Capstone Project
Machine Learning Engineer Nanodegree

Michael Boker
July 9th, 2017

Stock Market Predictions Based on News Headlines

Definition

Project Overview

During the early history of the stock market, stock traders made investment decisions based primarily on qualitative information. This information included reputations of entrepreneurs and employees at companies, products being produced by companies, and current events throughout the world. Since the 1950's, quantitative analysis has played an increasingly important role in these investment decisions[1]. Quantitative analysis involves mathematical models used to find predictive trends in the market, which are used to assist in making wise investment decisions. Computers are much more efficient than humans at processing numerical information, so they have been quite useful for quantitative analysis. Computers have been so dominant in this space, that many of the trader middle-men have been cut out, and trades have been automated. Some estimates say that 70 percent or more of trades are made automatically by computers, with no human involvement[2]. If we can use the qualitative information used by many human traders to augment the quantitative information currently used by the machines, we may be able to create more accurate prediction systems and thus more lucrative automated traders.

In this project, we will create a classifier which uses news headlines to predict behavior of the stock market. The classifier will be trained on a dataset from kaggle[3]. This dataset pairs the top 25 news headlines for a day with a binary value indicating the behavior of the Dow Jones Industrial Average (DJIA) for that day. The headlines were gathered by scraping the World News subreddit, taking the 25 most upvoted headlines. If the binary value is 1, the value of the index remained the same or rose; if the value is 0, the value of the index fell. This dataset contains points for just under 8 years.

Problem Statement

The goal is to create a classifier model, which will predict whether the DJIA will rise or fall on a day, based on the top news headlines for that day. The strategy will be to create a recurrent neural network for sentiment analysis. Because we are doing a binary classification, we can treat a label of 1 (the index rose or stayed the same) as a positive sentiment, and a label of 0 as a negative sentiment. Each of the headlines for a day will be concatenated into a single string, so the problem will be similar to predicting the sentiment of a message.

Evaluation Metrics

For evaluating the performance of the model, a combination of prediction accuracy and F score will be used. Accuracy will indicate the ratio of correct to incorrect predictions, but can be misleading if the data is skewed. F score seems to be a good metric for ensuring that the model truly learned from the

[1] McWhinney, James E. "A Simple Overview of Quantitative Analysis" Investopedia

[2] Salmon, Felix and Stokes, Jon "Algorithms Take Control of Wall Street" WIRED

[3] User Aaron7sun "Daily News for Stock Market Prediction" Kaggle

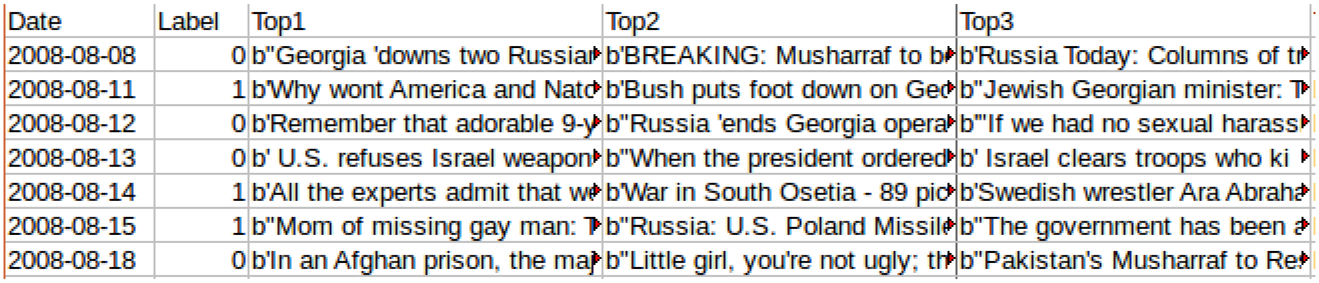
data, and did not just learn how to cheat from the data. If the values in the training data skew towards one label or the other, and the model learns to be biased towards that label, it will receive a low overall F score.

Preventing a learned bias is important for a task like stock market predictions. If the model learns to favor one prediction over another due to a skew in the data, this could be disastrous when the model is used on new data that may not be skewed the same way.

