Stock Analysis and Forecasting based on Airline Reviews

Milestone 1

By: Prajakta Gaydhani
Advisor: Dr. Carol Romanowski

Problem Statement

Opinion Mining of different customer reviews or feedback on major Airlines in the United States and analyzing the effect of customer's sentiments on the stock market of different airline companies.

Goal of the Project

- Correlation between customer reviews and stock price movement of different Airline Companies.
- Accurately forecast stock prices for different Airline Companies using customer reviews and stock market data.

Milestone 1 : Data Collection

Collected customer reviews and stock market data from Web

- Airlines: American, Delta, United, JetBlue and Alaska
- Dataset size: ~ 1 Million
- Collected per day data from Jan 2010 - Sept 2018.

Customer Reviews

Date	Tweets
1/1/18	""Delayed again? Seriously? You keep getting worst # AmericanAirlines @ AmericanAir""
1/2/18	""Delayed because of staffing... missing our connection which is last of the day... # lame # AA # americanairlines""

Stock Data

Date	Open	High	Low	Close	Adj Close
1/1/18	21.049999	21.4	19.1	19.299999	18.540155
1/2/18	19.299999	20.530001	19.200001	20.5	19.692907

Milestone 1 : Data Sources

~ 1M tweets

Twitter API

Used Python and Tweepy API to retrieve tweets

~ 80,000 reviews

Used Scrapy framework in Python and Selenium webdriver to extract feedbacks

~ 25,000 reviews

Got the dataset from data.world

From Jan 2010-Sept 2018

Used Ticker Symbols and Yahoo! Finance API to retrieve stock data

Milestone 1 : Data Pre-Processing

Cleaning Customer reviews

Used Regex in Python

a. Eliminated URLs, @usernames, emoticons, punctuations and special symbols
b. Removed #hashtags
c. Words like *haaapppy* are converted to *happy*
d. Contractions like *don't*, *should've* are replaced with *do not, should have*
e. Removed any additional whitespaces

Milestone 1 : Natural Language Processing

Used NLTK library in Python and WordNet lexical database of English

a. Tokenization
b. Removing stop-words
c. Lemmatization : converted words to their root forms based on WordNet lexical database
   - studies, studying ⟶ study
d. Parts of Speech Tagging (POS tagger)

Milestone 1 : Challenges Faced

- Limitations in collecting tweets from Twitter
  - Difficulty in getting tweets older than 7 days
  - Needed historical data for analysis
  - Made changes in the Tweepy search API to overcome this limitation
  - Rate limiting error for making too many requests
- Large amount of time was required for crawling data from web

Future Milestones

- **Milestone 2**
  - Opinion Mining of Customer reviews using dictionary based approach
  - Aggregating the daily sentiments scores
  - Integrating the sentiments with stock market data based on timestamp
  - Smoothing stock market data to remove seasonality and trends
- **Milestone 3**
  - Building conventional machine learning (Naive Bayes, Regression) and Deep learning models (LSTM/RNN)
  - Correlating stock market movement with sentiments scores
  - Evaluating the results

Thank You