Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Political Sentiment Analysis: Investigating the Impact of State Elections on Indian National Elections using print media

Ayush Singhania
[redacted]

May 3, 2020

Supervisor: Prof. Khurshid Ahmad

A Final Year Project submitted in partial fulfilment
of the requirements for the degree of
MAI (Computer Engineering)

Declaration

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

Signed: [redacted] Date: May 3, 2020

Abstract

There is an exponential rise in the volume of media content generated and simultaneously consumed by the general public through a wide range of channels from legendary media such as television, radio and print-based newspaper, magazines to developing online platforms such as social networks, podcast and mobile applications on the internet, in the 21st century. This aim of this research was to examine the newspaper generated by Indian print media for months before leading up to the State Legislative Election and 2019 General Election held in three states, Chhattisgarh, Punjab and Karnataka which served as the three case studies. Politics is deeply associated with the electorate's perception and the media plays a pivotal role in shaping it. Combating prejudice against a political leader or a party is a challenging task in current multicultural societies such as that of India. Mass media can decisively shape prejudice and electoral 'mood' because it often represents the main source of information for general masses. There is very little research that captures the causal effect of different media features on attitudes toward different important political events, such as granular state legislative elections and its effect on the General Parliamentary elections. This research work distinguishes it from other work conducted in a similar domain by using political sentiment expressed in news articles during several State Elections held in India and uses it as a proxy for observing the impact it had on the Indian General Election using the sentiment and political features extracted using sentiment Analysis.

The newspaper data was collected using an authoritative third-party data aggregator, LexisNexis. The sentiment analysis was performed using a dictionary-based lexicon approach in conjecture with a "bag-of-word" and vector space model. The Rocksteady affect analysis system was used to calculate negative sentiment affect time series data for all three case studies. Descriptive statistical analysis along with correlation heatmap and newspaper polarity were conducted using the political factor sentiment time-series generated by the program. The statistical significance of the features extracted from the news articles allowed this study to compare the results of two elections. The key finding from this study indicated the rise in overall negative sentiment communicated in the newspaper in the two elections. It observed that the winners of state elections saw a reduction in the negative sentiment expressed against them in subsequent general elections. A rise in coverage for the winning party in the state election was also noted across all case studies. This lower negative sentiment and higher media coverage do not necessarily translate to victory. Political features such as political leaders and parties in the news articles also had a considerable effect on the results. Although state election influenced the sentiment expressed in the English newspaper which subsequently shaped the electoral perception, with the given data and results observed for the scope of this study it is not reasonable to extrapolate the results of state election and predict the results of the national-level General Election in India.

Acknowledgements

I would like to take this opportunity to thank several people who made this dissertation possible. First and foremost, I would like to thank Professor Khurshid Ahmad and express my sincerest gratitude for his supervision in this project. His constant support and encouragement were a driving force throughout the study. His passion and expertise in this field are unmatched, his advice helped navigate through various hurdles presented during the research and his feedback proved remarkably valuable.

I would also like to express my gratitude towards Dr Mike Brady, the MAI Computer Engineer coordinator in the School of Computer Science and Statistics and my tutor Dr Michael Monaghan and Prof. Glenn Strong who were quick to respond to my queries throughout this journey. Next, I would like to thank my fellow classmates whom I interacted with throughout the year and the insightful discussion which proved very helpful. A big thanks to all my friends whose contact support and love were a huge encouragement through this difficult and unprecedented times which all of us are facing in our lives. This shall pass too.

I would also like to acknowledge Jasmeet Singh and Pranav Jain for sharing their experience and words of wisdom from their experience in the areas of study. I would like to also express my sincere gratitude towards all the front-line healthcare and other thousands of other workers who are keeping us all safe during this pandemic and allowing us to work in peace in the luxury of our homes. We are indebted forever. Gratitude must be expressed to the entire faculty of Trinity College and support staff for helping us smoothly transition to an online environment during these extremely difficult times.

Finally, I must express my heartfelt appreciation to my family, my mom, dad, sister, brother and my girlfriend for providing me with unequivocal and unparalleled love and support and continuous motivation. I am forever obliged to you all for being my support system. This research would have not been possible with the contribution of all the people and many others whom I could not mention but helped me steer through this journey. Thank you to all.

Contents

List of Figures

List of Tables

Nomenclature

BJP	Bharatiya Janta Party
BSE	Bombay Stock Exchange
CAGR	Compound annual growth rate
CG	Chhattisgarh
GE	General Election
IRS	Indian Readership Survey
INC	Indian National Congress
KA	Karnataka
MLA	Member of Legislative Assembly
MP	Member of Parliament
MRUC	Media Research Users Council
NLP	Natural Language Processing
PB	Punjab
TF-IDF	term frequency–inverse document frequency

1 Introduction

With the evolution of the world wide web and development of cheaper personal computer, laptops and smart mobile phones, people have changed their way of consuming knowledge and information. In this era of constant updates and this generations hunger to stay social connected and informed promting them to absorb as much news, from as many sources, as they possibly can, on matters that are essential to them or matters that catch their attention.

The origin and widespread deployment of the internet largely aided these breakthroughs. News publication and magazines brought content online, allowing for larger volumes of content to be published. New platforms produce enormous amounts of information such as social networking and microblogging website. Twitter, Facebook and Whatsapp to name a few. The democratization of the technology and advent of open source communities facilitated widespread public access to this information, thus resulting in an ocean of media content from both mainstream media sources and personal users alike. It enabled members of the common public to voice their opinions on a global stage revealed to the rest of the planet. Even as the move to digital media has challenged newspapers and broadcasters in better part of the world, the print and television news industry in India has continued to grow. With much of the population still offline, hundreds of millions of Indians still turn to newspapers, television and radio as their main sources of news [1].In fact, newspapers are the biggest form of print media in India with more than 425 million daily readers according to the Indian Readership Survey (IRS) [2].

Politics is deeply associated with the electorate's perception and the media plays a pivotal role in shaping it. We are looking at events with a large quantity of streaming information which is dynamic. Public generates emotions, opinions, sentiments, evaluations, appraisals and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [3]. General during our daily news coverage we observe heated correspondences between the opposition party and the ruling party where the ruling party is implicated of wrongdoing and later response from the government and it goes back and forth year-round.

Researchers from Johns Hopkins University found that the brain can assign value to an object in less than a tenth of a second [4]. With such high tendencies of humans to evaluate things, figuring out if something is quality or junk and assigns a value to that information. positive or negative [4]. Humans tend to assess any information projected at them very quickly and try to quantify it, with newspaper and media sentiment it's the same. Either a positive or negative sentiment is attached to it. Combating prejudice against a political leader or a party is a challenging tasks in current multicultural societies. Mass media can decisively shape prejudice and electoral 'mood' because it often represents the main source of information for general masses. Despite the impact of mass media on shaping public opinion, there is very little research that captures the causal effect of different media features on attitudes toward different important political events, such as granular state legislative elections and its effect on the General Parliamentary elections [5].

1.1 Media and Sentiments

1.1.1 Political Sentiment Analysis

The political landscape of any country is massively modelled through its media. Political parties reach its user base by leveraging the mass media and this connection linking the two becomes more profound during the campaigning period leading up to the polls. The term 'Political Sentiment Analysis' has been relatively newer. But, researchers interest the towards sentiment expressed in media coverage can influence how citizens think about their political leader and shape electoral perception is old. An extensive amount of research [6] [7] [8] has been done in this field to investigate sentiment expressed during political events such as election by various news publications, magazines, social media, TV etc. Event these event information flow changes and decisions made based on that information also changes. With the effect of such networks daily, we start recoiling our valuable decisions and actions with certain predefined notions created by others through their opinions expressed during these events in the media newspaper. Thus, it is both interesting and challenging to see what and how the trends in such dynamic political events change from time to time.

Most work on opinion mining has been carried out on subjective text types such as blogs and product reviews. Authors of such text types typically express their opinion freely. The major difference between news and product reviews is that the target of the sentiment is less concrete and is expressed much less explicitly. Another major difference is Newspaper articles is that they give an impression of objectivity to refrain from using positive or negative vocabulary. They resort to other means to express their opinion, like embedding statements in a more complex discourse, omitting facts that highlight some important people. For this reason sentiment analysis on news, the text is rather difficult compared to others.

1.1.2 Technical Challenge

The main technical challenges associated with such a kind study was identifying a kind of information to collect, in which language and why? How to harness such information and its management. Once the answer to these questions has been answered. How to analyse it? In this paper, we chose the challenging job of mining opinions on news articles which are related to politics as they are the cheapest means for people to acquire awareness on political scenarios and give their votes accordingly during polling.Many cases and situations exist, such as events where changes can be numerically quantified and are much easier to analyze and thus can be used to make inference and prediction. Examples include share prices, stock value and more which throughout the years have seen in the rise to research involving predicting stock market behaviour, But, many events also do exist in which information cannot be quantified. Like preference by an individual for a particular product or a candidate in an election. Typically for such information and events, few methods exist to gauge the 'mood' or preference of an individual. To perceive electoral perception, Opinion polls remain the popular choice for doing it.

The Election Commission of India (ECI) uses Electronic Voting Machines (EVM) to conduct elections. Once an electorate cast their vote on the date of the election. The posted ballots are only accessed on the day of counting the votes under the supervision of the Returning Officer (RO) who is responsible for conducting elections in a constituency, which also includes counting of votes. Before this process takes place there is no quantification to measure the "mood" of the electorate who cast their votes. Public opinion is often estimated using 'Opinions polls' to forecast electoral outcomes. An opinion poll is a survey of public opinion from a particular sample. Opinion polls are usually designed to represent the opinions of a population by asking a series of questions and then extrapolating generalities from responses in ratio or within confidence intervals. However, due to sampling issues, These offline techniques aren't a consistent estimator of the mass sentiment in countries like India with a population of over 1.2. Billion where manually opinion polling is general does not necessarily provide an accurate representation of nationwide sentiment. Media on the other hand shape's political opinion and is a better source for extraction of political sentiment. Thus, it is both interesting and challenging to see what and how the trends in such events change from time to time. Comprehensive investigations can lead to appropriate predictions. Opinion mining from Political, economic and social data, is a new need to make the huge amount of available information in an easily understood form to decision-makers in dedicated centres. This gives an edge to Text Mining and explores the area of Sentiment Analysis (Opinion Mining) which is designated as automatic processing of texts to detect opinions.

Mostly in the stock market when shares fluctuate dramatically, people claim it investor sentiment. In such events, alteration to public opinion and the actual asset impact the share of that company. In fixed sentiment analysis, people look at a newspaper and change in positive or negative affect score on a daily basis.

Typically Natural Language Techniques or commonly referred to as NLP are used to count affect related to a document. The newspapers are expected to reflect the opinion or news media may try to influence people. Analysing sentiment in news media is an indirect way of determining the elector sentiment opinion. The digital presence of news media has increased globally also at the same time, the journal circulation of news articles have decreased in all developed nations. But in developing countries such as India where a huge part of the population is still without an internet connection, the printed newspaper is seen as the primary source of information and therefore have seen a rise year over year. Once the sentiment is extracted, a benchmark is needed to verify that sentiment. The stock market has often been attributed with a fairly reliable proxy for benchmarking sentiment found in newspaper and media television. Various searches in the field of Political Sentiment Analysis have concluded the effect of one on the other with results from Roy[9] and Singh [10] both reaching similar conclusions which were that the stock market doesn't handle uncertainty well. From Roy's research it can be observed that in India after BJP won the election the stock market went down as it was uncertain which party was coming into power where from Singh analysis [10] of 2019 General Election of India BJP again won the election, but this time the market had more sure about this victory and the stock market went up. Although a proven method, reliable data source to gauge individual states final performance was tough to find in the limited scope of this study.

Another method was to use national elections as measure and use the sentiment expressed in local elections to match it with. The circulation figures from the Audit Bureau of Circulation show Hindi language newspapers covering the highest percentage of the market with almost 39% of total newspapers circulated in India. Circulation figures also show English newspapers at 18% at par with other regional languages newspapers and almost 1/3rd of Hindi [11]. However according to Reuters report, Online news generally (56%), and social media specifically (28%), have outpaced print (16%) as the main source of news among respondents under 35, whereas respondents over 35 still mix online and offline media to a greater extent [1]. In emerging Mobile-First markets like India digital lookups of English, newspapers are much higher with more people accessing it digitally through smartphones, laptops and tablets which subsequently leads to the much higher readership for English newspapers and therefore only English newspaper was included for this study.

1.1.3 Overview of Indian Political Landscape

The politics of India works within the framework of the country's constitution. India follows the dual polity system, i.e. a double government (federal in nature) that consists of the central authority at the centre and states at the periphery. The constitution defines the organisational powers and limitations of both central and state governments. Indian remains one of the most diverse countries The two main parties in India are the Bharatiya Janata Party, also known as the BJP, which is the leading right-wing party, and the Indian National Congress, commonly called the INC or Congress, which is the leading centre-left leaning party. These two parties currently dominate national politics, both adhering their policies loosely to their places on the left-right political spectrum. At present, there are seven national parties and fifty-one state parties and many more unrecognised parties. In Indian political eco-space, there are three party alliances regularly aligning on a national level in competing for Government positions. National Democratic Alliance (NDA) - Centre-Right coalition led by Bharatiya Janata Party (BJP) which is headed by current Prime Minister of India, Mr Narander Modi, United Progressive Alliance (UPA) - Centre-Left coalition led by Indian National Congress which is headed by Mr Rahul Gandhi, and Third Front - A coalition of parties which do not belong to any of the above camps. Despite the presence of this "Third front," and other seeming alternatives for those seeking options outside the INC or BJP, Indian politics, by and large, remains a de facto two-party system at the national level.

Major elections in the Republic of India include elections for Members of the Parliament in Lok Sabha also Known as the General Election of India. To elect 545 members of the parliaments or MP's election is divided into seven phases and each state had their allocated number of seats to be contested for Lok Sabha from that particular region, and another one to elect Members of State Vidhan Sabha or Legislative Assembly (MLA) known as State Legislative Assembly Elections held by all 29 states to form a government. Both the election requires any single or coalition to prove a two-third majority to form a government. Elections take place once every 5 years. The most recent 2019 General Election saw BJP-led National Democratic Alliance (NDA) win 353 seats, further increasing its substantial majority over 2014 and form the central government for a second consecutive time.

1.2 Research Objective

This research work specifically studies the news articles generated by Indian print media for months leading up to the State Legislative Assembly Election and 2019 General Election held in three different Indian states namely, Chhattisgarh, Punjab and Karnataka which served as the three case studies to examine the veracity of hypothesis that the results of state election for could be extrapolated to Indian General Election or not. The aim was to stream newspaper articles collected from authentic and authoritative sources created during elections where news flows remain the highest and create a statistical model to examine the electoral mood or opinion. This aims to create framework thought which opinion polling could be achieved using sentiment analysis. The research also further analyzes the negative sentiment expressed towards the political parties and examines whether any bias prevailed in the newspaper articles published during the course of the election.

This was accomplished through the pursuit of four main objectives. The first was to analyse text corpus collected from reputed sources specific to each state and to identify the presence of sentiment and quantify it form a proxy variable time series. The second was to build descriptive statistical analysis correlation analysis models for the political factor sentiment time-series for both the state and the general election for every three case-studies. The third is to apply these models on all case-studies and aid comparison and contract of results in all the case-studies between the state and the general election. The final was to make notable inference form the results found across different states and examine the truthfulness of the hypothesis.

1.2.1 Overview of Case-Studies

To conduct this research revolving around Indian politics, as mentioned in section 1.1.3. India's two biggest political parties BJP and INC currently dominate national politics. They assert regional dominance too and combined they also happen to rule in 19 of 29 states with BJP governing 14 and INC 5 as of March 2020. To successfully compare the result of the state elections in contrast with national election it was essential to select states with BJP and INC as its centre of political war. For the purpose of the research four scenarios were identified which would be included in the scope of this study. The selected states were also required to fill the following criteria:

- The last ruling government for the state was either BJP or INC.
- The most recent state election (before 2019) was won by either BJP or INC
- Majority of Lok Sabha (or Parliamentary) seats in the 2019 General Election was either won by INC or BJP from the chosen state.

From the list of states which fulfilled the criteria, they were divided into these 4 categories which serve as our case studies:

1. Case Study 1: Where INC won the state election but BJP won the Majority of Parliamentary seats in 2019 General Elections form the chosen state i.e. an observed change in public sentiment away from INC. Chhattisgarh was chosen for this case which saw INC winning the majority in 2018 State Assembly Election whereas BJP won 9 out of 11 seats from the state during the 2019 General Election.

2. Case Study 2: where INC won the majority in both the election and the public sentiment remained unchanged. Punjab was chosen for this case where INC won the 2017 State Assembly Elections and also a majority of 8 out of 13 seats during the 2019 General

Elections.

3. Case Study 3: where BJP won a majority in both the election and the public sentiment remained unchanged. Karnataka was chosen for this case as it won the 2018 State Election and also a majority of 25 out 28 seats in 2019 General Election

4. Case Study 4: where BJP won the state election but INC won the Majority of Parliamentary seats in 2019 General Elections form the chosen state i.e. an observed change in public sentiment away from BJP. Interestingly, no state in India observed this phenomenon.

The three selected states also geographically provided coverage of nationwide sentiment and mood with Punjab from the North, Chhattisgarh from Central and Karnataka from the South part of the country. This also allowed this study to encapsulate the behavioural difference associated with different regions as part of this diverse nation.

1.3 Structure of the Thesis

The remainder of the thesis is structured as follows: The following chapter describes the motivation behind this study. It is also concerned with a review of the methods and current literature which is available at this time that have employed by previous studies in the field of sentiment analysis and its application to print media new using statistical modelling in Chapter 2. Next, Chapter 3 which provides a detailed description of the employed mythology and the steps taken by this study throughout. It further outlines the implementation tools and techniques used to conduct this study. The results obtained are then presented for each of the three case studies observed during this project. Next, The results obtained are then presented for each of three case studies observed during the project and discuss key finding from each of them through different analytical models as mentioned in the previous chapter. Chapter 4 also offers an analysis of these results on a case by case basis and inferences deduced for each one. To end, the final chapter provides a summarized conclusion and suggests future work in Chapter 5.

2 Motivation and Literature Review

2.1 Introduction

This chapter first presents a 'Motivation' section which talks about the motivation for carrying this study and reason supporting the necessity of it. Next is the section called 'Literature Review' detailed description of previous works that have been done in this field and the state of the art already available in this field and how this research fits into it.

2.2 Motivation

With the development of the web and its offered services, a huge amount of data is generated [3] and thus additional needs emerge to take benefit from this information thesaurus. Social media platforms, including Facebook, Twitter, and Instagram, have become more ubiquitous. They have had an increasing role in social movements, elections, and everyday life around the world. Social science is well-positioned to explore the power and influence of social media economically, politically, and socially [12]. As mentioned earlier, politics is heavily dependent on electorate perception and opinion. Politics campaigns nowadays are so much more than rallies, flyers and big adverts on the billboards.

With changing time, political parties and leaders have started to embrace the 'Mobile-First' generation. India is emerging as an overwhelmingly mobile-first, and for many mobile-only, media market for internet use broadly, and for online news use specifically. According to a Reuters report 68% of the people identify smartphones as their main device for online news [1]. Political leaders have been quick to realise this. The popular social media platform Twitter has been evaluated as a space for political engagement by both voters and politicians. Political leaders use it to reach out to a much wider audience and express their ideas and shape public opinion. Narendra Modi during 2014 as well as 2019 Indian General Elections employed a huge team of social network specialists. In the western countries, Donald Trump embraced it too during his 2016 United State Presidential Campaign. In recent years a large number of research has been conducted to study how political parties leverage Twitter during an election to campaign [13] [12] [14] [15] [16]. The affect of social media on the outcome of the Indian

General Election of 2014 and 2019 has also been a special focus of many studies [17] [18] [19] [20] [21] [22]. Several reports have cemented Indian position as one of emerging mobile-first markets [1]. With one of the cheapest 4G tariffs in the world [23] and aggressive mobile pricing over 550 million Indian have access to internet [24]. But, the print and television news industry in India has continued to grow – though at a decreasing rate.

Figure 2.1 shows that the circulation of newspapers in India between 2015-16 and 2018-19 compared with other developed nation

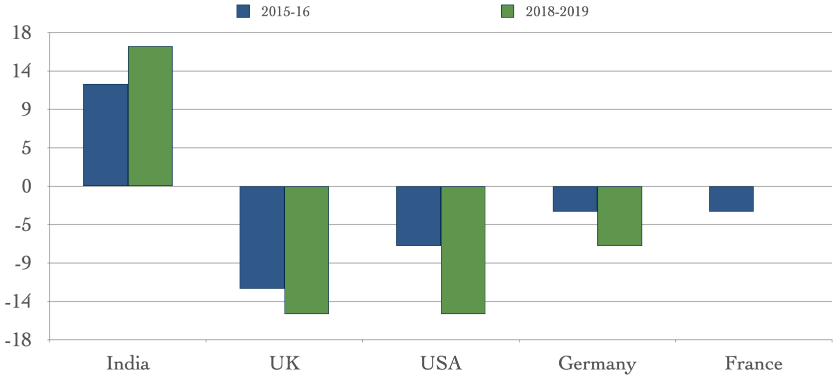


Figure 2.1: Newspaper Circulation Growth Between 2015-16 & 2018-19

While around the world or at least in developed nations of the west, the newspaper industry is facing imminent extinction as circulations are slumping and advertising revenue collapsing. But not in India. Newspaper circulation has grown significantly in India, from 39.1 million copies in 2006 to 62.8 million in 2016 – a 60% increase according to ABC [25]. With much of the population still offline, hundreds of millions of Indians still turn to newspapers television and radio as their main sources of news, with newspapers being the cheapest of the bunch. Disinformation and fake news have emerged as a pressing issue in India. Rueter reports over 57% of online users are concerned about what is fake on the internet [1]. That said, In countries such as India where newspapers remain the primary source of reliable information for millions which encapsulates the sentiment of the entire country, the lack of focus on this source of information remains the primary motive to choose print media articles for this study. The most widely used online news sources (beyond platforms) are generally the websites of leading legacy media including broadcasters and newspapers [1]. Therefore, for this study, the legacy media publications selected as sources for news articles was shortlisted under the criteria of publications with the highest circulation [11] and readership [2].

Although a huge number of research exists on the election and its effect on stock markets, oil prices and many other sectors, None exist. The lack of research which has tried to analyze the impact of the state election in India and its effect on the 2019 General Election using the sentiment from print media has been a major motivation behind this study. Hence, this study to analyze the sentiment captures during the state election gathered from authoritative sources and study the shift in behaviour, if one occurs in the sentiment expressed by the dallied as we approach the GE 2019 and compare the two through the use of multiple case-studies in multiple regions of the country to create a framework to digitally calculate the opinion of people as expressed and through statistical modelling and correlation analysis of the uni-variate sentiment proxy time series obtained from the sentiment analysis.

