University of Dublin

TRINITY COLLEGE

Sentiment Analysis to model stock price changes in the financial markets

Shouray Duggal
B.A.I Computer Engineering
Final Year Project April 2022
Supervisor: Professor Khurshid Ahmad

School of Computer Science and Statistics

O'Reilly Institute, Trinity College, Dublin 2, Ireland

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

First Last

Month Day, Year

Acknowledgements

First and foremost, I would like to sincerely thank Professor Khurshid Ahmad for his patience, time and supervision. I am truly grateful for his forever helping and guiding nature.

Next, I would like to thank my parents for always supporting me in all my pursuits in life. Having their support and encouragement made my Undergraduate education and thesis possible.

Lastly, I am grateful for all my friends who were always there in times of need.

Abstract

Finance theory states that a stock price is an amalgamation of its value and market noise. This research paper aims to explore the application of computational methods of sentiment analysis to model and forecast this noise.

The research paper employs a systematic approach for the extraction of pseudo sentiment from Hindi and English newspaper articles that cater to people from different socio-economic backgrounds. To estimate sentiment, we use the 'bag of words' approach with a finance and business domain-specific lexicon. The estimated emotion is modelled in conjunction with the stock market returns using linear and time series models.

We further explore methods of pan-linguistic sentiment analysis, by combining the pseudo sentiment from both languages to estimate the impact of the combined sentiment on Indian equities. The results of the various modelling approaches are then verified using several statistical tests.

Keywords: Sentiment Analysis, Bag of words, time series analysis, Linear regression

Table of contents

List of figures

List of tables

Chapter 1: Introduction

The oxford dictionary defines sentiment as "a feeling or an opinion, especially one based on emotions". Human beings have thoughts and feelings that they express as emotions. These emotions or sentiments about different products, people, and companies are well reflected in written documents such as blogs, social media posts and comments, and newspapers.

Newspapers have been in existence since the early 17th century and have been a vital source of information for human beings. We as a species make decisions and form opinions on different matters by consuming and analysing information available to us. With the boom of web 2.0, we have started drifting from newspapers towards e-news and online articles for this purpose. With the rise of the computer age and the betterment of technology available to us, we can now investigate how much of our decisions depend on the information we consume.
This research paper aims to answer this question in the context of business, investment and financial decisions by gathering business news articles in English and Hindi languages and deploying known computational methods of sentiment estimation to extract quantitative insights from the text.

The financial markets are free marketplaces where buyers and sellers interact and exchange assets such as stocks and shares. For every exchange or trade made on the financial markets, there must be a willing buyer, a willing seller and a specific price the two individual parties agree to exchange the commodity at.
Different participants in the markets evaluate the commodity price differently and hence the price changes with every agreeable trade made on the stock market.

This paper analyses how much of the change in the price of a stock can be attributed to the sentiment gathered from the news articles published pertaining to the stock on a given day.

We algorithmically estimate article sentiment from unstructured business news articles using the bag of words methodology. The bag of words is implemented using a domain-specific finance and business dictionary compiled by McDonald and Loughran (MCDONALD, 2011) in English. The dictionary has been translated to Hindi for use in Hindi articles.

The returns of a stock are modelled as a consequence of sentiment using several statistical techniques such as linear regression, multiple linear regression and vector autoregression. The paper performs Contingency table analysis, chi-squared distribution tests other statistical tests to examine the statistical significance of the results obtained.
This research paper also devises and implements techniques of combining sentiment estimated from different languages and studies the impact of the combined sentiment on stock returns.

Chapter 2 gives a detailed background of the process of sentiment estimation and explores various methods of sentiment analysis devised and implemented by various research papers. Chapter 2 also briefly talks about the financial markets and their basic mechanism. It also gives a brief introduction to the statistical and econometric techniques used in this research paper. Chapter 3 discusses the chosen approach for analysis and goes into technical and implementational depth. The chapter ends with all the statistical analyses performed. Chapter 4 Discusses the results achieved and their statistical significance. Chapter 5 summarises the entire research paper and draws final conclusions on the performed analysis. Chapter 5 also describes the scope of work and research on the topic going forward.

Chapter 2: Background and Literature review

2.1 Financial markets

The financial markets are open markets where buyers and sellers can engage and trade investment vehicles like bonds, commodity futures, stocks of publicly traded companies, etc. Every financial market trade or exchange requires a willing buyer, a willing seller, and a precise price at which the two parties agree to exchange an asset.
A stock or share of a company is fractional ownership of the company, when an individual buys a share of the company, they become part owners of the company business and all the assets and liabilities of the company.

The stock prices in a free market are dynamic and change every time a successful trade is executed on the market. That is to say; depending on the stock, a stock price changes between a couple hundred to a couple hundred thousand times during a trading session.

The fundamental law of supply and demand dictates that the price of a commodity increases when there are more willing buyers than sellers, and the price decreases when there are more willing sellers than buyers.
The law is valid for stock prices and other assets as well. If the number of buyers is greater than the number of sellers for a given stock, the stock price increases and if the number of sellers of a stock is greater than the buyers, the stock price decreases.

Amongst many theories surrounding stock prices, the most prominent is the efficient market hypothesis. The theory claims that a stock price reflects all the information available pertaining to it, and only reacts to the newly available information.
From the efficient market hypothesis, we can deduce that the market forces of supply and demand keep the stock prices in equilibrium.
The theory implies that it is impossible to consistently make risk-adjusted profits from the market.

2.1.1 Financial markets and sentiment

We as a species make decisions and from opinions based on the information available to us. When it comes to the financial markets, The main source of information is the news and the market itself. The news supplies investors and other market participants with information regarding the change in the specific industry, government policy or overall financial and economic state of the country and the world. News also brings in information about how the prices of stocks and shares are changing every day.

News is the source of information that aids market participants form opinions and therefore, may influence their investment decisions.
Implied from the law of supply and demand of commodities, if more people on a day decide to buy a given stock than the people who decide to sell, the price of the stock increases. Conversely, if the number of people who decide to sell the stock is more than the number of people who decide to buy the stock, the price decreases.
Thereby forming a plausible link or relation between news sentiment and stock prices.

2.2 Sentiment analysis

Human beings have feelings that they express using emotions. Sentiment analysis is a field of study that analyses human emotions, views, attitudes, and assessments of things. (Liu, 2012)
since the evolution of the web into web2.0, people have an open place to express and register their opinions and emotions regarding different entities like products, leaders, companies, etc.
There is no doubt that text information carries massive amounts of insight and understanding the 'emotions' and 'feelings' of the masses regarding an entity can impact the decision-making process substantially.
With the outreach of the internet expanding rapidly every day, according to Forbes, in 2019 456,000 tweets were sent over Twitter, every minute.(Forbes, 2019) It's impossible for human beings to manually process and derive insights from this data which is where automation comes in the picture.

The field of sentiment analysis is closely related to NLP and has been under research since the 1950s (MCDONALD, 2011), the objective of the approach can be classifying the document's polarity (positive, negative or neutral) and if required assign a score to this polarity.(Liu, 2012)

The task of sentiment extraction from a piece of text can be broadly categorised in the following ways: -

a) *Document-level***:**

As the name suggests the document level works on the topmost level, where the basic unit of information is considered to be a text document. The approach assumes that the entire document is concerning one entity and represents the view of one person.
Depending on the objective of the analysis, it can either be a classification problem if the expected result is categorical and if the expected result happens to be a score or an ordinal variable the solution can be achieved using regression. (Liu, 2012)

b) *sentence level:*

There is no fundamental difference between the sentence level analysis and document level analysis apart from the fact that the basic unit now is a sentence. A sentence can be interpreted as a mini-document for the analysis.
The analysis, however, can be used to classify sentences into objective sentences and subjective sentences. Which in some research papers has been used to remove objective sentences as researchers believe they carry no sentiment or opinion and are solely present to convey information.(Liu, 2012)

c)*Aspect level:*

For several applications document level or sentence level analysis may be inaccurate or may consist of a lot of noise. Since they do not excavate and assign sentiments to specific targets. When we talk about Aspect level analysis, we must define the opinion target entity and its different aspects. for example: 'the earphones are good but their battery life is short, the earphone is the general entity and its battery life is the aspect.
This approach can be fairly more complex than those discussed before and requires advanced NLP precision.(Liu, 2012)

2.2.1 NLP

Natural language processing is a field of computational linguistics that works with unstructured text data to interpret and extract actionable insights from it using both machine learning and lexicon-based approaches.

2.2.2 Machine learning

Machine learning is one of the buzzwords of the decade. Machines are fundamentally unintelligent, and the goal of machine learning is to train a machine to perform cognitive tasks like classification, clustering, or estimation; without hardcoding any specific instructions into it. To do so, learning is performed using several different sets of rules for problem-solving called algorithms, which can be decided by the practitioner depending on the underlying data and cognitive task.

For any machine learning task to be performed an algorithm must learn from past experience, that is from historical data and understand its underlying properties or features.

The learning stage is known as training where thousands and for some models, hundreds of thousands or even millions of data samples are used to train a model to perform the task at hand.

Based on the task at hand and the available data for training machine learning algorithms can be classified into supervised machine learning and unsupervised machine learning.

Supervised machine learning is where the training data is labelled, that is the attribute of the data we are trying to predict is available in the training dataset. Supervised machine learning algorithms are generally used to perform classification tasks.

Unsupervised machine learning is where the training data isn't labelled and the algorithms try to explore and learn different attributes of the underlying data. Unsupervised learning algorithms are generally used to perform clustering tasks where the data is clustered or divided into several groups based on its underlying attributes.

In the context of sentiment analysis, the goal of any machine learning model is to learn the cognitive process of comprehending the tone of a text.

The advantages of machine learning models are the flexibility to modify and construct trained models for a variety of applications and scenarios, whereas the disadvantages are the requirement of large data sets in order to train models and their low accuracy on new types of data. (D'Andrea, 2015)
Some machine learning models that may be used for sentiment classification: -

Multivariate Naïve Bayesian classifier

The supervised classifier can be used either on the sentence level or document level and works in the following way: -
let there be 3 classes Ci, where i belongs to {1,2,3} and 1= positive, 2 = negative and 3 = neutral. Using the pre-processed training data, a feature space is created, which corresponds to the probability of each word belonging to a given class. empirical probability is calculated for each word belonging to each class. $P(x_j/C_i)$; where j is the number of words or phrases we are examining.

Now,

$$P(Ci/X) = \prod_{j} \mathrm{p(xj/Ci)}$$

$P(Ci/X)$is calculated for all classes and the document or sentence is classified as belonging to the class with the highest probability.(Zhan, 2015)

SVM

Support Vector Machines are a supervised method of classification that separates two or more classes by drawing a hyperplane in the N-dimensional feature space, while there can be multiple hyperplanes to do that, SVM optimises this choice by choosing the one with maximum distance from either class. The key property of SVM is the kernel function that essentially converts the feature matrix to an N*N matrix, where N is the number of observations and is independent of the number of features.
Many kernel functions can be used according to the data being modelled and its feature space. Linear, Radial kernel and polynomial kernel functions happen to be some of the more frequently used examples.

Random forests

Random forests is a classification algorithm built over the decision tree algorithm that leverages two randomised processes, bootstrapping and random feature selection. The first step is bootstrapping which is randomly choosing N training data sets from the original training dataset for N different decision trees.

Now, for each of these N decision trees, features are randomly selected and each tree is trained using the randomly selected sub-section of features of a sub-section of the training data. After each tree is finished processing the input and returns a category, Aggregation on all the outputs is performed and the most frequent class is selected as the final output.

2.2.3 lexicon-based sentiment analysis methods

Before we get into lexicon-based methods let us break down a language into its fundamental constituents: words.
"The words of any language can be broadly broken down into two categories or classes Open-class words or Close-class words." (Murray)

2.2.3.1 Closed class words

In general, close-class words of a language are a family of words consisting of conjunctions, pronouns, articles and prepositions. This family in general doesn't accept new members or evolve, so to say, over time.
The function of these words is to add grammatical structure to a sentence. Prepositions mostly precede a noun or pronoun and specify a relation with another word in a sentence. Pronouns are words that refer to nouns and replace them in a sentence. conjunctions are words that are used to combine two or more phrases. (Nordquist, 2020)

2.2.3.2 Open class words

Open-class words in a language are a family of words comprising nouns, verbs, adjectives, etc. The family keeps accepting new members and keeps evolving with time. New words are added to the class as things are trending and become fashionable in the community.
Example: I will 'google' it tonight. Google is now a new addition to open-class words, a verb that means finding information on some topic over the internet.

