

Ontology Based Knowledge System of Sentiment Analysis for Indian Railways Tweets

A Thesis Report

submitted in partial fulfillment of the requirements for the award of

Doctor of Philosophy

in

Computer Science and Engineering

by

Rakesh Kumar Donthi

Roll No: 155CS11 Enrolment No: 151066

Under the Supervision of

Dr. Md. Tanwir Uddin Haider

Associate Professor



Department of Computer Science and Engineering
National Institute of Technology Patna
July 2020

CERTIFICATE from the SUPERVISOR

This is to certify that Mr. **Rakesh Kumar Donthi** Roll No. **155CS11** Enrolment No. **151066** is a registered candidate for Ph.D. Programme under department of **Computer Science & Enginnering** of National Institute of Technology Patna.

The undersigned certify that he has completed all other requirements for submission of the thesis and hereby recommend for the acceptance of the thesis entitled **Ontology Based Knowledge System of Sentiment Analysis for Indian Railways Tweets** in the partial fulfilment of the requirements for the award of Ph. D. Degree by National Institute of Technology Patna.

Date:

Dr. Md. Tanwir Uddin Haider
Supervisor

DECLARATION AND COPYRIGHT

I, **Rakesh Kumar Donthi**, Roll No. **155CS11** Enrolment No. **151066**, a registered candidate for Ph.D. Programme under department of **Computer Science & Engineering** of National Institute of Technology Patna declare that this is my own original work and that it has not been presented and will not be presented to any other University/Institute for a similar or any other Degree award.

Place:

(Rakesh Kumar Donthi)

Date:

This thesis is a copy right material protected under the Berne Convention, the copyright at 1999 and other International and National enactments, in that behalf, or intellectual property. It may not be reproduced by any means, in full or in part, except for short extracts in fair dealing, for research or private study, critical scholarly review or discouser with an acknowledgment, without written permission of the Department on both the author and NIT Patna.

ACKNOWLEDGMENTS

At the event of submitting the thesis for fulfilment of the requirement of the award of "Doctor of Philosophy", I take this opportunity to express my sincere gratitude to all the people, who are involved either directly or indirectly in completion of my PhD work.

I would never have been able to finish my thesis without the guidance of my guide, help from friends and faculty members, and support from my family.

It's my privilege and honour to work under the supervision of **Dr. Md. Tanwir Uddin Haider**, Associate Professor, Department of Computer Science & Engineering, National Institute of Technology Patna, Bihar, India. I would like to express my sincere gratitude to my supervisor Dr. Md. Tanwir Uddin Haider, for their excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. Their guidance has helped me at all times during my research and writing of this thesis. I could not have imagined having better mentor for my Ph.D. research. I am very thankful to him for their constant encouragement and support.

I am extremely grateful to our Head of Department and PhD Coordinator for his support and encouragement during the course of my research.

I am extremely grateful to **Prof. Ashok De**, Ex-Director National Institute of Technology Patna, who permitted me to register to the doctoral programme and would like to give special thanks to **Prof. P. K. Jain**, Director, National Institute of Technology Patna, for their constant support and encouragement throughout my career.

I would like to thank Ministry of Electronics & Information Technology (MeitY), Government of India for providing the financial assistance during my research work through "Visvesvaraya PhD Scheme for Electronics & IT". I would like to convey my special thanks to our Nodal officer for managing assistantship throughout the PhD programme.

I would like to thank my fellow research scholars at NIT Patna, who were always willing to help and give their best suggestions for research and while writing this thesis. My research would not have been possible without their help and support.

I would also like to thank all faculty members of our department, for their valuable suggestion during my research career.

I deeply acknowledge the love, cooperation, and the moral support extended by my supervisor, friends, relatives and family members right from the beginning of my Ph.D. study. My research work would not have been possible without the support of my Mother and Father in particular. Both of them kept me aloof from all the social responsibilities towards my family and relatives. They were always supporting me and encouraging me with their best wishes. I am really thankful to my family, friends and guide.

Finally, I must thank the almighty for guiding and blessing me in this endeavor.

Rakesh Kumar Donthi

SYNOPSIS

The emergence of social networking and virtual communities over the INTERNET has paved the way for people to express their opinions on products and services. Thus social media such as Facebook, Twitter, Micro-blogging are providing reviews or opinions. In other words, social media is rich in possession opinion of the public. With the invention of Online Social Networking (OSN) and other World Wide Web (WWW) applications that provide user interaction, there is an increased probability of obtaining and studying opinions of people. There is an unprecedented increase in the popularity of social media such as Google Plus, Twitter and Facebook over the Internet. There is also an increase in information sharing and instant communication among users across the globe, where people use synonyms of many different words to express their views or opinion. Twitter is one of the Social networks that exhibit exponential growth of users and tweets every year.

Tweets carry social feedback on different products or services catering to plenty of domains. Moreover, the opinions of people in the tweets do have the capability to influence the decision making of the public. For the opinion of the people in social media, we perform opinion mining or sentiment analysis. The sentiment analysis refers to the use of Natural Language Processing (NLP), text analysis and computational linguistics to identify and extract subjective information in the source material to study people's opinions, attitudes and emotions towards an entity with the computational study. It is widely applied to reviews over social media for a variety of applications ranging from marketing to customer services such as mobile users, movie reviews, postal services, and online shopping reviews, etc.

But in our research work, we have done sentiment analysis on Indian Railways (IR) data sets extracted from Indian Railways official twitter handler. The analysis has been done on different features of railways such as Punctuality, Staff Behaviour, Security, Cleanliness, and Food Quality. These analysis will help the passengers and organization of Indian Railways. Passengers will take decision while doing reservation for their respective journey not only on the availability of the seats but also based on the performance of above features whereas this analysis will also help the organization to improve the quality of the said features. In our research work first extraction of data has been done from twitter. For

extracting of dataset we have used the technique called Web Scraping in Python with BeautifulSoup.

Web Scraping is the defined process of downloading data from websites and extracting valuable information from that data. Whereas BeautifulSoup is a Python package for parsing HTML and XML documents (including malformed mark-up, i.e. non-closed tags). It creates a parse tree for parsed pages that can be used to extract data from XML and HTML, which is useful for Web Scraping. It generates the parse tree to provide navigating, searching and modifying. It is less time consuming so it saves programmer time. Next, the extracted data has been pre-processed using Natural Language Pre-processing (NLP) techniques. Pre-processing is the process of converting data to something that a computer can understand. Further, to bring data sets into a finer state for better performance we have introduced multilevel filtering and topic-based filtering approaches other than NLP techniques. Filtering helps in manifest pre-processing the final classification results with good opinion.

In NLP techniques we focused on operations like Tokenization, Lemmatization, Stemming, and Stop Words Removal. Tokenization is essentially splitting the sentence into smaller units of an entire document text. Each of these smaller units is called tokens. This is important because the meaning of the text could easily be interpreted by analysing the words present in the text. Lemmatization normally aiming to remove inflectional endings only and to return the base or dictionary form of a word which is known as the lemma. Whereas stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words know as stemma. For example, to find out the root of the word user, users, used, using the stem is *used*. It helps the removal of ambiguity in words. Thus Stemming and Lemmatization is to reduce inflectional forms and sometimes derivationally to related forms of the word to a common base form. Further, stop words are the most common words in any natural language to analyze text data, these stop words might not add much value to the meaning of the document. Some of the stop words are is, was, are, of, and, on, etc. Applying to stop word removal in NLP techniques takes less time for processing datasets. After applying the techniques of NLP, we further refine the dataset by applying Multilevel filtering approaches.

Our novel approach helps to refine according to our requirement of proper

dataset suitable for sentiment classification. This approach includes Inclusive Filter, Relevance Filter, and Opinion Filter. The inclusive filter includes the Indian railway's domain with the help of an official twitter handler. Relevance filter is used to filter the sentence with relevant features and their synonyms used in our research work. This filter used the WordNet lexicon to find the synonyms. Finally, the Opinion filter is used to predict the tweet containing opinion are not using SentiWordNet lexicon. But from the output of multilevel filtering we found some drawbacks such as some tweets are still having difficult to classify input to machine learning classification, there is an ambiguity and unclear to convert text data to numerical data, and also have some problems while selection of topics, collection of synonyms, extraction of concept, detection of subjectivity, checking of polarity, and errors in misclassification. Therefore, to overcome from these drawbacks we further introduced Topic-based filtering technique. This technique contains a module of Topic selection, Lexicon building, Concept extraction, Subjectivity detection, Polarity detection, and Ordinal classification. Topic selection means topics are to be understood by the system correctly, every word of every tweet should compare with topic selection and its synonyms. If it is matched with either topic selected or its synonyms, then it is not removed from the tweet list. Otherwise, it is removed. Lexicon is the list of stems and affixes, together with basic information about them. Lexicon is WordNet using Synsets. Lexicon building is developed by WordNet dictionary which supports sentiwordnet for sentiment analysis. The concept extraction phase is used to identify concepts based on the topics chosen. It is important to deconstruct given text or tweet into the concept for better semantic-aware analysis with WordNet synsets. The next phase is subjectivity detection. It is a NLP task that removes neutral or factual tweets from the given tweets. The factual content is also known as object content that does not reflect any opinion. This is important to increase the accuracy of the sentiment analysis. Polarity detection phase is crucial for sentiment analysis. In this phase, the tweets do have one of the given topics and sentiment is used as input to find the label of a tweet whether it is positive, negative, and neutral. Ordinal classification is also known as ordinal regression which finds error rate between tweets. As misclassification is costly measures such as the overall sentiment of the tweet, topic-based classification, and tweet quantification are used for evaluation. After performing topic-based filtering we created a feature vector

using Text Blog technique. A feature vector is in a numerical format that takes care of classification easier for good sentiment analysis. Where as Text Blog is a python inbuilt library for finding polarity values that is positive, negative and neutral in the numerical dataset as feature vector input target for sentiment classification. It provides a simple API for dividing into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. Our feature vector consists of values for features like train number, temporal feature, group of the tweet of a specific attribute, and opinion of the tweet. For Feature vector, we introduce priority-based system, such as ranking, and different train type. This system will help organisation and passenger to analyse and take decision easily while choosing train booking reservation. This vector is further given to machine learning classifiers for classification of feature selection. Classification will be done using machine learning and deep learning approaches using stacking. In machine learning, we have used Naïve Bayes and Support Vector Machine (SVM) classifier and for deep learning Long Short Term Memory (LSTM) classifier is used. For further enhancement of performance in classification, we have introduced an ensemble classifier with stacking that ensemble machine-learning algorithms with a deep learning algorithm. The ensemble classifier is constituted of Naïve Bayes with LSTM and SVM with LSTM. After classification, the tweets are finally classified into positives, negatives and neutral. Here, the positive means +1, negative as -1, and neutral as 0. The experimental results show that the F-Measure of single classifier i.e. Naïve Bayes is (80.5%), Support Vector Machine is (85.6%) and Long Short Term Memory the F-Measure is (88.1%). Further, the F-Measure of Ensemble classifier i.e. Naïve Bayes with LSTM is (90.1%) while the F-Measure of Support Vector Machine with LSTM is (92.1%). Next, we have compared the F-Measure of different single and ensemble classifier and the results explored that Support Vector Machine with LSTM is better. The empirical study revealed the utility of the proposed framework in ascertaining valuable insights besides understanding the additional value added by long short-term memory in the ensemble stacking approach towards better performance of sentiment classification.

Generally, we have observed from the literature review that the sentiment or opinion dataset is stored in the relational database. But there are certain demerits within these relational database. The demerits are that it is very dif-

difficult to address complex queries over the data and also it comes back with an answer at most once, lagging inherent properties such as transitivity or symmetry. It also has a closed world assumption i.e., what is not known to be true in the database is by default considered false. To overcome all these demerits we proposed and implemented knowledge based system (KBS) using ontology which is a semantic web technique for accessing sentiments data of Indian Railways tweets. Ontology is used in the knowledge-based system as a conceptual framework for providing, accessing, and comprehensively structuring information. To implement knowledge based system we first mapped the relational database (RDB) into relational database schema (RDS). Further, RDS is mapped into Resource Description Framework schema(RDFS). Then RDFS is mapped into Resource Description Framework (RDF). For mapping RDFS to RDF we have introduced and implemented an algorithm that has been used to map Database, Tables, Columns, and Constraints. Afterwards RDF is mapped into ontology which applies knowledge rules that have been developed to execute the query using Sparkle Protocol and RDF Query Language (SPARQL). This knowledge based system will act as a decision support system for Indian railways based on given features.

Keywords: Twitter Dataset, Pre-processing, Sentiment Analysis, Classification, Ontology, Knowledge based system, Indian Railways

Dedicated to

Family & Guru

Contents

Title Page	i
Certificate	ii
Declaration	iii
Acknowledgments	iv
Synopsis	vi
Dedication	xi
Table of Contents	xii
List of Figures	xv
List of Tables	xvii
1 Introduction	1
1.1 Sentiment Analysis	1
1.2 Social Networks	3
1.3 Web Scraping with BeautifulSoup Library	4
1.4 Pre-Processing of dataset	4
1.5 Machine Learning	6
1.6 Knowledge Based System	8
1.7 Motivation	10
1.8 Problem Definition	10
1.9 Scope of the Research	11
1.10 Objectives	11
1.11 Research Contributions	12
1.12 Organization of the Thesis	14
1.13 Summary	15
2 Literature Survey	16
2.1 SENTIMENT ANALYSIS	17
2.1.1 APPLICATIONS RELATED WITH SENTIMENT ANALYSIS	18
2.1.2 LEARNING SENTIMENTAL INFLUENCE	19
2.1.3 SENTIMENTAL POLARITY CLASSIFICATION	20
2.1.4 SENTIMENT ANALYSIS OF CONDITIONAL SENTENCES	21
2.1.5 LEXICON BASED METHODS	22
2.1.6 ROLE OF COMMON SENSE AND CONTEXT INFORMATION ON SENTIMENT ANALYSIS	23
2.1.7 TOOLS USED IN SENTIMENT ANALYSIS	24
2.1.8 MEASURES USED IN SENTIMENT ANALYSIS	24

2.2	PREPROCESSING	25
2.2.1	PRE-PROCESSING TWEETS	25
2.2.2	MORE ON PRE-PROCESSING TWEETS	26
2.3	MACHINE LEARNING AND DEEP LEARNING	29
2.4	SEMANTIC WEB ONTOLOGY	33
2.4.1	ONTOLOGY	33
2.4.2	SEMANTIC ANALYSIS	36
2.5	KNOWLEDGE BASED SYSTEM	36
2.6	DISCUSSION AND SUMMARY	39
3	PROPOSED FRAMEWORK FOR SENTIMENT ANALYSIS BASED ON INDIAN RAILWAYS TWITTER DATASET	42
3.1	FRAMEWORK FOR SENTIMENT ANALYSIS BASED ON IN- DIAN RAILWAYS TWEETS	43
3.1.1	WEB SCRAPING	45
3.1.2	PREPROCESSING	45
3.1.2.1	NATURAL LANGUAGE PROCESSING TECH- NIQUES	46
3.1.2.2	MULTILEVEL FILTERING	47
3.1.2.3	TOPIC BASED FILTERING	47
3.1.3	FEATURE SELECTION	48
3.1.4	CLASSIFICATION	49
3.1.5	SENTIMENT ANALYSIS	50
3.1.6	KNOWLEDGE BASED SYSTEMS USING ONTOLOGY	51
3.2	SUMMARAY	52
4	PRE-PROCESSING OF INDIAN RAILWAYS TWEETS FOR SENTIMENT ANALYSIS	54
4.1	FRAME WORK OF PRE-PROCESSING OF INDIAN RAILWAYS TWEETS FOR SENTIMENT ANALYSIS	54
4.1.1	WEB-SCRAPPING USING BEAUTIFUL SOUP LIBRARY	57
4.1.2	NATURAL LANGUAGE PROCESSING (NLP) TECH- NIQUES	58
4.1.3	MULTILEVEL FILTERING TECHNIQUES	62
4.1.3.1	TOPIC BASED FILTERING TECHNIQUES	70
4.2	SUMMARY	75
5	OPINION CLASSIFICATION OF PREPROCESSING DATASET FOR SENTIMENT ANALYSIS USING SINGLE AND ENSEM- BLE CLASSIFIER	79
5.1	PROPOSED METHODOLOGY FOR EFFECTIVE SENTIMENT ANALYSIS OF TWEETS OF IR	79
5.2	SINGLE CLASSIFIER APPROACH OF MACHINE LEARNING	82
5.2.1	Naïve Bayes	82
5.2.2	Support Vector Machine	84
5.3	SINGLE CLASSIFIER APPROACH OF DEEP LEARNING	84
5.4	ENSEMBLE CLASSIFIER APPROACH USING STACKING	87
5.4.1	Deep Learning based Stacking Ensemble	89

5.4.2	Evaluation Procedure	90
5.5	EXPERIMENTAL SETUP	91
5.5.1	Results of Individual Prediction Models	93
5.5.2	Results of Stacking Ensemble Models	94
5.5.3	Train wise Sentiment Analysis for Indian Railways	95
5.6	SUMMARY	98
6	KNOWLEDGE BASED SYSTEM OF SENTIMENT ANALYSIS DATASET OF INDIAN RAILWAYS	100
6.1	FRAME WORK OF KNOWLEDGE BASED SYSTEM OF SENTIMENT ANALYSIS DATASET OF INDIAN RAILWAYS	100
6.1.1	RELATIONAL DATABASE (RDB) SCHEMA OF SENTIMENT ANALYSIS DATABASE	103
6.1.2	RDB SCHEMA TO RESOURCE DESCRIPTION FRAMEWORK (RDF) SCHEMA	104
6.1.3	RDB to RDF Mapping (RRM) Algorithm	104
6.1.4	MAPPING RDF SCHEMA TO ONTOLOGY	105
6.1.5	FORMATION OF KNOWLEDGE RULES USED IN THE ONTOLOGY	106
6.1.6	FORMATION OF SPARQL QUERY	109
6.2	summarizes the chapter	111
7	Conclusion and Future scope	113
	References	117

List of Figures

3.1	Framework for sentiment analysis of Indian Railways Twitter posts	43
4.1	Frame work of Pre-processing of Indian Railways tweet for Senti- ment Analysis.	55
4.2	Snapshot of Indian Railways Twitter Handler	59
4.3	Snapshot of data extracted using beautiful soup library	59
4.4	Snapshot of dataset after removing unnecessary data for pre-processing	60
4.5	Snapshot of dataset after applying NLP techniques	61
4.6	Snapshot of dataset after having inclusive filter	62
4.7	Snapshot of dataset after having relevance filter	67
4.8	Snapshot of dataset after having opinion filter	69
4.9	Snapshot of results after applying MF algorithm	69
4.10	Proposed framework of Topic based filtering system	71
5.1	Proposed methodology for effective sentiment analysis of tweets of IR	81
5.2	Illustrates how hyperplane is formed for discrimination	85
5.3	Illustrates structure of LSTM [2]	86
5.4	Ensemble mechanism [1]	88
5.5	Shows confusion matrix	90
5.6	Feature vector with sentiment values helpful for labelling and pre- dicted values	93
5.7	Performance of individual sentiment prediction models	94
5.8	Performance of individual sentiment prediction models	96
5.9	Train wise sentiment analysis for Indian Railways	97
6.1	Framework for knowledge based system	101
6.2	Shows class names with the help of protégé tool with five features	105

6.3	Shows train selected in types of data property	106
6.4	onto graph using protégé tool	107
6.5	knowledge rules using protégé tool with names S1 and S2	109
6.6	Sample SPARQL query made	110

List of Tables

3.1	Shows Positive and Negative Words Used for Sentiment Analysis .	44
4.1	Shows Positive and Negative Words Used for Sentiment Analysis .	65
5.1	Performance comparison with precision, recall and F1-Measure . .	93
5.2	Results of ensemble methods	95
5.3	Train wise sentiment analysis for Indian Railways	97
6.1	An excerpt from dataset containing sentiment details for Indian Railways	103
6.2	An excerpt from dataset containing sentiment details for Indian Railways	108

Chapter 1

Introduction

SECTION 1.1 presents brief overview of the Sentiment Analysis. **SECTION 1.2** provides a brief overview of Social Network. **SECTION 1.3** presents a brief overview of Web Scraping with BeautifulSoup library, which has been used in this thesis to extract dataset from twitter. **SECTION 1.4** presents the brief overview of the Pre-processing of dataset. **SECTION 1.5** provides the brief overview of Machine Learning classification which has been used to classify the pre-processing dataset. **SECTION 1.6** presents brief overview of Knowledge Based System which has been developed for sentiment analysis dataset. **SECTION 1.7** explains the motivation behind the work presented in this thesis. **SECTION 1.8** discuss about problem definition. **SECTION 1.9** deals with scope of the research. **SECTION 1.10** deals about research objective. **SECTION 1.11** gives brief overview of the contribution of our work. **SECTION 1.12** outline the organization of thesis with a brief overview of topics covered in each chapter. Finally, **SECTION 1.13** summarizes this chapter.

1.1 Sentiment Analysis

Information gathering considers different aspects. However, finding what is the opinion of other people is an important aspect in information gathering. In the capacity of customers, people do think about what other people are thinking about a service or product. This aspect is even more important to organizations as they can make use of opinions of others to improve product or services. Moreover, the organizations can influence the users by improving their Quality of Services(QoS). Organizations want to know the events and the feedback of the

events towards product management and marketing. They wanted to have both traditional customer feedbacks in conventional channels and the social feedback through web based social networks. Thus a comprehensive business intelligence (BI) is expected by the organizations to ensure that their brands or services will give positive influences.

The technologies associated with Web 2.0 paved way for different aspects of higher importance in the contemporary era. They are known as social networking, tagging, broadcasting, social bookmarking, reviewing and blogging. With these there is increase in the resources giving opinion that bring about challenges and opportunities to organizations. This has led to significant attraction towards sentiment analysis or opinion mining. Therefore, the textual content with subjectivity, opinions or sentiments is given importance with computational treatment. Many tools came into existence for text analytics. With these tools specific to sentiment analysis such as WordNet and SentiWordNet, organizations are striving to gain timely social feedback that reflects opinions of customers on specific product or service [1]. By this BI, companies can take necessary steps and see that the customers' opinions will be positive and in turn influence other people as well.

In this thesis, we have performed the sentiment analysis on Indian Railways(IR) Tweets on respective features such as Punctuality, Staff Behaviour, Cleanliness, Security, and Food Quality. This analysis will help the railway organisation to improve their quality of services on above said features. As we know that Indian Railways is Asia's largest and the world's second largest rail network of India operated by the Ministry of Railways. It has more than 11,000 locomotives and over 70,000 passenger coaches. It transports around 2.5 crore passengers daily. IR carried around 8.26 billion passengers and 1.16 billion tonnes of freight in the fiscal year ending March 2018. As IR is the preferred transport to most of the Indians, it is observed that the Online Social Network (OSN) carry social feedback on IR. In addition to the direct feedback given by passengers in traditional approaches, they can also provide their feedback in the form of reviews, micro-blogging using social media platform like Twitter, Face book, and etc. Nevertheless, the tweets from Twitter carry the essence of social feedback given by passengers of IR. In this context, it is not wise to ignore social feed-

back. In fact, it is indispensable for any organization to consider opinions of public available in social media. IR is no exception to this. Swacch Bharat Abhiyan Prime Minister Organisation (PMO) considers the whole nation including IR to improve in various parameters for healthy approaches. Having understood about the importance of social feedback for IR, sentiment analysis is made on the tweets pertaining to Indian Railways so as to help it to get benefited for improving services and gain highly positive opinions on its services. We aimed to perform sentiment analysis on different features of IR and further to develop knowledge based sentiment analysis system that provides essential of social feedback on different features of IR such as Staff Behaviour, Punctuality, Cleanliness, Food Quality and Security. The research carried out in this thesis leads to an out of the box solution that can be adapted as part of Decision Support System (DSS) for IR. It also has impact on other stakeholders of IR.

1.2 Social Networks

A social network is a social made structure of a set of social individuals or organisation with social interactions between actors or entities. A social networking services is an online platform which people use to build social networks or social relationships with other people like Twitter. Before the emergence of World Wide Web (WWW) there was little possibility for leaving opinionated text made available to public. Even after invention of WWW, it was static initially used for information sharing without any interaction with user. People used to take opinions of other people in conventional way. After incorporation of the dynamic web contents with Web 2.0 [3], it became possible to have interactive web applications. This has led to different platforms over Internet emerged to share opinionated content. Those platforms include social media, forums and blogs. Thus various companies started their own facilities that help customers to leave their opinions. In addition to this online social networks such as Twitter became platforms for virtual communities to emerge. As of now there is large volumes of opinionated text being produced on social media every day. Most of the Internet based facilities for sharing opinions provide sufficient data to all stakeholders. This has become an important source for the research pertaining to sentiment analysis. The opinionated data generated by social networks such as Twitter

helps in opinion mining or sentiment analysis that leads to application of text analytics, computational linguistics and Natural Language Processing(NLP)[2].

1.3 Web Scraping with Beautiful Soup Library

Web Scraping is the process of acquiring data from running web applications. In fact, it is the technique that became very useful for generating data from web applications. The generated data can be saved to local storage or it can be saved to relational databases or kept in tabular format in spreadsheets as well. Web Scraping has many other names too such as web harvesting, web data extraction and screen scraping. The data presented by most of the web applications provide read only access to users through web browsers. Those applications do not provide any feature to save a copy of data for use personally. Users can only do it manually and that is a tedious task. To overcome this problem, web scraping is the technique that came handy. Web scraping software automatically extracts data from multiple pages and load it to the local storage. The web scraper software may be a generic software product like WebHarvy[4] or a custom built targeting specific web site. In this thesis, we have used web scraping with Beautiful Soup library to extract data form twitter handler pertaining to Indian railways for sentiment analysis. This library has been used as it helps better than Application Program Interface(API) like Rest API and Streamed API because these API has restriction of data like getting of only previous one week but according to our requirement we need streaming data that will be available only by using web scraping with beautiful soup library which is suitable and reliable for our research. This python library package is used for parsing HTML and XML documents which have malformed mark-up that is non closed tags[5]. It creates a parse tree for developing parsed pages that can be used to extract data form HTML, which is mainly useful for web scraping.

1.4 Pre-Processing of dataset

Dataset is a collection of records or tuples. In case of textual data such as Twitter posts, the dataset also contains opinions of people. They are also called sentiments and the analysis of such text is given paramount importance. However, the dataset when used for sentiment analysis, it results in poor performance

unless there is much needed pre-processing in the form of Natural Language Processing (NLP) techniques[6]. Data pre-processing is a data mining technique that involves transforming raw data into an every understandable format. Real nature world data is often inconsistent, incomplete or lacking in certain behaviours or trends, it is proven method of resolving such issues. The steps of data pre-processing are libraries import, data reads, checking for categorical data, transformation of data, and finally data spitting.

In this thesis, different NLP techniques has been used to pre-process the dataset. The techniques that has been used are Tokenization, Stemming, Lemmatization ,and Stop Word Removal. Tokenization is essentially splitting the sentence into smaller units of an entire document text. Each of these smaller units is called tokens. This is important because the meaning of the text could easily be interpreted by analysing the words present in the text. Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words know as stemma. For example, to find out the root of the word user, users, used, using the stem is *used*. It helps the removal of ambiguity in words. Lemmatization normally aiming to remove inflectional endings only and to return the base or dictionary form of a word which is known as the lemma. Thus stemming and lemmatization is to reduce inflectional forms and sometimes derivationally to related forms of the word to a common base form. Further, stop words are the most common words in any natural language to analyse text data, these stop words might not add much value to the meaning of the document. Some of the stop words are is, was, are, of, and, on, etc. Applying stop word removal in NLP techniques takes less time for processing the datasets. After applying the techniques of NLP, we have further introduced Multilevel filtering and Topic based filtering for better refining of dataset.

Multi-level filtering includes Inclusive Filter, Relevance Filter, and Opinion Filter. The inclusive filter includes the Indian railway's domain with the help of an official twitter handler. Relevance filter is used to filter the sentence with relevant features and their synonyms used in our research work. This filter used the WordNet lexicon to find the synonyms. Finally, the Opinion filter is used to predict the tweet containing opinion are not using SentiWordNet lexicon. Whereas

Topic based filtering includes Topic selection, Lexicon building, Concept extraction, Subjectivity detection, Polarity detection, and Ordinal classification. Topic selection means topics are to be understood by the system correctly, every tweet of every word should compare with topic selection and its synonyms. If it is matched with either topic selected or its synonyms, then it is not removed from the tweet list. Otherwise, it is removed. Lexicon is the list of stems and affixes, together with basic information about them. Lexicon building is WordNet using Synsets. Lexicon building is developed by WordNet dictionary which supports SentiWordNet for sentiment analysis. The concept extraction phase is used to identify concepts based on the topics chosen. It is important to deconstruct given text or tweet into the concept for better semantic-aware analysis with WordNet synsets. Whereas subjectivity detection is an NLP task that removes neutral or factual tweets from the given tweets. The factual content is also known as object content that does not reflect any opinion. This is important to increase the accuracy of the sentiment analysis. Polarity detection phase is crucial for sentiment analysis. In this phase, the tweets do have one of the given topics and sentiment is used as input to find the label of a tweet whether it is positive, negative, and neutral. Ordinal classification is also known as ordinal regression which finds error rate between tweets. After performing topic-based filtering we created a feature selection using Text Blog technique to give input to the classifier for machine learning classification for sentiment analysis.

