CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Sentiment Analysis of News Headlines for Comparing Performance of Governments of Pakistan

by

Saba Nawaz

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing

Department of Computer Science

2021

Copyright © 2021 by Saba Nawaz

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.

My work is devoted to My Parents, My Teachers, My Family, and My Friends. I have a special feeling of gratitude for My Parents and siblings. Special thanks to my supervisor whose support make me able to reach this milestone.

CERTIFICATE OF APPROVAL

Sentiment Analysis of News Headlines for Comparing Performance of Governments of Pakistan

by

Saba Nawaz

(MCS191013)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Mussarat Yasmin	COMSATS, Islamabad
(b)	Internal Examiner	Dr. M. Shahid Iqbal Malik	CUST, Islamabad
(c)	Supervisor	Dr. Nayyer Masood	CUST, Islamabad

Dr. Nayyer Masood
Thesis Supervisor
December, 2021

Dr. Nayyer Masood
Head
Dept. of Computer Science
December, 2021

Dr. M. Abdul Qadir
Dean
Faculty of Computing
December, 2021

Author's Declaration

I, **Saba Nawaz** hereby state that my MS thesis titled "**Sentiment Analysis of News Headlines for Comparing Performance of Governments of Pakistan**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Saba Nawaz)

Registration No: MCS191013

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Sentiment Analysis of News Headlines for Comparing Performance of Governments of Pakistan**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if i am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Saba Nawaz)
Registration No: MCS191013

Acknowledgement

I praise and worship my Allah who is all in all. He is perfect source of strength in my life. I want to thank my supervisor Dr. Nayyer Masood for this work. I have learned a lot from him. He motivated me to improve my work and helped me to understand the potential practical implications of the thesis. His insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. I want to show gratitude to my family including parents and siblings for their support and love. I would not be able to achieve anything without my family. I am thankful to my teachers. A debt of gratitude is also owed to Sir Ghulam Mustafa and Sir Omaid Ghayyur both has guided and helped me in a very good way.

(Saba Nawaz)

Abstract

A lot of unstructured textual data is available on the internet such as digital libraries, and news sources, and this data is mounting up every passing day. Availability of data in such abundance brings about new challenges with it. One of the challenges is to access, organize and summarize the relevant data to a task in an effective way. This helps in enhancing the knowledge, in the analysis of things or in building an opinion over an issue. News nowadays play an important role in shaping people's perceptions and opinions about any issue, such as political, economic, or arts etc. Because of the large amount of opinionated data available on various websites, sentiment analysis has become an important tool to get insight into what people think over an issue. Sentiment analysis is the process of extracting intent of a short message to be classified as one of the predefined classes. The focus of this work is to use sentiment analysis on news headlines to guage the individual and comparative performance of Pakistani governments. For the data collection, two major Pakistani newspapers have been selected, such as The Dawn and The Nation, and the data collected relates to the regimes of three political parties, that is, PPP, PMLN, PTI. In the literature, researchers have proposed text representation approaches that do not consider the context of terms. For textual representation, this study employs BERT, which captures the semantics and context of terms. Furthermore, the model includes feature reduction. In this case, the most appropriate feature selection can help to increase the overall accuracy of the sentiment analysis task. Experiments are done on two news headlines datasets, and the results of best-performing model BERT are compared to similar approaches proposed in literature i.e.N-gram, Tf-idf, Word2Vec using standard matrices for data analysis. The proposed model BERT outperformed the approaches that had previously been proposed with increase of 13% accuracy on Pakistani dataset and increase of 3% accuracy on base paper dataset. A keyword based approach is also presented to evaluate the performance of governements on the basis of provided keyword, where a user can type a keyword and it retrieves those headlines containing keywords from dataset of three regimes, and find its sentiment score. It calculates their impacts and ranks accordingly.

Contents

List of Figures

List of Tables

Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
NLP	Natural Language Model
SA	Sentiment Analysis
SVM	Support Vector Machine
TF	Term Frequency
TFIDF	Term Frequency Inverse Document Frequency

Chapter 1

Introduction

1.1 Background

People express their opinions in the form of reviews on a variety of topics, which are usually written in text format. People's habits of reading news have altered as a result of technological advancements. The majority of newspapers now publish in both print and electronic formats. The main advantage of ePapers is that they are easy to obtain and deliver up-to-date news if you have access to the internet and a suitable gadget. Today, newspapers and social media play an essential part in shaping people's attitudes on a variety of topics, such as social, political, and religious matters as well as any product. News articles that are published in newspapers can sometime create positive or negative impact on society on a large scale. It is well recognised that a lot of information can be gleaned from news websites but, if the material is properly processed, it may be used for decision-making and political campaigns. This information is extremely beneficial to a variety of organisations as well as political parties. Political concerns are an intriguing use of news data analysis, such as identifying user feelings regarding the current government's strategies or gathering opinions to anticipate the most desirable candidate and the leading political party for future elections. Analyzing news headlines could be valuable in determining the general sentiments of the users. Opinion mining or sentiment analysis is a task to identify human emotions (good, bad, neutral) from written text. As per Rameshbhai et, al.[1], Sentiment analysis is the process of classifying the result from text by using Natural Language processing (NLP). According to K C Ravishankar et al [4], The process of examining a piece of text to determine if the suggested expression is negative, positive, or neutral is known as sentiment analysis. Soonh et, al.[2] define sentiment analysis (SA) as the task of evaluating the views, emotions, and subjectivity by computing and is also known as Opinion Mining. Manual labeling of words that have sentiment is a time consuming process. In order to automate the process of sentiment analysis two main approaches are used: Machine learning and Lexicon based methods [2, 9]. Lexicon based approaches also known as unsupervised approaches it means there is no need to preprocess the data or train a classifier to predict polarity. Lexicon based methods consist of dictionary containing words have some polarity value. Words within documents are matched with dictionary words in order to find polarity. Opposed to Lexicon Based approaches, Machine learning approaches can either be supervised or unsupervised. There is a need of large amount of annotated data to be annotated by annotators in order to train a classifier. Researchers are combining the two approaches by use dictionaries and classifiers to determine sentiment polarity on three levels: Levels of Sentences, Documents, and Aspects [11, 13].

As per Dor [12], one can get idea of whole news content directly by reading only news headlines rather than going through whole article. Therefore, trivial headlines can also influence on large scale. Several machine learning techniques can be found in literature that are utilized to get insight into people's opinion i.e. linear regression, Naive bayes, Support Vector Machine. Researchers used SVM models to classify news content in context of a news dataset [1] and political social media messages [8] but they have some limitations. According to Gupta et, al.[8] there are two main reasons that can negatively affect model performance which are limited amount of training data and second is supervised ML models treat words separately as individual, without taking into account their context. To overcome these issues one can utilize Bert as it is proven to be beneficial for classifying political data due to the large amount of data it is trained on and it also keeps context from both left and right side. Figure 1.2 shows the overall flow of thesis.

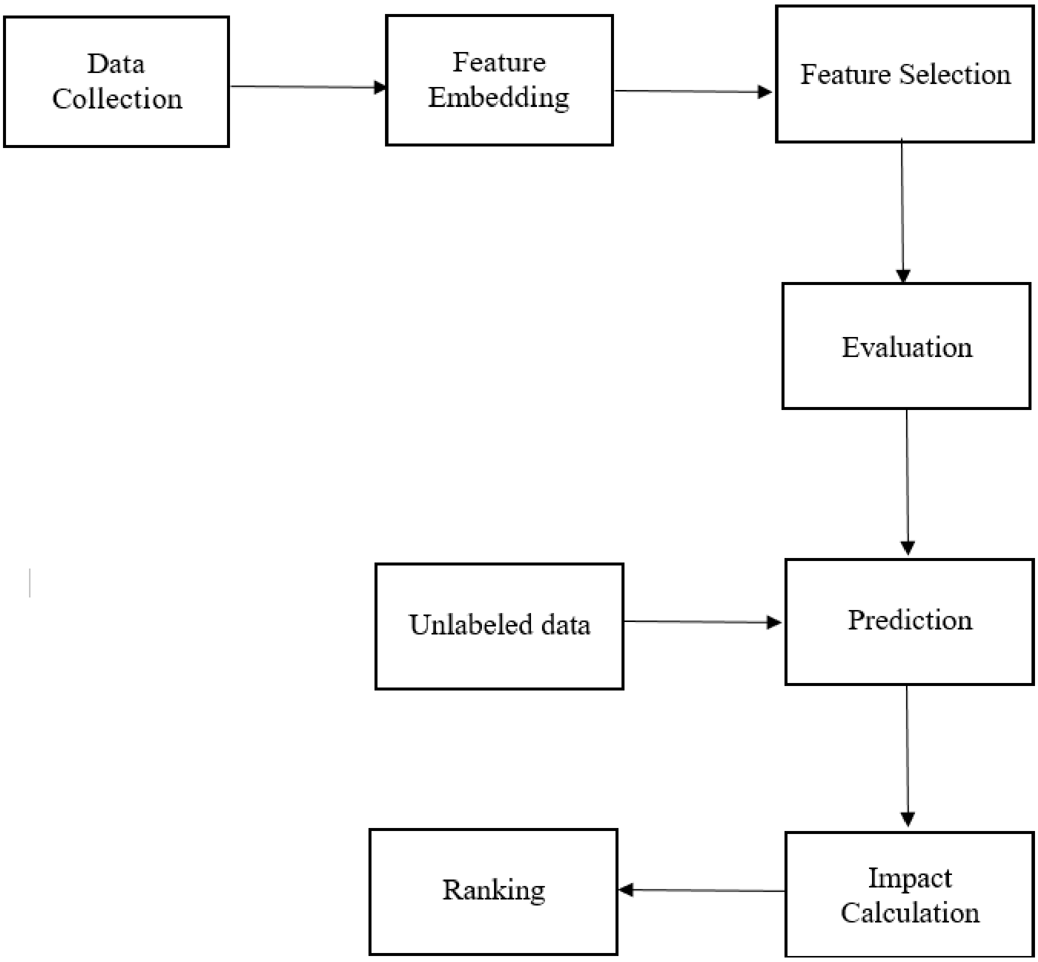


FIGURE 1.1: Overall flow of thesis

1.1.1 Classification

Due to the wide availability of text in a variety of formats, a large amount of unstructured data has been collected by researchers, who have discovered various methodologies in the literature to convert this unstructured data into a defined structured volume, this process is known as text classification [17]. A dataset is classified into one of the predefined classes during the classification process. The categorization works with both structured and unstructured data, predicting the class of a specific record in the dataset. Target, label, and categories are all terms used to describe the classes. Textual information is now available in almost every field for improved classification, text mining research is becoming increasingly popular. Text mining is utilised in a range of industries, such as image processing, health, finance, and many others, with the purpose of extracting useful information from semi-structured or unstructured text using techniques like supervised or unsupervised classification, or Natural Language Processing. There are two phases in supervised approach: training and testing. In training phase the training dataset is used to train the model, which then predicts on the testing dataset in phase two. Then it is easy to determine which class an unknown data will fall into.

1.1.2 Sentiment Analysis

Sentiment analysis also known as opinion mining. The goal of this task is to determine people's attitudes, whether they are positive, negative or neutral, based on written text or any statement such as reviews, headlines, and tweets. Soonh et, al.[2] defines sentiment analysis(SA) as the work of computing-based evaluation of viewpoints, opinions, emotions, and subjectivity.

Sentiment can be classed as either positive or negative in its most basic form, but it can also be expanded to include a wide range of values across multiple dimensions, such as fear, grief, rage, joy, and so on. This task may appear to be highly subjective, as various people perceive the same text in different ways. Although a human reader is the best judge of a text's sentiment, automated sentiment analysis can be effective when there is a large amount of content that human readers cannot handle.

Opinion mining primarily works with positive and negative sentiment rather than specific emotions (e.g., happiness, surprise), opinion mining primarily works with positive and negative sentiment. While some texts, such as those from classic literature, may express sentiment in a complex manner, most text data for social scientific study, such as news and online reviews, do so in very simple ways. As a result, sentiment analysis is frequently employed to examine social media posts and survey results. It also helps with corporate intelligence by swiftly summarizing thousands or millions of customer product reviews, recognizing movie popularity from various internet reviews and determining which elements of a vehicle are liked or hated by owners based on their remarks on a dedicated site or forum. Analyzing these sentiment containing statements could be valuable in determining the general sentiments of the users. According to Tyagi et, al.[20] sentiment analysis approaches are divided into two main categories: Machine learning and Lexicon based methods. Figure 1.2 shows the classification of sentiment analysis approaches. Machine learning approach is further subdivided into supervised learning, unsupervised and semi-supervised learning. Dictionary-based and corpus-based approaches are the two main approaches of lexicon based method. Lexicon based approaches can be useful in identifying sentiment polarity of data by using a dictionary of words labelled by sentiment. However, this approach has two drawbacks: first, those dictionaries do not contain all words in English or other languages, and second, it does not handle negation of feeling adequately. As an alternative, researchers have utilized Machine learning based approach. Machine learning approaches work on train test data. Using a sentiment-labeled training set, it train a machine learning model to recognize sentiment based on the words and their sequence. This is accomplished by extracting ”features” from the text, which are then used to predict a ”label.” Splitting the text into words and then using these words and their frequency in the text as features is an example of producing features.

This method is highly dependent on the text representation, algorithm and the quality of the training data. The most basic and widely used approaches to sentiment analysis are based on unigram(In terms of presence or the number of times they appear) [1, 3, 9] POS tags, term position [33], and syntactic dependencies [21], opinion words and sentences, negations. Support Vector Machine(SVM), Naive Bayes (NB) [7], and Maximum Entropy(ME) classifiers, as well as derived

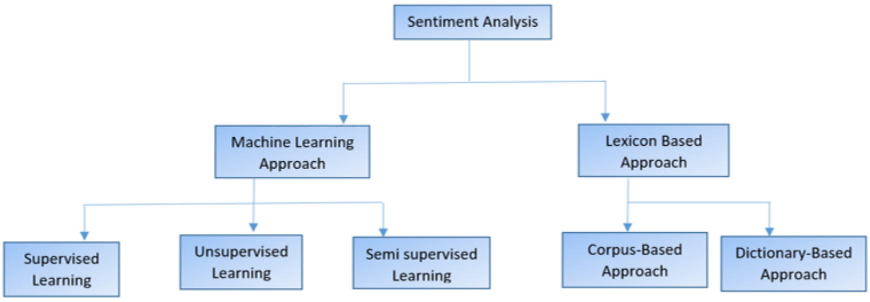


FIGURE 1.2: Classification of sentiment analysis approaches.

ensembles [3, 21, 26] have proven to be successful in text categorization. In comparison to naive Bayes, maximum entropy, and other approaches to solving the problem, support vector machine(SVM) is the most accurate [52]. Fortunately, the field of NLP has begun to evolve quicker and faster as a result of the combination of two primary techniques: word embeddings and deep learning-based models in recent years, and has become increasingly successful. Explained in the next sections.

1.1.3 News Headlines Sentiment Classification

From literature, it is observed that most of the work is performed in the field of text mining for news classification. Long-form news is categorized to a significant extent in the literature, but work on news headlines is minimal. News headlines are written by professionals and are free of typos or an informal text. Feature extraction is a crucial step in developing a precise model. It's crucial to extract features from the dataset since the stronger the feature, the more accurate the model could be. Our primary goal is to conduct an evaluation of news headline classification. Larger news stories are more difficult to classify since full text news categorization takes a lot of time and statistical calculations. Numerous scholars have proposed methods for classifying news headlines, with each new item being assigned to one of the pre-defined class. It is accomplished by identifying the most likely terms within the class.

In the past, Researchers have performed news headline classifications for various tasks such as emotions identification based on news headlines, opinion mining from new headlines, sentiment classification of Malaysia financial news , classification of short texts and classification of news headlines for presenting user-centered e-newspaper. Previous researches focuses on predicting opinion of people for election campaigning purposes by analyzing Facebook posts, twitter tweets. The obtained information after processing is extremely beneficial to a variety of organizations as well as political parties, identifying sentiments of users towards the present government's strategies, or finding opinions to predict the most ideal applicant and the leading political party for future elections. It is important to get insight about the opinions of the public towards any government or political party so that it can be analyzed which government gave good performance and next time they can select the best candidate. This can be done by doing sentiment analysis on news dataset in order to evaluate the performance of government. This performance accounting approach becomes important for each regime to outperform its predecessors, and it thus provides the foundation to suggest why it is vital to evaluate a government's governance. This is the main motivation behind the topic selection. The example of news headlines sentiment classification is as a particular headline could belong to 'positive' ,'negative' or 'neutral' class. Example is shown in figure 1.3.

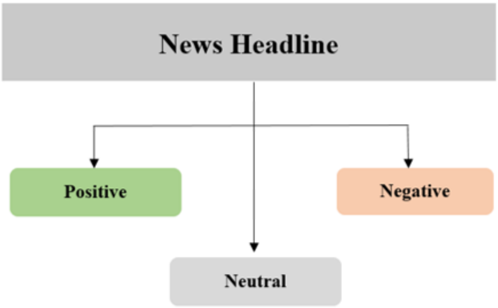


FIGURE 1.3: Classification of sentiment analysis approaches.

1.1.4 Text Representation

The majority of today's data is unstructured and consists primarily of text data. Understanding text data is critical since it contains a lot of information and may be used in a variety of applications. There have been several popular techniques to representing texts in past few decades. A distinction is frequently noted between count-based and prediction-based representations [27]. The primary goal of both approaches is the same as most of the Machine Learning algorithms require text to be converted in numerical representations. Therefore before performing classification or any other operation on text we need it to be first converted in vectors. Researchers have leveraged many feature representations which aims to represent unstructured text into numeric vector to make it mathematically computable. The word vectors in count-based representations are initialized based on the frequency of word occurrences in the text. The Bag of Words(BoW) method is an easy way to represent texts. The BoW method ignores grammar and word order, simply focusing on word count. The word counts represent the document's features.