2.2.3.3 Conclusion

It was established during the telegram times, and as observed in 'telegraphic' English that closed-class words can be removed easily without losing the intended message or meaning of a sentence.
Something that is also subconsciously achieved by most students when taking notes in class by using telegraphic English, and minimizing the word count without compromising the meaning.(Nordquist, Open class words in english grammar, 2020)

2.2.3.4 Lexicon-based methods

Lexicon-based approaches for natural language processing are algorithms that employ the aggregated frequencies of a lexicon including certain open class words used in the text to compute the polarity of the document.
Where polarity of a document is the sentiment it carries, which can be positive, negative or neutral
In comparison with machine learning methods, lexicon-based methods require no training using historical data and are easier to use and comprehend.
There are majorly two practical approaches to lexicon-based sentiment analysis:
a) corpus-based
b) dictionary-based
(D'Andrea, 2015)

2.2.3.4.1 Corpus-based approach

The corpus-based technique allows for the computerised generation of new sentiment lexicons or updating of existing lexica. The idea is to start with a small sentiment lexicon called a seed list and find new sentiment words by performing co-occurrence analysis on other words in the text corpus, Alternatively, syntactical patterns can also be exploited for the task (Mohammad Darwich, 2019) a sentiment-severity score or measure for the polarity can also be calculated using the distance between previously annotated sentiment words and the word under analysis. Using a large domain-specific corpus, this technique may be applied to create a domain-specific sentiment lexicon list from a general-purpose sentiment list.
However, because even in extremely domain-specific corpora, two different instances of the same phrase might convey different sentiments, this strategy can occasionally introduce noise into the process. As quoted in (Liu, 2012) , " 'This car is very quiet' is positive, but the sentence “The audio system in the car is very quiet” is negative."
reflecting quiet being associated with both positive and negative sentiment in the context of a car.

2.2.3.4.2 *Dictionary-based approach*

The dictionary-based approach is rather simple and works in the following way, one starts by manually collecting a small sample of seed words with known polarity and keeps on adding

to the lexicon by iteratively searching synonyms and antonyms from known language dictionaries. The synonym is assigned the same sentiment as the original world and the antonym is assigned the opposite polarity. The iterations may stop when no new word can be found. The prepared lexicon and its corresponding polarities must be manually authenticated before using it in any consequent classification or regression modelling.

Alternatively, the approach of calculating a polarity score based on WordNet relative distance was suggested in (Jaap Kamps, 2004). The paper calculates the polarity of a word by calculating its shortest relative distance from words with known polarities in the WordNet database. Where WordNet is a lexicon database established in the mid-1980s which aims to group and map words using semantic relationships between them.(university, n.d.)

2.2.4.3 Bag of words for text analytics

The notion of the approach is that we are only interested in the frequency of terms (one or greater than one word) in an artefact not in the specific order in which they appear in it or the specific context of each occurrence, as we believe in general the existence of each word is tied to reflecting a sentiment, which can be approximated by reading the frequency of some pre-defined, labelled words.

Each article is broken down into its constituent words or tokens using this lexicon-based method of analysis. The document is tokenized into a term-document matrix, with each term acting as a token. The term-document matrix is a matrix that represents the frequency of each term in each document.

$a_{i,j}$ = frequency of the j^{th} term in the i^{th} document.

Depending on the underlying data and analytical goals, several approaches like N-grams and skip-grams can be employed to vary the tokenization step. The N-grams technique tokenizes the document by using N consecutive words as the token, whereas the skip-grammes approach tokenizes the document by using N-words that are at a non-zero, constant distance from the predecessor. (Tom Martya, 2020)

For the purpose of extracting pseudo sentiment for our research, we are interested only in the frequencies of a select pre-determined, domain-specific dictionary of words. The words are labelled as conveying either positive or negative sentiment.

2.3 Econometric and Statistical methods

2.3.1 Linear regression:

"The most common parametric approach in the news analytics literature is multiple linear regression" (Tom Martya, 2020)
The method takes use of the correlation between the dependent and independent variables to estimate the dependent variable's value using the independent or predictor variables.

$Y = A_1x_1 + A2x_2 + ... + C$

In the formula above, y is the dependent variable and x_1, x_2, x_3, etc. are the dependent variables, whereas A_1, A_2, etc. are called the weights and C is called the bias term.
The approach uses Ordinary least squares or general least squares to find the optimal values of the parameters or weights.

OLS optimizes the weights by minimizing sum of least squares between the observed values and the predicted values.

2.3.2 Vector autoregression:
Vector autoregression is a variant of linear regression widely used in econometrics to realise the linkage between two interdependent time series variables.

The regressor uses the L-lagged values of each variable time series to estimate the current value of a time series in an equation.

Let X and Y be two time series. X(t) represents the value of X at time t and Y(t) the value of series Y at time t. The equations for a VAR model with lag 1 are as follows:

$Y_t = A_{1,1}Y_{t-1} + A_{1,2}X_{t-1} + E_1$

$X_t = A_{2,1}Y_{t-1} + A_{2,2}X_{t-1} + E_2$

Where E_1, E_2 are error terms for each model.

2.3.3 Z-scores:

A Z-score is the distance of an observation from the sample or population in terms of the standard deviation. The Z-score of an observation is defined as: -

$Z_i = (X_i - \mu) / \sigma$

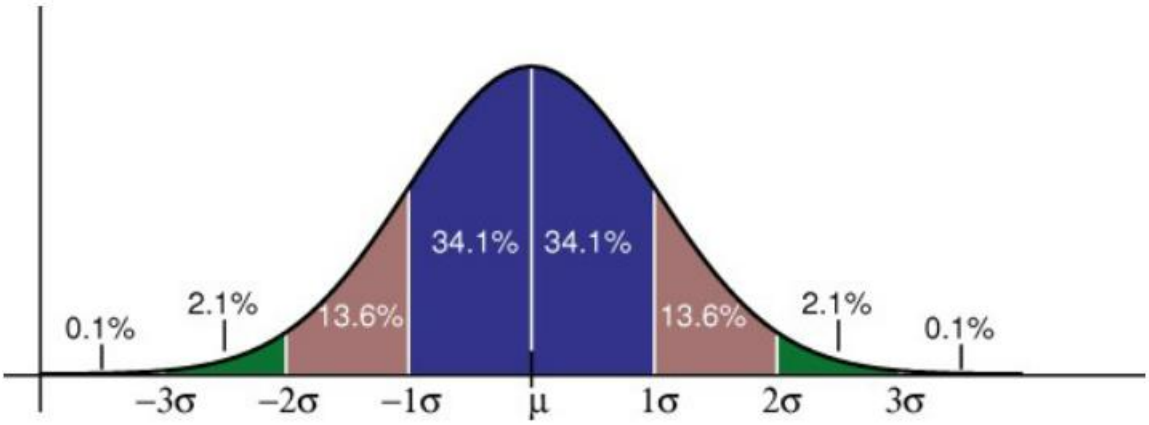
When describing quantities, it is often more reasonable to include a point of reference. Z-scores also aid in the investigation of the underlying distribution of the observations.

2.3.4 Normal Distribution:

The normal distribution is a bell-shaped probability distribution which is often observed in nature. For example: When measuring the length of a plant or tree, etc.

Any set of data is said to be normally distributed if: -

a) The probability of an observation to lie one standard deviation on either side of the mean is approximately 68 %
b) The probability of an observation to lie two standard deviations on either side of the mean is approximately 95 %
c) The probability of an observation to lie three standard deviations on either side of the mean is 99 %
d) It is statistically almost impossible for an observation to lie beyond five standard deviations from the mean.
e) The curve is symmetric around the mean.



(normal distribution, n.d.)

2.3.5 Statistical significance and hypothesis testing:

A result is said to be statistically significant if it can be determined that the result is based on some statistical property of the underlying data and isn't so because of absolute chance or randomness in the data.

To determine if a result is statistically significant, one must perform hypothesis testing, that is the process of setting a null and an alternate hypothesis. The null hypothesis is generally set to hypothesise the result to be statistically insignificant. We then calculate the probability of the data being the way it is assuming that the null hypothesis is true. This probability is known as the p-value and can be calculated using several different statistical tests depending on the underlying data.

P-value is then compared with the level of significance which is generally and arbitrarily set as 5%. If the p-value is less than the level of significance then we can confidently reject the null hypothesis and confirm that the results are not just because of chance.

2.3.6 Chi-squared test:

Chi-squared test is a statistical method of testing dependence between two categorical variables.

The null hypothesis if the test is that the variables are independent. The alternate hypothesis is that the two variables are dependent

The test calculates a chi-squared statistic using the formula:

$$X^2 = \sum \frac{\text{ObservedObserved(i)} - \text{Expected(i)}}{\text{Expected(i)}}$$

Where $\text{Expected}_{(i)}$ is the ith value in the expected distribution calculated assuming independence between the two variables.

Chapter 3: methodology and implementation

The chapter discusses the process chosen to test the hypothesis made in chapter 1 and the implementation of the chosen process.

The fig below gives a top view of the overall process and each subsequent subsection discusses each of the process steps shown in the figure with implementation of the same. Each subsection discusses the input received, output generated, the process followed in between and the need or reason for choosing the process.

Figure 3. 1: Implementation Work flow

The leaf nodes of the tree shown in the figure are models applied on the prepared dataset(s) and the results will be shared in the next chapter.

3.1 Data sources

This research paper works with 2 unstructured text news article datasets in English and Hindi languages and one structured market dataset.

3.1.1 Dainik Jagran

Dainik Jagran is a daily Hindi newspaper that published its first article back in 1942. According to the Audit Bureau of Circulation India, Dainik Jagran had an average half-yearly readership of 4.14 million as of 2019 (India, 2020) and is the second most-read paper by circulation. The newspaper is read prominently in the states of New Delhi, Punjab, Uttar Pradesh and other northern parts of India. The newspaper is a national paper and not a business paper. The paper, however, has a business section that summarises the few most relevant business events.

Even though the newspaper is highly decorated and well-read in India, the website is seriously outdated. The targeted search feature is obsolete and only allows targeted search using only one parameter. For the use case of this research paper, we could either target search business news or search news related to the company in question, it wasn't technically possible to perform a targeted search using complex search parameters.

The second concern was improper tagging of articles, since the paper is a national paper, the newspaper chose to tag different articles by their geolocation as well. In some cases, only a geolocation tag is available. In the context of this research paper, 'Jamshedpur' is a city in the eastern state of Jharkhand where the biggest factory of Tata motors is situated since Mr Ratan Tata is a great philanthropist, Jamshedpur also has a huge hospital operated by Tata motors. The news concerning the daily operations of the tata hospital and the factory were both tagged by the geotag 'Jamshedpur'.

The addressed problems added a severe amount of noise to the data set and automating the data collection process from this source proved inaccurate due to these challenges, which is why the data source was dropped from further analysis.

3.1.2 Amar Ujala

Amar Ujala is a daily Hindi newspaper that initially began publishing in 1948. As of 2019, Amar Ujala has an average half-yearly readership of 2.6 million according to the Audit Bureau of Circulation India (India, 2020) and is the fourth most-read Hindi paper and fifth most-read newspaper in India by circulation. The newspaper is widely read in the states of Haryana, Himachal Pradesh, New Delhi, Chandigarh, Punjab, Jammu & Kashmir, Uttarakhand, Uttar Pradesh and other northern parts of India.
It is important to note that the newspaper is a national paper and not a business paper. This newspaper, however, also has a business section that summarises the few most relevant business events.
We also observe the newspaper is not the most prominent in the financial capital of the country: Mumbai.

We observe this data source is representative of the general population of the country, and not of actual market participants.

We instil the assumption that if a piece of news is emotionally relevant, it will be covered by a major national newspaper as well. Even if the objective part of the news may be financially insufficient, the emotions will be reflected significantly in the text.

