#### Who drives the market? Sentiment analysis of financial news posted on Reddit and the Financial Times

Bachelor's Thesis by Michael Lubitz

January 23, 2018

### Overview

- Motivation
- Data Sources
  - Reddit
  - The Financial Times
  - Data Retrieval
- Experimental Set-up
- Sentiment Analysis
  - Using a dictionary
  - Using machine learning
- Evaluation

- Researchers found out that the sentiment of financial news articles have a certain power to predict falling or rising stock indices<sup>1</sup> as they are the main source for investors<sup>2</sup>
- The sentiment of social media posts (especially on twitter) has also been proven to have a even higher accuracy of predictions<sup>3</sup>
- Reddit combines both worlds

<sup>&</sup>lt;sup>1</sup> R. P. Schumaker and H. Chen, »Textual analysis of stock market prediction unsing breaking financial news«

<sup>&</sup>lt;sup>2</sup> X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, »News impact on stock price return via sentiment analysis«

<sup>&</sup>lt;sup>3</sup> J. Bollen and H. Mao, »Twitter mood as a stock market predictor«

- Creation of the Reddit dataset
  - Posts and news articles
- Creation of the Financial Times dataset
- Find suitable models for our predictions
- The task of this thesis was to compare the accuracy of predictions based on Reddit to the results of a classic news paper analysis



- Founded 2005
- Social News Aggregator
  - Each registered user can share links with the community (submission)
  - Other users can comment and vote on the relevance and importance of a submission

# Subreddits 1/2







## Subreddits 2/2

nie Reddit Baread of Decitorine Redet		
NEW RISING CONTROVERSIAL TOP GILDED WIKI	Want to join? Log in or sign u	up in secon
	search	
GoJS JavaScript diagrams. Dozens of samples to get your project started. (gojs.net)	username	
promoted by Northwoods_Software		
	password	
2017 State of the Subreddit Survey (self.Economics)	remember me reset passwo	ord
Image: Submitted 18 days ago by Interret [M] - announcement           16 comments           share save hide report		-
r/Fconomics Discussion Thread - 12 January 2018 (self Economics)		No.
submitted 7 days ago by AutoModerator [M] - announcement		-
24 comments snare save nide report		
The East India Company: How a trading corporation became an imperial ruler (historyextra.com)		
<ul> <li>Submitted 13 hours ago by aolataoldotcom</li> <li>6 comments share save hide report</li> </ul>	Neue Jobs in F	reibur
Polling back regulations often comes before a financial meltdown, according to the IME - Quartz ()	Wer hier nicht such selber schuld!	nt, ist
2 submitted 22 hours ago by data2dave	Submit a new	link
18 comments share save hide report	subscribe	
After All the Talk About a Skills Shortage in the U.S. Job Market, the Real Problem May Be an Employer Shortage (slate.com)	421,531 reader 228 users here n	rs 10W
submitted 21 hours ago by jonfla 7 comments share save hide report	Subreddit Bules	
In Venezuela, money has stopped working (werebingteneget over)	1. Discipline-Specific News, F	Research, & V
bubmitted 1 day ago by KareIIen	/r/Economics concerns prolife discourse pertaining to resea	eration and arch, news, aca
113 comments share save hide report	work, and academic summari perspective of economists.	ies from the
Births in China FALL despite it relaxing the one-child policy (dailymail.co.uk)	2. Economic-Specific Quality Submissions and/or commen	Standards nts tenuously r
submitted 5 hours ago by Bastet1 1 comment share save hide report	to economics, light on econor perspectives other than those be removed	mic analysis, e e of economis
Bank of Canada: Modeling Eluctuations in the Global Demand for Commodities (PDE) (Indexest - )	3. Source Standards	
submitted 16 hours ago by Central_Bank_Bot	Submissions must be from or Editorialization, memes, and	low-quality blo
comment share save hide report	acceptable. Source spam	iming is not
Common Property bostonreview.net (bostonreview.net)	4. User Benavior Standards Personal attacks and harassi	ment will resu
submitted 11 hours ago by hopeLB 3 comments share save hide report	removal of comments; multipl result in a permanent ban. Pl attacks, racism, miscourse, or	e infractions i lease report per
	or experience.	marassment )
[IO and Market Concentration] Craft Beer Is the Strangest, Happiest Economic Story in America (theatlantic.com) submitted 22 hours ago by gauchnomics	Images, charts, and/or video:	s including o
1 comment share save hide report	summary. Standards for multi left to moderator discretion	imedia are str
The Tether Conundrum: A look into how Bitcoin's price may currently be manipulated through counterfeit cryptocurrency (tonyarcieri.c	6. Off-Topic Discourse Guidel	lines
submitted 20 hours ago by bascule	Comments consisting of mere political comments, circleierk	e jokes, nakeo ling. or otherwi

Economics subreddit https://reddit.com/r/economics

# Why Reddit? 1/2

- Many different news from many different sources
  - Most articles of trustworthy newspapers
  - The quality of a source is essential for good predictions
- A news text has to be business related
  - We have chosen the subreddit /r/economics
  - The moderators of this subreddit ensure that there are sources related to this topic only