A more enhanced model of BoW representations proposed by Manning et, al.[29] is TF-IDF. In a document, the term frequency(TF) is multiplied by the inverse document frequency(IDF). The number of documents that contain the word is the document frequency. As a result, when compared to BoW, this technique gives unique words higher ratings than common terms. Numerous techniques i.e. Bag of Word (BOW), Term Frequency(TF), and Inverse Document Frequency (TFIDF), have been developed and used in the literature to represent text documents in numeric form. The latest approaches for sentiment analysis of news have employed these conventional statistical measures like TF, BOW, and TFIDF [1, 2, 7]. While they are effective methods for extracting features from text, the model's inherent nature causes to lose additional information such as semantics, structure, sequence, and context around neighbouring words in each text document. This study focuses on considering the semantic and contextual meaning of terms.

In literature, there are different semantic techniques for feature representation. A well-known technique that is utilized in a variety of domains is Word embedding. Word embeddings are a kind of word representation that assign similar representation to a word having similar meanings. Word embeddings started to gain popularity when Mikolov et, al.[28] introduced a great technique named as Word2Vec. Word2Vec is a statistical method for efficiently and effectively learning a standalone word embedding from a text corpus. It is one of the most common ways for learning word embeddings. It can be used with either the Continuous Bag of Words (CBOW) or continuous skip-gram models. The CBOW model tries to categorize a word by looking at the terms that surround it while by predicting context words depending on the target word, the skip-gram model seeks to achieve the reverse of CBOW. Both methods provide word embeddings that represent textual information. The neural network is used to train the algorithm.

A large corpus of text is used to train the model. The model may discover synonyms or suggest extra words for a partial sentence once the training phase is completed. Every word is represented by a vector. The cosine similarity between the vectors can be used to express the semantic similarity between the words. It only reads text from left to right while converting it to vectors. It is one of the drawback of Word2Vec that it cannot assign a proper value to terms that do not present in the corpus. To counter this Joulin et, al.[30] introduced FastText, a more modern static prediction-based technique. The problem is solved by vectorizing ngrams of characters rather than words. FastText, like Word2Vec, uses the CBOW approach to express n-grams in vector space. One of the fundamental drawbacks of these word representations is that the same word has the same vector value regardless of context. As Word2Vec does not take into account the context of the terms from right to left while converting.

ELMo [31] and Flair embeddings are two contemporary approaches for contextualized representations. To incorporate textual information into a context-vector, both methods require a bidirectional LSTM architecture. But these representations are limited due to their recurrent architecture as this architecture cannot be parallelized which makes it less efficient. Another approach called BERT counter its limitation by using Transformer architecture.

BERT is a Bidirectional Encoder Representations from Transformers. Unlike Word2Vec it converts text into vectors by using both left and right context of terms as it is a machine learning model based on transformers architecture. Transformers is a mechanism that learns context in both directions. It is a pretained model which is already trained on Wikipedia and book corpus by Google. It was developed by Jacob Devlin and Google researchers and published in 2018. There are basically two models of BERT which are BERT base and BERT Large. Both models are based on various number of layers, hidden units and having different number of heads. According to the requirement, these models can be utilized. The following is a description of both models: BERT base model has 12-layers, based on 768-hidden units, 12-heads and 110M parameter neural network architecture, as opposed to the BERT base, BERT LARGE model have 24-layers, 1024-hidden units, 16-heads and 340M parameter. The detailed working of this model is described in Chapter 3.

1.1.5 Feature Reduction

Data is being produced and collected at an ever-increasing rate in the modern era of technology. However, with machine learning, having too much unnecessary data can be bad. Because there is more data to be standardized, adding more characteristics or dimensions to a model might reduce its accuracy as not all the features are equally important and not every feature can represent the data very well, so the feature selection step is required. A technique for reducing a model's complexity and preventing over-fitting is dimensionality reduction. PCA is an unsupervised linear transformation approach that is widely used in a variety of domains, with feature extraction and dimensionality reduction being the most popular uses. Dimension of the features can be significantly large, and that causes the calculation to be too expensive so dimensionality reduction can help in many cases. Feature selection selects a subset of the original features, whereas feature extraction uses information from the feature set to create a new feature subspace. Because the complexity of the problem grows as the number of features grows, this study employs a machine learning technique for feature optimization. Principal Component Analysis(PCA) is a dimensionality reduction technique for reducing the dimensionality of large data sets by converting a large number of variables into a smaller one that preserves the majority of the data. Principal Component Analysis(PCA) is a statistical process that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principle components using an orthogonal transformation.

1.2 Problem Statement

Feature representation plays an important role in news headlines classification. Existing approaches for feature representation mainly treat words independently or in one direction without considering the context in which they are used. This is one of the reasons which negatively affects model performance in supervised machine learning. There is a need to build a model that considers the context from both left and right direction.

1.3 Research Question

This thesis have formulated the following research questions relying on the problem statement described above:

1. To what extent contextual model will be able to achieve better performance on news dataset?
2. How can we build a ranking system that compare/rank performance of government on the basis of provided keywords.

1.4 Scope

This thesis focuses on performing sentiment analysis approach in order to evaluate the performance of governments. Unlike previous researches which used TF-idf, uni-gram, bi-gram, our proposed context based technique will not count the frequency of the word but it keeps the context from both left and right of the word. The use of Contextual model considerably enhances the performance of Machine Learning Classifiers to predict the polarity of news headlines dataset. The purpose of this study is to perform sentiment analysis approach on news headlines and improve the accuracy of the Machine Learning Classifier by using contextual based model. This study also attempt to compare the performance of successive Pakistani regimes. The effort also represents a step toward creating a Pakistani news dataset from “The Nation” and “The dawn” news websites by applying annotation schemes for sentiment analysis. The reason for selection of this dataset is that there does not exist Pakistani news dataset to measure the performance of governments. Other countries can also apply this approach to evaluate the performance of their government and also can add more regimes as well as more headlines. It is concluded that Bert proved to be beneficial to classify political data due to the large training data it is trained on. As BERT is the first truly bidirectional representation model that keeps both left and right context. Traditional classification algorithms only keep left or right context. The significance of the proposed solution is based on two perspectives:

1. Technique: The use of BERT model may considerably enhance the performance of Machine Learning Classifiers to predict the polarity of news headlines dataset.
2. Dataset: The Pakistani news headlines dataset is created to evaluate the government's performance on scientific bases.

Furthermore, this technique could be expanded and implemented by some other countries to evaluate their government's performance on their own dataset.

1.5 Objective of the Research

The main objective of this research is to develop sentiment analysis approach to evaluate the performance of government. Research aims to improve the accuracy of machine learning classifier by using contextual embedding which were not used in previous researches. Also this study ranks the governments on the basis of provided keyword.

1.6 Summary

Data collection and preprocessing are the first steps in the proposed method. The Python language is used to collect data from various news sources. This study proposed a system to perform sentiment analysis based on just the news headlines without going through complete story. News headlines are collected from various news articles i.e. The Dawn and The Nation and then applied some preprocessing in order to clean the data. This study proposed a keyword base approach where a User provides a keyword and proposed system matches that key word with the headlines in the dataset and retrieves related headlines from dataset of three regimes. It then find their sentiment polarity. It provides analysis that in which Prime Minister's time period there were positive news and in which Prime Minister's time period there were more negative news. This work compare the performance of successive Pakistani governments by doing sentiment analysis on news dataset of their relevant time periods. Also proposed approach considerably improves the accuracy of sentiment analysis task by using Bert.

Chapter 2

Literature Review

Lot of research work has been done by many researchers on opinion mining in recent years Computational and natural language processing (NLP) techniques are primarily combined in sentiment analysis to extract, identify, classify, and categorize text information in order to determine people's attitudes and opinions. Positive and negative polarity are commonly used to categorize sentiment. It is primarily utilized in online review posts for movies, books, and consumer products, and it has a lot of potential in the business. Researchers are developing automated methods for sentiment classification of text i.e. classifying whether the text contains negative or positive sentiment [34]. Computer scientists are working on automatically classifying text related to different domains i.e. movie reviews [3], news dataset [1, 2, 4] and twitter data [5, 6, 9], product reviews and finding the opinions for predicting the leading party in elections etc. into one of the predefined classes. Sentiment classification task has carried out using two approaches which are machine learning approach and lexicon based approach. Researchers are utilizing both type of approaches by using dictionaries and classifiers to identify the sentiment polarity, it can be done at three levels: Sentence level, Document level and Aspect level [11, 13]. Sentiment classification can also be done using both approaches which result in hybrid approach.

1. Machine learning approach
2. Lexicon based approach

2.1 Machine Learning Approaches

Machine learning approaches can either be supervised or unsupervised. The basic aim is to create a classifier. The classifier requires training examples, which can be labelled manually or collected from an internet source that has been user-generated and labelled. Support Vector Machines (SVM), Naive Bayes classifier, and Multinomial Naive Bayes are the most commonly used supervised algorithms. The text to be analyzed must be represented as a feature for supervised approaches. The features are split into two parts: training and testing. First the classifier is trained on labeled data and then predict the class or labels on test data. To improve the sentiment analysis outcome, techniques such as feature selection, data integration, data cleansing, and crowd sourcing are required. This section elaborates the state-of-the-art Machine learning-based approaches:

For the first time, Pang and Lee introduced supervised classification to sentiment analysis. They used data from movie reviews to perform three machine learning approaches: Nave Bayes, Maximum Entropy Classification, and Support Vector Machine. It is stated that supervised techniques have been shown to outperform unsupervised techniques in terms of performance [33]. Savita et, al.[3] presented an ensemble classifier for sentiment classification of social media reviews. They Developed ensemble classifier which consist of SVM and ANN. Two feature sets: MPFS (most persistent feature set) formed by chi square and is further optimized using GA genetic algorithm to form optimized feature set. They developed baseline classifier models using naïve Bayes, MAXENT and SVM which are trained on bigrams as well as trigrams features. MPFS used to train ANN to produce ANNFS. SVMA2N2 uses both OFS and ANNFS to classify. Accuracy of model is compared with their baseline classifier models. Result concluded that SVMA2N2 achieves 97% accuracy. One of the reason of their improved accuracy is Feature optimization and the other reason is parallel processing of feature sets by SVM and ANN.

Maizatul et, al.[7] performed sentiment analysis on Malaysia financial news headlines using machine learning approaches. Authors have collected data from New

Straits Times web pages and applied some preprocessing steps i.e. removing stop words and stemming. After preprocessing they have stored their data in csv and then applied Opinion Lexicon-based algorithm and naïve bayes algorithm. They declared the resultant sentiments as positive and negative.

Gupta et al [8] proposed an approach for classifying political social media messages by using pretrained model BERT. Researchers have leveraged many supervised Machine Learning models to classify political campaigning content. Authors explored various ML algorithms i.e. linear regression and Naïve bayes but they have obtained best results by using SVM. Authors first applied SVM on Facebook posts and twitter tweets of US presidential elections and then used BERT model to further improve the accuracy. The outcome of the study revealed that by using Bert the authors have achieved improved performance by 9.0% for Twitter and 5.7% for Facebook.

Manish Munikar, Sushil Shakya and Aakash Shrestha [10] Used BERT model and fine tune it on the Standford Sentiment Treebank (SST) dataset to perform fine-grained sentiment classification task with five class labels. Dataset contains 11,855 one sentence movie reviews extracted from rotten tomatoes. Sentences are parced through Stanford consistency parcer in a tree structure. Preprocessing has been applied on dataset and special tokens [‘CLS’] and [‘SEP’] were added. Authors used BERT for computing sequence embedding and applied dropout regularization in training phase and then applied softmax layer with activation function which turns numbers into probabilities. Authors compute the accuracy and compare results with previous models and achieved 91.2 and 93.1% accuracy on SST-2 dataset by using BERT base and BERT large, 53.2 and 55.5% accuracy for SST-5 dataset.

Prashant Raina et, al.[37] proposed an approach for sentiment analysis in news articles using sentic computing. The objective of sentic computing is to make computers understand human emotions. Authors built their own engine specifically designed for polarity classification in news articles. Each sentence is given as an input to the semantic parcer.it extracts the set of common sense candidate concepts from each sentence and passed to sentiment analyzer which then match sentic vectors in senticNet and candidate concepts to classify each sentence as positive, negative or neutral. The performance is evaluated in terms of accuracy, precision, recall and F-measure. The reported accuracy level of this approach was 71%. Results concluded that this method is very reliable in identifying neutral sentences.

Chetashri Bhadanea, Hardi Dalalb and Heenal Doshic et al [38] proposed a two-step method for sentiment analysis. Authors also discussed various methods and their variants for text classification. Authors discussed two areas of sentiment classification machine learning and lexical approaches. Lexical approaches and its variants: baseline approach, stemming, POS tagging, WordNet, n-grams, Conjunction Rules, Stop Words and Negation method. Machine learning approaches: SVM and naive Bayes. The analysis of proposed method is performed in two parts: aspect identification and sentiment identification. For building a model for aspect classification. All lexicons from all reviews after preprocessing are used as features. All lexicons appearing in the reviews for that aspect after preprocessing (same as before) are chosen to create the models for polarity classification.
The use of word embedding in [53], high-dimensional word vectors that learn contextual information for words improves sentiment categorization accuracy. They employed word2vec's Skip-Gram Model to analyze sentiment in tweets about the US Military Base in Ghana. For training, the Random Forest classifier is utilized. They employed evaluation criteria like as accuracy, recall, precision, and F1-score metrics to assess the performance. The overall accuracy for the sentiment labels was 81%, indicating that the skip-gram model's quality word vectors aid in providing accurate sentiment labels.

Atul et, al.[52] compared different techniques to detect sarcasm. In order to train their model they used Twitter dataset, amazon product review and News headline dataset which are the most frequently used datasets, to remove inconsistencies they first preprocess the datasets and perform some feature engineering to train a better model with better performance. In order to predict weather the data is sarcastic or not different approaches can be applied.e.g. SVM, naïve Bayes, Random forest, Lexical method, neural networks. The study shows that Support vector machine (SVM) is the finest approach for sarcasm detection.

The work of Jinyan Li el, al.[39] demonstrated the hierarchical classification by applying multiple level of filtering on the semantic analysis algorithms to evaluate the performance. The main aim of research was to observe how these filtering levels contribute to the sentiment content and interception of text. Preprocess the data and applied three filters. Filters were polarity word filter, high frequency words filter and unique high frequency words filter. Filtering schemes were applied on three sets of articles where binary and multiclass classification are applied. Six Classification algorithms were validated using Tf-idf and SVM and then compared with each other. Results concluded that with SVM: Max entropy, Naive Bayes performed best, balanced winnow and c45 gave average performance, while decision tree and winnow gave worst performance. With Tf-idf: max entropy performed good and naïve Bayes gave bad performance.

Work of Ubale et,al.[40] this work presents analysis of news articles related to company. Candidate keywords are matched with positive and negative dictionary to classify it positive and negative using enhanced naïve bayes classifier. Laplacian smoothing solved the problem of zero probability. For handling duplication Bernoulli Naïve bayes is used. Handling negations was a major problem confronted during sentiment Classification and is solved by using N-gram. The algorithms performed very well as compare to other alternatives. And also reduced processing time. Algorithms Used for SA not only give better results than the other alternatives but also reduce the time required for processing.

Islam et,al.[36] proposed an approach to classify news articles based on structure of the sentence and dynamic dictionary. First it selects online news articles and then extracts the paragraphs sentence by sentence. Each sentence is then determined as simple, complex compound or complex compound. If it is simple sentence then it searches for positive and negative words and calculates the polarity of sentence. If the sentence is compound then divide the sentence into segments and determine the polarity for each and add their polarity to obtain sentence polarity. If it is a complex sentence then find the polarity for both dependent and independent clause and add them to obtain sentence polarity. Independent clause will obtain more weight than dependent. If a sentence is compound complex sentence then find the polarity of all clauses and add it to obtain polarity for sentence. Obtain the polarity for whole news article by adding polarities of each sentence. The resultant article will be classified as positive, negative or neutral. The performance of the proposed approach is evaluated by using confusion metric. It is concluded that it achieves 91.07% accuracy with 8.93% error margin.

Another approach was proposed by Agarwal, Sharma and Sikkaa [35] to performed sentiment analysis. They use two algorithms to evaluate the headlines. Algorithm 1 used to preprocess the words taken from news headlines. Preprocessing was carried out with the use of a POS tagger, lemmatization and stemming. The resultant words are then passed to SentiWordnet dictionary (Algorithm 2) as an input to identify positive, negative and neutral scores. Positive and negative words of headlines are then summed up separately. Algorithm 3 is the used for calculating sentiment scores day-wise. Headlines are also checked manually and sentiments scores are assigned to them. The algorithms have been applied ton 500 news headlines. Accuracy of classification is also tested by comparing it with manually classified news headlines. In future these headlines can be classified using SVM technique. The proposed algorithms can be applied on product reviews, movie reviews and comments on social media.
Based on this research, Chaudhary, et al.[1] developed a new approach for extracting opinions from news headlines, with the goal of improving the accuracy that prior algorithms lacked [35] by using different technique. NLTK tools and SVM classifier are employed to build the proposed method. News headlines are fetched and are classified manually either negative or positive and processed in coreNLP to build a classification model in order to perform SVM. CoreNLP process the data and gives the output as set of relative words. Count vectorizer is then used to convert these words from string to numeric values. Here three models are built. MODEL A: uni-gram bi-gram with SVM. Tf-idf on uni-gram bi-gram with SVM is used to build model B. Model C: The SGD is used to train the data for SVM. Model is evaluated by using confusion matrix. SGD and linear SVM models perform best for larger datasets and for smaller dataset TF-IDF and linear

SVM perform well according to the results and analysis. This method outperforms existing techniques while cannot handle sarcasm.

