

In this project we experimented with different types of state of the art models of domain on both unprocessed data (data before data cleaning) and processed data (data after data cleaning). Out of these models, Logistic Regression outperformed than other models on preprocessed data.
After observing this result we can conclude that Data Cleaning Procedures including Stopword removal, Lemmatization and Stemming are not effective in this case of sentiment analysis on financial text.
Future Recommendation: For future reference Data Augmentation technique can be use to overcome the problem of unbalanced dataset. Furthermore, statistical and lexical-based text features can be mined to improve the accuracy of the sentimental analysis.

5.1 Contributions:

My contributions in this project are:

- Text Pre-Processing: In this part I performed the Lemmatization technique and Data Visualization with these plots: Doc-length vs class plot, Bar Blot for text length, Doc freq vs class.
- State of the art models implementaion: For this I worked on Tokenization of bag of words with tf-idf and implemented the state of the arts models (Logistic Regression, Linear SVC, Multinomial NB, Decision Tree), compared the performance and selected the best working model for test data.
- I framed the structure of Model_1 and practically implemented the model and tuned it to get the best accuracy from the model. Calculated the wu-palmer similarity between the words which is principle component behind the working of model_1.

References

[1] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

[2] Abdulmohsen Al-Thubaity, Aali Alqarni, and Ahmad Alnafessah. Do words with certain part of speech tags improve the performance of arabic text classification? In *Proceedings of the 2nd International Conference on Information System and Data Mining*, pages 155–161, 2018.

[3] Xu Wang, Zhisheng Huang, and Frank van Harmelen. Evaluating similarity measures for dataset search. In *International Conference on Web Information Systems Engineering*, pages 38–51. Springer, 2020.