1.5 Machine Learning

Machine Learning (ML) is learning phenomenon used by a computer program to gain knowledge and make some important tasks like prediction or forecast. That is the rationale behind the fact that ML is part of Artificial Intelligence (AI). In fact, ML is used to automate data analysis and to build knowledge models that are used to make well informed decisions later on. ML techniques learn from data (gain knowledge), identify trends or patterns and make intelligent decisions with minimal or without human intervention[7]. In machine learning there are lot of classifiers which are used to classify the dataset such as Decision Tree Classifiers, Linear Classifier containing Support Vector Machine, Neural Network, Rule based Classifiers, Probabilistic Classifiers containing Naive Bayes, Bayesian

Network and Maximum Entropy which comes under supervised learning . But in this thesis we have used machine-learning classifiers such as Naive Bayes, Support Vector Machines (SVM) and Deep learning techniques such as Long Short Term Memory (LSTM) for classification of data set.

Naïve Bayes classifier is a family of simple probabilistic classifiers based on Bayes theorem with strong independence assumptions between the features of our Indian Railways Application. A Support Vector Machines(SVM) is a supervised models of machine learning with associated learning algorithms is to find a hyperplane in an N dimensional space (N – the number of features) that distinctly classifies the data points. Whereas Long Short Term Memory(LSTM) is an Artificial Recurrent Neural Network (A RNN) architecture used in the field of deep learning. LSTM cannot only process single data points but also entire sequence of data. For further enhancement of accuracy, we have also introduced ensemble classifier with stacking for classification of data set. Ensemble is nothing but combining two different classifiers with stacking. Stacked generalisation is an ensemble method where a new model learns how to best combine the predictions, decisions to form multiple existing models for training. In our research work the ensemble classifier is constituted of Naïve Bayes with LSTM and SVM with LSTM.

Unlike ML techniques in the past, the modern ML techniques are more accurate with new computing technologies. ML is interactive in nature as it needs to gain knowledge from the new data that is arrived. These techniques also learn from past computations to generate repeatable decisions reliably. In fact, it is the science that is widely used across industries now. ML algorithms are available for many years but of late they got ability to apply complex mathematical computations for solving many real world problems. For instance, self-driving Google car [8] is an example where ML is used. Another specific example of machine learning is the study of opinions that are in the form of Twitter posts. In this thesis, ML techniques with supervised learning are used to arrive at BI needed by Indian Railways. IR gets social feedback from tweets in order to improve its services in specific areas.

1.6 Knowledge Based System

Knowledge Based System(KBS) is a form branch of artificial intelligence(AI) [8] that aims to capture and decides the knowledge of human experts to support decision-making. Examples of knowledge-based systems include expert systems, which are so called because of their reliance on human expertise. The AI can be divided into two broad types: Knowledge based systems(KBS) and Computational Intelligence(CI) [9]. KBS use explicit representations of knowledge in the form of words and symbols. This explicit representation makes the knowledge more easily to read and understood by a human than the numerically derived implicit models in computational intelligence.

KBS include techniques such as rule-based, model-based, and case-based reasoning. They were among the first forms of investigation into AI and remain a major theme. Early research focused on specialist applications in areas such as chemistry, medicine, and computer hardware. These early successes generated great optimism in AI, but more broad-based representations of human intelligence have remained difficult to achieve.

After going through the literature survey of sentiment analysis or opinion mining we came across that the data set of sentiment analysis is being stored in the relational database (RDB). But there are certain drawbacks in storing the data in the relational database such as a relational database schema exists primarily to constraint and structure the data, therefore, it is very difficult to address complex queries over the data and also it come back with an answer at most one, further the relation in the database themselves do not have properties inherent in them such as transitivity or symmetry, and also it has closed world assumption this means that what is not known to be true in the database is by default considered false because knowledge of the world represented in the database is assumed to be complete.

Therefore, to overcome these drawbacks we have introduced and developed knowledge based system of sentiment analysis data set of features such as Cleanliness, Staff Behaviour, Punctuality, Security and Food Quality pertaining to Indian Railways using ontology and knowledge rules which come under rule based

technique. This system will be accessible to humans and also programs in heterogeneous Machine-to-Machine (M2M) environments.

In our research work we have developed knowledge based system by transforming relational data model to semantic data model. This model can be directly queried or queries through a program in an interoperable way. To implement this system we first converted the relational database (RDB) to relational database schema (RDS). Further, RDS is mapped into Resource Description Framework schema (RDFS). Next RDFS is mapped into Resource Description Framework (RDF). It is a standard format to store data. It forms Semantic web and Web Ontology Language (OWL) for interpretation of data. RDF has both literals and semantic meanings of the same. Thus it provides rich interoperable interface to a knowledge domain in fully automated and interactive fashion. RDF was introduced in order to handle situations where web data needs to be processed and exchanged by applications instead of just showing data to users. This ability of exchanging data between applications makes RDF very useful in the contemporary era. Semantic data models like RDF can help organizations of specific domain to organize domain BI or knowledge and share it through the enterprises. To map RDS to RDFS an algorithm has been proposed which has multiple procedures to map different objects of relational database schema to the equivalent objects in RDF schema.

Afterwards, RDF is mapped into ontology. Ontology is a semantic web technique used in the knowledge-based system as a conceptual framework for providing, accessing, and comprehensively structuring the information [10]. It is used to represent the knowledge which is made up of concepts and relationships among them. In ontology we define the Classes, Data Type Property, Object Property, and Knowledge rules to execute complex query. Therefore, we introduced and implemented an algorithm that has been used in relational database to map tables into classes, columns into data type property within domain and range, and constraints have been mapped into sub class of object property with restrictions. Here the ontology applies knowledge rule that have been developed to execute the complex query using Sparkle Protocol and RDF Query Language (SPARQL).

1.7 Motivation

IR is an indispensable service for Indian citizens. In fact, people of all walks of life prefer travelling through various kinds of trains run by IR. Since IR is the second largest train network, obviously it involves issues related to Punctuality, Staff Behaviour, Security, Cleanliness, and Food Quality. There is need for feedback from passengers from time to time in order to improve services of IR. Traditional feedback plays vital role in addressing problems. However, of late, social media became so powerful that the traditional feedback became inadequate. Unless social feedback is combined with traditional feedback, there is possibility of making incorrect or less than ideal decisions. When social feedback is used along with traditional one, it provides highly comprehensive means of garnering customer feedback that helps in making strategic decisions to improve services of IR. This research which is aimed to develop knowledge system of sentiment analysis for Indian Railways Tweets add value to IR. It will provide benefits to all stakeholders of the IR. This is the motivation behind taking up this research work.

1.8 Problem Definition

The research problem of this thesis is to develop knowledge based system (KBS) using ontology of sentiment analysis dataset based on different features of IR such as Punctuality, Staff Behaviour, Security, Cleanliness and Food Quality which will help IR to make a strategic decision to improve its services as well as it will also help the passengers to book their ticket of respective trains not only on basis of the availability of the reservation but also on the basis of the performance of the said features. The research problems can be further divided into four sub problems which are as follows:

- 1) Building Frame work for sentiment analysis based on Indian Railways Twitter data set.
- 2) Extraction of dataset from Twitter handler of Indian Railways afterwards Pre-processing of dataset using NLP techniques, Multilevel Filtering, and Topic based filtering.

- 3) Opinion classification of pre-processed dataset for sentiment analysis using single and ensemble classifiers approaches with stacking.
- 4) Finally, at last, designing and developing of Knowledge Based System using ontology of sentiment analysis dataset of IR to solve complex queries.

1.9 Scope of the Research

The scope is limited to develop a knowledge based system of sentiment analysis dataset using ontology for Indian Railways (IR) only. Further, the work is limited for only five features of IR such as Punctuality, Staff Behaviour, Security, Cleanliness and Food Quality. It involves novel methods for pre-processing of tweets including Natural language processing(NLP) techniques and filtering mechanisms, opinion classification using machine learning and deep learning techniques for effective sentiment analysis of tweets of IR and realization of knowledge based system using RDF and ontology for supporting SPARQL queries. In summary, the scope of the research is to build only a knowledge based system that mines social media (tweets) to ascertain essence of social feedback to take measures to improve services in IR.

1.10 Objectives

Our objective is to make use of the social media posts of Indian Railways (IR) for investigating the opinion of the customer using different techniques of Opinion Analysis that can improve the services of Indian Railways. Further, to investigate present state of the art on sentiment analysis and knowledge based systems that exploit social feedback for rendering services to organizations in the real world. Thus the main objectives are as follows:

- To propose a framework for knowledge based system of sentiment data of Indian Railways tweets that helps IR to know the essence of social feedback.
- To propose novel pre-processing techniques that leverage the proposed framework in garnering business intelligence from social feedback.
- To propose an ensemble classifiers using stacking for effective opinion classification of pre-processed tweets of IR.

- To design and develop a knowledge based system using ontology and knowledge rules that help IR to make well-informed decisions.
- Further, to evaluate the proposed system and draw conclusions.

1.11 Research Contributions

The section formulates the contribution of our work based on the given problem definition. A survey has been done in the area of the research of Sentiment analysis, Extraction of the dataset from the Twitter, Pre-processing of dataset, Classification using machine learning(ML) and deep learning approaches, and Knowledge based system (KBS). This survey helps in initiating the problem statement, and solution of this problem. The contributions are as follows:

1. A framework is proposed to realize knowledge based sentiment analysis for garnering social feedback from tweets of IR. The framework provides the module needed for novel pre-processing techniques, machine learning techniques like Naïve Bayes and SVM and deep learning technique such as LSTM, and ensemble classification techniques, knowledge based system using ontology and knowledge rules to visualize the knowledge and support for retrieval of knowledge with ease. Thus a knowledge based system is realized for IR to know the essence of social feedback to make well-informed decisions.

2. A novel pre-processing approach is defined and implemented. It includes different kinds of filtering techniques such as multi-level filtering and topic based filtering along with NLP techniques such as Tokenization, Stemming, Lemmatization, and Stop words removal. The filtering is realized with two algorithms namely Multi-Level Filtering (MF) algorithm and Representative Feature Selection (RFS) algorithm.

3. We have also introduced creation of feature selection using Text Blog technique. This vector will input to the machine learning and deep learning classification. In feature selection a novel approach has been introduced called priority based system. This system will helps ranking the most frequently occurred features in our twitter IR dataset.

4. After pre-processing and creation of feature selection different machine learning and deep learning algorithms are implied for effective classification of tweets of IR. We also performed classification using ensemble classifier with stacking to enhance the performance of the classification. The ML techniques used are Naïve Bayes and SVM. Later these techniques are used along with deep learning method known as LSTM to form ensemble classifier with stacking. Next, we have compared the F-Measure of different single and ensemble classifier and results explored that SVM with LSTM is better. The empirical study revealed the utility of the proposed framework in ascertaining valuable insights besides understanding the additional value added by LSTM in ensemble stacking approach towards better performance of sentiment classification.

5. The classification techniques used, as mentioned above, led to the results in the form of a relational data that contains train number and sentiment polarity for features like Punctuality, Food Quality, Cleanliness, Staff Behaviour and Security. But the demerits of this data base is that it comes back with an answer at most once and it is very difficult to address complex queries over the data, lagging inherent properties such as transitivity or symmetry. It also has closed world assumption i.e, what is not known to be true in the data base is by default considered as false. Therefore, to overcome these drawbacks a knowledge based system is made using ontology. The problem of deep web access in the contemporary era where knowledge needs to be represented and shared in quite natural and intuitive way among organizations and individuals is investigated. A framework is proposed to have a systematic approach to convert traditional RDB model to semantic data model in knowledge based intelligent system. This knowledge based system is accessible to humans and also programs in heterogeneous Machine-to-Machine (M2M) environments. It can also be used in decision support system (DSS) of Indian railways for effective decision-making as it conveys sentiments of Indian Railways tweets that is helpful for passengers, organization, and administrators to take further steps.

1.12 Organization of the Thesis

The research carried out in this thesis is organized into several chapters. Rest of the chapter within this thesis are organised as follows:

Chapter 2 reviews relevant literature pertaining to the study area. It covers present state of the art on sentiment analysis on Twitter tweets. It also throws light on different process used to pre-process the data sets extracted from online social networks (OSN) as well as it discussed different machine learning and deep learning algorithms used for classification of pre-processed dataset. Further, it also throws some light on semantic web and ontology and also presents reviews on Knowledge based systems. Finally, literature reviews has been discussed and summarized.

Chapter 3 presents the framework of sentiment analysis based on Indian Railways tweets. In this chapter the description of each and every module has been given which is used to construct the proposed framework such as description about Web Scrapping, Pre-processing of data sets, Natural Language Processing (NLP) techniques, Multilevel filtering, and Topic based filtering. It also gives the description regarding creation of feature selection, doing classification using machine learning, deep learning, and ensemble classifiers, and last it presents the description of knowledge based system of sentiment analysis data set using ontology. Finally, this chapter has been summarized.

Chapter 4 presents the pre-processing of dataset extracted from twitter handler of Indian Railways. The pre-processing includes NLP Techniques, Multi-level Filtering and Topic Based Filtering. The Multilevel Filtering includes Inclusive filter, Relevance filter, and Opinion filter. The relevance filter used WordNet Lexicon to find the synonyms. Finally, the opinion filter is used to predict the tweet containing opinion. Further, we used the Topic-based filtering technique. This technique contains a module of Topic selection, Lexicon building, Concept extraction, Subjectivity detection, Polarity detection, and Ordinal classification.

Chapter 5 presents the machine learning and deep learning classification. In Machine Learning we have used Naive Bayes and Support Vector Machine(SVM) classifier where as in Deep Learning we have used Long Short Term Memory(LSTM) classifier. For further enhancement of performance in classification we have intro-

duced a new approach called Ensemble Classifier with Stacking. This ensemble classifier is constituted as of Naive Bayes with LSTM and SVM with LSTM.

Chapter 6 provides details regarding development of Knowledge Based System(KBS) using ontology for sentiment analysis data set. It throws light on the mapping of relational database (RDB) to RDB schema and further RDB schema to RDF schema. For mapping RDB to RDF schema we have introduced and implemented an algorithm that has been used to map database, tables, columns, and constraints and further RDF schema is mapped into RDF and then RDF is mapped into ontology which apply knowledge rules that has been developed to execute the query using SPARQL.

Chapter 7 finally concludes the research besides providing directions for possible future scope of the research.

1.13 Summary

This chapter introduced the topic used in our research work. It also provided overview of the Problem statement, Research objectives, Scope of the work, and Research contribution with in the work. In the next chapter, we have presented the related work in the area of Sentiment Analysis, Pre-processing of the dataset, Machine learning and Deep Learning approaches, Semantic Web with Ontology, and Knowledge based system.

Chapter 2

Literature Survey

This chapter focuses mainly on those proposals that are related to the work presented in this thesis from the enormous amount of research that has been done in sentiment analysis in the recent years.

Sentiment analysis became an important area of research as there are business models that depend on social intelligence making well informed decisions. Expert decision making now needs to ascertain the trends in thinking of customers. The thinking patterns of customers are found in the bulk of data exponentially added to Online Social Networks (OSNs). Such data became a goldmine for researchers. In this chapter, keeping the importance of the subject, we reviewed literature to know the present academic thinking on sentiment analysis, its methods and applications. In the literature it is found that sentiment analysis provides ample scope to gain business intelligence (BI) by mining online reviews, news, social media data like tweets from Twitter and so on. Without considering sentiment analysis, the BI obtained with traditional methods become inadequate due to proliferation of OSNs and habits of people of all walks of life sharing their views over them. This review covers various methods in sentiment analysis, applications, tools used and measures used to evaluate methods. It is found that knowledge based approaches become valuable in the wake of the need for automated discovery of sentiments from given datasets. This chapter has been classified into five broad categories, namely sentiment analysis, pre-processing techniques, machine learning, semantic web ontology and knowledge based systems.

This chapter is organized as follows: SECTION 2.1 introduces the issues related to sentiment analysis. SUBSECTION 2.1.1 discusses the applications related with sentiment analysis. SUBSECTION 2.1.2 deals with the sentimental

influencing. SUBSECTION 2.1.3 introduces the issues related with sentimental polarity classification. SUBSECTION 2.1.4 focussed on conditional sentences for sentiment analysis. SUBSECTION 2.1.5 focussed on Lexicon Based Methods. SUBSECTION 2.1.6 introduces sentiment analysis based on common sense and context information. SUBSECTION 2.1.7 introduces some tools used in sentiment analysis. SUBSECTION 2.1.8 focussed on some measures used in sentiment analysis. SECTION 2.2 presents literature reviews on pre-processing of datasets extracted from online social networks (OSN). SECTION 2.3 presents literature reviews on classification using machine learning and deep learning for sentiment analysis. SECTION 2.4 throws light into semantic web and ontology. SECTION 2.5 presents literature reviews on Knowledge based systems. Finally, SECTION 2.6 summarises discussion and summary.

2.1 SENTIMENT ANALYSIS

This section reviews literature on various aspects of sentiment analysis. It throws light into its applications, methods, tools and measures used in the research of sentiment analysis. There are different aspects involved in sentiment analysis. Mowlaei et al. [?] employed aspect based lexicons in order to achieve sentiment analysis that exploits aspect-based approach. They proposed framework towards this end. It involves lexicon generation, lexicon generation using Genetic Algorithm (GA), chromosome function, create function, fitness function, mutate function, crossover function, lexicon fusion and classification. It has improved state of the art in aspect based sentiment analysis. However, it does not support identification of implicit aspects in the data which will make room for further research. Ozturk and Ayvaz [?] focused on the text mining approach for sentiment analysis considering a case known as Syrian Refugee Crisis (SRC). They used TF-IDF along with Turkish lexicon for sentiment analysis. Their research outcomes include the SRC and the reactions of different communities about it. Nakov et al.[?] explores the research and innovations made on sentiment analysis of Twitter tweets as part of SemEval-2016 challenge. They introduced two variants. The first variant talks about a Three-point scale in sentiment classification while the second variant deals with prevalence of the classes of interest. In future, they intended to improve their work for multi-lingual sentiment clas-

sification. Rosenthal et al. [104] used Twitter for sentiment analysis as part of SemEval-2017 challenge. They focused on the five-point scale of sentiments, quantification of distribution of sentiments and usage of other languages as well. With the innovations in the research, they intended to improve it further in future.

2.1.1 APPLICATIONS RELATED WITH SENTIMENT ANALYSIS

According to Duric and Song [15] there are many applications of sentiment analysis. They include online customer reviews, sentiment search, marketing and business intelligence, and detection of inflammatory text and cyber-bullying. Mizumoto et al. [16] proposed a methodology to build a polarity dictionary automatically in order to use it in the sentiment analysis of stock market news. They used semi-supervised learning method to achieve this. Zhang et al. [17] combined two models such as Bayesian classification model and system similarity model to have a prediction system for stock trends. Sakaki et al. [18] proposed an algorithm based on probabilistic spatiotemporal model for real time event detection (eg: Earthquake) by analysing Twitter tweets. Each user of Twitter is considered as sensor and particle filtering is applied to have event detection and location estimation. In [19] event prediction is explored by combining causal rules extraction and temporal sentiment analysis.

Gupta and Shalini [20] used sentimental orientation calculator proposed in [21] for sentiment analysis targeting Tweets pertaining to Indian Railways to improve user experience. Modelling and recognizing situations is studied in [22] for different scenarios. Sentiment analysis can help in understanding the opinion of people on a product or service. This can help in providing contextual advertising in social networking sites. Fan and Chang [23] focused on contextual advertising. Towards this end, they proposed a framework named Sentiment-Oriented Contextual Advertising (SOCA). Their methodology has two important phases such as textual advertise matching and sentiment analysis to have more utility in advertising.

Wei and Gulla [24] proposed Hierarchical-Learning Sentiment Ontology Tree (HL-SOT) for understanding sentiments on product reviews. Their approach showed better performance as it helped in opinion mining of real world reviews.

E-learning is also possible with sentiment analysis. Ortigosa et al. [25] proposed sentiment analysis method on Facebook and its utility in e-learning process. They implemented the method in face book application known as SentBuk. They extracted users' sentiment polarities and modelled them to understand their emotional changes and used them in e-learning application. A survey of sentiment analysis methods used for education domains is found in [26]. Ji et al. [27] proposed a real time sentiment analysis method known as Epidemic Sentiment Monitoring System (ESMOS) for monitoring disease sentiments and make well informed decisions. Singhal et al. [28] are used to model Indian general elections using political Twitter data. They made use of political tweet in order to model general elections and extract knowhow from it. Ali et al. [29] explored an on-line review classification system by combining SVM and fuzzy domain ontology (FDO). Saif et al. [30] built an application known as SentiCircle based on lexicon which makes use of contextual and semantic meaning of words in corpus. Karanasou et al. [31] presented a system that is scalable and real time in sentiment analysis of Twitter data. They found it to be highly scalable and accurate as well. Nakov et al. [32] provides sentiment classification in Twitter with binary and ordinal granularity in recognizing sentiments. Azar and Lo [33] explored prediction of stock markets to understand and analyze stocks that increase asset prices and yield more returns. Besides, they planned to have Federal Open Market Committee (FOMC) meetings through Twitter feeds. This has revealed the importance and wisdom of Twitter crowds.

2.1.2 LEARNING SENTIMENTAL INFLUENCE

Wu and Ren (2011) [34] proposed models that help in discovering sentimental influencing probabilities and influenced probabilities for Twitter users. They proposed an unsupervised lexicon based approach for achieving this. They considered three kinds of user actions namely mention, reply, and retweet. Mention is identified with the presence of @username in a tweet which is known as an influencing action. Reply begins with @username which is considered an influencing action. Retweet is also considered an influenced action. They represented sentiment score as in Eq. (2.1).

$$score_t = count_t(pos.word) - count_t(neg.word) \quad (2.1)$$

Based on the three actions, a probability model is defined as shown in Eq.

(2.2) for learning sentimental influence. The influence probability model considers two users such as u and v showing probability of user u influencing user v .

$$P_{u2v} = \alpha \frac{(\text{count}(\text{mention}_{u2v}))}{(\text{count}(\text{all_tweets}_u))} + (1 - \alpha) \frac{(\text{count}(\text{reply}_{v2u}) + \text{count}(\text{retweet}_{v2u}))}{(\text{count}(\text{all_tweets}_v))} \quad (2.2)$$

Where α denotes influential factor which is between 0 and 1. The Eq. (2.2) is based on Bernoulli distribution model. The research revealed that users with high influencing/influenced probabilities tend to have either positive or negative influence able/influential. On the other hand, the users with positive/negative influence able/influential always worth nothing in the subsequent step in the analysis.

$$P_{u2S} = 1 - \Pi_{(v \in S)}(1 - P_{u2v}) \quad (2.3)$$

$$P_{S2u} = 1 - \Pi_{(v \in S)}(1 - P_{v2u}) \quad (2.4)$$

The influencing probability can result in either positive or negative value. They have not focused on the sentiment propagation over social networks with dynamic computing of probabilities.

Fersini [35] studied in the influence of words used by the users in order to find the importance of influence probability. The opinions in general will have their influence on others. That influence can be measured in different ways. For instance, in virtual communities, influencing factors play important role and the researchers used it for making benefits out of it. Organizations used it in order to have better results by influencing the people over virtual communities in making decisions intended by them.

2.1.3 SENTIMENTAL POLARITY CLASSIFICATION

Liu et al. [36] analysed different linguistic features for ascertaining sentimental polarity classification. They found different features related to polarity classification. They are known as Lexical Features (LF), Polarized Lexical Features (PL), Polarized Bigram Features (PB), and Transform Word Features (T). These features with different combinations are used to have polarity classification results. They found issues with idiomatic expressions, ironic writing style and background knowledge and required further research.

Rosenthal et al. [37] focused on sentiment analysis on Twitter. They used sentiment score that ranges from 0 to 1 showing intermediate values from lowest

score (zero) to highest score (1). They used in their research the degree of polarity in order to make decisions on the sentiment classification. They also considered contextual polarity disambiguation for better decision making. In addition, they also used message polarity and topic polarity and overall polarity for improving the state of the art in sentiment classification. When polarity is given qualitatively, it ends up with positive and negative. But when the range of values or probabilities are used, it becomes quantitative and fuzzier in nature to improve accuracy in classification.

Cambria et al. [38] given importance to sentiment polarity, polarity identification or detection. Sentiment polarity is extracted from the textual content. Polarity bearing states such as FIX, BREAK and INTACT are used for better management. Polarity is considered in multi-word expressions that are part of text corpus. Identifying linguistic patterns and inferring polarity is given paramount importance in their research. Sentence-level polarity and overall polarity are considered in order to have better accuracy in classification.

2.1.4 SENTIMENT ANALYSIS OF CONDITIONAL SENTENCES

Narayanan et al. [39] focussed on conditional sentences for sentiment analysis. Conditional statements are used in any language. For instance, if it rains, take umbrella while going out is one such sentence in English. They used different phases in the methodology. They are feature construction and classification strategies. Different classification strategies they followed include clause-based classification, consequent-based classification and whole-sentence based classification. Al-Smadi et al. [40] used conditional approaches for sentiment classification. They used a conditional random field based classification method associated with deep learning. Especially it is based on the LSTM neural network. Conditional approach helped in fining the class labels based on the aspects availability and opinion target expression (OTE). Thus the conditional random field used on top of LSTM could improve performance in classification.

Al-Ayyoub et al. [42] investigated on Arabic language for sentiment analysis. They also used sentiment analysis to be carried out on conditional statements. They used conditional random fields (CRFs) in order to have better classifiers in their research. For sentiment labelling, their approach helped in a better way.

Al-Smadi et al. [43] investigated on the sentences that have conditions as well. They used LSTM neural networks to train and achieve aspect based sentiment analysis. They used aspect sentiment polarity identification as well for improving accuracy in classification.

2.1.5 LEXICON BASED METHODS

Taboada et al. [44] proposed a lexicon based solution known as Semantic Orientation Calculator (SO-CAL) for sentiment analysis. They used the notion of Sentiment Orientation (SO) value for making a calculator. Wilson et al. [45] used a lexicon of positive and negative words in order to recognize contextual polarity while finding sentiments at phrase-level. Combining lexicon based approaches and learning-based methods is studied in [46] for Twitter sentiment analysis. It was an entity level sentiment analysis and exploits lexicon based knowledge to improve it further.

Fersini et al. [47] explored different methods that are based on lexicon for sentiment analysis. Lexicon is used for both supervised learning and also unsupervised learning methods. Sentiment lexicons are constructed in order to arrive at prediction of sentiment labels. They also discussed about lexicon-corpus based learning approach for sentiment analysis. The lexicon approach is also combined with other approaches for better accuracy. For instance, lexicon based learning and corpus based learning are combined to have a hybrid approach. Alaei et al. [48] on the other hand used lexicon in tourism domain for sentiment analysis. They used both lexicon and rule based methods in order to develop a framework for sentiment analysis. Comprehensive lexicons are combined with fine-tuned rules for semi-automatically detecting labels pertaining to sentiments. Jeong et al. [49] studied different approaches such as deep learning based methods, text classification based methods and lexicon based methods for sentiment classification. The lexicon based method exploits pre-defined dictionaries that enable defining sentiment words. The tools like WordNet and SentiWordNet play important role in sentiment classification methods. Chaturvedi et al. [50] found that there are different methods to make use of lexicons in the research of sentiment analysis. Lexicons can be generated with different approaches using neutral words and detection of subjectivity is also given importance in the research. Word-emotion association is also used for lexicon construction and usage. There

are manually created lexicons and automatically generated lexicons for the utility in sentiment analysis.

2.1.6 ROLE OF COMMON SENSE AND CONTEXT INFORMATION ON SENTIMENT ANALYSIS

Agarwal et al. [51] proposed a method for sentiment analysis based on common sense and context information extracted from ConceptNet based ontology. This kind of ontology contains domain specific concepts used to extract important features to make well informed decisions. The ConceptNet ontology is also used in [52] for sentiment aggregation for more effective sentiment analysis. Smeureanu et al. [53] presented a solution for business ontology that is aimed at evaluating Corporate Social Responsibility (CSR). Study of emotions play vital role in understanding people and context besides making strategic decisions. Sentiment oriented information retrieval is studied in [54] besides working on sentiment based characterization. Chaturvedi et al. [55] explored the role of context in deriving sentiments. Context of discussion and subjectivity analysis provide better results in sentiment analysis. The sentences contain contextual information that has significance in polarity classification. Without contextual information, it is not easy to find whether the opinion is given as good or bad really. Thus it leads to difficulties in finding sentiments correctly. Context of words used in a sentence, therefore, is essential to be understood for better accuracy in sentiment classification. Hussein [56] studied challenges involved in sentiment analysis. In the process, the researcher identified the importance of using context in which expressions are made. Therefore, they have given importance to contextual polarity instead of using polarity in general sense of the word. In the process, preference is given to lexical dictionaries that provide contextual information as well.

Jianqiang et al. [57] investigated the discovery of latent contextual semantic relationships that provide useful information for better classification. With respect to Twitter tweets, context information is used for finding good information features that are used for better labelling. Semantics in the given tweet needs to be considered in the given context for more meaningful analysis prior to classification. Vectors are used for contextual representation of the tweets in order to analyse in better way.

2.1.7 TOOLS USED IN SENTIMENT ANALYSIS

SentiWordNet is a publicly available lexical resource that is used to achieve opinion mining (OM). This tool provides synsets with three numerical scores such as positive score, negative score and neutral score [58]. Protégé is used in [59] for building ontology required for sentiment analysis. SentiWordNet is used in [60] for polarity identification. ConceptNet has explored in [61] is used to have semantic network of concepts for sentiment analysis. Montejo-Raez et al. [62] used lexical dictionaries like WordNet and SentiWordNet for producing a ranked WordNet graph for building a classifier to analyze sentiment polarities. SentiWordNet is also used in [63], [64], [65] for sentiment classification. SentiWordNet tool is made available for many Indian languages [66].

Polarity detection tools are used in sentiment analysis as explored by Chaturvedi et al. [67]. There is optimization of such tools in order to distinguish from positive sentiments from negative sentiments. Hussein [68] on the other hand identified lexicon tools for sentiment analysis. SentiWordNet, WordNet, OpenNLP, Chinese Lexicon and Microblogging lexicon are used for better sentiment analysis. Jianqiang et al. [69] used Twitter corpora and used sentiment analysis tools like GloVe in order to improve classification of sentiments.