Their dataset contains 1472 Indian news headlines with their respective sentiment score manually annotated by humans. These 1472 news headlines comprise on 302 negative and 1170 positive headlines. After exploring the dataset get the positive and negative number of records we observe that data is imbalance. There is a class imbalance when observation in one class is higher than observation in other classes. To classify data into positive and negative classes, for example. here, positive records number roughly 302 whereas negative records number around 1170. In machine learning, class imbalance is a typical issue, particularly in classification problems. Imbalance data can significantly reduce the accuracy of a model. Several machine learning approaches work well when the number of samples in each class is nearly equivalent. As most algorithms are supposed to maximize accuracy while minimizing errors. If the data set is uneven, however, one can anticipate the majority class with high accuracy, but can miss the minority class, which is normally the goal of constructing the model in the first place.

Their dataset also contains duplicate records. After experimentation we observe that there are 170 records that are duplicate. Out of 170 there are 28 negative records and 142 positive. Data cleaning is critically an important step in any machine learning task. By identifying and eliminating rows containing duplicate data, machine learning algorithms perform better. Duplicate rows cause misleading performance. If we use a train/test split, for example, a duplicate row or rows may exist in both the train and test datasets, and any model evaluation based on these rows should be correct. As a result, an optimistically skewed estimate of performance on unseen data will be produced. According to Thelwall et, al.[25] the loss of classification accuracy can be strongly correlated with the contamination level. Rows of duplicate data should be deleted from dataset before modeling.

2.2 Lexicon Based Approaches

Opposed to Machine learning approaches, Lexicon based approaches also known as unsupervised approaches it means there is no need to preprocess the data or train a classifier to predict the polarity. Lexicon based methods consist of dictionary containing words have some polarity value. Words within documents are matched with the dictionary words in order to find polarity. Lexicon based method includes Dictionary-based and corpus-based approaches. This section elaborates the state-of-the-art Lexicon based approaches.

Soonh Taj, Baby Bakhtawer Shaikh and Areej Fatemah Meghji [2] proposed a Lexicon based Approach for Sentiment Analysis of News Articles. BBC news dataset is utilized to perform experiments. Preprocessing has been done by using the Rapid Miner tool and important words are identified using Tf-idf. Sentiment score is assigned to those discovered words using wordNet dictionary. First it finds the polarity of individual words, phrases or sentences and then combine to predict the polarity of whole document. News articles were classified in to either positive, negative or neutral by observing their total sentiment score. From Results it is observed that most of the news articles fell into the positive or negative categories with a minor percentage of articles having neutral sentiments. This method only uses English news articles from one source for sentiment analysis.

K C Ravishankar et al [4] presented a system for modeling news articles in to topics and compute the sentiment scores of those topics to determine whether the discussion is positive, negative or neutral. They performed sentiment analysis on news articles using a probabilistic topic modeling technique called latent dirichlet allocation and evaluated sentiment scores by sentiWordnet. Discussion showed that there is some correlation between their results and real world scenarios that existed at the time of publications.
Muqeem Ahmed et al [5] used twitter data from September 2019 to October 2019 .Crawler is used to gather the data comprising seven categories from various news websites. Articles are downloaded in html format and stored in a text file and

then preprocessing is performed on that data. Tf-idf is applied on data to identify the important words. Sentiment score is given to those discovered words by using wordnet lexical dictionary. Polarity of single words phrases and sentences is calculated and then combining all the polarities, sentiment score is calculated for each document. Text having polarity score +1,-1 and 0 are considered positive, negative and neutral. Articles are classified into positive, negative and neutral. Results concluded that most of the articles belong to either positive or negative category. Using this approach it is noticed that sentiment analysis focused on English words and other languages like Arabic, Italian. In Future this approach could be implemented for other languages.

Kavya Bharathan and Deepasree Varma P [6] proposed an approach of topic modeling for summarizing large documents and for automatically extracting sentiments. They have collected data based on three events using twitter API and news feed. They preprocess the data and apply Latent Semantic Analyzer for extracting mostly discussed topics and visualize them by using Word Cloud. For sentiment collection they used Valence Aware Dictionary and Semantic Reasoner and results are visualized by using bar chart. It is concluded that their approach achieved accuracy of 78.3%.

In this Ankita Sharma and Udayan Ghose [9] presented an approach to perform Sentiment Analysis on tweets in order to get insight towards people's opinion concerning general elections in India. Authors have collected tweets of two months from Jan 2019 to Mar 2019 by using twitter API, preprocessing has been done and extracted Named entities from tweets. At the end authors performed sentiment analysis based on two approaches: Lexicon based approach (LBA) and NRC dictionary based approach. Using LBA approach tweets were classified into positive, negative and neutral classes while using NRC dictionary based approach tweets were classified into 8 classes. Two popular political candidates were considered for analysis. Experimentation have been done on set of real tweets in order to validate the results of the proposed approaches. It is concluded that results obtained by proposed approaches was in full compliance with the actual results obtained in 2019 elections.

2.3 Feature Selection

The salient aspects of text or documents are represented as feature vectors since most sentiment analysis systems rely on machine learning techniques. Feature selection, or data processing to eliminate the least useful n-grams, has been proven to enhance classification performance marginally, for example, by selecting a limited set of features (e.g., 3000) that score highest on a measure like information gain [41] or log probability [42]. Small improvements can also be made when employing n-grams by trimming the feature set of features that are subsumed by simpler features with higher information gain values [41]. Bi-grams and tri-grams, according to [43], give improved product-review polarity classification. [44] Found that terms that appear only once in a corpus are effective indicators of high-precision subjectivity. Machine learning techniques are significantly more powerful, but because of the high number of features, they are much more computationally expensive to use. Feature selection can be employed to limit the amount of features. Despite its widespread use, little study has been done on the effects of alternative approaches of feature selection in sentiment analysis [45, 46]. Recently [47] proposed a method especially developed for reducing dimensionality of feature space in sentiment analysis (SA) problems. [48] Proposed a new approach for Feature Reduction using Principal Component Analysis for Opinion Mining.

2.4 Critical Analysis of Literature Review

Researchers have leveraged many embedding techniques for feature representations on various datasets. Creating features from text is not problematic by itself as it is the foundation for different models, like Term Frequency Inverse Document Frequency (TF-IDF), Uni-gram, Bi-gram but they can be problematic by removing words from the context in which they were originally situated. Homonyms and other forms of textual polysemy are not distinguishable by supervised algorithms. Similarly, supervised algorithms do not have any prior knowledge of a word's definition and/or synonyms, which can cause problems for classification models

Table 2.1: Critical Analysis of Literature Review

Title	Dataset	Approaches Representaion Technique	Result	Analysis of Literature/ Limitation
Sentiment Analysis for Financial News Headlines using Machine Learning Algorithm (2018)	Financial news headlines	Bag of Words, Opinion Lexicon, Naīve Bayes	N/A	Cannot capture context.
Sentiment Analysis with word embedding (2018)	Tweets dataset	Skip gram of word2Vec Random Forest	Accuracy1: 81%	Consider context in one direction
Opinion mining on newspaper headlines using SVM and NLP (2019)	News paper headlines	TFIDF, Uni-gram, bi-gram representation, 1.Tf-idf+SVM 2. SVM+SGD	Accuracy1: 88.13% Accuracy2: 91.52%	Statistical method for feature selection, can't consider context of words
Sentiment Analysis of News Articles: A Lexicon based Approach (2019)	News Articles	Word net Tf-idf	N/A	Can't consider context of words , statistical method for feature selection, limited word coverage
Sentimental Analysis of Twitter Data with respect to General Elections in India (2020)	Tweets dataset	Uni-gram NRC dictionary	N/A	Limited word coverage, used unigram , Cannot consider context of words

Table 2.1-Continued from previous page

Title	**Dataset**	**Approaches Representaion Technique**	**Result**	**Analysis of Literature/ Limitation**
Sentiment classification of social media reviews using an ensemble classifier (2019)	Movie reviews	ANN MPFS,SVM ,Genetic Algo Uni-gram, Bi-gram,Tri-gram	Accuracy 97.4%	Cannot consider the context of words w.r.t whole sentence

when commonly used terms are replaced with synonyms that the algorithm does not recognize. The following are the primary observations from literature review that motivated and signify our proposed framework. To fill the gaps of previous researches we focus on the use of contextual language model. The brief overview of most related techniques to our work proposed in literature is given in Table 2.1.

Count based techniques for feature representation belong to a family of bag of words models, like Tf-idf, N-gram and so on while these methods are useful for extracting features from text, we lose additional information such as semantics, sequence, and context of words because they are just a bag of unstructured words. This forms enough motivation for us to explore models which can capture this information and give us context based features in the form of vector representation.

Chapter 3

Research Methodology

3.1 Introduction

The critical analysis of already proposed approaches in literature review dictates that the researchers has proposed different approaches to find the sentiments on news datasets into binary and multiple categories. The following are the main findings from the literature review that motivated and defined the proposed framework: 1) to the best of our knowledge, there is no study that has evaluated the performance of governments by using sentiment analysis. 2) There does not exist Pakistani news dataset to measure the performance of governments. 3) There is no study that has employed a contextual model for text representation that considers left and right context of the terms on news dataset. As a result of these primary observations this study has proposed a technique to overcome the issues discussed above. This chapter elaborates details of techniques that adopted in this thesis based on the background, literature review and state of the art from the former two chapters. It is stated that Machine Learning algorithms require text to be converted in numerical representations. Researchers have leveraged many feature representations for this purpose. Recently, TF-IDF, Uni-gram and Bi-gram is used by a researcher for text to numeric representation [1] but it lacks the context while converting words from text to numeric as it counts the frequency of the words appear in document. This study has proposed pre trained model BERT which keeps both left and right context of the words. It assigns different vector representations to the words which are having semantically different meanings. BERT (Bidirectional Encoder Representations from Transformer) a deep learning model that consider both left and right context of a word. It is pre-trained on wikipedia and book corpus. Its attention based mechanism process all input tokens in parallel. One additional layer is required to fine-tune it to obtain state of the art results. Recently[8] study revealed that the improved performance of machine learning classification on political content can be achieved by use of BERT. In this research work, our aim is to classify news headlines in order to compare the performances of various regimes in Pakistan by doing sentiment analysis. This study proposed a system where a user provides a keyword and on the basis of that keyword it searches related headlines from dataset of three regimes and find their sentiments. It provides analysis on the basis of positive and negative impact of three regimes w.r.t keyword.

3.2 Methodology

This method begins with data collection. First, data i.e. news headlines and dates are extracted from newspaper websites by using python language and stored in CSV format. Annotation Scheme is proposed to assign labels in order to train the classifier. Preprocessing i.e. tokenizing, stop words removal, stemming is done in python using NLP toolkit. Data is then converted from text to numeric using BERT(Bidirectional Encoder Representations from Transformer). In machine learning, having too much data can be bad. Because there is more data to be standardized, adding more characteristics or dimensions to a model might reduce its accuracy as not all the features are equally important and not every feature can represent the data very well, so the feature selection step is required. Here, PCA is used for dimensionality reduction to reduce the model's complexity. After that, Machine learning classifier is used to predict the opinions in the text. Sentiment results are then calculated. As evaluation measures we use F1 score and Accuracy. Figure 3.1 represent the block diagram of proposed system.

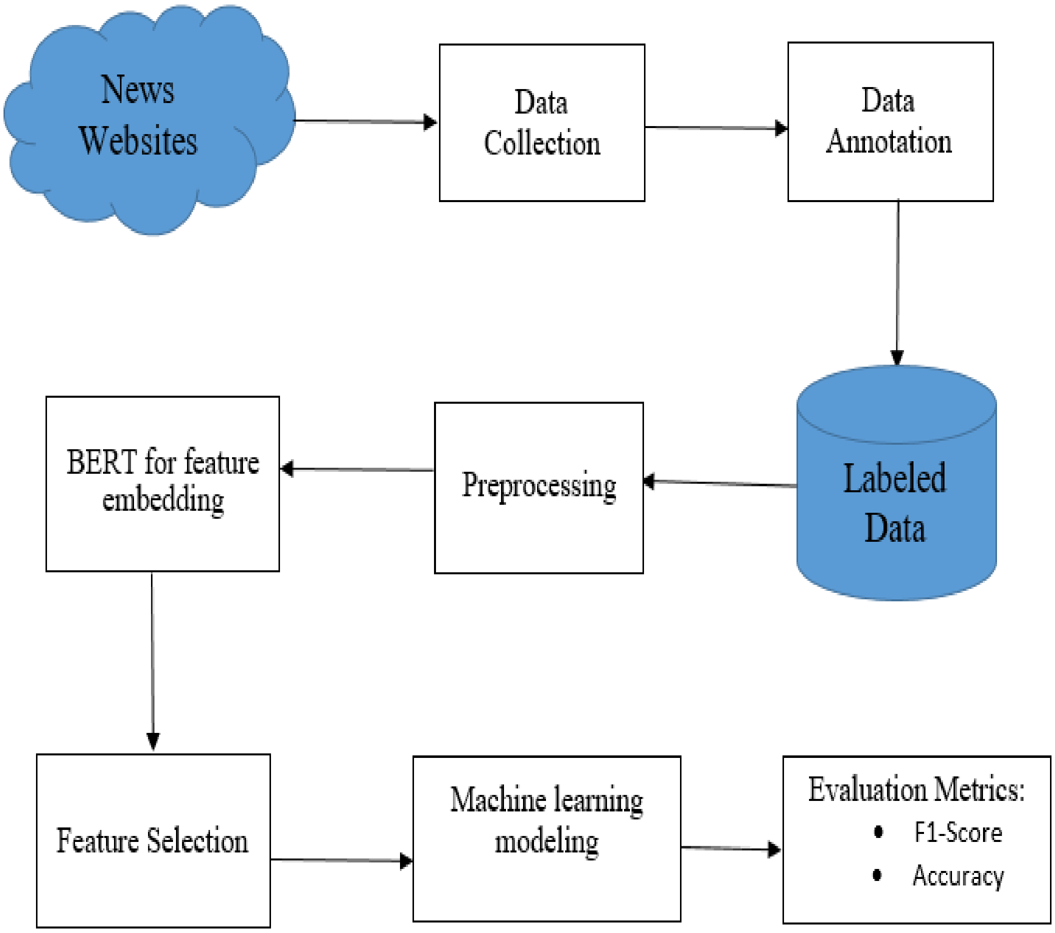


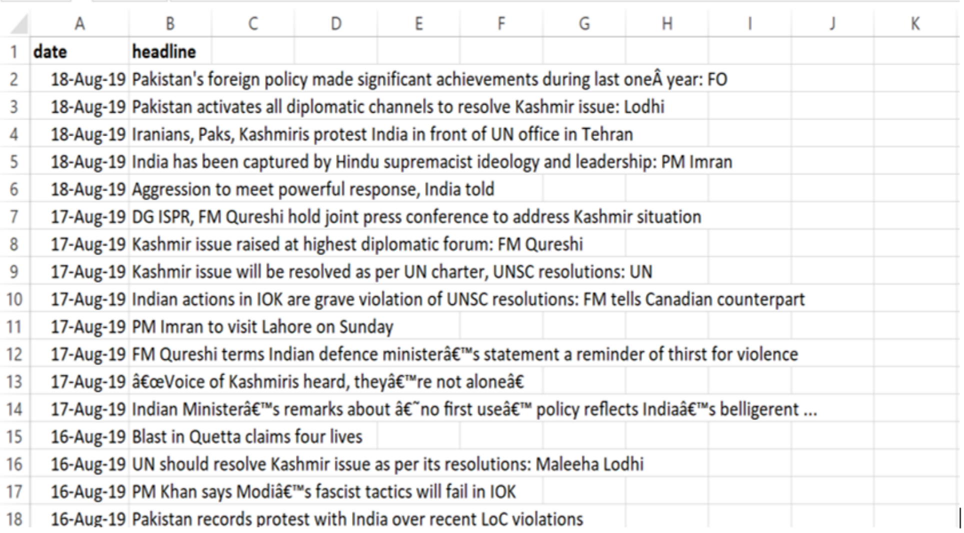
FIGURE 3.1: Methodology diagram of Proposed Solution

3.2.1 Dataset

We have used two datasets of news headlines for experimentation, implementation and evaluation of our proposed approach.1) Pakistani news Dataset, 2) Base-paper news Dataset [1].

3.2.2 Data Collection

To evaluate the proposed system, dataset has been created that contains news headlines from "The Nation" and "The Dawn" news websites ranges from 2008 to 2020. The reason for choosing this dataset is that there isn't exist a Pakistani news dataset that can be used to assess government performance. Scraper is developed to scrape news headlines. These headlines are then stored in CSV file format. Each row of news headline is associated with its publishing date. We limited the data collection to active government time periods for specific years. Figure 3.2 presents example set of news headlines.



