Indian Institute of Information Technology Vadodara
Government Engineering College, Sector-28, Gandhinagar,
Gujarat - 382028

Design Project Report

on

Statistical Learning Stock Price Predictor

Submitted by

Vikash Choudhary (201851144)
Himanshu Bhadu (201851048)
Ritik Rawat (201851102)
Noorul Hasan Ali (201851078)

under the supervision of

Dr. S.K. Patra

Abstract- We have tried to connect social media's influence with the stock market using machine learning algorithms. We took data from twitter and processed it to determine the positivity and negativity of the sentiment and then took it as one of the feature with the standard features to train the model.

INTRODUCTION

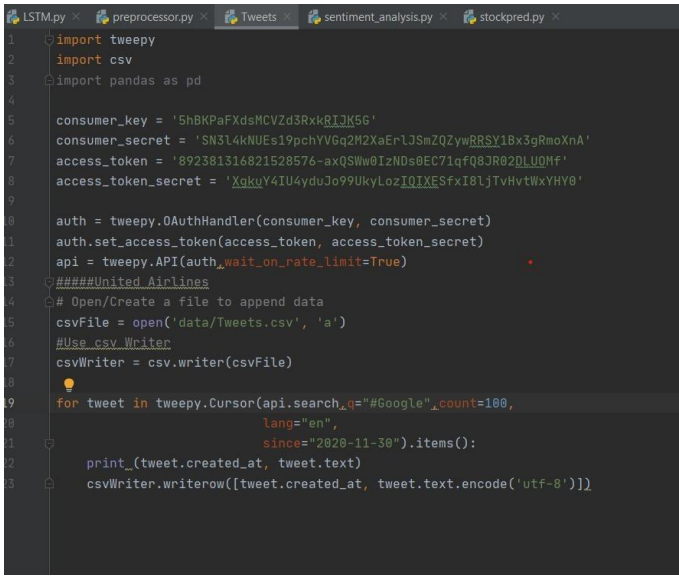
Prices of stocks are very much determined by the sentiments of the people and the news that is being circulated regarding the company and its policies.
Today social media is one destination where people share their views on the news circulated in the market. No matter what the news is people are often divided in support or against it but the majority is always on one side.
Thus, we have tried to analyse how this affects the stock prices, that is we compared the results by training and testing a data without considering the sentiments and with sentiments.
We have tried and tested various machine learning algorithms with the twitter API for various stocks (e.g. Google).

LITERATURE SURVEY

Through various research papers and blogs, we found that people have tried this but with different algorithms and techniques. Some of them had around 70% accuracy some 80% and we have tried to get more than this.

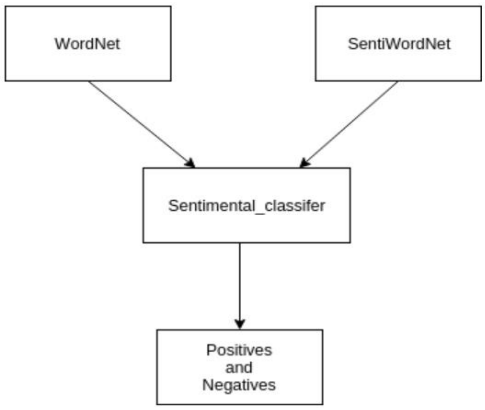
THE PRESENT INVESTIGATION

- We first analyzed how stocks get effected with different features like date, open, high, low, last, close, etc using machine learning algorithms such linear regression and random forest. We got error value of 1.325 and 0.868 respectively.
- We needed to add sentiments so for that we used Twitter. Collecting data of past one day in csv file we processed it using twee-py.



- Now to add features to include sentiments we used wordnet and senti-wordnet. We added two features negativity and positivity of a tweet.