2.3 Literature Review

2.3.1 Data Retrieval

For the retrieval of textual print media data various data retrieval techniques have been employed in the past by the researchers. These techniques can be broken down into three broad categories. Web-scraping, application programming interface (API) offered data sources and third-party services and database systems.

Kelly used the first approach for this study by using popular python frameworks "Scrapy" and "Beautiful Soup" to gather media content through web scraping [26]. The data gathered is often required to be preprocessed using suitable techniques to make it suitable for analysis. Web-scraping is generally not the first choice due to reliability issues and also concerns over legality and ethicality of this approach [27] and therefore is often only employed in cases where data cannot be accessed using first-party API or third-party services.

A more suitable approach would be gathering data by using the APIs directly made available to the public by the publications or information providers. Data retrieved using this approach usually comes with a formatted coherent structure making it easier to work with. The problem with this approach is when the data is required to be collected from multiple sources, contributing to added work. Also, several legacy media publications don't provide API access or have no API infrastructure, promoting the use of web scraping with added work.

The third and final approach suggests building a data corpus using third-party aggregator services such as LexisNexis, Factiva, ProQuest, Acxiom and Westlaw etc. Each of these services produces large amounts of data expanding across multiple sources and serve different domains. Kelly, Zhao and Ahmad, Roy and Singh, all have conducted studies in the past requiring data from multiple sources to perform sentiment analysis [28] [29] [9] [10]. All the aforementioned studies have employed the services of LexisNexis which circumvents the problems associated with other techniques when dealing with multiple sources. It also ensures the authenticity of the data collected. These results encourage me to use LexisNexis to retrieve data required for this study.

2.3.2 Political Sentiment and Print Media

Political preferences of an individual are entirely based on one's opinion. This bias can change the way that person perceives news information daily. However, this preference can be heavily influenced by news media and their coverage [30]. Harvard's Lerner in a series of studies explains how emotions can influence decision-making their choices in her research [31]. Many psychologists have studied the relationship between various forms of emotional communication and its effect on an individual's judgement [32]. Iyengar in her book pointed to the impact of news in framing the public's preferential accountability for social, political and economic issues [33].

As mentioned before, politics is a practice of influencing people at an individual level and in this, all forms of media play a very important role as they often form opinions. Traditional media sources such as radio, television and newspaper as well as various social media platforms cover all kinds of political events. While social media platforms have gained significant traction, newspapers accessed via physical copies and digital lookups through smartphones remains a key figure in the spread of political news [1]

Newspapers usually embed writers' bias and perspective in their articles but are largely controlled by the organizations they work for. Although journalists are refrained from using clearly positive or negative vocabulary. Studies observed that they resort to other means to express their opinion by placing the statement in more complex discourse or omitting facts [34]
citebalahur2013sentiment. Reports in the UK suggest the "unashamedly partisan" nature of multiple news publications [35]. Rueter reports of 2018 shows 51% of their respondents express concern over hyper-partisan content by Indian news publication [1]. That said, print media plays a significant role in moulding the political sentiment of their readers.

In a highly politicised context, a country with a history of communal violence and characterised by explosively growing access to news media and low trust in news, disinformation has emerged as a pressing issue in India. Reuters reported that 57% of their respondents said they are concerned about what is real and what is fake on the internet, a number comparable to levels of concern in Turkey and the United States [1]. Growing numbers of cases have been sighted where new contents internationally feed to their belief of a particular sector and lobbyist to maximize financial gains from the commercialization of a viral news trend or through advertisements [36].

One way of measuring the effect of newspaper coverage is to look at the voting intentions of readers just before elections [35]. In most cases, it isn't too much of a surprise. According to an Ipsos/Mori poll in the UK, 65% of Telegraph readers said they would vote Tory in 2001 and 2005, while 67% of Mirror readers pledged to vote Labour in 2005. In the same year, 57% of Daily Mail readers said they intended to vote Tory while 22% promised to vote Labour and 14% Lib Dem [37]. A similar study conducted by Brandenburg also found media bias [38] in the UK. He also stated Irish election of 2002 also witnessed negative sentiment projected towards political figures by certain news publications [39]. Ahmad et al. also showed media bias as well as gender bias in the Irish election of 2011 [40].

Indian General Election of 2014 and 2019 have also been of particular interest among the scholars and private agencies alike. Many studies have examined the political orientation associated with English newspapers leading up to the election. Padmaja et al., Barclay et al and Roy [41] [42] [9] have all analysed the election using a print media and have been able to predict the outcome of elections with a high level of accuracy. Singh also followed a similar methodology to predict the Indian General Election of 2019 with success [10].

Although a vast amount of research has been conducted at national level politics in India covering national sentiment, the effect on the financial market and more [29] [26] [10] [9]. None have given attention to the more granular level and the very important State Assembly Election in the 29 States of India and question, could the results of one influence the results of another? Are they correlated? Is the conclusion cause? The lack of research on the impact of various state elections on the results of Indian General Election motivated me to study the political orientation in newspapers from different regions of the country leading up to both their respective state and national elections.

2.4 Sentiment Analysis

Sentiment Analysis is a process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral as defined by the New Oxford American Dictionary [43]. Also known as opinion mining or emotion artificial intelligence, sentiment analysis refers to the use of techniques such as natural language processing (NLP), text analysis and computational linguistics to systematically identify and categorize the sentiment expressed by the author and study affective states and analyze the subjective information better. It is often used to capture opinion and sentiment of the public and media which plays a very important role and can decisively shape bias or prejudice [5].

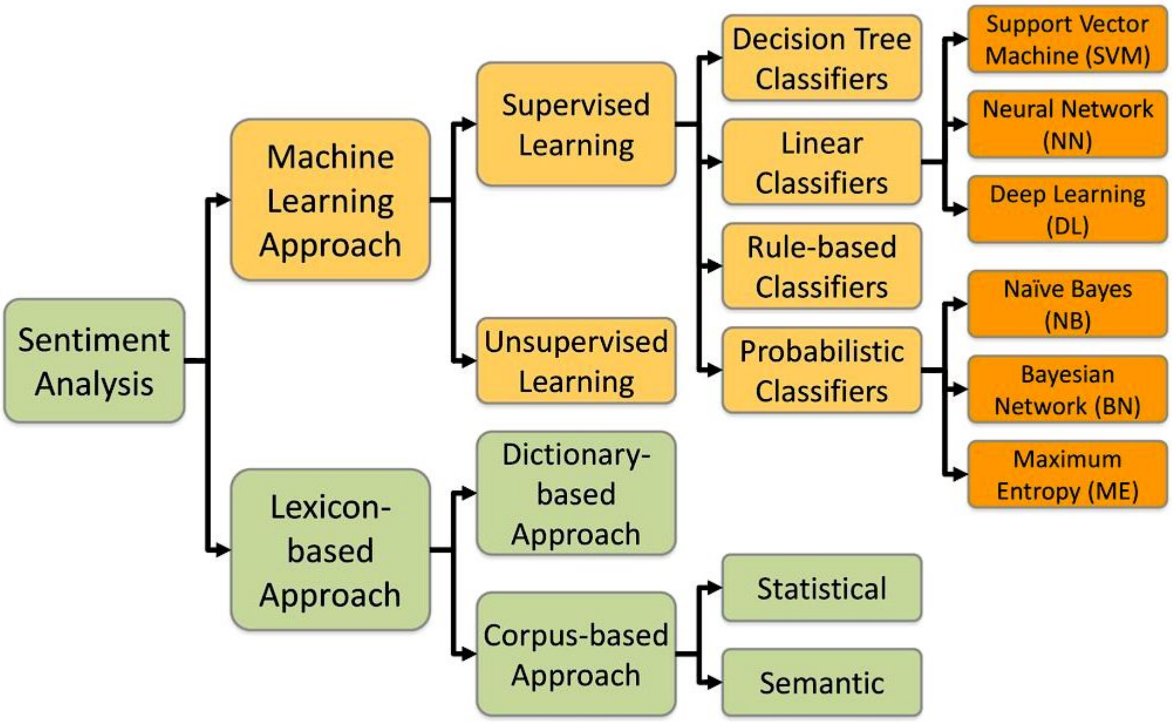


Figure 2.2: Sentiment Analysis Classification Techniques

For the purpose of this study, sentiment analysis refers to the examination of print media, intending to extract, identify and quantify the subjective content expressed by the media. A vast number of sentiment classification techniques exist but all of them can be broadly classified into two main categories: machine learning approach or lexicon-based approach. Figure provides a hierarchical overview of the several techniques which fall under these two approaches.

2.4.1 Machine Learning Approach

Machine Learning approach typically employs algorithms that often rely on supervised classification approaches, requiring labelled data corpus to train classifiers algorithms which can learn the affect of a specific term or text. The supervised techniques are applied mostly for recognizing the sentiment of complete documents. Mullen and Collier [44], Finn and Kushmerick[45], Aua and Gamon [46], Pang and Lee [47] developed their classifier using a support vector machine (SVM). Pang and Lee further extend their study and use a step-wise approach for the classification of texts by using Naive Bayes classifier which was also employed by Tan et al. [48] and Lewis [49]. Other researchers have chosen maximum entropy-based classifier [50]. Ribeiro et al. take a more holistic approach by presenting a benchmark comparison of twenty-four popular sentiment analysis methods [51]. Although Machine learning approach produces good results, it heavily relies on the availability of pre-labelled data to train the model and large volume of data to feed into the system for validation and testing which is difficult to get and therefore it is a major drawback for this approach.

2.4.2 Lexicon-Based Approach

The lexicon-based approach semantic orientation (SO) of the document is calculated by summing the semantic orientation of the words and phrases in the document[52]. The approach is further broken down into two types: a corpus-based approach and a dictionary-based approach. The Corpus-based approach helps to solve the problem of finding opinion words with context-specific orientations. Its methods depend on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus [53]. It begins with a "seed list" of opinion words/phrases and calculated sentiment by searching or other similar contextual words and phrases [54]. There are further two methods in the corpus-based approach: Statistical Approach and Semantic approach. Using part-of-speech (POS) patterns, statistical approach classifies the text by extracting the bigrams. PMI is then calculated by using the polarity score for each bigram. Seed opinion words can be found using statistical techniques. Semantic approach assigns similar sentiment values to semantically close words. These Semantically close words can be obtained by getting the list of sentiment words and iteratively expanding the initial set with synonyms and antonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word [55]. Ortiz and Cruz used this approach to expand a system initially designed for general language text into domain-specific sentiment analysis system [56]. The study although showed promising results felt short as the seed list remained static, but the specialization level of the corpus increased.

The dictionary-based approach involves using predefined dictionary words (including synonyms and antonyms of a word) where each word is associated with a specific sentiment polarity strength. The approach employs a 'bag-of-words' (BOG) model where the text is broken down into single-words token or uni-grams maintaining a frequency count of each while removing context [57]. The entries in the dictionaries are matched with the words from the text to compute the sentiment. A General Enquiry Dictionary which is a combination of Harvard IV-4 and Laswell dictionaries is generally used across the community to create a comprehensive general English glossary to be used as a base dictionary for sentiment analysis [58]. SentiNet-World [59] based on WordNet [60] is another popular choice developed by Esuli and Sebastian [59].

Several studies create a custom domain-specific dictionaries in addition to a base directory to take into account the degree of specialization involved in their work. This is to avoid the misclassification of terms by using a generic dictionary. Ahmad at al. and Kelly employed domain-specific dictionaries along with base general dictionaries in their studies in the domain of finance [61] [26] [28] [29]. Roy investigated the Indian election of 2014 and its relation with the stock market by creating a domain-specific directory in addition to the general English language dictionary [9].

2.4.3 Existing Sentiment Analysis Systems

Sentiment analysis has become an extremely popular tool, applied in several analytical domains, especially on the Web and social media. Searches on Google for the query 'Sentiment Analysis' has also seen steady growth over the last decade. In the last few years alone, more than 7000 scientific papers have investigated sentiment analysis, several startups that measure opinions on real data have emerged and a number of innovative products related to this theme have been developed. Due to the enormous interest and applicability, there has been a corresponding increase in the number of proposed sentiment analysis systems in the last few years by both independent researchers and in commercial capacity alike. In 2014 Bloomberg added social sentiment analytics to its trading terminal[26]. Infotrie "FinSentS" can scan millions of sources: websites, blogs, and business news publications in real-time and analyse more than 50,000 stocks, topics, commodities people and other assets through [27]. Lexalytics Semantria [28] is a paid tool that employs multilevel analysis of sentences and is used by Thomson Reuters News Analytics which then have been used by a plethora of studies[29] [30]

From a research perspective, several sentiment analysis systems also have been developed for academic purposes. SentiWordNet [59] [62] is a tool used for construction of a lexical resource for Opinion Mining based on WordNet [60] developed by Esuli and Sebastiani based on a machine learning approach for classification. The tool groups adjectives, nouns, etc and associates three polarity scores (positive, negative and neutral) for each one in synonym sets (synsets). Linguistic Inquiry and Word Count (LIWC)is a popular text analysis tool to evaluate emotions, cognitive and structural components of a given text [63]. It utilizes a glossary with words classified into categories such as health, education, leisure, anxiety etc. LIWC updated its tool in 2015 with better performance. It is a paid tool with limited access to students available for academic purposes.

The Sentiment Orientation Calculator (SO-CAL) [64] developed by Taboada et al. creates a new lexicon with unigrams (adjectives, adverbs, noun, verbs) and multi-gram (intensifier and phrasal verbs) hand ranked with a scale from positive five (strongly positive) to negative five (strongly negative) [51]. Dictionaries can be created manually or automatically using seed words to expand the list of words in the lexicon. Similar to SO-CAL, Valence Aware Dictionary and Sentiment Reasoner (VADAR)[65] is a sentence-level sentiment analysis method. It combines a lexicon and the processing of the sentence characteristics to determine a sentence polarity. This rule-based approach makes use of a series of intensifiers, punctuation transformation, emoticons, and many other heuristics [51]. VADER was created from a generalizable, valence-based, human-curated gold standard sentiment lexicon [51].VADAR is becoming widely used, being even implemented as part of the well-known NLTK python library [66].Sentiment140 [67], SASA [68], PANAS-t [69] and AFINN - a new ANEW [70], are few other sentiment analysis method developed specifically for Twitter and opinion mining in the platform.

Rocksteady is an affect analysis system developed at Trinity College Dublin by Ahmad and Zemánková that also employs a dictionary-based approach combined with a 'bag-of-words' model for sentiment analysis [71]. The system allows the user to select any number of base and domain-specific specialists dictionaries for the analysis process. It offers a library of general language dictionaries as well as a few domains specific glossaries. Various studies have been conducted in the past such by Ahmad et al., Zhao at al. and Kelly [26] [28] [29] making use of rocksteady affect analysis capabilities. In the past, Ahmad et al. [], Gupta [], Roy [9] and Singh [10] all have employed Rocksteady to study the impact of sentiment expressed in newspapers on the financial market promoting the need for new specialised dictionaries along with base English glossary. The results have been promising and are remarkably closer to real observations [40].

3 Method & Implementation

3.1 Introduction

This chapter gives a detailed discussion of the research carried out and the methods, tools and technique used throughout the project. It aims to provide an overview of the steps taken from the start of this study, until its completion. As outlined in Chapter 1, the scope of research done is divided into three phases: the acquisition and curation of print media pieces (mainly newspapers), extracting sentiment from the acquired data through Sentiment Analysis using RockSteady affect analysis system and finally performing descriptive statistical analysis. Thus, the first three sections of the chapter provide a comprehensive overview of the methodology employed by each phase. The last two sections discuss the techniques used to visualize the data and finally the implementation for this study.

During initial phases of study, the methodology also consisted to Figure 3.1. provides a high-level flow chart presenting an overview of the developed system architecture of this framework, delineating the methodology for this research.

3.2 Data Acquisition

To extract real-world sentiments expressed in the form of emotion and opinion, it was required to create a data corpus of newspaper articles which were relevant to our areas of research. Data corpus is a well-organised collection of linguistic data (text) which is representative with regards to the research hypothesis [72]. The focus of this study was to look for physical print media content published by leading and formal media outlets. An excess of news sources exists with over 100,000 newspaper publication registered with Registrar of Newspapers for India as of 2018 [73], and hence appropriate attention was given to ensure that the collected sample would serve as an accurate representation of people sentiment available as expressed in the news articles and widely available to the public and market.

The data were collected from a few months before and after a State Assembly Elections and Indian General election 2019 from the selected state/territory according to different case studies.

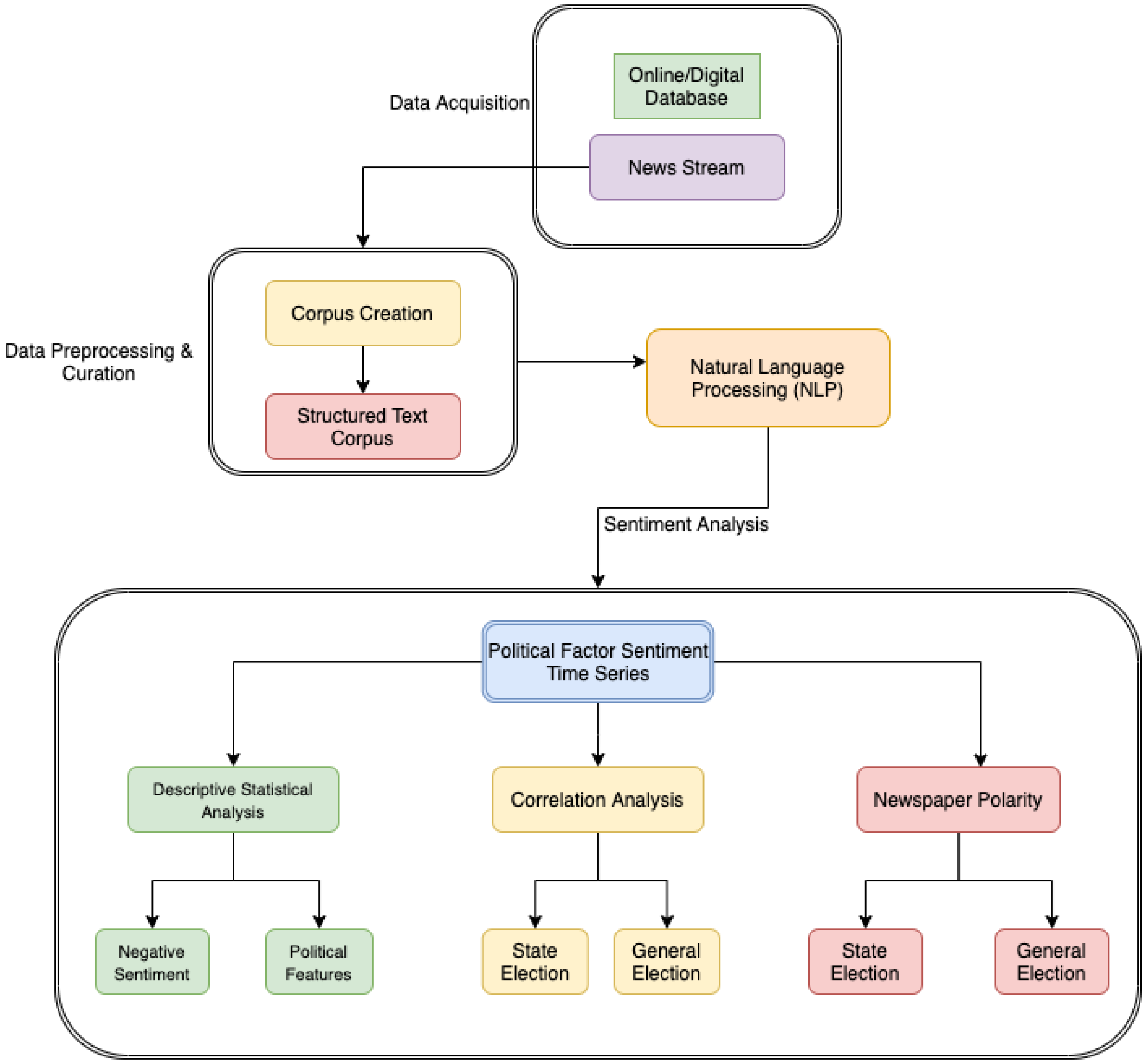


Figure 3.1: A high-level flow chart of the methodology used in the study

In India, print media publishes content in over 50+ language with Hindi taking the largest share in the market with 39% followed by English at 18% and so on [11]. Even with Hindi being more popular out of two, the study focuses only on English newspapers as I had access to no credible and authentic source which would have allowed me to acquire Hindi news articles. Therefore, the top selling and highly circulated English Indian newspaper publication were identified as reported by the Audit Bureau of Circulation (ABC) [11]. Table 3.1 shows the names of the top English newspaper in India by circulation.

Table 3.1: Daily circulation figures of top English newspaper publication in India

Sr. No	Publication	Daily Figures
1	The Times of India	2,640,770
2	The Hindu	1,404,901
3	Hindustan Times	945,221
4	The Indian express	495,618
5	The Telegraph	401,083

The study decided to use the Lexis Nexis News & Business online service to gather news articles as done similarly in the past by Kelly, Zhao, Ahmad, Roy and Singh [26][28][9][10]. The research was only concerned with articles which had information related to elections in different states as and when they took place and therefore the downloaded newspaper articles were refined by searching for query term(s) which would appear only in a particular context. Lexis Nexis also provides narrowing the search by providing a number of available criteria: such as anywhere in the text body, in the headline, in the name of the company, at the start, a high similarity with three or more mentions, low similarity or none.

Table 3.2 presents the query terms and search criteria which were used by different case studies for collecting the data using Lexis Nexis News & Business.

Table 3.2: Lexis Nexis Query terms and Search Criteria

Case Study	Query Term	Search Criteria
1	Election AND Chhattisgarh	High Similarity,
2	Election AND Punjab	i.e. 3 or more mentions in
3	Election AND Karnataka	the text body or in the headline

3.3 Data Preprocessing & Curation

Once the data was acquired, it was pre-processed to convert from downloaded Rich text format (.rtf) to format which was compatible with Sentiment Analysis system, Rocksteady used by this study. It was done using Python script and standard text file template (available at appendix) (.txt) which would parse all the collection of document present in a directory and process it into a single text file in a compatible format.

Data curation is the active and ongoing management of data through its life cycle of interest and usefulness [74]. It is processes of collecting data from diverse sources and integrating it into repositories that are many more times more valuable than the independent parts. Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle. Data curation not only means creation, management, and maintaining of data, but may also be involved in determining best practices for working with that data. Data curation often involved presenting data in a visual format such as a chart, dashboard or report. The principal objective of data curation is three folds: Preserving: Collecting and taking care of research data, Sharing: Revealing data's potential across domains, Discovering: Promoting the re-use and new combinations of data. It is to facilitate future research which could retrieve and reuse the existing data from the present repositories.

3.4 Text Analysis

After data collection and preprocessing, the next phase was to recognise, quantify, and then begin extracting sentiment from the news articles for each of the three case studies. Previous work in this field, as mentioned in the literature review, Ahmad et al. [61], Kelly [26], Roy [9]), Singh [10] and Tetlock [75] are few of the names who focused largely on the negative sentiment attached to the news articles.