	A	B
1	**date**	**headline**
2	18-Aug-19	Pakistan's foreign policy made significant achievements during last oneÂ year: FO
3	18-Aug-19	Pakistan activates all diplomatic channels to resolve Kashmir issue: Lodhi
4	18-Aug-19	Iranians, Paks, Kashmiris protest India in front of UN office in Tehran
5	18-Aug-19	India has been captured by Hindu supremacist ideology and leadership: PM Imran
6	18-Aug-19	Aggression to meet powerful response, India told
7	17-Aug-19	DG ISPR, FM Qureshi hold joint press conference to address Kashmir situation
8	17-Aug-19	Kashmir issue raised at highest diplomatic forum: FM Qureshi
9	17-Aug-19	Kashmir issue will be resolved as per UN charter, UNSC resolutions: UN
10	17-Aug-19	Indian actions in IOK are grave violation of UNSC resolutions: FM tells Canadian counterpart
11	17-Aug-19	PM Imran to visit Lahore on Sunday
12	17-Aug-19	FM Qureshi terms Indian defence ministerâ€™s statement a reminder of thirst for violence
13	17-Aug-19	â€œVoice of Kashmiris heard, theyâ€™re not aloneâ€
14	17-Aug-19	Indian Ministerâ€™s remarks about â€˜no first useâ€™ policy reflects Indiaâ€™s belligerent ...
15	16-Aug-19	Blast in Quetta claims four lives
16	16-Aug-19	UN should resolve Kashmir issue as per its resolutions: Maleeha Lodhi
17	16-Aug-19	PM Khan says Modiâ€™s fascist tactics will fail in IOK
18	16-Aug-19	Pakistan records protest with India over recent LoC violations

Figure 3.2: Pakistani News Dataset

The detailed description about the dataset and the brief discussion of components which have been done for the process of Sentiment classification is provided as below:

3.2.3 News Headlines Dataset

News headlines are collected from “The nation” and “The dawn” news websites. For evaluation of government performance we have selected first year data of three regimes i.e. Mr.Yousaf Raza Gillani, Mr.Nawaz Sharif and Mr.Imran Khan. Typically, the time duration when an interim government is formed (due to a politically unstable scenario or a power transfer procedure following new elections), the goal for socioeconomic problems is very low or non-existent [16]. Therefore we are not evaluating the performance on the basis of actual time period a regime has reigned but by the starting era for each regime. For example, Mr.Mir Zafarullah Khan Jamali’s era commenced on November 2002 and ended in June 2004 and Mr.Shaukat Aziz took charge in 28 August 2004. The 1 Month, 27 Days’ time period, from 26 June 2004 to 30 June 2004. An interim government was in charge at the time. It is highly likely that the interim government did not have time to offer any significant

Table 3.1: Time duration of Various Governments in Pakistan

	PPP	PMLN	PTI
Time Periods/ Duration	25-march-2008 to 25- mar-2009	5-june-2013 to 5-june-2014	18-aug-2018 to 18-aug -2021

changes in social policy over these nearly two months. In addition, the first two to three months of a newly elected regime are extremely hectic in terms of forming and naming ministries. This shows that policy changes on social concerns are either non-existent or minor. The resulting corpus of Dawn news contains almost 3747 headlines from particular years. The dataset of The Nation News contains almost 15572 headlines from particular years. Different time frames are chosen for data collection so that the comparison of three regimes could be done by the data of their relevant time periods. The time period of each government is summarized in Table 3.1.

3.2.4 Word Cloud

It's also important to have knowledge of the most commonly used words in headlines. They can be used to learn about the structure and type of dataset. The bigger the word in the word cloud, the more common it is in the corpus. Here, News headlines are visualized with the help of "word cloud" package. Word clouds for one year data of all the three regimes are generated using this procedure. Fig 3.3 shows the word clouds for three parties: PPP, PMLN and PTI.

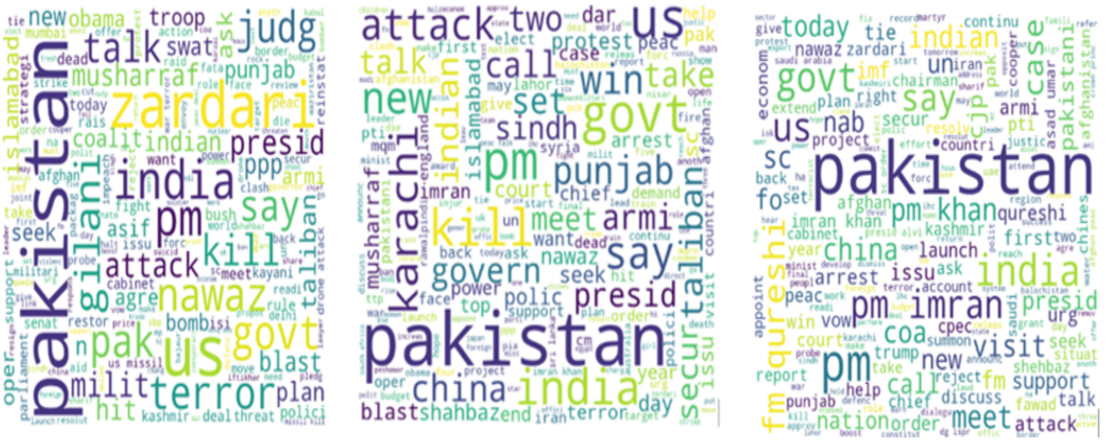


Figure 3.3: Pakistani News Dataset

3.3 Annotation Overview

Data i.e. news headlines, reviews, tweets and sentences for sentiments are annotated by simply asking people to label them as positive, negative and neutral. It works well when the sentiments or emotions are expressed clearly however, for the complex type of data it is difficult for the people to annotate with the true labels [15]. After data collection the next task is to tag the data/news headlines with labels i.e. positive or negative. For this, numerous annotation schemes can be found in literature. Complex or some other type of sentences are challenging for the respondent to annotate. There is need to provide enough information that will not leave the annotator in uncertainty that how to label it. These challenges could be addressed at some extent by providing instructions to annotators on how these type of instances are to be labeled. Therefore, annotation scheme is developed and provided to the annotators. Dataset is manually classified by 41 human annotators and than it is validated by 3 more human annotators. Following is the annotation scheme which has been followed by the annotators.

3.3.1 Annotation Scheme

Label 1

1. Read Text/headline if there are positive and negative adjectives:
2. Positive: Identify that the majority positive adjectives are inclined towards particular government than tag positive.
3. Negative: If majority adjectives are negative and inclined towards particular government than tag negative.
4. Neutral: If headline contain equal number of positive and negative adjectives or not contain any than tag neutral.OR
5. If the respondent is unable to identify the adjectives then consider that the overall impact of headline is positive or negative or neutral towards that particular government and tag accordingly.

Label 2

1. Read headline and state yes or no in label-2 column.
2. YES: If the headline contains government entities and you think it will contribute in evaluating government performance.
3. NO: If there is no government entity and you think it will not contribute in evaluating government performance.

A corpus of 1500 news headlines from the dawn and the nation was carefully evaluated and tagged as negative and positive for the evaluation of our approach. These headlines may be good for one and bad for others here we are taking it generically. After getting the labeled dataset we have discarded those instances which were tagged “NO” in Label-2 column. As we have to evaluate performance of governments therefore we only kept the instances that were tagged as “YES” in Label2 column. After eliminating the un-necessary records there are 984 news headlines in total. Contains 487 positive and 461 negative news headlines. This dataset is further used to evaluate the classifier performance.

3.4 Preprocessing

Preprocessing is a data mining technique that involves preparing the raw data to make it suitable for building and training machine learning classifier. Generally, real world datasets are incomplete, inaccurate or inconsistent. Because the data was extracted from an online source, hence it contains noise that is why, it cannot be used directly for analysis purpose, and therefore its pre-processing is required. For this purpose data is preprocessed in different steps i.e. tokenization, noise Removal, stop word’s removal, duplication removal and stemming. Data preprocessing is the first and extremely important step in data Science. A good result can be derived from well-executed data preprocessing, and vice versa [51]. Figure 3.2 contains the preprocessed sample of dataset. Let us discuss these steps one by one.

3.4.1 Tokenization

The first step in preprocessing is Tokenization. In this process, text can be divided into set of meaningful pieces known as tokens e.g. document can be divided into set of sentences, or set of sentences could be break down into tokens or pieces. In our work we break news headlines into set of tokens.

3.4.2 Stop Word's Removal

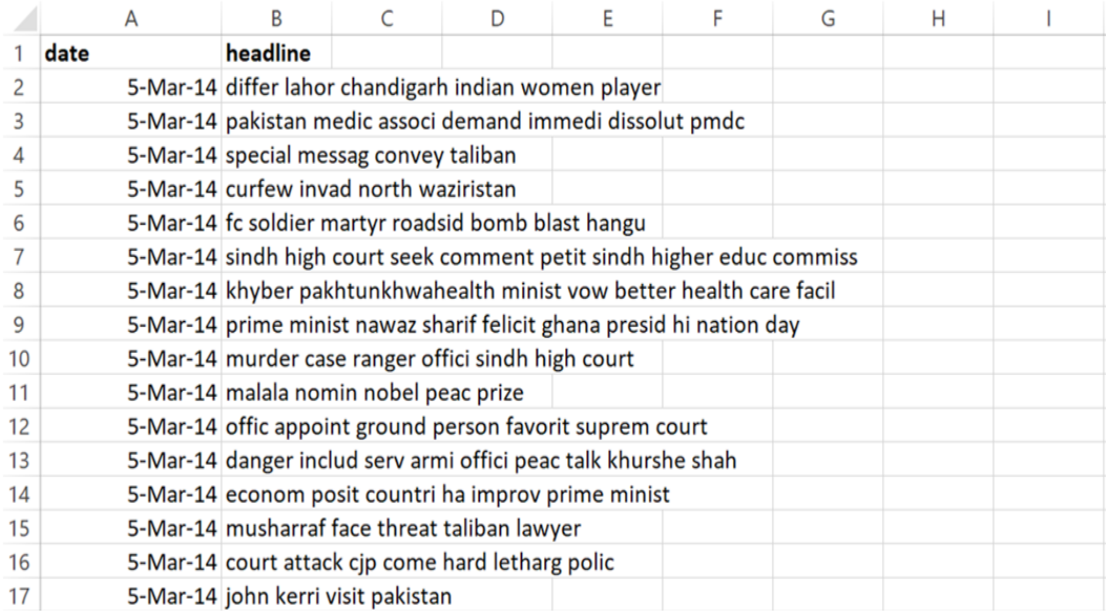
Stop words are the most common words in the language. Such as "a, an, and, are, as, at, he, in, is, its, of, on, for, from, has, that, the, to etc. These words are not much meaningful therefore can be excluded from the text for improved results. We have used the NLTK for stop word removal. NLTK is a library containing list of stop words. It matches the stop words with the tokenized list and removes them from the text.

3.4.3 Stemming

Stemming is the process of generating the root word of the inflected words. It is like cutting the branches of tree to its stems. Advantage of stemming is that it reduces the size of vocabulary. For example the stem or root form of the words "eating", "eats" and "eaten" is eat. We have performed stemming by using porter stemmer algorithm, it converts all the terms of a text into their root terms. The stemming algorithm is applied on whole dataset. The preprocessed dataset is shown in Figure 3.4.

3.4.4 Deduplication

Duplicates were removed from data. There are two types of duplicated data: partial duplicate and exact duplicate. Only exact duplicates were removed from the data. Effect of duplication on results is described in 3.5.3 section.



	A	B	C	D	E	F	G	H	I
1	**date**	**headline**							
2	5-Mar-14	differ lahor chandigarh indian women player							
3	5-Mar-14	pakistan medic associ demand immedi dissolut pmdc							
4	5-Mar-14	special messag convey taliban							
5	5-Mar-14	curfew invad north waziristan							
6	5-Mar-14	fc soldier martyr roadsid bomb blast hangu							
7	5-Mar-14	sindh high court seek comment petit sindh higher educ commiss							
8	5-Mar-14	khyber pakhtunkhwahealth minist vow better health care facil							
9	5-Mar-14	prime minist nawaz sharif felicit ghana presid hi nation day							
10	5-Mar-14	murder case ranger offici sindh high court							
11	5-Mar-14	malala nomin nobel peac prize							
12	5-Mar-14	offic appoint ground person favorit suprem court							
13	5-Mar-14	danger includ serv armi offici peac talk khurshe shah							
14	5-Mar-14	econom posit countri ha improv prime minist							
15	5-Mar-14	musharraf face threat taliban lawyer							
16	5-Mar-14	court attack cjp come hard letharg polic							
17	5-Mar-14	john kerri visit pakistan							

Figure 3.4: Preprocessed Pakistani News Dataset

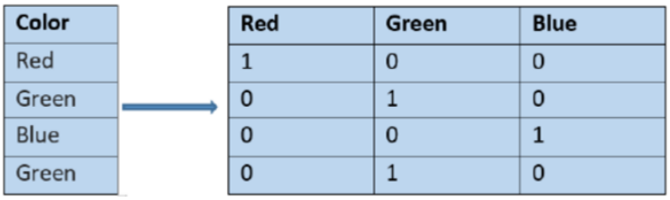
3.5 Conversion of Text to Numeric

Most of the Machine Learning algorithms require text to be converted in numerical representations. Therefore before performing classification or any other operation on text we need it to be first converted in vectors. Researchers have leveraged many feature representations which aims to represent unstructured text into numeric vector to make it mathematically computable. For this, numerous techniques have been presented in literature. Word vectors have evolved over the years to know the difference between “record the play” vs “play the record”. These techniques are mainly divided into two categories: Count based techniques and semantic based techniques.

3.5.1 One-hot-encoding

One-hot encoding is the most common and a most basic way to turn a text into a vector. In this strategy, each word is converted into a binary value 1 or 0, which indicate the word appear in a document or not. Suppose there is a feature ‘Color’ have taken values ‘Red’, ‘Green’, and ‘Blue’. One hot encoding generates new (binary) columns that indicate the presence of all possible data values. It converts the ‘Color’ feature to three features, ‘Red’, ‘green’, and ‘Blue’ which all are binary. The graphical representation is shown in Figure 3.5. Although this is very simple strategy to implement but it has some disadvantages such as:

1. This method does not consider the position of a terms therefore it become difficult to examine the context of a word.
2. Does not consider the frequency information of a terms.
3. Vector representation size grows as the vocabulary size grows.



Color
Red
Green
Blue
Green

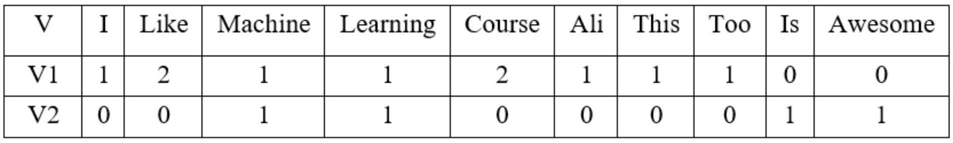
Red	Green	Blue
1	0	0
0	1	0
0	0	1
0	1	0

FIGURE 3.5: An example of one-hot encoding

3.5.2 Bag of Word (BOW) or Term Frequency (TF)

Using a BoW on word vectors is the traditional way to build a document vector for tasks such as classification. Traditional NLP does not focus on understanding the words/text that it is processing. Word embeddings with say a bag-of-words approach can turn a sentence or a document into a short dense numerical vector. In BOW the document vector is a weighted sum of the numerical vectors of the words making up the document. We count that how many times the word appears(Tf) in a document and assign the corresponding vector. Following is the example of BOW having simple text documents: D1: “I Like Machine learning course. Ali like this course too” D2: “Machine Learning is awesome”.

Fig 3.6 shows the encoding of two text documents which are v1 and v2. Here, the vocabulary V = I, like, Machine, Learning, course, Ali, this, too is, awesome. If the word appears in certain news headlines, the value for that feature is 1, and if the word appears twice, the value for that feature is 2. The value is determined by the frequency of the word in the headline. This approach solves the frequency issue of one hot encoding. However, this approach treats words like tokens with no meaning and no relationships with other words. Moreover this approach does not take into account the context of words. It's another variation is N-gram.

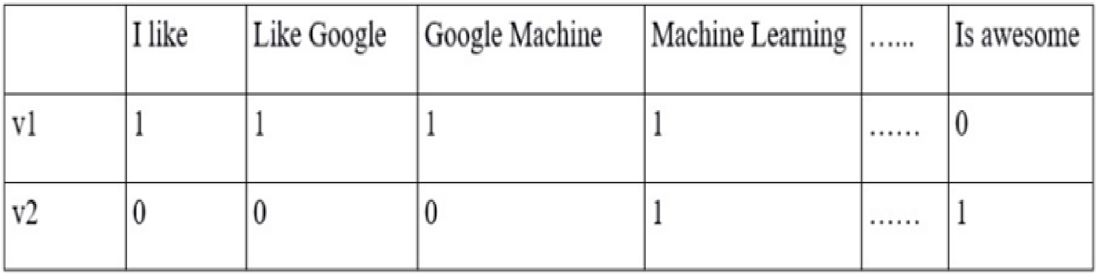


V	I	Like	Machine	Learning	Course	Ali	This	Too	Is	Awesome
V1	1	2	1	1	2	1	1	1	0	0
V2	0	0	1	1	0	0	0	0	1	1

Figure 3.6: TF representation of given vocabulary

3.5.3 N-gram

N-gram is a language model in NLP. In n-gram the probability of the occurrence of a word in a sequence of words is measured it could be done by using uni-gram, bi-gram or tri-gram etc. Following figure 3.7 shows the examples of bi-gram. D1: "I Like Google Machine learning course" D2: "Machine Learning is awesome" The bi-gram model enhances data representation by determining occurrences by a sequence of two words rather than by a single word. Bigrams in a BoW are more powerful than a single word BoW. Despite this, the use of n-grams leads in a large number of irrelevant entries[50].



	I like	Like Google	Google Machine	Machine Learning	……	Is awesome
v1	1	1	1	1	……	0
v2	0	0	0	1	……	1

Figure 3.7: Bi-gram representation of given vocabulary

3.5.4 Term Frequency and Inverse Document Frequency (TFIDF)

Inverse Document Frequency(IDF) and Term Frequency (TF)(IDF). Term Frequency is a metric that measures how frequently a term appears in a document.