2.1.8 MEASURES USED IN SENTIMENT ANALYSIS

Kendall's tau coefficient [70] is used to know influenced probabilities and count of friends and influencing probabilities and count of followers. The former is known as out-degree while the latter is called in-degree in Twitter. The measure is as in Eq. (2.4).

$$\tau = \frac{(\sum_{i,j}[(i,j)_{in_same_order}] - (\sum_{i,j}[(i,j)_{in_different_order}])}{n(n-1)} \quad (2.5)$$

where n is the total number of users. All pairs of neighbours are considered for summation. The Kendall's tau is thus used to know both influencing and influenced probabilities in Twitter datasets. Bhaden et al. [71] explored measuring opinions used for sentiment analysis. In [72] precision and recall are the measures are used for evaluating event prediction mechanism with sentiment analysis. They measured causal rule prediction using two measures named Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Measure of Concern (MOC) is another measure used in sentiment analysis to determine the severity

of sentiments [73]. The evaluation metrics used for sentiment analysis in [74] are accuracy, precision, recall and F-measure. Cambria et al. [75] used context embedding and also used measures like Generalized Average Precision (GAP) to know the accuracy of the classification. There is another measure used named as Subsampled Randomized Hadamard Transform (SRHT) for determine training samples when the number of features are more than a threshold. SenticNet is the tool used to find polarity and that polarity is used in order to have better results. There are many similarities based measures as well for finding sentiment classification.

2.2 PREPROCESSING

This section reviews literature on pre-processing of Twitter tweets or the state of the art on text pre-processing in general.

2.2.1 PRE-PROCESSING TWEETS

Tweets have become goldmine for researchers and enterprises as they can help retrieve intelligence for a variety of business models. In this section, we provide insights of the filtering approach followed in this thesis and discuss its merits and demerits in connection with other research contributions found in the literature. From the flood of tweets, filtering interesting tweets based on user interests is the focus of the research in [76]. Filtering of tweets in this thesis is based on the query of user like the process in [76] where evolutionary timeline is used for filtering with given input as

$Q = q_1, q_2, \dots, q_{|Q|}$ and output as $\varepsilon = T_1^Q + T_2^Q, \dots, T_{(|\varepsilon|)}^Q$ respectively. Set of queries is denoted as Q . Output contains chronologically ordered tweets where S is denoted as tweet stream, ξ is the number of tweets in the evolutionary timeline while T_i^Q denotes a filtered tweet and associated with published date t_i^p and filtered date t_i^f . Our research on the other hand is multi-level filtering those results in tweets ready for processing. Twitter data analysis presented in [76] which is graph based and meant for finding in-degree centrality with respect to retweeting users is not suitable for multi-level filtering process of this thesis. Even Eigenvector centrality and between centrality are not directly suitable for our filtering.

The filtering work presented in [77] is close to our work in this thesis. They

used crowd-enabled process for filtering tasks while we used lexical analysis for semantic meaning of words for effective filtering process. Moreover, the filtering is made at three levels achieving highly accurate and desired tweets based on the brand or service for given attributes to be analysed. Labelling done in [77] for having ground truth looks not effective provided the millions of tweets that come from different sources with different style in writing. Semantic meaning of words with lexical dictionary for different levels of purposeful filtering is approach followed in this thesis. Event detection approaches explored in [78], [79] and [80] are intended to detect a pre-defined or ad hoc event in the tweets. The approaches used for event detection include unsupervised approaches, supervised approaches, and hybrid approaches. Our work comes under unsupervised machine learning approach based on the words used in filtering and the analysis made using semantic meanings of words taken from lexical dictionary. The content based filtering using classification approaches explored in [81] and distance and connectivity features used in [82] are techniques not viable for our work as our work focuses on identifying tweets related to given brand and attributes. The filtering approaches with respect to privacy in [83] with the help of representative words have certain similarity with our work. However, the work in this thesis is to filter tweets at different levels where some sort of intelligence is given to each filter to perform its intended task. The filtering process explored in [84] deals with temporal features. However, in our work, we have no concern with temporal information as of now. It deals straight with relevancy and possession of sentiments in the tweets. The filtering process based on categories, post formats and tags presented in [85] and based on public statuses as used in [84] are not strictly mandatory for our work.

2.2.2 MORE ON PRE-PROCESSING TWEETS

Taboada et al. [86] studied lexicon-based approach and proposed semantic orientation calculator. Semantic Orientation (SO) is a measure to find subjectivity or opinion in given text. We used lexical dictionary based approach to compute scores such as positivity, negativity and objectivity for synsets of given words programmatically. In this sense, the pre-processing work where we find whether a tweet has sentiment appears close to the work in [86] as far as dependence on lexicon-based approach. Wilson et al. [87] proposed an approach

for contextual polarity detection in given phrase prior to disambiguating actual polarity values. In our work during pre-processing of twitter tweets, we have not used polarity detection. However, we used lexicon-based procedure to know whether sentiment exists in a given tweet. Esuli and Sebastiani [88] explored SentiWordNet which is an extension to lexical dictionary WordNet. They used it to obtain three numerical sentiment scores for given synset such as Obj(s), Pos(s) and Neg(s) representing positive neutral score, positive score and negative score respectively. Thus they could use the statistics for quantitative analysis. In this thesis we also gained such scores by using Text blob technique in python programming to the advanced lexical dictionary prior to recognizing a tweet to have sentiment. it is an easy to use interface and is quite easy for a beginner to understand. If u want to work on basic Natural language tool kit(NLTK), Text blob is open source software, in fact text blob performs better than NLTK for textual analysis.

Lin and He [87] proposed a modelling framework based on Latent Dirichlet allocation (LDA). They called it as Joint Sentiment/Topic (JST) model. Their approach was supervised learning for making a model for sentiment analysis. Our focus is on pre-processing of tweets and finding tweets that have sentiments. Our approach is also unsupervised without the need for having corpora with pre-classified labels. Duwairi and El-Orfali [88] employed different pre-processing strategies based on k-NN, Naive Bayes, and SVM. They used RapidMiner tool in order to perform the pre-processing. In our thesis we found that we do not need to classify the tweets as the scope is limited to multi-level filtering and finally choosing tweets that are highly relevant to Indian railways with given attributes and having sentiment.

Yi et al. [89] proposed an analyzer to find sentiments in given text based on NLP techniques containing topic specific feature term extraction, sentiment extraction and relationship analysis. It makes use of sentiment lexicon and sentiment pattern database. The work is limited to have filtering and pre-processing in order to find tweets that has sentiment and most relevant to given brand or service. Therefore, our work is close to [89] only as far as the usage of lexicon database as concerned. The pre-processing techniques employed in [90] include replacing emoticons, uppercase identification, lower casing, URL extraction, detection of pointers, identification of punctuations, removal of stop words, removal

of query term, compression of words, and removing skewness dataset. While preparing dataset, we followed certain strategies such as converting to lower case, removal of URLs and so on. Angiani et al. [91] compared many pre-processing techniques used for sentiment analysis in Twitter. Data collection, pre-processing of data and finding attributes related to a service or brand besides containing given attributes is done in this thesis too as did in [91]. But there is no supervised learning process. In other words, there is no learning step with training data here as the tweets are pre-processed directly prior to sentiment analysis without the need for having a model built. Devika et al. [92] explored many sentiment analysis methods such as machine learning approach with SVM, n-gram sentiment analysis, Naive Bayes, Maximum Entropy (ME) classifier, kNN, multilingual, feature-driven, rule based ones, and lexical based approach. The focus is limited to filtering tweets and finding highly relevant tweets with given attributes and sentiment ready for sentiment analysis. Therefore, it only deals with lexicon-based approach for pre-processing which also includes natural language processing confined to obtaining tweets that are relevant to given brand or service containing sentiment. Munkova et al. [93] explored data pre-processing for text mining. They used stop words approach as pre-processing to sentiment analysis and analysed its impact on quantity of extracted rules, removal of inexplicable rules, and impact on quality of extracted rules. As stop words can reduce time and space complexity, this is used in the pre-processing. Haddi et al. [94] studied the role of pre-processing in sentiment analysis. In all their experiments, they found that pre-processing improved accuracy, precision, recall and F-measure of text categorization. Seeing the above facts, we considered pre-processing for reducing space and time complexity so as to increase quality of sentiment analysis later.

Advantage of pre processing It is a used in machine learning and data mining to make input data easier to work with dataset format. The advantages of data pre-processing are cleaning, editing, reduction, transformation. Drawbacks are The system is built for a single and specific task only; it is unable to adapt to new domains and problems because of limited functions and also the system may not be able to provide the correct answer it the question that is poorly worded

2.3 MACHINE LEARNING AND DEEP LEARNING

Machine learning and deep learning techniques are widely used in solving real world problems. This section reviews literature on the machine learning and deep learning for sentiment analysis. Wawre and Deshmukh [95] used Support Vector Machine (SVM) and Naive Bayes classification technique for sentiment analysis. They found that Naive Bayes showed better performance than SVM. Deng et al. [96] proposed a supervised term weighting scheme for sentiment analysis. It has two basic factors like importance of a term in a document (ITD) and for expressing sentiments (ITS). Khan et al. [97] proposed a hybrid classification scheme for opinion mining as part of a framework. They found that prior works on the sentiment analysis focused on classification accuracy, sarcasm and data sparsity. However, they classified many tweets incorrectly. To overcome this problem, they presented a hybrid method known as TOM (Tweet of Mining). Their methodology contains phases like acquiring tweets using Twitter API, pre-processing, classification and evaluation. The evaluation metrics used are accuracy, precision, recall and F-measure.

Narayanan et al. [98] proposed an enhanced classifier model based on Naive Bayes for sentiment classification. They found linear training and testing complexities with respect to execution time. Troussas et al. [99] on the other hand used Facebook statuses to learn a classifier using Naive Bayes for sentiment analysis. They explored the sentiment analysis for language learning. Dey et al. [100] proposed a hybrid method using k-NN classifier and Naive Bayes classifier to have sentiment analysis on review datasets. They found Naive Bayes to be better than k-NN. Preety and Dahiya [101] explored SVM and Naive Bayes algorithms to have sentiment analysis. Li [102] studied the significance of Naive Bayes algorithm in sentiment classification of Twitter tweets. MapReduce programming paradigm is used to have better performance. Matharasi and Senthilrajan [103] explored Naive Bayes using unigram approach on Twitter data to classify sentiments.

Yan et al. [104] improves SVM to work with Apache Spark framework that supports parallel processing. In addition to parallelizing SVM, they introduced Radial Basis Function (RBF) as kernel function for improving performance in terms of speed and accuracy. Kalarani and Brunda [105] proposed an ensemble

method using ANN and SVM for sentiment classification. They found the ensemble method to show higher accuracy than individual learning algorithms. Sinha et al. [106] used SVM and Naive Bayes classification of sentiments on ICON aircraft log. Sun et al. [107] proposed a multi-objective optimization function on SVM for sentiment classification using data from financial domain. They proposed Multi-Objective Genetic Algorithm (MOGA) that works in tandem with the SVM for performance improvement. Paramesha and Ravishankar [108] proposed a method for sentence level sentiment classification by exploiting dependency relations. They used SVM as classifier for achieving binary classification of sentiments. Sun et al. [109] used Fuzzy SVM to have sentiment classification. Their methodology reduced sensation of noise points besides making it to handle outliers while performing classification. Quan et al. [110] compared semantic similarity method and SVM classification for sentiment analysis. Similar kind of work is carried out by Ye et al. [111]. They employed SVM for emotion classification and found higher accuracy. Xia et al. [18] used SVM classification approach for analysing customer reviews to make sentiment classification. Shein and Nyunt [112] proposed a methodology that makes use of ontology and SVM classifier for sentiment classification.

Decision Tree [113] is used to classify data, usually into two groups per step, each of them being a class, which is defined based on a training data set provided. From the root of the tree, which is the node that contains all the training data set, the Decision Tree algorithm will go over all the features of that dataset and try to find which feature is the best to be used to split the data set into two (for binary split) or multiple subsets (for multiway split). The way Decision Tree algorithm defines the best feature is according to which feature gives the best information gain across all the split sets, Decision Tree can aid in the process of classifying a new data to determine whether it belongs to this class or the other by going down the tree with the feature of the new data. Should the new data that are being classified have a feature value not used or are new to the decision tree, there are two popular approaches to this, either (1) put all new data down to the child node with the highest number of instances, which is used in CART; or (2) put the new data down to all the child nodes with a weight proportional to the number of instances of each child, which is used in C4.5. CART and C4.5 are two popular decision tree algorithms.

As explored in [114], kNN is widely used for classification. It supports pattern classification and non-parametric in nature. It is simple but effective classification method. It has no need to know about data priori and needs no assumptions on the data as well. It is meant for finding k-nearest data points in the given training set. It is widely used in applications like loan disbursement, image recognition, healthcare, finance, political science, hand writing recognition, credit ratings and so on. It works based on feature similarity approach. In the name K-NN, the K means number of nearest neighbours which is the determining factor in the classification process. K-NN is widely used for prediction of class labels.

There are three important phases of the algorithm. They are known as computing distance from given point, finding the neighbours that are closest and vote for labels. The data point which gets more votes will be the class label for the newly arriving unlabelled instance. It is best used when number of features are limited. When number of dimensions is increased, it results in overfitting. To overcome this problem Principal Component Analysis (PCA) kind of for selecting required features. Determining number of features is not easier. It depends on the application in hand.

Random Forest (RF) [115] is another classification algorithm which follows ensemble classification approach. It is made up of many DTs. It was first developed in 1995 and the name was coined by Tin Kam Ho. RF combines the random selection of features and also the bagging idea of Breiman. Each decision tree which is part of RF is an individual learner. When they are combined, they become random forest. Data exploration is one of the common approaches for which RF is widely used. An example for decision tree used for RF is Classification and Regression Tree (CART). It follows a recursive, top down and greedy approach to divide the feature space into many regions. Then the stop condition is verified. If stop condition is satisfied, it ends after computing prediction error. If the stop condition is not satisfied, it builds the next split that is subjected to a series of operations namely choosing variable subset followed by an iterative process to choose the best split.

Li and Qian [116] proposed RNN language based LSTM for capturing sequencing information effectively and achieving multi-classification for attributes related to emotions in textual data. LSTM based approach is proved to be better than traditional RNN in terms of accuracy. Saraclar and Ozgur [117] explored

recurrent neural networks and SVM based LSTM to predict sentiments in political content. Miyazaki and Komachi [118] proposed tree structured LSTM with a mechanism that pays attention to sub-trees. It is used to achieve Japanese sentiment classification. Xu et al. [119] on the other hand explored cached LSTM (CLSTM) neural networks for performing sentiment classification at document level. This mechanism divides memory into multiple parts to enable the network to retain sentiment information in a better way in a recurrent unit. They used three datasets namely YELP 2013, YELP 2014 and IMDB for document level sentiment analysis. LSTM recurrent neural networks are also studied by Scherer et al. [120]. They combined bidirectional LSTM and Gated Recurrent Unit (GRU) to achieve this. Al-Smadi et al. [121] focused on aspect based sentiment analysis using LSTM deep neural networks on Arabic reviews. Based on LSTM, they defined two classifiers for achieving this. The first one is known as Bi-directional LSTM with CRF and the second one is aspect-based LSTM. The former is used to classify aspect opinion target expressions while the latter is used to have aspect sentiment polarity classification. Ain et al. [122] proposed deep learning techniques for sentiment analysis. Different neural network techniques are studied for useful insights. They found that deep learning techniques perform better than SVMs. Araque et al. [123] proposed Recurrent Neural Network (RNN) based on LSTM for sentiment classification of Spanish tweets. The combination of features improved the performance of sentiment classification. Liu et al. [124] also used RNN for multi-task learning. It learns across many tasks that are related. Training is made on the entire network jointly to have improved performance. Tang et al. [125] studied proposed two target dependent LSTMs where automatic consideration of target information could lead to effective sentence representation resulting in higher classification accuracy. From the literature it is understood that Naive Bayes, SVM and LSTM are widely used classifiers for sentiment analysis. Zhang and Liu [126] explored deep learning based methods for sentiment analysis. In the process, they analysed the LSTM method as a special kind of RNN method. They studied LSTM architecture and its functioning with Gated Recurrent Unit (GRU). They found that LSTM model for sentiment analysis is growing in usage in machine learning arena. Duric and Song [125] explored different feature selection methods that make use of Content and Syntax models to learn features from review documents automatically. The selected

features are then used for sentiment analysis. Hassan et al. [127] proposed a Bootstrap Ensemble framework for Twitter sentiment analysis. It was based on time-series data to produce strong positive and negative sentiment results. Tag based user profiles and resource profiles are used in [128] for incorporating sentiments in personalized searching. Bravo-Marquez et al. [129] used meta-features and sentiment dimensions in order to have Twitter sentiment classification with polarities, emotions and strengths. Service reviews online play vital role in decision making. Feature based approach for sentiment analysis is made in [130] for service reviews. It made use of Term Frequency – Document Inverse Frequency (TF-IDF) and linear regression to achieve this. Latent Dirichlet allocation (LDA) approach is employed in [131] for making an ontology tree to have better mining of opinions on online reviews. Since the LDA is probabilistic approach, it results in more robust classifications

2.4 SEMANTIC WEB ONTOLOGY

This section throws light into semantic web and ontology. It assumes significance as the semantic web technology plays crucial role in sentiment analysis.

2.4.1 ONTOLOGY

Thakor and Sasi [132] proposed a methodology named as Ontology-based Sentiment Analysis Process for Social Media content (OSAPS). It makes use of negative sentiments as well. They extracted Twitter tweets automatically and identified tweets containing negative sentiments. Then data clean-up is performed using GATE software. Afterwards, nouns and verbs are extracted besides removing duplicates and non-qualified verbs and nouns. Then ontology model is built using Protégé. They also proposed model retrieval of information from ontology model besides performing sentiment analysis. Penalver-Martinez et al. [133] used ontology to achieve feature-based opinion mining. Dataset containing user opinions is subjected to Natural Language Processing (NLP) that includes tokenization, splitting sentences, POS-Tagging, and Lemmatization. The resultant data is further subjected to ontology-based feature identification process that includes features research and score evaluation. Afterwards polarity identification is made using SentiWordNet before performing actual mining of opinions into

positives, negatives and neutral.

Kontopoulos et al. [134] performed ontology based sentiment analysis on Twitter posts. They found that text-based sentiment classifiers were inefficient due to non-consistent words in texts. They explored ontology-based techniques to have more efficient sentiment analysis. Instead of simply giving sentiment score to tweets, they used the concept of sentiment grade for each distinct notion in the given tweet. Zheng et al. [135] proposed an ontology based approach for video sentiment analysis. Their ontology approach is known as SentiPairSequence which helps in opinion mining. On the other hand, SentiBank is another ontology model proposed in [136] for detection of sentiment in visual contents. It is used for live sentiment prediction in video content. Lau et al. [137] proposed an ontology mining algorithm known as fuzzy product ontology mining for aspect-oriented sentiment analysis. An aspect is nothing but a product feature of consumer comment.

Sam and Chatwin [138] proposed an ontology-based solution for sentiment analysis of customer reviews pertaining to electronic products. Yasavur et al. [139] on the other hand proposed a methodology based on Named-Entity Recognizers (NERs) for analysing behavioural health. Lau et al. [140] proposed a novel ontology extraction method for understanding contextual sentiment knowledge. Product features can be extracted from domain ontology in order to identify relationships in a better way. Kherwa et al. [141] proposed an approach for comprehensive sentiment analysis to understand the hidden information from product reviews.

Polpiniz and Ghose [142] proposed ontology based sentiment analysis method for online customer reviews. Zhou and Chaovolit [143] proposed a methodology known as Ontology Supported Polarity Mining (OSPM). They used both supervised and unsupervised learning methods for achieving it. Cotfas et al. [144] built an application known as TweetOntoSense for semantic social media analysis based on ontology for sentiment analysis of Twitter tweets. It was able to provide complex feelings such as sadness, surprise, affection and so on. Ptaszynski et al. [145] proposed a robust ontology model for representing emotion objects. Ontology and Cases are used for making a model for sentiment analysis in [146]. Case based reasoning strategy was used to utilize historical cases for more efficient sentiment analysis. Patel and Madia [147] studied ontology based information

retrieval systems to know the utility of ontology in information systems. Liu et al. [148] proposed a system known as Complura for exploring large scale visual ontology that can be used for effective discovery of sentiments in visual contents. Izhar et al. [149] used big data to structure goal-based ontology for effective query processing in social media. NodeXL is used to analyze data collected from Twitter.

Logesh [150] explored automatic construction of topic ontology. It depends on acquisition of concepts and identifying semantic relationships. Towards acquisition of concepts a topic mapping algorithm is proposed. There is grouping of similar concepts using an algorithm known as semantic similarity clustering. Salatino et al. [151] studied ontology pertaining to computer science. They developed an algorithm towards this end. It considers concepts and relationships and generates graphical construction of ontology that reflects concepts and the relationship among the concepts. The constructed ontology is at basic level and its needs further enhancement.

Schriml et al. [152] developed a methodology for building human disease ontology that is used to analyse the diseases, to classify diseases and so on. This ontology construction helps in programmatically expansion and also data analytics in order to gain required business intelligence. Cheyer [152] opined that there needs to be a novel method for searching active ontology. His work is US patented as it has apparatus and method for performing search operations on active ontology. It makes use of semantic representation as part of the search.

Ontology has different applications in the real world. Subramaniaswamy et al. [154] proposed an IoT based healthcare system that is meant for generating recommendations pertaining to food with ontology-driven personalization. Their application is known as ProTrip which is a recommender system. It has many filtering strategies and personalization so as to meet the specific needs of individual users. It has Complex Event Processing (CEP) mechanisms and web ontology technologies and linked streamed data being used. It can process streaming and temporal data that is used to generate food recommendations. In future, they intended to work on social media and user level feedback for improving recommender system.

2.4.2 SEMANTIC ANALYSIS

Semantic analysis of tweets is made in [155] for understanding real time events and detecting the location at which event occurred. First of all, they used SVM classification of tweets before applying methods for event detection and location estimation. ConceptNet is a large semantic network that contains common sense concepts [156]. It is also used in semantic analysis of sentiments. Saif et al. [157] proposed a novel approach to identify similar contextual semantics and sentiments in tweets. Brooke [158] proposed a semantic approach for automatically analysing sentiments. The combination of Semantic analysis and machine learning approaches are used in [159] for Twitter data analysis to discover sentiments. They used different classification algorithms like SVM, Maximum entropy and Naive Bayes. Gonzalez et al. [160] exploited semantics in tweets to have more fine-grained sentiment analysis. They used NLP and semantic representation of data to achieve this. Wang et al. [161] proposed an access model known as Inverted XML access control which is based on semantic dependency analysis. Integrity constraints are considered along with semantic dependency. With semantic dependency among concepts, it became easier to explore the relationships among them. They developed an architecture with three layers namely data source layer, semantic filter layer and users layer. Ontology and semantic analysis are employed to achieve access control mechanism. They intend to improved it in future to make it flexible for dynamic update and prevent inference attacks.

Cheyre [162] developed a method and apparatus for the purpose of search ontology where semantic analysis is made for better results. The method receives a search string, split it into number of tokens, tokens are matched with active ontology concepts, then semantic representations are generated, database is searched with semantic representations and finally the desired results are obtained and they are presented to the end user.

2.5 KNOWLEDGE BASED SYSTEM

Knowledge based systems play pivotal role in the contemporary era. The technological advances help in realizing such systems. There has been significant effort towards it. Due to proliferation of various Internet applications as virtual platforms, knowledge is made available for enterprises over such media. For in-

stance, opinions of people can be used to build a knowledge based system. Gupta and Shalini [163] explored improvisation of experience in terms of gaining access to knowledge by defining a polarity dictionary for sentiment analysis. They quantified sentiments for analysis. However, they have not exploited semantic web components. Knowledge can be represented in the form of ontology. Ontology based techniques are proposed by Kontopoulos et al. [164] for sentiment analysis. By defining domain ontology, they could load and represent knowledge for enabling interface to M2M applications. However, a fully ontology-based solution is not realized in their work. Borth et al. [165] exploited ontology further to have large scale ontology with visualization of sentiments and emotions. They used more than 1200 concepts and a classifier known as SentiBank. Each concept is associated with Adjective Noun Pair (ANP) for stronger visualization of knowledge leading to multi-modal visualization.

Ontology can be used in building knowledge based system in any domain. With respect to customer reviews, Sam and Chatwin [166] proposed a knowledge model for ontology based opinion mining. Named-Entity Recognizers (NERs) became important for effective retrieval of knowledge. Polpinij and Ghose [167] on the other hand used lexical variable ontology for knowledge representation of consumer reviews. Yasavur et al. [168] proposed one such system based on ontology for healthcare domain. With NER, the system can automatically tag important words in sentences. However, dynamic distance thresholds for the system are not yet studied. Thakor and Sasi [169] proposed a novel framework named Ontology-based Sentiment Analysis Process for Social Media (OSAPS). It incorporates negative sentiments. Ontology based knowledge retrieval system is proved to be effective with ontology representation. The rationale behind this is that it is interoperable and machine readable.

Polarity refers to the associated sentiments that can be quantified. Zhou and Chaovalit [170] proposed a method known as Ontology-supported polarity mining (OSPM). This system enables effective retrieval and mining of polarities based on the concepts built using ontology. However, they were yet to explore multiple-property assignments with different level of membership. In the wake of unprecedented growth of micro-blogging services, ontology based opinion mining became essential. In this context, Cotfas et al. [171] proposed semantic social media analysis framework known as TweetOntoSense which is used rep-

resent knowledge using ontology in terms of sadness, anger, surprise, affection and happiness. There is notion of business ontology as well. Smeureanu et al. [172] proposed ontology named as business ontology to represent knowhow of a company's Corporate Social Responsibility (CSR). However, the system was not yet supporting automated access to knowledge.

Human emotions are also reflected in the form of opinions. In other words, opinions reflect human emotions also. Ptaszynski et al. [173] proposed a methodology to build ontology based solution to analyse human emotions. However, they are yet to explore formal objects of emotions for standardization. Thus it is understood that ontology based solutions make the systems to gain access to knowledge with interoperability. It is also reflected in the review made by Patel and Media [174]. Knowledge based approach is essential for different real world applications. Healthcare is one such domain where it is essential. For instance, knowledge based system might provide public health index and details across regions of the country. Ji, Chun and Geller [175] proposed such system for public disease monitoring. This kind of system is made accessible across the platform as it is made interoperable. The proposed intelligence model is known as Epidemic Sentiment Monitoring System (ESMOS). It is meant for automatically measuring disease concerns, measure of concern (MOC). This will help in understanding disease outbreak and help officers to make strategic decisions.

Apart from knowledge based systems, it is also possible to have ensemble approaches for sentiment analysis. Ensemble is nothing but union of multiple classifiers in order to have highly accurate analysis. Ankit and Saleena [176] proposed an ensemble classification system which can complement well to knowledge based systems that make use of ontology for machine readability. They opined that study of neutral tweets also provides some kind of intelligence. However, it was not yet explored by them. Such knowledge based system can be realized with ontology platforms. Arp et al. [177] explored and presented the way of building formal ontology which can help researchers to build knowledge based systems with M2M communication provision. Smith [178] provided the next version of basic formal ontology to enhance user experience. It is understood from the review of literature that knowledge based systems can be built using ontology. However, the reverse is also possible. The fact is that ontology can be built from knowledge based systems. Kharbat et al. [179] have demonstrated the same.

They opined that the business intelligence gained from data mining algorithms can be represented in the form of ontology. Thus they called it so. There are some contributions where ontology based systems are built for rail domain (transport). The system built in [180] provides access to ontology based knowledge related to railways. When data is stored in databases, generally they go inside a deep web. However, with knowledge based systems it is possible to query such data as well. Mei et al. [181] explored usage of RDF and OWL ontologies to realize such phenomenon. They studied ontology query answering on databases with SPARQL. This is somewhat closer to the work of this research. Like SPARQL, there are query languages that can work on ontology. Zhang and Miller [182] explored different languages that support queries on SWRL, OWL, RDFS and RDF. Out of these RDF is studied along with SPARQL for realizing knowledge based systems. Ortiz [183] also investigated the process of making queries on ontology based knowledge representations. The existing knowledge based systems found in the literature are useful in making interoperable queries and machine readable knowledge representations. However, there is need for having a complete system that converts and builds knowledge based systems from legacy systems or new systems to support machine readability and interoperable querying.

Tarus et al. [184] explored ontology – based e-learning system or recommendations. In fact, they investigated on the knowledge based recommendations that are more useful to end users. Towards this end, they employed ontology with collaborative filtering. Different hybridization techniques are used in order to get better results. They found that knowledge based recommendations play vital role in real world applications. For instance, e-Learning applications need knowledge based recommendations to ensure that the time taken for decisions is less and the decisions are made more accurately. They intend to study more hybridized methods in future for further analysis.

