FUNDAMENTALS OF NLP

GROUP PROJECT

Cyril Kancel | Nandini Gantayat | Sai Sandesh Nagarur

IESEG SCHOOL OF MANAGEMENT

NLP - Final Project - Financial Phrases Analysis

Table of Contents

Introduction

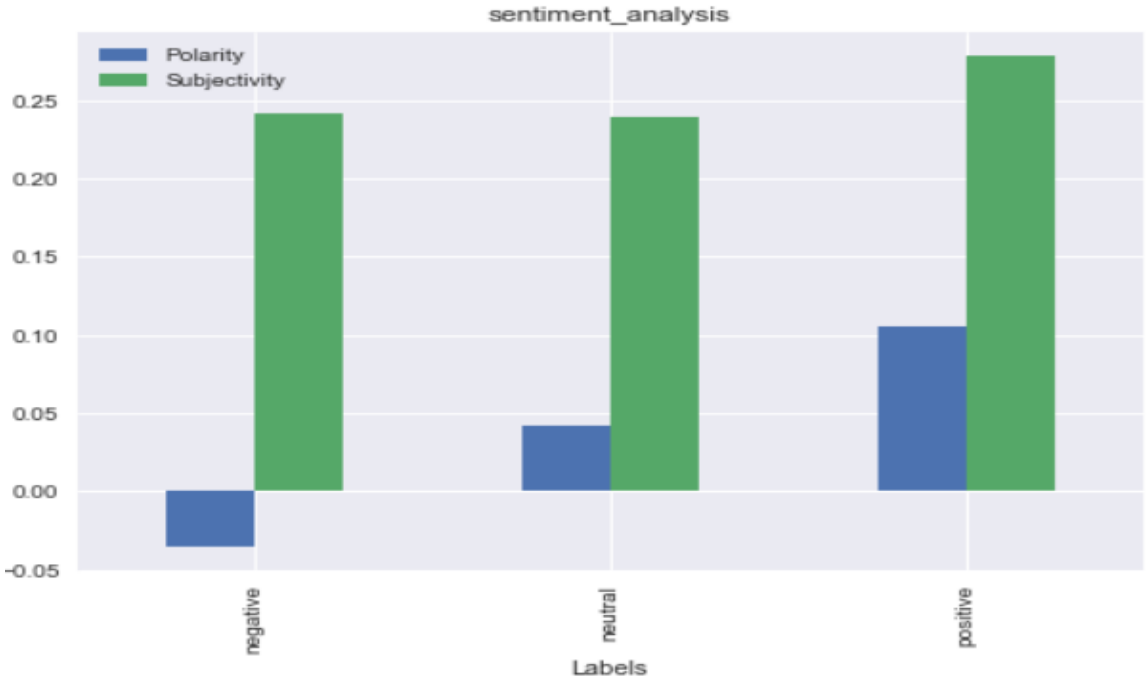
As a retail investor, analysing the market is a really time-consuming task that requires a lot of attention. In fact, when we want to monitor the trends whether in terms of technologies, company news or stock prices, automating this task can lead to a significant time saving. NLP has significant potential to help retailers make optimal decisions.
The Financial Phrasebank is a dataset of pre-defined financial phrases that can be used for sentiment analysis and other NLP tasks. In a study by Khan et al. (2019), the authors used the Financial Phrasebank to analyze customer reviews of retail products. The authors found that their NLP approach was effective in identifying key topics and sentiments expressed in the reviews, such as product quality and customer service. This information can be used by retailers to make data-driven decisions regarding product design, pricing, and customer engagement strategies.

The analysis elaborated in this report was conducted on a dataset of sentences from financial news. This dataset consists of 4846 phrases, all related to the field of finance. All these sentences are labelled as "Positive", "Neutral" or "Negative" according to the sentiment associated with the sentence.

Polarity and Subjectivity Analysis of the phrases

As a first step towards exploring the data we run a quick polarity and subjectivity analysis on the phrases. Polarity analysis helps to determine the sentiment of a text such as positive or negative and subjectivity analysis on the other hand helps determine the amount of opinions in the phrase rather than simply stating facts.

