Predicting the Effects of News Sentiments on Stock Exchange.

Dataset Format:

Date News Article Headline Market Trend

Where market trend is:

- 1: when the market value rises or stays the same.
- 0: when the market value decreases.

Machine Learning Models:

Bag of Embeddings:

- Take the pre-trained word embeddings of all words in the news headline and average them.
- Take the average embedding of news headline and perform the linear transformation to predict the market trend:

Long Short Term Memory Networks (LSTMs):

- This method uses many-to-one architecture.
- Take the embeddings of news headline and pass the word embeddings to LSTM cells one by one in the sequence.
- Take the output of final cell and perform some activation function to predict the market trend.

Convolutional Neural Networks (CNNs):

- Take the word embeddings of all words in the news headline and stack them in matrix form where each row represents the word embedding of a word.
- Convolve the matrix with kernels of of size hxw where:
    - h : height of kernel represents no of words considering during a single convolution.
    - w : width of kernel is equal to the dimensions of word embedding vector (100/200/300)
- Perform max pooling or average pooling on the result of convolution.

- Finally apply some activation function to predict the market trend.

Available Dataset:

https://www.kaggle.com/aaron7sun/stocknews
There are two channels of data provided in this dataset:

1. News data: 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. (Range: 2008-06-08 to 2016-07-01).
2. Stock data: Dow Jones Industrial Average (DJIA) is used to "prove the concept". (Range: 2008-08-08 to 2016-07-01).

DataFormat:

1. **RedditNews.csv**: two columns The first column is the "date", and the 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. This data is modified as:
    a. *"1" when DJIA Adj Close value rose or stayed the same;*
    b. *"0" when DJIA Adj Close value decreased.*
3. **Combined_News_DJIA.csv**: To make things easier for my students, I provide this combined dataset with 27 columns. The first column is "Date", the second is "Label", and the following ones are news headlines ranging from "Top1" to "Top25".

Australian Stock Exchange Dataset (needs to be prepared).

A million news headlines from ABC News (Australian Broadcasting Corporation):

This contains data of news headlines published over a period of 17 years.
This includes the entire corpus of articles published by the ABC website in the given time range. With a volume of two hundred articles each day and a good focus on international news, we can be fairly certain that every event of significance has been captured here.
Digging into the keywords, one can see all the important episodes shaping the last decade and how they evolved over time. For example: financial crisis, iraq war, multiple elections, ecological disasters, terrorism, famous people, local crimes etc.
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/SYBGZL
Format: CSV ; Single File

1. **publish_date**: Date of publishing for the article in yyyyMMdd format
2. **headline_text**: Text of the headline in Ascii , English , lowercase

Start Date: 2003-02-19 ; End Date: 2019-12-31

Australian Stock Exchange Data (S&P/ASX 200)

https://au.finance.yahoo.com/quote/%5EAXJO/history/
Data format:
Date Open High Low Close Volume

The ‘A Million News Headlines’ and ‘Australian Stock Exchange Dataset’ can be combined together to analyze the effect of global events on the stock exchange.