2.6 DISCUSSION AND SUMMARY

This chapter has presented a comprehensive survey highlighting the current progress, emerging research directions, potential new research areas and novel classification of state-of-the-art approaches in the field of Sentiment Analysis (SA). These analysis was carried out over five key aspects: (1) data source,

which is the collection of dataset used for sentiment analysis, (2) pre-processing methodology, which describes the pre-processing techniques applied on datasets received from the twitter post before giving it to the classification procedure, (3) classification procedure, which discusses the classifier algorithms used for classification using machine learning and deep learning techniques, (4) Semantic web approaches, which discuss several methods to make the domain ontology, and Finally (5) Knowledge based system to make the systems as well-informed business decisions. Social media has started generating unprecedented data which became a goldmine for researchers and business organizations. The proliferation of Online Social Networking (OSN) applications paved the way for people to socialize and opine freely. Thus the data on OSN assumed importance as the opinion of people can influence businesses or help them to know the sentiment of people. This knowhow can help organizations to have better strategies to promote customer satisfaction with increased Quality of Service (QoS). Sentiment analysis thus became an indispensable part of the decision-making systems of enterprises as it can provide required Business Intelligence (BI). Though there has been considerable research on this area, it is still open to opportunities, possibilities, and optimizations. It throws light into the present state-of-the-art of sentiment analysis and finds gaps in the research. It tries to cover the landscape of the subject in terms of lexicon-based methods, knowledge-based approaches, machine learning techniques, classification strategies, ontology-based methods, aspect oriented method and semantic approaches found in the literature in the study of sentiment analysis. This review also includes study on sentiment analysis in social media such as Twitter and Facebook, real-time applications of sentiment analysis besides various measures employed as part of sentiment analysis methodologies. It provides useful insights on various aspects of sentiment analysis that leverage Information Retrieval (IR), Data Mining (DM) and expert systems where BI is extracted and interpreted for making well-informed business decisions. This chapter has provided insights pertaining to sentiment analysis and other associated concerns. From the review of literature, it is understood that there is need for efficient pre-processing of textual documents, need for machine learning and deep learning techniques with feature selection, need for usage of ontology and make an efficient knowledge based system for sentiment analysis of Indian Railways (IR).

From the review of the literature, it is understood that there have been machine learning techniques widely used for sentiment analysis. Sentiment analysis using supervised learning methods like SVM. Text corpus used in sentiment analysis needs pre-processing. Many pre-processing approaches are found in the literature. There are filtering methods as well. Filtering methods are used to filter the tweets so as to find only the tweets that are suitable for pre processing or those are related to the scope of the research. There is ontology based approaches existing in the literature for representing concepts and relationships among them. An important research gap found in the literature, when it comes to the current research in the thesis, is that the pre-processing of Twitter tweets pertaining to IR needs to be improved. It does mean that pre-processing framework needs to be built for efficient pre-processing. There is need for intelligent filtering as well for finding tweets that are interest in this research. There is need for making ensemble of classifiers and the use of deep learning for more efficient model. On top of ontology and the concepts and relationships, knowledge based sentiment analysis is highly desired as the existing systems lack in this aspect. These research gaps are addressed in this thesis and the research carried out is presented in the subsequent chapters. Chapter 3 presents the proposed framework for achieving a knowledge based sentiment analysis system for IR tweets.

Chapter 3

PROPOSED FRAMEWORK FOR SENTIMENT ANALYSIS BASED ON INDIAN RAILWAYS TWITTER DATASET

This chapter is organized as follows. SECTION 3.1 presents the proposed framework for sentiment analysis based on Indian Railways Tweets. SECTION 3.1.1 presents the description of Web Scrapping. SECTION 3.1.2 presents the description of Pre-processing. SUBSECTION 3.1.2.1 presents the description of Natural Language Processing (NLP) techniques. SECTION 3.1.2.2 presents the description of Multilevel filtering. SECTION 3.1.2.3 presents the description of Topic based filtering. SECTION 3.1.3 presents the creation of Feature Selection. SECTION 3.1.4 presents the detail description of Classification. SECTION 3.1.5 presents the description of Sentiment Analysis. SECTION 3.1.6 presents the description of knowledge based system of sentiment analysis dataset using ontology. SECTION 3.2 summarizes this chapter.

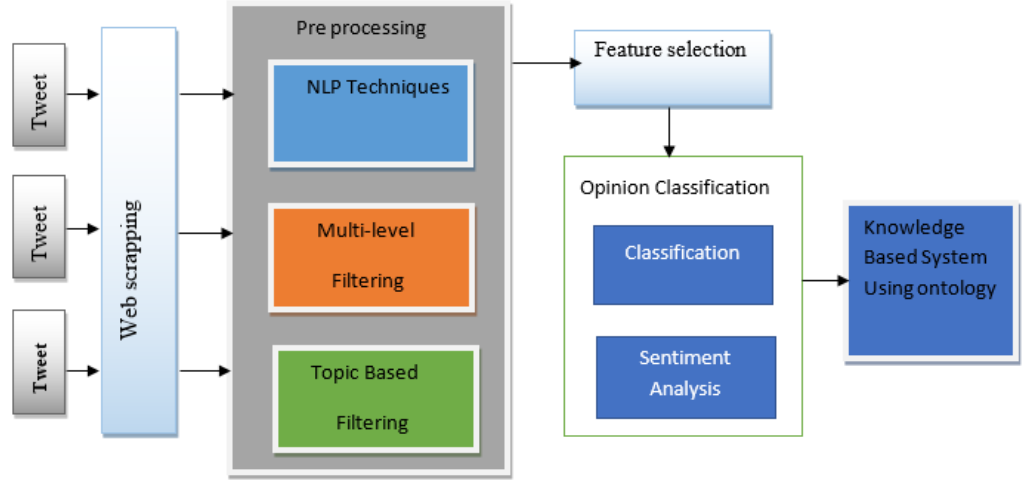


Figure 3.1: Framework for sentiment analysis of Indian Railways Twitter posts

3.1 FRAMEWORK FOR SENTIMENT ANALYSIS BASED ON INDIAN RAILWAYS TWEETS

Basically, our research work is an application based research in which we are implementing a knowledge based sentiment analysis for Indian Railways tweeter dataset. Therefore, to implement this a framework has been proposed which is shown in Figure 1. The framework facilitates to have various operations involved for sentiment analysis. The inputs for the framework are tweets that are from Twitter which is one of the famous social networking and virtual platform where people across the globe can participate in social networking activities. The tweets from Twitter contain wealth of information hidden. The information may be of anything including public opinion on politics, Indian railways, healthcare, etc. Therefore, the tweets provide ample scope to discover opinion of people.

Figure 3.1: Framework for sentiment analysis of Indian Railways Twitter posts The overall process of the framework is to take twitter tweets as input and discover knowledge and present it in such a way that it can help in making strategic decisions or to comprehend the current situation. The tweets obtained are subjected to the process of mining or sentiment analysis in order to produce meaningful propositions. The functionalities are the framework is logically grouped into various modules such as Web Scraping, Pre-Processing (within pre-processing sub-modules are Natural Language Processing (NLP) Techniques, Multilevel Filtering, and Topic Based Filtering), Classification, Sentiment Anal-

Table 3.1: Shows Positive and Negative Words Used for Sentiment Analysis

Positive Words	Negative Words
Good, okay, fine, appropriate, nice, lovely, clean, tasty, great, punctual, awesome, superb, fast, excellent, fantastic, marvellous, best.	Sad, late, bad, poor, ugly, dirty, unhygienic, damaged, rude, terrible, horrible, worst, useless, disastrous, pathetic, improper.

ysis, and Knowledge based System using Ontology. Its results in having tweets dataset that can be used for empirical study. For instance, when application in question is to perform sentiment analysis on the tweets related to Indian railways this module filters accordingly and provides tweets pertaining to Indian railways. For instance, with respect to Indian railways the attributes such as punctuality, cleanliness, staff behaviour, security and quality of food. These attributes are used to identify most relevant tweets. There is an issue here as the tweets might have different words with meaning of the attributes. For instance, the attribute punctuality is missing in a tweet. But the tweet has its synonym such as promptness. This is challenging and here there is need for a lexical dictionary to obtain synonyms of attributes on the fly. WordNet is the famous lexical dictionary used in this module to achieve the task. The result of this module is the removal of unnecessary words from tweets and identifying all tweets that contain words that are semantically similar to the attributes provided. Ontology is well known for its ability to represent knowledge in the form of concepts and the relationships among them. This module is responsible to construct ontology that can be used programmatically for sentimental analysis

Sentiment analysis is responsible to analyse the tweets and the knowledge in the tweets represented in the form of ontology. It makes use of the attributes semantically and finds various words that convey opinion or sentiment related to the attributes. It makes use of the positive and negative words provided in Table 1. This module results in finding statistics pertaining to sentiment or opinions. Classification is responsible to classify the tweets into three categories namely positives, negatives and neutral.

3.1.1 WEB SCRAPING

Web Scraping is the process of acquiring data from running web applications. In fact, it is the technique that became very useful for generating data from web applications. The generated data can be saved to local storage or it can be saved to relational databases or kept in tabular format in spreadsheets as well. Web Scraping has many other names too such as web harvesting, web data extraction and screen scraping. The data presented by most of the web applications provide read only access to users through web browsers. Those applications do not provide any feature to save a copy of data for use personally. Users can only do it manually and that is a tedious task. To overcome this problem, web scraping is the technique that came handy. Web scraping software automatically extracts data from multiple pages and load it to the local storage. The web scraper software may be a generic software product like WebHarvy [4] or a custom built targeting specific web site. In this thesis, we have used web scraping with Beautiful Soup library to extract data form twitter handler pertaining to Indian railways for sentiment analysis. This library has been used as it helps better than Application Program Interface(API) like Rest API and Streamed API. These API has restriction of data like getting of only previous one week and other dynamic through online streaming but according to our requirement of dataset web scraping with beautiful soup is suitable and reliable. It creates a parse tree for developing parsed pages that can be used to extract data form HTML, which is mainly useful for web scraping. This python library package is used for parsing HTML and XML documents which include having malformed mark-up that is non closed tags are so named after tag soup as Beautiful Soup [5]. Advantages of Beautiful soup are decent speed in parser, very fast in mark-up as “lxml”, only currently supported XML parser and it creates valid HTML5. Disadvantages not as fast as lxml, less lenient than html5lib, External on C, Python languages dependency.

3.1.2 PREPROCESSING

Before analysis of sentiment of dataset, it first pre-proceed. It is a used in machine learning and data mining to make input data easier to work with dataset format. The advantages of data pre-processing is cleaning, editing, reduction, transformation During literature survey we came across some techniques that is

used to pre-process the dataset extracted from online social networks in which NLP is one of them. During survey we analyse that NLP approach is essential to process the dataset but some other approach should also be added with NLP so as to refine the dataset for well analysis of sentiment. We have also found some drawbacks while using only NLP approach. The drawbacks are as follows. The system is built for a single and specific task only; it is unable to adapt to new domains and problems because of limited functions and also the system may not be able to provide the correct answer if the question that is poorly worded. Therefore, to refine the dataset and also to minimize the above said drawbacks we added two new concepts in this pre-processing modules i.e. multilevel filtering and topic based filtering. The details are placed at subsection 3.1.2.2 and 3.1.2.3.

3.1.2.1 NATURAL LANGUAGE PROCESSING TECHNIQUES

It has been used to pre-process the dataset. There are lot of techniques that come under NLP. such as Grammar induction, Lemmatization, Part of speech tagging, parsing, Stemming, Tokenization, Named Entity recognition etc. But the techniques that has been used in our research work are Tokenization, Stemming, Lemmatization, and Stop Word Removal. After removing unnecessary tags, Tokenization is essentially splitting the sentence into smaller units of an entire document text. Each of these smaller units is called as tokens for example “waiting for train 05036 late by 4 hours” its tokens are waiting,for,train,05036,late,by,4, hours. This is important because the meaning of the text could easily be interpreted by analysing the words present in the text. Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words know as stemma. For example, to find out the root of the word in waits, waiting, waited the stem is wait. It helps the removal of ambiguity in words. Lemmatization normally aiming to remove inflectional endings only and to return the base or dictionary form of a word which is known as the lemma as “Waiting”. Thus Stemming and lemmatization is to reduce inflectional forms and sometimes derivationally to related forms of the word to a common base form. Further, stop words are the most common words in any natural language to analyse text data, these stop words might not add much value to the meaning of the document. Stop words for above example are “for”, “by”. Applying to stop word removal in NLP techniques takes less time for processing datasets.

3.1.2.2 MULTILEVEL FILTERING

We further process the output data of NLP techniques using multilevel filtering approach which we have introduced in our research work. As the dataset getting from NLP Techniques have some drawbacks such as the system is built for a single and specific task only; it is unable to adapt to new domains and problems because of limited functions and also the system may not be able to provide the correct answer within the question that is poorly worded. Therefore, multilevel filtering has been introduced to minimize the said drawbacks. Multilevel filtering includes Inclusive Filter, Relevance Filter, and Opinion Filter. The inclusive filter includes the Indian railway's domain with the help of an official twitter handler. Whereas Relevance filter is used to filter the sentence with relevant features and their synonyms used in our research work. Finally, the Opinion filter is used to predict the tweet containing opinion are using Senti WordNet lexicon. Considering the above example of NLP techniques, "train number" in above tweets tells about the inclusive filter, whereas "late" in above tweets tells about relevance filter, finally "waiting" in above tweets gives opinion of the Indian railways tweet as late which belongs to feature of punctuality comes under opinion filter.

3.1.2.3 TOPIC BASED FILTERING

Further the output dataset of multilevel filtering is again processed using topic based filtering which we have introduced. As multilevel filtering drawbacks are being one type of data pre-processing still some tweets are difficult to classify input to machine learning classification, there is an ambiguity and unclear to convert text data to numerical data.

To minimize the above said drawbacks we now introduced topic based filtering. It includes Topic selection, Lexicon building, Concept extraction, Subjectivity detection, Polarity detection, and Ordinal classification. Topic selection means topics are to be understood by the system correctly, every tweet of every word should compare with topic selection and its synonyms. If it is matched with either topic selected or its synonyms, then it is not removed from the tweet list. Otherwise, it is removed. Lexicon is the list of stems and affixes, together with basic information about them, lexicon building is WordNet using Synsets. Lexicon building is developed by WordNet dictionary which supports Sentiwordnet for sentiment analysis. The concept extraction phase is used to identify concepts

based on the topics chosen. It is important to deconstruct given text or tweet into the concept for better semantic-aware analysis with WORDNET synsets. The next phase is subjectivity detection. It is an NLP task that removes neutral or factual tweets from the given tweets. The factual content is also known as object content that does not reflect any opinion. This is important to increase the accuracy of the sentiment analysis. Polarity detection phase is crucial for sentiment analysis. In this phase, the tweets do have one of the given topics and sentiment is used as input to find the label of a tweet whether it is positive, negative, and neutral. Ordinal classification is also known as ordinal regression which finds error rate between tweets. As misclassification is costly measures such as the overall sentiment of the tweet, topic-based classification, and tweet quantification are used for evaluation. Example “waiting for train 05036 late by 4 hours” after pre-processing as NLP techniques we get “waiting”, “train”, “05036”, “late”, “4”, “hours” then further pre-processing as multilevel filtering where it includes Inclusive filter as “train”, Relevance filter as “late” finally “waiting” as Opinion filter. Thereafter, applying topic based filtering where it includes topic selection as “train”, polarity as “negative”, concept extraction as “late”, subjectivity detection as “waiting” further we find as ordinal classification as error rate in above tweet.

3.1.3 FEATURE SELECTION

Feature selection is the process where you automatically or manually those select features which contribute to most of our decision variable or output in which we are interested in. Having irrelevant features in your data can decrease our accuracy of the models and make our model learn based on features irrelevant. Carrying above example of dataset of Indian railways. After applying feature selection on pre-processing dataset we get numerical format of dataset such as domain of railways as 1, train number as 05036, attribute as punctuality as 1, opinion as 1, sentiment as negative (-1). In this thesis, entropy and gain based representative feature selection is carried out. Entropy and Gain are the two statistical measures used to make decision pertaining to removal of irrelevant features and choosing relevant features that are relevant to the chosen concept. These two are also used to determine correlation between two features in a given data set. Entropy characterizes impurity of a feature while the Gain

is the expected reduction in entropy where examples are partitioned according to given feature. Features are many synonyms based on Indian railway tweets, we apply classification on top ranking features which gives better results for our input data set for sentiment analysis. We created a feature selection using text blob technique. It is a simplified text processing python library for pre-processing textual data. It provides a simple API for diving into common natural language processing tasks such as part of speech tagging, noun phrase extraction, value of sentiment analysis, translation and more. In feature selection by applying text blob the above tweet “waiting”, “late” gives negative opinion of the tweet in punctuality domain with polarity value as “-1” which further given input to the classifier for machine learning classification.

3.1.4 CLASSIFICATION

Machine Learning (ML) is learning phenomenon used by a computer program to gain knowledge and make some important tasks like prediction or forecast. That is the rationale behind the fact that ML is part of Artificial Intelligence (AI). In fact, ML is used to automate data analysis and to build knowledge models that are used to make well informed decisions later on. ML techniques learn from data (gain knowledge), identify trends or patterns and make intelligent decisions with minimal or without human intervention [7]. In machine learning there are lot of classifiers which are used to classify the dataset such as Decision Tree Classifiers, Linear Classifier containing Support Vector Machine, Neural Network, Rule based Classifiers, Probabilistic Classifiers containing Naive Bayes, Bayesian Network and Maximum Entropy which comes under supervised learning . But in this thesis we have used machine-learning classifiers such as Naive Bayes, Support Vector Machines (SVM) and Deep learning techniques such as Long Short Term Memory (LSTM). because our dataset suitable for this classifiers with results, performance, concepts, accuracy after running application we dedicated to go further with below classifier. Naïve Bayes classifiers a family of simple probabilistic classifiers based on applying Bayes theorem with strong independence assumptions between the features of my Indian Railways Application. A Support Vector Machines(SVMs) is a supervised models of machine learning with associated learning algorithms is to find a hyperplane in an N dimensional space (N – the number of features) that distinctly classifies the data points. Deep learning

technique as Long Short Term Memory(LSTM) is an Artificial Recurrent Neural Network (A RNN) architecture used in the field of deep learning. LSTM cannot only process single data points but also entire sequence of data. We have also introduced ensemble classifier with stacking for classification of dataset. Ensemble is nothing but combining two different classifiers with Stacking. Stacked generalisation is an ensemble method where a new model learns how to best combine the predictions, decisions from multiple existing models of training. In our research work the ensemble classifier is constituted of Naïve Bayes with LSTM and SVM with LSTM. We have ensemble supervised learning with deep learning which is of heterogeneous mixture where paralleled programming is executed. We used only this algorithm because after analysis of our dataset we found these algorithms are suitable. Unlike ML techniques in the past, the modern ML techniques are more accurate with new computing technologies. ML is interactive in nature as it needs to gain knowledge from the new data that is arrived. ML techniques also learn from past computations to generate repeatable decisions reliably. In fact, it is the science that is widely used across industries now. ML algorithms are available for many years but of late they got ability to apply complex mathematical computations for solving many real world problems. For instance, self-driving Google car [8] is an example where ML is used. Another specific example of machine learning is the study of opinions that are in the form of Twitter posts. In this thesis, ML techniques with supervised learning are used to arrive at BI needed by Indian Railways. IR gets social feedback from tweets in order to improve its services in specific areas.

3.1.5 SENTIMENT ANALYSIS

The proliferation of micro-blogging websites and social media has brought opportunities for businesses to know the opinion of people on their products and services. Apart from traditional feedback, the opinion of people over social media became important to understand the needs of customers and ensure customer loyalty. In today's scenario, twitter has been reflected as the most popular dataset for research where we can extract short and authentic information. Nowadays, almost every organization manages twitter accounts to get the opinion of people about its services in which Indian Railways is one of them. In the Indian railway, there are a lot of services that are provided to the people while travel-

ing. However, in our research work, we have explored different services such as punctuality, security, cleanliness, food quality, and staff behavior for sentiment analysis. Sentiment analysis refers to the use of Natural Language Processing (NLP), text analysis and computational linguistics to identify and extract subjective information in source material to study of people's opinion, attitudes and emotions towards an entity with the computational study. Sentiment Analysis is widely applied to reviews over social media for a variety of applications ranging from marketing to customer service.

3.1.6 KNOWLEDGE BASED SYSTEMS USING ONTOLOGY

The results are grouped into clusters based on five attributes pertaining to IR such as Cleanliness, Staff Behaviour, Punctuality, Security and Food Quality. These clusters are updated from time to time to reflect up to date social feedback. However, the problem with the existing system is that, its accessibility and ease of use to stakeholders is not easy, very difficult to address complex queries over the data and also it comes back with an answer at most once, lagging inherent properties such as transitivity or symmetry. It also has closed world assumption i.e., what is not known to be true in the database is by default considered as false because knowledge represented in the database is assumed to be complete. To overcome this problem, we proposed a knowledge based system which will represent clusters for a universal and interoperable data representation that is ontology based RDF schema. This system is accessible to humans and also programs in heterogeneous Machine-to-Machine (M2M) environments. The knowledge-based system has been developed using an ontology, which is a semantic web technique. Ontology is used in the knowledge-based system as a conceptual framework for Providing, Accessing, comprehensively structuring information. We proposed and implemented a knowledge-based system using an ontology, for accessing sentiments data of Indian Railways tweets. It has been developed by transforming relational data model to semantic data model. In other words, it is responsible to convert Business Intelligence (BI) into a knowledge based, accessible or programmable system that can be directly queried or queries through a program in an interoperable way. Before we create ontology firstly we have to convert Relational Database schema (RDS) to Resource Description

Format Schema (RDFS) which is a standard format to store data. It forms semantic web and Web Ontology Language (OWL) for interpretation of data. RDF schema has both literals and semantic meanings of the same. Thus it provides rich interoperable interface to a knowledge domain in fully automated and interactive fashion. The concept of ontology has made this possible as ontology is the knowledge representation which is made up of concepts and relationships among them. RDF schema was introduced in order to handle situations where web data needs to be processed and exchanged by applications instead of just showing data to users. This ability of exchanging data between applications makes RDF very useful in the contemporary era. Semantic data models like RDF and frameworks based on them can help organizations of specific domain to organize domain data or knowledge and share it throughout the enterprise. Such models ensure that semantic means of accessing knowledge leads to fully automated systems in distributed environments. We have also introduced and implemented an algorithm that has been used in Database to map Tables, Columns, and Constraints. As Tables have been mapped into classes, whereas Columns have been mapped to Data Type Property within domain and range, and Constraints have been mapped as a subclass of Object property with Restrictions. Further this schema is mapped into ontology which applies knowledge rules that have been developed to execute the query using Sparkle Protocol and RDF Query Language (SPARQL). The detail design with implementation of knowledge based system on sentiment dataset of railways is given in chapter 6.

3.2 SUMMARY

This chapter introduced the modules used to design the framework of sentiment analysis based on Indian Railways twitter dataset. The framework is constituted with the modules called Web scrapping, pre-processing, feature selection, classification, sentiment analysis, and knowledge based system using ontology. Web scrapping is used for collection of twitter dataset of Indian railways, pre-processing module is used for cleaning of dataset. This module is further divided into submodules called NLP techniques, multilevel filtering, and topic based filtering which is further used for data reduction, and transformation. Feature selection gives input to the classification to find sentiment analysis, and finally,

knowledge based system using ontology used for decision support system for Indian railways. In next chapter, we have presented the detail description with results about pre-processing of the dataset that has been extracted from twitter for sentiment analysis.

Chapter 4

PRE-PROCESSING OF INDIAN RAILWAYS TWEETS FOR SENTIMENT ANALYSIS

This chapter is organized as follows. SECTION 4.1 presents the frame work of pre-processing of Indian Railways tweets for Sentiment Analysis. SUBSECTION 4.1.1 presents the web-scrapping procedure on tweets with results using Beautiful Soup python library. SUB SECTION 4.1.2 presents the Natural Language Processing (NLP) techniques applied on extracted tweets with results. SUBSECTION 4.1.3 presents the description with results of multilevel filtering techniques applied on NLP output. SUBSECTION 4.1.4 presents the description with results of Topic based filtering techniques applied on the output of multi-level filtering. Finally, SECTION 4.2 summarizes this chapter.

4.1 FRAME WORK OF PRE-PROCESSING OF INDIAN RAILWAYS TWEETS FOR SENTIMENT ANALYSIS

Sentiment analysis based on Twitter tweets is a significant research area identified. As sentiment analysis is widely used for extracting business intelligence from social media content, that is why it became an emerging research area. However, there is a problem with generating quality data sets that are domain specific for effective sentiment analysis. Assume that a data set D has thou-

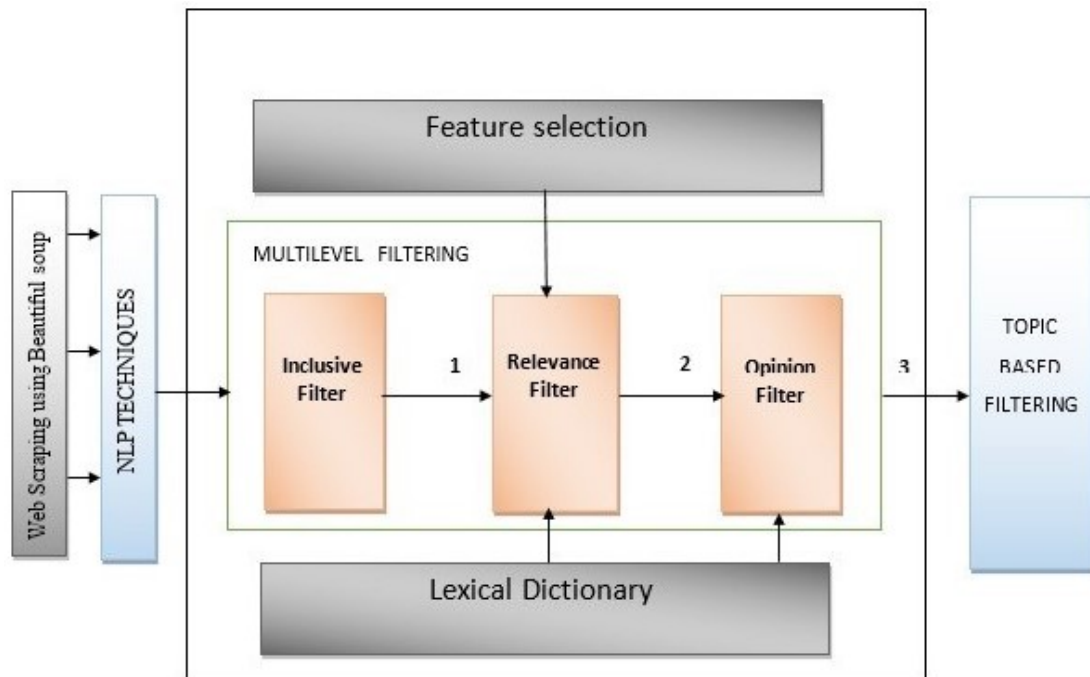


Figure 4.1: Frame work of Pre-processing of Indian Railways tweet for Sentiment Analysis.

sands of tweets. It is collected from Twitter streams. Unless, the data set D is subjected to pre-processing, it is not possible to have quality inputs suitable for sentiment analysis. It is tedious task to produce D' from D after pre-processing. The problem of automatic generation of quality data set from Twitter tweets is non-trivial and hence it needs methodology for it. Domain specific data set creation for sentiment analysis is a challenging problem for meaningful and comprehensive mining for discovering useful information. In the contemporary era, social networking became predominant and traditional business intelligence is inadequate. The traditional intelligence coupled with the sentiment analysis or opinion mining made on the related data associated with social networking sites can provide more comprehensive business intelligence that can be used effectively for business strategies. Instead of reinventing the wheel, this chapter throws light how to generate domain specific qualitative data sets for sentiment analysis. For this we first have to pre-process the dataset that has been extracted from twitter.

As we are performing sentiment analysis on Indian Railways dataset therefore we have proposed the framework of pre-processing of Indian Railways tweet for sentiment analysis as shown in Figure 4.1.

In the above framework the first module is Web scraping which is used for

extraction of data from twitter. For web scraping beautiful soup library using python techniques has been used. Twitter web scraping using beautiful soap library helped to obtain domain specific tweets by only considering Indian Railways Twitter accounts such as @railminindia and @Indian railways. The result of such process contains tweets pertaining to Indian Railways with specific period. After collecting tweets, they are subjected to module called Natural Language Processing (NLP) that includes tokenization, stemming, lemmatization, and stop words removal. Tokenization is used to split the sentence into smaller units of an entire document text. Each of these smaller units is called tokens. This is important because the meaning of the text could easily be interpreted by analysing the words present in the text, whereas stemming is used to reduce a word to its word stem that affixes to suffixes and prefixes or to the roots of words know as stemma. For example, to find out the root of the word user, users, used, using the stem is used.

It helps the removal of ambiguity in words. It also reduces inflected words and considers base or root of a word, further lemmatization is the process of grouping words to have representative ones according to dictionary. Both stemming and lemmatization can reduce space and time complexity. Finally, stop word removals removes standard stop words from the tweets that do not have any bearing in the sentiment analysis, removing words like ‘a’, ‘an’, ‘about’, ‘across’ and so on can reduce search space and thus reduces space and time complexity in text mining. But there are some drawbacks of NLP techniques as the system is built for a single and specific task only it is unable to adapt to new domains and problems because of limited functions and also the system may not be able to provide the correct answer of the question that is poorly worded. Therefore, we have proposed the module called Multilevel filtering which minimizes the above said drawbacks. After NLP techniques, the resultant tweets are subjected to multilevel filtering.

This filtering modules includes inclusive filter, relevance filter and opinion filter. The inclusive filter includes the Indian railway’s domain with the help of an official twitter handler. Relevance filter is used to filter the sentence with relevant features and their synonyms used in our research work. This filter used the WordNet lexicon [12] to find the synonyms. Finally, the opinion filter is used to predict the tweet containing opinion are not using Sentiwordnet Lexi con [11]. After multilevel filtering, the resultant tweets are subjected to topic based

filtering module for further pre-processing of twitter tweets. This module has been proposed because the tweets coming from multilevel filtering are difficult to classify input to machine learning classification, there is also an ambiguity and unclear to convert text data to numerical data.