3.1.2.1 Web scraping

For any research or analysis, the most fundamental and crucial step is data collection. When it comes to data, for any research the more the merrier. While sometimes it may be essential, it isn't advisable to collect data manually since it is a tediously dull process and consumes a lot of time. While most websites like Twitter, Facebook, etc. have their own APIs that allow users to retrieve structured data, this is not the case for most websites on the internet.
Web scraping is the process employed to automate the data collection process for the data available on the internet. Web scraping can be broken down further into two processes: Crawling and scraping.
Crawlers are programs that browse the web to search for relevant data. While scrapers are used to get the required data from unstructured HTML web pages.
This is essentially done by writing sophisticated code scripts or can be done using multiple available frameworks.

For example, Scrapy is an open-source and collaborative framework that allows people to create their own crawlers with minimal programming. The framework was written in python and is portable across all operating systems.
Similarly, 'Beautifulsoup' is a python library that can be deployed over HTML or XML documents and makes it relatively easier to iterate on, modify and retrieve information from such documents.

3.1.2.2 Web Scraping Amar Ujala: a 2-step approach

For aggregating Hindi news article data, we used a hybrid approach of web scraping where we manually performed the crawling process by implementing a targeted search using the keyword ' टाटा मोटर्स ' (Tata Motors) on the Amar Ujala website and retrieved the link of the search results.
Search results were of an infinitely scrollable web page, the web page keeps refreshing the visible data as the user keeps scrolling.

Next, using the manually retrieved link and the requests library in python, we created a web connection between the script and the Amar Ujala website and downloaded the content of one scroll per cycle. The web page data is stored in semi-structured HTML objects.
We inspected the webpage manually and spotted the tags that were used to store the links of articles shown in the search results. In the HTML object, we also located the published date and the metatag associated with each article.
Using ‘beautifulsoup’, we extracted the link, published date and meta tag and saved them in tabular form. The entire cycle was performed 150 times with each cycle bringing in the data of 19 links on average.
This was performed only 150 times because cycles after 150th kept returning the data received from the last cycle.
Remember the first step of the targeted search was just targeting articles with the company name and getting the date, article link and metatag of the article.
The fig below shows the metatag distribution of the retrieved results after the first step.
The total entries retrieved are: 1819

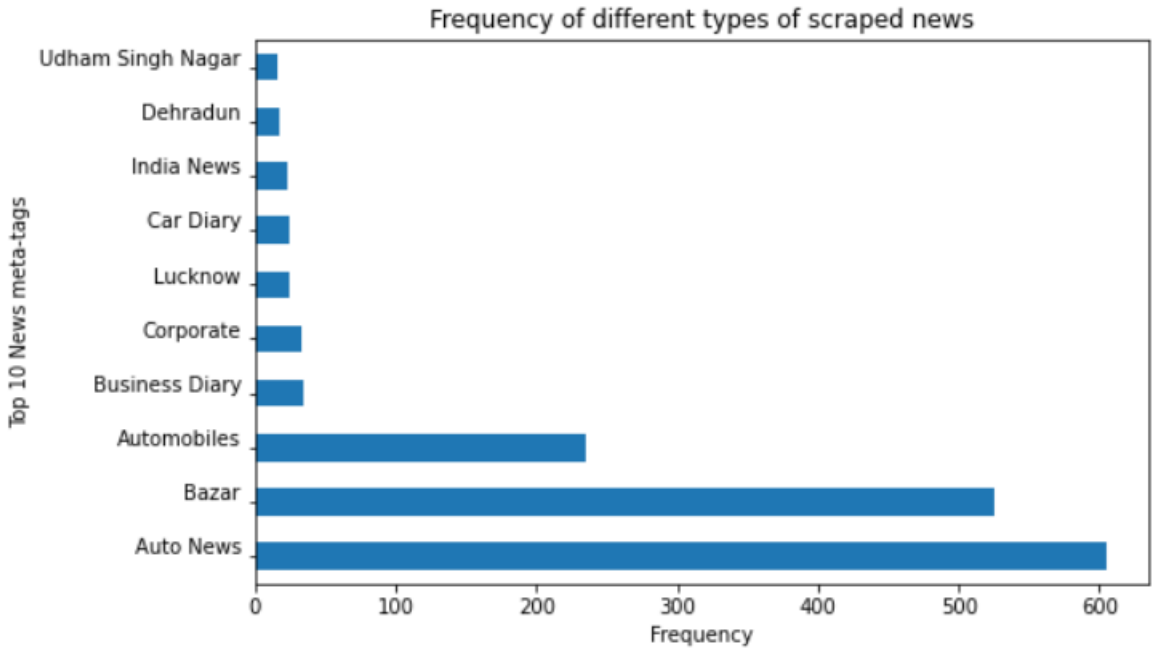


Figure 3.1. 1 Scraped data targeted Search: step 1

Next, we only further scrape the articles with meta tags being 'Bazar', 'Business Diary', 'Corporate' and 'Business'.
To complete the second step of the targeted search, we accessed the said category links to get the complete article data.

Ensuring we have *business* news articles concerning *Tata Motors* from Amar Ujala

As a result of this process, we now have a tabular data set of articles comprising: the headline, article text, date of publishing and meta tag.

Source name	Start date	End Date	Total articles	Total Words (Raw Data)
Amar ujala	2017-07-05	2022-03-14	609	389,205

Table 3.1. 1 Hindi data description

3.1.3 Lexis data

LexisNexis is a data solutions corporation founded in 1970 that provides analytics and database services; Amongst many online databases offered by the company is the 'news and business database. This database comprises business news published in multiple languages from all over the world and also contains articles from newswires, industry trade press, magazines, journals, blogs, etc.

We performed a Targeted search on the platform using five parameters: the keyword 'Tata motors', the time frame set to the past 5 years, the publication type set to only newspapers, The language of publication was set to English, the country was set to India.

The time frame was set to past 5 years to maintain uniformity between Hindi and English language datasets. To comply with the hypothesis in chapter 1, we are only concerned with sentiment generated by newspaper articles and not blogs, newswires, etc. The country was set to India since we are trying to understand the impact of sentiment on Indian equities.

While Web scraping of English news was also a possibility, the Lexis database was preferred, since multiple well-established business newspapers like Financial Times have their own paid API services and do not permit web scraping of their data. Secondly, the data is reliable because it is in the organization's business self-interest to supply accurate and reliable data. Lexis was chosen as the data source also as a matter of convenience because of LexisNexis corporation's association with the Trinity College Dublin Library. A brief analysis of the sources of the news resulted in the following

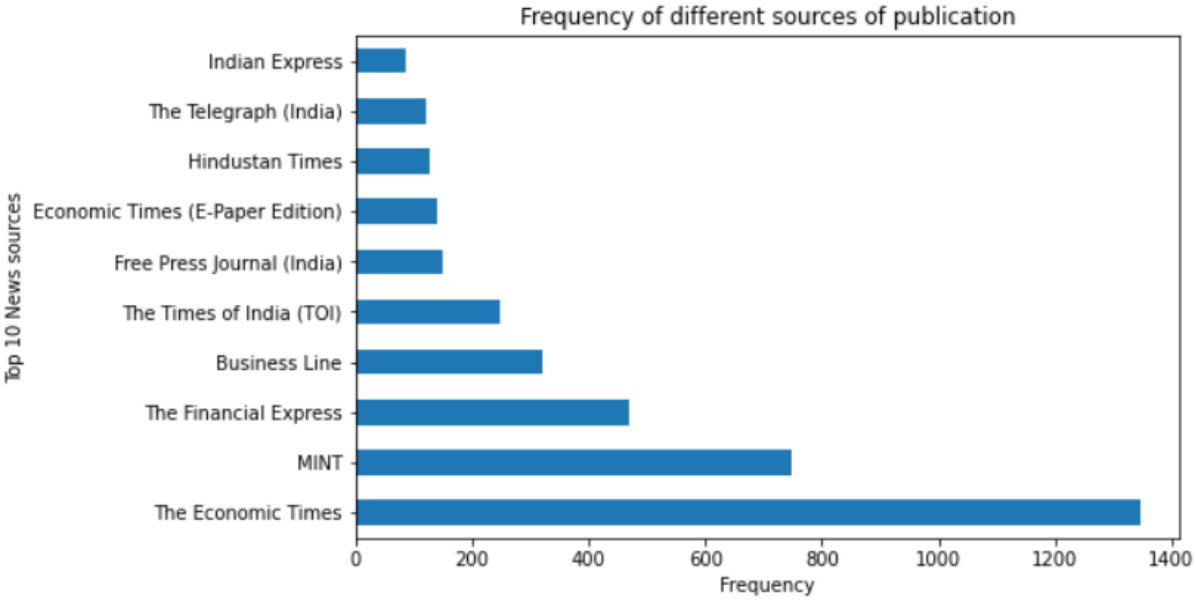


Figure 3.1. 2 English data sources

3.1.4 Financial Data

Tata Motors: Tata Motors is a leading auto manufacturer active in all sub-domains of the industry from trucks and buses to cars and SUVs. The company also has a strong presence in the defence industry in India, outside of India Tata motors owns JLR (Jaguar Land Rover) among other subsidies and has a substantial market share in the world auto market.

We chose this stock mainly because of two reasons:
Its market capitalization is approximately 1.3 trillion USD, with 765.83 million outstanding shares and an average daily trading volume of 1.35 million shares, this shows statistically it's almost impossible to manipulate this stock. Hence the data would reflect the true nature of the market. Secondly, Tata as a conglomerate and Tata motor as an auto manufacturer are very old, famous and prominent in India, therefore a substantial amount of news data would be available for the objectives of this paper.

Since Yahoo finance is a market data provider chosen by many finance research papers (Berkley), 2020) and market participants. Whilst many more accurate and real-time service

providers like Bloomberg exist, they charge a fee for their services, whereas yahoo finance provides historical data of most equities and ETFs for free of cost. Since this research paper does not require real-time data and only analyses the impact of sentiment data on historical data, Yahoo finance was chosen as the data source.

Daily market OHLC data was collected that contained the stock's opening and closing prices, along with the highest and lowest prices the stock traded at during the day, the total number of shares traded during the trading hours and the final adjusted closing price for the day.

This research paper aims to model the adjusted closing prices, which are shown in the figure 3.1.3 shown below

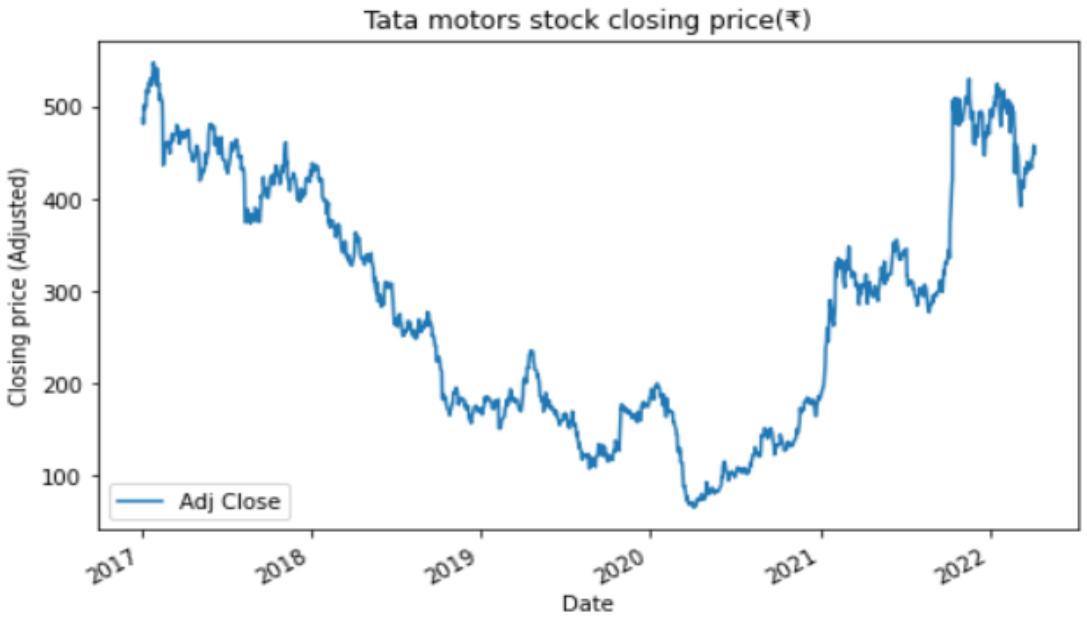


Figure 3.1. 3 Financial data: Closing prices

We summarise the quantitative information of all the Data gathered in table no: 3.1 below, marking the end of the data collection process.

