Sentiment Analysis Applied to Stock Prediction

Domain Background

Sentiment Analysis refers to the use of machine learning to identify the emotional reaction to an event, document or topic [1]. One of the possible applications of sentiment analysis is for predicting stock market movements. The internet is full of sources that represent the public opinion and sentiment about current events. Studies in [2] shows that the aggreagate public mood can be correlated with Dow Jones Industrial Average Index (DJIA).

Problem Statement

This capstone seeks a model which uses the top daily news headlines from Reddit (/r/worldnews) to predict stock market movement. A dataset with 8 years of daily news headlines and their respective DJIAs is available in Kaggle [3]. The stock market movement will be modeled into a binary classification problem, where:

- 1 when DJIA Adj Close value **rose or stayed as the same**
- 0 when DJIA Adj Close value **decreased**.

Datasets and Inputs

Two channels of data are provided for this dataset:

1. **Features:** Historical news headlines from Reddit WorldNews Channel (/r/worldnews). They are ranked by reddit users' votes, and only the top 25 headlines are considered for a single date.
2. **Target Variable:** Stock data from Dow Jones Industrial Average (DJIA). The index is converted binary values where:
    - 1 when DJIA Adj Close value rose or stayed as the same
    - 0 when DJIA Adj Close value decreased.

Three data files are provided on Kaggle in *.csv* format:

1. **RedditNews.csv:** two columns The first column is the "date", and second column is the "news headlines". All news are ranked from top to bottom based on how *hot* they are. Hence, there are 25 lines for each date.
2. **DJIA_table.csv:** Downloaded directly from Yahoo Finance: check out the web page for more info.
3. **Combined_News_DJIA.csv:** This is a combined dataset with 27 columns. The first column is "Date", the second is the target variable (DJIA), and the following ones are news headlines ranging from "Top1" to "Top25".

The model will be implemented using only the file **Combined_News_DJIA.csv**. The range of the data is from **2008-06-08** to **2016-07-01** with a total of **3973 rows**. The most recent two years of the dataset (about 20%), from 2015-01-02 to 2016-07-01, is going to be reserved for testing.

Solution Statement

First, the text data from the 25 features is going to be cleaned (some HTML tags are still present in the original data). Next, the text is going to be grouped and processed into feature vectors. The method **bag of words** [5] is going to be used to represent the text as numerical feature vectors. The bag of words model is going create a vocabulary of **tokens** from the headlines data and then counted. Also, the relevancy of words is going to be accessed using the method **term frequency-inverse document frequency (tf-idf)** [6]. Machine Learning algorithms from sklearn are going to be evaluated. The specific models are still going to be defined, but probably the first approach is going to be Logistic Regression and SVM (Stochastic Gradient Descent if it is too slow).

Benchmark Model

Since this dataset is from a Kaggle kernel, there is no 'official' benchmark available. Below are the scores from some Kaggle users who used the very same metric and test set that is going to be implemented in this project (AUC metric and 2 last years as test set):

Kaggle User	**Method**	**AUC score**	**Reference**
Aaron7sun*	CNN and LSTM	62-63%	Link
Kate	Bernoulii Naive Bayes	59%	Link
Dan Offer	Unknown	56%	Link
Kate	Logistic Regression	49%	Link

* This was the user who provided the original database.

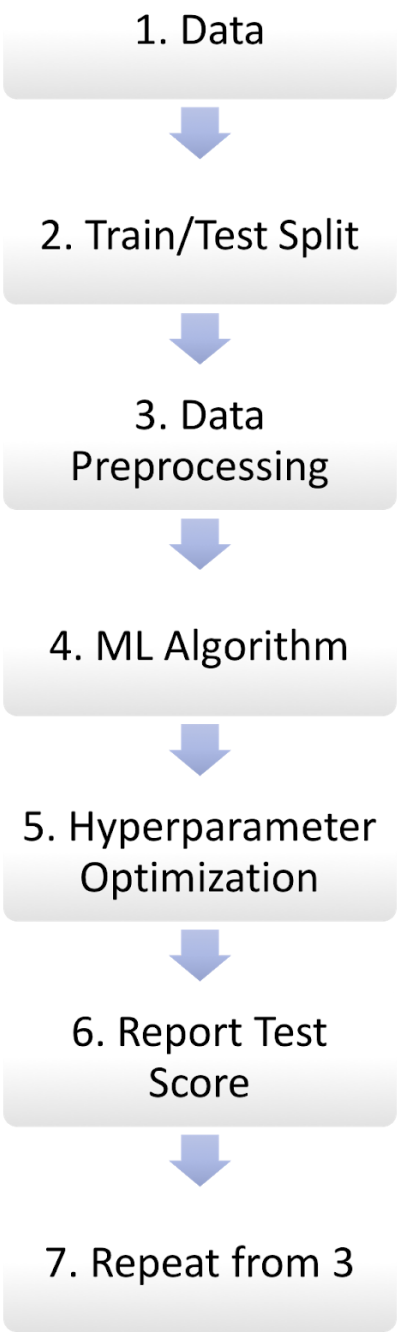
However, this benchmark will be used as a secondary benchmark. I have the following list to be considered as primary benchmark:

1. A 'dummy' classifier with random output
2. A dummy classifier with all 1 as output

Evaluation Metrics

The evaluation metric to be used is Area Under the Curve (AUC) which is a metric derived from receiver operating characteristic (ROC) curve [4]. The most recent two years of the dataset (about 20%), from 2015-01-02 to 2016-07-01, is going to be reserved for testing.

Project Design



1. A dataset with 3973 samples and 25 features is used as input.
2. The most recent two years are going to be used as test set and the previous 6 years of data as training set.
3. The preprocessing goes as follows:

    a. The text from the features is going to be cleaned (some HTML tags are still present in the original data).
    b. Next, the text is going to be grouped and processed into feature vectors using bag of words.
4. As ML algorithms, it is going to be evaluated Logistic Regression, SVM (Stochastic Gradient Descent in case it is too slow), maybe XGBoost, and some other might be considered as well.
5. Besides the hyperparameters from the ML algorithms, different preprocessing methods are also going to be included in the hyperparameter optimization for evaluation. For example, the inclusion or not of the tokenizer function PorterStemmer[7]. Also, the the usage or not of **stop-words**[8] filters in the news headlines. Also the usage or not of **tf-idf**. In case GridSearchCV is too slow, RandomizedSearchCV is going to be used.
6. The model with the highest validation score is going to be used to evaluate the test set. The score is going to be reported
7. The process from 3+ are repetaded for evaluating new ideas thet might come.

References

1. https://en.wikipedia.org/wiki/Sentiment_analysis
2. https://arxiv.org/pdf/1610.09225.pdf
3. https://www.kaggle.com/aaron7sun/stocknews
4. https://en.wikipedia.org/wiki/Receiver_operating_characteristic
5. https://en.wikipedia.org/wiki/Bag-of-words_model
6. https://en.wikipedia.org/wiki/Tf%E2%80%93idf
7. http://www.nltk.org/_modules/nltk/stem/porter.html
8. http://www.nltk.org/book/ch02.html