# 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].

## 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. Different metrics were used by each user. Since this work is going to use the metric AUC, below it is shown some scores reported by users using the very same metric:
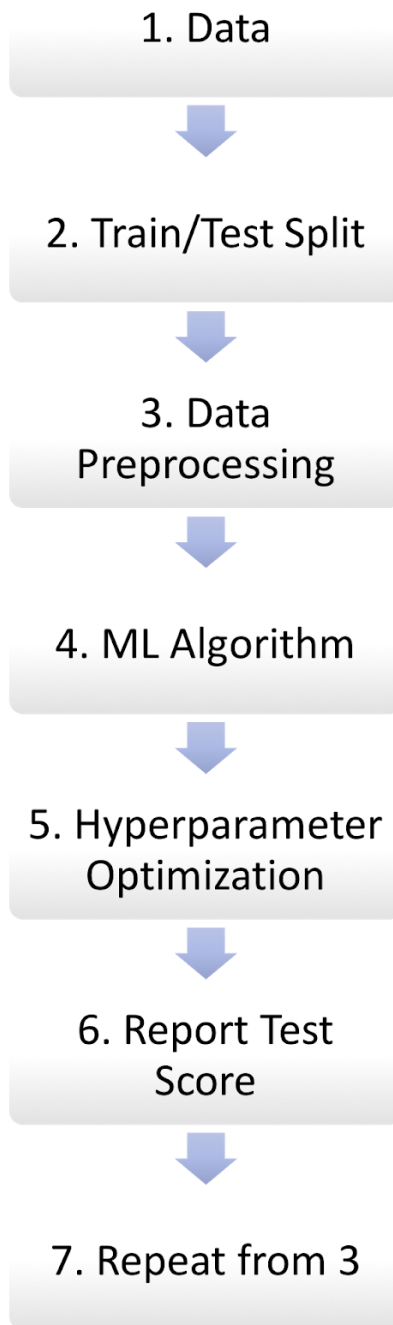
| Kaggle User | Method | AUC score | Reference |
|---|---|---|---|
| Aaron7sun* | CNN and LSTM | 62-63% | [Link] |
| Kate | Bernoulii Naive Bayes | 59% | [Link] |
| Dan Offer | Unknown | 56% | [Link] |
| Kate | Logistic Regression | 49% | [Link] |

* This was the user who provided the original database.

## 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. Data

⬇

2. Train/Test Split

⬇

3. Data Preprocessing

⬇

4. ML Algorithm

⬇

5. Hyperparameter Optimization

⬇

6. Report Test Score

⬇

7. Repeat from 3

# 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