Data source	Start date	End date	Total observations/ articles	Total words (Raw Data)	Data type
Amar Ujala	2017-07-05 (5th July, 2017)	2022-03-14 (14th March, 2022)	609	389,205	Text(articles): Hindi
Lexis	2017-04-09 (9th April, 2017)	2022-04-08 (8th April, 2022)	4513	1,944,905	Text(articles): English
Yahoo finance	2017-01-01 (1st January, 2017)	2022-04-09 (8th April, 2022)	1302	NA	Market data (OHLC)

Table 3. 1 Total data collected

3.2 Text Pre-processing:

As one might be able to estimate from Table 3. 2 the number of words calculated in the raw form of Amar Ujala Hindi articles is 389,205 which corresponds to an average article length of 639 words which appears to be high. This is reflective of the fact that raw web scraped data might be unsuitable for any type of further analysis. After further thorough and manual inspection we observed the data contains a lot of anomalies including but not limited to Html tags from the website, English lexicon metatags that were a part of the article on the website, misplaced newline('\n') and space (' ') characters, conjoined words around punctuations when switching paragraphs, etc.
As mentioned in the previous sub-section, the headlines and the body of the article were also saved as separate attributes.

For the scope of this research paper, we are interested in the sentiment conveyed by all the text information and will be weighing sentiment across the entire text equally, alternatively, we could measure the pseudo sentiment by adding more weight to the sentiment conveyed by the headline and less to the sentiment in the article's body.(Tom Martya, 2020)

3.2.1 Amar ujala Hindi text

To begin the text pre-processing, we conjoined the headline and the article text to form a new attribute and then removed all the numeric characters, punctuations and English alphabet, replacing each of these with a blank space character (' '). Next, we replaced all the double or more consecutive space characters with a blank (''). This was necessary since our word-counting function counts the number of space characters.

3.2.2 Lexis English text

The data received from LexisNexis was rather structured and the text was clean and did not require much cleaning. We just removed punctuations and converted all the text to upper case. This wasn't done for the Hindi text as there is no concept of a case in Hindi, there are no capital or short hand letters.

3.3 Financial data processing

3.3.1 Removing covid crisis from data

Between the 20th of February, 2020 and 23rd of March, 2020 the financial markets all over the world lost value massively and for many consecutive trading days. figures 3.3.1 and 3.3.2 shown below track the price movement of S&P 500 index and the price movement of Tata motors stock.

Figure 3.3. 1 S&P500 index during covid-19

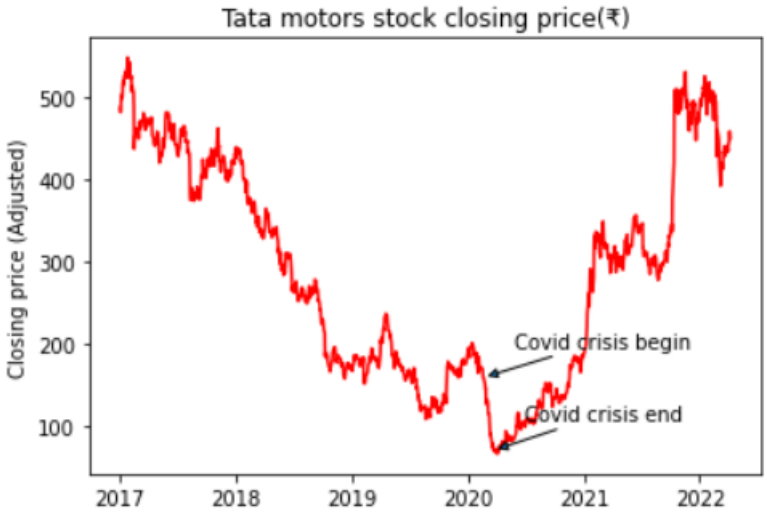


Figure 3.3. 2 Tata stock: Covid-19 crisis

This is an outlier behaviour where the informed traders and Uninformed traders both traded in the selling direction causing a market wise bust (Ahmad, 2021) This behaviour is not frequently observed in the markets and was removed from the data set.

3.3.2 Calculating log returns

As we Noticed in figure 3.3.1 and figure 3.1.3**,** the stock closing prices are highly correlated and are absolute in nature. The autocorrelation of the presented data was 0.998 which is severely high.
The goal of this research paper is to examine the impact of sentiment on stock prices, which cannot be done by looking at individual stock prices in isolation.
To understand the impact of sentiment on the stock prices we must study the prices relative to one another, to quantify the associated change.

This is why log-returns were calculated using the adjusted close attribute.

$$\text{Log return} = \log\left(\frac{\text{Adjusted close price(t)}}{\text{Adjusted close price(t}-1)}\right)$$

Following the transition from prices to log-returns, the autocorrelation dropped significantly to 0.023 as seen in the figure below, the data is also seen to be mean-reverting in nature, which implies that the returns always return to the long run mean.

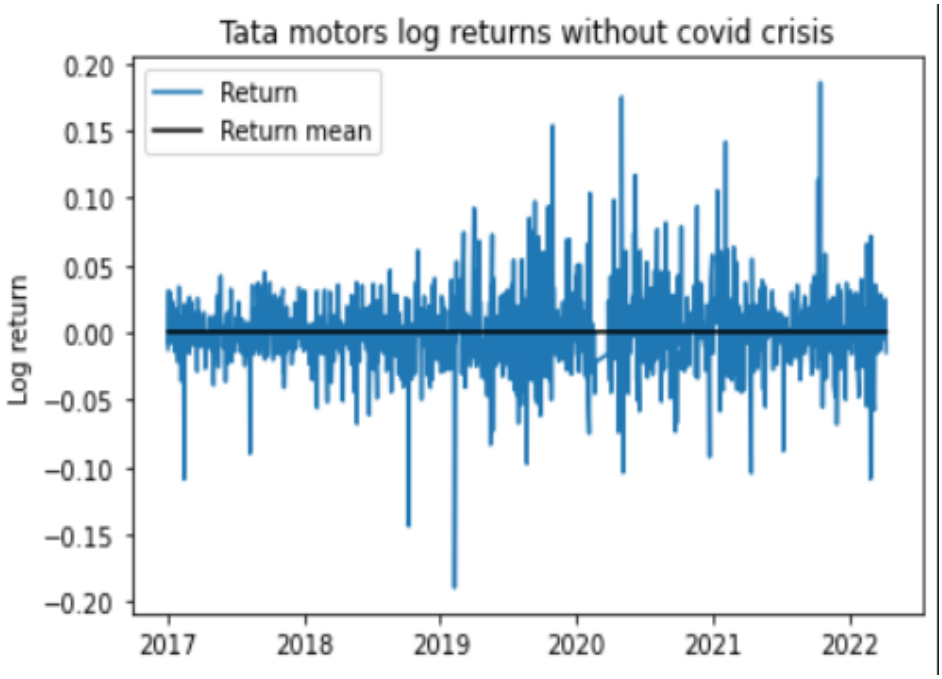


Figure 3.3. 3: Log returns vs time (Without covid-19 crisis)

3.3.3 Exploratory Data Analysis and descriptive statistics

Descriptive statistics:

Statistic	*Value*
Mean	0.00057653
Standard Error	0.000811478
Median	0.000434212
Standard Deviation	0.029032323
Sample Variance	0.000842876
Kurtosis	6.752175505
Skewness	0.506073977
Minimum	-0.189674882
Maximum	0.185860152
Count	1280

Table 3. 3 Descriptive statistics

To analyse the distribution of log returns we have binned the values of the returns between N and N-1 standard deviations away from the mean where N belongs to Z and [-8,8]:

Standard deviation from the mean	Value	No. of data points in bin	Percent data	Cumulative
-8	-0.23159	0	0	0
-7	-0.20257	0	0	0
-6	-0.17355	1	0.078125	0.078125
-5	-0.14453	0	0	0.078125
-4	-0.11551	1	0.078125	0.15625
-3	-0.08649	8	0.625	0.78125
-2	-0.05747	13	1.015625	1.796875
-1	-0.02844	120	9.375	11.171875
0	0.000577	505	39.453125	50.625
1	0.029598	498	38.90625	89.53125
2	0.058618	92	7.1875	96.71875
3	0.087639	28	2.1875	98.90625
4	0.11666	8	0.625	99.53125
5	0.145681	2	0.15625	99.6875
6	0.174702	2	0.15625	99.84375
7	0.203723	2	0.15625	100
8	0.232744	0	0	100

Table 3. 4: distribution of log returns

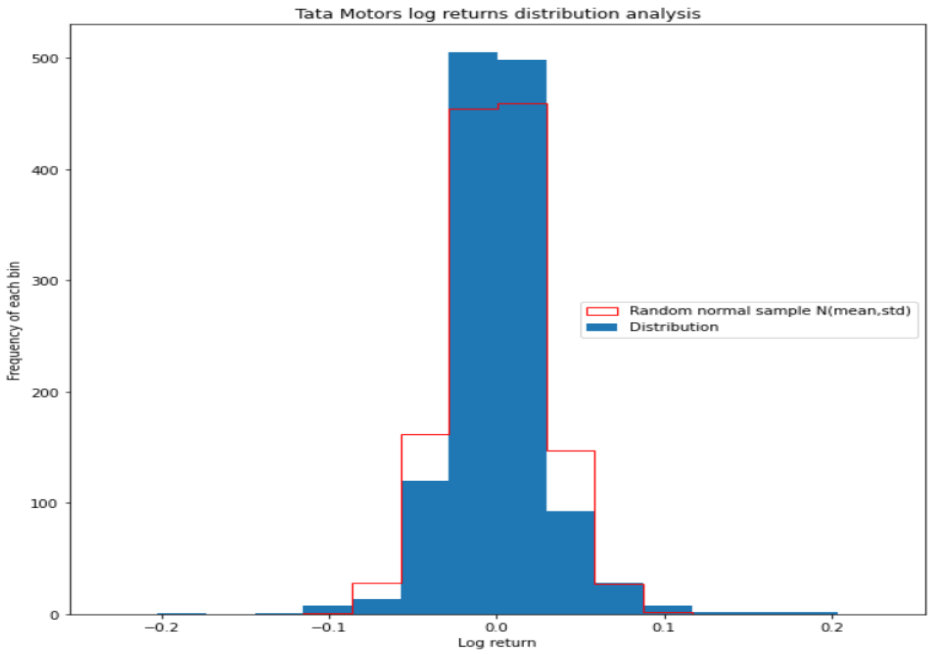


Figure 3.3. 4 Histogram of log returns

The data is *not* normally distributed as a lot of values lie beyond Five standard deviations away from the mean and the observed kurtosis of the data is double that of a normal distribution. However, we can observe that maximum values lie one standard deviation away from the mean in either direction and form Table 3. 5 observe 50% of values lie below the mean and 50% above the mean statistically proving the conjecture of mean reversion observed in figure 3.3.3.

3.4 Estimating sentiment

This sub-section describes the key task of extracting quantitative insights from qualitative, unstructured text data. We can't say we extracted the precise sentiment because sentiment is subjective and may differ for each individual, but we can say we estimated the sentiment represented by each article.

To estimate the sentiment for each article, the 'Bag of words' approach was applied. As explained in chapter 2, bag of words is a dictionary-based method of sentimental analysis. The approach converts text data into a Term-document matrix and employs the frequencies of a select few pre-assembled and categorised words to estimate sentiment.
For this approach to be effective, the choice of the word dictionary is of paramount importance as "English words have many meanings, and a word categorization scheme derived for one discipline might not translate effectively into a discipline with its own dialect." - (MCDONALD, 2011)

For example, the word ‘Share’ is associated with positive sentiment and ‘debt’ is associated with negative sentiment in general-purpose English dictionaries but when it comes to business and finance texts, ‘share’ doesn’t really convey any sentiment. It means a ‘piece’ and conveys no sentiment. Similarly, ‘Debt’ cannot be associated with negative sentiment as investment funds and banks invest in debt all the time.
These examples are representative of a wide array of words that are categorised as conveying some sentiment but rather do not convey any sentiment or worst convey the opposite sentiment as categorised.
This inaccurate categorization of words adds severe noise to sentiment estimation and by extension to regressive models estimating its impact. (MCDONALD, 2011)

3.4.1 McDonald and Loughran Dictionary

As a solution to the addressed challenges McDonald and Tim Loughran compiled a domain-specific dictionary that can be broken down into Fin-pos and Fin-neg lists of words, compiled using 10-K reports filed from 1994 to 2008.
The Fin-neg is an exhaustive list of negative sentiment words, exhaustive in a manner that if an informed managerial decision to avoid words in the list is made, it would be very difficult to do so. The list covers almost all forms of words where the root is the same.