Analysis

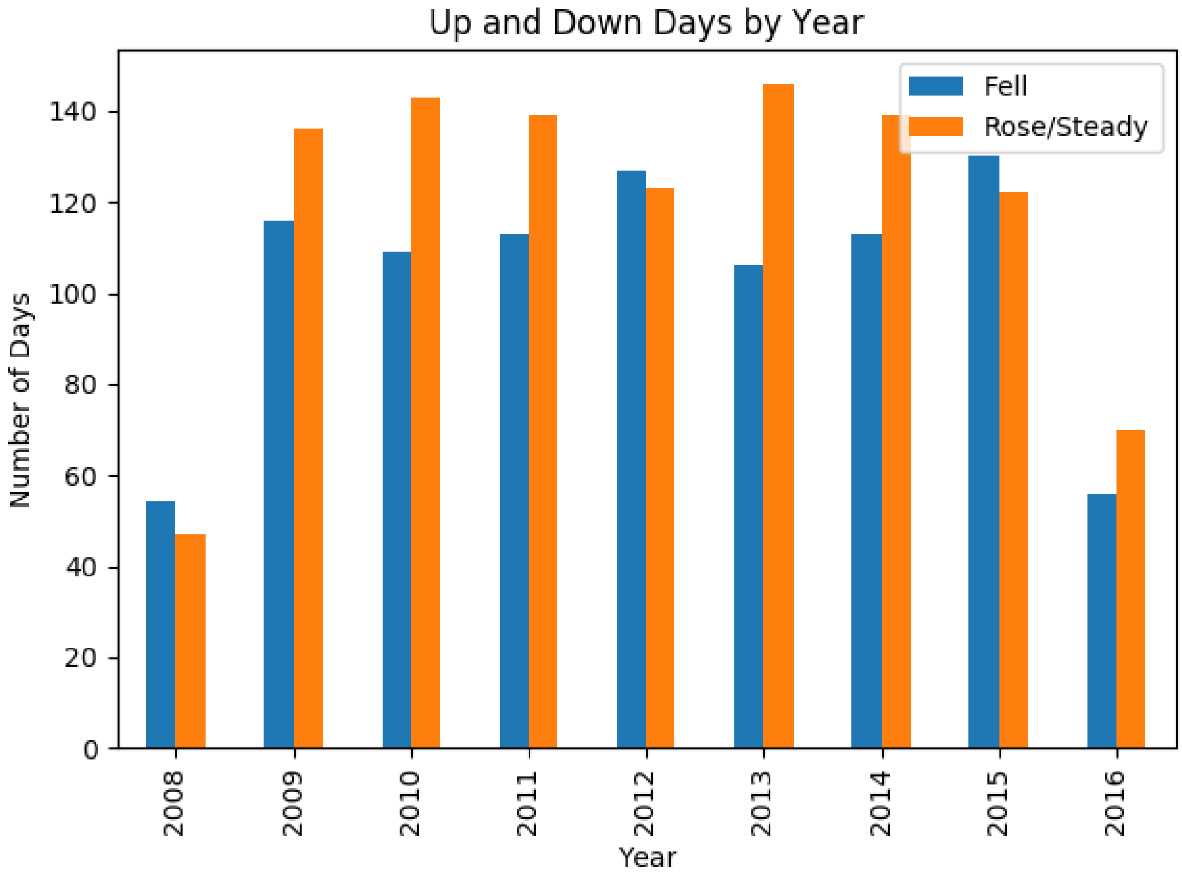
Data Exploration and Visualization

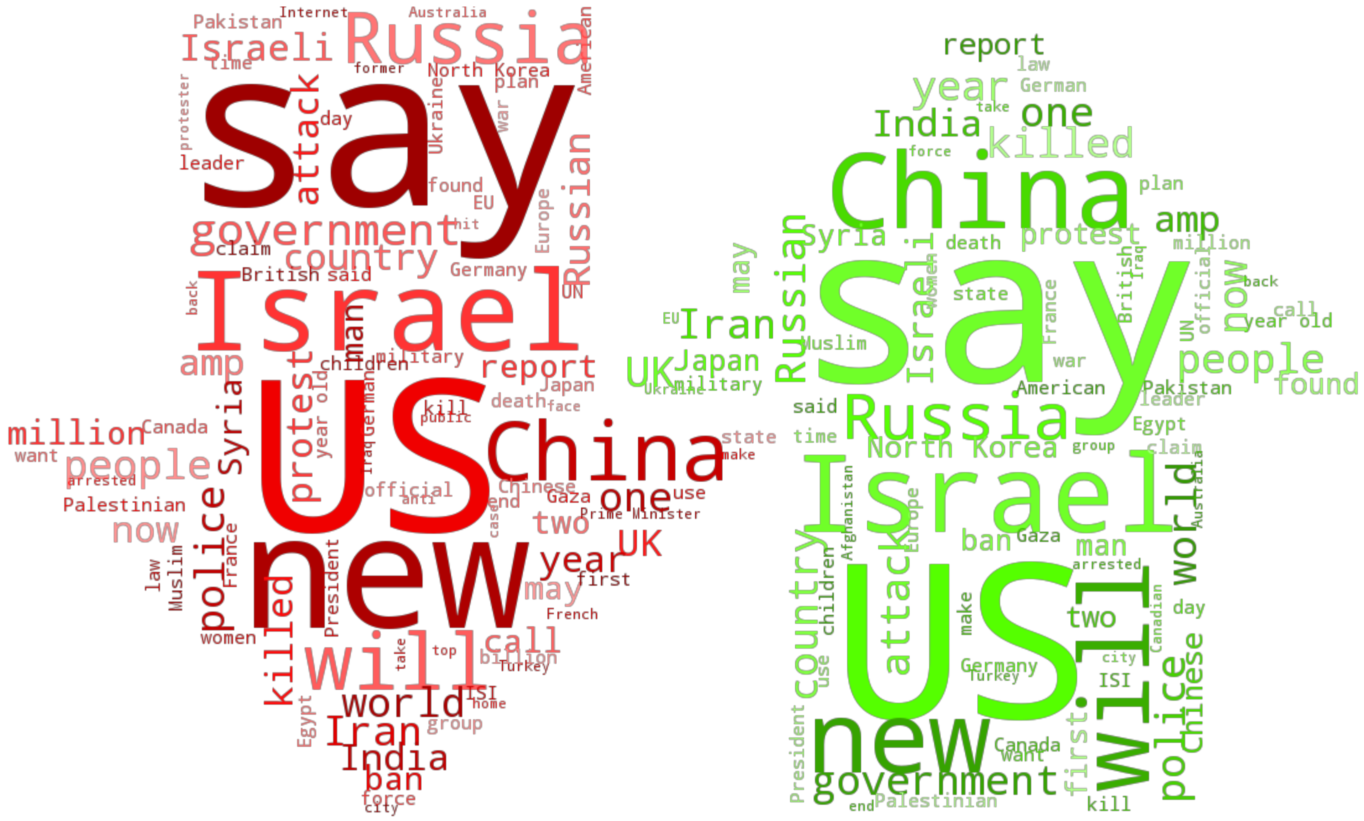
The dataset consists of 1889 points, each with 27 features. Each point corresponds to a day. The first feature, named “Date,” is the date for that day, and the second feature, named “Label,” is the binary value indicating the performance of the DJIA that day. The remaining 25 features are “Top1” through “Top25,” and they contain the top 25 headlines for the day, in descending order of popularity. Below is a snippet of the spreadsheet containing this data, showing 7 points, and the first 5 features, with each of the headline features abbreviated to fit on the page.



Date	Label	Top1	Top2	Top3
2008-08-08	0	b"Georgia 'downs two Russiar	b'BREAKING: Musharraf to b	b'Russia Today: Columns of tr
2008-08-11	1	b'Why wont America and Natc	b'Bush puts foot down on Gec	b"Jewish Georgian minister: T
2008-08-12	0	b'Remember that adorable 9-y	b"Russia 'ends Georgia opera	b'"If we had no sexual harass
2008-08-13	0	b' U.S. refuses Israel weapon	b"When the president ordered	b' Israel clears troops who ki
2008-08-14	1	b'All the experts admit that we	b'War in South Osetia - 89 pic	b'Swedish wrestler Ara Abraha
2008-08-15	1	b"Mom of missing gay man: T	b"Russia: U.S. Poland Missile	b"The government has been a