The results of this analysis are displayed in the below table:

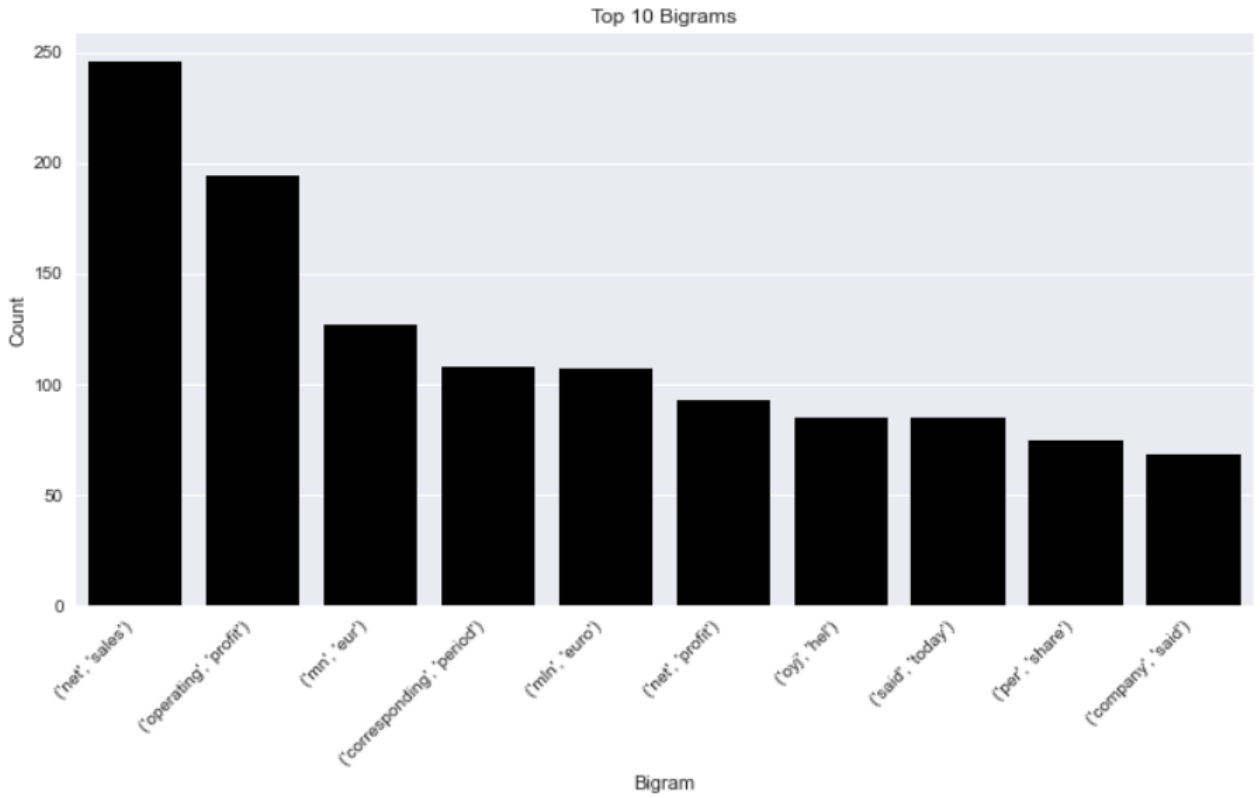


There is a clear distinction in polarity based on the labels of the phrases. The negatively labelled phrases have a negative polarity (-0.03), the neutral phrases have a very slight positive polarity (0.04) and the positively labelled phrases have a polarity north of 0.1. But when it comes to subjectivity there is hardly anything to distinguish them by and that is absolutely alright as it would only mean that all the phrases are more or less equally opinionated.

Bigrams and Word Clouds

Next we take a look at the most commonly occurring words and bigrams in the entire dataset. We also include bigrams as it helps us to capture phrases that single words would miss. Let's start with the bigrams.