Because every document is varied in length, a word may appear significantly more frequently in large documents than in shorter ones. As a result, the word frequency is usually divided by the length of the document (for normalization).The importance of a word is measured using the inverse document frequency (IDF). When calculating TF, all terms are given equal weight. Tf-idf can be calculated as follows: **TF**: It keeps track of how many times a term appears in a document. Because the length of each document varies, a word may appear much more frequently in longer documents than in shorter ones. As a result, the term frequency is commonly normalised by dividing it by the length of the document (i.e., the total number of terms in the document).

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document} \tag{3.1}$$

IDF: It determines the significance of a term when calculating TF, some terms, such as "is," "of," and "that," are well-known for appearing frequently despite having little significance. As a result, must scale down the frequent terms while scaling up the rare ones by computing:

$$IDF(t) = \log_e \frac{Number\ of\ documents}{Number\ of\ documents\ containing\ term(t)} \tag{3.2}$$

Example: Let us consider a document with 100 words and three occurrences of the word apple. The apple word frequency (tf) is then (3 / 100) = 0.03. Let's say we have ten million documents, and one thousands of them contain the term apple. After that, log (10,000,000 / 1,000) = 4 is used to determine the inverse document frequency (idf). As a result, the Tf-idf weight is equal to the product of these numbers: 0.03 * 4 = 0.12.

TF-Idf is an effective method and is widely used in literature [1, 2, 7] for determining the importance of a term in a document However, being a bag of unstructured words we lose additional information such as semantics, structure, sequence, and context around neighboring words in each text document. Therefore, in the next section we'll look at an alternative to the above-mentioned methodologies for capturing the semantic and contextual information of terms, which is extensively utilized in several areas and is producing high accuracy.

3.5.5 Word Embedding

Because the meaning of a phrase varies depending on the context, we must know the semantics and context of a term before we represent it. Let's take the term 'bank' as an example. The word 'bank' has several different meanings. A financial entity or a land alongside the river. So, a similar word is depicting different meanings in different contexts. There is a need to represent a word with different vectors by looking at their context or neighboring words. There are different semantic techniques for feature representation in literature. After performing an in-depth analysis, we identified a well-known technique that is utilized in a variety of domains called Word embedding. Word embeddings are a kind of word representation that assign similar representation to a word having similar meanings.

3.5.5.1 Word to Vec

Word embeddings began to gain popularity when a great technique named as Word2Vec [28] was introduced. Word2Vec is a statistical method for quickly and successfully learning a solitary word embedding from a text corpus. It is one of the most common ways for learning word embeddings. homas Mikolov is the developer of the Word2Vec technique, it was published in 2013. We train a single hidden layer neural network to predict a target word based on its context (neighboring words) with Word2Vec. It only reads text from left to right while converting it to vectors. It is one of its drawback that it does not take into account the context of the terms from right to left while converting. There is just one vector (numeric) representation for each word. For example, The term "bank" is used in two separate contexts: a) as a financial entity, and b) as a piece of land along a river (geography). For both phrases, word2Vec will generate the same single vector for the word bank. It is one of the drawback of Word2Vec that it cannot assign a proper value to terms that do not present in the corpus.

3.5.5.2 FastText

To counter the problem of out of vocabulary words which was faced in WordtoVec [30] introduced FastText, a more modern static prediction-based technique, The

problem is solved by vectorizing n-grams of characters rather than words by using FastText. It was released by Facebook in 2016. FastText uses the CBOW technique to represent n-grams in vector space, similar to Word2Vec. The earliest approaches used words for word embedding. FastText, on the other hand, considers the components of words and characters in addition to the words themselves. It gives a word its meaning, and the parts that make up the word are now characters. So, for example, if the word ”system” has n=2, The representation of FastText is: < S,sy,ys,st,te,em,m> where the angle brackets represents the beginning and ending of word.
This method offers two benefits:

1. It solves the problems of generalization.
2. More training examples aren’t required.

This approach has some disadvantages:

1. Emphasis on the syntactic analogy rather than the semantic.
2. Requires High memory.

3.5.5.3 BERT

BERT stands for Bidirectional Encoder Representations from Transformers. Transformer is a mechanism that learns contextual relations between words in a text. BERT makes use of a Transformer. One of the main advantages of using BERT is, The Transformer encoder reads the complete sequence of words at once, unlike directional models that read the text input sequentially (left-to-right or right-to-left). As a result, it is regarded as bidirectional. BERT is considering the ”context”, not just the single word (the ”window word” and n-grams) as FastText. In a few words, BERT considers the sentence and the sentences around that sentence. It allows the model to understand the particular word within a sentence, and the sentence itself within a period. An important advantage of BERT over the first generation of word embedding models is the capacity of embedding the same word with a different meaning. Let’s take an example, suppose there are two sentences:

1. I checked my bank account.

2. We went to the river bank.

The model generates embeddings for the word based on the context it appears thus generating slightly different embeddings for each of its occurrences. In our case, the bank will have a different vectors according to the different context. As first sentence is having word “bank” in a context of financial entity. In the second sentence word “bank” is used in a context of land alongside the river. There would be a mistake in guessing the meaning of ”bank” in any of these two sentences if both sentences are read from only one side, sentence one from left side and sentence two from right side. Therefore, we have to read these from both sentences in left and right direction so that we can keep the context from both sides. In standard word embeddings such as Glove, Fast Text or Word2Vec each instance of the word bank would have the same vector representation in both contexts. BERT enables NLP models to better disambiguate between the correct senses of a given word. This study revealed that the BERT model is the best at capturing the semantic and contextual information of terms after a detailed analysis of count and semantic based techniques. As a result, in this thesis we have employed the BERT model to vectorize a text.

3.6 Dataset 1: (1472)

Pre-trained BERT model is used here to process news headlines. Then the output of that model is used to classify the headlines into positive and negative classes. here the aim is to build a model that takes a sentence (exactly like the ones in the dataset) and outputs either a 1 (meaning a positive sentence) or a -1 (showing a negative sentence). It can be seen as shown in the below figure 3.8: The model is made up of two models shown in figure 3.9 below: one for feature extraction and the other is for classification. Model-A takes the headlines as an input and extract some information and pass it to Model-B for classification. The data passed between the two models is a vector of size 768. This vector can be thought of as a sentence embedding that can be used for classification.

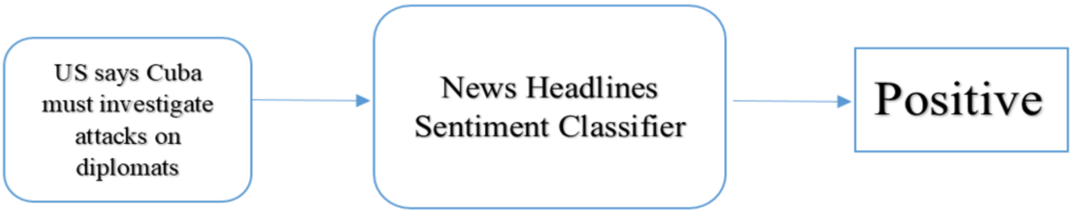

US says Cuba must investigate attacks on diplomats
News Headlines Sentiment Classifier
Positive

Figure 3.8: Sentiment Classifier

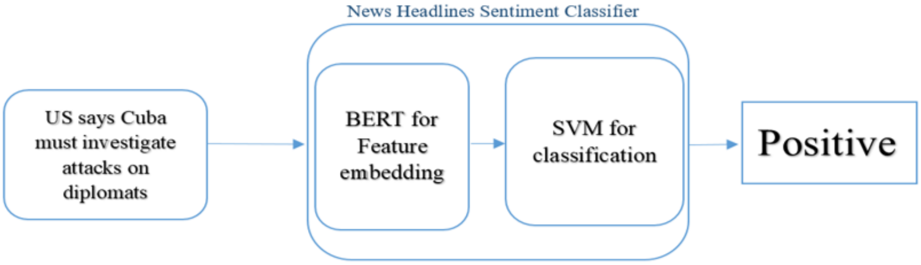

News Headlines Sentiment Classifier
US says Cuba must investigate attacks on diplomats
BERT for Feature embedding
SVM for classification
Positive

Figure 3.9: Two model sentiment classifier

3.6.1 Dataset

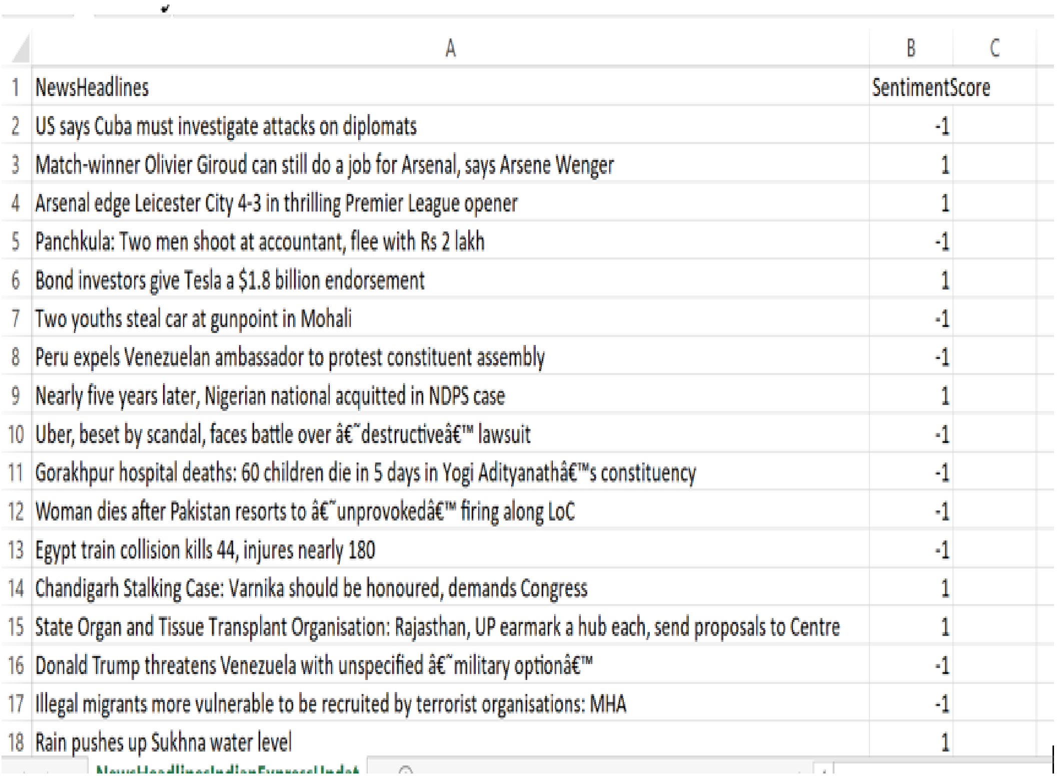
The dataset used here is based on news headlines, which contains 1472 number of news headlines, each one is tagged as either positive (value 1) or negative (has the value -1) manually annotated by humans. figure 3.10 shows the example set of base dataset.

3.6.2 Preprocessing

Preprocessing has been done on news headlines dataset in different steps i.e. tokenization, noise removal, Stop word's removal and stemming. Below is the preprocessed sample of dataset in figure 3.11.

3.6.3 Deduplication

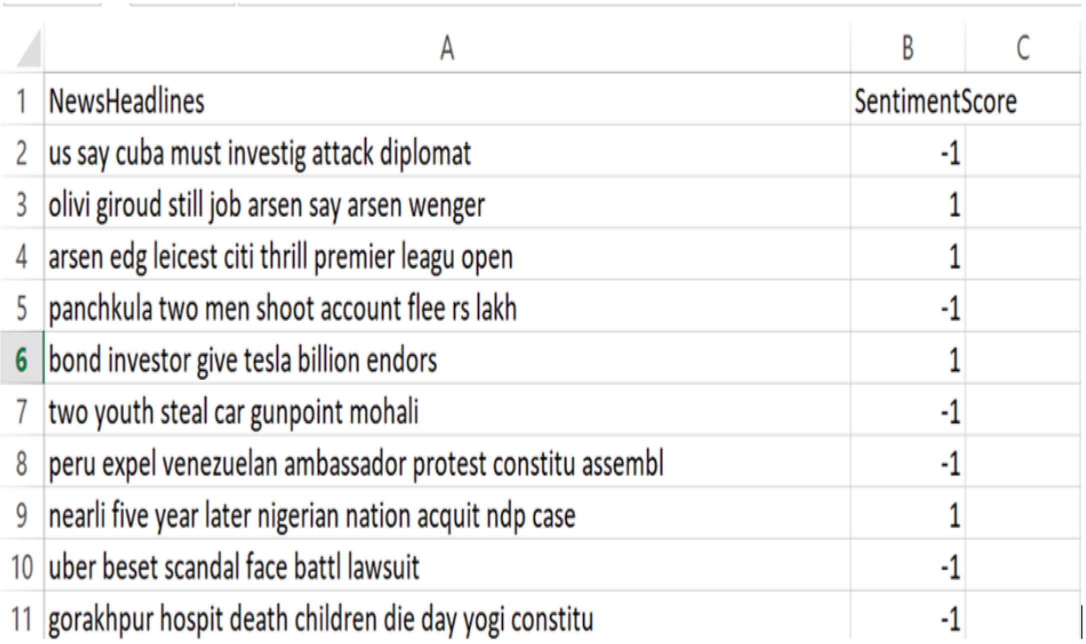
Their [1] dataset also contains duplicate records. After experimentation we observe that there are 170 records that are duplicate. Out of 170 there are 28 negative records and 142 positive. Duplication simply indicate that our dataset contains repeated data. This could happen due to data entry problems, data collection



	A	B	C
1	NewsHeadlines	SentimentScore	
2	US says Cuba must investigate attacks on diplomats	-1	
3	Match-winner Olivier Giroud can still do a job for Arsenal, says Arsene Wenger	1	
4	Arsenal edge Leicester City 4-3 in thrilling Premier League opener	1	
5	Panchkula: Two men shoot at accountant, flee with Rs 2 lakh	-1	
6	Bond investors give Tesla a $1.8 billion endorsement	1	
7	Two youths steal car at gunpoint in Mohali	-1	
8	Peru expels Venezuelan ambassador to protest constituent assembly	-1	
9	Nearly five years later, Nigerian national acquitted in NDPS case	1	
10	Uber, beset by scandal, faces battle over â€˜destructiveâ€™ lawsuit	-1	
11	Gorakhpur hospital deaths: 60 children die in 5 days in Yogi Adityanathâ€™s constituency	-1	
12	Woman dies after Pakistan resorts to â€˜unprovokedâ€™ firing along LoC	-1	
13	Egypt train collision kills 44, injures nearly 180	-1	
14	Chandigarh Stalking Case: Varnika should be honoured, demands Congress	1	
15	State Organ and Tissue Transplant Organisation: Rajasthan, UP earmark a hub each, send proposals to Centre	1	
16	Donald Trump threatens Venezuela with unspecified â€˜military optionâ€™	-1	
17	Illegal migrants more vulnerable to be recruited by terrorist organisations: MHA	-1	
18	Rain pushes up Sukhna water level	1	

Figure 3.10: Base dataset

procedures or could be due to intentionally inserted. If we use a web scraper, for example, we might scrape the same webpage twice or the same information from two separate pages. Whatever the reason behind the cause of duplication, it might lead to erroneous conclusions by making us assume that some observations are more prevalent than they are. According to [25] the loss of classification accuracy can be strongly correlated with the contamination level. Duplication increases data redundancy, which can lead to statistical bias and affect the outcomes of experiments [49]. Data cleaning is critically an important step in any machine learning task. Machine learning algorithms performs better by identifying and removing rows with duplicate data. Duplicate rows results in misleading performance. For example, if we are using a train/test split, then a duplicate row or rows may exist in both the train and test datasets, and any model evaluation based on these rows should be accurate. As a result, an optimistically skewed estimate of performance based on unseen data will be produced. Rows of duplicate data should probably be deleted from dataset prior to modeling. Therefore different experiments have been performed on the original data (contains duplication) and on new data (having no duplication). Results of these experiments are presented



	A	B	C
1	NewsHeadlines	SentimentScore	
2	us say cuba must investig attack diplomat	-1	
3	olivi giroud still job arsen say arsen wenger	1	
4	arsen edg leicest citi thrill premier leagu open	1	
5	panchkula two men shoot account flee rs lakh	-1	
6	bond investor give tesla billion endors	1	
7	two youth steal car gunpoint mohali	-1	
8	peru expel venezuelan ambassador protest constitu assembl	-1	
9	nearli five year later nigerian nation acquit ndp case	1	
10	uber beset scandal face battl lawsuit	-1	
11	gorakhpur hospit death children die day yogi constitu	-1	

Figure 3.11: preprocessed dataset

in Chap 4.

3.6.4 Imbalance

These 1472 news headlines have 302 negative and 1170 positive headlines. After exploring the dataset and getting the positive and negative number of records we observe that data is imbalance. There is a class imbalance when observation in one class is higher than observation in other classes. To classify data into positive and negative classes, for example. It can be clearly seen in the graph below, positive records number roughly 302, whereas negative records number around 1170. In machine learning, class imbalance is a typical issue, particularly in classification problems. Imbalance data can significantly reduce the accuracy of a model. When the number of samples in each class is roughly equal, most machine learning methods work well. As most algorithms are supposed to maximize accuracy while minimizing errors, this is the case. However, if the data set is unbalanced, High level of accuracy can be simply achieved by forecasting the majority class, but it will miss the minority class, which is usually the objective of developing the model in the first place. Data imbalance is illustrated in Figure 3.12 and 3.13.