As discussed earlier in the literature review (Chapter 2), there are multiple approaches in which one can lean into, to perform sentiment analysis on a given data corpus. For this study, a lexicon-based combined with bag-of-word (BOW) text categorization method used. Lexicon based approach has the advantage of avoiding the hard-working step of labelling training data. Moreover, several standard lexicons have been developed by the influential research in the past which have also been made publicly available for use covering a wide range of sentiments attached to the varying context. Figure **??** is a flow diagram which given an overview of the Natural Language processing techniques employed by the Rocksteady system which are explained further in subsection 3.4.2

3.4.1 Sentiment Analysis using Rocksteady

Rocksteady affect analysis system, developed at Trinity College Dublin by Ahmad and Zemánková [71] was used for performing sentiment analysis on the collected data corpus. The system calculates an affect score for a text corpus, to facilitate analysis of the sentiment inherent to that text corpus using a bag-of-words (BOW) model with combination with vector space model [76]. The program analyzes a user-provided corpus of texts and quantifies the level of sentiment present across it in time series manner as used by Ahmad et al., Zhao et al, Kelly, Das and Singh [61] [26] [9]) [10] [75]. The program provides flexibility to the user by providing a plethora of functionality and options such as the ability to group articles in several ways (e.g. by source, by year, by month, by week, by day, etc.), date range filters, the option to exclude duplicated news articles using string comparison techniques.

Data was aggregated using multiple grouping options such as by day, by month and by

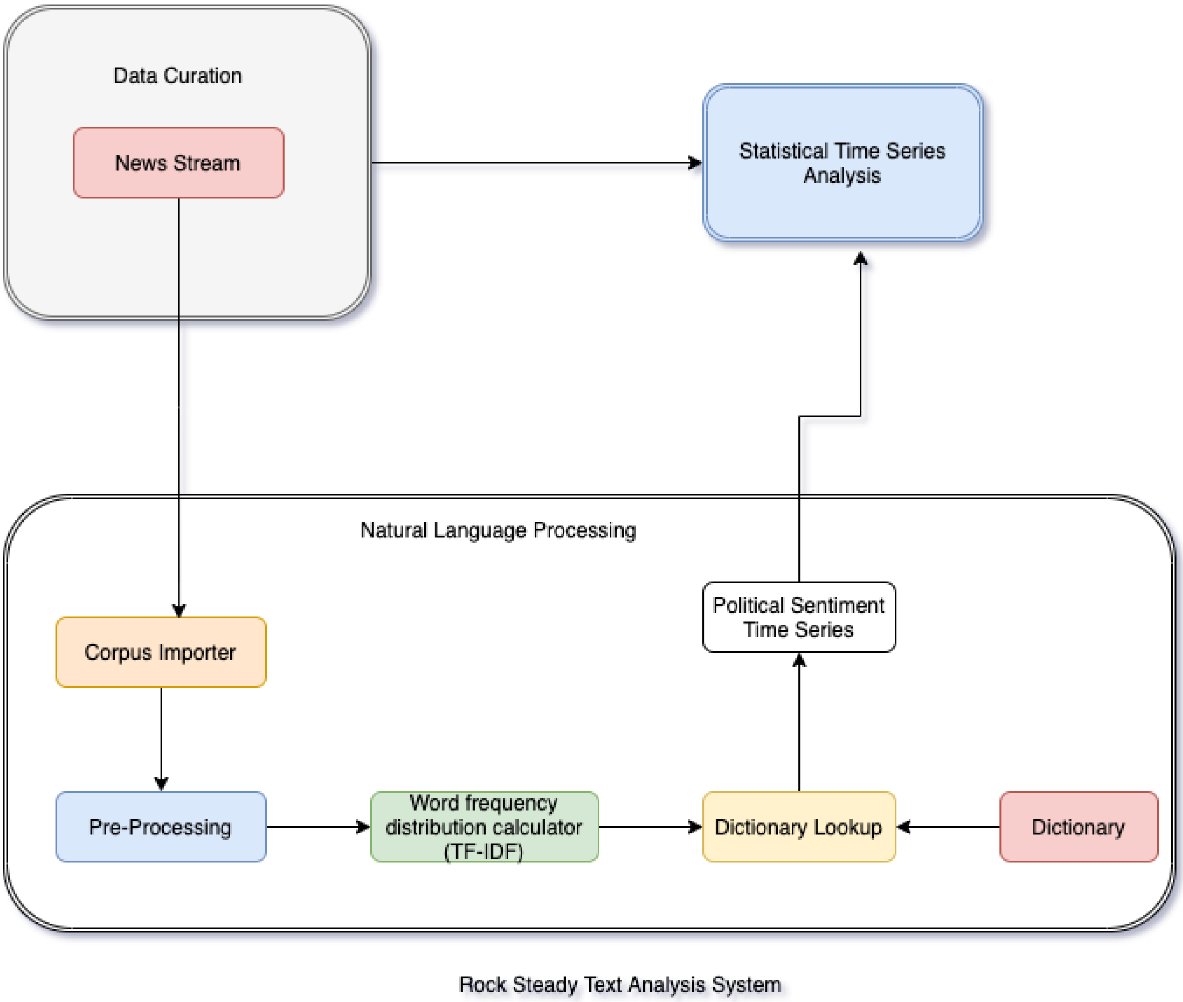

Data Curation
News Stream
Statistical Time Series Analysis
Natural Language Processing
Corpus Importer
Political Sentiment Time Series
Pre-Processing
Word frequency distribution calculator (TF-IDF)
Dictionary Lookup
Dictionary
Rock Steady Text Analysis System

Figure 3.2: A high-level flow chart NLP techniques used by the Rocksteady

the source to compute three different kinds of statistical time-series analysis on each of them.

Rocksteady uses a BOW model where a document gets a representation as a vector or "bag" of words in Euclidean space where each word is independent of others. This bag of individual words is commonly called a collection of unigrams or a token. The independence of unigrams means that the appearance of one unigram in the text will not influence the appearance of any other unigram [76]. Rocksteady discards the context, the definite ordering of terms while it maintains a word frequencies count.

Sentiment in any given text is calculated by matching the unigrams or token against a pre-determined dictionary of affect-laden words. A term frequency-inverse document frequency (tf-idf or TF-IDF) score, which is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus is used to maintain a final affect score of the document.

Table 3.3 and 3.4 shows an example of how the system uses this aforementioned technique to calculate the level of negative sentiment present in a sample text. A frequency count is maintained for the broken unigrams or word token (Table 3.3). The work tokens are then matched against a glossary of terms carrying negative implication and then a simple calculation tells us the percentage of total terms in the text which is considered to exhibit a negative sentiment which effectively is the negative sentiment % score for that text (Table 3.4).

Table 3.3: Bag-of-word Representation

Terms	Frequency
Sensex	1
dips	1
sharply	1
down	1
91	1
points	1
Total	6

Sample Text: "Sensex dips sharply, down 91 points."

Table 3.4: Calculation of Negative Sentiment Score

Glossary hits for Negative Sentiment	2
Total Number of Terms	6
Negative Sentiment % Score	**33.33%**

Note: sample text is an excerpt from a June 3rd 2009 article in the Hindustan Time

3.4.2 Natural Language Processing

Tokenization or Bag-of-Words Creation

The incoming string is broken into tokens or a bag-of-words known as tokenization: which are un-ordered collections of words and other elements in that document. It is called a "bag" of words because any information about the order or structure of words in the document is discarded and contains no context [77]. The model is only concerned with whether known words occur in the document, not wherein the document and it is maintained by keeping a frequency count of each term. As the vocabulary size increases, so does the vector representation of documents. For a very large corpus, such as thousands of books, that the length of the vector might be thousands or millions of positions. Further, each document may contain very few of the known words in the vocabulary.

This results in a vector with lots of zero scores called a sparse vector or sparse representation. Sparse vectors require more memory and computational resources when modelling and the vast number of positions or dimensions can make the modelling process very challenging. Also, all words present are not significant and therefore require some form for pre-processing to reduce the noise and before sentiment can be extracted.To decrease the size of the vocabulary when using a bag-of-words model. There are simple text cleaning techniques that can be used as a first step:

- **Lowering the Case**: All the token are converted to lower-case
- **Stemming & Lemmatization**:
  For grammatical reasons, documents are going to use different forms of a word, such as organize, organizes, and organizing. Additionally, there are families of derivationally related words with similar meanings, such as democracy, democratic, and democratization. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set.

  The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance:
  am, are, is –> be
  car, cars, car's, cars'–> car

  The result of this mapping of text will be something like:
  the boy's cars are different colors –>
  the boy car be differ color
  However, the two words differ in their flavor

  Stemming: Stemming refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes [78]. Rocksteady uses Porter Stemmer to stem all words.

  Lemmatization: Lemmatization refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma [78].

Once these pro-processing steps are finished, the corpus is passed on to the vector space mode.

Vector Space Model (VSM) The Vector Space Model (VSM) is based on the notion of *similarity*. The model assumes that the relevance of a document to query is roughly equal to the document-query similarity. Both the documents and queries are represented using the bag-of-words model. For a document collection, we first determine a set of terms (i.e., vocabulary) and order the terms. Next, documents are represented as n-dimensional vectors, where each dimension corresponds to a term. Terms are weighted using tf-idf or BM25 method [79].

Rocksteady uses term frequency-inverse document frequency (tf-idf or TF-IDF) as a scoring mechanism.

Calculating TF-IDF

A problem with scoring word frequency is that highly frequent words start to dominate in the document (e.g. larger score), but may not contain as much "informational content" to the model as rarer but perhaps domain-specific words.
One approach is to rescale the frequency of words by how often they appear in all documents so that the scores for frequent words like "the" that are also frequent across all documents are penalized. This approach to scoring is called Term Frequency – Inverse Document Frequency, or TF-IDF for short, where:

- TF, Term Frequency: is a scoring of the frequency of the word in the current document. It measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more often in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization [80]:

  $TF_{(t)}$ = (Number of times term t appears in a document) / (Total number of terms in the document).

- IDF, Inverse Document Frequency: is a scoring of how rare the word is across documents. It measures how important a term is. While computing TF, all terms are considered equally important. However, it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scaling up the rare ones, by computing the following [80]:

  $IDF_{(t)}$ = log_e(Total number of documents / Number of documents with term t in it).

Combining the definitions of term frequency and inverse document frequency, to produce a composite weight for each term in each document. The tf-idf weighting scheme assigns to term t a weight in document d given by:
Equation,
tf-idf weight of a term is obtained by a product of tf and idf as shown in the equation above. The higher the TF*IDF score (weight), the rarer the term and vice versa. It is lowest when the term occurs in virtually all documents.

Rocksteady facilitates the use of single or multiple dictionaries at the same time. They are divided into two categories: base dictionaries and specialist (domain) dictionaries. Base dictionaries consist of general terms which can be used in multiple contexts while the specialist

(domain) dictionary contains domain-specific terms which are relevant to that particular study. If there is conflict and a term exists in both the glossaries, the terms take on the attributes defined in the specialist dictionary.
A General English language dictionary was selected as a base dictionary. For the purpose of this study, an Indian General Election 2019 domain-specific dictionary was created and coupled with the base one.

3.4.3 Domain-Specific Dictionary Creation

A comprehensive list of terms associated with the expression of sentiment was already included in the general English language base dictionary used for this study. The principal objective of creating a new domain-specific dictionary was to ensure that terms and certain phrases which are normally affiliated with one kind of sentiment in general language, would take on an entirely different connotation in the area of this study and would prompt the misclassification of terms in the sentiment analysis process. For instance, the phrase "the restaurant is will close down next month" is often linked with a negative sentiment, with the word "close" expressing the negative sentiment in this sentence, where in the world of wall street and finance, "close price of stock" it simply refers to the price of the stock at the closing bell in a given day. To avoid such misclassification, and in the reach of this study, a new domain-specific glossary was constructed covering all political features specific to the Indian General Election of 2019. Political features such as names of political parties(national and regional), political leader, slogans were incorporated in the dictionary.

Assembling a list of political parties and leaders was relatively easier and was collected by querying from DBpedia [81] which allows users to semantically query relationships and properties of Wikipedia resources and extract its structured information repositories. While the campaign slogan was not, As it was not available in any structured form and therefore was done by scraping through numerous website and digital media publications and manually collecting and storing them in our CSV files format.

Table 3.5 shows the structure of the specialist dictionary created for this study. Every political term was mapped to their relevant associated feature, forming a i x j matrix. For example, "Narender Modi" who is the current prime minister and party leader of BJP will be mapped to features such as BJP which is a feature in this dictionary and "PolitLead" which identifies that he is a political leader.

Table 3.5: Mapping structure for Domain specific glossary

Term Entry	Feature 1	Feature 2	Feature 3	...	Feature n
Term 1	Feature 1	Feature 2		...	
Term 2	...	...	Feature 3	...	
...	...		...		
Term i		...	Feature 3	...	Feature j

The numbers against each of the features indicated the frequency of the appearance of words mapped to that feature. Using these numbers, the tf-idf scores are calculated for all data corpus corresponding to three chosen states and exported as a time-series data-set that included the affect scores.

3.5 Statistical Analysis

Time-series is collection of observations x_t, each one being recorded at time t. Time series analysis is a statistical technique that deals with time-series data, or trend analysis. Time series data means that data is in a series of particular time period or intervals. Once we have obtained the required time-series from the Rocksteady affect analysis system, time-series can now be analysed using typical statistical techniques as described in this section. To have a better insight and understanding of the features of the dataset and summarize it, descriptive statistics were calculated for different time-series.

To check for any possible connection between the variables and the strength of those relationships, Correlation analysis was performed. These are further described in the following subsection below.

3.5.1 Descriptive Statistics

Descriptive statistics involves summarizing and organizing the data so they can be easily understood. Descriptive statistics, unlike inferential statistics, seeks to describe the data, but does not attempt to make inferences from the sample to the whole population. Here, we typically describe the data in a sample. This generally means that descriptive statistics, unlike inferential statistics, is not developed on the basis of probability theory [82].

Descriptive statistics involves summarizing and organizing the data so they can be easily understood. Descriptive statistics, unlike inferential statistics, seeks to describe the data but does not attempt to make inferences from the sample to the whole population. Here, we typically describe the data in a sample. This generally means that descriptive statistics, unlike inferential statistics, is not developed on the basis of probability theory. Descriptive statistics are broken down into two categories. Measures of central tendency and measures of variability (spread).

Measure of Central Tendency Central tendency refers to the idea that there is one number that best summarizes the entire set of measurements, a number that is in some way "central" to the set. These include mean/average, median and mode [82].

Measure of Spread / Dispersion Measure of Spread refers to the idea of variability within your data. It includes the standard deviation, variance, range, percentile, kurtosis, skewness and more [82].

Outlined below are few of the more sophisticated statistical measures and techniques which were used in the study. Some more basic concepts have been listed in Appendix A1.1 for reference.

Skewness Skewness, in statistics, is the degree of distortion or asymmetry from the symmetrical bell curve in a probability distribution..Distributions can exhibit right (positive) skewness or left (negative) skewness to varying degrees. Besides positive and negative skew, distributions can also be said to have zero or undefined skew. In a perfect normal distribution, the tails on either side of the curve are exact mirror images of each other. Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution.

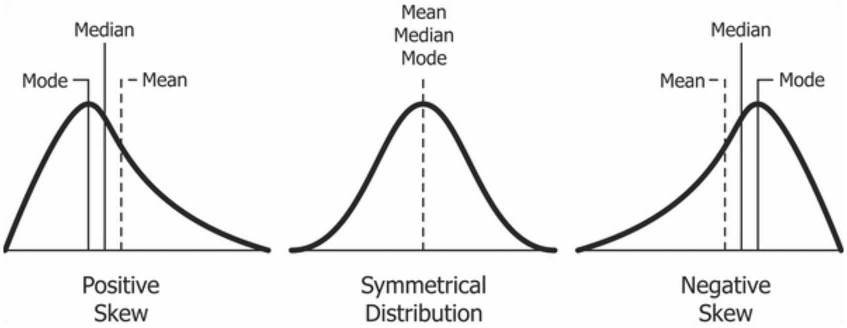


Figure 3.3: Skewness

When a distribution is skewed to the left, the tail on the curve's left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode. This situation is also called negative skewness.

When a distribution is skewed to the right, the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode. This situation is also called positive skewness.

Skewness is considered a significant metric as it considers the extremes of the data set rather than focusing solely on the average

Kurtosis kurtosis is a statistical measure that is used to describe the distribution. It's about the existence of outliers. kurtosis measures extreme values in either tail. Distributions with large kurtosis exhibit heavy-tailed (profusion of outliers) data exceeding the tails of the normal distribution. Distributions with low kurtosis exhibit light-tailed (lack of outliers) data that are generally less extreme than the tails of the normal distribution. Kurtosis is sometimes confused with a measure of the peakedness of a distribution. However, kurtosis is a measure that describes the shape of a distribution's tails in relation to its overall shape. A distribution can be infinitely peaked with low kurtosis, and a distribution can be perfectly flat-topped with infinite kurtosis. Thus, kurtosis measures "tailedness," not "peakedness."

The main difference between skewness and kurtosis is that the skewness refers to the degree of symmetry, whereas the kurtosis refers to the degree of presence of outliers in the distribution.

3.5.2 Correlation Analysis

Correlation is a statistical technique that measures the degree to which two variables move in relation to each other. It can show whether and how strongly pairs of variables are related. Correlation measures association, but does not tell you if x causes y or vice versa, or if the association is caused by some third (perhaps unseen) factor. It is computed as the correlation coefficient (r), which has a value that must fall between -1.0 and +1.0. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movement of the two variables. Values of r between 0 and 1 reflect a partial correlation, which can be significant or not [83]. A positive correlation is a relationship between two variables in which both variables move in tandem—that is, in the same direction [84]. A negative correlation is a relationship between two variables in which one variable increases as the other decreases, and vice versa [85].

Correlation analysis is a statistical method used to evaluate the strength of the relationship between two quantitative variables. A high correlation means that two or more variables have a strong relationship with each other [86]. For the purpose of this study, all the correlation coefficients are multiplied by 100 and presented as a percentage for better visualization and understanding. Therefore, the (r) lies between -100% to 100%.

However, "Correlation does not imply Causation". It statistics it refers to the inability to legitimately deduce a cause-and-effect relationship between two variables solely on the basis of an observed association or correlation between them [87]. It means that correlations should be investigated to determine a cause. However, correlations should not be interpreted as evidence of one variable causing a change in another variable. Complex business environments often present many complex causes and related data with variable correlations lacking causation.

For example, an increase in consumer spending and revenue may occur at the same time as positive media coverage, but it may have a different cause, such as movement into a new emerging market [88]. Correlations between two things can be caused by a third factor that affects both of them called "Confounder" [89]. So the less the information we have the more we are forced to observe correlations. Similarly, the more information we have the more transparent things will become and the more we will be able to see the actual casual relationships [90].

3.6 Data Visualization

3.6.1 Microsoft Excel

Microsoft Excel is a spreadsheet program developed by Microsoft that is used to record and analyse numerical data. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. Excel supports charts, graphs, or histograms generated from specified groups of cells. The generated graphic component can either be embedded within the current sheet or added as a separate object. These displays are dynamically updated if the content of cells changes [91].

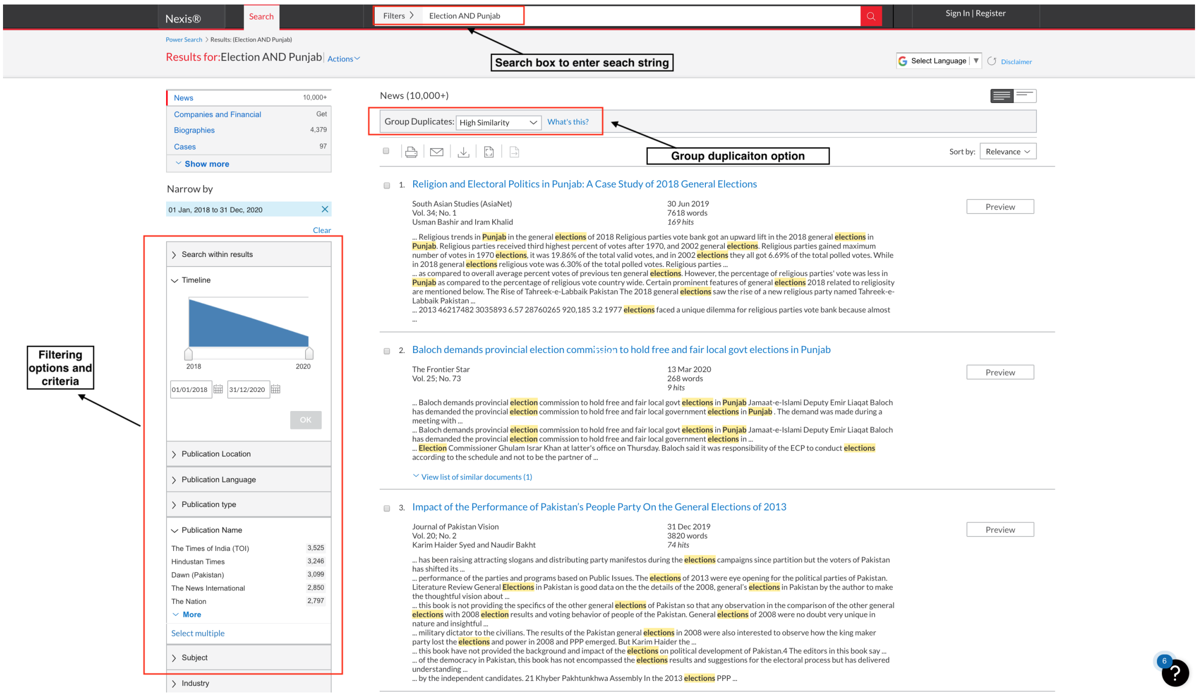
Microsoft Excel has more DataViz capabilities than people may realize. The time-series obtained was exported to CSV file format and used using Excel. It was used as the main visualization tool for the purpose of this entire research. Additional features are available using add-ins. For example, Analysis ToolPak, It Provides data analysis tools for statistical and engineering analysis. It was used to find the correlation coefficient and further create heatmap out if it. A heat-map is a graphical representation of data where values are depicted by multiple colors. Visualizing complex data using heatmap makes it easy to understand it at a glace.

3.7 Implementation

3.7.1 LexisNexis News & Business

LexisNexis is a corporation providing computer-assisted legal research (CALR) as well as business research and risk management services. During the 1970s, LexisNexis pioneered the electronic accessibility of legal and journalistic documents. As of 2006, the company had the world's largest electronic database for legal and public-records related information [92].

More specifically, LexisNexis News & Business service was used to download the electronic version of newspaper articles to be used in this study for extracting sentiment. It is an authoritative and recommend research tool For news, companies and markets Insights, multiple legal practice areas, and Business and Science biographies by. LexisNexis News & Business can be accessed using TCD credential provided to students for accessing other Trinity college student portals and other services. I or by accessing it from anywhere in the college campus which then uses network IP to auto-login the user. Figure 3.4 shows the user interface of the LexisNexis tool. It provides multiple filtering and query criteria to choose from.