The following table contains the 10 most common bigrams:

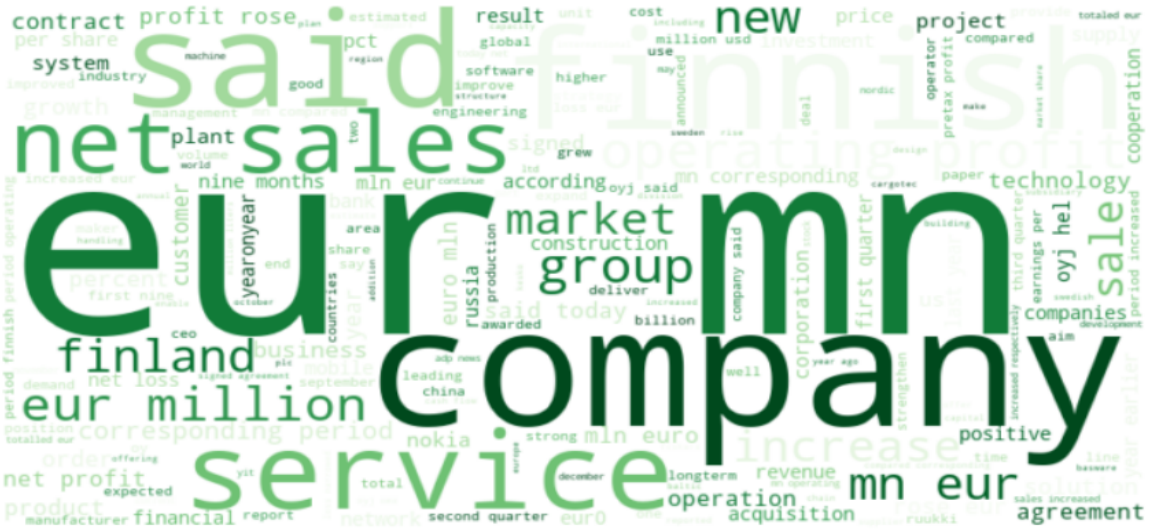


The bigrams 'net sales' and 'operating profits' occur nearly 200 hundred times throughout the dataset.

Now let us take a look at the single words also known as unigrams. The most commonly occurring words are represented in the following word cloud:

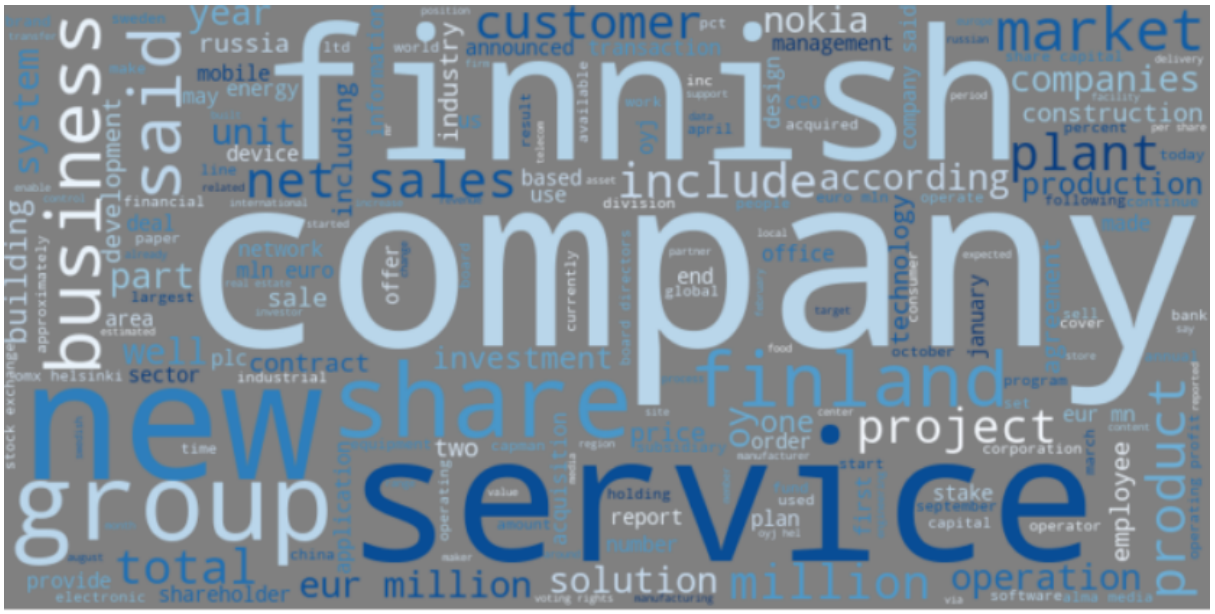
We can clearly see as stated in the introduction that this is a financial dataset and now we can infer that a huge part of the data is related to the country FInland. This word cloud gives a good overall picture of the words that are used in the phrases but let's go one step further and see what kind of unique words distinguish the negative from the neutral phrases and the neutral from the positive phrases. To do that we create 3 separate word clouds, one for each label.