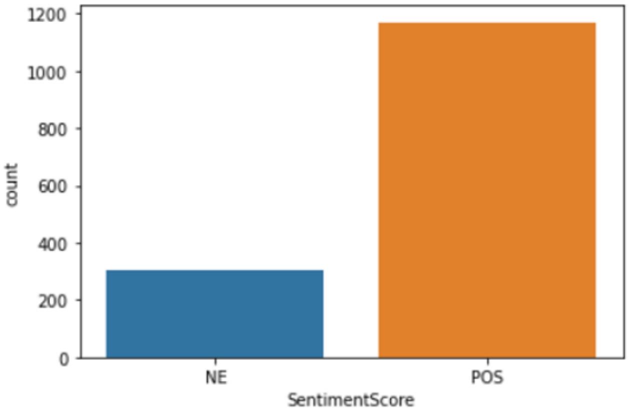


FIGURE 3.12: Original dataset

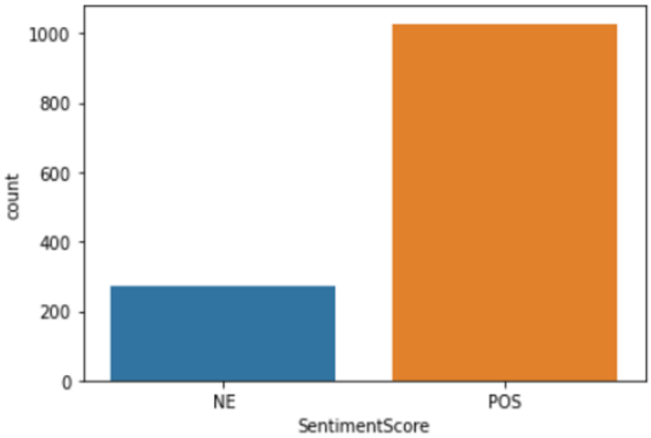


FIGURE 3.13: New Dataset (without duplication)

3.6.5 Data Balancing

One of the most typical approaches of dealing with an imbalanced dataset is to resample the data. Undersampling and oversampling are the two most common ways for this. Oversampling techniques are preferred over undersampling techniques in most circumstances. The reason for this is that when we undersample data, we tend to remove instances that may contain important information. SMOTE (Synthetic Minority Oversampling Technique) is an oversampling technique in which synthetic samples for the minority class are generated. This approach aids in overcoming the problem of overfitting caused by random oversampling. It concentrates on the feature space in order to produce new examples by interpolating between minority class instances that are close together. The total number of oversampling observations, N, is put up first. It is usually chosen so that the binary class distribution is 1:1. However, depending on the situation, this could be tuned. The iteration then begins with a random selection of a minority class instance. The KNNs (by default 5) for that instance are then retrieved. Finally, N of these K instances are chosen as the basis for creating new synthetic instances. To do so, the difference in distance between the feature vector and its neighbors is calculated using any distance metric. This difference is now multiplied by any random number in the range (0,1) and added to the prior feature vector. Oversampling is illustrated in Fig 3.14.

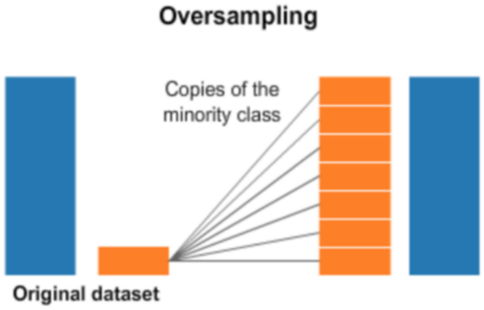


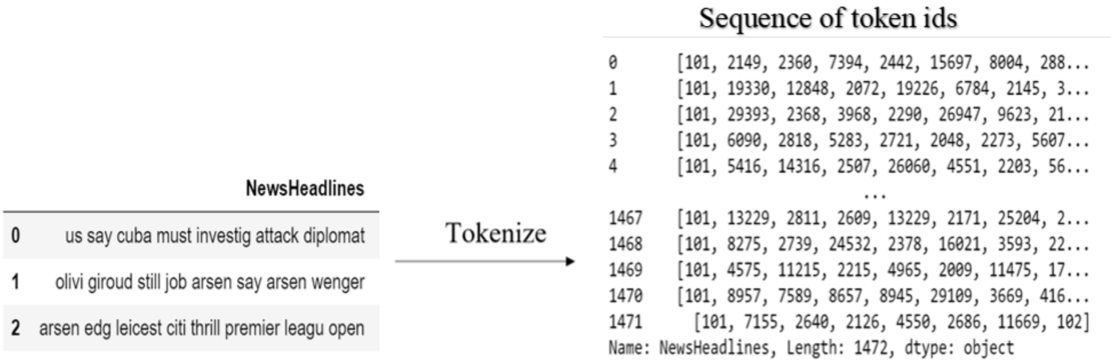
Figure 3.14: Oversampling

3.6.6 BERT

The first step is to divide the word into tokens using the BERT tokenizer. The special tokens required for sentence categorization are then added (these are [CLS] inserted at the first position, and [SEP] token inserted at the end of the sentence). Each token is replaced by an id from the embedding table, which is a component acquired with the trained model, in the third phase. Input sentences have now been transformed into the right shape to be provided to Bert's model, shown in figure 3.15.

The dataset is a pandas Series or a list of lists. Before BERT can process this as input, all of the vectors must be the same size, which can be accomplished by padding shorter words with the token id 0. It now has a matrix/tensor that can be provided to BERT after the padding. Matrix after padding is shown in below figure 3.16.

After running the BERT model, the output is a tuple with the shape number of headlines, maximum number of tokens in the sequence, and number of hidden

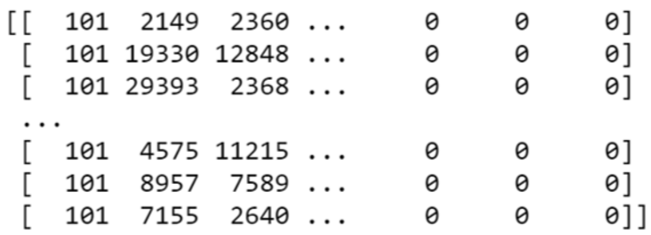


	NewsHeadlines
0	us say cuba must investig attack diplomat
1	olivi giroud still job arsen say arsen wenger
2	arsen edg leicest citi thrill premier leagu open

Sequence of token ids

```
0       [101, 2149, 2360, 7394, 2442, 15697, 8004, 288...
1       [101, 19330, 12848, 2072, 19226, 6784, 2145, 3...
2       [101, 29393, 2368, 3968, 2290, 26947, 9623, 21...
3       [101, 6090, 2818, 5283, 2721, 2048, 2273, 5607...
4       [101, 5416, 14316, 2507, 26060, 4551, 2203, 56...
                              ...
1467    [101, 13229, 2811, 2609, 13229, 2171, 25204, 2...
1468    [101, 8275, 2739, 24532, 2378, 16021, 3593, 22...
1469    [101, 4575, 11215, 2215, 4965, 2009, 11475, 17...
1470    [101, 8957, 7589, 8657, 8945, 29109, 3669, 416...
1471       [101, 7155, 2640, 2126, 4550, 2686, 11669, 102]
Name: NewsHeadlines, Length: 1472, dtype: object
```

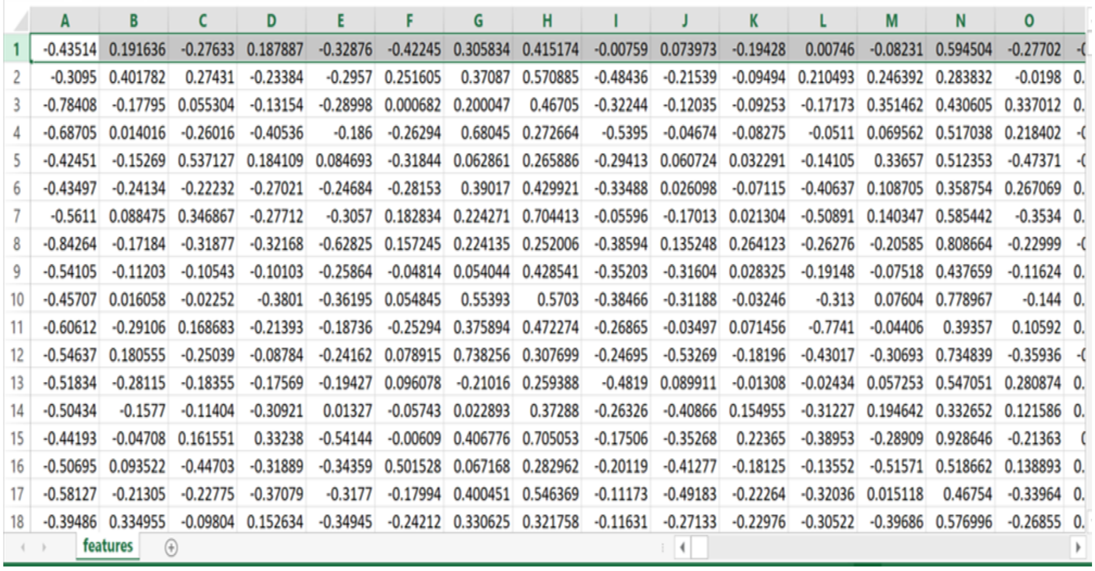
FIGURE 3.15: BERT tokenization



```
[[  101  2149  2360 ...     0     0     0]
 [  101 19330 12848 ...     0     0     0]
 [  101 29393  2368 ...     0     0     0]
 ...
 [  101  4575 11215 ...     0     0     0]
 [  101  8957  7589 ...     0     0     0]
 [  101  7155  2640 ...     0     0     0]]
```

FIGURE 3.16: Padding token-id

units. In this output, these are 1472 headlines, 37 (which is the number of tokens in the longest sequence from the 1472 examples) and 768 (the number of hidden units in the BERT model). 3d tensor is sliced to get the 2d tensor. And now features is a 2d numpy array containing the sentence embeddings of all the news headlines in our dataset. The features obtained by BERT are shown in figure 3.17 below.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	-0.43514	0.191636	-0.27633	0.187887	-0.32876	-0.42245	0.305834	0.415174	-0.00759	0.073973	-0.19428	0.00746	-0.08231	0.594504	-0.27702	-(
2	-0.3095	0.401782	0.27431	-0.23384	-0.2957	0.251605	0.37087	0.570885	-0.48436	-0.21539	-0.09494	0.210493	0.246392	0.283832	-0.0198	0.
3	-0.78408	-0.17795	0.055304	-0.13154	-0.28998	0.000682	0.200047	0.46705	-0.32244	-0.12035	-0.09253	-0.17173	0.351462	0.430605	0.337012	0.
4	-0.68705	0.014016	-0.26016	-0.40536	-0.186	-0.26294	0.68045	0.272664	-0.5395	-0.04674	-0.08275	-0.0511	0.069562	0.517038	0.218402	-(
5	-0.42451	-0.15269	0.537127	0.184109	0.084693	-0.31844	0.062861	0.265886	-0.29413	0.060724	0.032291	-0.14105	0.33657	0.512353	-0.47371	-(
6	-0.43497	-0.24134	-0.22232	-0.27021	-0.24684	-0.28153	0.39017	0.429921	-0.33488	0.026098	-0.07115	-0.40637	0.108705	0.358754	0.267069	0.
7	-0.5611	0.088475	0.346867	-0.27712	-0.3057	0.182834	0.224271	0.704413	-0.05596	-0.17013	0.021304	-0.50891	0.140347	0.585442	-0.3534	0.
8	-0.84264	-0.17184	-0.31877	-0.32168	-0.62825	0.157245	0.224135	0.252006	-0.38594	0.135248	0.264123	-0.26276	-0.20585	0.808664	-0.22999	-(
9	-0.54105	-0.11203	-0.10543	-0.10103	-0.25864	-0.04814	0.054044	0.428541	-0.35203	-0.31604	0.028325	-0.19148	-0.07518	0.437659	-0.11624	0.
10	-0.45707	0.016058	-0.02252	-0.3801	-0.36195	0.054845	0.55393	0.5703	-0.38466	-0.31188	-0.03246	-0.313	0.07604	0.778967	-0.144	0.
11	-0.60612	-0.29106	0.168683	-0.21393	-0.18736	-0.25294	0.375894	0.472274	-0.26865	-0.03497	0.071456	-0.7741	-0.04406	0.39357	0.10592	0.
12	-0.54637	0.180555	-0.25039	-0.08784	-0.24162	0.078915	0.738256	0.307699	-0.24695	-0.53269	-0.18196	-0.43017	-0.30693	0.734839	-0.35936	-(
13	-0.51834	-0.28115	-0.18355	-0.17569	-0.19427	0.096078	-0.21016	0.259388	-0.4819	0.089911	-0.01308	-0.02434	0.057253	0.547051	0.280874	0.
14	-0.50434	-0.1577	-0.11404	-0.30921	0.01327	-0.05743	0.022893	0.37288	-0.26326	-0.40866	0.154955	-0.31227	0.194642	0.332652	0.121586	0.
15	-0.44193	-0.04708	0.161551	0.33238	-0.54144	-0.00609	0.406776	0.705053	-0.17506	-0.35268	0.22365	-0.38953	-0.28909	0.928646	-0.21363	(
16	-0.50695	0.093522	-0.44703	-0.31889	-0.34359	0.501528	0.067168	0.282962	-0.20119	-0.41277	-0.18125	-0.13552	-0.51571	0.518662	0.138893	0.
17	-0.58127	-0.21305	-0.22775	-0.37079	-0.3177	-0.17994	0.400451	0.546369	-0.11173	-0.49183	-0.22264	-0.32036	0.015118	0.46754	-0.33964	0.
18	-0.39486	0.334955	-0.09804	0.152634	-0.34945	-0.24212	0.330625	0.321758	-0.11631	-0.27133	-0.22976	-0.30522	-0.39686	0.576996	-0.26855	0.

FIGURE 3.17: BERT features

3.6.6.1 BERT Working Example

- Sentence: US says Cuba must investigate attacks on diplomats
- [CLS] us say cuba must investig attack diplomat [SEP]
- [’[CLS]’, ’us’, ’say’, ’cuba’, ’must’, ’invest’, ’ig’, ’attack’, ’diplomat’, ’[SEP]’]
- [101, 2149, 2360, 7394, 2442, 15697, 8004, 2886, 11125, 102]
- [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
- Number of layers: 13 layers ,12 are BERT layers with 1 initial embedding
- Number of batches: it is 1 because it has one sentence
- Number of tokens: 10, as it contains 10 tokens
- Number of hidden units: 768
- Final sentence embedding vector of shape: torch. Size([768])10

As mentioned above there are two basic models of BERT. Which are BERT base and BERT large. Bert-base is used in this research. There are many strategies to utilize Bert layers for achieving good results like using all layers, taking average of second to last hidden layers but he results depend on the type of task and type of data. This thesis have experimented with multiple strategies to utilize layers but the best results achieved by concatenating last four hidden layers.

3.6.7 Classifier

Now that the output of BERT is ready, the dataset is assembled there is need to train the model. The features are in 768 columns, and the labels are from the initial dataset. Initially, supervised learning algorithms i.e. Gaussian Naive bayes, Decision tree, Random forest, K-neighbours and SVM were employed to build classification models, but the best results were obtained by SVM. Experiments are also done by using SVM with different kernels i.e. Polynomial, RBF, and linear kernel the best results were obtained by linera SVM. Results are shown in chapter 4. In order to develop the model, 80 percent of the data is used for the training set and 20 percent for the evaluation set. As the SVM model is declared to train it against the dataset after doing the standard train/test split of machine learning. The kernel is linear, hence SVM generates a linear hyper plane that separates words. It seperates negative words from positive words. SVM model is trained on the training set. Now that the model is trained, it can be scored against the test set.

3.6.8 PCA

PCA is a method for performing unsupervised linear transformations that is frequently utilized in many fields as a feature extraction and dimensionality reduction application. The dimensions of the features could be quite large, making calculations incredibly expensive. In many cases, dimensionality reduction can help. Whereas we derive information from the feature set to create a new feature subspace. The goal of principal component analysis is to reduce a high-dimensional input to a low-dimensional one. That low-dimensional input will eventually be used in a model, which we will train with the training data and test with the testing data. It is sufficient to include enough principal components to represent approximately (70-80%) of the data variation [18]. The reduced dimension dataset makes it simple for users to interpret, analyze, and handle data. To begin, we'll select how many of the 768 principal components we wish to keep in our model. The cost and accuracy of a system with more components rises. We did various experiments by iterating n-components e.g. the value of n-component changes at each iteration and on each iteration it shows the results i.e. accuracy, precision, recall and F1-score for both classes. At some point its results are repeating then we stop, and select the value of n-component which gives the highest accuracy with precision, recall and F1-score. Initially experiments were done by using PCA and wrapper method. The best results were obtained by PCA. Results are shown in chapter 4.

3.7 Research Methodology for RQ2

This thesis, built a system to compare the performance of governments and rank them on the basis of provided keyword. The study attempt to compare the performances of various regimes in Pakistan by doing sentiment analysis on news dataset of their relevant time periods. This dataset is divided into three files. 1) Containing headlines from PPP government, 2) containing headlines from PMLN government,3) containing headlines from PTI government.

An interface is being built where a user provides a keyword and system searches for that keyword in the dataset and matches it with headlines. After getting headlines it finds the sentiment polarity. It provides analysis that in which Prime Minister's time period there were more positive news and in which Prime Minister's time period there were more negative news on that particular topic and rank accordingly.

Chapter 4

Result and Evaluation

The proposed methodology was explained in detail in Chapter 3. The proposed methodology is used to obtain the findings. The results of data extraction, preprocessing, vector representation and feature reduction have been presented in this chapter. The supervised approach, on the other hand, require extensive preprocessing of the training data. It requires both the training and test data to be represented in some way. The representation could be taken in the form of a bag of words, or a feature-based representation. We have used Bert for feature representation.

4.1 Supervised Approach

Supervised machine learning approach is employed in this section. A three-step process was employed for experimentation with proposed approach as well as the identification of factors that affect results. Data collection, preprocessing training the data is the initial stage. The second step is data representation and feature extraction. The third step is to train and test classifier. By altering the preprocessing steps, feature reduction, and then machine learning techniques, this approach allows us to experiment with different possibilities. Each of them are explained in the following subsections.

