

THE USE OF NLP AND DEEP LEARNING ON NEWS HEADLINES TO PREDICT STOCK MOVEMENTS

Tayfun Ayazma, Ph.D.

May 20, 2020



Outline

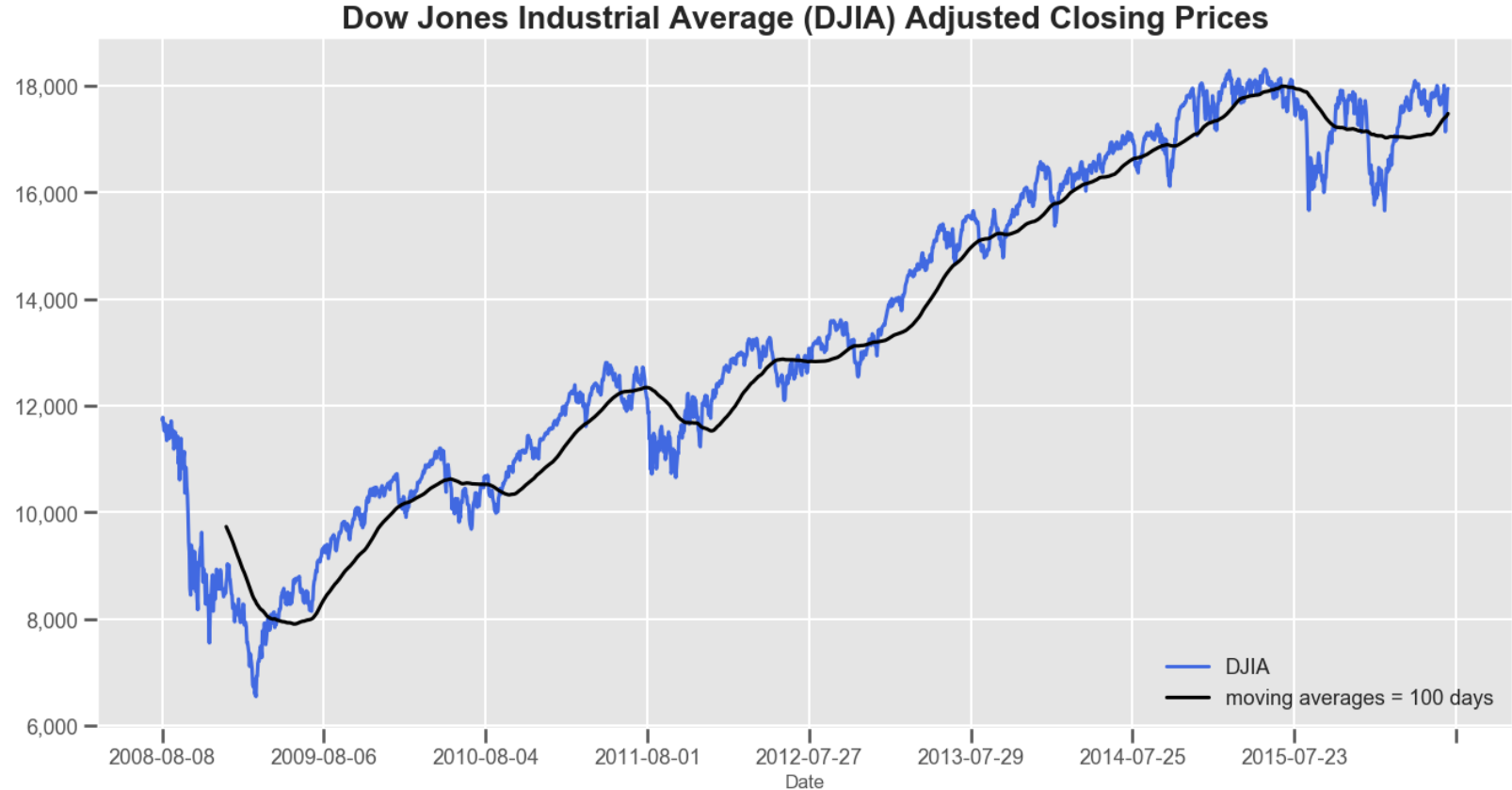
- I. Problem Statement
- II. Dow Jones Industrial Average (DJIA)
- III. Data
- IV. Data Preparation and Text Preprocessing
- V. Model Building
- VI. Most Important Features
- VII. Results
- VIII. Business Recommendations