1. Religion and Electoral Politics in Punjab: A Case Study of 2018 General Elections

South Asian Studies (AsiaNet)
Vol. 34; No. 1
Usman Bashir and Iram Khalid

30 Jun 2019
7618 words
169 hits

... Religious trends in Punjab in the general elections of 2018 Religious parties vote bank got an upward lift in the 2018 general elections in Punjab. Religious parties received third highest percent of votes after 1970, and 2002 general elections. Religious parties gained maximum number of votes in 1970 elections, it was 19.86% of the total valid votes, and in 2002 elections they all got 6.69% of the total polled votes. While in 2018 general elections religious vote was 6.30% of the total polled votes. Religious parties ...
... as compared to overall average percent votes of previous ten general elections. However, the percentage of religious parties' vote was less in Punjab as compared to the percentage of religious vote country wide. Certain prominent features of general elections 2018 related to religiosity are mentioned below. The Rise of Tahreek-e-Labbaik Pakistan The 2018 general elections saw the rise of a new religious party named Tahreek-e-Labbaik Pakistan ...
... 2013 46217482 3035893 6.57 28760265 920,185 3.2 1977 elections faced a unique dilemma for religious parties vote bank because almost
...

2. Baloch demands provincial election commission to hold free and fair local govt elections in Punjab

The Frontier Star
Vol. 25; No. 73

13 Mar 2020
268 words
9 hits

... Baloch demands provincial election commission to hold free and fair local govt elections in Punjab Jamaat-e-Islami Deputy Emir Liaqat Baloch has demanded the provincial election commission to hold free and fair local government elections in Punjab . The demand was made during a meeting with ...
... Baloch demands provincial election commission to hold free and fair local govt elections in Punjab Jamaat-e-Islami Deputy Emir Liaqat Baloch has demanded the provincial election commission to hold free and fair local government elections in ...
... Election Commissioner Ghulam Israr Khan at latter's office on Thursday. Baloch said it was responsibility of the ECP to conduct elections according to the schedule and not to be the partner of ...

View list of similar documents (1)

3. Impact of the Performance of Pakistan's People Party On the General Elections of 2013

Journal of Pakistan Vision
Vol. 20; No. 2
Karim Haider Syed and Naudir Bakht

31 Dec 2019
3820 words
74 hits

... has been raising attracting slogans and distributing party manifestos during the elections campaigns since partition but the voters of Pakistan has shifted its ...
... performance of the parties and programs based on Public Issues. The elections of 2013 were eye opening for the political parties of Pakistan. Literature Review General Elections in Pakistan is good data on the the details of the 2008, general's elections in Pakistan by the author to make the thoughtful vision about ...
... this book is not providing the specifics of the other general elections of Pakistan so that any observation in the comparison of the other general elections with 2008 election results and voting behavior of people of the Pakistan. General elections of 2008 were no doubt very unique in nature and insightful ...
... military dictator to the civilians. The results of the Pakistan general elections in 2008 were also interested to observe how the king maker party lost the elections and power in 2008 and PPP emerged. But Karim Haider the ...
... this book have not provided the background and impact of the elections on political development of Pakistan.4 The editors in this book say ...
... of the democracy in Pakistan, this book has not encompassed the elections results and suggestions for the electoral process but has delivered understanding ...
... by the independent candidates. 21 Khyber Pakhtunkhwa Assembly In the 2013 elections PPP ...

Figure 3.4: LexisNexis News & Business user interface

The user needs to specify the query terms(s) and can join terms using AND, OR, WITHOUT etc. operators and make them more relevant to the study. Table 3.2 lists the query terms used for this study for different case studies. The search box is located at the top of the page. Several filtering options such as date range and publication type, location, language etc. can be used to make our search more refined. When viewing news results, it provides the option to choose whether you want to use similarity analysis to process your search results. Similarity analysis analyzes a results list, identifies news stories that have similar content, and groups the similar news articles together. It was set to High similarity where news articles must be nearly identical for the service to include them in the same group of similar articles. Due to service restrictions, it was only possible to select and download a maximum of 100 hundred articles at any given time from the Lexis Nexis tool. While downloading, LexisNexis also providing various formatting option such as bold, italics, etc. and file type option such as pdf, RTF etc. The files collected this study were downloaded in Rich text format (.rtf) file format with all formation removed form it.

3.7.2 Python

Python is an interpreted, high-level, general-purpose programming language. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. It supports structured programming, object-oriented programming as well as the functional programming language and others. Python's unique attribute and is easy to use when it comes to quantitative and analytical computing.

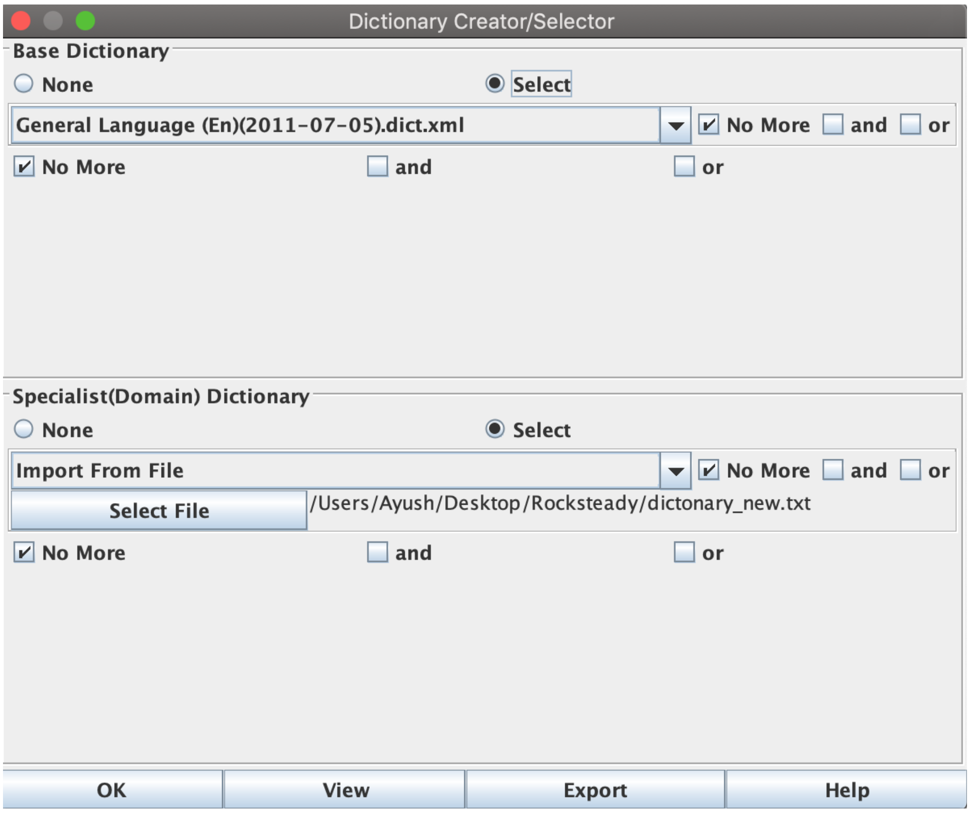
The downloaded RTF files were not compatible with the Rocksteady analysis system. It was because in February 2020, LexisNexis transitioned its database services to the Amazon Web Services cloud architecture and shut down its legacy mainframes and servers and therefore few structural changes were made to the database.

To make it compatible with Rocksteady, Every downloaded RTF document was passed and processed to be converted into a proper format using a template text (.txt) file which was created specifically for this conversion process. The structure of the template has been included for reference in Appendix A1.2.

The final processed file which consists of all documented in the precise format and was then concatenated, thus creating a single corpus file containing all the news articles, from all selected sources.

3.7.3 Rocksteady

Rocksteady is an affect analysis system developed at Trinity College Dublin by Ahmad and Zemánková [71] was used to calculate the affect score of a data corpus as described in detail in subsection 3.4.1. The system uses a General English language base dictionary and allows the user to import a custom domain specific dictionary as mentioned in subsection 3.4.3 Figure 3.5 here shows the Dictionary selector dialog box which is promoted the first time program runs.



Base Dictionary

○ None ● Select

General Language (En)(2011-07-05).dict.xml ☑ No More □ and □ or

☑ No More □ and □ or

Specialist(Domain) Dictionary

○ None ● Select

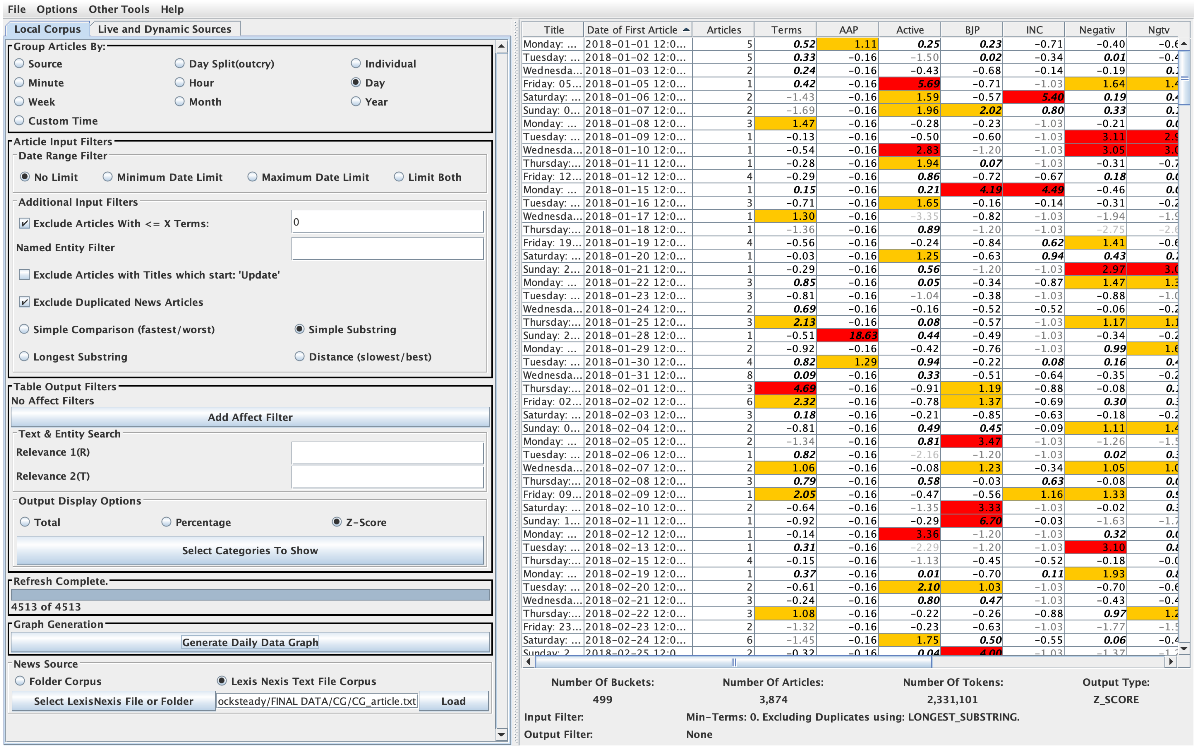
Import From File ☑ No More □ and □ or

Select File /Users/Ayush/Desktop/Rocksteady/dictonary_new.txt

☑ No More □ and □ or

Figure 3.5: Rocksteady Dictionary Selector

The selector provides an option to select multiple dictionaries using an option such as AND, OR or if only single then NO MORE. Once the dictionary preferences are saved. Figure 3.6 shows the user interface of the Rocksteady program.



Group Articles By:
○ Source ○ Day Split(outcry) ○ Individual
○ Minute ○ Hour ● Day
○ Week ○ Month ○ Year
○ Custom Time

Article Input Filters
Date Range Filter
● No Limit ○ Minimum Date Limit ○ Maximum Date Limit ○ Limit Both

Additional Input Filters
☑ Exclude Articles With <= X Terms: 0
Named Entity Filter
☐ Exclude Articles with Titles which start: 'Update'
☑ Exclude Duplicated News Articles
○ Simple Comparison (fastest/worst) ● Simple Substring
○ Longest Substring ○ Distance (slowest/best)

Table Output Filters
No Affect Filters
Text & Entity Search
Relevance 1(R)
Relevance 2(T)

Output Display Options
○ Total ○ Percentage ● Z-Score

Refresh Complete.
4513 of 4513

Graph Generation

News Source
○ Folder Corpus ● Lexis Nexis Text File Corpus
ocksteady/FINAL DATA/CG/CG_article.txt

Title	Date of First Article ▲	Articles	Terms	AAP	Active	BJP	INC	Negativ	Ngtv
Monday: ...	2018-01-01 12:0...	5	0.52	1.11	0.25	0.23	-0.71	-0.40	-0.6
Tuesday: ...	2018-01-02 12:0...	5	0.33	-0.16	-1.50	0.02	-0.34	0.01	-0.4
Wednesda...	2018-01-03 12:0...	2	0.24	-0.16	-0.43	-0.68	-0.14	-0.19	0.1
Friday: 05...	2018-01-05 12:0...	1	0.42	-0.16	5.69	-0.71	-1.03	1.64	1.4
Saturday: ...	2018-01-06 12:0...	2	-1.43	-0.16	1.59	-0.57	5.40	0.19	0.4
Sunday: 0...	2018-01-07 12:0...	2	-1.69	-0.16	1.96	2.02	0.80	0.33	0.1
Monday: ...	2018-01-08 12:0...	3	1.47	-0.16	-0.28	-0.23	-1.03	-0.21	0.0
Tuesday: ...	2018-01-09 12:0...	1	-0.13	-0.16	-0.50	-0.60	-1.03	3.11	2.9
Wednesda...	2018-01-10 12:0...	1	-0.54	-0.16	2.83	-1.20	-1.03	3.05	3.0
Thursday:...	2018-01-11 12:0...	1	-0.28	-0.16	1.94	0.07	-1.03	-0.31	-0.7
Friday: 12...	2018-01-12 12:0...	4	-0.29	-0.16	0.86	-0.72	-0.67	0.18	0.0
Monday: ...	2018-01-15 12:0...	1	0.15	-0.16	0.21	4.19	4.49	-0.46	0.0
Tuesday: ...	2018-01-16 12:0...	3	-0.71	-0.16	1.65	-0.16	-0.14	-0.31	-0.2
Wednesda...	2018-01-17 12:0...	1	1.30	-0.16	-3.35	-0.82	-1.03	-1.94	-1.9
Thursday:...	2018-01-18 12:0...	1	-1.36	-0.16	0.89	-1.20	-1.03	-2.75	-2.6
Friday: 19...	2018-01-19 12:0...	4	-0.56	-0.16	-0.24	-0.84	0.62	1.41	-0.6
Saturday: ...	2018-01-20 12:0...	1	-0.03	-0.16	1.25	-0.63	0.94	0.43	0.2
Sunday: 2...	2018-01-21 12:0...	1	-0.29	-0.16	0.56	-1.20	-1.03	2.97	3.0
Monday: ...	2018-01-22 12:0...	3	0.85	-0.16	0.05	-0.34	-0.87	1.47	1.3
Tuesday: ...	2018-01-23 12:0...	3	-0.81	-0.16	-1.04	-0.38	-1.03	-0.88	-1.0
Wednesda...	2018-01-24 12:0...	2	0.69	-0.16	-0.16	-0.52	-0.52	-0.06	-0.2
Thursday:...	2018-01-25 12:0...	3	2.13	-0.16	0.08	-0.57	-1.03	1.17	1.1
Sunday: 2...	2018-01-28 12:0...	1	-0.51	18.63	0.44	-0.49	-1.03	-0.34	-0.2
Monday: ...	2018-01-29 12:0...	2	-0.92	-0.16	-0.42	-0.76	-1.03	0.99	1.6
Tuesday: ...	2018-01-30 12:0...	4	0.82	1.29	0.94	-0.22	0.08	0.16	0.4
Wednesda...	2018-01-31 12:0...	8	0.09	-0.16	0.33	-0.51	-0.64	-0.35	-0.2
Thursday:...	2018-02-01 12:0...	3	4.69	-0.16	-0.91	1.19	-0.88	-0.08	0.1
Friday: 02...	2018-02-02 12:0...	6	2.32	-0.16	-0.78	1.37	-0.69	0.30	0.3
Saturday: ...	2018-02-03 12:0...	3	0.18	-0.16	-0.21	-0.85	-0.63	-0.18	-0.2
Sunday: 0...	2018-02-04 12:0...	2	-0.81	-0.16	0.49	0.45	-0.09	1.11	1.4
Monday: ...	2018-02-05 12:0...	2	-1.34	-0.16	0.81	3.47	-1.03	-1.26	-1.5
Tuesday: ...	2018-02-06 12:0...	1	0.82	-0.16	-2.16	-1.20	-1.03	0.02	0.3
Wednesda...	2018-02-07 12:0...	2	1.06	-0.16	-0.08	1.23	-0.34	1.05	1.0
Thursday:...	2018-02-08 12:0...	3	0.79	-0.16	0.58	-0.03	0.63	-0.08	0.0
Friday: 09...	2018-02-09 12:0...	1	2.05	-0.16	-0.47	-0.56	1.16	1.33	0.9
Saturday: ...	2018-02-10 12:0...	2	-0.64	-0.16	-1.35	3.33	-1.03	-0.02	0.3
Sunday: 1...	2018-02-11 12:0...	1	-0.92	-0.16	-0.29	6.70	-0.03	-1.63	-1.7
Monday: ...	2018-02-12 12:0...	1	-0.14	-0.16	3.36	-1.20	-1.03	0.32	0.0
Tuesday: ...	2018-02-13 12:0...	1	0.31	-0.16	-2.29	-1.20	-1.03	3.10	0.8
Thursday:...	2018-02-15 12:0...	4	-0.15	-0.16	-1.13	-0.45	-0.52	-0.18	-0.0
Monday: ...	2018-02-19 12:0...	1	0.37	-0.16	0.01	-0.70	0.11	1.93	0.8
Tuesday: ...	2018-02-20 12:0...	2	-0.61	-0.16	2.10	1.03	-1.03	-0.70	-0.6
Wednesda...	2018-02-21 12:0...	3	-0.24	-0.16	0.80	0.47	-1.03	-0.43	-0.4
Thursday:...	2018-02-22 12:0...	3	1.08	-0.16	-0.22	-0.26	-0.88	0.97	1.2
Friday: 23...	2018-02-23 12:0...	2	-1.32	-0.16	-0.23	-0.63	-1.03	-1.77	-1.5
Saturday: ...	2018-02-24 12:0...	6	-1.45	-0.16	1.75	0.50	-0.55	0.06	-0.4
Sunday: 2	2018-02-25 12:0	2	-0.32	-0.16	0.04	4.00	-1.03	-1.37	-1.2

Number Of Buckets:	Number Of Articles:	Number Of Tokens:	Output Type:
499	3,874	2,331,101	Z_SCORE

Input Filter: Min-Terms: 0. Excluding Duplicates using: LONGEST_SUBSTRING.
Output Filter: None

Figure 3.6: Rocksteady's user interface

The program offers numerous option and features including group article by options, date filter, string matching to remove duplicate articles, output's scoring option and generating daily data graphs from extracted sentiment features. The flexible nature of the program then allows the user to export the extracted sentiment in various file format. For the purpose of this study, all the output data was exported to CSV file format which could then be used to perform statistical analysis on the data and visualize it.

4 Results & Discussion

4.1 Introduction

This chapter provides a comprehensive overview on the three different case studies which was chosen for this study which focused on three different Indian states namely, Chhattisgarh, Punjab and Karnataka and discusses the results and outcomes from different phases of work using the methodologies outlined in the previous chapter (3). Table 4.1 lists some of the facts which would be essential to comprehend the different case-studies and the results and discussion which follow them.

Table 4.1: Fact and useful information required for case studies

Winner of 2019 Indian General Election (GE)	BJP
Prime Minister of India	Narendra Modi
Bharatiya Janata Party (the ruling Party)	BJP
Indian National Congress (the opposition party)	INC
Winner of 2018 Chhattisgarh Legislative Assembly election	INC
Winner of Majority Seats in Chhattisgarh during GE 2019	BJP
Winner of 2017 Punjab Legislative Assembly election	INC
Winner of Majority Seats in Punjab during GE 2019	INC
Winner of 2018 Karnataka Legislative Assembly election	BJP
Winner of Majority Seats in Karnataka during GE 2019	BJP

The first section described the data acquired which was used in this study. The section covers 'Sentiment Analysis' using Rocksteady on the collected data corpus which is further divided into three subsections for each of the case studies, each comprising of 'Negative Sentiment Overview', 'Descriptive Statistical Analysis', 'Correlation Analysis', 'Newspaper Polarity Analysis' with a comparison and contrast between the Assembly Election and Indian General Election help in that chosen territory (or state). Also, not all of the results were striking or significant as is to be expected with this type of analysis. Therefore there will be a significant focus on the more meaningful results in this section and few others are present at Appendix for more insight if required.

4.2 Data Acquisition

This section offers an overview of the accumulated text corpus of Indian newspaper articles from three different states for a time period of 3 months before and 3 months after the Assembly Elections and General Election in the respective regions and reflects the approach taken for retrieval.

4.2.1 Text corpus of news articles

Amongst the top English newspaper in the country, these five English newspapers (Table 3.1 were selected from which the newspaper articles were collected to address different case studies. These newspapers were chosen using the newspaper circulation data from the Audit Bureau of Circulation (ABC) [11] for Indian English newspapers and in conjecture with INDIAN READERSHIP SURVEY 2019 Q1 [2] conducted by Media Research Users Council (MRUC) [93]. Figure 4.1 shows the year-wise circulation figures of each publication.

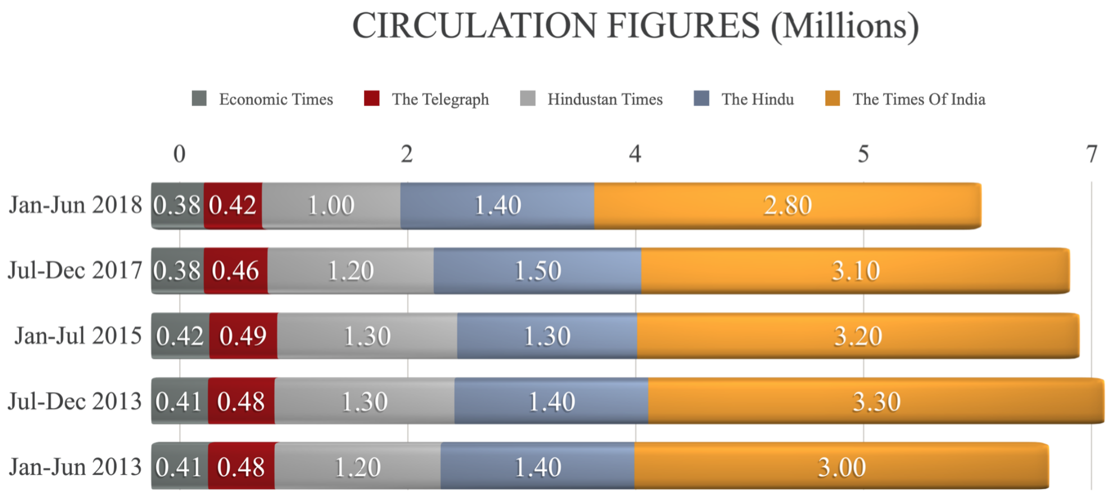


Figure 4.1: Circulation figure for top five Indian English Newspaper

Table 4.2 shows the newspaper with most readership according to the Indian Readership Survey (IRS) [2]. For daily circulation figures, refer to Table 3.1 in section 3.2.