The evaluation of the proposed technique is based on the dataset. The details

Table 4.1: Base dataset details

	Original dataset	New dataset (without duplicates)
Count	1472	1302
Positive	1170	1028
Negative	302	274

about the dataset are already explained in chapter 3. For experimentation, the base paper dataset (dataset-1) and the Pakistani news headlines dataset(dataset-2) were chosen.

4.1.1 Dataset-1

There are 1472 news headlines in the base paper dataset manually annotated by humans. While analyzing the dataset it is observed that it contains 170 duplicate records. Out of 170 records there are 28 negative records and 142 positive. Duplicates effect model performance, to test it, this study devised a controlled experiment. This might be done by comparing the performance of the raw dataset to the dataset with duplicates removed. Therefore those duplicates have been removed from the dataset. It is experimented on both variations with original data as well as on the dataset with removed duplicates to see the effect on results. Also it is noticed that the dataset is imbalanced. This study balanced the original dataset as it have 1170 positive instances and 302 negative instances. We have also balanced the new dataset having no duplicates. To see the effect of class imbalance on results, we experimented on both variations with original data and on the dataset with eliminated duplicates. The count of headlines is illustrated in Table 4.1.

4.1.2 Dataset-2

For experimentation purpose, the dataset of the first two months of three different regimes are used. There were 1500 unlabeled news headlines which were selected

Table 4.2: Pakistani News headlines dataset details

	Data Annotated	After eliminating ($\text{Label}_2 = NO$)
Count	1500	948
Positive	643	487
Negative	857	461

for annotation. An annotation scheme have been derived to provide enough information, for the process of manual labeling. By following the provided annotation scheme four human annotators had labeled the dataset. There were two columns to tag against each headline. Details are given in chap 3.

After getting the labeled dataset it has discarded those instances which were tagged "NO" in Label2 column. As we have to evaluate performance of governments therefore we only kept the instances that were tagged as "YES" in Label2 column. After cleaning the dataset there are 984 news headlines in total. Contains 487 positive and 461 negative news headlines, starting of two month durations of three governments and their count is illustrated in table 4.2.

4.1.3 Pre-processing

Next step is to pre-process the data because the data needs to be cleaned. These steps are performed in preprocessing:

- The first step is to use the panda's package to read the dataset from a CSV file.
- Then, convert whole dataset in lower case.
- Use the NLTK Library to tokenize the text of headlines, i.e. tokenize.
- Eliminating noise from all headlines, such as punctuation, numbers, and so on.
- Using NLTK Stop words, remove all stop words from all headlines
- NLTK stemmer, i.e. Porter Stemmer, is used to stem the text of all headlines.
- Reassemble the words and save them in a file.

4.1.4 Text Representation

The data in a CSV file has been successfully pre-processed, and it is now ready for vector representation. The vectorization of text in this study is done with BERT, which takes into account the semantic meaning of terms. This process includes the following steps: **Input:** In comparison to other techniques, the BERT's input is slightly different.

- The first step is to divide the word into tokens using the BERT tokenizer.
- The special tokens required for sentence categorization are then added these are [CLS] and [SEP] tokens. [CLS] is inserted at the beginning of the sentence and [SEP] at the end.
- [CLS] is the token that appears at the start of the sentence, and this token is especially used to perform the tasks of classification.
- Each token is replaced by an id from the embedding table, which is a component acquired with the trained model.
- Input sentences are now in the correct format to be delivered to Bert.

4.1.5 BERT

This thesis was successfully able to convert the text of all features into vectors using this approach. Different combinations of BERT layers can be utilized for embedding it depends on the task and the type of data. This study has used embeddings of second to last layer.

4.1.6 Classification

After vector representation, classification is the next step. To develop classification models, supervised learning algorithms such as Gaussian Naive bayes, Decision tree, Random forest, K-neighbours and SVM were used at first, but SVM produced the best results. Experiments were also conducted using SVM with various kernels, including polynomial, RBF and linear kernels. With linear SVM it yields the best results. Results are shown in table 4.3. The SVM classifier is used here for classification. First data is being divided into two parts, one is training set and the other is testing set. Classifier first trains the model on training set and when the training phase is over it predicts the labels on test data. Here, for building the model, 20% data is considered for model evaluation and 80% data is considered for training the model.

TABLE 4.3: Result comparison of different classifiers

Classifier	**Accuracy**
SGD Classifier	0.72%
Guassian NB Classifier	0.73%
K-Neighbors Classifier	0.85%
Decision Tree Classifier	0.74%
Random Forest Classifier	0.82%
SVM: Kernel(linear)	0.75%
SVM: Kernel(poly)	0.82%
SVM: Kernel(rbf)	0.82%
SVM: linear SVC()	0.88%

4.1.7 Feature Reduction

This study have to choose how many of the 768 main components to include in a model. A lot of feature subset are possible which define different combinations of features. A system with more components increases in cost and accuracy. Several experiments were conducted by iterating n-components, in which the value of the n-component varies with each iteration, and the outcomes, i.e. accuracy, precision, recall, and F1-score for both classes, are displayed for each iteration. When the results start to repeat themselves, it stops and choose the n-component value that yields the best accuracy, precision, recall, and F1-score. Here the best accuracy is achieved when the number of n-components is 26. Figure 4.1 shows how accuracy of the algorithm fluctuate as the number of iterations increases.

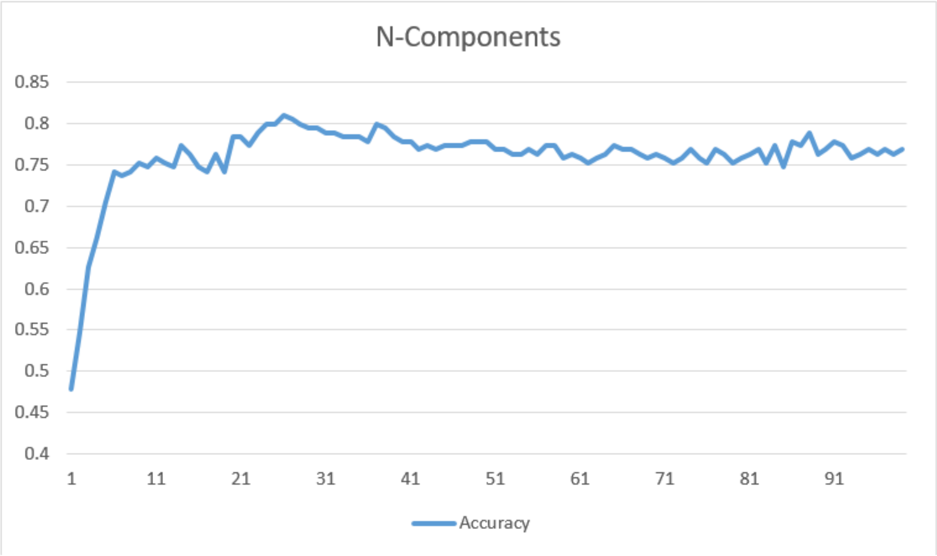


FIGURE 4.1: Number of iterations vs accuracy

4.2 Evaluation Methods and Metrics

Classification is a commonly used approach in machine learning problems with a wide range of industrial applications, ranging from facial recognition to medical diagnosis. Text classification, hate speech detection on twitter or sentiment analysis. Some of the most prominent categorization models include support vector machines (SVM), logistic regression and decision trees. A classification model can be evaluated in a number of ways.

4.2.1 Classification Accuracy

The most important classification metric is accuracy. It is simple and easy to understand. And it's well-suited to both binary and multiclass classification problems. Accuracy is defined as the proportion of true outcomes among the total number of instances analysed. Here a question arises, when to use it? Well the answer is: Accuracy is a good choice of evaluation for classification problems that are well balanced and not skewed, or have no class imbalance. Although a model can be reasonably accurate, but it is useless if the target class is very small.

$$Accuracy = \frac{True\,Positive + True\,Negative}{True\,Positive + True\,Negative + False\,Positive + False\,Negative} \tag{4.1}$$

4.2.2 F1-Score

This is best evaluation metric for classification Models. The F1 score is the harmonic mean of precision and recall, and it ranges from 0 to 1. It's used when we need a model this is both having good precision and good recall. In Simply words, for classifier the F1 score maintains a balance between precision and recall. The F1 score is low if precision is low, and F1 score is also low if recall is low.

$$F1 - Score = \frac{2(Precision * Recall)}{Precision + Recall} \tag{4.2}$$

4.2.3 Precision

Precision and recall are two metrics that are used to examine the performance of classification or retrieval systems. The fraction of relevant instances among all retrieved instances is defined as precision. It refers to how precise/accurate a model is in terms of how many of those predicted positives are actually positive. It is calculated as:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{4.3}$$

4.2.4 Recall

The fraction of retrieved instances among all relevant examples is known as recall, or sensitivity. Precision and recall are both equal to 1 in a good classifier. It is calculated as:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{4.4}$$

4.3 Evaluation Metrics and Methods for Solution

Evaluation Measures we use for evaluating proposed solution are F1-score and accuracy which are the key classification metrics. The F1-score is the harmonic mean of precision (the number of correctly identified positive results divided by the total number of positive results including those that are incorrectly identified) and recall (the number of positive results correctly identified divided by the total number of positive results including those that are incorrectly identified) (the number of correctly identified positive results divided by the number of all elements that are really positive). It has a best score of 1 and a worst score of 0. Precision and recall both contribute equally to the F1-score, whereas accuracy is defined as the ratio of correct predictions to total input components.

4.3.1 Proposed Approach: Dataset 1

Original dataset [1], total number of headlines 1472. Contains 1170 positive instances and 302 negative instances. It is observed that it contains 170 duplicate records. After removing duplicate records total number of headlines are 1302. It has 1028 positive records and 247 negative records. It is also observed that the dataset is imbalance, both datasets containing (1472) and (1302) news headlines were balanced for experimentation.

4.3.2 Proposed Approach: Dataset 2

This dataset contains 984 Pakistani news headlines. Contains 487 positive and 461 negative news headlines. Approach of Rameshbhai et, al.[1] have applied uni-gram, bi-gram. They claim that they have achieved 87 to 91% accuracy on their approach. After implementing the base approach result obtained from both models MODEL A and MODEL B accuracy achieved is 84 to 85%.

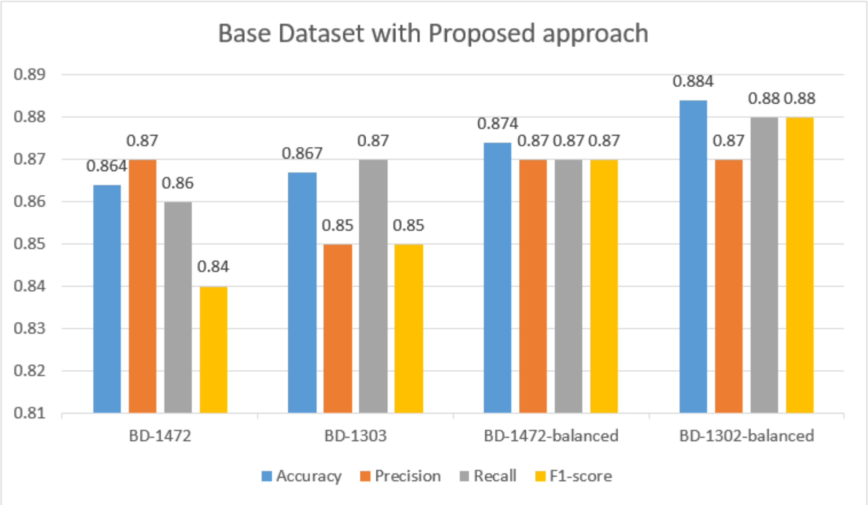


FIGURE 4.2: Experiments: Dataset 1

4.3.2.1 Result Discussion

Base dataset contains headlines with duplicate records. It might lead to erroneous conclusions by making us assume that some observations are more dominant than

Table 4.4: Base Dataset Results

	Dataset	**Result**	
Base approach	News headlines (1472)	ModelA Unigram: 63% Bigram: 65%	ModelB Unigram: 65% Bigram:66%
BERT with PCA	News Headlines (1472)	0.864%	
BERT with PCA	No Duplication (1303)	0.867%	
BERT with PCA	Balanced (1472)	0.874%	
BERT with PCA	Balanced (1303)	0.884%	
Word2Vec with PCA	No Duplication (1303)	0.862%	
BERT with Wrapper method	No Duplication (1303)	0.77%	

they are. Duplication increases data redundancy, which can lead to statistical bias and affect the outcomes of experiments. Data cleaning is critically an important step in any machine learning task. By recognizing and eliminating rows containing duplicate data, machine learning algorithms perform better. Duplicate rows results in misleading performance. Proposed approach out performs base results, in addition experiments are performed by removing duplication it results in increase of approximately 3% accuracy. By analyzing the data more deeply it is observed that data is imbalance. Due to that, high level of accuracy can be simply achieved by forecasting the majority class, but it misses the minority class, when the class distribution is similar, accuracy can be employed, but F1-score is a better statistic when there are imbalanced classes, like in this situation. It can be clearly seen through the figure 4.2. In figure 4.2 accuracy is approximately equal when the dataset is imbalance as well as balanced but F1-score is highest when the dataset is balanced and have low value for dataset imbalance. Results are computed based on base dataset with BERT embeddings and PCA. Experiments are also performed

Table 4.5: Pakistani News dataset Results

	Dataset	**Model A**	**Model B**
Base approach	News headlines ofthree regimes (948)	Unigram: 63% Bigram: 65%	Unigram: 65% Bigram: 66%
Our Approach	News headlines ofthree regimes (948)	81%	

by using BERT embeddings with wrapper method. Here we used 5 fold cross validation with feature combination set between 1 to 30. Table 4.3 shows the results of above mentioned experiments done on base dataset. Moreover, Pakistani news dataset is used to evaluate the performance of governments. Since the dataset has roughly balanced number of samples of all classes, this study directly use the accuracy measure to evaluate the performance of proposed model and compare it with other models. The proposed approach is compared with the approaches proposed in the literature. In the literature, researchers have utilized count based approaches for feature representation. Figure 4.3 illustrates the comparison. Results on Pakistani dataset is shown in table **??**. The results of proposed approach outperformed the previous techniques by achieving accuracy 88% on base dataset and 81% on Pakistani news headlines dataset with increase of 15% accuracy.

4.3.3 Performance Ranking of Governments

To evaluate the performance of governments, data of first two months of each regime is used. After cleaning the data this study only considered those instances that were tagged “YES” in label two column during Annotation. As these instances contribute in evaluating government performance. Percentages are calculated of instances having positive and negative sentiment for each regime. After experimentation on pakistani news dataset of three regimes it is indicated that PPP(Pakistan Peoples Party) have 40.50% positive and 59.48% negative instances. PMLN(Pakistan Muslim League Noon) have 51.20% positive and 48.79% negative instances while PTI(Pakistan Tehreek Insaaf) have the highest positive score 63.15% and lowest negative score 36.80%. Governments are ranked by computing the impact of positive and negative scores such as adding positive score of PPP to negative score the value is (-18.98) and adding positive score of PMLN to its negative score the value is(2.41) and the impact value for PTI is (26.35) which is the highest among all. Therefore, PTI is ranked 1, PMLN is ranked 2 and PPP is ranked 3 as it has more negative impact as compared to other two regimes. The ranking comparison is shown in the Figure 4.4.