Topic based filtering includes Topic selection, Lexicon building, Concept extraction, Subjectivity detection, Polarity detection, and Ordinal classification. Topic selection means topics are to be understood by the system correctly, every word of every tweet should compare with topic selection and its synonyms. If it is matched with either topic selected or its synonyms, then it is not removed from the tweet list. Otherwise, it is removed. Lexicon is the list of stems and affixes, together with basic information about them, lexicon building is WordNet using Synsets. Lexicon building is developed by WordNet dictionary which supports Sentiwordnet for sentiment analysis. The concept extraction phase is used to identify concepts based on the topics chosen. It is important to deconstruct given text or tweet into the concept for better semantic-aware analysis with WordNet synsets. The next phase is subjectivity detection. It is an NLP task that removes neutral or factual tweets from the given tweets. The factual content is also known as object content that does not reflect any opinion. This is important to increase the accuracy of the sentiment analysis. Polarity detection phase is crucial for sentiment analysis. In this phase, the tweets do have one of the given topics and sentiment is used as input to find the label of a tweet whether it is positive, negative, and neutral. Whereas ordinal classification is used to find error rate between tweets. It is also known as ordinal regression. As misclassification is costly measures such as the overall sentiment of the tweet, topic-based classification, and tweet quantification are used for evaluation.

4.1.1 WEB-SCRAPPING USING BEAUTIFUL SOUP LIBRARY

Web Scraping is the process of acquiring data from running web applications. In fact, it is the technique that became very useful for generating data from web applications. The generated data can be saved to local storage or it can be saved to relational databases or kept in tabular format in spreadsheets as well. Web Scraping has many other names too such as web harvesting, web data extraction and screen scraping. The data presented by most of the web

applications provide read only access to users through web browsers. Those applications do not provide any feature to save a copy of data for use personally. Users can only do it manually and that is a tedious task. To overcome this problem, web scraping is the technique that came handy. Web scraping software automatically extracts data from multiple pages and load it to the local storage. The web scraper software may be a generic software product like WebHarvy [4] or a custom built targeting specific web site. In this thesis, we have used web scraping with Beautiful Soup library to extract data form twitter handler pertaining to Indian railways for sentiment analysis.

This library has been used as it helps better than Application Program Interface(API) like Rest API and Streamed API. These API has restriction of data like getting of only previous one week and other dynamic through online streaming but according to our requirement of dataset web scraping with beautiful soup is suitable and reliable. It creates a parse tree for developing parsed pages that can be used to extract data form HTML, which is mainly useful for web scraping. This python library package is used for parsing HTML and XML documents which include having malformed mark-up that is non closed tags are so named after tag soup as Beautiful Soup [5]. There are certain advantages of Beautiful soup library such as decent speed in parser, very fast in mark-up as “lxml”, only currently supported XML parser and it creates valid HTML5. Twitter web scraping using beautiful soap library helped to obtain domain specific tweets by only considering Indian Railways Twitter accounts such as @railminindia and @Indian railways. The snapshot of Twitter Handler of Indian Railways on which web scrapping techniques using beautiful soup library has been applied for data extraction are shown in figure 4.2 and snapshot of data extracted using beautiful soup library are shown in figure 4.3.

In the above snapshot we are getting tweets after web scraping using beautiful soup library. In this tweets the people are commenting on attributes like cleanliness, staff behaviour, food quality, punctuality, and security.

4.1.2 NATURAL LANGUAGE PROCESSING (NLP) TECHNIQUES

The resultant tweets contain one of the attributes such as staff behaviour, cleanliness, food quality, security, punctuality or their synonym semantically



Figure 4.2: Snapshot of Indian Railways Twitter Handler

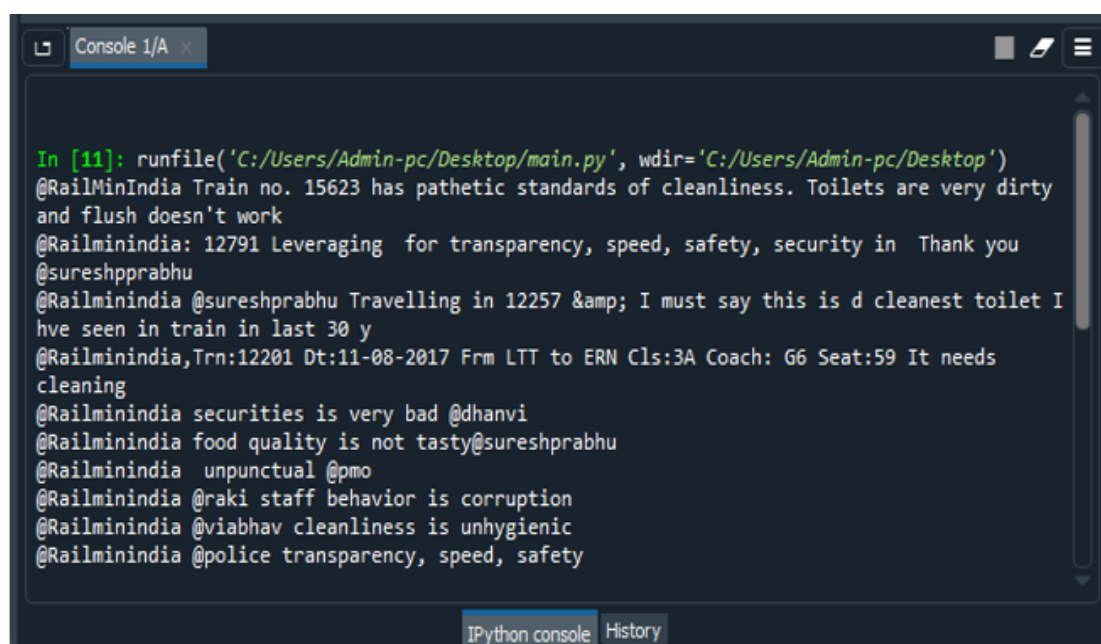
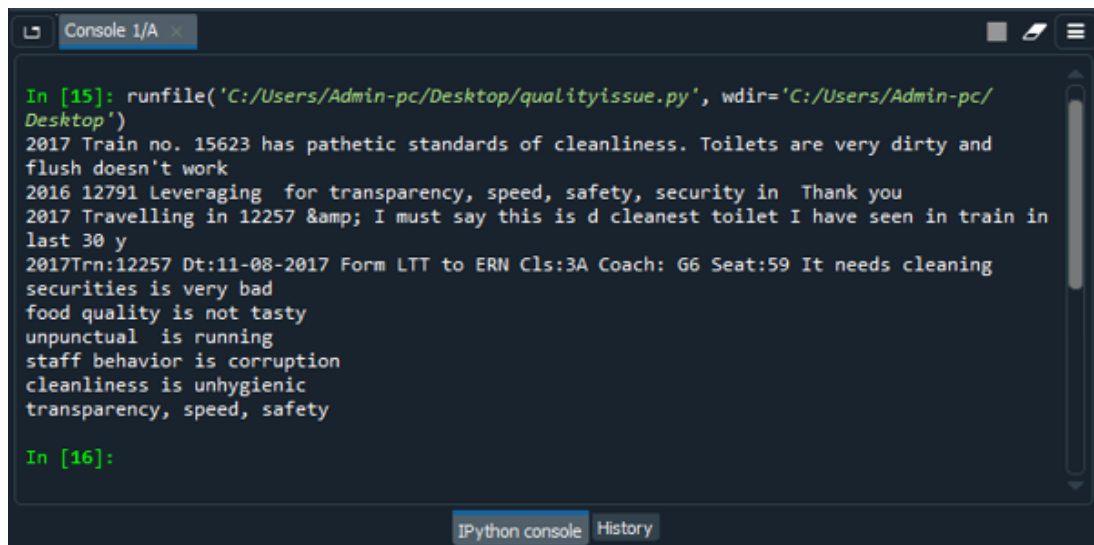


Figure 4.3: Snapshot of data extracted using beautiful soup library



```
In [15]: runfile('C:/Users/Admin-pc/Desktop/qualityissue.py', wdir='C:/Users/Admin-pc/Desktop')
2017 Train no. 15623 has pathetic standards of cleanliness. Toilets are very dirty and flush doesn't work
2016 12791 Leveraging for transparency, speed, safety, security in Thank you
2017 Travelling in 12257 & I must say this is d cleanest toilet I have seen in train in last 30 y
2017Trn:12257 Dt:11-08-2017 Form LTT to ERN Cls:3A Coach: G6 Seat:59 It needs cleaning
securities is very bad
food quality is not tasty
unpunctual is running
staff behavior is corruption
cleanliness is unhygienic
transparency, speed, safety

In [16]:
```

Figure 4.4: Snapshot of dataset after removing unnecessary data for pre-processing

meaningful words. We removed unnecessary data like tagging, hyperlinks, and emoticons as we are not considering. The snapshot of dataset after removing unnecessary data for pre-processing is shown in figure 4.4.

The resultant tweets after removing unnecessary data is further pre-processed by using NLP techniques. The techniques that has been applied on the extracted dataset are Tokenization, Stemming, Lemmatization and Stop word removal. Tokenization helps essentially splitting the sentence into smaller units of an entire document text. Each of these smaller units is called as tokens. While stemming reduces inflected words and considers base or root of a word. Thereafter, lemmatization is the process of grouping words to have representative ones. Both stemming and lemmatization can reduce space requires it's not costly now but time has costly time complexity in further processing. Finally, stop word removal help to removes standard stop words from the tweets. It reduces computational and time complexity while the tweets are used for further processing. Stop words removal is very simple algorithm that has an iterative process to simply remove specific words. Yet it is not part of actual filtering of the tweets.

```

In [17]: runfile('C:/Users/Admin-pc/Desktop/tokens.py', wdir='C:/Users/Admin-pc/Desktop')
2017 Train no. 15623 has pathetic standards of cleanliness. Toilets are very dirty and flush doesn't work
TOKENS are train number,15623,pathetic,standards,cleanness,very,dirt,flush,doesnot,work
Stemma are standard,clean,Lemma are standards,cleanness,dirt,flush,work,Stop word removal
has,of,are,and
2016 12791 Leveraging for transparency, speed, safety, security in Thank you
TOKENS are 12791, leveraging,transparency,speed,safety,security,Thankyou,Stemma are
leverage,transparency,speed,safety,security,Lemma are leveraging,Stop word removals for,in
2017 Travelling in 12257 I must say this is cleanest toilet I have seen in train in last
30 y,Tokens are Travelling,12257, say, cleanest, toilet, last,30y,Stemma are
travel,clean,Lemma are Travelling,cleanest,stop word removal in,I,must,thisis,have,seen
2017 Trn:12257 Form LTT to ERN Cls:3A Coach: G6 Seat:59 It needs cleaning,Tokens are
12201,LTT,ERN,Cls 3A,coach,G6 seat 59, needs, cleaning,Stemma are need,cleaning,Lemma are
needs,cleaning,Stop words removal form,It,to
securities is very bad ,Tokenssecurities,very ,bad,Stemmasecurity,Lemmasecurities,Stop word
removal is
food quality is not tasty ,tokens food quality,not,tasty,stemmataste,lemmatasty,stop word
removal is
unpunctual is running, tokens unpunctual,running,stemmaruns,lemma running ,stop word
removal is
staff behavior is corruption, token staff
behavior,corruption,stemmacorrupt,lemmacorruption,stop word removal is
cleanliness is
unhygienic,tokenscleanliness,unhygienic,stemmaclean,unhygiene,lemmacleanliness,unhygienic,s
top word removal is

```

Figure 4.5: Snapshot of dataset after applying NLP techniques

Algorithm 1 Stop Words Removal (SWR)**INPUT:** Tweets T , stop words S **OUTPUT:** Tweets without stop words

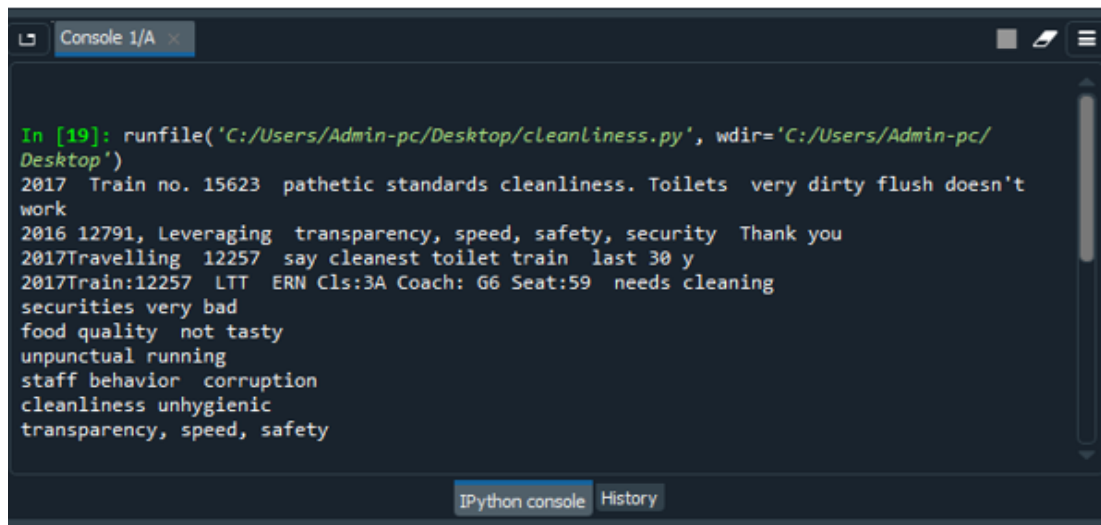
```

for each tweet  $t$  in  $T$  do
  for each word  $s$  in  $S$  do
    if  $s$  is found in  $t$  then
      Remove that word from  $t$ 
    end if
  end for
end for

```

As shown in Algorithm-1, the tweets obtained from Twitter accounts of Indian Railways are taken and stop words are removed from each tweet. The snapshot of resultant dataset after applying NLP techniques is shown in figure 4.5.

The above snapshot reflects the result of NLP techniques after applying Tokenization, Stemming, Lemmatization, and Stop word removal. In this snapshot, tokens of first tweet are train number,15623, pathetic, standards, cleanliness and etc., whereas stemma are standard, clean, and lemma are standards, cleanliness,



```

In [19]: runfile('C:/Users/Admin-pc/Desktop/cleanliness.py', wdir='C:/Users/Admin-pc/Desktop')
2017 Train no. 15623 pathetic standards cleanliness. Toilets very dirty flush doesn't
work
2016 12791, Leveraging transparency, speed, safety, security Thank you
2017Travelling 12257 say cleanest toilet train last 30 y
2017Train:12257 LTT ERN Cls:3A Coach: G6 Seat:59 needs cleaning
securities very bad
food quality not tasty
unpunctual running
staff behavior corruption
cleanliness unhygienic
transparency, speed, safety

```

Figure 4.6: Snapshot of dataset after having inclusive filter

dirty, flush, work and finally stop words are has, of, are, and which has been removed as these stop words are unnecessary consuming time and space complexity of pre-processing.

4.1.3 MULTILEVEL FILTERING TECHNIQUES

The resultant tweets after applying NLP techniques is then subjected to multilevel filtering which has three levels as Inclusive filter, Relevance filter, and Opinion filter. The details of each filter with their results are as follows:

INCLUSIVE FILTER MODULE

The Level 1 Inclusive filter takes three inputs. They are tweets collected from twitter streams, product/service for which opinion mining needs to be done and a list of inclusive keywords related to the product/service. The Level 1 filter makes use of the keywords and service/product (e.g.: Indian Railways) and returns tweets that satisfy the filtering criterion. The resultant tweets are named as included tweets. Inclusive filter is the basic filter in multilevel filtering if our tweet is related to our domain called Indian railways finally its shows 1 if it is not present its shows 0. The snapshot of dataset after applying inclusive filter is shown in figure 4.6.

The above snapshot shows the only collection of Indian Railways domain tweets like in first tweet it is showing train number 15623 which reflect that it

belongs to Indian railways domain.

RELEVANCE FILTER MODULE

Relevance filtering module on the other hand actually starts filtering tweets based on the given features. We considered Indian Railways features to test the relevance filtering on Cleanliness, Staff behaviour, Punctuality, Security and Food quality. In its simplest form, the relevance filter removes the tweets that do not have the aforementioned features. However, in English or any language for that matter (in linguistics) a word or phrase expressed differently may convey similar meaning. For instance, the feature cleanliness has other words like hygiene and sanitation giving same meaning. Therefore, it is important to consider semantic meanings of phrases or words in relevance filter. Lexical resource such as WordNet is employed to achieve this. WordNet provides programmable interface and semantic meaning of words to improve quality of functioning of relevance filter.

Relevance filter as described above is used to produce meaningful domain specific tweets. However, it is improved further using feature selection process. Each feature may have many synonyms. By considering representative features, the time and space complexity can be reduced further. For instance, with respect to food, response time is one of the features that show quality of food service. This way for every feature there might be number of features that can be grouped into representative features. An algorithm named Representative Feature Selection (RFS) based on entropy and gain has been built for this purpose. More details of the algorithm are provided later in this section. With feature selection algorithm, the tweets subjected to relevance filter are pruned based on representative features satisfying filter criterion such as matching with given features. With RFS algorithm the quality of filtering is achieved. Further, in relevance filtering, there are possibilities to incorporate domain specific queries based on train number with or without specific time period. These two aspects, in this thesis, are called query and temporal information. These are elaborated later in this thesis.

We applied various types of filtering like relevance for features and opinion for sentiments and further clustered for better sentiment analysis on Indian railways application. Relevance filter is aimed at filtering tweets based on given attributes such as cleanliness, punctuality, staff behaviour, security and food quality. In simple terms, it removes tweets that do not have the specified attributes. To over-

come problems related to semantic meanings, this filter exploits lexical resource like WordNet. The inputs provided by the NLP processing module are given input to this filter. The input is denoted as $T = t_1, t_2, t_3, \dots, t_n$. Let the words in each tweet be denoted as $W = w_1, w_2, w_3, \dots, w_n$. Set of attributes based on which the filter needs to perform its filtering task is denoted as $A = a_1, a_2, a_3, \dots, a_n$. This filter also considers query and temporal information which is denoted by Q .

$$\begin{aligned} Q &= q_1, q_2, q_3, \dots, q_n \\ T &= t_1, t_2, t_3, \dots, t_n \\ W &= w_1, w_2, w_3, \dots, w_n \end{aligned} \tag{4.1}$$

$$A = a_1, a_2, a_3, \dots, a_n$$

$$(\forall(T, W) \epsilon t_n \times w_n.T_{t_n w_n} == q_n) \&\& (\forall(T, W) \epsilon t_n \times w_n.T_{t_n w_n} == \forall A_{a_n}) \tag{4.2}$$

$$\begin{cases} 1 & \text{if } T_{t_n w_n} \epsilon Q, T_{t_n w_n} = A_{a_n} \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

$(\forall(T, W) \epsilon t_n \times w_n.T_{t_n w_n})$ means first word in first tweet is verified with query which is selected by user. This process continues until all words in tweet are verified with query. With respect to temporal information and capturing, if year is not match with our tweet it just eliminates the tweet and move onto next tweet. If it matches with year it will move onto second comparison i.e. $(\forall(T, W) \epsilon t_n \times w_n.T_{t_n w_n} == \forall A_{a_n})$ first word in first tweet compares with attribute set. This process continues until words in tweets are compared with all attribute set. If it is found in middle of tweet it won't require checking remaining words in tweet.

$(\forall(T, W) \epsilon t_n \times w_n.T_{t_n w_n} == \forall A_{a_n})$ means comparing every word in tweet with attributes. Here 0 indicates that it is not relevant tweet whereas 1 indicates it is relevant tweet.

In relevance filter we want relevant features for our parameters that we can find using Feature selection that can improve the performance of the propose methodology. In this thesis, entropy and gain based representative feature selection is carried out. Entropy refers to disorder or uncertainty. Entropy and Gain are the two statistical measures used to make decision pertaining to removal of irrelevant features and choosing relevant features that are relevant to the chosen concept. These two are also used to know correlation between two attributes in given dataset. Entropy characterizes impurity of an attribute while the Gain

Table 4.1: Shows Positive and Negative Words Used for Sentiment Analysis

Notations	Description
$E(T)$	Entropy of one attribute
I	X Two different attributes
C	Total number of attributes
$P(c)$	Probability of attribute
$E(C)$	Entropy of attribute
P_i	Probability of particular attribute
$E(T,X)$	Entropy of two attributes

is the expected reduction in entropy when examples are partitioned according to given attribute. The formula to find the entropy and gain are given below whereas Table 4.1 gives the annotation used in the entropy and gain with their descriptions.

$$Entropy E(T) = \sum_{i=1}^c -P_i \log_2 P_i$$

$$E(T, X) = \sum_{c \in X} P(c) E(C)_{(\text{for two attribute})}$$

$$Gain(T, X) = E(T) - E(T, X)$$

As shown in the above algorithm entropy, gain and ranking of features are computed on the tweets and representative features are selected for reducing computational space and improve efficiency of the proposed methodology. The feature selection algorithm is aimed at selecting representative features from a set of features to reduce space and time complexity further. For instance, cleanliness is an attribute that may have associated features like timeliness, response time and it is evident that there are number of synonyms and by considering there is probability of improving performance of relevance filter in terms of accuracy. The filtering process is thus supported by lexical resource like WordNet, feature selection algorithm as discussed. Snapshot of resultant dataset after applying relevance filter is shown in figure 4.7.

Algorithm 2 Representative Feature Selection Algorithm

INPUT: X : Tweets dataset D

OUTPUT: X_{OPT} = Feature subset FS

Initialize: $gt, et, FS, \mathbf{F}, \mathbf{A}, \mathbf{c}, \mathbf{T}$

Extract attributes of \mathbf{D} into \mathbf{A} based on \mathbf{c}

Extract Relevant Features

repeat

$X' = \text{search_strategy}(a, A)$

$hX_{OPT} = \text{For attribute compute gain, entropy, } a = (\text{weight}, a)$

if $e > et$ and $g > gt$ **then**

$X_{OPT} = a$ ($a = a + F$)

end if

until stop criteria is found

Construction of Tree

repeat

$X' = \text{search_strategy}(f, F)$

$X_{OPT} = f + F$

until stop criteria is found

Find Representative Features

repeat

$X' = \text{search_strategy}(\text{node}, T)$

$X_{OPT} = \text{find nodes in } T$

if feature pair is correlated **then**

$X_{OPT} = FS \leftarrow f$

end if

until stop criteria is found

The above snapshot shows that the tweets coming after relevance filter is relevant to our defined features such as cleanliness, security, punctuality and relevance of related features of tweets with the help of WordNet Dictionary for further filtering.

```

In [21]: runfile('C:/Users/Admin-pc/Desktop/Transparency.py', wdir='C:/Users/Admin-pc/
Desktop')
2017 Train no. 15623 pathetic standards cleanliness. Toilets very dirty flush doesn't
work
2016 12791 Leveraging transparency, speed, safety, security Thank you
2017 Travelling 12257 say cleanest toilet train last 30 y
2017 Train:12257 LTT ERN Cls:3A Coach: G6 Seat:59 needs cleaning

In [22]:

```

Figure 4.7: Snapshot of dataset after having relevance filter

OPINION FILTER MODULE

This module is the final step in the filtering process. The datasets required by sentiment analysis need to have sentiments essentially. In simple terms the function of this filter is to remove tweets from the output of relevance filter that do not have either positive or negative sentiments. This filter also needs lexical resources such as Sentiwordnet to compute sentiment score of each tweet. The sentiment score is compared with a threshold value that is determining factor to filter tweets from the collection of tweets. The tweets that do not possess real sentiments or opinion of public are not useful for sentiment analysis. When such tweets are pruned, the processing pertaining to sentiment analysis will yield high quality results besides reducing time and space complexity. The result of opinion filter is collection of tweets (dataset generated to say appropriately) that are domain specific and ready to be used for sentiment analysis.

As shown in algorithm 3, it is evident that there are three modules for filtering. This filter takes input from relevance filter. Then it finds whether each tweet has any sort of sentiment. Sentiwordnet is used to compute sentiment score for each tweet. Sentiwordnet is a lexical resource that is useful for opinion mining. In this thesis, it is used to know whether a tweet has sentiment or not. The rationale behind this is that the thesis is aimed at making Tweet datasets for sentiment analysis and actual procedure for knowledge based sentiment analysis is deferred for our future work. Sentiwordnet has all information required in the form of Sentiwordnet _3.0.0_20130122.txt file. The snapshot of dataset after applying

opinion filter is shown in figure 8. The prototype application built in this thesis gets the information from the file and computes sentiment score for each tweet. Based on the information obtained from Sentiwordnet the application determines whether there is sentiment in the given tweet or not. Opinion filter is responsible to find whether each tweet has opinion or not.

Relevance filter's output is taken as input of this filter. The input is denoted as $T = t_1, t_2, t_3, \dots, t_n$. Let the words in each tweet be denoted as $W = w_1, w_2, w_3, \dots, w_n$. Set of opinion words or sentiment score vectors based on which the filter needs to perform its filtering task is denoted as $O = o_1, o_2, o_3, \dots, o_n, T = t_1, t_2, t_3, \dots, t_n, W = w_1, w_2, w_3, \dots, w_n$.

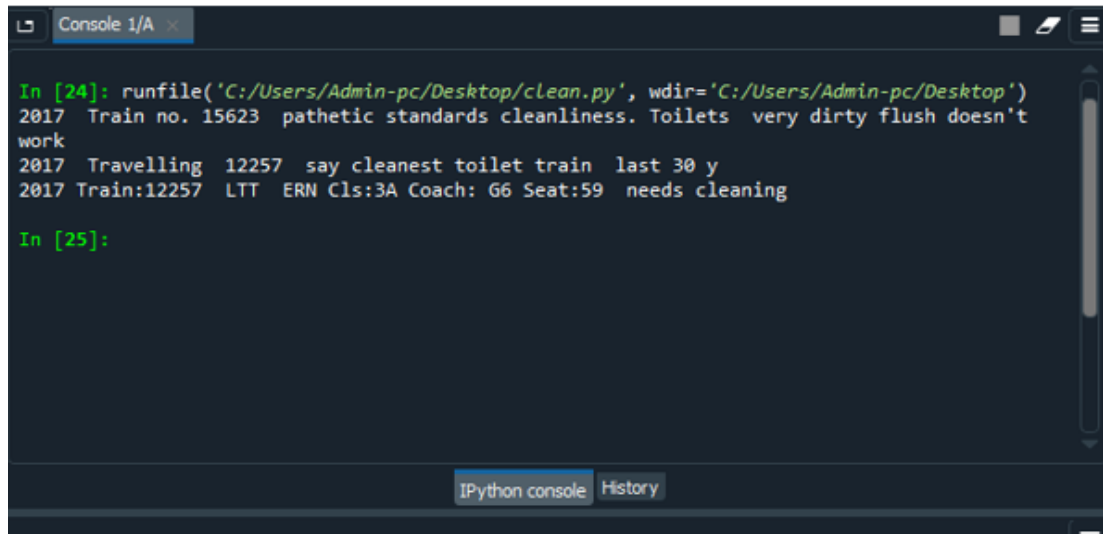
$$(\forall(T, W) \epsilon t_n \times w_n. T_{t_n w_n} == O_n) \& \& (\forall(T, W) \epsilon t_n \times w_n. T_{t_n w_n} == \forall O) \quad (4.4)$$

$$\begin{cases} 1 & \text{if } T_{t_n w_n} \in Q, T_{t_n w_n} = O_{a_n} \\ 0 & \text{otherwise} \end{cases} \quad (4.5)$$

$(\forall(T, W) \epsilon t_1 \times w_1. T_{t_1 w_1} == O_n$ first word in first tweet compares with Opinion set. This process continues until words in tweet compare with all opinion word set. If it is found in middle of tweet it won't require checking remaining words in tweet. $(\forall(T, W) \epsilon t_1 \times w_1. T_{t_1 w_1} == O_n$ means comparing every word in tweet with opinion set. In this filter we filtered opinion of tweets. Finally, we get required tweets. For example, we took one tweet and we checked every word in the tweet with opinion set, which one is match with opinion set, those were our required tweets. According to mathematical model 0 indicates failure and 1 indicates success. The figure 4.8 shows the snapshot of datasets after having opinion filter which contains opinions of the tweets.

The above snapshot reflects about the tweets related to opinion or not with the help Sentiwordnet such as in first tweet showing opinion of toilets that it is very dirty and flush does not work. This shows the negative opinion of the tweet. To implement multilevel filtering an algorithm has been designed called **Multi-Level Filtering (MF)** Algorithm. This algorithm is meant for achieving pre-processing of Twitter tweets by employing multi-level filtering to generate domain-specific tweets dataset for sentiment analysis. It has mechanisms for the three modules aforementioned.

The following reason we get the MF algorithm explained

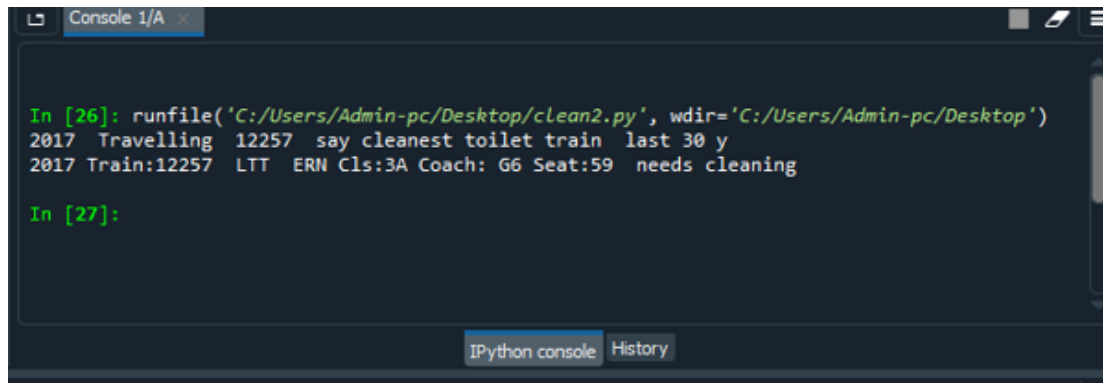


```

In [24]: runfile('C:/Users/Admin-pc/Desktop/clean.py', wdir='C:/Users/Admin-pc/Desktop')
2017 Train no. 15623 pathetic standards cleanliness. Toilets very dirty flush doesn't
work
2017 Travelling 12257 say cleanest toilet train last 30 y
2017 Train:12257 LTT ERN Cls:3A Coach: G6 Seat:59 needs cleaning

In [25]:
    
```

Figure 4.8: Snapshot of dataset after having opinion filter



```

In [26]: runfile('C:/Users/Admin-pc/Desktop/clean2.py', wdir='C:/Users/Admin-pc/Desktop')
2017 Travelling 12257 say cleanest toilet train last 30 y
2017 Train:12257 LTT ERN Cls:3A Coach: G6 Seat:59 needs cleaning

In [27]:
    
```

Figure 4.9: Snapshot of results after applying MF algorithm

1. Tweets of Indian Railways.
2. Tweets of Indian Railway with given Attributes.
3. Tweets of Indian Railway with given attributes containing opinion.
4. Tweets of Indian railways with query and temporal dimension. Snapshot of tweets of Indian railways after applying query and temporal dimension are shown in figure 4.9.