Table 4.2: Top Six most read English Dailies in India (Fig in 000's)

Name of Publication	IRS 2017	IRS 2019 Q1
The Times of India	13045	15236
Hindustan Times	6847	7675
The Hindu English	5300	6226
The Economic Times	3103	3701
Mumbai Mirror	1813	2165
The Indian Express	1599	1855

A quick glance at both the table and the chart above clearly asserts the dominance of Time of India (TOI) over all other English Dailies in terms of both circulation and readership. India of readership number, circulation figures were chosen as the criteria for newspaper selection because a bigger circulation number would suggest a higher reach and wider audience and subsequently larger readership, whereas if readership numbers are considered, it might also be due to the popularity of a certain publication in a higher population density region. For instance, According to IRS Report Q1 2019 [2]. Mumbai Mirror and Economic Time both ranked higher than Indian Express and The Telegraph, but circulation figures suggested otherwise. It can be inferred from the fact that Mumbai Mirror is a compact newspaper available in the city of Mumbai, Maharashtra with an estimated population of over 24 Million and thus higher readership, but it does capture the sentiment of other states and nation as a whole and thus was excluded.

Another noticeable discrepancy is the exclusion of Economic Times from the selected list of publications for this study, even though it was top 5 in both circulation and readership rankings. This particular decision was taken as it is a business newspaper and its main content is based on the Indian economy, international finance, share prices, prices of commodities as well as other matters related to finance. They cover general new items too but they are far less in comparison. Instead, the next most popular news publication is chosen, The Indian Express, with 6th largest reader accordion to IRS 2019 Q1 [2].

The retrieval of news articles for the individual state's corpus of text data was performed by using the Lexis Nexis News & Business aggregation service as mentioned in subsection 3.7.1 (chapter 3). The news publications source were selected based on circulation (figure 4.1). The query term used to search and gather the text articles pertaining to each of the three states was the term "Election" followed by "AND" (footer, 2 or more words anywhere in the document) operator and "Name of the state", requiring the database to look for articles containing both. Additionally, the query criteria also restricted to search for articles with major mentions and remove duplicate with High Similarity option (recall Table 3.2)

Table 4.3 shows the breakdown of the articles obtained from each publication across the three states during their State Assembly Election respectively and Table 4.4 articles retrieved during the General Election 2019 focusing on three individual states.

Table 4.3: Breakdown of Articles as retried by publication during Assembly State Election

Chhattisgarh		**Punjab**		**Karnataka**	
Name	Count	Name	Count	Name	Count
The Times of India	631	The Times of India	1153	The Times of India	1529
Hindustan Times	827	Hindustan Times	1960	Hindustan Times	1103
The Hindu	3	The Hindu	16	The Hindu	12
The Indian Express	424	The Indian Express	828	The Indian Express	739
The Telegraph	187	The Telegraph	105	The Telegraph	324

Table 4.4: Breakdown of Articles as retried by publication during General Election 2019

Chhattisgarh		**Punjab**		**Karnataka**	
Name	Count	Name	Count	Name	Count
The Times of India	267	The Times of India	1076	The Times of India	645
Hindustan Times	352	Hindustan Times	943	Hindustan Times	561
The Hindu	19	The Hindu	54	The Hindu	131
The Indian Express	154	The Indian Express	462	The Indian Express	240
The Telegraph	43	The Telegraph	67	The Telegraph	88

4.2.2 Domain specific glossary creation

A new specialised domain-specific dictionary covering all political features specific to Indian General Election and three state election was created for use in combination with base English General Dictionary for sentiment Analysis using Rocksteady as specific in detail in section 3.4.3. The "Indian Election" glossary created by Roy [9] and used by Singh [10] was used as a foundation and was modified to include the names of political leaders, slogan, regional parties and more for better sentiment extraction and avoid any misclassification due to the state level political terms. Table A1.1 in Appendix A1.4 provides the breakdown of the number of political terms added to the new glossary which were mapped to the various political features.

4.3 Sentiment Analysis

The Rocksteady effect analysis system (recall 3.4.1) was used to perform sentiment analysis on each of the region-specific corpora gathered. The specialised domain-specific dictionary mentioned in the previous section (4.2.2) in conjunction with a base General English dictionary was used for sentiment classification and calculating affect score as mentioned in 3.7.3 (chapter

3). Imported articles were grouped by day and proxy time series for the negative sentiment was obtained for each of three case studies, the results of which are discussed in detail in the following subsection.

4.3.1 Case Study: Chhattisgarh

This case study deals with a Central-Indian state of Chhattisgarh, which saw INC, the ruling party of the state as of 2018 State Assembly Election, lost majority of Parliamentary seats to the opposition party BJP in General Election of 2019. This saw a shift for INC, from a majority in the state assembly to minority in General election.

4.3.1.A Descriptive Statistical Analysis

i) Negative Sentiments

Figure 4.2 shows a proxy effect time series for negative sentiment during Chhattisgarh State Election from a period of 1st August 2018 to 28th February 2018 and Figure 4.3 shows the same for General Elections 2019 from 1st March 2019 to 31st August 2019.

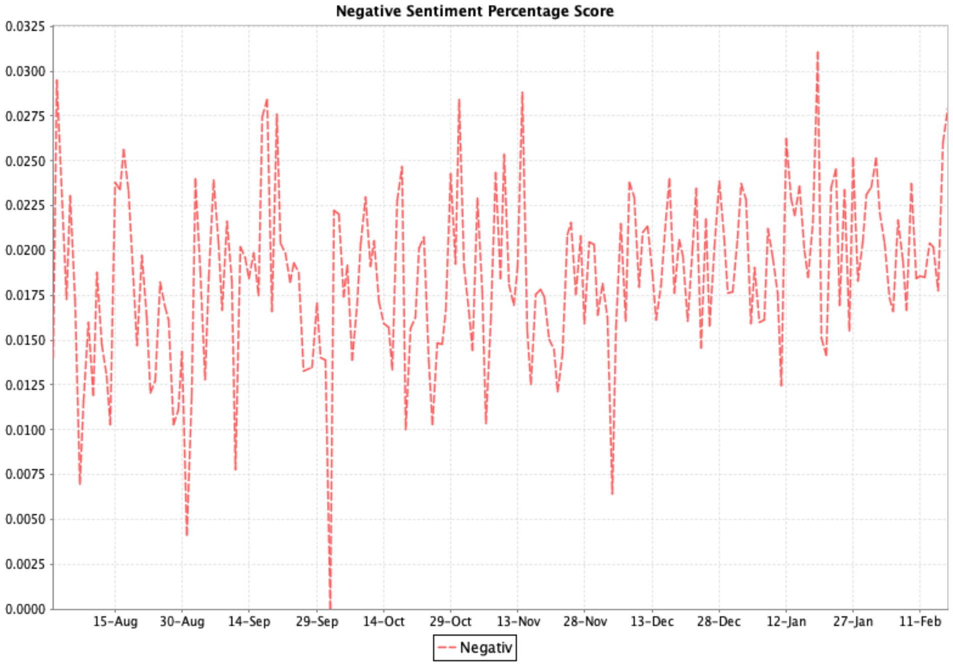


Figure 4.2: CG: Negative Sentiment Affect Time Series (State Elections)

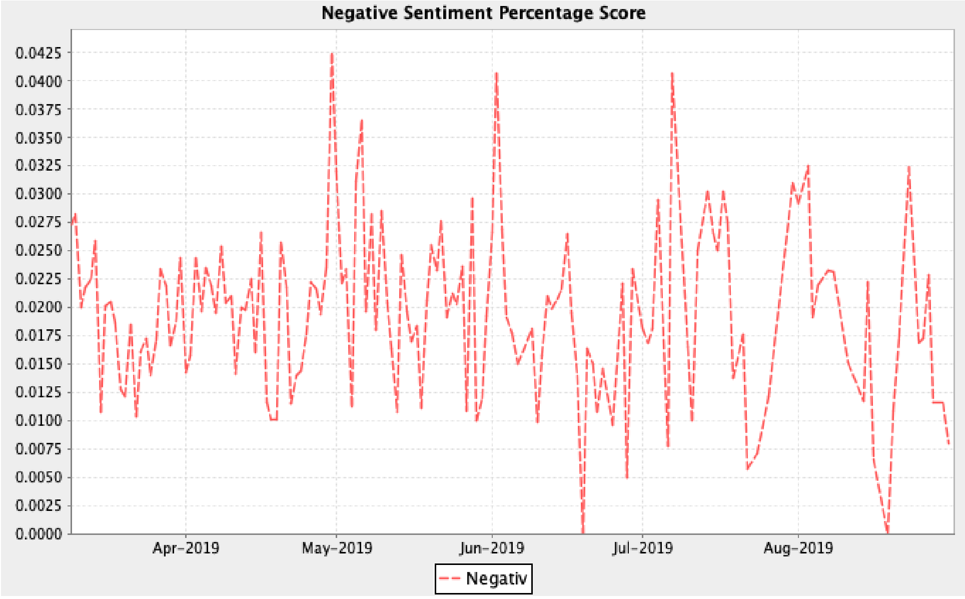


Figure 4.3: CG: Negative Sentiment Affect Time Series (General Elections)

It can be observed in Figure 4.2 that the negative sentiment is almost constant throughout with few outliers whereas in Figure 4.3 it the same but there is drop is negative sentiment post (May 2019) completion of the election. Table 4.5 presents the Descriptive statistic for Negative Sentiment expressed in newspaper articles during both State Assembly Election and General Election held in Chhattisgarh.

Descriptive Statistics	**Negative Sentiment during 2018 State Election**	**Negative Sentiment during 2019 General Election**
Minimum	0	0
Maximum	4.62	4.24
Range	4.62	4.24
Mean	1.84	1.93
Standard Deviation (σ)	0.55	0.72
Kurtosis	3.54	0.72
Skewness	0.59	0.28

Table 4.5: Descriptive statistics of Negative Sentiment expressed in newspapers from Chhattisgarh

The skewness value of 0.59 and 0.28 indicate that the distribution was symmetrical for both state and general elections respectively. With a Kurtosis of 3.54 during State Elections indicated that during some days there was a higher negative sentiment expressed in the newspaper. It can be attributed to the fact that during election political party often projects the rival in a negative light. With a higher mean of 1.93, during GE 2019 as compared to 1.84 during State Elections indicate a rise in the negative sentiment expressed in the article during later.

The descriptive statistical analysis for Positive sentiment shows a higher mean attached to positive sentiment. Figure A1.1 in Appendix A1.3 showcase the comparison. From the comparison it is evident that the newspaper articles had a higher positive sentiment attached to them. This can be attributed to the fact that in general day to day communicate, the underlying tone is considered to be more positive than negative and hence higher positive sentiment.

ii) Political Features

The Descriptive Statistics for a few of the essential political features which were extracted using Rocksteady for both Chhattisgarh State Assembly Election listed in Table 4.6 and General Election in Table 4.7 are presented below. Although the number of political features extracted was much larger as listed in Appendix A1.4 Table A1.1, only a few key ones are discussed here. The following selected features are Number of Articles (Articles), Political Parties (PolitP), Political Leader (PolitLead), National Parties (PolitNat), Regional Parties (PolitReg), Political Campaign (PolitCamp), BJP and INC. The term mentioned within the bracket in the corresponding identity used a feature in the special glossary. For the other two case studies later in this chapter, the features are addressed by the features name as presented in Appendix A1.4 Table A1.1. All the case-studies uses the same aforementioned political features in their respective subsections.

Table 4.6: Descriptive statistics of Political Features for Chhattisgarh State Election expressed during Aug 2018 to Feb 2019

Series	Min	Max	Range	Mean	(σ)	Kurtosis	Skewness
Number of Articles	1	123	122	9.74	11.26	53	6.23
Political Parties	0	4.35	4.35	1.26	0.68	1.61	0.88
Political Leaders	0	2.31	2.31	0.72	0.35	2.07	0.96
National Parties	0	3.21	3.21	0.98	0.54	1.51	0.93
Regional Parties	0	1.75	1.75	0.27	0.28	5.74	2.04
Political Campaign	0	0.21	0.21	0.01	0.03	12.59	3.23
BJP	0	1.26	1.26	0.32	0.19	2.82	1.28
INC	0	1.77	1.77	0.27	00.25	8.35	2.29

Table 4.7: Descriptive statistics of Political Features for Chhattisgarh in General Election expressed during Mar 2019 to Aug 2019

Series	Min	Max	Range	Mean	(σ)	Kurtosis	Skewness
Number of Articles	1	49	48	5.94	5.89	17.51	3.13
Political Parties	0	4.16	4.16	1.02	0.66	2.51	0.99
Political Leaders	0	5.29	5.29	0.75	0.61	20.09	3.59
National Parties	0	2.16	2.16	0.80	0.49	-0.47	0.23
Regional Parties	0	2.97	2.97	0.21	0.37	29.62	4.76
Political Campaign	0	0.16	0.16	0.01	0.02	11.07	3.18
BJP	0	1.76	1.76	0.32	0.31	7.29	2.32
INC	0	3.53	3.53	0.37	0.45	18.04	3.59

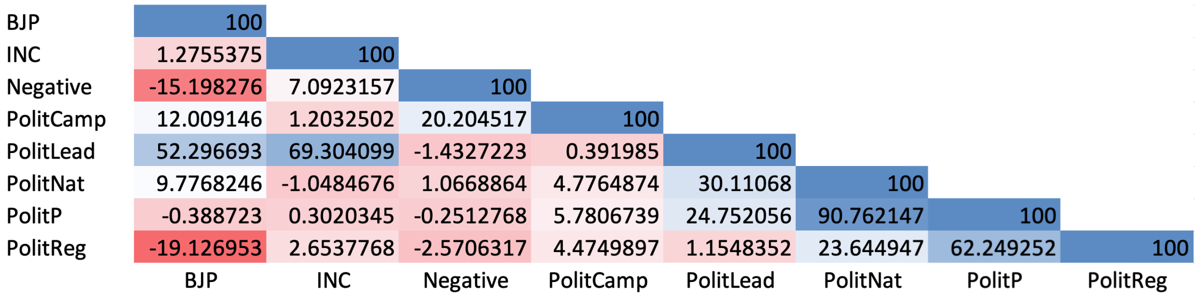
The figures from both the tables look very similar to each other in many political features. The higher positive kurtosis of 53 and 17.51 and skewness of 6.23 and 3.31 respectively from both table 4.6 and table 4.7 for Number of articles indicates the asymmetrical distribution in newspapers during the election period and that as we approach the election month, their in sudden influx and more converge to the election by the newspapers. This is confirmed by drawing a daily affect series graph for the political feature (Articles) which shows the peaks and spikes as we approach election dates.

Interesting to note is the mean for BJP which was much higher at 0.32 than that of INC suggesting more media coverage for BJP as compared to INC. This is notable due to the fact that INC went on to win the election even with less coverage. The mean for BJP remains the same at 0.32 during 2019 General Election too, whereas INC sees a significant jump from 0.27 in 2018 to 0.37 in 2019 indicating much larger media coverage, surpassing that of BJP. This can be attributed to the fact that INC won the 2018 election in the states and even which higher converge they lost the majority of seats in 2019 GE. The higher coverage can be used to make an assumption that INC ruling states and power in the state helped them gather more coverage, although not necessarily to their advantage. Another interesting observation is that the Kurtosis of INC of 8.25 and 18.04 as compared to 2.82 and 7.29 highlights that INC has gotten more coverage on specific days and events and bought under the limelight a lot more.

4.3.1.B Correlation Analysis

This section shows results for the correlation analysis made between specific key political features and negative sentiment. Figure 4.4 shows the heat-map illustrating the correlation coefficients between selected political features and negative sentiment for 2018 Chhattisgarh State Legislative Assembly Election.

The heatmap here shows the relative intensity correlation values by assigning colour. Heatmaps representing the correlation between features and negative sentiment in State Elections throughout this study (Figure 4.8 and Figure 4.12) follows the same colour-scheme (blue-red). Those that show the highest positive correlation are given a “cool” blue colour, while those with negative correlation are represented by "hot" red colour. The lighter shades of each colour, blue and red represent their relative values to their highest correlation coefficient values of 100% and -100% respectively.

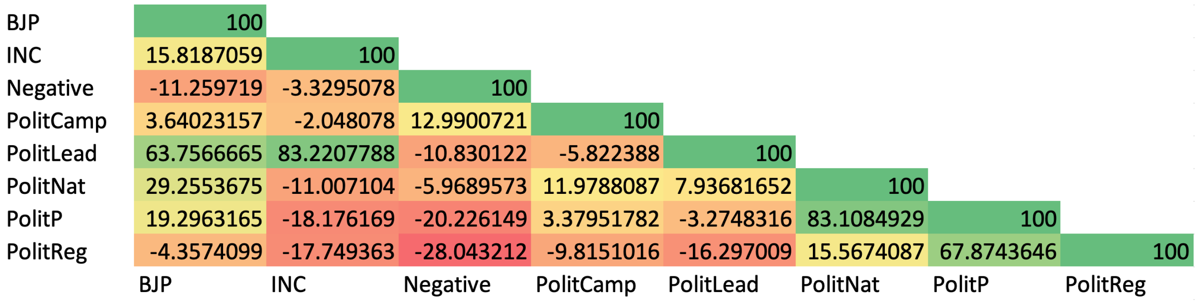


	BJP	INC	Negative	PolitCamp	PolitLead	PolitNat	PolitP	PolitReg
BJP	100							
INC	1.2755375	100						
Negative	-15.198276	7.0923157	100					
PolitCamp	12.009146	1.2032502	20.204517	100				
PolitLead	52.296693	69.304099	-1.4327223	0.391985	100			
PolitNat	9.7768246	-1.0484676	1.0668864	4.7764874	30.11068	100		
PolitP	-0.388723	0.3020345	-0.2512768	5.7806739	24.752056	90.762147	100	
PolitReg	-19.126953	2.6537768	-2.5706317	4.4749897	1.1548352	23.644947	62.249252	100

Figure 4.4: Heat-map presenting the correlation (multiplied by 100) between selected political features and negative sentiment for 2018 Chhattisgarh State Legislative Assembly Election

Leading up to the 2018 CG State Assembly elections, BJP is less likely to be mentioned in articles containing all political parties or regional political parties in comparison to INC with correlation equal to -0.38% and -19% for BJP and 0.30% and 2.65% for INC respectively. Also, articles had much higher mentions of Political leader from INC with a correlation between the two at 69.30% then for a politician from BJP at 52.29%. Consistent with their superb campaign and catchy slogan, BJP will have much higher mentions of their political agenda and company slogan with a correlation of 12% compared to 1.20% for INC. Interestingly BJP had a negative correlation of -15% with the negative sentiment expressed during the elections as compared to the much higher 7.09% for INC for the same time period.

Figure 4.5 shows the heat-map illustrating the correlation coefficients between selected political features and negative sentiment for parliamentary seats contested in the 2019 General Election from Chhattisgarh Region. Heatmaps representing the correlation between features and negative sentiment in General Elections throughout this study (Figure 4.9 and Figure 4.13) follows the same colour-scheme (green-yellow-red). Those that show the highest positive correlation are given a “vibrant” green colour, while those with negative correlation are represented by "hot" red colour. The "yellow" represents the values in between the 0 and 100% where the "orange" represents values from 0 to -100%. The lighter shades of each colour, green and red represent their relative values to their highest correlation coefficient values of 100% and -100% respectively.



	BJP	INC	Negative	PolitCamp	PolitLead	PolitNat	PolitP	PolitReg
BJP	100							
INC	15.8187059	100						
Negative	-11.259719	-3.3295078	100					
PolitCamp	3.64023157	-2.048078	12.9900721	100				
PolitLead	63.7566665	83.2207788	-10.830122	-5.822388	100			
PolitNat	29.2553675	-11.007104	-5.9689573	11.9788087	7.93681652	100		
PolitP	19.2963165	-18.176169	-20.226149	3.37951782	-3.2748316	83.1084929	100	
PolitReg	-4.3574099	-17.749363	-28.043212	-9.8151016	-16.297009	15.5674087	67.8743646	100

Figure 4.5: Heat-map presenting the correlation (multiplied by 100) between selected political features and negative sentiment for 2019 General Election from Chhattisgarh Region

Although receiving higher overall mention's (recall table 4.6), INC is less likely to be mentioned in articles contain all political parties, national or regional parties in comparison to BJP with correlation equal to -18.17%, -11.00% and -17% for INC and 19.29%, 29.25% and -4.25% for BJP. Similar to the 2018 State election, articles had much higher mentions of Political Leader from INC with a correlation of 83.22% and 63.75% for BJP and the again comparable results with BJP's political campaigns being more favours in the news articles with a correlation of 3.6% and -2.04% for INC. INC saw a drop in the negative sentiment expressed with -3.32% correlation compared to 7.09%, which could be credited to their status as a ruling party.

Table 4.8 shows the overview of comparison between the two elections with rows featuring INC presented with the color green and BJP with saffron.

Correlation	2018 Chhattisgarh State Assembly Election	2019 General Election for Chhattisgarh Region
INC-PolitP	0.30%	-18.17%
BJP-PolitP	-0.38%	19.29%
INC-PolitNat	-1.04%	-11.00%
BJP-PolitNat	9.77%	29.25%
INC-PolitReg	2.65%	-17.00%
BJP-PolitReg	-19.12%	-4.25
INC-PolitLead	69.30%	83.22%
BJP-PolitLead	52.29%	63.75%
INC-PolitCamp	1.20%	-2.04%
BJP-PolitCamp	12.00%	3.6%
INC-Negative	7.09%	-3.32%
BJP-Negative	-15%	-11.25%

Table 4.8: Summary of correlations comparison between features for 2018 State Election and 2019 General Election from Chhattisgarh

4.3.1.C Newspaper Polarity

This section discusses the affect score calculated by Rocksteady for individual newspaper and presents the results for the coverage given to BJP and INC along with the overall negative and positive sentiment expressed by the dallies during the 2017 State and 2019 general elections in Chhattisgarh.