Positive Word Cloud:



Negative Word Cloud:

Neutral Word Cloud:



Unfortunately the most common words in all three situations are pretty much the same which means that it is not possible to determine the polarity of the phrases simply based on the choice of words. Therefore let's try to find out other ways to do it. But before we do that there are a few other things we'll do.

Similar Words

The first of them is to find all the words that are similar to suggested words. For that we use the vector representation of words to graph them. The interpresentation of this graph is fairly simple, the closer two words are to each other the more similar they are supposed to be.

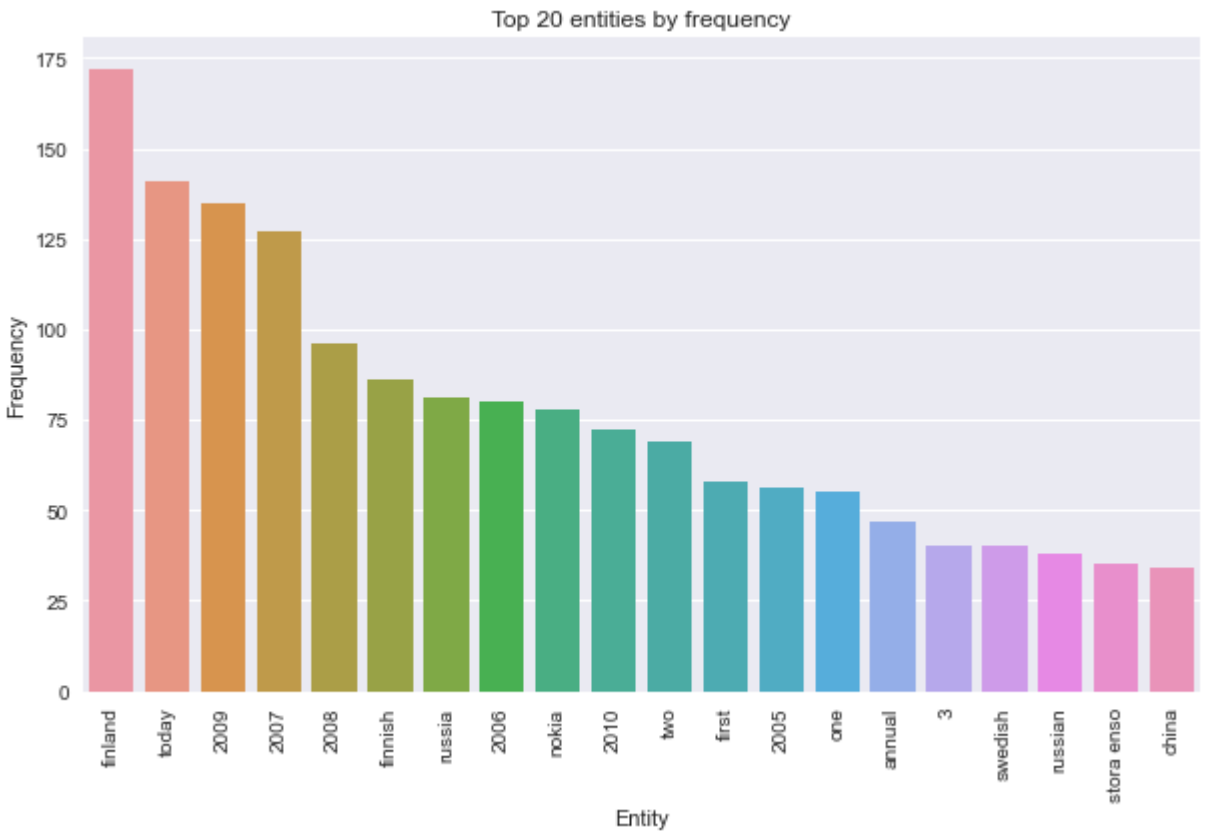
Below is a graph of similar words for the words, 'technology','nokia','finland','stocks':