Problem Statement

- ❑ Vast amount of resources dedicated to quantify and analyze qualitative data
- ❑ Complexities of modeling stock market dynamics
- ❑ A variety of resources:
 - ❑ SEC reports (8-K and 10-K)
 - ❑ Financial news articles (Financial Times and Wall Street Journal)
 - ❑ Media Outlets (Twitter, Facebook, and Reddit)
- ❑ NLP and Deep Learning to predict Dow Jones stock movements

Dow Jones Industrial Average (DJIA)

- ❑ Stock market index
- ❑ Tracks the stock performance of 30 large companies
- ❑ Components of DJIA



Data

- ❑ Reddit historical news headlines
 - ❑ Only the top 25 news headlines for a single date
- ❑ DJIA stock data downloaded from the Yahoo Finance
- ❑ Target feature - the “Label” column:
 - ❑ 1 if DJIA adjusted close value stayed the same or went up
 - ❑ 0 if it dropped

Data Preparation and Text Preprocessing

- ❑ Inspect for quality (outliers, missing value, data types)
- ❑ Preprocessing the textual data (NLTK library):
 - ❑ Lowercasing
 - ❑ Removing stopwords, punctuation and numbers
 - ❑ Lemmatizing words
- ❑ Combine text columns
- ❑ Splitting the data into train and test sets:
 - ❑ Train set → 08-08-2008 to 12-31-2014 (80%)
 - ❑ Test set → 01-01-2008 to 07-01-2016 (20%)

Model Building

☐ Six supervised machine learning models:

- ☐ Logistic Regression
- ☐ Multinomial Naïve Bayes
- ☐ Random Forest Classifier
- ☐ Support Vector Machine Classifier
- ☐ XGBoost Classifier
- ☐ Bernoulli Naïve Bayes

☐ Deep learning neural network architectures:

- ☐ A fully connected multilayer perceptron (MLP)
- ☐ Recurrent neural network (RNN) – LSTM
- ☐ Convolutional neural network (CNN)

Most Important Features


	Word	Coefficient
6042	nigeria	0.838646
4705	jew	0.713932
7247	record	0.664518
8131	since	0.656317
8474	state	0.654680
7556	right	0.653375
5825	mubarak	0.640013
7963	set	0.621422
8857	tear	0.617964
7852	scrap	0.600945

	Word	Coefficient
4201	hour	-0.684081
5633	military	-0.702586
3852	hacking	-0.703074
9798	without	-0.739373
2014	criminal	-0.761785
5259	low	-0.770794
7750	sanction	-0.794457
1970	country	-0.804450
4595	iran	-0.811626
7670	run	-0.820194

- ❑ For logistic regression model, the most important features both positive and negative
- ❑ “Military”, “Hacking”, “Criminal” , “Low” and “Sanction”

Results

- ❑ RNN – LSTM network with a 58% accuracy
- ❑ Model tuning for hyper-parameters for LSTM and SVM Classifier
- ❑ The complexity of modeling stock market dynamics
- ❑ Limited data

	Accuracy	ROC_AUC	Precision	Recall	F1
Model					
 LSTM	0.582011	0.581989	0.589474	0.583333	0.586387
SVM Classifier	0.537037	0.533350	0.530686	0.765625	0.626866
MLP	0.515873	0.514701	0.520737	0.588542	0.552567
Logistic Regression	0.515873	0.509241	0.512968	0.927083	0.660482
XGBoost	0.510582	0.508233	0.514286	0.656250	0.576659
Multinomial Naive Bayes	0.507937	0.500000	0.507937	1.000000	0.673684
BernoulliNB	0.502646	0.495800	0.505682	0.927083	0.654412
CNN	0.497354	0.494792	0.504000	0.656250	0.570136
Random Forest Classifier	0.484127	0.479503	0.494983	0.770833	0.602851

Business Recommendations

- ❑ Textual data to provide informed decisions
- ❑ The potential of deep learning to revolutionize the stock market trading
- ❑ Changes in stock movements in days following the release of news headlines

Questions?