Table 4.9: Affect score for political features as expressed by chosen news paper during 2018 CG State Election

Source	BJP	INC	Negative	Positive
The Times of India (TOI)	0.35	0.36	1.84	3.12
Hindustan Times	0.30	0.30	1.82	4.05
The Hindu	0.49	0.23	1.66	2.73
The Telegraph	0.33	0.24	2.37	3.59
The Indian Express	0.31	0.23	2.06	3.16

Table 4.9 shows that BJP has a higher affect score for all the newspapers in comparison to INC indicating BJP received more coverage by all the publications during the state elections. Also as discussed before, all the newspaper papers had higher positive sentiment as compared to negative with greater affect score. The top two most circulated and read English newspaper, TOI and Hindustan Times almost showed no bias towards any political party with almost similar values and also both had a similar amount of negative sentiment expressed in their news articles.

The Hindu showed the most inclination towards BJP with value 0.49 compared to 0.23 for INC. Also, The Hindu had the least both negative and positive sentiment in their newspaper suggesting a more neutral tone with BJP at its centre. The Telegraph was with the most negative attached to its newspaper.

Table 4.10: Affect score for political features as expressed by chosen news paper during 2019 CG General Election

Source	BJP	INC	Negative	Positive
The Times of India (TOI)	0.34	0.34	1.83	3.08
Hindustan Times	0.31	0.28	1.95	3.35
The Hindu	0.27	0.19	1.62	2.78
The Telegraph	0.29	0.31	2.29	3.58
The Indian Express	0.26	0.36	2.25	3.12

Table 4.10 shows that BJP notably only has a higher affect score only in Hindustan Times and The Hindu with 0.31 and 0.27 for BJP and 0.29 and 0.19 for INC respectively. While The Telegraph and The Indian Express mentioned INC more with 0.31 and 0.36 in comparison to

0.29 and 0.29 for BJP. TOI remained neutral with 0.34 for both. Again all news dallies had higher positive sentiment as opposed to negative. In comparison to the 2018 election, INC saw a rise in coverage by The telegraph and The Indian Express, while BJP saw it decline in the same.

4.3.2 Case Study: Punjab

The second case study deals with a North-Indian state of Punjab, which saw INC, the ruling party of the state as of 2017 State Assembly Election, also winning the majority of Parliamentary seats from the state during the 2019 General Election. The result remained a constant, With INC winning both the election by a majority and replicating its success of 2017 in 2019.

4.3.2.A Descriptive Statistical Analysis

i) Negative Sentiments

Figure 4.2 shows a proxy affect time series for negative sentiment during Punjab State Election from a period of 1st November 2016 to 31 March 2017 and Figure 4.3 shows the same for General Elections 2019 from 1st March 2019 to 31st August 2019.

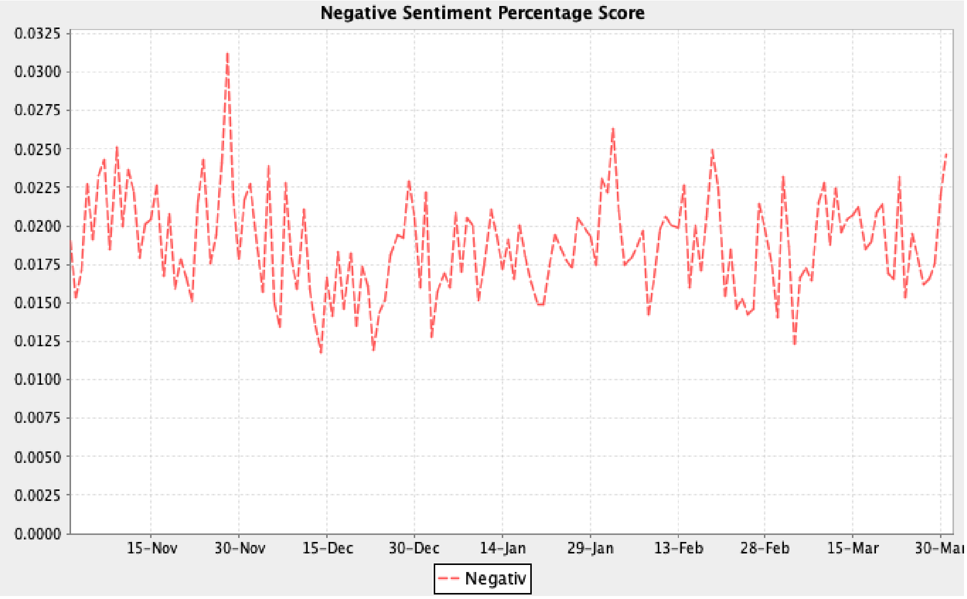


Figure 4.6: PB: Negative Sentiment Affect Time Series (State Elections)

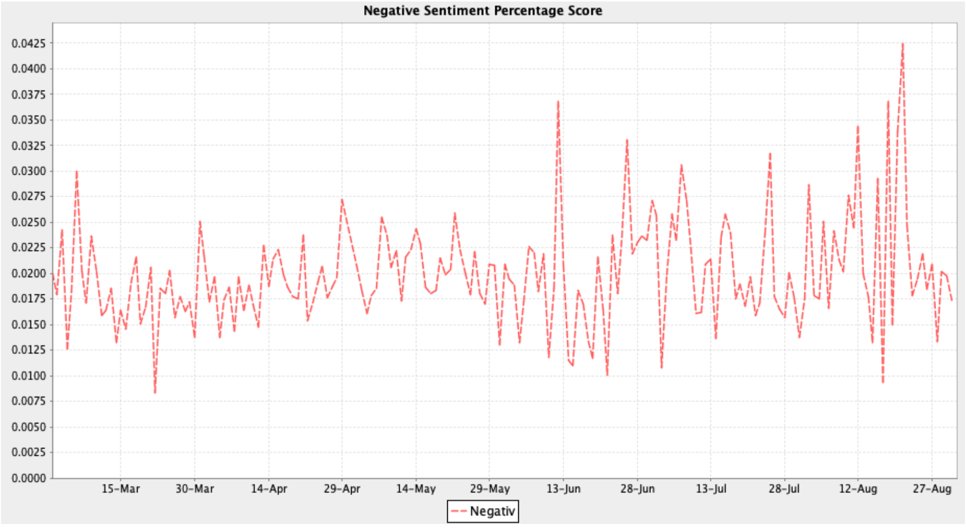


Figure 4.7: PB: Negative Sentiment Affect Time Series (General Elections)

It can be observed in Figure 4.6 that the negative sentiment is almost constant throughout election with whereas in Figure 4.7 it is the same but there is rise in negative sentiment post (May 2019) completion of the election.

Descriptive Statistics	Negative Sentiment during 2017 State Election	Negative Sentiment during 2019 General Election
Minimum	1.17	0.83
Maximum	3.12	4.23
Range	1.94	3.40
Mean	1.88	1.99
Standard Deviation (σ)	0.32	0.51
Kurtosis	0.39	2.71
Skewness	0.41	1.07

Table 4.11: Descriptive statistics of Negative Sentiment expressed in newspapers from Punjab

Kurtosis of 0.39 and skewness 0.4 indicates it is symmetrical distribution during state election, whereas during 2019 with a higher kurtosis of 2.71 and skewness of 1.07 suggests that certain days and events saw a spike in negative sentiment in the news articles. It is confirmed by figure 4.7 which shows spikes in the graph during certain time intervals. With a higher mean of 1.99 during GE209 as compared to 1.88 during the 2018 state election meant a rise in the negative sentiment expressed in news articles.

The descriptive statistical analysis for Positive sentiment shows a higher mean attached to positive sentiment. Figure A1.2 in Appendix A1.3 showcase the comparison, on comparison it is evident that the newspaper articles had a higher positive sentiment attached to them.

ii) Political Features

The Descriptive Statistics for a few of the essential political features which were extracted using Rocksteady for both Punjab State Assembly Election listed in Table 4.12, and General Election in Table 4.13 are presented below.

Table 4.12: Descriptive statistics of Political Features for Punjab State Election expressed during Nov 2016 to Mar 2017

Series	Min	Max	Range	Mean	(σ)	Kurtosis	Skewness
Number of Articles	4	118	114	26.88	16.65	5.20	1.75
Political Parties	0.15	1.98	1.83	1.05	0.30	0.59	0.34
Political Leaders	0.08	1.49	1.41	0.70	0.26	0.40	0.25
National Parties	0.05	1.20	1.15	0.35	0.19	3.55	1.56
Regional Parties	0.05	1.45	1.39	0.71	0.24	0.96	0.43
Political Campaign	0	0.18	0.18	0.01	0.01	39.23	5.16
BJP	0	0.65	0.65	0.13	0.08	13.32	2.82
INC	0.01	0.93	0.92	0.32	0.17	0.66	0.63

Table 4.13: Descriptive statistics of Political Features for Punjab in General Election expressed during Mar 2019 to Aug 2019

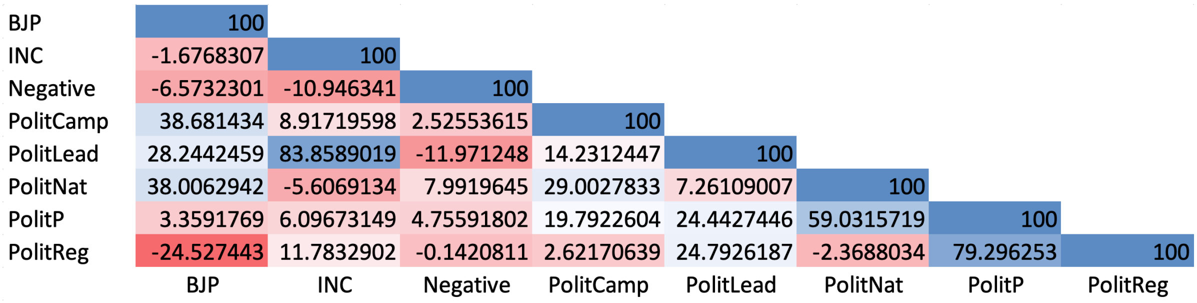
Series	Min	Max	Range	Mean	(σ)	Kurtosis	Skewness
Number of Articles	2	52	50	14.29	10.2	0.39	0.95
Political Parties	0	2.18	2.18	0.80	0.41	-0.11	0.21
Political Leaders	0.04	1.68	1.64	0.64	0.27	0.88	0.58
National Parties	0	1.23	1.23	0.31	0.21	1.96	1.06
Regional Parties	0	1.34	1.34	0.49	0.29	-0.49	0.15
Political Campaign	0	0.07	0.07	0.01	0.01	15.69	3.43
BJP	0	0.63	0.63	0.13	0.10	4.02	1.61
INC	0.02	1.60	1.58	0.39	0.24	4.36	1.71

The figures from both the tables look very similar to each other in almost all political features present. Following a similar trend as observed in the previous case study (subsection 4.3.1), the number of articles has a higher kurtosis value of 5.20 and Skewness of 1.75 indicating selected periods consists more articles which had mentioned regarding elections and various features associated with it. This behaviour is expected as news media publications often shift their entire coverage to elections and associated events as the voting day gets close.

The larger kurtosis for BJP at 13.32 suggested them being in the spotlight more on certain key days and events (mostly closer to the election) as compared to INC which was symmetrical throughout. The mean for BJP remained constant at 0.13 during both the elections. INC with a noticeable higher mean of 0.32 during the 2017 state election and even higher number at 0.39 during GE2019, clearly indicated the much higher coverage was received by INC in comparison to BJP during both the periods. Similar to the trend seen in case study 1, The jump in mean for INC after 2017 can be credited to the fact that INC was the state ruling party and was expected to be more in news.

4.3.2.B Correlation Analysis

This subsection shows results for the correlation analysis made between specific key political features and negative sentiment. Figure 4.8 shows the heat-map illustrating the correlation coefficients between selected political features and negative sentiment for 2017 Punjab State Legislative Assembly Election.



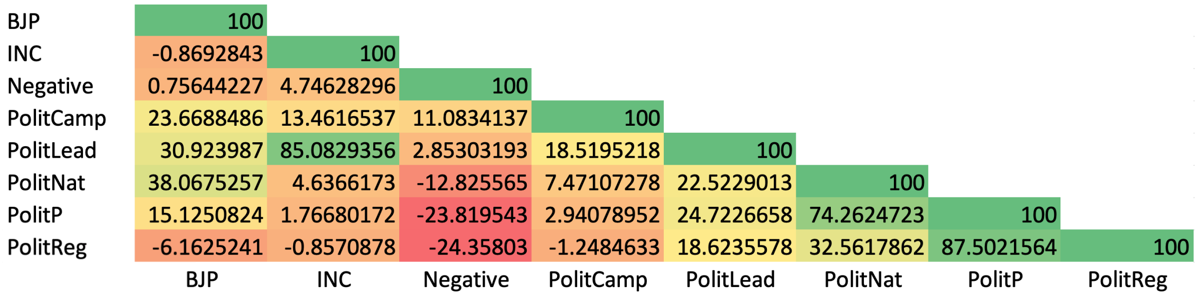
	BJP	INC	Negative	PolitCamp	PolitLead	PolitNat	PolitP	PolitReg
BJP	100							
INC	-1.6768307	100						
Negative	-6.5732301	-10.946341	100					
PolitCamp	38.681434	8.91719598	2.52553615	100				
PolitLead	28.2442459	83.8589019	-11.971248	14.2312447	100			
PolitNat	38.0062942	-5.6069134	7.9919645	29.0027833	7.26109007	100		
PolitP	3.3591769	6.09673149	4.75591802	19.7922604	24.4427446	59.0315719	100	
PolitReg	-24.527443	11.7832902	-0.1420811	2.62170639	24.7926187	-2.3688034	79.296253	100

Figure 4.8: Heat-map presenting the correlation (multiplied by 100) between selected political features and negative sentiment for 2017 Punjab State Legislative Assembly Election

Leading up to the 2018 PB State Assembly elections, the following observations are made. BJP is less likely to be mentioned in articles containing all political parties or regional political parties in comparison to INC with correlation equal to 3.35% and -24.78% for BJP and 6.09% and 11.78% for INC respectively. This asserts INC regional dominance as the more prominent out of the two, whereas BJP had a considerably greater correlation with articles containing the mention of National Parties at 38.00% in contrast to -5.6% for INC, suggesting it more popular in context to the national sentiment.

Another striking observation was the significant higher correlation between INC and political leaders with a correlation of 83.85% much larger than that of BJP at 28.24%. Such disparity clearly indicated the prevalence of the INC political leaders which are much more probable to be discussed in articles in contrast to that of BJP. True to its strength in campaigning and with similar results to that observed in case study 1, BJP lead the way in higher positive correlation at 38.38% for political campaigns and slogan in contrast to INC at 8.91%. INC also showed a lower negative correlation of -10.94% with the negative sentiment expressed during the state election as compared to -6.57% for INC.

Figure 4.9 shows the heat-map illustrating the correlation coefficients between selected political features and negative sentiment for parliamentary seats contested in the 2019 General Election from Punjab Region.



	BJP	INC	Negative	PolitCamp	PolitLead	PolitNat	PolitP	PolitReg
BJP	100							
INC	-0.8692843	100						
Negative	0.75644227	4.74628296	100					
PolitCamp	23.6688486	13.4616537	11.0834137	100				
PolitLead	30.923987	85.0829356	2.85303193	18.5195218	100			
PolitNat	38.0675257	4.6366173	-12.825565	7.47107278	22.5229013	100		
PolitP	15.1250824	1.76680172	-23.819543	2.94078952	24.7226658	74.2624723	100	
PolitReg	-6.1625241	-0.8570878	-24.35803	-1.2484633	18.6235578	32.5617862	87.5021564	100

Figure 4.9: Heat-map presenting the correlation (multiplied by 100) between selected political features and negative sentiment for 2019 General Election from Punjab Region

Following another pattern from that of Chhattisgarh in case study 1 (4.3.1), Although articles during 2019 had higher mention of political feature INC as a whole then BJP's (recall mean value from table 4.13), INC is much less likely to be mentioned in articles which talked about all political parties and national parties in comparison to BJP with correlation equal to 1.7% and 4.63% for INC and, 15.12% and 38.06% for BJP. Similar to the 2017 State election, articles had much greater mentions of Political Leader from INC with a positive correlation of 83.22% for INC and 30.96% for BJP. This again insinuates the notably higher popularity of INC leaders in contrast with BJP. Interestingly, Both INC and BJP observed a rise in the negative sentiment direct to them with 4.74% for INC and 0.75% for BJP compared to state elections. This indicates a rise in the overall negative sentiment communicated in the news articles.

Table 4.14 shows the overview of comparison between the two elections with rows featuring INC presented with the color green and BJP with saffron.

Correlation	2017 Punjab State Assembly Election	2019 General Election for Punjab Region
INC-PolitP	6.09%	1.7%
BJP-PolitP	3.35%	15.12%
INC-PolitNat	-5.6%	4.63%
BJP-PolitNat	38.00%	38.06%
INC-PolitReg	11.78%	-0.85%
BJP-PolitReg	-24.78%	-6.16%
INC-PolitLead	83.85%	85.08%
BJP-PolitLead	28.24%	30.92%
INC-PolitCamp	8.91%	13.46%
BJP-PolitCamp	38.38%	23.67%
INC-Negative	-10.94%	4.74%
BJP-Negative	-6.57%	0.75%

Table 4.14: Summary of correlations comparison between features for 2018 State Election and 2019 General Election from Punjab

4.3.2.C Newspaper Polarity

This section discusses the affect score calculated by Rocksteady for individual newspaper and presents the results for the coverage given to BJP and INC along with the overall negative and positive sentiment expressed by the dallies during the 2017 State and 2019 general elections in Punjab.

Table 4.15: Affect score for political features as expressed by chosen news paper during 2017 PB State Election

Source	BJP	INC	Negative	Positive
The Times of India (TOI)	0.11	0.33	1.89	2.84
Hindustan Times	0.12	0.32	1.85	3.77
The Hindu	0.17	0.21	2.64	3.32
The Telegraph	0.21	0.36	2.19	3.44
The Indian Express	0.12	0.35	1.97	2.61

Table 4.15 shows that INC has a much higher affect score in all the newspaper in contrast to BJP. This translated to higher coverage and mentions for INC. It also portrays INC s high momentum going into the state elections with all media houses providing significantly higher coverage. All the newspapers again had higher positive undertone in comparison to negative sentiment. The Hindu had the highest negative sentiment attached to its newspapers whereas, Hindustan Times shows the least negative while also expressed highest positive value.

Table 4.16: Affect score for political features as expressed by chosen news paper during 2019 PB General Election

Source	BJP	INC	Negative	Positive
The Times of India (TOI)	0.14	0.36	1.91	3.00
Hindustan Times	0.11	0.39	1.82	3.97
The Hindu	0.26	0.48	2.03	3.44
The Telegraph	0.23	0.30	2.68	3.43
The Indian Express	0.14	0.36	2.03	2.95

Table 4.16 shows that all the newspaper witnessed identical behaviour during the 2019 General Election to that observed in the 2017 State Election. INC surpassed BJP in coverage and mentions in all the newspapers by a notable margin showcasing its dominant presence in the news articles from every publication. Every other attribute also showed an identical trend.

4.3.3 Case Study: Karnataka

The third and final case study deals with an Southern Indian state of Karnataka, which saw BJP, the ruling party of the state as of 2018 State Assembly Election, also winning the majority of Parliamentary seats from the state during the 2019 General Election. The result remained a constant, With BJP winning both the election by a majority and replicating its success of 2018 in 2019.

4.3.3.A Descriptive Statistical Analysis

i) Negative Sentiments

Figure 4.10 shows a proxy effect time series for negative sentiment during Punjab State Election from a period of 1st February 2018 to 31st August 2018 and Figure 4.11 shows the same for General Elections 2019 from 1st March 2019 to 31st August 2019.

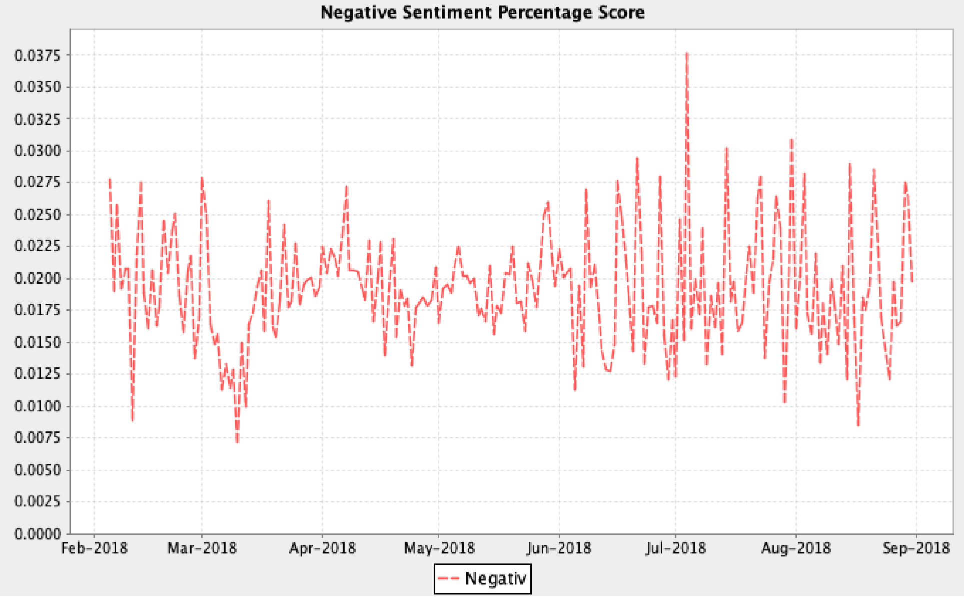


Figure 4.10: KA: Negative Sentiment Affect Time Series (State Elections)

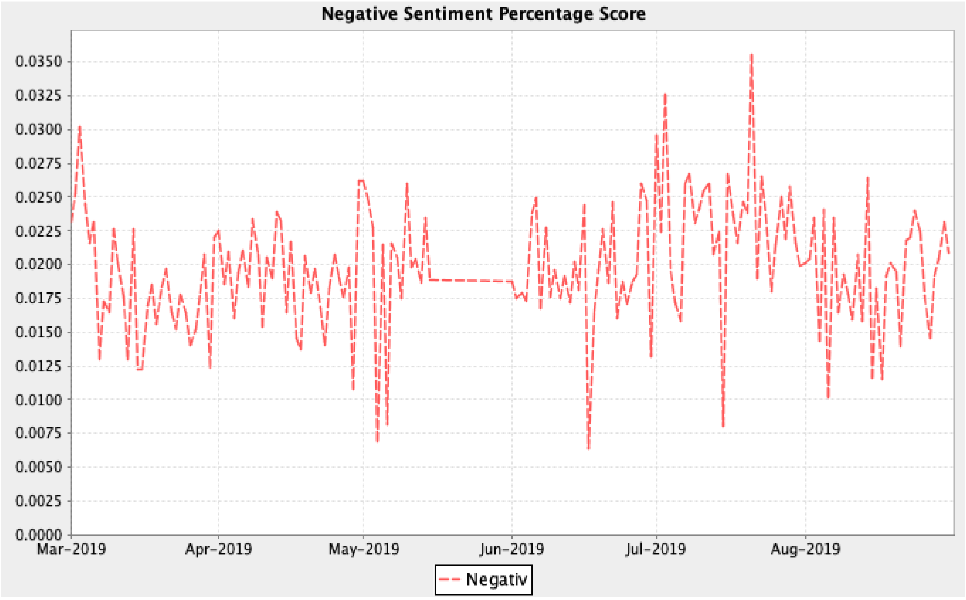


Figure 4.11: KA: Negative Sentiment Affect Time Series (General Elections)

It can be observed in Figure 4.10 that the negative sentiment is almost constant throughout election with whereas in Figure 4.11 it is the same but there is rise in negative sentiment post (May 2019) completion of the election.