Semantic orientations

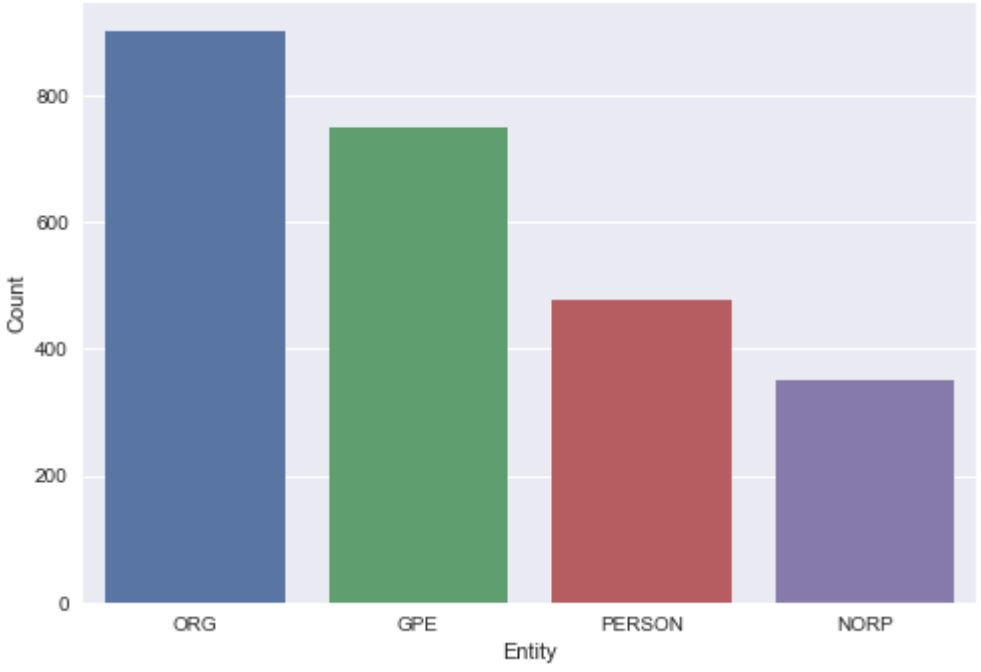
Semantic orientation is often used in sentiment analysis. It helps in automatically identifying the sentiment expressed in our phrases. Here we use it to find the ambivalent and ironic phrases. Ambivalent phrases express both positive and negative emotions therefore it is hard to classify them. Ironic phrases say something but mean something else, and that is something very hard to classify too. Therefore once we identify these phrases, we will eliminate them before proceeding for further analysis.

Named-entity recognition

In order to locate and classify named entities in the different sentences of the dataset, we use a named-entity recognition algorithm. This step of the analysis will help us scan in the dataset which entity is mentioned with a high frequency compared with others. The following graph shows the 20 entities mentioned the most in the dataset.



