SENTIMENT ANALYSIS OF STOCK-RELATED NEWS ARTICLES

Project 2

Students: *Albert Kang, Guanru Peng, Peilin Wang, Zhaohui Man*
Site Supervisors: *Sabarish Karuppiah, Tassos Sabane, Anthony Sourial*
Faculty Advisors: *Ray Wang, Jaime Arguello*

Problem Statement

For project 2, we investigate how to use articles relating to suspended stocks and correlating historical stock price to predict sentiment (negative, neutral, positive)

- This will allow analysts to search relevant article more efficiently and spend more time to understand the stock positions
- Provides a methodology for analysts to rank news articles by importance and determine which ones to focus on
- Stock sentiments could provide more insight to capture the volatility of the stock price shed more information on the corporate derivative portfolio

Data Acquisition & Labeling

DATA ACQUISITION: Get the list of tickers——Use google news rss as the entrance of scraping——Create specific scraper for Bloomberg——Scrape according to the target sites using python library ‘newspaper 3k’ and ‘selenium’

LABELING: Scrape Yahoo Finance for historical data for each ticker

↓

For each article: i. Obtain 10 days worth of stock closing price prior to article publish date, find mean

ii. Obtain closing price of stock for the day after article publish dateawethc

iii. Find the percent change of the final closing price compared to the mean

↓

percent change > 4% positive

-4% < percent change < 4% neutral

percent change < -4% negative

Model Building:Try Everything

FEATURE EXTRACTION: TfIdfVectorizer, **CountVectorizer**, WordVEC, Doc2VEC, fastText, Pre trained GloVe, Pre Trained Google News Word2Vec, LSTM, SpaCy

N-gram: Unigram, Bigram, Trigram and **combination of them**

CLASSIFIERS: **Logistic Regression**, Linear SVC, LinearSVC , Multinomial NB, Bernoulli NB, Ridge Classifier, AdaBoost, Extreme Boosting Gradient, Perceptron,Passive-Aggressive, Nearest Centroid, SVM, Randomforest, Ensemble, NN, CNN, RNN

DIMENSION REDUCTION: **CHI2**, PCA

WINNER

Part of the Results (chi2, trigram)

	Auto Labelling						Manual Labelling (ten articles)					
	LR	RF	SVM	Ens	CNN	RNN	LR	RF	SVM	Ens	CNN	RNN
CountVectorizer	**0.65**	0.56	0.51	0.59	0.53	0.49	0.51	0.46	0.49	0.48	0.38	0.37
TfIdfVectorizer	0.60	0.55	0.48	0.58	0.54	0.48	0.50	0.48	0.47	0.45	0.37	0.35
W2Vec	0.53	0.49	0.48	0.55	0.49	0.45	0.44	0.43	0.47	0.46	0.40	0.41
Glove	0.52	0.51	0.49	0.53	0.48	0.44	0.46	0.47	0.42	0.48	0.33	0.39
LSTM	0.58	0.55	0.53	0.59	0.53	0.46	0.56	0.49	0.50	0.51	0.40	0.42
spacy	0.60	0.58	0.55	0.60	0.55	0.49	**0.61**	0.50	0.51	0.52	0.41	0.43

Limitation

1. Labeling Method: A Compromise
- Manually labeling is time-consuming; lexicon-based approach cannot guarantee labeling accuracy
- Data loss: 30,000+ scraped news articles → 7000 after preprocessing and labeling
- For future work, manually label articles (or use labeling services e.g. Amazon Mechanical Turk)

2. Scraping Method
- Newspaper3k cannot retrieve the content of every news article via the url fetched from Google News RSS
- Impossible to build one scraper for each website - restriction on scraping

3. Modeling
- Insufficient data
- For future work, deploy models on cloud computing platform e.g. AWS to increase computing power

Thank You

Please feel free to ask questions.