2008-08-18	0	b'In an Afghan prison, the maj	b"Little girl, you're not ugly; th	b"Pakistan's Musharraf to Re

Table 1: Snippet of Dataset


Up and Down Days by Year
Fell
Rose/Steady
Number of Days
0
20
40
60
80
100
120
140
Year
2008
2009
2010
2011
2012
2013
2014
2015
2016


Pakistan
Internet
Australia
Israeli
Russia
time
former
North Korea
plan
American
say
protester
day
attack
Ukraine
war
Russian
leader
found
EU
government
hit
Europe
claim
country
British said
Germany
back
Israel
UN
amp
man
children
military
report
Japan
US
protest
year old
Iraq German
kill
public
death
face
million
Canada
want
people
Syria
arrested
state
make
China
official
anti
Chinese
Palestinian
end
Gaza
one
use
police
case
Prime Minister
two
now
France
President
law
Muslim
new
UK
year
may
first
French
killed
women
will
call
take
top
billion
Turkey
ISI
home
world
group
Egypt
Iran
India
ban
force
city
report
law
year
German
take
India
one
force
killed
China
plan
amp
Syria
death
protest
million
may
Russian
say
Israeli
women
state
British
Iraq
back
official
now
call
EU
Iran
France
UN
year old
Muslim
UK
Japan
war
people
Ukraine
military
American
Pakistan
found
said
leader
Russia
Egypt
time
North Korea
group
claim
make
Australia
Israel
Afghanistan
world
Europe
ban
Gaza
man
country
children
arrested
attack
make
US
day
two
use
Canadian
city
Germany
Turkey
ISI
will
Chinese
police
President
new
Canada
want
first
government
end
Palestinian
kill