It is also possible to check which type of entity is the most recurrent in the dataset. As the following graph shows, in this dataset, the most common types of entity appearing are “ORG” (companies, agencies, institutions), “GPE” (geopolitical entity), “PERSON” (person) and “NORP” (Nationalities or religious or political groups).

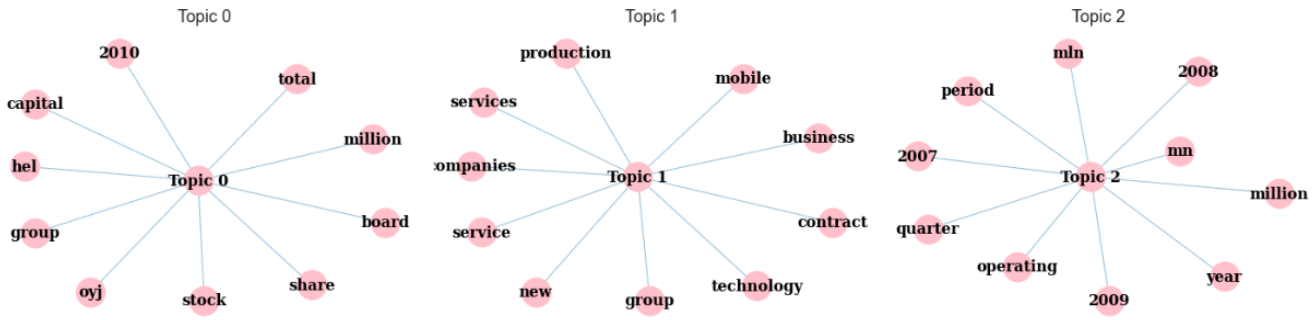


With the aim of deepening the analysis, it is also possible to list the named-entity and order them by frequency, as shown for the “ORG” entities in the next figure.

nokia	76
eur0	20
microsoft	8
matti	6
metso	6

Topic Modelling

In order to discover the different topics present in the corpus, LDA (Latent Dirichlet Allocation) can be a very useful tool. The objective is to extract the main topics, represented by a set of words, that appear in a corpus of documents. In this particular case, as shown in the following graph, three main topics of interest stand out.