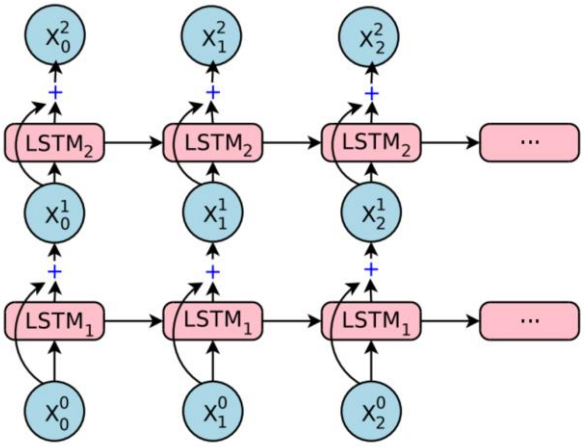
Indian Institute of Information Technology Vadodara
Government Engineering College, Sector-28, Gandhinagar,
Gujarat - 382028




```python
def clean_text(text):
    text = text.replace("<br />", " ")
    #text = text.decode("utf-8")

    return text


def swn_polarity(text):

    text = clean_text(text)

    positive_sent = 0
    negative_sent = 0
    raw_sentences = sent_tokenize(text)
    for raw_sentence in raw_sentences:
        tagged_sentence = pos_tag(word_tokenize(raw_sentence))

        for word, tag in tagged_sentence:
            wn_tag = penn_to_wn(tag)
            if wn_tag not in (wn.NOUN, wn.ADJ, wn.ADV):
                continue

            lemma = lemmatizer.lemmatize(word, pos=wn_tag)
            if not lemma:
                continue

            synsets = wn.synsets(lemma, pos=wn_tag)
            if not synsets:
                continue
```

```python
import nltk
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('wordnet')

from nltk.stem import WordNetLemmatizer
from nltk.corpus import wordnet as wn
from nltk.corpus import sentiwordnet as swn
from nltk import sent_tokenize, word_tokenize, pos_tag
import pandas as pd

lemmatizer = WordNetLemmatizer()


def penn_to_wn(tag):

    if tag.startswith('J'):
        return wn.ADJ
    elif tag.startswith('N'):
        return wn.NOUN
    elif tag.startswith('R'):
        return wn.ADV
    elif tag.startswith('V'):
        return wn.VERB
    return None


def clean_text(text):
    text = text.replace("<br />", " ")
    #text = text.decode("utf-8")
```

- Along with the previous features, we encoded each tweet using a variation of the Word2Vec model. we took weighted averages of the tweets belonging to one-time span.
- Now features we tried various algorithms but got best results with SLTM.



- Optimization: rmsprop
  Epochs: 500
  Activation: tanh

Result And Discussion

Applying various techniques with different algorithms we were able to get better results than previous research papers. on an average we had around 90% accuracy with the actual stock prices.

Conclusions and Future Work

We found that public sentiment highly influences the behaviour of stocks and social media plays huge part in it. we were able to predict the prices of the stocks, not highly accurate but at least the behaviour.

Data for a longer period and multiple companies can be used to improve accuracy and an application can be developed which can help people to stay alert and reduce the risk with trade.

Indian Institute of Information Technology Vadodara
Government Engineering College, Sector-28, Gandhinagar,
Gujarat - 382028

ACKNOWLEDGMENT

The completion of this project would not have been possible without the guidance of our mentor **Dr. S.K. Patra**. He not only helped us with his knowledge but also connected to other people who had more knowledge in this field. We would like to thank him for his support and time.
We would also like to thank IIIT- Vadodara for this wonderful opportunity.

REFERENCES

[1] Illustrated Guide to LSTM's and GRU's: A step by step explanation | by Michael Phi | Towards Data Science
[2] sigproc-sp.dvi (stanford.edu)
[3] A Python script to download all the tweets of a hashtag into a csv · GitHub.
[4] https://towardsdatascience.com/machine-learning-for-stock-prediction-a-quantitative-approach-4ca98c0bfb8c