Descriptive Statistics	Negative Sentiment during 2018 State Election	Negative Sentiment during 2019 General Election
Minimum	0.71	0.63
Maximum	3.76	3.55
Range	3.05	2.92
Mean	1.91	1.97
Standard Deviation (σ)	0.47	0.46
Kurtosis	0.87	0.88
Skewness	0.48	-0.08

Table 4.17: Descriptive statistics of Negative Sentiment expressed in newspapers from Karnataka

The skewness value of 0.48 and -0.08 and identical almost kurtosis value of 0.87 and 0.88 respectively for 2018 state election and 2019 general elections, indicates that the distribution was symmetrical with a very low degree of outliers for both. Also, standard deviation remained nearly the same with values 0.47 and 0.46. Similar to both the previous case studies, the negative sentiment expressed in the newspaper rose in 2019 with a mean of 1.97 as compared to the 2018 election where the mean is at 1.91.

The descriptive statistical analysis for Positive sentiment shows a higher mean attached to positive sentiment. Figure A1.3 in Appendix A1.3 showcase the comparison, on comparison it is evident that the newspaper articles had a higher positive sentiment attached to them.

ii) Political Features

The Descriptive Statistics for a few of the essential political features which were extracted using Rocksteady for both 2018 Karnataka State Assembly Election listed in Table 4.18, and 2019 General Election in Table 4.19 are presented below.

Table 4.18: Descriptive statistics of Political Features for Karnataka State Election expressed during Feb 2018 to Aug 2019

Series	Min	Max	Range	Mean	(σ)	Kurtosis	Skewness
Number of Articles	1	232	231	17.83	24.19	31.66	4.57
Political Parties	0	3.13	3.13	1.10	0.58	0.72	0.83
Political Leaders	0	1.63	1.63	0.64	0.28	1.02	0.56
National Parties	0	2.11	2.11	0.85	0.42	-0.04	0.43
Regional Parties	0	1.49	1.49	0.24	0.28	4.19	1.99
Political Campaign	0	0.13	0.13	0.01	0.02	10.9	2.85
BJP	0	0.67	067	0.22	0.12	0.86	0.69
INC	0	1.31	1.31	0.32	0.20	2.98	1.10

Table 4.19: Descriptive statistics of Political Features for Karnataka in General Election expressed during Mar 2019 to Aug 2019

Series	Min	Max	Range	Mean	(σ)	Kurtosis	Skewness
Number of Articles	1	43	42	9.98	6.79	3.34	1.54
Political Parties	0.03	3.27	3.24	0.94	0.47	2.71	1.02
Political Leaders	0	2.71	2.71	0.66	0.37	8.99	2.08
National Parties	0.03	3.27	3.24	0.76	0.40	7.28	1.56
Regional Parties	0	1.09	1.09	0.17	0.20	4.75	3.64
Political Campaign	0	0.13	0.13	0.01	0.02	11.04	3.15
BJP	0	2.23	2.23	0.27	0.24	23.56	3.62
INC	0	2.33	2.33	0.29	0.24	29.53	4.04

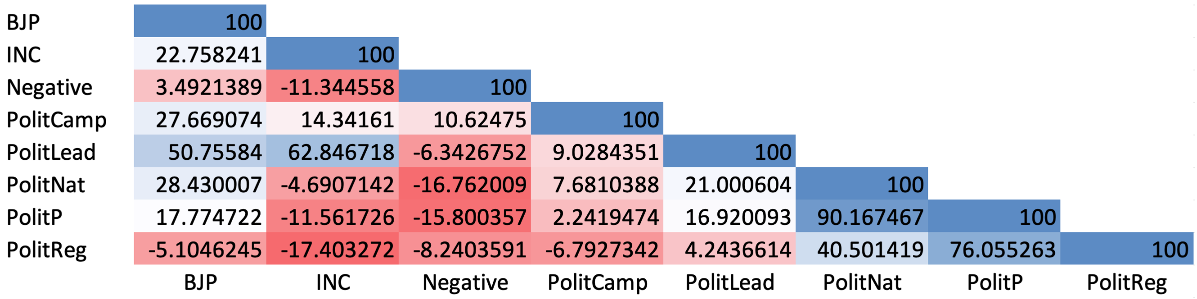
The numbers from both the tables look similar to each other in many of the political features with symmetrical distribution, except for few. With the same trend as that of shown by the previous two case studies, Number of articles shows a much higher kurtosis of 31.66 and 3.34 and skewness of 4.57 1.54 during the 2018 state election and 2019 general elections respectively. As already discovered from the previous two cases, this behaviour is foreseen. During the 2018 State Election INC remained in more spotlight on certain days and events with higher kurtosis of 2.98 and skewness of 1.10 compared to 0.86 and 0.69 for BJP. For GE2019 both parties saw massive jumps in kurtosis values of 23.56 for BJP and 29.56 for BJP, highlighting more featured news coverage and articles covering both, although still mentioning

INC little more.

It was quite interesting to look at the mean value for both BJP and INC during the two elections. During the 2018 state elections, Although BJP won the election, INC received much more media coverage with a higher mean value of 0.32 compared to 0.22 for BJP. This was similar to what we saw in Chhattisgarh in Case-study 1 (4.3.1) where the winning party in the state election received less coverage. Both also had one thing in common, which was that both of them were existing ruling parties from their respective states which received higher coverage, BJP in case of Chhattisgarh and INC in case of Karnataka. The same behaviour followed during the GE2019 in Karnataka where the mean for BJP got higher at 0.22 while decreasing for INC at 0.29. The following results, along with previous two case studies, clearly show a trend with ruling parties influencing or at least affecting the coverage received by a political party during any election, state or general towards themselves.

4.3.3.B Correlation Analysis

This subsection shows results for the correlation analysis made between specific key political features and negative sentiment. Figure 4.12 shows the heat-map illustrating the correlation coefficients between selected political features and negative sentiment for 2018 Karnataka State Legislative Assembly Election.

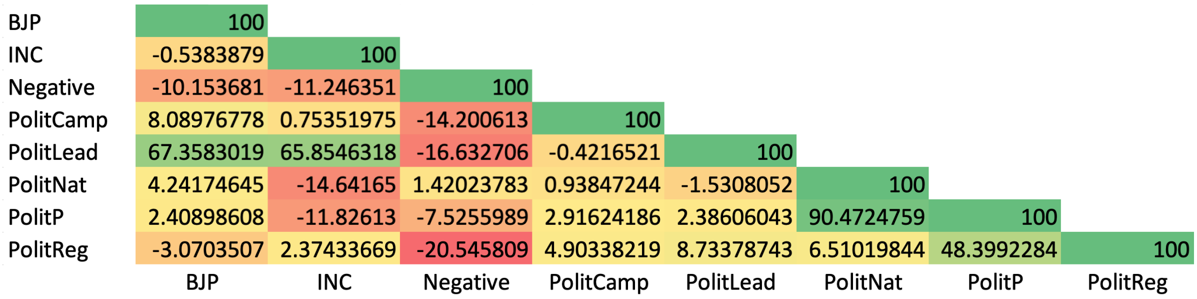


	BJP	INC	Negative	PolitCamp	PolitLead	PolitNat	PolitP	PolitReg
BJP	100							
INC	22.758241	100						
Negative	3.4921389	-11.344558	100					
PolitCamp	27.669074	14.34161	10.62475	100				
PolitLead	50.75584	62.846718	-6.3426752	9.0284351	100			
PolitNat	28.430007	-4.6907142	-16.762009	7.6810388	21.000604	100		
PolitP	17.774722	-11.561726	-15.800357	2.2419474	16.920093	90.167467	100	
PolitReg	-5.1046245	-17.403272	-8.2403591	-6.7927342	4.2436614	40.501419	76.055263	100

Figure 4.12: Heat-map presenting the correlation (multiplied by 100) between selected political features and negative sentiment for 2018 Karnataka State Legislative Assembly Election

Leading up to the 2018 KA State Assembly elections, the following observations are made from the correlation heat-map. INC is significantly less likely to be mentioned in articles containing all political parties, national or regional political parties in comparison to BJP with correlation equal to -11.56%, -4.69% and -17.40% for INC and 17.74%, 28.43% and -5.10% for BJP respectively. Both the parties have a similar correlation with articles containing the names of the political leader at about 50.75% for BJP and 62% for INC, with INC being the higher with slightly higher coverage. Similar to the trend from the previous two case-studies, BJP has a higher correlation with articles containing political campaign slogans and phrases with a correlation of 27.66% compared to 14.31% for INC. Notably, BJP has a greater correlation with negative sentiment at 3.49% whereas INC with a smaller correlation of -11.34%.

Figure 4.13 shows the heat-map illustrating the correlation coefficients between selected political features and negative sentiment for parliamentary seats contested in the 2019 General Election from Karnataka Region.



BJP	100							
INC	-0.5383879	100						
Negative	-10.153681	-11.246351	100					
PolitCamp	8.08976778	0.75351975	-14.200613	100				
PolitLead	67.3583019	65.8546318	-16.632706	-0.4216521	100			
PolitNat	4.24174645	-14.64165	1.42023783	0.93847244	-1.5308052	100		
PolitP	2.40898608	-11.82613	-7.5255989	2.91624186	2.38606043	90.4724759	100	
PolitReg	-3.0703507	2.37433669	-20.545809	4.90338219	8.73378743	6.51019844	48.3992284	100
	BJP	INC	Negative	PolitCamp	PolitLead	PolitNat	PolitP	PolitReg

Figure 4.13: Heat-map presenting the correlation (multiplied by 100) between selected political features and negative sentiment for 2019 General Election from Karnataka Region

During the 2019 General Elections, INC is much less likely to be mentioned in articles which talked about all political parties and national parties in comparison to BJP with correlation equal to 1.7% and 4.64% for INC and, 15.12% and 38.06% for BJP. Interestingly, INC observed a rise in correlation during the GE2019 compared to state election between itself and articles containing the names of the political leader to 85.08% whereas BJP saw a dip in it at 30.92%. Both the parties saw a change of 20% in either direction. Another noticeable difference was fall in BJP correlation with the negative sentiment expressed in the article at 0.75% in 2019 compared to rise for INC at 4.74% which saw 15% increase.

Table 4.20 shows the overview of comparison between the two elections with rows featuring INC presented with the color green and BJP with saffron

Correlation	2018 Karnataka State Assembly Election	2019 General Election for Karnataka Region
INC-PolitP	-11.56%	1.7%
BJP-PolitP	17.74%	15.12%
INC-PolitNat	-4.69%	4.63%
BJP-PolitNat	28.43%	38.06
INC-PolitReg	-17.40%	-0.85%
BJP-PolitReg	-5.10%	-6.16%
INC-PolitLead	62.84%	85.08%
BJP-PolitLead	50.75%	30.92%
INC-PolitCamp	14.31%	13.46%
BJP-PolitCamp	27.66%	23.67%
INC-Negative	-11.34%	4.74%
BJP-Negative	3.49%	0.75%

Table 4.20: Summary of correlations comparison between features for 2018 State Election and 2019 General Election from Karnataka

4.3.3.C Newspaper Polarity

This section discusses the affect score calculated by Rocksteady for individual newspaper and presents the results for the coverage given to BJP and INC along with the overall negative and positive sentiment expressed by the dallies during the 2018 State and 2019 general elections in Karnataka.

Table 4.21: Affect score for political features as expressed by chosen news paper during 2018 KA State Election

Source	BJP	INC	Negative	Positive
The Times of India (TOI)	0.22	0.35	1.88	3.19
Hindustan Times	0.26	0.36	1.76	4.18
The Hindu	0.28	0.30	1.99	3.83
The Telegraph	0.27	0.30	2.39	3.68
The Indian Express	0.31	0.34	2.02	3.38

Table 4.21 shows that INC has a much higher affect score in all the newspaper in contrast to BJP. This translated to higher coverage and mentions for INC. All the newspapers again less negative tone to higher positive sentiment. Hindustan Times shows the least negative sentiment while also expresses highest positive value.

Table 4.22: Affect score for political features as expressed by chosen news paper during 2019 KA General Election

Source	BJP	INC	Negative	Positive
The Times of India (TOI)	0.24	0.23	1.81	3.07
Hindustan Times	0.28	0.35	1.89	3.81
The Hindu	0.30	0.31	2.01	3.21
The Telegraph	0.32	0.37	2.40	3.55
The Indian Express	0.28	0.31	2.09	3.22

Table 4.22 shows the change in coverage after BJP came into power in 2018. INC sees a drop in vales by TOI, Hindustan Times and The Indian Express while BJP saw a rise in coverage by every publication except The Indian Express. The newspaper also had less difference between coverage received by INC and BJP showing a rise in overall coverage for BJP while a decrease for INC. The Top two newspaper, TOI and Hindustan Times again shows moved neutral tone.

5 Conclusion

This final chapter serves as a summary of the work outlined throughout the study. It provides an overview of key finding across all the case studies. Further, it discusses the learning outcomes from this study and the challenges faced across the entirety of this study. Finally, the limitations of the methodology employed by this study and recommendation for future work building on top of this work are outlined.

5.1 Key Findings

Descriptive statistical analysis of negative sentiment extracted from all three case studies shows the rise in mean values of negative sentiment from state election to the general election. This is a consistent behaviour among all three case-studies showing an increase in the overall negative sentiment expressed by the newspaper articles during the 2019 Indian General Election.

All the newspapers have a higher mean value for their positive sentiment comparison to mean values for the negative sentiment (see table A1.2 in appendix A1.5). This was again a consistent behaviour across all the news publications. This trend could be since the general tone expressed during every conversation is considered to be positive which is confirmed by the above results.

Descriptive statistical analysis of political features showed patterns followed across all the case-studies. For all the three states the party which won the state election saw an increase in mentions of their political party in the news during the general election. It could have been due to the political party coming in power and influencing the media with its privileges or, It could also be due to it more in news due to it being the governing party and change to policies and governance would include the concerning political party. Nonetheless, it was a notable figure.

In two out of three case-studies, namely Chhattisgarh where INC won and Karnataka where BJP won during their respective state elections. Both INC in CG and BJP in KA have less overall mentions in newspapers compared to BJP and INC respectively leading up to the state election. It could be due to BJP being the previous ruling party in CG and INC in KA favouring them. As observed from previous results, it has been observed that the ruling party gets more coverage in the newspaper due to their ruling status. Punjab saw no such pattern.

The correlation analysis showcased BJP's strong political campaign around the country. The correlation value between BJP-PolitCamp was always higher in comparison to INC-PolitCamp across all case studies and the two elections. This could be due to the exorbitant amount of money spent by BJP for its political campaign.

Another interesting key finding was that the correlation heatmaps for all three states showed a decrease in negative sentiment during the 2019 GE for the political party which won their respective state election. For example, INC-Negative correlation value decreased from 7.09% in 2018 state election to -3.32% during the 2019 GE. Similar results followed in other states too. This can again be combined with all other previous results be attracted to the changing sentiment for the ruling party after it came to power.

Noticeably the correlation between negative sentiment expressed in the article and BJP was always lower in comparison to INC-Negative for all the case-studies. Irrespective of election results. This could be due to collective national sentiment in favour of BJP.

The overall newspaper polarity calculated by combining all the newspaper by source showcase Hindustan Times with the most positive tone and interestingly the least negative one too (see tableA1.1)

5.2 Learning Outcome

This research is conducted to assess the impact of political sentiment expressed in the multiple print media during the State Assembly Election using three case studies before the 2019 General Election and study the effect of those state elections on the results of General Election from that region using sentiment analysis. Due to no or very little amount of research conducted in the areas, the outcome of the project suggests more work to be done in this domain.

This research has been a great source of knowledge and incredible learning experience, both socially and technically. It helped me expand my horizon from a political standpoint and me to look at the problems of society may it be political, economical or social and helped me change my perspective toward those issues and gain a new one as an Engineer and Computer Scientist and to develop meaningful solutions keeping my personal bias aside. Examining the related work of other people for the literature review allowed me to learn numerous ways to look at problems and their respective approach to the problem. Excellent reports from various sources helped me understand the Indian political socio-space. The research also allowed me to learn more about my own country and Politics associated with it both on a national and regional level.

From a technical point of view, the research allowed me to explore the field of sentiment analysis with great depth and multiple approaches associated with it. This study introduced me to the concept of lexicon-based 'bag-of-words' model using a dictionary-based approach towards sentiment analysis for specialized domains-specific studies. The research allowed me to learn the data curation and management techniques associated with good research practices as well as several accessing third party electronic databases to collect data corpus for specific topics by formulating queries as part of the data retrieval process. Python programming language was also used to preprocess the collected data and to explore the machine learning approach to this problem at the initial stages of this study. From a statistical standpoint, this study allowed me to perform in-depth descriptive statistical analysis and learn several new terms associated with it. The study allowed me to use Microsoft Excel in more ways than I did previously, a powerful tool which was used to perform data analysis as well as a visualization tool to create correlation heatmaps.

5.3 Challenges

Through the course of this research, there were few challenges which presented itself. The first and foremost was Data Acquisition and Retrieval. For the purpose of this research spanning across three different states and two separate elections, a huge amount of data was required. Thousands of newspaper articles were required to be collected for each individual state for a period of 3 more before and after the election month for both state and general election. Greater the size of data corpus better the results. This process was repeated thrice to accomplish this. Data retrieval was the most time-consuming process as LexisNexis servers had a download limit of 100 articles. This restriction meant the data could only be collected 100 at a time and doing it for multiple times was very time-consuming. The limit from 500 to 100 was recently introduced after they shifted their servers to amazon web services. Data management was also one of the key challenges, to manage all the data collected and organize it by day, source, monthly and yearly for all three states and different elections proved a challenge too.

The initial phase of the study also prompted me to look for articles from different languages, other than English. Particularly Hindi, which has the highest market share in the print media in India. The challenge associated with using Hindi Articles was the authenticity of the data sources available. For any meaningful result, the authenticity of the data source is foremost. I could not find any third-party data aggregator such as LexisNexis which would allow me to gather the data form Indian Hindi News Publication which was accessible by me. Other retrieval methods such using API was also not possible as most of the publication didn't employ one. Web-scraping was discarded due to legality and ethicality associated with this technique. Thus, Hindi or any other regional newspaper articles were not included for this study.

The unfortunate event which we all are witnessing currently due to the COVID-19 pandemic, prompting the closure of colleges initially and then the entire country also proved to be unexpected additional challenges to this research which were not expected by any means. It was successfully overcome by the support of my supervisor and all the faculty which contributed to a seamless online transition of several resources. The computation resources required to run Rocksteady was one them, without the high-specs, high-performing PC's from the college, I was forced to run the program in a relatively modest machine which would often either slow down the computer significantly or crash. The computation time also increased significantly due to the large size of the corpus and the dictionary forced me to run the same process several times requiring much greater time to complete the same process. The learning curve associated with this study was also one of the challenges. I lacked the knowledge required for sentiment and statistical analysis for this study and I was relatively novice to both these concepts and it required the help of my supervisor to learn a few key fundamental concepts essential for this kind of study.

5.4 Future Work

This research has plenty of scope for improvement and room for extension. The case studies could be extended to accommodate multiple states for each of them. This would allow for better comparison between each case study and provide more information and better analysis. Another extension for this could be to look at the 2014 general election and compare the results of state elections held across the same states as this study and see if the results showed any similarities with this one. Another suggested improvement could be the process of data collection. Although the newspaper articles were collected using the query "election" AND "name of the state" which provided good sample collection. It did cover all the news articles related to politics from that particular region. The query methodology could be improvised to accommodate more keywords specific to a region to extract more articles with rich political commentary. Another approach would be to look for regional English newspapers specific to the selected state providing specific coverage to the chosen state.

The dictionary-based lexicon approach used for sentiment analysis required a specialized domain-specific dictionary. Although the dictionary used for this study was detailed and exhaustive. For this study, different dictionaries could be constructed mapping to each case study and their respective states including more local leaders and details. This would provide a much better affect score. Another approach would be to try and employ a machine learning approach for the same problem and see if the results are similar. This would help understand the difference between the two sentiment analysis approaches.

This study could be expanded to include financial data as a proxy for the sentiment. It would be by looking up for financial metrics that would be impacted by state-level election and can be collected from an authentic source. This is particularly difficult to find and if possible would serve as an excellent measure of people's sentiment. As English is by no means the most spoken or most read the language in India. As previously mentioned Hindi and other regional languages occupy greater share in the print media industry than English in India. As of 2019 Q1, Therefore only 8% of total newspaper readers read English newspapers (exclusive digital lookups) whereas 44% of where Hindi and rest regional. In future, the articles could be collected for Hindi newspapers to get much better and more comprehensive views of the sentiment expressed in the Media and its impact of the opinion and sentiment expressed by a much bigger population.

5.5 Conclusion

This aim of this research was to examine the newspaper generated by Indian print media for months before leading up to the State Legislative Assembly Election and 2019 General Election held in three states, Chhattisgarh, Punjab and Karnataka which served as the three case studies. We tried to analyze the sentiment expressed in the newspapers and see if would extrapolate the results of the state election by extracting the negative sentiment from the newspaper during the state election and use it as a proxy for electoral sentiments and analyze the impact it had on the 2019 General Election. The idea was to create a framework to analyse dynamic events with a large stream of news to extract the sentiment and opinion of the general public as expressed in the English newspaper and predict the outcome of the General Election. The study provides us with few key observations and it identified patterns that would point at the change in electoral sentiment after a state election and how it translated to the general election.

Numerous factors influence election results, the complexity of election requires more evidence to make a prediction which is a notoriously difficult exercise. It was observed that winners of state elections were able to reduce the negative sentiment against them by the general election. It was also observed that the ruling party in the state after winning the state election had much more coverage than the predecessor. This lower negative sentiment and higher media coverage does not necessarily translate to victory every time during the general election for the given party but gives an idea about the opinion and bias news publication holds for the government in power and the way its shape elector perception. The results from two states Chhattisgarh and Karnataka very similar and analysis of both would allow us to make reasonable predictions with greater accuracy, however, in the case study involving Punjab, the descriptive statistics would probably suggest a victory for BJP which was not the case.

The Indian voters have historically distinguished between state-level and national-level issues with parties cloning on different issues and different focuses. This is also confirmed by this study which shows BJP being favoured during the general election irrespective of winner pointing towards collective national sentiment and bias towards BJP at the central level. The state election results also have disproportionately favoured state parties and smaller political movements specific to the state, who are more likely to be ignored and crowded out in the general elections when fewer seats are up for contest. Finally, I would like to conclude to saying that although state election sways the sentiment communicated in the newspaper which in turn might change and shape the electoral perception, with the given data and results observed for the score of this study it is not reasonable to extrapolate the results of state election and predict the results of the national-level General Election in India.