The figure below shows the distribution of our domain-specific dictionary:

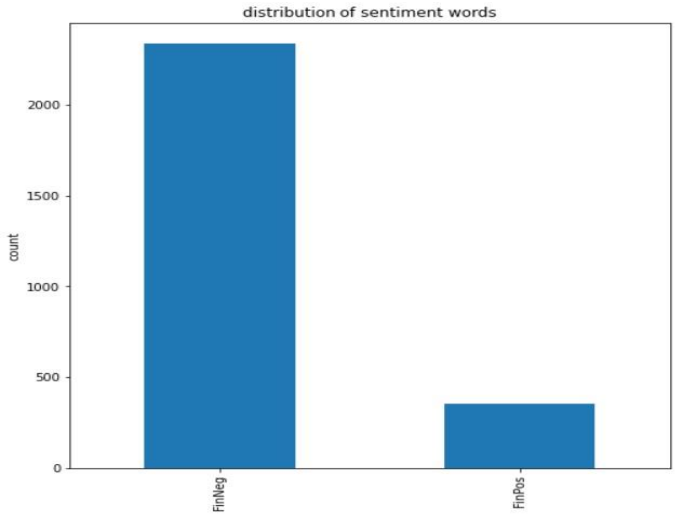


Figure 3.4. 1 Distribution of the dictionary

As one may observe, the Fin-pos list is significantly shorter than the Fin-neg list, owing to the fact that it is common in practice to frame negative news using positive words, whereas the opposite is not true.

The Fin-pos list contains only words of positive potential tone to ensure minimal noise is added to the sentiment estimation process. The authors claim that the Fin-pos list may not be as resourceful in capturing the tone of the document.

Before we began the bag of words analysis, we excluded a set of closed-class words from the analysis by replacing them with a blank char (''), this was done in accordance with the belief that closed-class words do not convey any sentiment and are used to maintain the grammatical structure of a sentence.

3.4.2 Amar Ujala Hindi text: bag of words

After the pre-processing of raw text data in the previous sub-section, a list of 242 stop words was collected from Kaggle, which is an online machine learning and data science community portal (Jha, N, Shenoy, & K R, 2018)Since the authenticity of the dataset was ambiguous, it was thoroughly checked and verified manually by the author and several additional words such as ' फीसदी ' (percentage), 'कार' (car),'कारें' (cars), 'एसयूवी' (SUV), 'वाहन' (Vehicle), etc. were added to the general-purpose list to fit it more appropriately to the text dataset concerning the auto manufacturer.

Before estimating the sentiment, the English financial dictionary was converted to Hindi using the Google translate API and the "os" library in python.
The python script requests the API with translation words at a very high frequency, which the unpaid version of the API does not accommodate and causes runtime errors. To overcome this challenge, the thread making translation requests was forced to sleep for one second after every five translation requests were made.

Now that the Hindi financial and business dictionary was ready, a python script written using the base strings library and pandas library was executed to count the frequencies of each positive and negative word in the dictionary for every article.

Figure showing most frequent positive and negative words

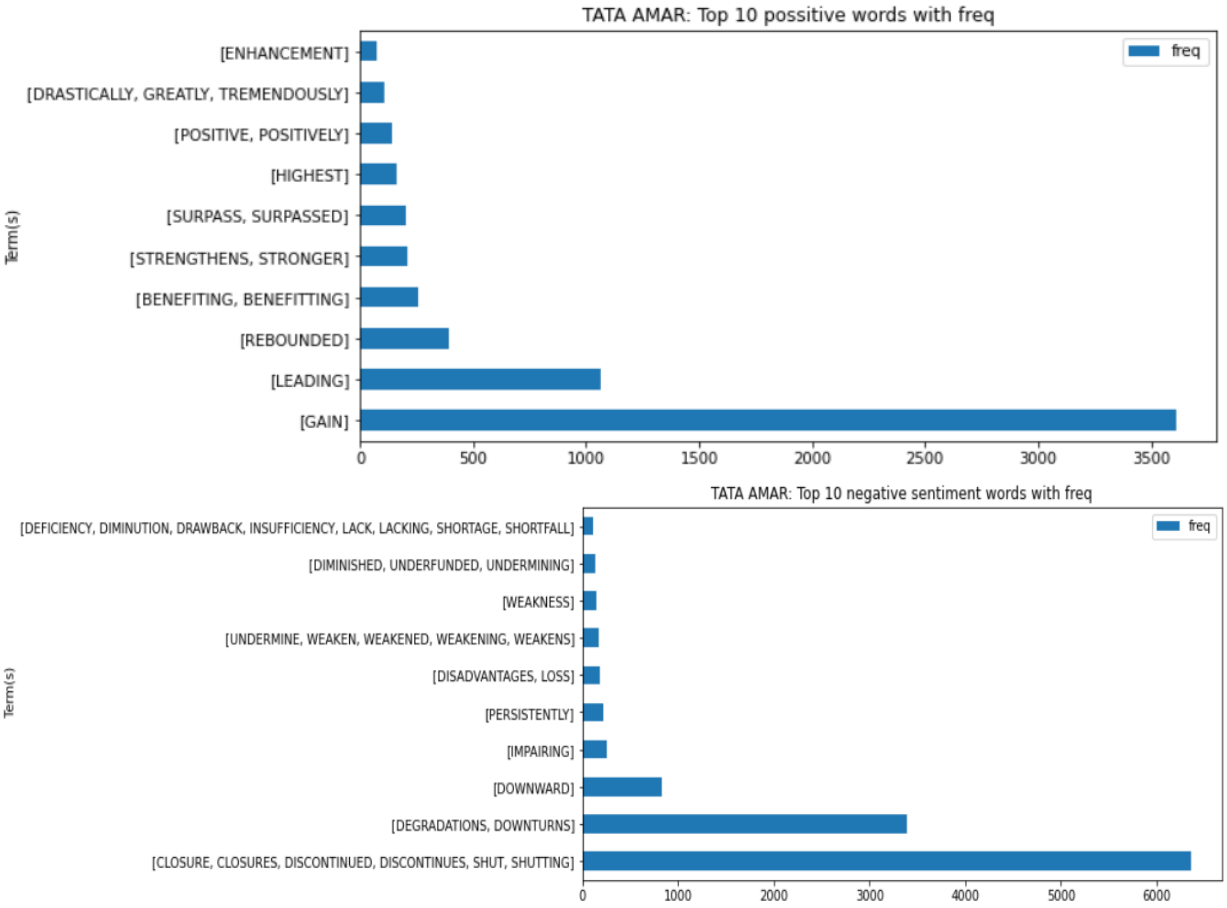


Figure 3.4. 2 Top 10 Hindi positive and negative words by frequency

We can observe a list of terms in place of expected single words(1-gram) in the figure 3.4.2**.** This is due to the fact that a single word in Hindi translates to multiple words in English.

To understand the positivity and negativity reflected in the article text in a relative manner, the percentages of positive and negative words for each article were calculated.

$Percent\ positive = \frac{\sum_{i=0}^{n} Frequency(i)}{N}$ where n = all positively categorized words found in the article and N is the total word count.

$Percent\ negative = \frac{\sum_{i=0}^{k} Frequency(i)}{N}$ where k = all negatively categorized words found in the article and N is the total word count.

After performing the said analysis, the distribution spread of percent positive and percent negative words is shown in the figures shown below:

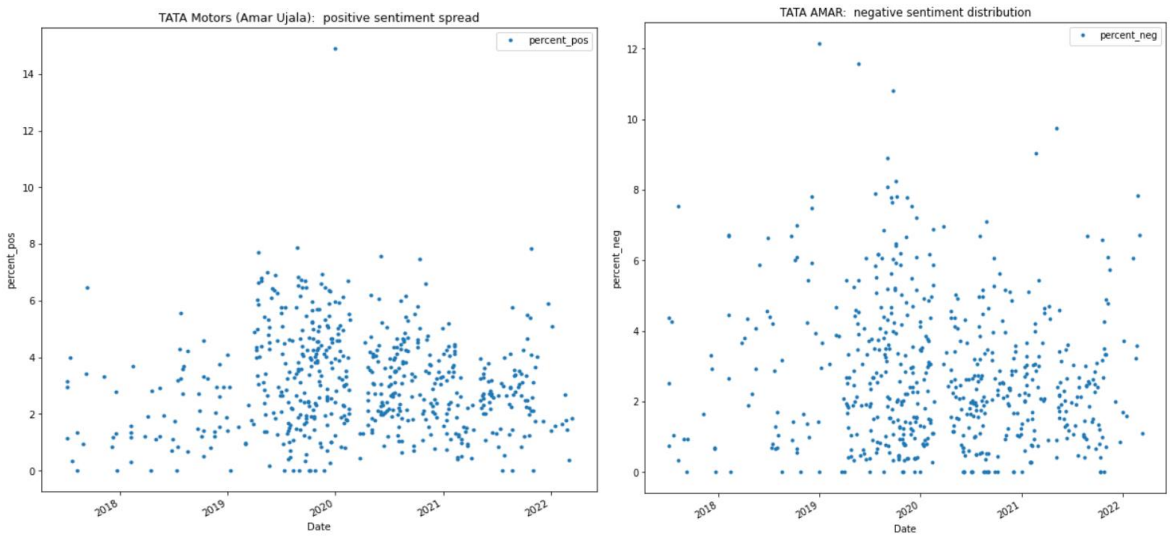


Figure 3.4. 3 Percent of Hindi positive and negative words vs time

3.4.3 Sentiment using company name

This paper also applies an alternate method of sentiment extraction, A variant of work shown in (Tom Martya, 2020), The authors of (Tom Martya, 2020) discuss the importance of estimating sentiment concerning a company on a given day by using the number of articles published on the given day. Based on the assumption that when a company is in upheaval or experiencing remarkable growth, the name of the company appears more frequently in the news.
In a similar vein, instead of using the number of publications, this report uses the frequency of the company's name (Tata Motors) to estimate the severity of the sentiment on a given day.

For the English news text analysis, both the approaches of estimating sentiment have been deployed, a general-purpose English language list of close class words was used and an identical analysis was performed. As a consequence, the results are also similar and will be shared in the next sections.

3.5 Merging datasets

Since the goal of this research is to determine the impact of sentiment on stock prices, the next logical step is to combine the estimated sentiment from news articles with the corresponding market returns based on the publication dates of both. However, before we do that, we must first aggregate the sentiment of all articles published on the same day. To do so, the positive, negative, and total words of articles published on the same date were put together, and net positive and negative percentages were calculated.

The following table shows the number of observations left after the said merger.

Data set	Total Observations	Total words under review	Total articles under review
Amar ujala (Hindi)	421	226079	558
Lexis news (English)	986	1492875	3439

Table 3.5. 1: Final time series data on merging

We observe a sharp downfall in the number of total observations mainly because multiple sentiment observations were aggregated to form a single observation but one must also observe that the data sets are merged at the intersection of dates, which implies that there were many articles published on weekends which had to be dropped because there is no market activity on weekends and conversely the market data had to be dropped on days where there was no article about the company in the news.

