Stock Sentiment Analysis using News Headlines

Shibasish Shaw
212123053
shibasishshaw@iitg.ac.in

Indian Institute of Technology Guwahati

April 29, 2023

Table of Contents

1. Introductions
2. The Dataset
    - Data Preprocessing
    - Exploratory Data Analysis
    - Converting the Textual Data into Vectors
3. Applying Random Forest
4. Applying Naïve Bayes
5. Applying Convolutional Neural Network
6. Conclusion

Table of Contents

1. Introductions
2. The Dataset
   - Data Preprocessing
   - Exploratory Data Analysis
   - Converting the Textual Data into Vectors
3. Applying Random Forest
4. Applying Naïve Bayes
5. Applying Convolutional Neural Network
6. Conclusion

Sentiment Analysis

Sentiment analysis is a technique that involves analyzing a piece of text to determine its emotional tone, which could be positive, negative, or neutral. It is commonly performed various ML algorithms. It has a variety of applications in fields like social media monitoring, customer feedback analysis, and brand reputation management.

Sentiment Analysis and Stock Price Variation

Fluctuations in a stock's price can be caused by several factors, one of which is sentiment. Sentiment is typically influenced by news related to the company, such as positive or negative earnings reports, product launches, missed targets, or the departure or death of a key figure. These events can affect the demand and price of shares in the stock market.

The Objective

The main aim of this project is to demonstrate how the price of a stock fluctuates as a result of relevant human sentiment using Machine Learning algorithms. Also, we have performed the experiments with and without the preprocessing part for the textual data.

Table of Contents

1. Introductions

2. The Dataset
    - Data Preprocessing
    - Exploratory Data Analysis
    - Converting the Textual Data into Vectors

3. Applying Random Forest

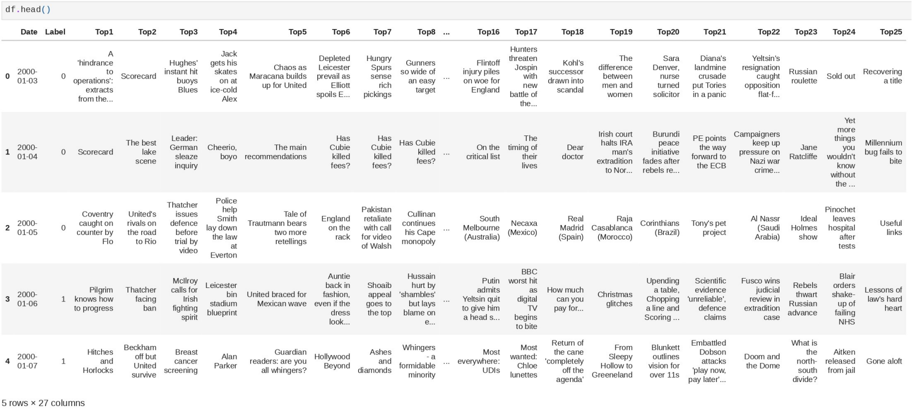
4. Applying Naïve Bayes

5. Applying Convolutional Neural Network

6. Conclusion

How the Dataset looks like

Let us have a look at the dataset that has been taken from Kaggle: `https://www.kaggle.com/code/rohit0906/stock-sentiment-analysis-using-news-headlines`. [1]



```
df.head()
```

	Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	...	Top16	Top17	Top18	Top19	Top20	Top21	Top22	Top23	Top24	Top25
0	2000-01-03	0	A 'hindrance to operations': extracts from the...	Scorecard	Hughes' instant hit buoys Blues	Jack gets his skates on at ice-cold Alex	Chaos as Maracana builds up for United	Depleted Leicester prevail as Elliott spoils E...	Hungry Spurs sense rich pickings	Gunners so wide of an easy target	...	Flintoff injury piles on woe for England	Hunters threaten Jospin with new battle of the...	Kohl's successor drawn into scandal	The difference between men and women	Sara Denver, nurse turned solicitor	Diana's landmine crusade put Tories in a panic	Yeltsin's resignation caught opposition flat-f...	Russian roulette	Sold out	Recovering a title
1	2000-01-04	0	Scorecard	The best lake scene	Leader: German sleaze inquiry	Cheerio, boyo	The main recommendations	Has Cubie killed fees?	Has Cubie killed fees?	Has Cubie killed fees?	...	On the critical list	The timing of their lives	Dear doctor	Irish court halts IRA man's extradition to Nor...	Burundi peace initiative fades after rebels re...	PE points the way forward to the ECB	Campaigners keep up pressure on Nazi war crime...	Jane Ratcliffe	Yet more things you wouldn't know without the ...	Millennium bug fails to bite
2	2000-01-05	0	Coventry caught on counter by Flo	United's rivals on the road to Rio	Thatcher issues defence before trial by video	Police help Smith lay down the law at Everton	Tale of Trautmann bears two more retellings	England on the rack	Pakistan retaliate with call for video of Walsh	Cullinan continues his Cape monopoly	...	South Melbourne (Australia)	Necaxa (Mexico)	Real Madrid (Spain)	Raja Casablanca (Morocco)	Corinthians (Brazil)	Tony's pet project	Al Nassr (Saudi Arabia)	Ideal Holmes show	Pinochet leaves hospital after tests	Useful links
3	2000-01-06	1	Pilgrim knows how to progress	Thatcher facing ban	McIlroy calls for Irish fighting spirit	Leicester bin stadium blueprint	United braced for Mexican wave	Auntie back in fashion, even if the dress look...	Shoaib appeal goes to the top	Hussain hurt by 'shambles' but lays blame on e...	...	Putin admits Yeltsin quit to give him a head s...	BBC worst hit as digital TV begins to bite	How much can you pay for...	Christmas glitches	Upending a table, Chopping a line and Scoring ...	Scientific evidence 'unreliable', defence claims	Fusco wins judicial review in extradition case	Rebels thwart Russian advance	Blair orders shake-up of failing NHS	Lessons of law's hard heart
4	2000-01-07	1	Hitches and Horlocks	Beckham off but United survive	Breast cancer screening	Alan Parker	Guardian readers: are you all whingers?	Hollywood Beyond	Ashes and diamonds	Whingers - a formidable minority	...	Most everywhere: UDIs	Most wanted: Chloe lunettes	Return of the cane 'completely off the agenda'	From Sleepy Hollow to Greeneland	Blunkett outlines vision for over 11s	Embattled Dobson attacks 'play now, pay later'...	Doom and the Dome	What is the north-south divide?	Aitken released from jail	Gone aloft

5 rows × 27 columns

Figure: Top 5 rows of the Dataset

How the Dataset looks like

- Data ranges from 2008 to 2016 and the data from 2000 to 2008 was scrapped from Yahoo finance.
- There are 25 columns of top news headlines for each day in the data frame.
- Class 1- the stock price increased.
- Class 0- the stock price stayed the same or decreased.

Data Preprocessing

- We have split the data set into **train** and **test** datasets.
- Apart from “a”to “b” and “A”to “B”, we have removed everything.
- We have also converted the sentences into lower case and then joined all of the sentences in a row together.
- Next we have removed the **stopwords** that is inbuilt in the **NLTK** library.
- Further on, we have performed **Lemmatization** on the data.

Is the Dataset Balanced?

Here, we have plotted a pie chart to check if the dataset is balanced or not.

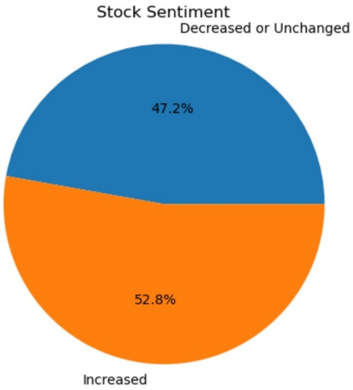


Figure: The Dataset is More or Less Balanced

Output vs Negative Word Count

Here, we see how the outputs are distributed with regard to the number of negative word count in the corresponding input. We have used the builtin list of negative words from **NLTK**.

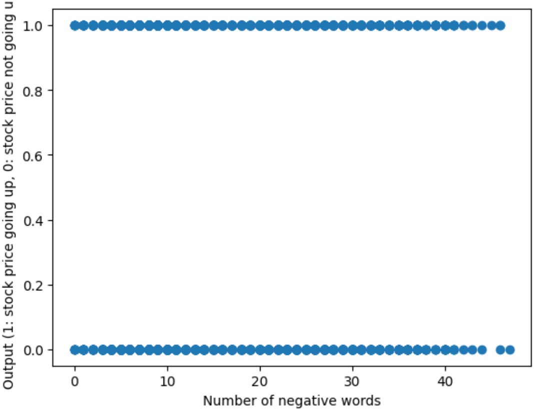


Figure: Output vs The Number of Negative Words

The Sentiment Distribution

We have calculated the sentiment score of each of the headlines (without the preprocessing), using **SentimentIntensityAnalyzer()**, which we have imported from **nltk.sentiment.vader**. Here is the sentiment distribution