The above tweets show the query and temporal dimension such as in both the tweets above its temporal dimension i.e. year is 2017 and we are applying query on train number 12257 within all tweets to be filter.

4.1.3.1 TOPIC BASED FILTERING TECHNIQUES

Further the output dataset of multilevel filtering is again processed using topic based filtering which we have introduced. In multilevel filtering there are certain drawbacks such as some tweets are still having difficult to classify input to machine learning classification, there is an ambiguity and unclear to convert text data to numerical data, and also have some problems while selection of topics, collection of synonym, extraction of concept, detection of subjectivity, checking of polarity, and in ordinal classification. Therefore, to overcome these above drawbacks Topic based filtering module has been proposed. The proposed methodology is illustrated in figure 4.10. It takes the help of lexical dictionary and also polarity detection for efficient ordinal classification of tweets on three-scale basis. Tweets dataset is collected from Twitter accounts of Indian Railways. Sentiment analysis can be made with or without topics. For instance, it is possible to find whether a tweet expresses positive, negative or neutral sentiment without using any topic. On the other hand, topic-based sentiment analysis classifies sentiments conveyed towards the given topic. In this thesis, our final level of pre-processing is topic based filtering we chose it to be topic based and the topics used are punctuality, security, cleanliness, staff behaviour and food quality. This is important to increase the accuracy of the sentiment analysis. Polarity detection phase is crucial for sentiment analysis. In this phase, the tweets do have one of the given topics and sentiment is used as input to find the label of a tweet whether it is positive, negative, and neutral and topic based filtering techniques are implemented below with framework, mathematical model, results which is useful for next level of machine learning classification.

In the above framework the first module is topic selection which is used to make it domain-independent. Since the topics are to be understood by the system correctly, lexicon building is made with WORDNET synsets. Topic selection and lexicon building are done as follows.

$$T = t_1, t_2, \dots, t_n$$

$$W = w_1, w_2, \dots, w_n$$

T is set of tweets. W is set of words in tweet. $Ts_i = ts | \forall b_n b_n$ is set of synonyms (lexicon vector) for every Topic selection concatenated to relevant topic selection. Set of topic selection is denoted as in Eq. 4.6.

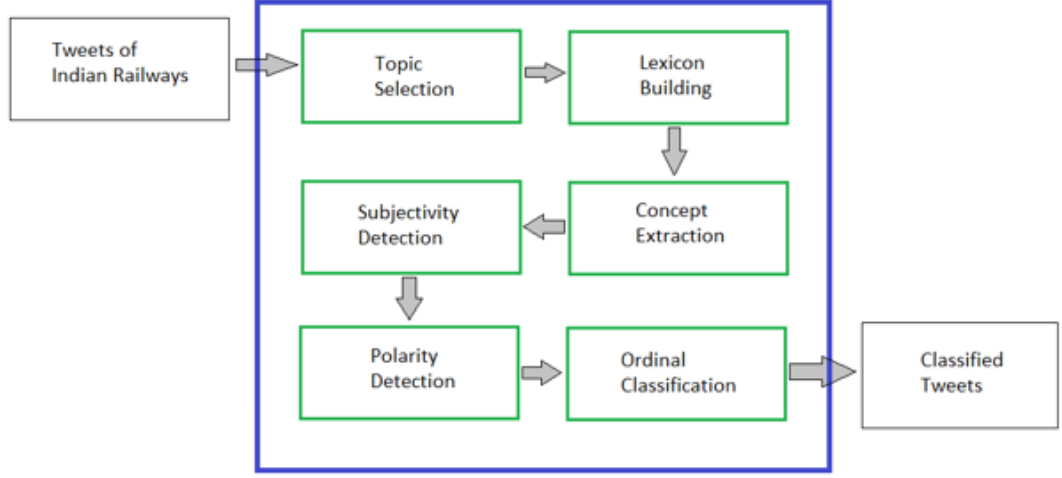


Figure 4.10: Proposed framework of Topic based filtering system

$$TS = U_i TS_i = \{TS_1, TS_2, \dots, TS_n\} \quad (4.6)$$

Where $TS_1 = b_1, b_2, \dots, b_m$ means set of synonyms of topic selection $TS_2 = b_1, b_2, \dots, b_m$ means set of synonyms of topic selection $TS_n = b_1, b_2, \dots, b_m$ means set of synonyms of topic selection.

The concept extraction phase is used to identify concepts based on the topics chosen. It is important to deconstruct given text or tweet into concept for better semantic-aware analysis. The concept extraction can make use of term wait and other approaches in case of document analysis. However, in this work, we considered the words or phrases associated with given topics as concepts. That way concept extraction is made straight forward and simple. This phase also helps in eliminating tweets that do not have one of the chosen topics. For this task to complete the lexicon constructed in the previous phase is utilized.

$$(\forall(T, W) \in t_n \times w_n. T_{t_n w_n} == TS) \quad (4.7)$$

As shown in Eq. 4.7, it indicates that every word of every tweet should compare with Topic selection and its synonyms. If it is matched with either topic selected or its synonyms, then it is not removed from tweet list. Otherwise it is removed. The next phase in the framework is subjectivity detection. It is an NLP task which removes neutral or factual tweets from the given tweets. The factual content is also known as object content that does not reflect any opinion. This

is important to increase the accuracy of the sentiment analysis. The removed tweets are saved for further evaluation procedures. It is modelled as follows.

$$SD = sd_1, sd_2, \dots, sd_n$$

SD is set of subject detection words (synsets vector).

$$(\forall(T, W) \epsilon t_n \times w_n. T_{t_n w_n} == SD) \quad (4.8)$$

Eq. 4.8 indicates that every word of every tweet should compare with Subject detection words. If it is matched with Subject detection word, then it is not removed from tweet list. Otherwise it is removed. Next is polarity detection phase which is very crucial for sentiment analysis. In this phase, the tweets do have one of the given topics and sentiment is used as input. It results in providing polarity values that are used for ordinal classification. Sentiwordnet is used to extract polarity values. The algorithm has been designed to implement topic based filtering called Topic-based Ordinal Classification of Tweets (TOCT). The details are as follows:

Topic-based Ordinal Classification of Tweets (TOCT) Algorithm

This algorithm is used to perform sentiment analysis. It takes tweets dataset and a set of topic as input and produce classification of them with 3-scale labels such as positive, negative, and neutral.

As shown in Algorithm 4, the procedure in the sentiment analysis covers all phases provided in the framework. An important phases include lexicon building for given topics, concept extraction for removing tweets that do not have topics associated, subjectivity detection to remove neutral tweets, polarity detection for finding polarity of the tweets in three-point scale and ordinal classification for final classification of tweets of Indian railways according to the mappings

Polarity value checking:

Polarity checking difference between system generated value with manual values of tweets of figure 8 is shown in below box which is useful in finding error rate for misclassification of tweets.

According to our conditions:

total score for -0.9383803551002399

1. 905372737836863488, 2017:RT Train no. 15623 has pathetic standards of cleanliness. Toilets are very dirty and flush doesn't work

polarity value for above tweet is -0.9383803551002399 which is greater than -1 . So it is negative.

total score for 0.15711861522415682

2. 894460386753208321, 2017RT Leveraging #IT4Rail for transparency, speed, safety, security in Thank you

polarity value for above tweet is 0.15711861522415682 which is less than 1 so it is positive.

total score for 0.6942945806653511

3. 897048502710042624,RT Travelling in 12257 & I must say this is d cleanest toilet I hve seen in train in last 30 y

Data write succefully.....

BUILD SUCCESSFUL (total time: 1 second)

Misclassification of the labels is a big mistake. Therefore, error rate is to be computed. Equation 4.9 is used to compute error rate.

$$Errorrate = \frac{1}{|Te|} \sum_{y_i} |(x_i) - y_i| \quad (4.9)$$

$y_i = \{-1, 0, 1\}$ means present polarity value set.

(x_i) =predicted polarity value set.

$|Te|$ = magnitude of difference of present and predicted values of y_i

Error rate is calculated by using difference of polarity value set and predicted value set and its magnitude value of polarity value and predicted value set.

Error rate:

$y_i = \{-1, 0, 1\}$

$h(x_i) = \{-0.9383803551002399, 0.15711861522415682, 0.6942945806653511, 0.25\}$

$Errorrate = \frac{1}{|Te|} \sum_{y_i} |(x_i) - y_i|$

$|Te|_{[21]} = \sqrt{((y1)^2 - (x1)^2 + (y2)^2 - (x2)^2 + (y3)^2 - (x3)^2 + (y4)^2 - (x4)^2}$

$$\begin{aligned}
 |T_e| &= \sqrt{((-1)^2 - (0.9383)^2 + (1)^2 - (0.15)^2 + (1)^2 - (0.694)^2 + (1)^2 - (0.25)^2)} \\
 &= \sqrt{0.1195 + 0.9775 + 0.5183 + 0.9383} \\
 &= 1.5977
 \end{aligned}$$

$$\begin{aligned}
 \text{For first tweet } Errorrate &= \frac{1}{|T_e|} \sum_{y_i} |(x_i) - y_i| \\
 &= \frac{1}{.5977} (1 + 0.9383)^2 \\
 &= \frac{1}{.5977} (0.0038) \\
 &= 0.0023
 \end{aligned}$$

$$\begin{aligned}
 \text{For second tweet } Errorrate &= \frac{1}{|T_e|} \sum_{y_i} |(x_i) - y_i| \\
 &= \frac{1}{.5977} (1 + 0.1571)^2 \\
 &= \frac{1}{.5977} (0.0.7104) \\
 &= 0.444
 \end{aligned}$$

$$\begin{aligned}
 \text{For third tweet } Errorrate &= \frac{1}{|T_e|} \sum_{y_i} |(x_i) - y_i| \\
 &= \frac{1}{.5977} (1 + 0.6942)^2 \\
 &= \frac{1}{.5977} (0.0935) \\
 &= 0.058
 \end{aligned}$$

$$\begin{aligned}
 \text{For fourth tweet } Errorrate &= \frac{1}{|T_e|} \sum_{y_i} |(x_i) - y_i| \\
 &= \frac{1}{.5977} (1 + 0.25)^2
 \end{aligned}$$

$$= \frac{1}{.5977}(0.5625)$$

$$= 0.352$$

If the error rate is below 0.1[21] then it shows low error rate which identifies that it comes under negative and if it is above 0.1 then it is high which says that it comes under positive. Therefore, the first and third tweets have low error rate and the second and fourth tweets are at high error rate.

4.2 SUMMARY

This chapter introduced the modules used to pre-processing the Indian Railways tweets for sentiment analysis. The modules used for pre-processing of the data set are NLP techniques, Multilevel filtering, and Topic based filtering. Firstly, we get tweets form web scraping using beautiful soup library. Afterwards, NLP techniques such as Tokenization, Stemming, Lemmatization and Stop words removal has been applied on the resultant tweets. Tokenization helps to find the tokens whereas Stemming helps to perform root stem word, Lemmatization helps to find lemma from dictionary and finally, stop words removal removes unnecessary text in dataset. But NLP techniques has some drawbacks as it cannot filter according to inclusive words, relevant features, and opinion sentiments. To overcome this drawback, we proposed multilevel filtering techniques. This technique contains inclusive filter that define the present tweet is in domain are not, whereas relevance filter helps to filter according to the given attributes and features which are available in tweets are not, and finally opinion filter helps to find sentiment within the tweets. But after performing multilevel filtering we came to know that there are still some drawbacks left such as selection of topics, collection of synonym, extraction of concept, detection of subjectivity, checking of polarity, ordinal classification. Finally, to overcome these above drawbacks Topic based filtering module has been proposed. This module contains six submodules which are Topic selection, Lexicon Building, Concept Extraction, Subjectivity Detection, Polarity Detection, and Ordinal Classification. Topic selection means topics are to be understood by the system correctly, every word of every tweet should compare with topic selection and its synonyms. If it is matched with either topic

selected or its synonyms, then it is not removed from the tweet list. Otherwise, it is removed. Lexicon is the list of stems and affixes, together with basic information about them, lexicon building is WordNet using Synsets. Lexicon building is developed by WordNet dictionary which supports Sentiwordnet for sentiment analysis. The concept extraction phase is used to identify concepts based on the topics chosen. It is important to deconstruct given text or tweet into the concept for better semantic-aware analysis with WordNet synsets. The next phase is subjectivity detection. It is an NLP task that removes neutral or factual tweets from the given tweets. The factual content is also known as object content that does not reflect any opinion. This is important to increase the accuracy of the sentiment analysis. Polarity detection phase is crucial for sentiment analysis. In this phase, the tweets do have one of the given topics and sentiment is used as input to find the label of a tweet whether it is positive, negative, and neutral. Ordinal classification is also known as ordinal regression which finds error rate between tweets. The low error rate shows the tweets are negative and high error rate shows the tweet are positive. As misclassification is costly measures such as the overall sentiment of the tweet. Finally, at last we get the well refined resultant tweets that can be used as input as feature selection vector which is given to machine learning classification. Thus in the next chapter we have presented the detail description with results about different classification algorithm used to classify the dataset that has been pre-processed.

Algorithm 3 Multi-Level Filtering (MF) Algorithm

INPUT: Twitter Tweets Dataset \mathbf{D} , attributes \mathbf{A} , query information and capturing \mathbf{q} , temporal information and capturing t_q , Representative Features of Attributes \mathbf{F} (result of RFS algorithm)

OUTPUT: Tweets that are highly relevant with given attributes and having opinion

NLP Technique

Initialize tweets output vector D'

Apply SWR algorithm, Porter Stemmer algorithm and lemmatization on D

Relevance Filtering

$D = D'$

$D' = \text{empty}$

For each tweet t in D

For each attribute a in A and F

Use WordNet to get synonyms for a and t

if $a \in t$ and t satisfies q and t_q (including synonyms) **then**

Assign t to D'

else

Discard t from D

end if

Opinion Filtering

$D = D'$

$D' = \text{empty}$

For each tweet t in D

Compute sentiment score ss of t from S

if t has sentiment based on s_s **then**

Assign t to D'

else

Discard t from D

end if

Output D'

Algorithm 4 Topic-based Ordinal Classification of Tweets (TOCT) Algorithm

INPUT: Tweets dataset TD , topics T

OUTPUT: Classified tweets in three-point scale

```
Initialize lexicon vector  $V$ 
Initialize synsets vector  $s$ 
Initialize vector for factual tweets  $N$ 
 $T = \text{load Topics } ()$ 
for each topic  $t$  in  $T$  do
     $s = \text{getSynsets } ()$ 
    Add  $s$  to  $V$ 
end for
for each tweet  $td$  in  $TD$  do
     $s = \text{getSynset}(V)$ 
    if  $s$  not found in  $td$  then
        Remove  $td$ 
    end if
end for
for each tweet  $td$  in  $TD$  do
     $p = \text{get Polarity } ()$ 
    if  $p = 0$  then
        Remove  $td$  and add to  $N$ 
    end if
end for
for each tweet  $td$  in  $TD$  do
     $p = \text{get Polarity } ()$ 
    if  $p = -1$  then
        Classify  $td$  as Negative
    else if  $p = 0$  then
        Classify  $td$  as Neutral
    else if  $p = 1$  then
        Classify  $td$  as Positive
    end if
end for
```

Chapter 5

OPINION CLASSIFICATION OF PREPROCESSING DATASET FOR SENTIMENT ANALYSIS USING SINGLE AND ENSEMBLE CLASSIFIER

SECTION 5.1 presents the proposed methodology for effective sentiment analysis of tweets of Indian Railways (IR). SECTION 5.2 presents the description of single classifier approach of machine learning. SECTION 5.3 presents the description of single classifier approach of deep learning. SECTION 5.4 presents the description of ensemble classifier approach with stacking. SECTION 5.5 presents the results of classifiers. Finally, SECTION 5.6 summarizes this chapter.

5.1 PROPOSED METHODOLOGY FOR EFFECTIVE SENTIMENT ANALYSIS OF TWEETS OF IR

Sentiment analysis based on tweets is well known research area. However, efficient and novel pre-processing of tweets prior to sentiment analysis makes it different from state of the art. In the preceding chapter, pre-processing of tweets

associated with Indian Railways is made with the proposed methodology. The pre-processing includes multi-level filtering of tweets followed by topic based filtering. The resultant tweets contain sentiment besides attributes such as staff behaviour, cleanliness, food quality, security and punctuality. Lexicon is used to exploit semantic meaning of the attributes to improve quality of filtering. Based on the results of pre-processing of tweets, this chapter has mechanisms for classification of tweets with traditional machine learning approaches and also deep learning. Different supervised learning techniques such as Naïve Bayes, Support Vector Machine (SVM) and Long Short Term Memory (LSTM) besides an ensemble learning approach using stacking. The usage of deep learning has made the classification with higher level of accuracy. Further enhancement in producing quality sentiments is made using ensemble learning comprising of machine learning and deep learning approaches. More details of the research carried out for efficient classification of sentiments associated with Indian Railways are provided in subsequent sections. The results of empirical study are provided to show proof of the concept.

The sentiment analysis methodology for tweets of Indian Railways is designed to be more effective, novel and efficiency. Its novelty lies in ensemble classification besides its predecessor step described in preceding chapter. As there are number of classification algorithms available, some of them are carefully chosen for leveraging classification performance. In addition to the usage of machine learning (ML) techniques like SVM and Naïve Bayes, deep learning based method known as LSTM is also used. Afterwards, the ensemble approach with stacking method is used to improve classification accuracy. Since the classification algorithms have different capabilities, they are exploited by the ensemble method which is given significance due to its ability to arrive at higher level of accuracy. Though the classifier such as SVM and NB are proved to be good at classification, there is ever possibility that these algorithms may produce better results. The rationale behind this is that the quality of training has its influence on the accuracy of classification. Therefore, the main focus of this thesis is to have novel approach for efficient pre-processing which is discussed in Chapter 4. Unlike traditional ML techniques, deep learning methods like LSTM exhibit more depth in training process. Thus, in this research, deep learning as well as traditional ML techniques are used to improve performance. In either case (ML and deep learning),

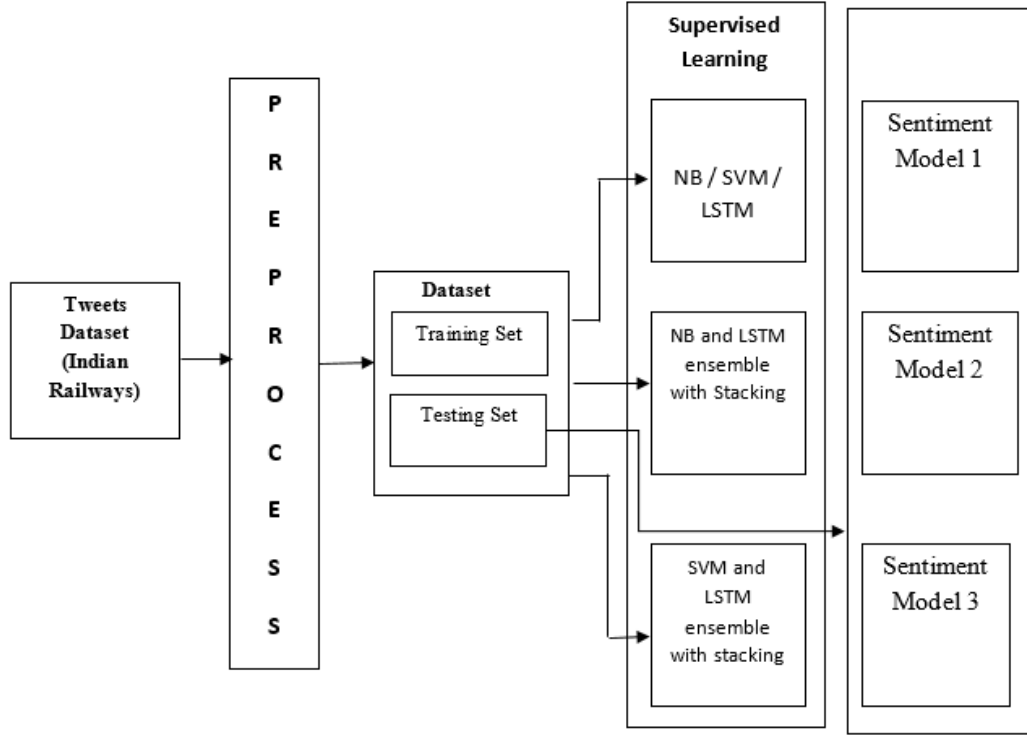


Figure 5.1: Proposed methodology for effective sentiment analysis of tweets of IR

pre-processing is preceded by classification. For this reason, it is expected that, the methodology for sentiment analysis shown in Figure 5.1 is more efficient provided the novel pre-processing method that includes multi-level filtering and topic based filtering (discussed in Chapter 4).

Both single and ensemble classifiers for sentiment analysis are used. The ML techniques used are NB and SVM and deep learning based classification method used in the empirical study is LSTM. Then the LSTM is ensemble with NB and SVM separately to leverage classification performance. These classifiers are employed after pre-processing of textual data tweets with feature extraction. Naive Bayes uses text data for labelling, training and testing purpose but for both Support Vector Machine, and Long Short Term Memory we used feature selection for better performance.

As presented in Figure 5.1, it is to be understood that the input datasets are pertaining to Indian Railways. Moreover, they are not raw in nature. They are the result of pre-processing with novel filtering techniques as presented in Chapter 4. After pre-processing, the data divided into training and testing sets. The training set is used to have quality training to the supervised learning methods

while the testing set is used to evaluate the classification methods in determining correct class labels. Based on the ML or deep learning method, classifier is built. After training process is completed a model is built that has required knowledge to classify tweets based on sentiments. In other words, the classifiers use three class labels such as positive, negative and neutral. The learning process involves usage of individual classifiers such as NB, SVM and LSTM. Afterwards, the deep learning method LSTM is ensemble with NB and SVM separately with using stacking phenomenon. The following subsections provide more details of the underlying learning mechanism.

5.2 SINGLE CLASSIFIER APPROACH OF MACHINE LEARNING

In this section we have used two classifiers called Naïve Bayes (NB) and Support Vector Machine (SVM) which comes under supervised learning method of machine learning. Supervised learning is the machine learning task of learning a function that maps an input to an output based on model of input and output pairs. A supervised learning algorithm analyses the training data and produces an inferred function, which can be used for mapping new examples. The details description of Naïve Bayes and Support Vector Machine applied on the pre-processed dataset of Indian Railways tweets are as follows:

5.2.1 Naïve Bayes

Naïve Bayes is one of the widely used classifiers. It is a probabilistic machine learning model that's used for classification task. The crux of the classifier is that it is based on the Bayes theorem. It is used in many real world applications for classification. For instance, it is used in heart disease prediction, spam filtering, classifying cancer diseases, segmenting documents and predicting sentiments in online reviews. The Naïve Bayes classification technique is based on the Bayes theory. The features it uses are independent and hence the name naïve. It does mean that when a value of a feature is changed, it does not affect other feature directly. This algorithm is found faster as it is probabilistic. It is also scalable in nature and suitable for applications where scalability is in demand. It has

its associated concepts like Bayes Rule and conditional probability. The Bayes theorem is as in Eq. 5.1.

$$p(A|B) = p(B|A) \frac{P(A)}{P(B)} \quad (5.1)$$

Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the features are independent. That is presence of one particular feature does not affect the other. Considering playing golf problem, for example, we classify whether the day is suitable for playing golf, given the features of the day. The columns represent these features and the rows represent individual entries. If we take the first row of the dataset, we can observe that is not suitable for playing golf if the outlook is rainy, temperature is hot, humidity is high and it is not windy. We make two assumptions here, one as stated above we consider that these predictors are independent. That is, if the temperature is hot, it does not necessarily mean that the humidity is high. Another assumption made here is that all the predictors have an equal effect on the outcome. That is, the day being windy does not have more importance in deciding to play golf or not. According to this example, Bayes theorem can be rewritten as in Eq. 5.2.

$$p(y|X) = p(X|y) \frac{P(y)}{P(X)} \quad (5.2)$$

The variable y is the class variable (play golf), which represents if it is suitable to play golf or not given the conditions. Variable X represent the parameters/features. It is computed as in Eq. 5.3.

$$X = (X_1, X_1, X_2, X_3, ..., X_N) \quad (5.3)$$

present the features, i.e. they can be mapped to outlook, temperature, humidity and windy. By substituting for X and expanding using the chain rule we get the proposition as in Eq. 5.4.

$$P(y|X_1, ..., X_n) = \frac{(P(x_1|y)P(x_2|y)...P(x_n|y)P(y))}{(P(x_1)P(x_2)...P(x_n))} \quad (5.4)$$

Now, you can obtain the values for each by looking at the dataset and substitute them into the equation. For all entries in the dataset, the denominator does not change, it remains static. Therefore, the denominator can be removed and a proportionality can be introduced as in Eq. 5.5.

$$P(y|X_1, \dots, X_n) \propto P(y) \prod_{i=1}^n P(X_i|y) \quad (5.5)$$

In this case, the class variable(y) has only two outcomes, yes or no. There could be cases where the classification could be multivariate. Therefore, we need to find the class y with maximum probability as in Eq. 5.6.

$$y = \operatorname{argmax}_y P(y) \prod_{i=1}^n P(X_i|y) \quad (5.6)$$

Using the above function, we can obtain the class, given the predictors.

5.2.2 Support Vector Machine

Support Vector Machine (SVM) is one of the widely used machine learning techniques used for prediction purposes like heart disease prediction. It can be used in many real world applications like hand-written character recognition, cancer prediction, protein classification, image classification and text categorization to mention few. It is a discriminative classifier which predicts class labels based on training knowledge. It achieves the classification formally with a definition of hyperplane. Thus it can provide largest minimum distance known as maximum margin associated with training samples. With this SVM can minimize the margin of the data used for training.

The optimal hyperplane with maximum margin is able to separate or discriminate two classes. In many research papers, SVM is found to have superior performance over its counterparts as found in the literature. It is used as part of the proposed methodology in this research. It makes use of the hyperplane illustrated in Figure 5.2. SVM is basically a binary classifier which contains two class labels. In order to have multiple class labels, it is used with appropriate kernel parameter.