We see the final time series comprises unequidistant observations, the distance was observed as follows:

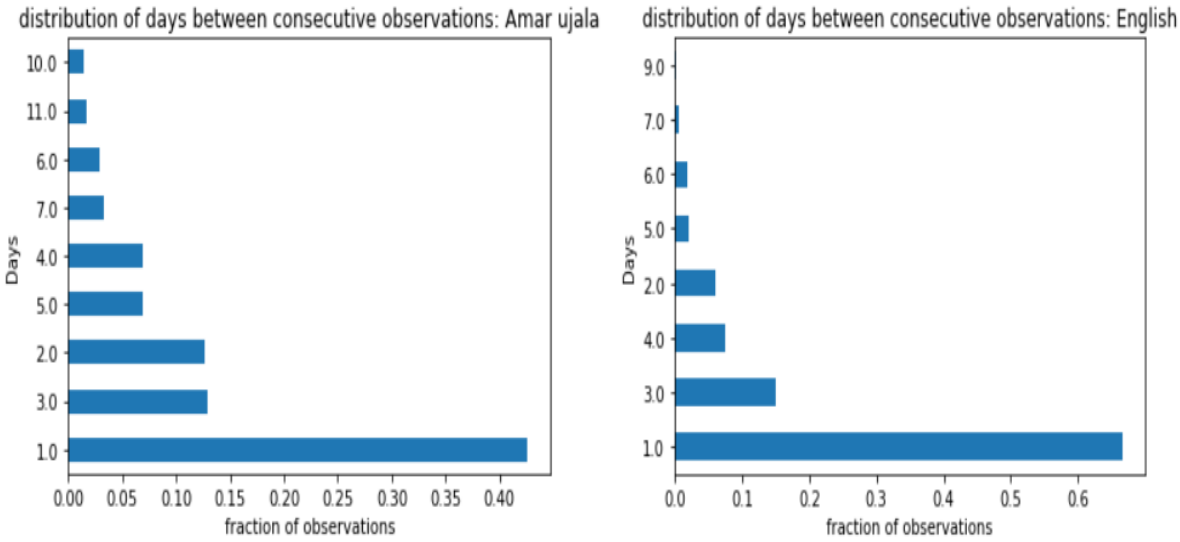


Figure 3.5. 1 Analysing the final time series

We observe that maximum number of observations are 1 day apart in both the data sets but a significant number of observations are more than one day apart. This may bestow additional noise to econometric models that assume equidistant time series while solving for model coefficients.

3.6 Statistical analysis and results

Before we begin with statistical analysis, we calculated the individual z-scores of the features of interest, namely the percentage of positive and negative words and the log returns. This was done to study the impact of change in sentiment on the change in return.
Further analysis using Z-scores gave us a chance to examine how sentiment values on one end of the spectrum influence the returns. Each statistical analysis has a

Analysing the distribution of Z-scores for both the datasets:

Dataset	Z score returns min	Z-score returns max	Z-score negative words percent min	Z-score negative words percent max	Z-score positive words percent min	Z-score negative words percent max	Z-score name frequency words min	Z-score name frequency words max
English	-6.25	6.04	6.01	-1.176	-1.51	5.33	-0.87	9.339
Hindi	-2.83	4.65	-1.36	5.02	-1.87	3.042	-1.26	10.34

Table 3.6.1. 1 Distribution of Z-scores

3.6.1 Linear regression

3.6.1.1 Studying Z-score of log returns vs Z-score of percent negative words in a day

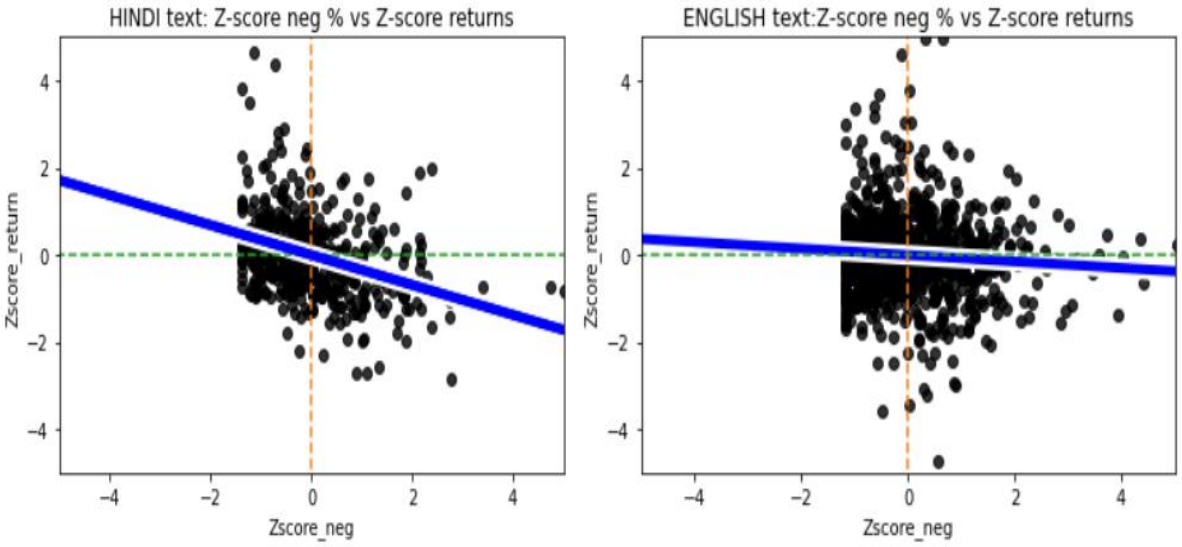


Figure 3.6. 1 Hindi and English Negative words vs Returns (Z-scores)

Dataset	Model slope	Model intercept	Correaltion coefficient(r value)	associated p-value
English	-0.074	0 (approx)	-0.074	0.0196
Hindi	-0.34	-5.45 (10^{-17})	-0.34	4.13(10^{-13})

Table 3.6.1. 2 Z-score of log returns vs Z-score of percent negative words

In figure 3.6.1 shown above we see a negative slope between the returns and negative percentage of words.

The negative correlation essentially means that when the Z-score returns are possitive, the z-score for negative words tends to be negative and when the Z score retrurns are negative, the z-score for negative returns tends to be possitive.

The visualizations shown above are split into four quadrants, the coefficient of correlation highlights the quadrants of interest are the top left quadrant and the bottom right quadrant.

Creating a normalized contingency table for the variables under analysis for the Hindi text we see that:

Z-score negative words	Negative	Positive	Total
Z-score of returns			
Negative	0.2327 (Quadrant- III)	0.2755 (Quadrant-IV)	0.50831
Positive	0.3444 (Quadrant-II)	0.1472 (Quadrant-I)	0.4916
Total	0.577	0.422	1.00

Table 3.6.1. 3 Contingency table for returns vs negative Hindi text

We see that more points lie in quadrant-II and quadrant-IV than the rest. Next, we calculate the conditional probability of returns being positive given that the z-score of negative words was negative.

P (z-score returns = pos | z-score negative = neg)

P (z-score returns =pos | z-score negative= neg) $= \frac{P((Z\text{-}score\ returns = pos) \cap (Z\ score\ negative\ words = neg))}{P(Z\ score\ negative = neg)}$

= 0.344/0.577

P (z-score returns = pos | z-score negative = neg) = 0.6 **>** 50%

P (z-score returns = neg | z-score negative = pos) = 0.65 **>** 50%

Performing a Chi-squared distribution test of independence using the contingency table, we found the p-value to be 7.91×10^{-7}.

Similarly calculating a contingency table between z-scores of returns and negative words for English text

Z-score negative words	Negative	Positive	Total
Z-score of returns			
Negative	0.271805 (Quadrant- III)	0.235294 (Quadrant-IV)	0.507099
Positive	0.3052 (Quadrant-II)	0.1876 (Quadrant-I)	0.4929

Total	0.577	0.422	1.00

Table 3.6.1. 4 Contingency table for Returns vs negative English text

We see that more points lie in quadrant-II and quadrant-IV than the rest. Next, if we calculate the conditional probability of returns being positive given that the z-score of negative words was negative

P (z-score returns = pos | z-score negative = neg) = 0.52 **>** 50%

P (z-score returns = neg | z-score negative = pos) = 0.556**>** 50%

Performing a Chi-squared distribution test of independence using the contingency table **1456656**, we found the p-value to be 0.0097.

3.6.1.2 Studying Z-score of log returns vs Z-score of percent positive words in a day

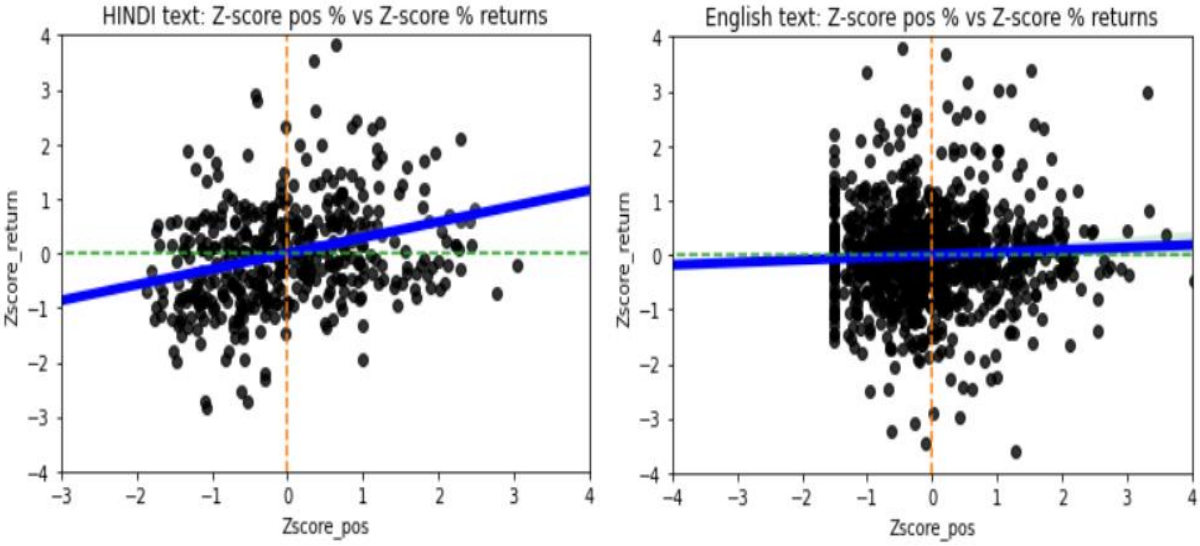


Figure 3.6. 2Hindi and English positive words vs Returns (Z-scores)

Dataset	Model slope	Model intercept	Correaltion coefficient(r value)	associated p-value
English	0.046	6.5 (10^{-19})	0.046	0.1475
Hindi	0.288	-7.287 (10^{-17})	0.288	1.63(10^{-09})

Table 3.6.1.2. 1 Regression of Z-score Returns vs Z-score of positive words

In the figures shown above we see a positive slope between the returns and positive percentage of words.

The positive correlation essentially means that when the Z-score returns are possitive, the z-score for positive words tends to be positive as well and visa versa.

The visualizations shown above are split into four quadrants, the coefficient of correlation highlights the quadrants of interest are the Ist and the IIIrd quadrant.

Creating a normalized contingency table for the variables under analysis for the Hindi text we see that:

Z-score positive words	Negative	Positive	Total
Z-score of returns			
Negative	0.318290 (Quadrant- III)	0.190024 (Quadrant-IV)	0.508314
Positive	0.230404 (Quadrant-II)	0.261283 (Quadrant-I)	0.4916
Total	0.548694	0.45130	1.00

Table 3.6.1.2. 2 Contingency table for returns vs positive Hindi text

P (z-score returns = neg | z-score positive = neg) = 0.57 **>** 50%

P (z-score returns = pos | z-score positive = pos) = 0.57 **>** 50%

Performing a Chi-squared distribution test of independence using the contingency Table 3.6.1.2. 3, we found the p-value to be 7.91×10^{-07}

Similarly, calculating a contingency table between z-scores of returns and positive words for English text

Z-score negative words	Negative	Positive	Total
Z-score of returns			
Negative	0.2870 (Quadrant- III)	0.22 (Quadrant-IV)	0.507099
Positive	0.2687 (Quadrant-II)	0.2241 (Quadrant-I)	0.4929
Total	0.555	0.444	1.00

Table 3.6.1.2. 4 Contingency table for returns vs positive English text

P (z-score returns = neg | z-score positive = neg) = 0.517 **>** 50%

P (z-score returns = pos | z-score positive = pos) = 0.504**>** 50%

Performing a Chi-squared distribution test of independence using the contingency Table 3.6.1.2. 5, we found the p-value to be 0.55.