# Why Reddit? 2/2



<sup>1</sup> C. Buntain and J. Golbeck, »Identifying social roles in reddit using network structure«

#### Stock Index: S&P 500

- One of the most important stock indices worldwide
- Grounded on 500 companies listed on American stock exchanges
- According to Reddit 54% of its visitors come from the United States



Source: finance.yahoo.com

# Framework



### **Data Retrieval**



- Python crawler
- Time period: January 2008 to July 2017

#### Financial Times:



## Preprocessing 1/3

- 1. Remove special characters from a news text (from now on called document *d*)
- 2. Remove common stop words (e.g. and, the, for)
- Convert text to a list of words (bag of words) / a feature vector

Problem: The natural language has different word forms and linked words which is not respected by list of words and feature vectors

# Preprocessing 2/3

- Noun phrases:
  - We used OpinionFinder to extract nouns from the documents
  - Idea: Only use nouns to reduce noise (especially machine learning)
- Stemming:
  - Not used in combination with machine learning
  - Stemming reduces words to their stem (get rid of different word forms)
  - We used the Porter 2 stemming algorithm (Python package stemming) on documents and dictionary

# Preprocessing 3/3

- Doc2Vec:
  - Used in combination of machine learning
  - Problem with bag of words approach: Words are usually related or linked to other words in a text
  - Doc2Vec tries to find such related words and stores them as vectors
    - Example: Paris <-> France; Paris <-/-> desert
  - Python package gensim

# Labeling with dictionaries

- Each word in the dictionary is labeled with either positive or negative, or in a numeric representation +1 or -1
- For each document d we count the number of positive words p and negative words n

Word	Sentiment	
accomplish	positive	
effective	positive	
perfect	positive	
bankruptcy	negative	
failure	negative	

Dictionary example

Then, we calculate a sentiment score s<sub>d</sub> as follows:

 $s_d = (p - n) / (p + n)$ ;  $-1 \le s_d \le 1$ 

If s<sub>d</sub> < 0 we consider document d as negative and otherwise

# Labeling with machine learning

- Naive Bayes and Random Forests work with probabilistic models to classify documents either as positive or negative
- Support Vector Machines (SVM) is a large margin classifier which uses a decision boundary to separate the documents into positive and negative
- Training set:
  - Collection of already labeled financial news texts
- The sentiment score will be either -1 or +1

- Weighting the sentiment value s<sub>d</sub> of a document with the
  - 1. Number of votes v
  - 2. Number of comments c
  - 3. Number of votes and comments
- Multiply the sentiment value with max{log(v+1), v<sub>min</sub>} or / and

max{log(c+1), c<sub>min</sub>}

where v<sub>min</sub> and c<sub>min</sub> are minimum weights (found through optimization)

# Scoring

- Not in combination with machine learning
- Use of BM25 scoring to put emphasis on the relevance of words
- We calculate the BM25 score for each word in a document d that also appears in the dictionary
- Determine the sentiment of d (positive or negative) as before and replace the sentiment score with the BM25 score (positive or negative)

- Group all documents by its publishing date
- Sum up all sentiment scores of a day and normalize it by the number of documents on that day



#### Results 1/2

Baseline (guessing based on majority class): 54.20%

- Reddit ———

	Scoring	Financial Times	Unweighted	Votes weighted	Comments weighted	Votes and comments weighted
Bag of words	None	54.19	54.59	54.68	54.80	54.76
	BM25	54.83	54.76	54.30	54.26	53.84
Stems	None	54.20	54.39	54.47	54.43	54.55
	BM25	54.20	54.30	54.30	54.30	53.84
Nouns	None	54.24	54.45	<b>55.45</b>	<b>55.56</b>	<b>55.56</b>
	BM25	54.02	<b>54.83</b>	54.41	54.52	54.41

Sentiment Analysis based on a dictionary

#### Results 2/2

Baseline (guessing based on majority class): 54.20%

Reddit ——

	Feature representation	Financial Times	Unweighted	Votes weighted	Comments weighted	Votes and comments weighted
Naive Bayes	Bag of words	54.20	54.85	54.43	54.43	54.72
	Nouns	54.24	55.24	<b>55.45</b>	<b>56.39</b>	<b>56.49</b>
	Doc2Vec	45.80	46.27	46.48	46.23	46.31
Random Forest	Bag of words	54.20	54.51	54.42	54.43	54.43
	Nouns	54.24	<b>55.45</b>	<b>55.45</b>	55.45	55.45
	Doc2Vec	47.89	50.50	51.13	51.93	51.63
Support Vector Machines	Bag of words Nouns Doc2Vec	54.20 54.24 54.20	54.85 55.42 53.86	55.02 <b>55.45</b> 53.77	54.89 55.45 54.07	54.97 55.56 53.52

Sentiment Analysis based on a machine learning

#### Conclusion

- Almost all of our results outperform baseline and Financial Times
- Therefore we state that Reddit has a certain power to predict stock index changes
- Generally the use of votes and comments for weighting purposes is reasonable