5.3 SINGLE CLASSIFIER APPROACH OF DEEP LEARNING

LSTM is used in this research as deep learning based classifier. LSTM helps in semantic representation of textual data. Effective LSTM models human thought

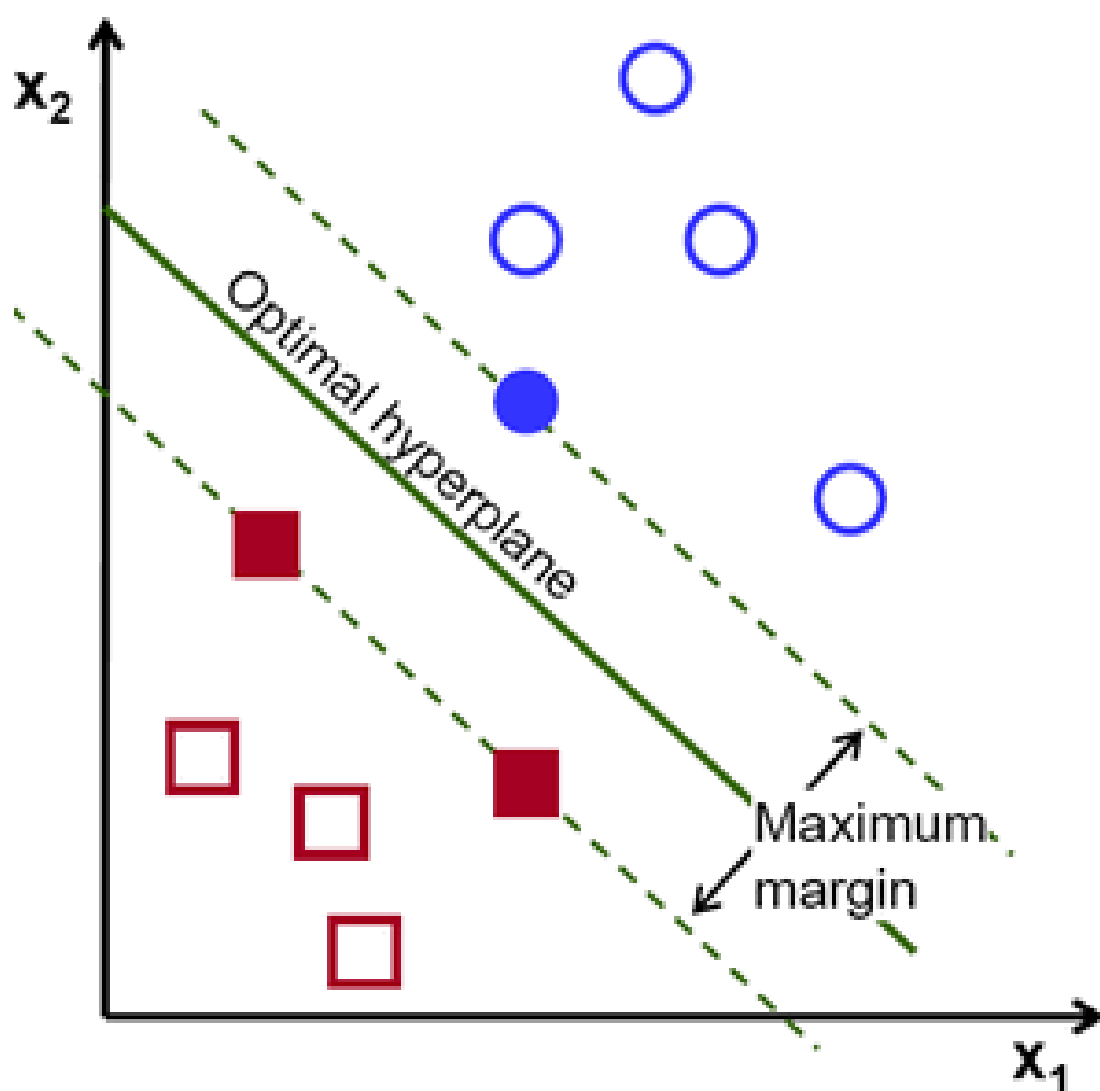


Figure 5.2: Illustrates how hyperplane is formed for discrimination

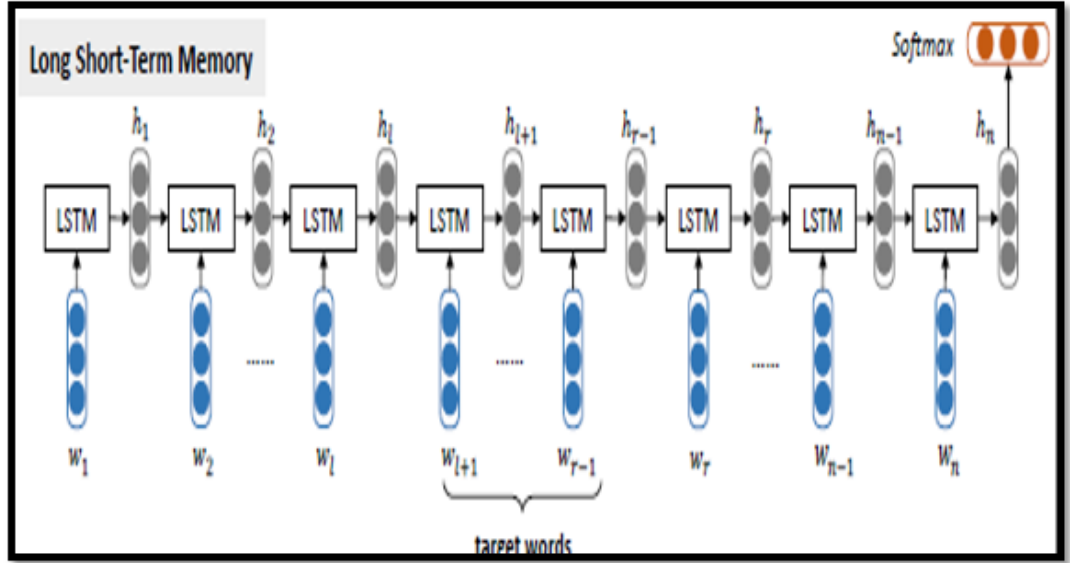


Figure 5.3: Illustrates structure of LSTM [2]

process as closely as possible. When human reads any data, his understanding of the data depends on the previous learning. Traditional neural networks do not have this capability of doing this. This drawback is overcome with deep learning based method known as LSTM. LSTM is a special kind of Recurring Neural Network (RNN) that can remember long term dependencies. In other words, LSTM overcomes the problem of long term dependency. RNN is good for remembering information for long time by default.

As presented in Figure 5.3, LSTM network has a form containing a chain of modules of neural network. There is simple structure associated with each repeating module. Each cell in the LSTM network has its state which changes as needed. First, it is determined what information needs to be thrown from the cell state. Afterwards, the information to be added to the cell state is determined. The structure has different gates such as input gate, output gate and forget gate. There is relation between the time and the cells remembering values. LSTM units are trained with pre-processed data (tweets of IR). This approach is known as supervised learning. The training sequences are used for prediction of results.

LSTM mode is built using sequential model Embedding Layer, LSTM Layer, Dropout Layer, and Dense layer. The Embedding layer is defined as the first hidden layer of a network. It must specify 3 arguments input_dim, output_dim, and input_length. With respect to LSTM layer, it is essential to provide the number of nodes in the hidden layers with the LSTM cell. In the empirical study

50 units are used in the layer. However, it can be fine-tuned any time. Dropout is a regularization technique, which aims to reduce the complexity of the model to prevent overfitting by randomly turns-off the activations of some neurons in the LSTM layer. A dense layer is a classic fully connected neural network layer where each input node is connected to each output node. Once the model is built, it is subjected to prediction process that comes up with results of sentiment analysis.

5.4 ENSEMBLE CLASSIFIER APPROACH USING STACKING

Individual classification mechanisms explored in the Section 5.3 have their capabilities. However, exploiting capabilities of multiple supervised learning methods yields better performance in sentiment classification. Based on this proposition, one of the individual classifiers such as NB and SVM is combined with LSTM with stacking approach to realize ensemble of the methods.

As presented in Figure 5.4, the ensemble model is made up of multiple classification algorithms. In this research, the deep learning model LSTM is combined with ML model such as NB or SVM. These models are part of Artificial Intelligence (AI). The input data is divided into many parts. Each slice of data is subjected to classification process using different algorithm. Two ensemble cases are used for the empirical study. In the first case, $LSTM + NB$ is used while in the second case $LSTM + SVM$ is used. The two prediction models are finally combined into a single model. This process is carried out as per the stacking algorithm used for ensemble approach. By combining predictions of different classifiers, stacking produces prediction results that are better than the results of individual classifier.

Ensemble models have been used extensively in credit applied based scoring applications resources and other areas because they are considered to be more stable with more space and execution time but have more other advantages compared to it. And, more importantly, predict better than single classifiers. They are also known to reduce model complexity. The objective of the ensemble method is to compare predictive accuracy using two ensemble classifiers and compare results with three single classifiers to evaluate the proposition that says “ensemble method performs better than individual prediction model”.

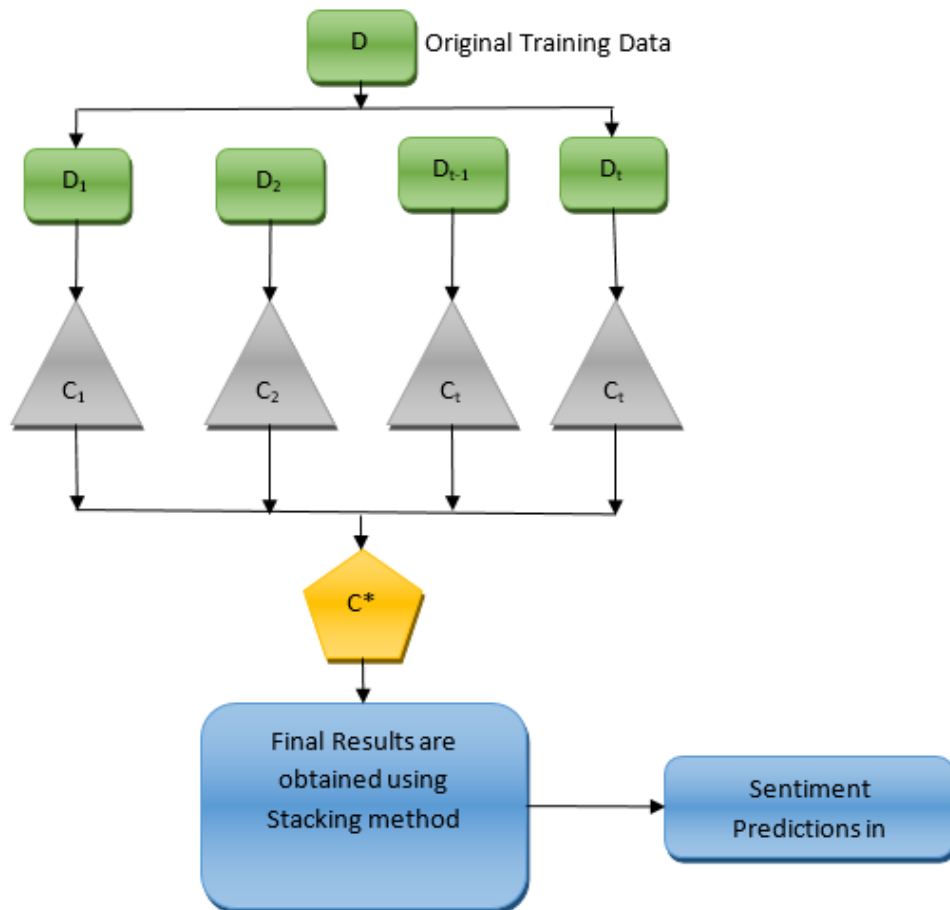


Figure 5.4: Ensemble mechanism [1]

5.4.1 Deep Learning based Stacking Ensemble

Two ensemble experiments are made in the empirical study. The first one is to stack Naïve Bayes with Long Short Term Memory while the second one is to stack Support Vector Machine with Long Short Term Memory. The reason behind this is that Long Short Term Memory is a deep learning method that leverages training quality with depth in learning.

Algorithm 5 Deep Learning based Stacking Ensemble

INPUT: Training data set D

Initialize training data set matrix TD

Initialize testing data set matrix $TD1$

algorithm Stack 1 = [Naïve Bayes, Long Short Term Memory]

algorithm Stack 2 = [Support Vector Machine, Long Short Term Memory]

for each i , algorithm in algorithm Stack **do**

for trainix, testix in split(D , $k=10$) **do**

$TD = \text{algorithm.Fit}(\text{train}[\text{trainix}], \text{target}[\text{trainix}]).\text{predict}(\text{train}[\text{testix}])$

end for

$TD1[i] = \text{algorithm.Fit}(\text{train}).\text{predict}(\text{test})$

 Final Predictions = Stacked Algorithm. Fit(TD , target).predict($TD1$)

end for

Return final Predictions

As presented in Algorithm 1, the algorithm takes dataset which is then split into training and testing set. Step 1 is used to initialize training data denoted as TD . The Step 2 initializes test data denoted as $TD1$. In Step 3 and Step 4, two algorithm stacks are created, one with LSTM and NB and other with LSTM and SVM. In Step 5, an iterative process is started. For each ensemble algorithm, another iterative process starts on Step 6 through Step 8 for training the ensemble model. In Step 9 and 10 there are mechanisms to get final predictions with stacking approach. Step 12 returns the final predictions made. Stacking is used as that often considers heterogeneous weak learners, learns them in parallel and combines them by training a meta model to output a prediction based on the different weak models predictions. Very roughly, we can say that bagging will mainly focus at getting an ensemble model with less variance than its components whereas boosting and stacking will mainly try to produce strong models less

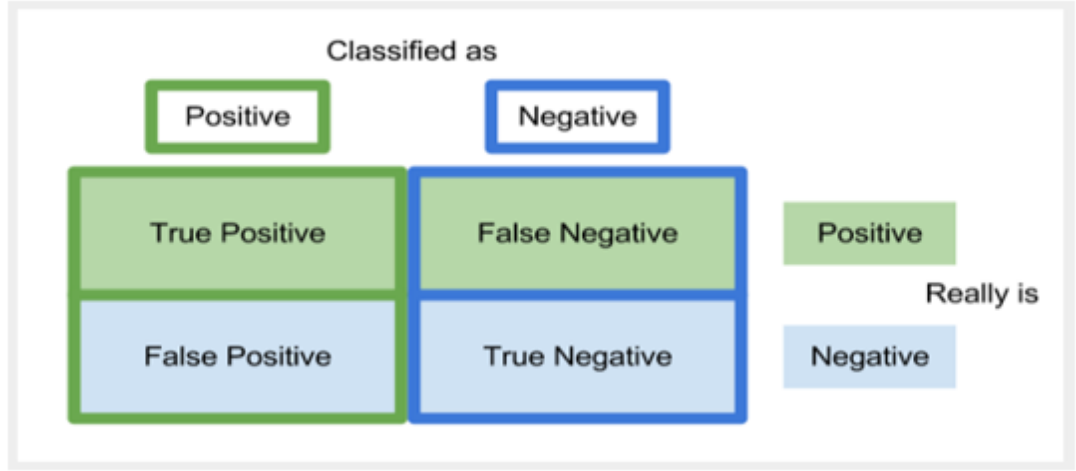


Figure 5.5: Shows confusion matrix

biased than their components even if variance can also be reduced.

5.4.2 Evaluation Procedure

The proposed algorithms are evaluated using a standard approach based on confusion matrix. Confusion matrix helps in deriving multiple metrics in machine learning approaches. They are used in process mining for prediction problems as well.

As shown in Figure 5.5, confusion matrix provides different cases like TP, FP, FN and TN. These are used to derive performance metrics like precision, recall and F1 score as in Eq. (5.7), Eq. (5.8) and Eq. (5.9) respectively. The predicted positives are made of actual positives then it is further considered as True Positive (TP). Second case is that predicted positives made of actually but negatives. Such results are known as error as False Positives (FP). The third case is predicted negative but other label actually positive. It is known as error shown False Negative (FN) in sentiment analysis if it is right result as positive but given result as other classification as negative or neutral. The fourth case is that predicted negative but always actually negative.

$$Precision = \frac{TP}{TP + FP} \quad (5.7)$$

$$Recall = \frac{TP}{TP + FN} \quad (5.8)$$

$$F_1 = 2 \times \frac{\text{precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5.9)$$

The measures like precision, recall and F1-score are widely used for evaluation of machine learning algorithms. All these measures are based on the TP, FP, TN and FN. The precision indicates the ratio of number of correctly classified tweets to the total number of total tweets. High precision relates to the low false positive rate. Similarly, recall is the ratio that indicates the number of correctly classified tweets to the total number of correct matches. F measure indicates the ratio of multiplication of Precision and Recall to Sum of Precision and Recall. F measure is used to test accuracy. Precision, Recall, F measure of state-of-art techniques are compared to the proposed approach and it is observed that both our single as well as ensemble classifier using stacking shows good results for sentiment analysis classification for our Indian Railways Application of tweets.

5.5 EXPERIMENTAL SETUP

Experiments are made with the prototype built using Python data science platform. Tweets from Indian Railways are used for empirical study. The ratio for training and testing is 80 : 20. It does mean that 80% of data is used for training while 20% is used for testing. The rationale behind this is that the more in training data leads better quality in training.

Train no. 15623 has pathetic standards of cleanliness. Toilets are very dirty and flush doesn't work also 06060 train number -1
 12791 Leveraging for transparency, speed, safety, security in Thank you other train 13407 1
 Travelling in 12257 & amp; I must say d cleanest toilet seen train last 30 y other 12403 0
 "Trn:12201 Dt:11-08-2017 Frm LTT to ERN Cls:3A Coach: G6 Seat:59 It needs cleaning other train 18234 -1
 12164 securities is very bad -1
 12403 food quality is not tasty -1
 12486 unpunctual -1
 05036 train number staff behavior is corruption train number
 07092 cleanliness is unhygienic -1
 Train 06042 transparency, speed, safety 1

Listing 1: An excerpt from training set

As shown in the training data, there is class label when observed carefully against every tweet. The class label is 0 if the tweet has neutral sentiment. The class label is 1 for positive and -1 for negative sentiment.

Train no. 15623 has pathetic standards of cleanliness. Toilets are very dirty and flush doesn't work also 06060 train number
12791 Leveraging for transparency, speed, safety, security in Thank you other train 13407
Travelling in 12257 & I must say d cleanest toilet seen train last 30 y other 12403
"Trn:12201 Dt:11-08-2017 Frm LTT to ERN Cls:3A Coach: G6 Seat:59 It needs cleaning other train 18234
12164 securities is very bad
12403 food quality is not tasty
12486 unpunctual
05036 train number staff behavior is corruption train number
07092 cleanliness is unhygienic
Train 06042 transparency, speed, safety

Listing 2: An excerpt from training set

As shown in the testing data, there is no class label. The class label is predicted by the sentiment prediction models. The ML models such as NB and SVM are used as individual prediction models. LSTM is the deep learning based prediction model used for experiments. Afterwards, the NB + LSTM and SVM + LSTM are used as stacking ensemble sentiment prediction models.

In the above Figure 5.6 results of feature vector of 6 Attributes first 0,1 indicates domain are not Indian railways, next is the train number and third tweet identification number It will be pre-processed further, again carries features number (1 punctuality, 2 cleanliness, 3 food quality, 4 staff behaviour, 5 security) with sentiment values which helps for training, testing, labelling for predicted values of sentiment values. Finally, we have tag person of tweets.

RESULTS AND DISCUSSION

Experimental results of the aforementioned prediction models for sentiment analysis are presented in this section. The three measures such as precision, recall and F1-score are used for evaluation. The results are provided for observations in two sets of experiments. In the first set of experiments, individual prediction models are used along with the pre-processing mechanism presented in Chapter 4. Afterwards, the models are compared with the state of the art as well.

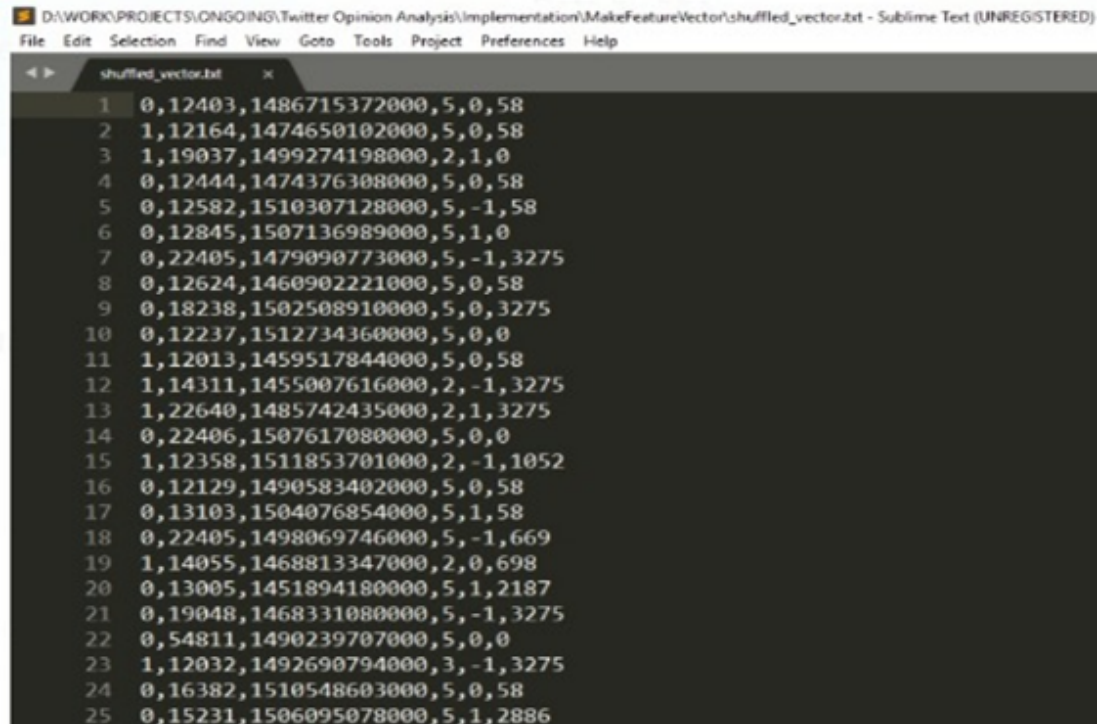


Figure 5.6: Feature vector with sentiment values helpful for labelling and predicted values

5.5.1 Results of Individual Prediction Models

The prediction models made up of NB, SVM and LSTM are used for the empirical study in the first set of experiments. Python data science platform is used for making experiments. Tweets data set pertaining to IR that has been subjected to pre-processing such as multi-level filtering, as explored in Chapter 4, is used for the experiments.

As presented in Table 5.1, observations on the sentiment analysis performance in terms of precision, recall and F1-score for three individual prediction models

Table 5.1: Performance comparison with precision, recall and F1-Measure

Prediction Models	Precision	Recall	F1- Measure
Naïve Bayes	83.3	78	80.5
Support Vector Machine	88.5	83	85.6
Long Short Term Memory	91.5	85	88.1

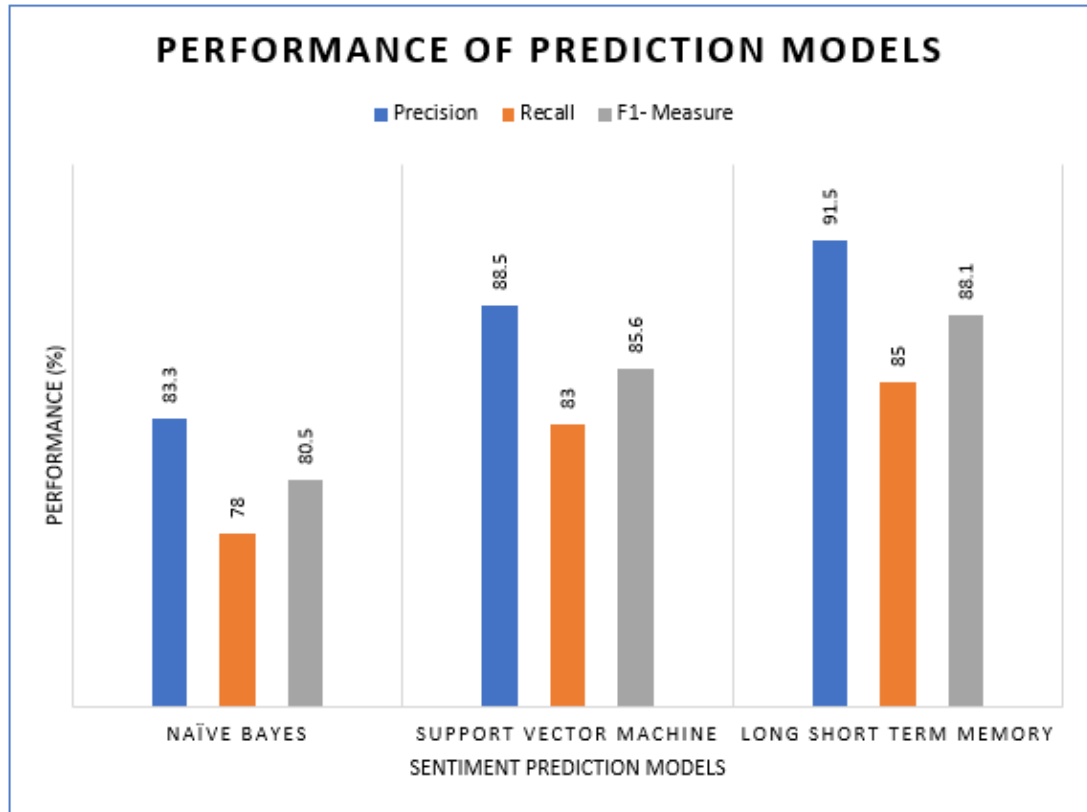


Figure 5.7: Performance of individual sentiment prediction models

mentioned are provided.

As presented in Figure 5.7, the horizontal axis shows the prediction models like NB, SVM and LSTM. The vertical axis shows the performance in terms of precision, recall and F1-Measure. The results revealed that each prediction model has shown different performance. Between the two ML algorithms NB and SVM, SVM is proved to be better in prediction of sentiments with 88.5% precision, 83% recall and 85.6% F1-Measure. SVM’s performance is better than that of NB. However, both the ML algorithms showed poor performance when compared with deep learning based LSTM. LSTM showed highest performance with 91.5% precision, 85% recall and 88.1% F1-Measure. From the empirical study, therefore, it is concluded that deep learning model has better performance over its ML counterparts.

5.5.2 Results of Stacking Ensemble Models

The hypothesis “ensemble models provide better performance over individual prediction models” is evaluated and the results are presented in this subsection.

Table 5.2: Results of ensemble methods

Prediction Models	Precision	Recall	F1-Measure
Naïve Bayes and Long Short Term Memory (stacking)	93.5	87	90.1
Support Vector Machine and Long Short Term Memory (stacking)	95.5	89	92.1

Stacking is the model used to ensemble the individual models such as NB and SVM.

As presented in Table 5.2, observations on the sentiment analysis performance in terms of precision, recall and F1-score for three individual prediction models aforementioned are provided. The stacking ensemble models such as *NB+LSTM* and *SVM + LSTM* are used.

As presented in Figure 5.8, the horizontal axis shows the stacking ensemble prediction models like NB + LSTM and SVM + LSTM. The vertical axis shows the performance in terms of precision, recall and F1-Measure. The results revealed that each prediction model has shown different performance. Out of the two ensemble models, the SVM+LSTM model showed better performance with 95.5% precision, 89% recall and 92.1% F1-score. Therefore, it is understood that SVM along with LSTM stacking ensemble can be used for performance improvement for sentiment analysis of tweets of IR. Naïve Bayes exhibited 80.5% F1-score, but its ensemble with LSTM has increased the same to 90.1%. In the same fashion, SVM as independent prediction model showed 85.6% F1-score. However, its stacking ensemble with LSTM showed 92.1% F1-score. Thus, the results revealed that the ensemble prediction models showed significantly better performance over individual prediction models.

5.5.3 Train wise Sentiment Analysis for Indian Railways

The previous subsections provided performance details of individual and ensemble prediction models. However, this section throws light on the actual sentiment analysis of Indian Railways in terms of punctuality, food quality, cleanliness, staff behaviour and security for different trains. The trains are identified with a

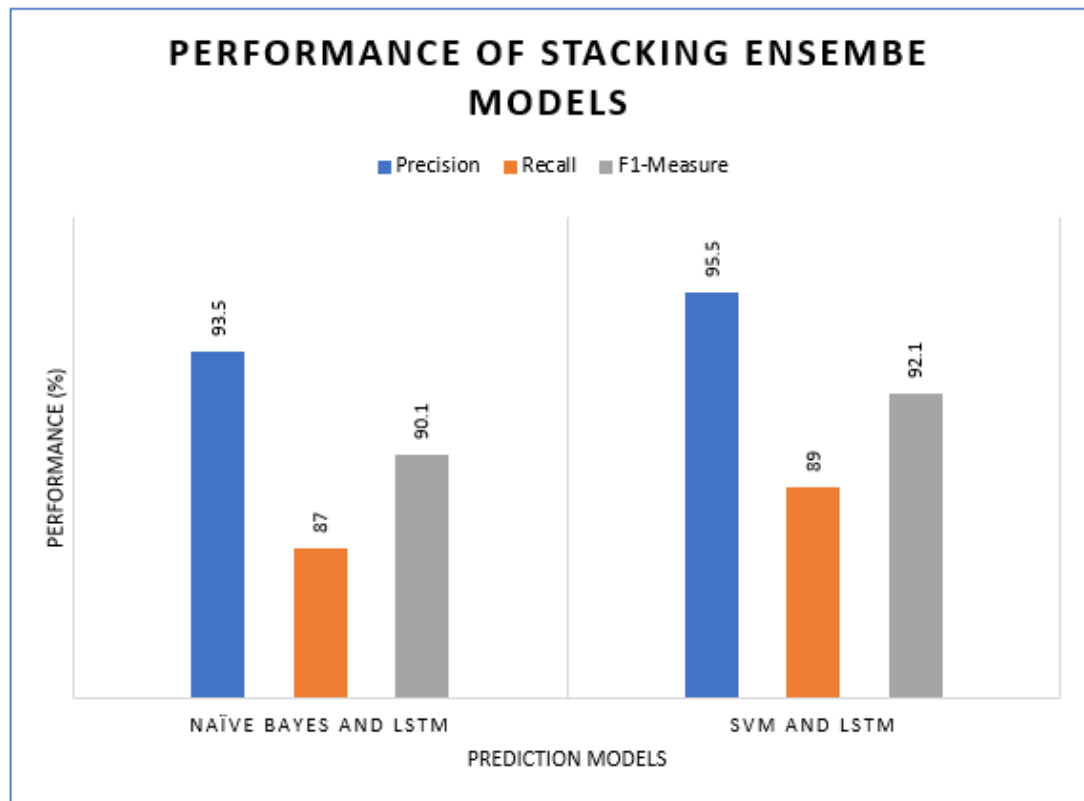


Figure 5.8: Performance of individual sentiment prediction models

unique ID. The sentiment results are considered as follows. For each train five attributes have observed sentiments. Positive sentiment is represented by the value 1, negative sentiment by -1 and neutral is by 0. The attributes used are punctuality, food quality, cleanliness, security and staff behaviour.

As presented in Table 5.3, it is understood that for each train sentiment score is provided for the aforementioned attributes that are important to IR to assess social feedback. This feedback when interpreted provides required business intelligence to IR for making well informed decisions.