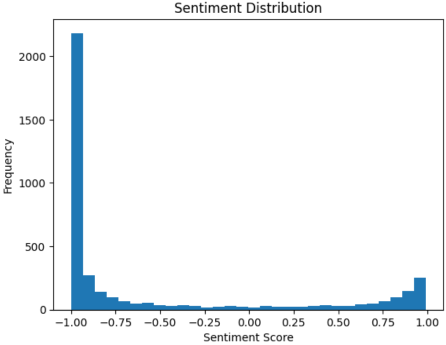


Figure: Sentiment Distribution of News Headlines using VaderSentiment

Using Bag of Words

We have used **Bag of Words** in order to create the input vectors. After the application, After applying Bag of Words, we get a matrix representation of the corpus of text documents where each row corresponds to a document and each column corresponds to a word in the corpus. The matrix contains the frequency of occurrence of each word in the corresponding document.

Table of Contents

1. Introductions
2. The Dataset
   - Data Preprocessing
   - Exploratory Data Analysis
   - Converting the Textual Data into Vectors
3. Applying Random Forest
4. Applying Naïve Bayes
5. Applying Convolutional Neural Network
6. Conclusion

Random Forest

Random Forest is an ensemble learning method used for classification and regression. It uses **Bagging** and builds multiple decision trees using randomly selected features and samples, and aggregates their predictions to make a final prediction. The algorithm reduces overfitting, which is a problem with Decision Trees.

The Experiment With Removing Stopwords and Lemmatization

The accuracy is 0.51 with the support of 378.

	precision	**recall**	**f1-score**	**support**
0	0.50	0.51	0.50	186
1	0.51	0.51	0.51	192
macro avg	0.51	0.51	0.51	378
weighted avg	0.51	0.51	0.51	378

Table: Classification Report using RF with Removing Stopwords and Lemmatization

The Experiment With Lemmatization only

The accuracy is 0.84 with the support of 378.

	precision	**recall**	**f1-score**	**support**
0	0.94	0.72	0.82	186
1	0.78	0.96	0.86	192
macro avg	0.86	0.84	0.84	378
weighted avg	0.86	0.84	0.84	378

Table: Classification Report using RF with Lemmatization only

The Experiment Without Removing Stopwords and Lemmatization

The accuracy is 0.87 with the support of 378.

	precision	**recall**	**f1-score**	**support**
0	0.95	0.77	0.85	186
1	0.81	0.96	0.88	192
macro avg	0.88	0.87	0.87	378
weighted avg	0.88	0.87	0.87	378

Table: Classification Report using RF without Removing Stopwords and Lemmatization

Table of Contents

1. Introductions
2. The Dataset
   - Data Preprocessing
   - Exploratory Data Analysis
   - Converting the Textual Data into Vectors
3. Applying Random Forest
4. Applying Naïve Bayes
5. Applying Convolutional Neural Network
6. Conclusion

Naïve Bayes

Naive Bayes is a probabilistic algorithm used for classification tasks in machine learning. It works by assuming that the presence or absence of a feature is independent of the presence or absence of any other feature, hence the name "naive". It calculates the probability of each class given a set of input features and selects the class with the highest probability as the output.

The Experiment Without Removing Stopwords and Lemmatization

The accuracy is 0.85 with the support of 378.

	precision	**recall**	**f1-score**	**support**
0	0.93	0.74	0.83	186
1	0.79	0.95	0.86	192
macro avg	0.86	0.84	0.84	378
weighted avg	0.86	0.85	0.84	378

Table: Classification Report using Naïve Bayes

Table of Contents

1. Introductions
2. The Dataset
   - Data Preprocessing
   - Exploratory Data Analysis
   - Converting the Textual Data into Vectors
3. Applying Random Forest
4. Applying Naïve Bayes
5. Applying Convolutional Neural Network
6. Conclusion

Convolutional Neural Network

Convolutional Neural Network (CNN) is a type of neural network used for image, audio and text analysis. It consists of input, convolution, activation, pooling, fully connected and output layers. The convolution layer applies filters to the input matrix to extract features, while pooling layers reduce dimensionality to avoid overfitting.