Bibliography

[1] Zeenab Aneez, Rasmus Nielsen, Antonis Kalogeropoulos, and Taberez Neyazi. Reuters institute india digital news report, 03 2019.

[2] Indian readership survey q1 2019. URL https://https://mruc.net/uploads/posts/8e428e54a95edcd6e8be593a7021a185.pdf.

[3] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[4] Science X staff. Researchers find that the brain can assign value to an object in less than a tenth of a second, Feb 2018. URL https://medicalxpress.com/news/2018-02-brain-assign-tenth.html.

[5] Sylvie Graf, Pavla Linhartova, and Sabine Sczesny. The effects of news report valence and linguistic labels on prejudice against social minorities. *Media Psychology*, 23(2): 215–243, 2019. doi: 10.1080/15213269.2019.1584571.

[6] Jakob-Moritz Eberl, Markus Wagner, and Hajo G. Boomgaarden. Are perceptions of candidate traits shaped by the media? the effects of three types of media bias. *The International Journal of Press/Politics*, 22(1):111–132, 2016. doi: 10.1177/1940161216674651.

[7] James N. Druckman and Michael Parkin. The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, 67(4):1030–1049, 2005. doi: 10.1111/j.1468-2508.2005.00349.x.

[8] Adnan Qayyum, Zafar Gilani, Siddique Latif, and Junaid Qadir. Exploring media bias and toxicity in south asian political discourse. *2018 12th International Conference on Open Source Systems and Technologies (ICOSST)*, 2018. doi: 10.1109/icosst.2018.8632183.

[9] Nabanita Roy. *Media, Political Sentiment and its Impact on Emerging Market: A Case-Study of India's General Election of 2014*. PhD thesis, 2018.

[10] Jasmmet Singh. *POLITICAL SENTIMENT ANALYSIS : IMPACT OF PRINT MEDIA POLITICAL SENTIMENT ON STOCK MARKET DYNAMICS*. PhD thesis, 2019.

[11] Language wise certified circulation figures for the audit period january june 2019. URL http://www.auditbureau.org/.

[12] Dhiraj Murthy and Laura R. Petto. Comparing print coverage and tweets in elections: A case study of the 2011–2012 u.s. republican primaries. *Social Science Computer Review*, 33(3):298–314, 2015. doi: 10.1177/0894439314541925. URL https://doi.org/10.1177/0894439314541925.

[13] G. Kavitha, B. Saveen, and N. Imtiaz. Discovering public opinions by performing sentimental analysis on real time twitter data. In *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, pages 1–4, 2018.

[14] Nima Dokoohaki, Filippia Zikou, Daniel Gillblad, and Mihhail Matskin. Predicting swedish elections with twitter. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM 15*, 2015. doi: 10.1145/2808797.2808915.

[15] Adam Bermingham and Alan Smeaton. On using twitter to monitor political sentiment and predict election results. *Proceedings of the Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP 2011)*, 13, 05 2012.

[16] Luca Buccoliero, Elena Bellio, Giulia Crestini, and Alessandra Arkoudas. Twitter and politics: Evidence from the us presidential elections 2016. *Journal of Marketing Communications*, 26:1–27, 08 2018. doi: 10.1080/13527266.2018.1504228.

[17] B. R. Naiknaware and S. S. Kawathekar. Prediction of 2019 indian election using sentiment analysis. In *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on*, pages 660–665, 2018.

[18] Shelly Ghai Bajaj. The use of twitter during the 2014 indian general elections. *Asian Survey*, 57(2):249–270, 2017. doi: 10.1525/as.2017.57.2.249.

[19] Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. volume 10, 01 2010.

[20] Kokil Jaidka, Saifuddin Ahmed, Marko Skoric, and Martin Hilbert. Predicting elections from social media: a three-country, three-method comparative study. *Asian Journal of Communication*, 29(3):252–273, 2018. doi: 10.1080/01292986.2018.1453849.

[21] Nugroho Prasetyo and Claudia Hauff. Twitter-based election prediction in the developing world. pages 149–158, 08 2015. doi: 10.1145/2700171.2791033.

[22] Kokil Jaidka and Saifuddin Ahmed. The 2014 indian general election on twitter. *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development - ICTD 15*, 2015. doi: 10.1145/2737856.2737889.

[23] Worldwide mobile data pricing league: Cost of 1gb in 230 countries. URL https://www.cable.co.uk/mobiles/worldwide-data-pricing/.

[24] Statista. Topic: Internet usage in india. URL https://www.statista.com/topics/2157/internet-usage-in-india/.

[25] Audit bureau of circulations report shows there is no stopping regional press as print soars, May 2017. URL https://indianexpress.com/article/explained/no-stopping-regional-press-as-print-soars-4648582/.

[26] Stephen Kelly and Khurshid Ahmad. *News, Sentiment, and Financial Markets: A Computational System to Evaluate the Influence of Text Sentiment on Financial Assets*. PhD thesis, October 2016.

[27] Vlad Krotov and Leiser Silva. Legality and ethics of web scraping. 09 2018.

[28] Zeyan Zhao and Khurshid Ahmad. Qualitative and quantitative sentiment proxies: Interaction between markets. *Intelligent Data Engineering and Automated Learning – IDEAL 2015 Lecture Notes in Computer Science*, page 466–474, 2015. doi: 10.1007/978-3-319-24834-9_54.

[29] Zeyan Zhao. A computational account of investor behaviour in chinese and us market. *International Journal of Economic Behavior and Organization*, 3(6):78, 2015. doi: 10.11648/j.ijebo.20150306.11.

[30] Jakob-Moritz Eberl, Markus Wagner, and Hajo G Boomgaarden. Are perceptions of candidate traits shaped by the media? the effects of three types of media bias. *The International Journal of Press/Politics*, 22(1):111–132, 2017.

[31] Christine Ma-Kellams and Jennifer Lerner. Trust your gut or think carefully? examining whether an intuitive, versus a systematic, mode of thought produces greater empathic accuracy. *Journal of personality and social psychology*, 111(5):674, 2016.

[32] Jennifer S Lerner and Dacher Keltner. Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & emotion*, 14(4):473–493, 2000.

[33] Shanto Iyengar. *Is anyone responsible?: How television frames political issues*. University of Chicago Press, 1994.

[34] Lara Logan. Political bias is destroying people's faith in journalism, Feb 2019. URL https://nypost.com/2019/02/26/political-bias-is-destroying-peoples-faith-in-journalism/.

[35] Roy Greenslade. Election 2010: What influence do newspapers have over voters?, May 2010. URL https://www.theguardian.com/media/2010/may/03/election-2010-newspapers-influence-over-voters.

[36] Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

[37] Political monitor archive. URL https://www.ipsos.com/ipsos-mori/en-uk/political-monitor-archive?oItemId=2476&view=wide.

[38] Heinz Brandenburg. Party strategy and media bias: A quantitative analysis of the 2005 uk election campaign. *Journal of elections, public opinion and parties*, 16(2):157–178, 2006.

[39] Heinz Brandenburg. Political bias in the irish media: A quantitative study of campaign coverage during the 2002 general election. *Irish Political Studies*, 20(3):297–322, 2005.

[40] Khurshid Ahmad, Nicholas Daly, and Vanessa Liston. What is new? news media, general elections, sentiment, and named entities. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 80–88, 2011.

[41] S. Padmaja, S. S. Fatima, S. Bandu, P. Kosala, and M. C. Abhignya. Comparing and evaluating the sentiment on newspaper articles: A preliminary experiment. In *2014 Science and Information Conference*, pages 789–792, 2014.

[42] Francis Barclay, Chinnaswamy Pichandy, and A. Venkat. Indian elections, 2014: Political orientation of english newspapers. *Asia Pacific Media Educator*, 24:7–22, 06 2014. doi: 10.1177/1326365X14539215.

[43] Angus Stevenson and Christine A. Lindberg. sentiment analysis, 2011. URL https://www.oxfordreference.com/view/10.1093/acref/9780195392883.001.0001/m_en_us1445769.

[44] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 412–418, 2004.

[45] Aidan Finn, Nicholas Kushmerick, and Barry Smyth. Genre classification and domain transfer for information filtering. 02 2002. doi: 10.1007/3-540-45886-7_23.

[46] Anthony Aue and Michael Gamon. Customizing sentiment classifiers to new domains: A case study. 01 2005.

[47] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[48] Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. Adapting naive bayes to domain adaptation for sentiment analysis. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, editors, *Advances in Information Retrieval*, pages 337–349, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[49] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.

[50] Kamal Nigam, John Lafferty, and Andrew McCallum. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67. Stockholom, Sweden, 1999.

[51] Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):23, Jul 2016. ISSN 2193-1127. doi: 10.1140/epjds/s13688-016-0085-1. URL https://doi.org/10.1140/epjds/s13688-016-0085-1.

[52] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011. doi: 10.1162/COLI_a_00049. URL https://doi.org/10.1162/COLI_a_00049.

[53] Rahul Rajput and Arun Kumar Solanki. Review of sentimental analysis methods using lexicon based approach. *IJCSMC*, 5(2):159–166, 2016.

[54] Aditya Bhardwaj, Yogendra Narayan, Maitreyee Dutta, et al. Sentiment analysis for indian stock market prediction using sensex and nifty. *Procedia Computer Science*, 70: 85–91, 2015.

[55] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd international conference on computational linguistics*, pages 806–814. Association for Computational Linguistics, 2010.

[56] Antonio Moreno-Ortiz and Javier Fernández-Cruz. Identifying polarity in financial texts for sentiment analysis: a corpus-based approach. *Procedia-Social and Behavioral Sciences*, 198:330–338, 2015.

[57] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[58] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer: A computer approach to content analysis. 1966.

[59] Andrea Esuli and Fabrizio Sebastiani. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2006/pdf/384_pdf.pdf`.

[60] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[61] Khurshid Ahmad, JingGuang Han, Elaine Hutson, Colm Kearney, and Sha Liu. Media-expressed negative tone and firm-level stock returns, Dec 2015. URL `https://www.sciencedirect.com/science/article/pii/S0929119915001637`.

[62] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. volume 10, 01 2010.

[63] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29 (1):24–54, 2010.

[64] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[65] M Bastian, S Heymann, and M Jacomy. International aaai conference on weblogs and social media. *North America*, 2009.

[66] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, Mar 2011. ISSN 1877-7503. doi: 10.1016/j.jocs.2010.12.007. URL `http://dx.doi.org/10.1016/j.jocs.2010.12.007`.

[67] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[68] CD Manning, M Surdeanu, J Bauer, JR Finkel, S Bethard, and D McClosky. Acl (system demonstrations). 2014.

[69] Pollyanna Gonçalves, Fabrício Benevenuto, and Meeyoung Cha. Panas-t: A pychometric scale for measuring sentiments on twitter, 2013.

[70] Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs, 2011.

[71] IE) Zemánková Andrea (Dublin IE) Ahmad, Khurshid (Dublin. Methods and system for calculating affect scores in one or more documents, September 2014. URL http://www.freepatentsonline.com/y2014/0278375.html.

[72] what is a corpus, Sep 2016. URL https://courses.helsinki.fi/sites/default/files/course-material/4433684/070916part2.pdf.

[73] office of registrar of newspapers for india, url=http://rni.nic.in/, journal=OFFICE OF REGISTRAR OF NEWSPAPERS FOR INDIA.

[74] Michelle Knight. What is data curation?, Sep 2019. URL https://www.dataversity.net/what-is-data-curation/.

[75] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *SSRN Electronic Journal*, 2005. doi: 10.2139/ssrn.685145.

[76] Olga Kolchyna, Thársis Souza, Philip Treleaven, and Tomaso Aste. *Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination*. 07 2015.

[77] Jason Brownlee. A gentle introduction to the bag-of-words model, Aug 2019. URL https://machinelearningmastery.com/gentle-introduction-bag-words-model/.

[78] URL https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

[79] Vector space models. URL https://www.sciencedirect.com/topics/computer-science/vector-space-models.

[80] idf :: A single-page tutorial - information retrieval and text mining. URL http://www.tfidf.com/.

[81] Online access. URL https://wiki.dbpedia.org/OnlineAccess.

[82] Sarang Narkhede. Understanding descriptive statistics, May 2019. URL https://towardsdatascience.com/understanding-descriptive-statistics-c9c2b0641291.

[83] Akhilesh Ganti. Correlation coefficient definition, Feb 2020. URL https://www.investopedia.com/terms/c/correlationcoefficient.asp.

[84] Adam Hayes. Understanding positive correlation, Feb 2020. URL https://www.investopedia.com/terms/p/positive-correlation.asp.

[85] Elvis Picardo. Negative correlation definition, Jan 2020. URL https://www.investopedia.com/terms/n/negative-correlation.asp.

[86] Correlation analysis. URL https://www.sciencedirect.com/topics/medicine-and-dentistry/correlation-analysis.

[87] Correlation does not imply causation, Apr 2020. URL https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation.

[88] Investopedia. How should i interpret a negative correlation?, Jan 2020. URL https://www.investopedia.com/ask/answers/040815/how-should-i-interpret-negative-correlation.asp.

[89] Nathan Green. Correlation is not causation | nathan green's s word, Jan 2012. URL https://www.theguardian.com/science/blog/2012/jan/06/correlation-causation.

[90] Seema Singh. Why correlation does not imply causation?, Jan 2019. URL https://towardsdatascience.com/why-correlation-does-not-imply-causation-5b99790df07e.

[91] The Editors of Encyclopaedia Britannica. Microsoft excel, Jan 2019. URL https://www.britannica.com/technology/Microsoft-Excel.

[92] Lexisnexis, Apr 2020. URL https://en.wikipedia.org/wiki/LexisNexis.

[93] Indian readership survey. URL https://mruc.net/studies.

A1 Appendix

A1.1 Basic Statistical techniques

Mean/Average: Mean or Average is a central tendency of the data i.e. a number around which a whole data is spread out. In a way, it is a single number which can estimate the value of whole data set.

Median: Median is the middle number in a sorted list of numbers. It is the value which divides the data in 2 equal parts i.e. number of terms on right side of it is same as number of terms on left side of it when data is arranged in either ascending or descending order (cite tds). The median can be used to determine an approximate average, or mean. The median is sometimes used as opposed to the mean when there are outliers in the sequence that might skew the average of the values. The median of a sequence can be less affected by outliers than the mean (cite investopedia)

Mode: Mode is the term appearing maximum time in data set i.e. term that has highest frequency (cite tds).A set of numbers may have one mode, more than one mode, or no mode at all.In statistics, the mode is the most commonly observed value in a set of data (cite investo).

Standard Deviation The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation. A low standard deviation indicates that the data points tend to be close to the mean of the data set (Cite) The Formula for Standard Deviation

Variance: Variance (σ^2) in statistics is a measurement of the spread between numbers in a data set. That is, it measures how far each number in the set is from the mean and therefore from every other number in the set.Variance is calculated by taking the differences between each number in the data set and the mean, then squaring the differences to make them positive, and finally dividing the sum of the squares by the number of values in the data set.

A1.2 Custom Rocksteady template

```
                             {}
                              All Rights Reserved
                               {} of {} DOCUMENTS
                                   {}
                               {}
{}
BYLINE: {}
SECTION: {}
LENGTH: {}
{}
LOAD-DATE: {}
LANGUAGE: {}
PUBLICATION-TYPE: {}
```

A1.3 Statistical Analysis

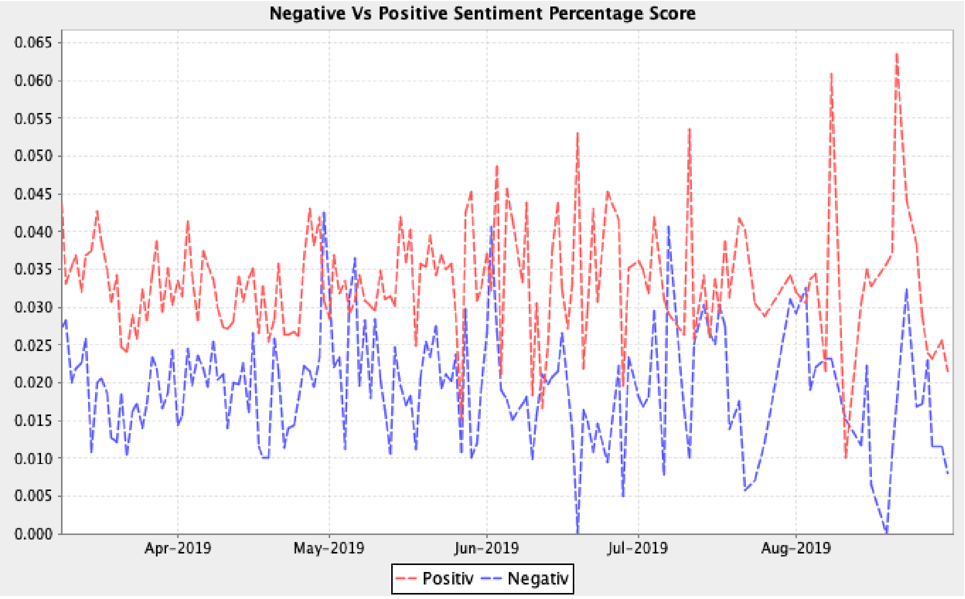


Figure A1.1: Negative vs Positive Sentiment Affect time series for CG

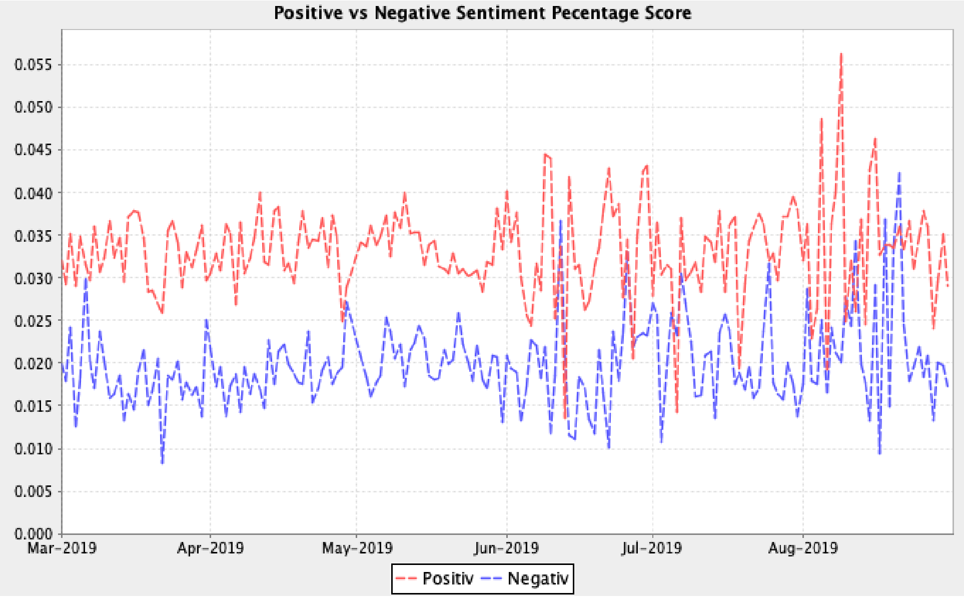


Figure A1.2: Negative vs Positive Sentiment Affect time series for PB

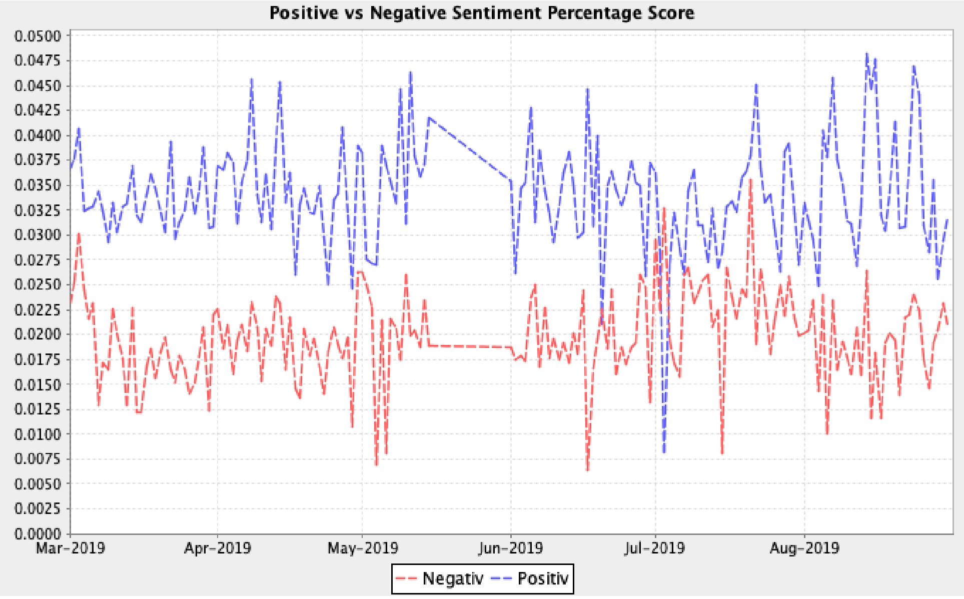


Figure A1.3: Negative vs Positive Sentiment Affect time series for KA

A1.4 Domain Specific Dictionary

Table A1.1: Domain Specific Dictionary and its features

Features	Description
Entry	**Terms specific to each category**
Exclude	Excluded terms from the General English Dictionary which did not contain affect words related to this study
AITC, BSP, BJP , CPI, CPI(M), INC, AAP, NCP, SAD	Mapping of terms with are related to the following Indian Political parties: All Indian Trinamool Congress, Bahujan Samaj Party, Bharatiya Janata Party, Communist Party of India, Communist Party of India (Marxist), Indian National Congress, Aam Aadmi Party and National Congress Party, Shiromani Akali Dal.
PolitP , PolitWing, PolitLead	Key words associated with Political Parties: Mapping of terms which are related to Political Party, Political Wing, Political Leader.
PolitNat, PolitReg, RNorth, RSouth, REast, RNorthEast, RCentral, RWest	Key words associated with regional politics: Mapping of Political Parties and Leaders which are National Party, Regional Party belonging to North Region, South Region, East Region, North-East region, Central Region and West Region (arranged in in order from top to bottom)
PolitCamp, Religion, Tech, Edu, Health, Religion, Develop	Keywords associated with Political Campaigning: Mapping of terms related to Religion, Technology, Education Health and Development during the election campaign.

A1.5 Overall Newspaper polarity

Table A1.2: Newspaper polarity calculated for all combined newspaper

Source	BJP	INC	Negative	Positive
The Times of India (TOI)	0.25	0.33	1.89	3.05
Hindustan Times	0.23	0.34	1.83	3.89
The Hindu	0.30	0.29	1.99	3.10
The Telegraph	0.28	0.32	2.39	3.55
Indian Express	0.24	0.34	2.07	3.13