To see how Two News websites “The Dawn” and “The Nation” rank these governments an other experiment has been done. It is indicated that both Websites have more positive scores about PTI than negatives. Both news websites have high negative score for PMLN as compare to positive ones. There are more negative scores and less positive scores for PPP than PMLN. After computing the impact of positive and negative scores such as adding positive score of PPP to negative score for both news websites results are (-8.33 and -20.61). The values (-4.85 and -1.59) are computed by adding positive score of PMLN to its negative score. The resultant values for PTI are (32.11 and 11.63) which are the highest values among

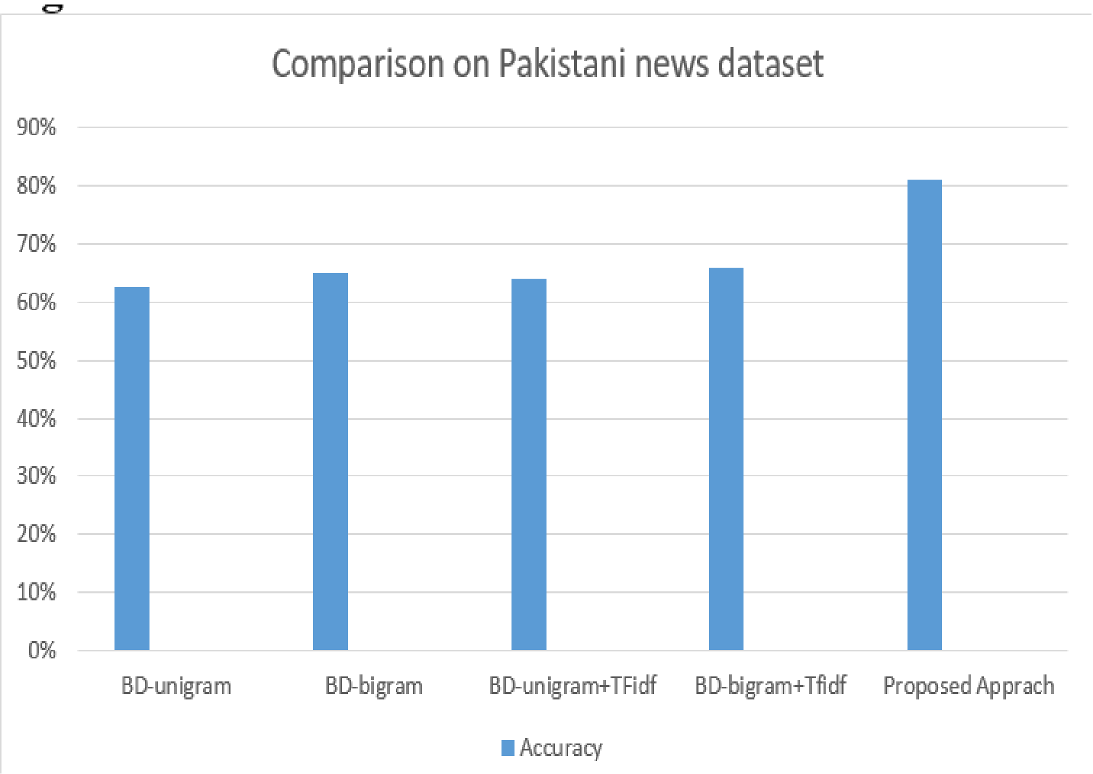


Figure 4.3: Comparison on Pakistani Dataset

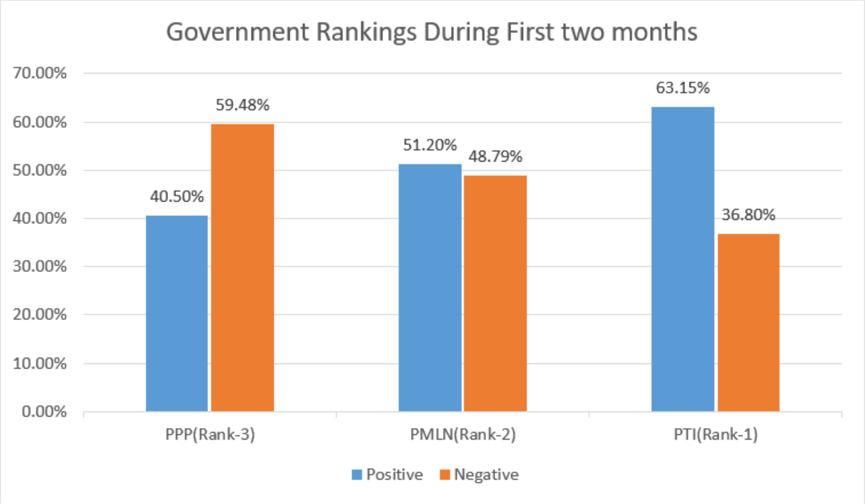


FIGURE 4.4: Ranking

all. Therefore, PTI is ranked 1, PMLN is ranked 2 and PPP is ranked 3 as it has more negative impact as compared to other two regimes. The ranking is shown in the Figure 4.5.

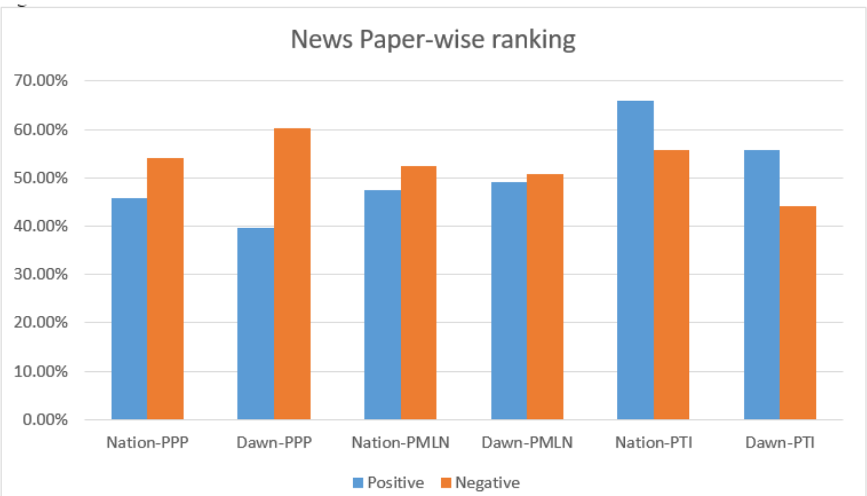


FIGURE 4.5: Ranking

After experimentation regime wise as well as News paper wise it is concluded that PTI have the highest positive rate which lead this study to Rank it 1 and PMLN is ranked 2, PPP is ranked as 3. To summarize our findings, the rankings under sentiment analysis methodology reflects that during first two months of successive governments, the PTI government had better impression or positive coverage in the National press. There could be different reasons for this, like, due to a better pre-election compaign of PTI, or due to bad performance of the previous government, or due to personality charisma of PTI leader or may be some other reason. The worst first two months coverage is for the PPP government, that can possibly be due to the bad situation of law and order in the country, murder of Benazir Bhutto, or dictatorial regime previously or may be due to any other reason. The picture can be further improved if we could process the data of later months as well. Thus by ranking government's performance based on sentiment polarity, various political regimes can be evaluated and ranked. These analyses and rankings serve as benchmarks for future governments to improve. The idea is very useful it can be useful to guide or educate the people on different aspects of government. It has ability to generalize in multiple directions. More news headlines from different news websites can be added. It can be used as an international tool.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Sentiment analysis of news headlines is a big challenge for researchers. It is very important for retrieving relevant information, and for analyzing the sentiments of public towards governments. In literature, there are many techniques proposed for classifying the news headlines into one of the predefined categories. Mostly, researchers have performed sentiment analysis on twitter, Facebook and news headlines. The main advantage of using news headlines dataset is one can get the idea of the main areas such as entertainment, politics, sports, and technology by only analyzing news headlines without going through the whole story/paragraphs or articles.

The beauty of this study is that it combines text mining and sentiment analysis, which are normally treated as separate topics; however, this work has combined the two into one. This study employed a systematic approach. It includes a text analysis of news headlines before doing sentiment analysis on them. The data set that was used in the study is created and interpreted carefully. There are two types of approaches to sentiment analysis. Lexicon based approaches can be useful in identifying sentiment polarity of data this approach highly rely on pre-define list or dictionary of sentiment words built by others that associate words to their sentiments. But there is also a limitation of this approach that those dictionaries do not cover all the words in the English or other languages and the other limitation is it cannot manage properly negation of sentiment. Therefore it is important to know source and design of the dictionary to be used to decide whether it is appropriate for dataset and purposes or not. As an alternative way, researchers have utilized machine learning based approach. Machine learning approaches work on train test data. Using a sentiment-labeled training set, it train a machine learning model to recognize sentiment based on the words and their sequence. This is accomplished by extracting ”features” from the text, which are then used to predict a ”label.” Splitting the text into words and then using these words and their frequency in the text as features is an example of producing features. This method is highly dependent on the text representation, algorithm and the quality of the training data. The news headlines are in the form of text data. For classification of the headlines, text representation is an important task. In literature, researchers have utilized the statistical measures like TFIDF. These measures gain information using the frequency of terms. All these techniques do not take into account the semantics of the text to counter this, this study uses BERT model for feature representation.

In literature, most of the researchers have used all features. But with more number of features, its complexity increases. These all problems which are mentioned above led us to propose a solution. Data is extracted from two news websites “The Nation” and The Dawn” ranges from 2008 till 2021. This dataset is based on particular years of three regimes: Pakistan Tehreek Insaaf, Pakistan People’s Party, Pakistan Muslim league noon. After data extraction, preprocessing is performed on the dataset. The datasets were pre-processed before being sent to the machine learning classifier. The steps involved in preprocessing are 1) tokenize all the text into words, then 2) remove all the stop words, 3) and at last do stemming of the words. Now, the dataset is preprocessed, the next step is to do text representation. For representation of text, frequency based techniques have been used in literature but this study takes into account the semantics and context of the term used in the text. This study used Bert model for text representation and it is a bidirectional model as it reads text from both sides of the sentence. It is already pre-trained. This study has used pre-trained model and it generates vectors of the dataset.

Now the dataset is in the form of vectors and is ready for classification task. This study has performed classification using svm classifier. After classification of the news headlines, the next step is to do feature reduction. The PCA is used for reducing the features. The experimentation has done by evaluating all the features and the results achieved having accuracy 83.3%. After doing feature reduction by using PCA, the results achieved having accuracy 88.12% for the features subset. This study compared the proposed approach with the approach proposed by [1]. This approach used context based features and achieved an accuracy of 81.05% on the Pakistani news dataset. Moreover, for text representation both the semantics and context of the text are considered. The feature reduction is done for reducing its complexity.

5.2 Future Work

The results were mentioned in the previous section. However, when analyzing the data, there are certain important limitations to consider. Furthermore, the limitations of this thesis lay the groundwork for future research. As a result, this section elaborates the constraints and relates them to research suggestions for the future work. This study has identified some of the work which is to be done in future and it is described below:

1. The work of this study can be extended by evaluating proposed approach on large dataset.

2. This technique could be expanded and implemented by some other countries to evaluate their government's performance on their own dataset.

3. The scope is restricted to textual news data gathered from news websites. Additional data, such as photographs, sounds, internet articles, multimedia, and so on, could be added to this data collection in the future.

4. This study plans to investigate the datasets in depth in the future and compare them to other machine learning algorithms.

5. This study could be extended to make a semantic model which is trained on news headlines dataset.

Bibliography

[1] Rameshbhai, C. J., & Paulose, J. (2019). Opinion mining on newspaper headlines using SVM and NLP. International Journal of Electrical and Computer Engineering (IJECE), 9(3), 2152-2163.

[2] Taj, S., Shaikh, B. B., & Meghji, A. F. (2019, January). Sentiment analysis of news articles: a lexicon based approach. In 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) (pp. 1-5). IEEE.

[3] Sangam, S., & Shinde, S. (2019). Sentiment classification of social media reviews using an ensemble classifier. Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), 16(1), 355-363.

[4] Gaurav M Pai, Paramesha K, K C Ravi Shankar. (2018). Sentiment Analysis of News Articles using Probabilistic Topic Modeling. International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 5(4), 2394-2320.

[5] Ahmed, J., & Ahmed, M. (2020). A Framework for Sentiment Analysis of Online News Articles. International Journal on Emerging Technologies, 11(3): 267-274.

[6] Bharathan, K., & Varma, P. D. (2019, November). Polarity Detection Using Digital Media. In 2019 9th International Conference on Advances in Computing and Communication (ICACC) (pp. 181-187). IEEE.

[7] Shuhidan, S. M., Hamidi, S. R., Kazemian, S., Shuhidan, S. M., & Ismail, M. A. (2018, March). Sentiment analysis for financial news headlines using machine learning algorithm. In International Conference on Kansei Engineering & Emotion Research (pp. 64-72). Springer, Singapore.

[8] Gupta, S., Bolden, S., Kachhadia, J., Korsunska, A., & Stromer-Galley, J. (2020, October). PoliBERT: Classifying political social media messages with BERT. In Social, Cultural and Behavioral Modeling (SBP-BRIMS 2020) conference. Washington, DC.

[9] Sharma, A., & Ghose, U. (2020). Sentimental analysis of twitter data with respect to general elections in india. Procedia Computer Science, 173, 325-334.

[10] Munikar, M., Shakya, S., & Shrestha, A. (2019, November). Fine-grained sentiment classification using BERT. In 2019 Artificial Intelligence for Transforming Business and Society (AITB) (Vol. 1, pp. 1-5). IEEE.

[11] Rathod, S. L., & Deshmukh, S. N. (2016). Sentiment Analysis Using SVM and Maximum Entropy. International Research Journal of Engineering and Technology (IRJET), 3, 453-458.

[12] Dor, D. (2003). On newspaper headlines as relevance optimizers. Journal of pragmatics, 35(5), 695-721.

[13] Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using bert. In Proceedings of the 22nd Nordic Conference on Computational Linguistics (pp. 187-196).

[14] Kotelnikova, A. V. (2020, October). Comparison of Deep Learning and Rule-based Method for the Sentiment Analysis Task. In 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon) (pp. 1-6). IEEE.

[15] Mæhlum, P., Barnes, J. C., Øvrelid, L., & Velldal, E. (2019). Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian. In Linköping Electronic Conference Proceedings (pp. 121-130).

[16] Rasheed, F., Ahmad, E., & Kazmi, A. A. (2006). An Evaluation of the Performance of Government of Pakistan [with Comments]. The Pakistan Development Review, 831-841.

[17] Rana, M. I., Khalid, S., & Akbar, M. U. (2014, December). News classification based on their headlines: A review. In 17th IEEE International Multi Topic Conference 2014 (pp. 211-216). IEEE.

[18] Salem, N., & Hussein, S. (2019). Data dimensional reduction and principal components analysis. Procedia Computer Science, 163, 292-299.

[19] Kirange, D. K., & Deshmukh, R. R. (2012). Emotion classification of news headlines using SVM. Asian Journal of Computer Science and Information Technology, 5(2), 104-106.

[20] Tyagi, E., & Sharma, A. K. (2017). Sentiment analysis of product reviews using support vector machine learning algorithm. Indian J. Sci. Technol, 10(35), 1-9.

[21] Perikos, I., Hatzilygeroudis, I. (2017). Aspect based sentiment analysis in social media with classifier ensembles. In 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS) (pp. 273-278). IEEE.

[22] Mukherjee, S., Joshi, S. (2014). Author-Specific Sentiment Aggregation for Polarity Prediction of Reviews. In LREC (pp. 3092-3099).

[23] Diamantini, C., Mircoli, A., Potena, D. (2016, October). A negation handling technique for sentiment analysis. In 2016 international conference on collaboration technologies and systems (cts) (pp. 188-195). IEEE.

[24] Pota, M., Esposito, M., De Pietro, G. (2016). A forward-selection algorithm for SVM-based question classification in cognitive systems. In Intelligent Interactive Multimedia Systems and Services 2016 (pp. 587-598). Springer, Cham.

[25] ThelWall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A. (2011). Sentiment in Short Strength Detection Informal Text (vol 61, pg 2544, 2010). Journal of the American Society for Information Science and Technology, 62(2), 419-419.

[26] Berger, A., Della Pietra, S. A., Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. Computational linguistics, 22(1), 39-71.

[27] Levy, O., Goldberg, Y., Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. Transactions of the association for computational linguistics, 3, 211-225.

[28] Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[29] Schütze, H., Manning, C. D., Raghavan, P. (2008). Introduction to information retrieval (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.

[30] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

[31] Yang, Y., Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In Icml (Vol. 97, No. 412-420, p. 35).

[32] Akbik, A., Blythe, D., Vollgraf, R. (2018, August). Contextual string embeddings for sequence labeling. In Proceedings of the 27th international conference on computational linguistics (pp. 1638-1649).

[33] Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of EMNLP, pp. 79–86.

[34] Pang, B., & Lee, L. (2008, August). Using very simple statistics for review search: An exploration. In Coling 2008: Companion volume: Posters (pp. 75-78).

[35] Agarwal, A., Sharma, V., Sikka, G., & Dhir, R. (2016, March). Opinion mining of news headlines using SentiWordNet. In 2016 Symposium on Colossal Data Analysis and Networking (CDAN) (pp. 1-5). IEEE.

[36] Islam, M. U., Ashraf, F. B., Abir, A. I., Mottalib, M. A. (2017, December). Polarity detection of online news articles based on sentence structure and

dynamic dictionary. In 2017 20th International Conference of Computer and Information Technology (ICCIT) (pp. 1-5). IEEE.

[37] Raina, P. (2013, December). Sentiment analysis in news articles using sentic computing. In 2013 IEEE 13th International Conference on Data Mining Workshops (pp. 959-962). IEEE.

[38] Bhadane, C., Dalal, H., Doshi, H. (2015). Sentiment analysis: Measuring opinions. Procedia Computer Science, 45, 808-814.

[39] Li, J., Fong, S., Zhuang, Y., Khoury, R. (2016). Hierarchical classification in text mining for sentiment analysis of online news. Soft Computing, 20(9), 3411-3420.

[40] Swati, U., Pranali, C., Pragati, S. (2015). Sentiment analysis of news articles using machine learning approach. In Proceedings of 20th IRF International Conference,2, 114-116.

[41] Riloff, E., Patwardhan, S., & Wiebe, J. (2006, July). Feature subsumption for opinion analysis. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 440-448).

[42] Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics (pp. 841-847).

[43] Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528).

[44] Liu, B. (2011). Opinion mining and sentiment analysis. In Web Data Mining (pp. 459-526). Springer, Berlin, Heidelberg.

[45] Vinodhini, G., & Chandrasekaran, R. M. (2013). Effect of Feature Reduction in Sentiment analysis of online reviews. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2(6), 2165-2172.

[46] Meng, J., Lin, H., & Yu, Y. (2011). A two-stage feature selection method for text categorization. Computers & Mathematics with Applications, 62(7), 2793-2800.

[47] Shah, S., Shabbir, H., Rehman, S., & Waqas, M. (2020). A comparative study of feature selection approaches: 2016-2020. International journal of scientific and engineering research, 11(2), 469.

[48] Jotheeswaran, J., Loganathan, R., & Madhu Sudhanan, B. (2012). Feature reduction using principal component analysis for opinion mining. International Journal of Computer Science and Telecommunications, 3(5), 118-121.

[49] Che, L. (2019). Sentiment-based spatial-temporal event detection in social media data.

[50] Goldberg, Y. (2017). Neural network methods for natural language processing. Synthesis lectures on human language technologies, 10(1), 1-309.

[51] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

[52] Jain, T., Agrawal, N., Goyal, G., & Aggrawal, N. (2017, August). Sarcasm detection of tweets: A comparative study. In 2017 Tenth International Conference on Contemporary Computing (IC3) (pp. 1-6). IEEE.

[53] Deho, B. O., Agangiba, A. W., Aryeh, L. F., & Ansah, A. J. (2018, August). Sentiment analysis with word embedding. In 2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST) (pp. 1-4). IEEE.