Herald: Stock Movement Predict using NLP

Andong Ma, Angel Zhang, Denise Chen, Changsheng Yan

Simon Fraser University

Motivation

- Stock movement prediction has long attracted both investors and researchers. It is a challenging problem: the market is highly volatile and it is affected by numerous factors.
- Discover whether and how news articles and social media would affect the stocks market trend.

Goals

- Use NLP methods to do feature extraction on news and twitter dataset to predict stock price.
- Optimize the deep learning model with time-series dataset to predict the stock price trend or stock price
- Build an interactive platform to acquire latest news article, tweets, and stock price for several tickers and Nasdaq Index and predict with real-time processing.

Approach

- Scrape financial news and twitter dataset to find correlation with stock price using API and BeautifulSoup and total dataset size reaches 700 GB.
- Find topic distribution in news article through LDA and look for keywords in separate categories.
- Build baseline model with Linear regression, Naïve Bayes, Linear SVM, logistic regression, and random forest.
- Use Doc2Vec, and Glove to do feature engineering on news article and input words embedding to the deep learning models.
- Use time-series dataset to fit into RNN LSTM model with multiple approaches to optimize the model performance including Convolutional and Maxpooling layer.

Exploratory Data Analysis

Learning & Future Work

Learning:

- Data mining is open ended and LDA helps understand the latent topics of corpus.
- NLP methods like Doc2Vec, Glove to do feature engineering on word embedding can boost model performance.
- But difficult to measure the performance of embedding.

Future Work:

- Visualize high dimensional embedding vectors for Doc2Vec.
- Combine topic models for downstream work with unsupervised method like K-means clustering.
- Fine-tune and build more complex model to fit the embedding text data.

Pipeline

Data Collection and Integration

Storage: Json and pkl files

Data Sources: New York Times, Bloomberg, CNN, The Washington News, Reuters, The Guardian, BBC

APIs: Quandl, Alpha Vintage, NYT, The Guardian, Twitter

Time Frame: 2014-2019

Data Preprocessing Exploratory Data Analysis

Text Preprocessing:

- Tokenize, remove stop words, lemmatize

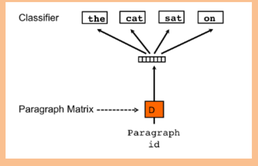
EDA:

- Plot the stock trend with the positive and negative news count
- Use LDA to do topic modeling on news content
- Find the news distribution in different news categories

Feature Extraction and Engineering

Methods:

- Bag of words into N-gram model
- Words embedding into Glove model on news headline
- Words embedding with Doc2Vec on news content
- Use normalized methods to transform stock price including MinMaxScaler



Predictive Modeling

Base Model (~52% Acc):

- Naïve Bayes, random forest, linear regression, and logistic regression

LSTM Classification Model (54.5% Acc):

- Trained a recurrent neural network (LSTM) to predict whether stock price rises or drops in next following days

LSTM Regression Model:

- Perform regression to predict the stock price in next few days given the sentiment on news headlines and minmax featured stock price

Visualization on User Interface

Web App:

- An interactive frontend with SpringBoot and Bootstrap
- Chart.js and HighCharts to plot the real-time Nasdaq market index and historical stock price
- Real-time predict by extracting the latest news and tweets related to stocks ticker

Model Plot and Web Interface