3.6.1.3 Studying Z-score of log returns vs Z-score of Tata motors frequencies in a day

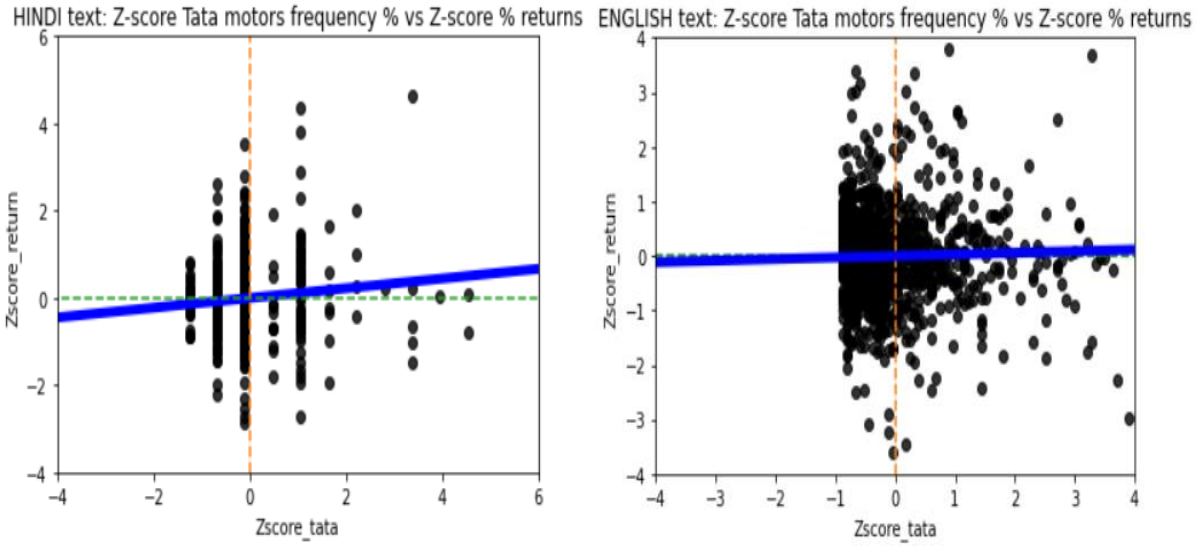


Figure 3.6. 3 Hindi and English 'Tata motors' frequency vs Returns (Z-scores)

Dataset	Model slope	Model intercept	Correaltion coefficient(r value)	associated p-value
English	0.028	-4.15 (10^{-19})	0.028	0.36
Hindi	0.11	-2.06 (10^{-17})	0.11	0.022

Table 3.6.1.3. 1 Z-score of log returns vs Z-score of Tata motors frequencies

In the figures shown above we see a positive slope between the returns and the company name frequency. For the English text

Similar to before, We'll investigate it further using contingency tables.

Z-score Tata Motors frequency	Negative	Positive	Total
Z-score of returns			
Negative	0.4133 (Quadrant- III)	0.095 (Quadrant-IV)	0.5083
Positive	0.3871 (Quadrant-II)	0.104 (Quadrant-I)	0.4916
Total	0.8004	0.199	1.00

Table 3.6.1.3. 2 Contingency table for returns vs Hindi text Tata frequency

P (z-score returns = neg | z-score tata frequency = neg) = 0.516 **>** 50%

P (z-score returns = pos | z-score tata frequency = pos) = 0.52 **>** 50%

Performing a Chi-squared distribution test of independence using the contingency Table 3.6.1.3. 3, we found the p-value to be 0.66.

Similarly, calculating a contingency table between z-scores of returns and Tata motors frequency for English text

Z-score Tata motors frequency	Negative	Positive	Total
Z-score of returns			
Negative	0.335 (Quadrant- III)	0.171 (Quadrant-IV)	0.507099
Positive	0.320 (Quadrant-II)	0.172 (Quadrant-I)	0.4929
Total	0.656	0.343	1.00

Table 3.6.1.3. 4 Contingency table for returns vs English text Tata frequency

P (z-score returns = neg | z-score tata motors frequency= neg) = 0.51 **>** 50%

P (z-score returns = pos | z-score tata motors frequency = pos) =0.501 ≈ 50%

Performing a Chi-squared distribution test of independence using the contingency table 3.6.1.3. 5, we found the p-value to be 0.74.

From the Chi-squared test values of both positive and negative sentiment for English and Hindi datasets we observe that return is not independent of both the percentage of positive and negative words.

Next, we utilise this property and model a regression function where returns are a linear a combination of both percentage of positive words and negative words published on a day.

3.6.2 VAR model

Now we begin with the econometric modelling of the two time series, the sentiment (positive and negative) on a given day and the market returns. We aim to inspect the impact of lag 1 values of sentiment and lag 1 values of returns on the current return.

Before we apply the model, the VAR function in the statsmodels library expects the input series to be stationary. We check the stationarity of the series using the augmented dickey-fuller test for stationarity**.**

The results are as follows:

Dataset	Variable	Test statistic	Associated p-value
Amar Ujala (Hindi)	Positive words (Z-score)	-11.11	$3.66(10^{-20})$
	Negative words (Z-score)	-17.51	$4.291(10^{-30})$
	Returns (Z-score)	-7.3990	$7.906(10^{-11})$
Lexis news (English)	Positive words(Z-score)	-6.5	$7.1(10^{-09})$
	Negative words (Z-score)	-6.83	$1.82\ (10^{-09})$
	Returns (Z-score)	-31.477	0(approx.)

Table 3.6.2. 1 augmented dickey-fuller test results

Analysing the test statistics and associated p-values we can claim all the series are stationary and do not require any further differencing or application of other methods of conversions.

Moving forward and applying a VAR (lag 1) to model the returns as a consequence of variables shown in Table 3.6.2. 2

3.6.2.1 Returns vs negative Hindi sentiment lag(1)

$Returns_t = A_{1,1}Returns_{t-1} + A_{1,2}(\text{negative sentiment estimate})_{hindi,t-1} + E_1$

$\text{Negative sentiment estimate}_t = A_{2,1}\,Returns_{t-1} + A_{2,2}\,\text{Negative sentiment estimate}_{t-1} + E_2$

The results obtained for the returns using this model are shared in the next chapter in Table 3.6.2. 3

3.6.2.1 Returns vs positive Hindi sentiment lag(1)

$Returns_t = A_{1,1}Returns_{t-1} + A_{1,2}(\text{positive sentiment estimate})_{hindi,t-1} + E_1$

$\text{Positive sentiment estimate}_t = A_{2,1}\,Returns_{t-1} + A_{2,2}\,\text{positive sentiment estimate}_{t-1} + E_2$

The results obtained for the returns using this model are shared in the next chapter in Table 3.6.2. 4

3.6.2.1 Returns vs negative English sentiment lag(1)

$Returns_t = A_{1,1}Returns_{t-1} + A_{1,2}(\text{negative sentiment estimate})_{English,t-1} + E_1$

$\text{Negative sentiment estimate}_t = A_{2,1}\,Returns_{t-1} + A_{2,2}\,\text{Negative sentiment estimate}_{t-1} + E_2$

The results obtained for the returns using this model are shared in the next chapter in Table 3.6.2. 6

3.6.2.1 Returns vs Positive English sentiment lag(1)

$Returns_t = A_{1,1}Returns_{t-1} + A_{1,2}(\text{Positive sentiment estimate})_{English,t-1} + E_1$

$\text{Negative sentiment estimate}_t = A_{2,1}\, Returns_{t-1} + A_{2,2}\, \text{Negative sentiment estimate}_{t-1} + E_2$

The results obtained for the returns using this model are shared in the next chapter in Table 3.6.2. 8

3.6.3 Pan linguistic analysis

This section aims to assess the impact of positive and negative sentiment across Hindi and English languages in unison. Two approaches have been used to model this.

Since negative sentiment from both the datasets disproved the hypothesis of independence in chi-squared distribution, we fit a VAR (1) model and a multiple regression model with returns as the dependant variable and the negative sentiments from both datasets as independent variables.

In the second approach we calculate the positivity and negativity by combining the datasets and re calculating the positive and negative percentages of words followed by a z-score analysis following the work shown in the preceding sub-sections.

On merging the two data sets on date, the resultant data is shown in the following table.

Total observations (Days)	Language	Number of articles	Number of words
365	Hindi	554	195,570
	English	3439	581,508
	Total	3993	777,078

Table 3.6.3. 1 Pan-lingual Dataset

Approach 1: Treating sentiment from different languages as separate variables

Using negative sentiment z-score for both English and Hindi articles as predictor variables a multiple linear regression model was fit on the pan-lingual data set.

Constants	Values	p-values

Intercept	0.0012	0.98
Negative English	-0.0367	0.458
Negative Hindi	-0.3399	1.495086e-11

Table 3.6.3. 2 Multiple regression coefficient values

Associated R^2 Value for the model: 0.12

VAR (lag 1) model with three variables: lagged returns, lagged negative English score and lagged negative Hindi score.

The results are shared in the subsequent section.

Approach 2: calculating aggregate positive and negative words

We calculated aggregate percentage of positive and negative words printed on a day and further converted them into z-scores. We also calculated the z-scores of the new dataset, it's important to note that the percentage of change or log returns on the same date were the same but the Z-scores of returns varied as the population mean for English and Hindi datasets was not the same.

	Z-score negative words	Z-score positive words	Z-score returns
Max value	4.5	3.88	4.61
Min value	-1.06	-1.73	-2.85

Table 3.6.3. 3: Pan-linguistic regression models

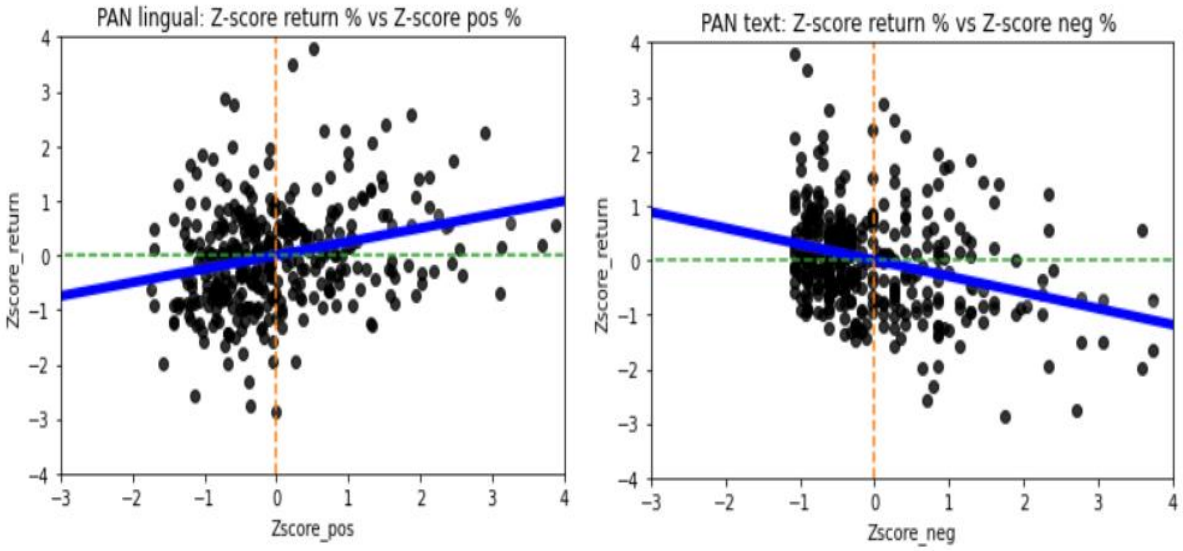


Figure 3.6. 4 Pan-linguistic aggregate positive and negative words vs returns(z-scores)

model	intercept	slope	r-value	Associated p-value
Return vs negative words	$2.05(10^{-17})$	-0.295	-0.295	$8.7(10^{-09})$
Return vs positive words	$-6.01(10^{-17})$	0.248	0.248	$1.50(10^{-06})$

Table 3.6.3. 4 Coefficients of pan-linguistic regression models

Observing the polarity of the r-vlaues of the regression models we further focus on quadrants I and III and Quadrants II and IV for negative and possitive models respectively.