The Experiment Without Removing Stopwords and Lemmatization

The accuracy is 0.82 with the support of 378.

	precision	**recall**	**f1-score**	**support**
0	0.84	0.79	0.81	186
1	0.81	0.85	0.83	192
macro avg	0.82	0.82	0.82	378
weighted avg	0.82	0.82	0.82	378

Table: Classification Report using CNN

Table of Contents

1. Introductions
2. The Dataset
   - Data Preprocessing
   - Exploratory Data Analysis
   - Converting the Textual Data into Vectors
3. Applying Random Forest
4. Applying Naïve Bayes
5. Applying Convolutional Neural Network
6. Conclusion

The Confusion Matrices

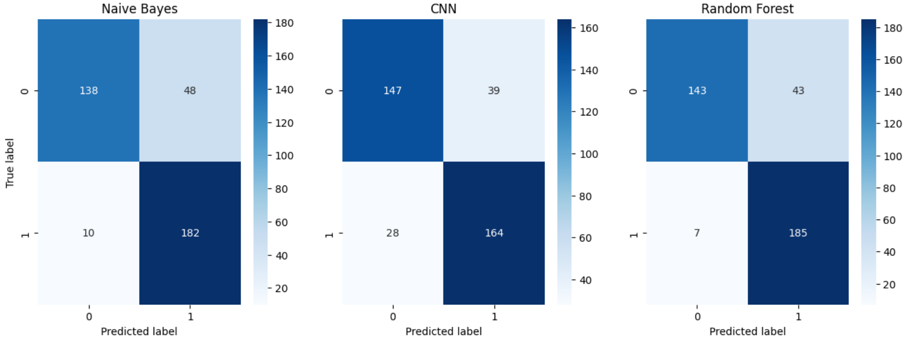


Figure: Confusion Matrices

Comparing the Performances

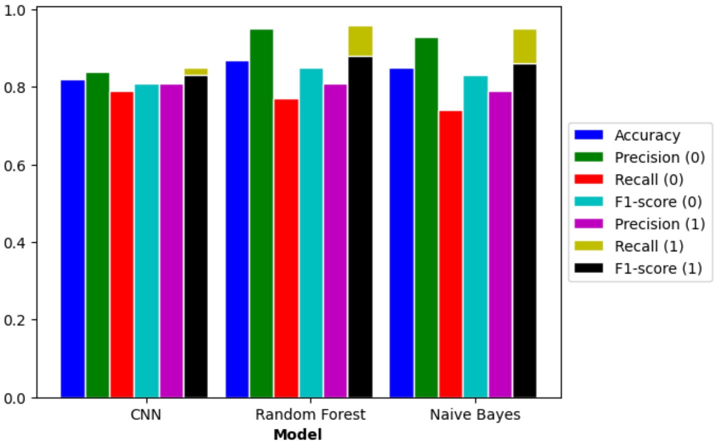


Figure: Comparing the Performances

What We Learned?

- Standard textual data cleaning may not work for all sentiment analysis.
- Removing stopwords can delete relevant keywords, so one needs to be careful.
- Stemming/Lemmatization may mix words with different contexts and reduce accuracy.
- Naive Bayes, CNN, and Random Forest work well for text classification.
- In this use case, Random Forest > Naive Bayes > CNN in terms of performance.

[1] K. Andrews and B. Rajiv. On some applications of eigenvalues of Toeplitz matrices. *Journal of Mathematical Analysis and Applications*, 56(2):237–239, 2007.

[2] C. C. Chang. Algebraic analysis of many valued logics. *Transactions of the American Mathematical Society*, 88:467–490, 1958.

[3] Vijay Choubey. Text classification using CNN. `https://medium.com/voice-tech-podcast/text-classification-using-cnn-9ade8155dfb9` , July 2020.

[4] Runella Gerla. Automata over MV-algebras. In *ISMVL '04: Proceedings of the 34th International Symposium on Multiple-Valued Logic*, pages 49–54, Washington, DC, USA, 2004. IEEE Computer Society.

[5] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Second Edition. The Johns Hopkins University Press, 1989.

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Convolutional networks. In *Deep Learning*, chapter 9, pages 267–313. MIT Press, Cambridge, MA, 1st edition, 2016.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Random forests. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, chapter 15, pages 587–604. Springer, New York, 2nd edition, 2009.

[8] Kevin P. Murphy. Naive Bayes. In *Machine Learning: A Probabilistic Perspective*, chapter 3, pages 77–94. MIT Press, Cambridge, MA, 1st edition, 2012.

[9] Rohit Sharma. Stock sentiment analysis using news headlines. `https://www.kaggle.com/code/rohit0906/stock-sentiment-analysis-using-news-headlines`, May 2021. Accessed on April 22, 2023.

[10] Harshdeep Singh. Understanding random forests. `https://medium.com/@harshdeepsingh_35448/`

`understanding-random-forests-aa0ccecdbbbb` , March 2019.

That Concludes the Presentation, Thank You!