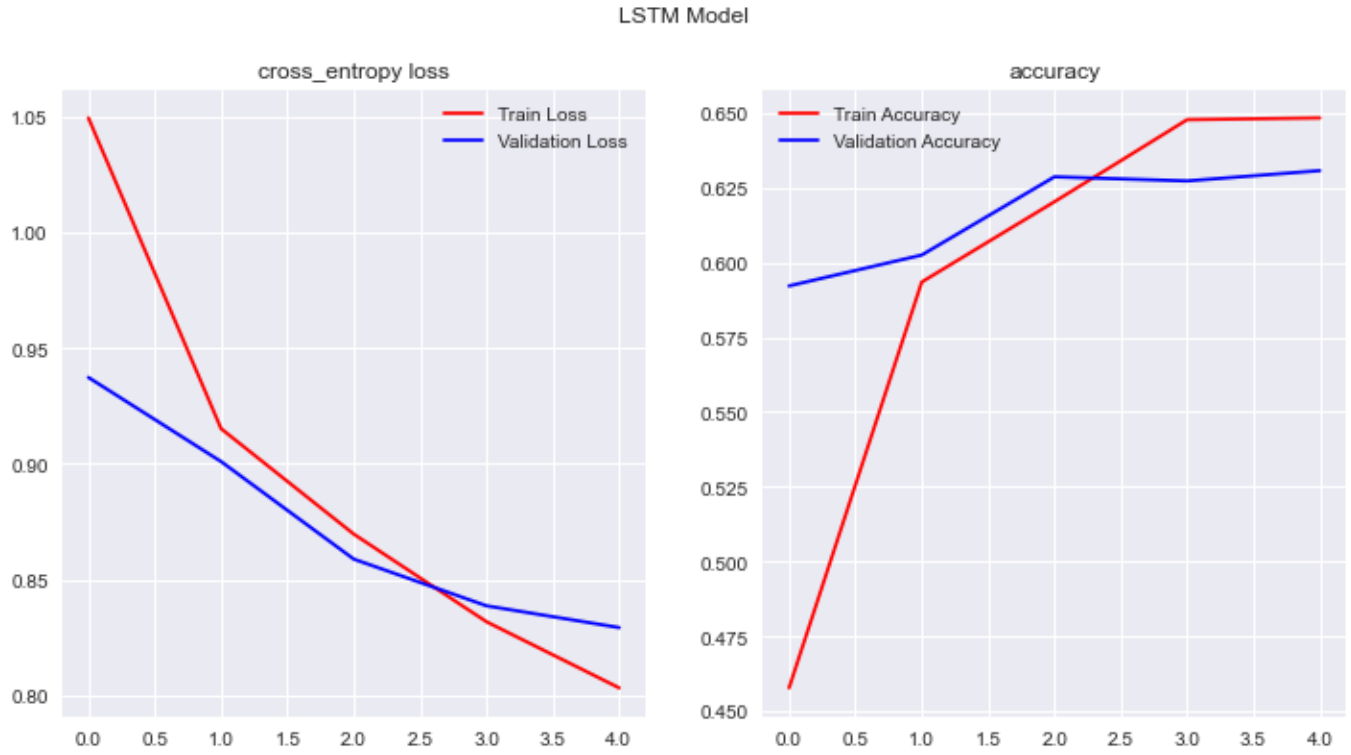


The first topic (“Topic 0”), is mostly composed of words related to the finances of the organisation, its capital, stock etc. The second topic (“Topic 1”) regroups words associated with the operational side of the company, its productions, services etc. The third and last topic (“topic 2”) regroups mostly numbers related to the company, such as dates and periods. This last topic is less informative than the other two.

Sentiment Analysis

In our analysis we constructed a LSTM rnn model which is deemed highly efficient for text classification. A common approach with nlp which we discovered is using GloVe a word embedding technique that represents words as dense vectors in high-dimensional space according to their similarities.

Our model was trained for 5 epochs with a batch size of 1200.
The model slightly improves with each epoch as the loss seems to decrease and accuracy increases equally on both training and validation data with no overfitting.



During the first epoch, the model achieved a loss of 1.0497 and an accuracy of 0.4578 on the training data. On the validation set, the loss was 0.9374 and the accuracy was 0.5922.

By the end of the fifth epoch, the model had a training loss of 0.8031 and a training accuracy of 0.6483. On the validation set, the loss was 0.8292 and the accuracy was 0.6307. However, it may be beneficial to further tune the model's hyperparameters or adjust the architecture to potentially improve its performance further.

This kind of analysis can be useful if we want to add more data that is not labelled yet.

Conclusion

NLP is recently making a breakthrough in the financial field. Our analysis gives surface level insights on how implementing NLP in this domain can yield optimized decision making and extracting knowledge from unstructured data.

Challenges

The main challenge encountered in the elaboration of this project is this particular dataset. In fact, finding an appropriate dataset was a hard task. The main issues related to this dataset are the lack of variety of the news covered, as well as the imbalance nature of the data.

Room for Improvement

As of now we have a code notebook which contains everything including some interactive variables. There is a scope for improvement here as we could create an interactive dashboard/app in the future so that it can be a more user friendly platform for anyone trying to gain insights using our analysis.

References

https://www.sciencedirect.com/topics/computer-science/sentiment-orientation

https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news

https://nlp.stanford.edu/projects/glove/
https://www.kaggle.com/code/samarthsarin/simple-guide-for-lstm-and-glove-embeddings

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know

Malo, P., Sinha, A., Takala, P., Korhonen, P. and Wallenius, J. (2014): "Good debt or bad debt: Detecting semantic orientations in economic texts." Journal of the American Society for Information Science and Technology.