Z-score Negative words	Negative	Positive	Total
Z-score of returns			
Negative	0.2904 (Quadrant- III)	0.2191 (Quadrant-IV)	0.5095
Positive	0.3397 (Quadrant-II)	0.1506 (Quadrant-I)	0.4904
Total	0.6301	0.3698	1.00

Table 3.6.3. 5 Contingency table for returns vs pan linguistic sentiment (negative)

Performing a Chi-squared distribution test of independence using the contingency table 3.6.3. 6, we found the p-value to be 0.02.

Z-score Tata Motors frequency	Negative	Positive	Total
Z-score of returns			
Negative	0.3479 (Quadrant- III)	0.1616 (Quadrant-IV)	0.5095
Positive	0.2575 (Quadrant-II)	0.2328 (Quadrant-I)	0.4904
Total	0.6054	0.3945	1.00

Table 3.6.3. 7 Contingency table for returns vs pan linguistic analysis(positive)

Performing a Chi-squared distribution test of independence using the contingency table 3.6.3. 8, we found the p-value to be 0.00208.

Chapter 4: Results

This chapter dwells on the inference gathered from the statistical analysis performed in previous chapter and discusses the statistical significance of each of the results.

4.1 Linear regression

We fitted linear regression models between three separate variables and the returns.

4.1.1 Z-score Negative words vs Returns

We saw a negative slope and a negative Pearson correlation coefficient for both the datasets. The P-values for the coefficients were less than 0.05 hinting to the results being statistically significant. Which essentially implies that higher Z-scores of returns are observed when lower z-scores of negative words are found. That is to say on a given day, if the number of negative words is less than the population average, we see returns higher than average.

Next using consistency tables, Table 3.6.1. 5 and Table 3.6.1. 3 we calculated conditional probabilities and observed the probabilities were higher than 50% reflecting the same fact. Calculating P-values using the contingency tables and chi-squared test for independence we observed the p-values to be less than 0.05, hence, we rejected the null hypothesis claiming negative words and the stock returns on a given day are independent.

However, we must note that the magnitudes of both the probabilities and the magnitude of coefficient of correlation were higher for the Hindi dataset when compared with the English dataset.

4.1.2 Z-score positive words vs Returns

We observed a positive correlation coefficient and a positive slope for both the datasets. The p-value for the Hindi coefficient was less than 0.05, making the correlation coefficient statistically significant. However, the p-value for the English dataset was 0.14, making it greater than 0.05, and not statistically significant.

The positive slopes essentially mean that when greater z-scores of positive terms are identified, higher Z-scores of returns are observed. That is, if the number of positive words published on a given day exceeds the population average, we generally get higher-than-average returns.

We then calculated conditional probabilities of getting higher than average return, given that we had more than average number of positive words in the news. Both the probabilities came out to be greater than 50% for the Hindi dataset, but were close to 50% for the English dataset hinting that the number of positive words on a given day in English news, may not influence the stock returns.

Further we calculated chi-squared distribution tests for both the datasets using the constructed contingency tables (Table 3.6.1.2. 6 and Table 3.6.1.2. 7) and saw the p-value for the Hindi dataset was far less than level of significance and the p-value for the English data set was 0.55 significantly higher than 0.05 and therefore we may conclude that the results for the Hindi dataset are significant and the results for English dataset are not statistically significant.

4.1.3 Z-score 'Tata motors' frequency vs Returns

We observed a positive correlation coefficient and a positive slope for both the datasets. The p-value for the Hindi coefficient was less than 0.05, making the correlation coefficient statistically significant. Whereas the p-value for the English dataset was substantially greater

than the significance level, we can not say the result is statistically significant for the English dataset.

We then calculated conditional probabilities of getting higher than average return, given that the company name was mentioned more than average number of times in the news and the probability of getting a lower-than-average return, given that the company name in the news was less than usual. Both the probabilities came out to be very close to 50% for both the datasets, hinting that the frequency of the company name on a given day, may not influence the stock returns

Further we calculated chi-squared distribution tests for both the datasets using the constructed contingency tables (Table 3.6.1.3. 6 and Table 3.6.1.3. 7) and saw the p-values for both the datasets were far greater than level of significance, therefore, we can conclude that the frequency of the company name for these datasets does not have an impact on the returns and will not be used in any subsequent analysis.

4.2 Var models

Lag 1 VAR models were implemented using both positive words and negative words Z-scores to model the returns.

Constant	Coefficient	Std. error	t-stat	p-value
Returns(Z-score) with lag (1)	-0.00208	0.052175	-0.040	0.968
Negative words (Z-score) with lag (1)	-0.01785	0.0522	-0.342	0.732
Additive constant	-0.000116	0.048963	-0.002	0.998

Table 3.6.2. 3 VAR summary table for Returns vs Hindi text analysis (Negative)

Table 3.6.2.2 shows the VAR model summary for Hindi negative words. The p-values for all the coefficients are significantly higher than the level of significance. We can conclude that the results for the model are not statistically significant.

Constant	Coefficient	Std. error	t-stat	p-value
Returns(Z-score) with lag (1)	-0.0202997	0.050980	-0.398	0.691
Positive words (Z-score) with lag (1)	0.084684	0.051007	1.660	0.097
Additive constant	-0.000293	0.048809	-0.006	0.995

Table 3.6.2. 4 VAR summary table for Returns vs Hindi text analysis (positive)

Table 3.6.2. 5 shows the VAR model summary for Hindi positive words. The p-values for all the coefficients are higher than the level of significance. We can conclude that the results for the model are not statistically significant.

Constant	Coefficient	Std. error	t-stat	p-value
Returns(Z-score) with lag (1)	-0.00557	0.031994	-0.174	0.862
negative words (Z-score) with lag (1)	-0.020916	0.031992	-0.654	0.513
Additive constant	-0.000444	0.031901	-0.014	0.989

Table 3.6.2. 6 VAR summary table for Returns English text analysis (negative words)

Table 3.6.2. 7 shows the VAR model summary for English negative words. The p-values for all the coefficients are higher than the level of significance. We can conclude that the results for the model are not statistically significant

Constant	Coefficient	Std. error	t-stat	p-value
Returns(Z-score) with lag (1)	-0.006010	0.031916	-0.188	0.988
Positive words (Z-score) with lag (1)	-0.043687	0.031912	-0.188	0.851
Additive constant	-0.000463	0.031877	-0.015	0.988

Table 3.6.2. 8 VAR summary table for Returns vs English text analysis (positive words)

Table 3.6.2. 9 shows the VAR model summary for English positive words. The p-values for all the coefficients are higher than the level of significance. We can conclude that the results for the model are not statistically significant

4.3 Pan-linguistic modelling

4.3.1 Approach 1: *separate languages from different languages as separate variables*

We applied multiple linear regression by using the estimated negative sentiments of both the languages as independent variables to estimate returns. We observed that p-values associated with the calculated coefficients were greater than the level of significance except for the Hindi negative sentiment coefficient.

The coefficient of determination for the model was 0.12, which means the model approximately explains only 12% of the variance in returns.

Further we applied a lag 1 VAR model, using negative sentiment estimates from both the datasets.

Constant	Coefficient	Associated p-value
Additive constant	-0.000651	0.99
Return (Z-score) (lag =1)	-0.18740	0.74

Negative Hindi words (Z-score) (lag =1)	0.019194	0.73
Negative English words (Z-score) (lag =1)	-0.008468	0.873

Table 3.6.3. 9 VAR model coefficient values

Table 3.6.3. 10 shows the VAR model summary. The p-values for all the coefficients are higher than the level of significance. We can conclude that the results for the model are not statistically significant

4.3.2 Approach 2: using aggregate positive and negative words

From Table 3.6.1.3. 4 we observe the slope values for positive words and negative words are positive and negative respectively. Which is similar to the individual analysis. The P-values for the coefficients were less than 0.05 hinting to the results being statistically significant.

Calculating P-values using the contingency tables and chi-squared test for independence we observed the p-values to be less than 0.05, hence, we rejected the null hypothesis claiming negative words and positive words, and the stock returns on a given day are independent.

Chapter 5: Conclusions and future work

5.1 Conclusions

This research paper applies the bag of words technique to estimate sentiment from Hindi and English news articles collected using web-scraping and from a corporate data provider, respectively.

We look at articles consumed by different socio-economic communities to model stock price changes. We establish that sentiment indeed impacts stock market returns.

we observe, that the more the negative words published in a day, the less the stock returns on that day.

We observe that both positive and negative sentiments estimated from Hindi words statistically significantly influence stock prices, whereas only negative sentiment estimation from English articles statistically significantly impacts the stock returns. Although we successfully established a relationship between sentiment and returns on the same day, we couldn't establish any statistically significant relationship between returns and lagged stock returns and lagged estimated sentiment.

For the English analysis, the statement claiming “The tone of negative words has a much more pervasive effect” by Loughran and McDonald has proven true.

We looked at two ways of combining estimated sentiment for pan-linguistic analysis and observed both estimated positive and negative pan-linguistic sentiments and return have a statistically significant impact. We observed that modelling accumulative positive and negative sentiment on a given day gives statistically significant results resonating with the individual findings of the paper whereas pan-linguistic modelling using negative sentiment words from different languages doesn't give statistically significant results for these datasets

5.2 Future Work

Lately, there has been a cultural shift in the Hindi language, where the informal language has adopted a lot of English words, to the extent that the general public and even some newspaper articles use Hinglish (English + Hindi) in their writing. This research paper doesn't account for Hinglish words, that is English words written using Hindi alphabets, which may have led to missed signals. Any future study can examine the significance of such terms in their analysis.

To generate a Hinglish dictionary, one may create a corpus level dictionary and use co-occurrence analysis to gather words. The created dictionary can be cross-validated with a domain-specific dictionary by any native speaker of the languages.

We applied two approaches to combine sentiment estimated from different sources and languages, going forward more research can be done to devise and experiment with ways of pan-linguistic analysis.

We have studied the impact of sentiment only on stock returns, future research can also study the impact of other financial and economic factors in conjunction with estimated sentiment to model stock returns.

References

Ahmad, K. (2021). *Did The Market Move For You? Artificial Intelligence, Financial, & Commodity Trading.* Z/Yen Group.

Berkley), S. Y. (2020). *Stock Price Forecasting Using Information from Yahoo Finance and google trends.*

D'Andrea, A. &. (2015). Approaches, Tools and Applications for Sentiment Analysis Implementation. *International Journal of Computer Applications*.

Forbes. (2019). *https://www.forbes.com/sites/bernardmarr/2018/05/21how much data do we create every day the mind blowing stats everyone should read*. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=64ba024360ba

India, A. B. (2020). *Audit Bureau of Circulation India*. Retrieved from http://www.auditbureau.org/files/JJ2018%20Highest%20Circulated%20amongst%20ABC%20Member%20Publications%20(across%20languages).pdf

Jaap Kamps, M. M. (2004). Using WordNet to Measure Semantic Orientations of Adjectives.

Jha, V., N, M., Shenoy, P. D., & K R, V. (2018). *Hindi language stop words*. Retrieved from https://data.mendeley.com/datasets/bsr3frvvjc/1

Liu, B. (2012). *Sentiment Analysis-and-OpinionMining.*

MCDONALD, T. L. (2011). When is a liability not a liability. *Journal of finance*.

Mohammad Darwich, S. A. (2019). Corpus-Based Techniques for Sentiment Lexicon Generation: A Review. *Journal of Digital Information Management*.

Murray, T. (n.d.). *The structure of English.*

Nordquist, R. (2020, 08). *"Closed Class Words."*. Retrieved from thoughtco: thoughtco.com/what-is-closed-class-words-1689856

Nordquist, R. (2020, 08 25). *Open class words in english grammar*. Retrieved from ThoughtCo: thoughtco.com/open-class-words-term-1691454

normal distribution. (n.d.). Retrieved from statisticshowto: : https://www.statisticshowto.com/probability-and-statistics/normal-distributions/#whatisND

Tom Martya, B. V. (2020). News media analytics in finance: a survey.

university, p. (n.d.). *what is wordnet?* Retrieved from https://wordnet.princeton.edu/

Zhan, X. F. (2015). Sentiment analysis using product review data. *Journal of Big Data*.