As presented in Figure 5.9, the sentiments exhibited by followers of IR over Twitter social media are analysed. The horizontal axis shows train numbers for which analysis made. The vertical axis shows the sentiment value that may be either 0, 1 or -1 as discussed earlier. According to the analysis, the train with ID 6060 has got positive social feedback with respect to security, food quality and punctuality. It has to improve in cleanliness as it has negative social feedback. The IR got neutral feedback for this train with respect to staff behaviour. In case of the train with ID 4403, cleanliness is good as per the social feedback. However, its staff behaviour and punctuality are to be improvised when social feedback is

Table 5.3: Train wise sentiment analysis for Indian Railways

Train #	Punctuality	Food quality	Cleanliness	Staff Behaviour	Security
06060	1	1	-1	0	1
04403	-1	0	1	-1	0
13407	0	-1	1	1	0
05036	1	1	0	-1	1
12486	0	-1	1	1	1
06056	-1	0	-1	1	1
11487	1	1	1	-1	0
07092	-1	0	1	1	-1
06042	0	-1	0	1	1

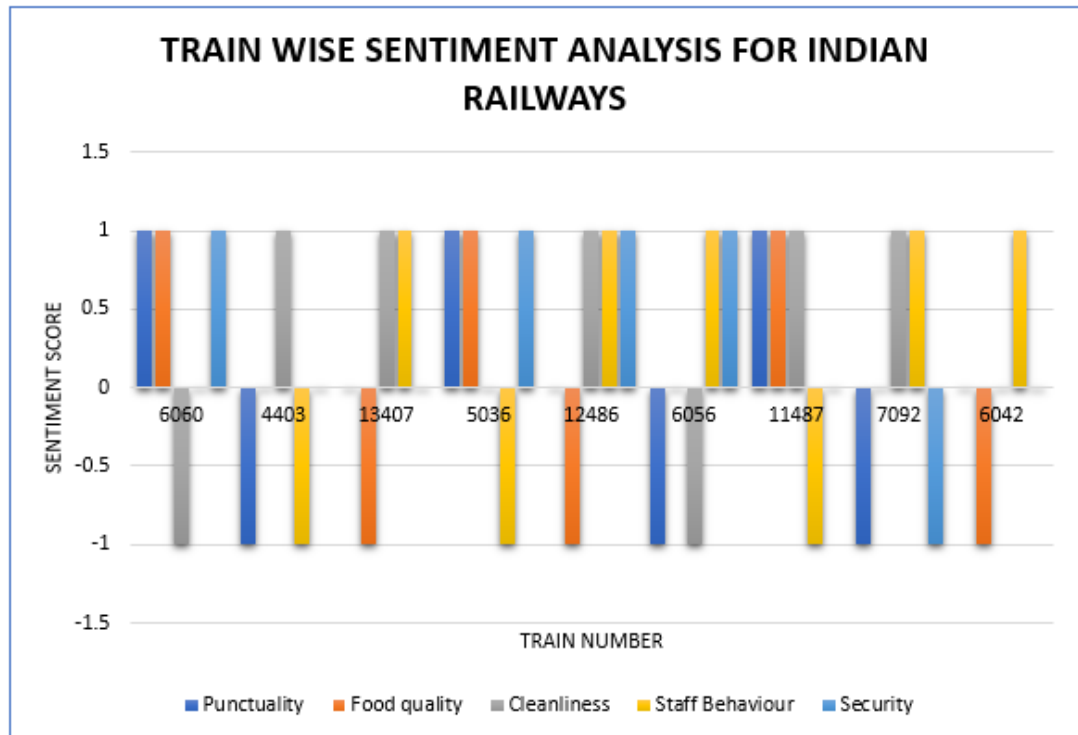


Figure 5.9: Train wise sentiment analysis for Indian Railways

considered. This train has neutral feedback on food quality and security.

The train with ID 13407 has got good social feedback on cleanliness and staff behaviour. However, its food quality is not good. This train has got neutral social feedback on punctuality and security. The train with ID 5036 has got good social feedback on punctuality, food quality and security. However, its staff behaviour is not good. This train has got neutral social feedback on cleanliness. The train with ID 12486 has got good social feedback on cleanliness, staff behaviour and security. However, its food quality is not good. This train has got neutral social feedback on punctuality. The train with ID 6056 has got good social feedback on security and staff behaviour. However, its punctuality and cleanliness are not good. This train has got neutral social feedback on food quality. The train with ID 11487 has got good social feedback on punctuality, food quality and cleanliness. However, its staff behaviour is not good. This train has got neutral social feedback on security. The train with ID 7092 has got good social feedback on cleanliness and staff behaviour. However, its punctuality and security are not good. This train has got neutral social feedback on food quality. The train with ID 6042 has got good social feedback on staff behaviour and security. However, its food quality is not good. This train has got neutral social feedback on punctuality and cleanliness. From the sentiment analysis, it is understood that IR needs to concentrate on the negative social feedback for each train with highest priority. Afterwards, it may improve the attributes for which neutral feedback is given. For all attributes that acquired positive social feedback also need focus so that it is possible to sustain that for long time and also replicate same with respect to other attributes.

5.6 SUMMARY

This chapter has covered the sentiment analysis methodology that is based on machine learning and deep learning approaches. The work presented has its connection to the outcome of the pre-processing framework discussed in the preceding chapter. The machine learning algorithms used for sentiment analysis are NB and SVM while deep learning method used for the empirical study is known as LSTM. These three models have different capabilities to produce sentiment classifications. However, it is observed that their performance is improved sig-

nificantly with the proposed pre-processing that involves multi-level filtering and NLP. In addition to this, the classification methods used for the experiments are combined using stacking approach for making an ensemble of classifiers to improve classification accuracy. Two ensemble models are used for experiments. The first model involves NB and LSTM while the second model involves SVM and LSTM. The results revealed that the deep learning model has shown superior performance over its counterparts. In the same fashion, the results also exhibited the fact that SVM+LSTM ensemble has better performance over NB+LSTM. This chapter also focused on the train wise sentiment analysis and discussion on the social feedback given for each train against the investigated attributes such as cleanliness, food quality, security, staff behaviour and punctuality. Chapter 6 covers the knowledge based system for business intelligence for Indian Railways.

Chapter 6

KNOWLEDGE BASED SYSTEM OF SENTIMENT ANALYSIS DATASET OF INDIAN RAILWAYS

SECTION 6.1 presents the frame work of knowledge based system of sentiment analysis dataset of Indian Railways. SUBSECTION 6.1.1 presents the procedure of creating Relational Database (RDB) schema of sentiment analysis database. SUBSECTION 6.1.2 presents the mapping procedure of RDB schema to Resource Description Framework (RDF) schema. SUBSECTION 6.1.3 presents the mapping procedure of RDF schema to Ontology. SUBSECTION 6.1.4 presents the formation of knowledge rules used in the Ontology. SUBSECTION 6.1.5 presents the formation of complex query in sparkle protocol and RDF query language (SPARQL) with results. Finally, SECTION 6.2 summarizes this chapter.

6.1 FRAME WORK OF KNOWLEDGE BASED SYSTEM OF SENTIMENT ANALYSIS DATASET OF INDIAN RAILWAYS

The proposed framework is for knowledge based system by transforming relational data model to semantic data model automatically. In other words, it is responsible to convert business intelligence into a knowledge based, accessible or

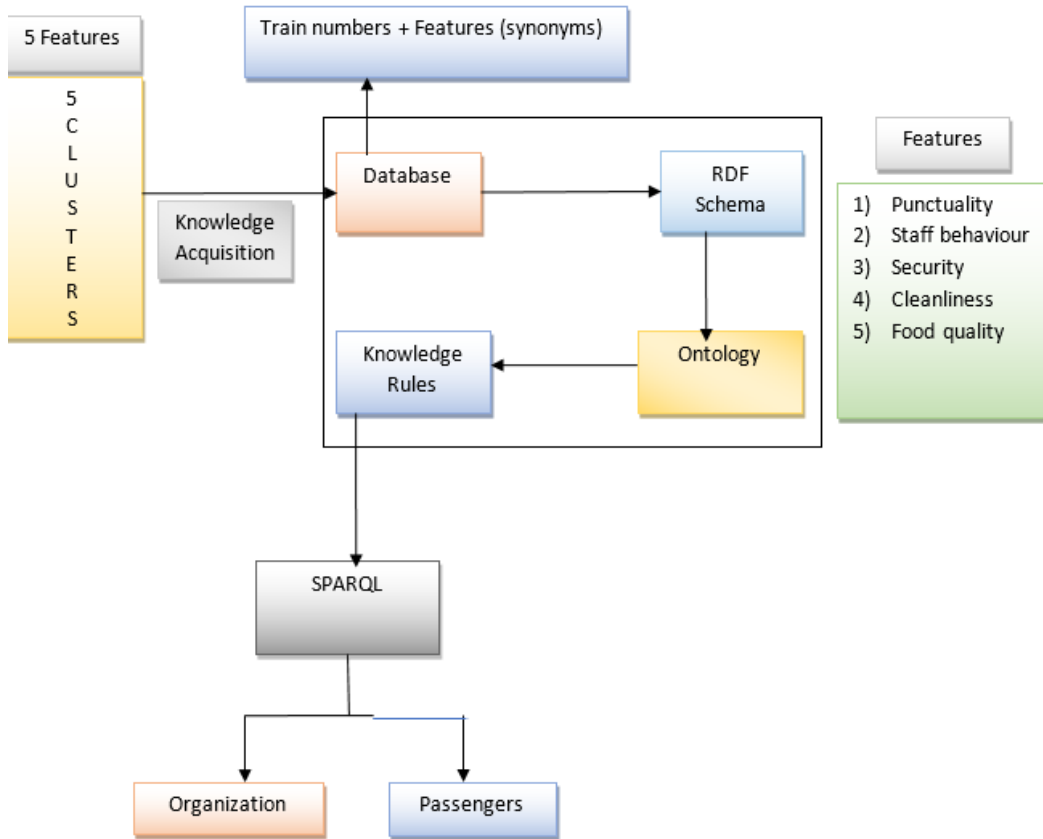


Figure 6.1: Framework for knowledge based system

programmable system that can be directly queried or queries through a program in an interoperable way. This is quite interesting and useful phenomenon which is taken care of by the framework shown in Figure 6.1. Input dataset is given to the framework. It is nothing but the Business Intelligence (BI) of Indian Railways that reflects social feedback on different operational trains across the country. The BI generated through sentiment analysis is in the form of 5 clusters. Each cluster provides BI pertaining to an important attribute associated with given train. The five attributes thus associated with each train are punctuality, staff behaviour, security, cleanliness and food quality. These five clusters are there, for instance, in a legacy system where data is stored in a relational model. When data is in relational model, there is limitation to its accessibility. The rationale behind this is that the data goes behind deep web. Therefore, transforming it to a semantic data model and making it accessible through interface that is interoperable and machine readable is supported by the framework.

The conventional database is converted into an RDF schema to facilitate access to deep web. It is the Resource Description Framework (RDF) which

resurfaces deep web with semantic interpretation and further processing of the BI. When BI is behind the deep web, it has limitations in accessibility. When the same is in the form of RDF, it has become interoperable with machine queries and human queries. Thus the proposed framework narrows down the gap between the traditional deep web and the surface web. Resource Description Framework Schema (RDFS) is a standard framework to store data. It forms semantic web and Web Ontology Language (OWL) for interpretation of data. RDF schema has both literals and semantic meanings of the same. Thus it provides rich interoperable interface to a knowledge domain in fully automated and interactive fashion. The concept of ontology has made this possible as ontology is the knowledge representation which is made up of concepts and relationships among them.

RDF schema was introduced in order to handle situations where web data needs to be processed and exchanged by applications instead of just showing data to users. This ability of exchanging data between applications makes RDF very useful in the contemporary era. Semantic data models like RDF and frameworks based can help organizations of specific domain to organize domain BI or knowledge and share it throughout the enterprise. Such models ensure that “semantic” means of accessing knowledge leads to fully automated systems in distributed environments. Many experts with systems are constructed with from knowledge with extracted in whole or in part from of databases. With increased to amount of the knowledge stored in the databases, the acquisition with of such knowledge is becomes more than difficult. Knowledge engineers are of human professionals who are able to the communicate with expert’s system and consolidate with knowledge from various sources to be build a valid knowledge based system. They can use computers systems and special methods where to overcome the difficulties in the knowledge engineering. Validation is very much a critical process with in the system whole of knowledge-based system lifecycle process. A knowledge based incorporated with into such systems has to be verified validated. There have been many of approaches to the develop the specialised procedures and techniques, aimed at assuring the highest level of the knowledge equality.

Table 6.1: An excerpt from dataset containing sentiment details for Indian Railways

Train #	Punctuality	Food quality	Cleanliness	Security	Staff Behaviour
06060	1	1	-1	0	1
04403	-1	0	1	-1	0
13407	0	-1	1	1	0
05036	1	1	0	-1	1
12486	0	-1	1	1	1
06056	-1	0	-1	1	1
11487	1	1	1	-1	0
07092	-1	0	1	1	-1
06042	0	-1	0	1	1
04912	1	1	-1	0	-1

6.1.1 RELATIONAL DATABASE (RDB) SCHEMA OF SENTIMENT ANALYSIS DATABASE

The input dataset is in the form of clusters. It is stored in relational database. The dataset contains any kind of relational data. However, we have considered the data of Indian Railways. It is in the form of sentiment value associated with different train numbers. For each train five attributes have been observed which are considered as sentiments. Positive sentiment is represented by the value 1, negative sentiment by -1 and neutral is by 0. The attributes used are punctuality, food quality, cleanliness, security and staff behaviour. These attributes are used to obtain synonyms using lexical dictionary like WordNet. Therefore, every attribute has n-number of synonyms. All of them can be used while transforming and while making SPARQL queries. Thus more flexible knowledge based data retrieval system will be in place. Table 6.1 shows the input data given to the knowledge based transformation representation.

Each train has associated sentiments obtained from sentiment analysis as part of our prior work [24]. The train 06060 has positive sentiments related to three attributes such as punctuality, food quality and staff behaviour. It has neutral sentiment with respect to security. In case of cleanliness, it has negative sentiment. In the same fashion different trains do have the social feedback resulted from the sentiment analysis. When this data is given to the proposed framework,

it automatically stores in relational database. Up to this extent it is required for applications that are specific to accessing relational databases using SQL queries. It is important to convert from relational schema to RDF schema. In other words, it is the conversion from relational data model to semantic data model.

6.1.2 RDB SCHEMA TO RESOURCE DESCRIPTION FRAMEWORK (RDF) SCHEMA

Mapping from relational data model to semantic data model (RDF in this case) is done with three layers associated with the relational model use case. The first layer is known as relational database area of knowledge. The second layer is domain data knowledge and the third layer is known as application specific knowledge.

Database Mapping Procedure

In case of database having multiple tables, the database mapping procedure is an iterative and incremental approach. It uses all tables denoted as **T** and all attributes of each table denoted as **A** in an associated schema denoted as **S**. The tables in relational model need to have corresponding classes in semantic data model. Therefore, a table **t** is mapped to class **c**. The columns in table are mapped to properties of corresponding class. The column data types are mapped to responding data types of properties. The primary key, foreign key and other constraints associated with a relational table are mapped to corresponding keys in semantic data model RDF. Then different relationships are mapped to equivalent items in RDF. It considers different aspects like disjointness, transitive chain of relation and the degree of relationship.

6.1.3 RDB to RDF Mapping (RRM) Algorithm

This algorithm plays crucial role in transforming traditional deep web models into modern semantic web models. Stated differently, it helps to convert traditional knowledge store (RDBMS) into RDF schema that is machine readable and interoperable besides making web applications to represent and exchange knowhow. It takes all relational tables as input and performs various procedures (complexities such as constraints, relationships and so on) to have a compati-

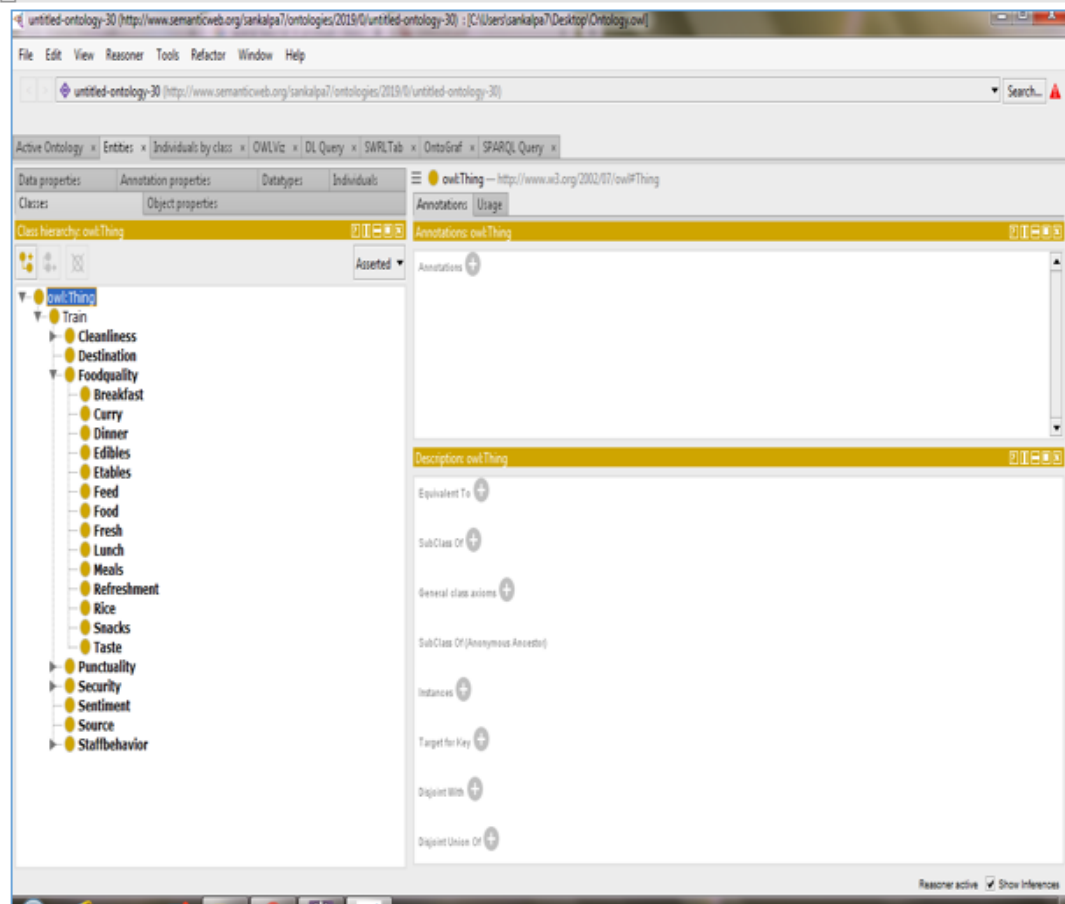


Figure 6.2: Shows class names with the help of protégé tool with five features

ble RDF schema and data is mapped from relational model to knowledge based model with underlying mechanisms of the algorithm.

As shown in Algorithm 1, the mapping of different database objects is carried out. They include the whole database, tables, columns, constraints and relationships. This algorithm is implemented to have knowledge based system that can be used to have interoperable and machine readable queries.

6.1.4 MAPPING RDF SCHEMA TO ONTOLOGY

The implementation of the algorithm is based on the BI available for Indian Railways that is in the form of social feedback obtained through sentiment analysis. The application has provision to execute mapping procedures to build a knowledge based system. The algorithm described in the previous section is implemented to achieve this. Before showing the application, the rules pertaining to the proposed empirical study are as follows:

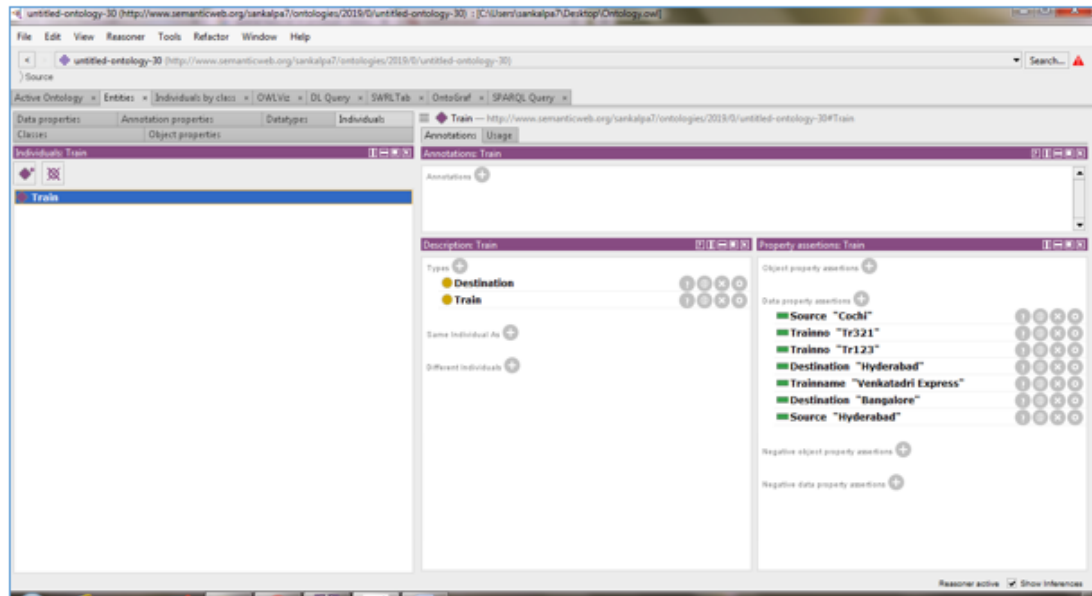


Figure 6.3: Shows train selected in types of data property

As shown in Figure 6.2, class hierarchy of OWL is built. It is compatible with the Indian Railways case study for building knowledge based system. Each class may have sub classes in the semantic domain. The classes reflect all the attributes associated with each train for querying BI.

Above figure 6.3 shows about the we used in train class with types destination and trains and contain the data properties as source, train number, destination. Where in ontology mapping classes are mapped into object property and columns are mapped into data property with characterised functions and individuals with respect of domain and range.

6.1.5 FORMATION OF KNOWLEDGE RULES USED IN THE ONTOLOGY

knowledge-base was to describe one of the two sub-systems of a knowledge-based system. A knowledge-based system consists of a knowledge-base that represents facts about the world and inference engine that can reason about those facts and use rules and other forms of logic to deduce new facts or highlight inconsistencies. Knowledge rules are made in order to obtain required information based on knowledge based Intelligence. When such queries are made from different applications, the proposed implementation could realize the expected knowledge based system that shares to BI to different applications or individuals

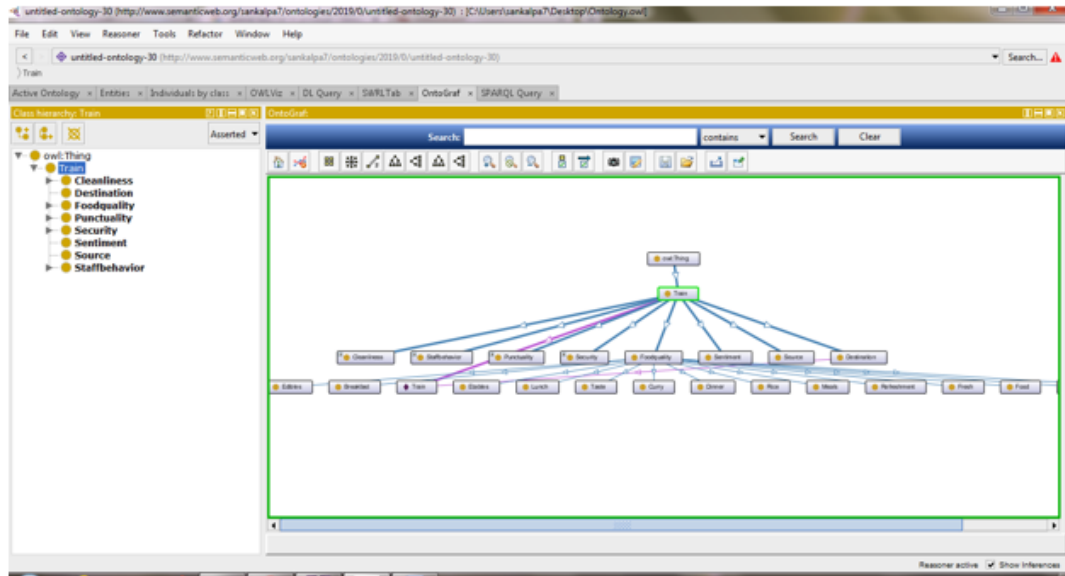


Figure 6.4: onto graph using protégé tool

or query based interface onto graph with train as class and ontology owl as main class with sub classes with object and data properties with individuals. Where we find the object oriented methods, properties and it looks like a tree or graph based. Where we can find differences between database and ontology structure.

As shown in above table 6.2 we have combination of features like punctuality, cleanliness, food quality, security, staff behaviour which able to give knowledge on our sentiment values which are positive, negative, neutral the rules contain the essence of sentiment analysis that is based on each train. The attributes considered are used in the formation of rules. As presented in Figure 6.3, the Train is the chosen item under the individuals. Its description and property assertions are made available.

As presented in Figure 6.5, knowledge rules are made in order to obtain required information based on knowledge based Intelligence. When such queries are made from different applications, the proposed implementation could realize the expected knowledge based system that shares to BI to different applications or individuals or query based interface. Sample of two rules are executed and checked with Term Reasoned.

Table 6.2: An excerpt from dataset containing sentiment details for Indian Railways

IF (Train No. "is" 06060) and (Food Quality "is" 1) and (Punctuality "is" 1) THEN (Sentiment "is" Positive)	Knowledge rule with train number explain the train food quality and punctuality sentiment values as positive
IF (Train No. "is" 04403) and (Food Quality "is" 0) and (Punctuality "is" 0) THEN (Sentiment "is" Neutral)	Knowledge rule with train number explain the train food quality and punctuality sentiment values as neutral
IF (Train No. "is" 13407) and (security "is" 0) and (Punctuality "is" 0) THEN (Sentiment "is" Neutral)	Knowledge rule with train number explain the train security and punctuality sentiment values as neutral
IF (Train No. "is" 05036) and (Food Quality "is" 1) and (Punctuality "is" 1) THEN (Sentiment "is" Positive)	Knowledge rule with train number explain the train food quality and punctuality sentiment values as positive
IF (Train No. "is" 12486) and (cleanliness "is" 1) and (Punctuality "is" 0) THEN (Sentiment "is" Positive)	Knowledge rule with train number explain the train cleanliness and punctuality sentiment values as positive
IF (Train No. "is" 06056) and (Food Quality "is" 0) and (security "is" 1) THEN (Sentiment "is" Positive)	Knowledge rule with train number explain the train food quality and security sentiment values as positive
IF (Train No. "is" 11487) and (Food Quality "is" 1) and (Punctuality "is" 1) THEN (Sentiment "is" Positive)	Knowledge rule with train number explain the train food quality and punctuality sentiment values as positive
IF (Train No. "is" 07092) and (Food Quality "is" -1) and (Punctuality "is" 0) THEN (Sentiment "is" Negative)	Knowledge rule with train number explain the train food quality and punctuality sentiment values as negative
IF (Train No. "is" 06042) and (staff behaviour "is" -1) and (Punctuality "is" 0) THEN (Sentiment "is" Negative)	Knowledge rule with train number explain the train staff behaviour and punctuality sentiment values as negative
IF (Train No. "is" 04912) and (Food Quality "is" 1) and (Punctuality "is" 1) THEN (Sentiment "is" Positive)	Knowledge rule with train number explain the train food quality and punctuality sentiment values as positive

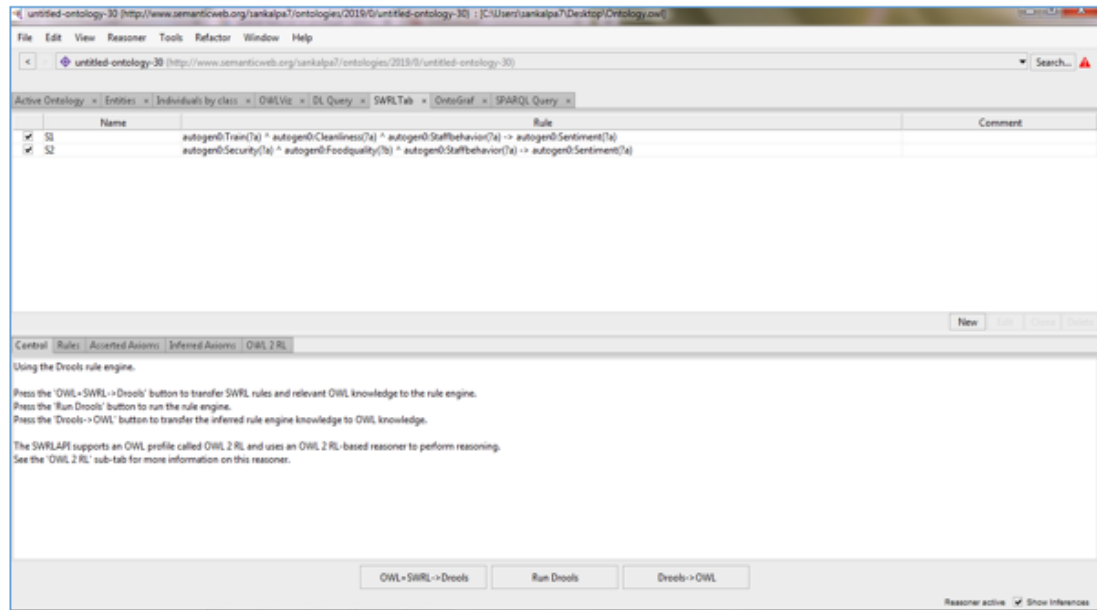


Figure 6.5: knowledge rules using protégé tool with names S1 and S2

6.1.6 FORMATION OF SPARQL QUERY

In the case of the queries that are read data from the database, the SPARQL language specifies four different to query variations for different purposes

SELECT query used to extract raw values from a SPARQL end point, the results are returned in the table format

CONSTRUCT query used to extract information from the SPARQL endpoint and to transform the results into valid of RDF

ASK query used to provide a simple with True/False result for the query on a SPARQL end point

DESCRIBE query used to extract of an RDF graph from the SPARQL as endpoint, the content of the which is the left to the final point to decided, based on what the maintainer makes as useful information

As presented in Figure 6.6, sample SPARQL query is made on the knowledge based system built by transforming RDB schema into OWL. SPARQL is an RDF query language which has its syntax and semantics. In other words, it is semantic query language for knowledge databases. Retrieving and manipulating data present in the form of RDF is made using SPARQL.

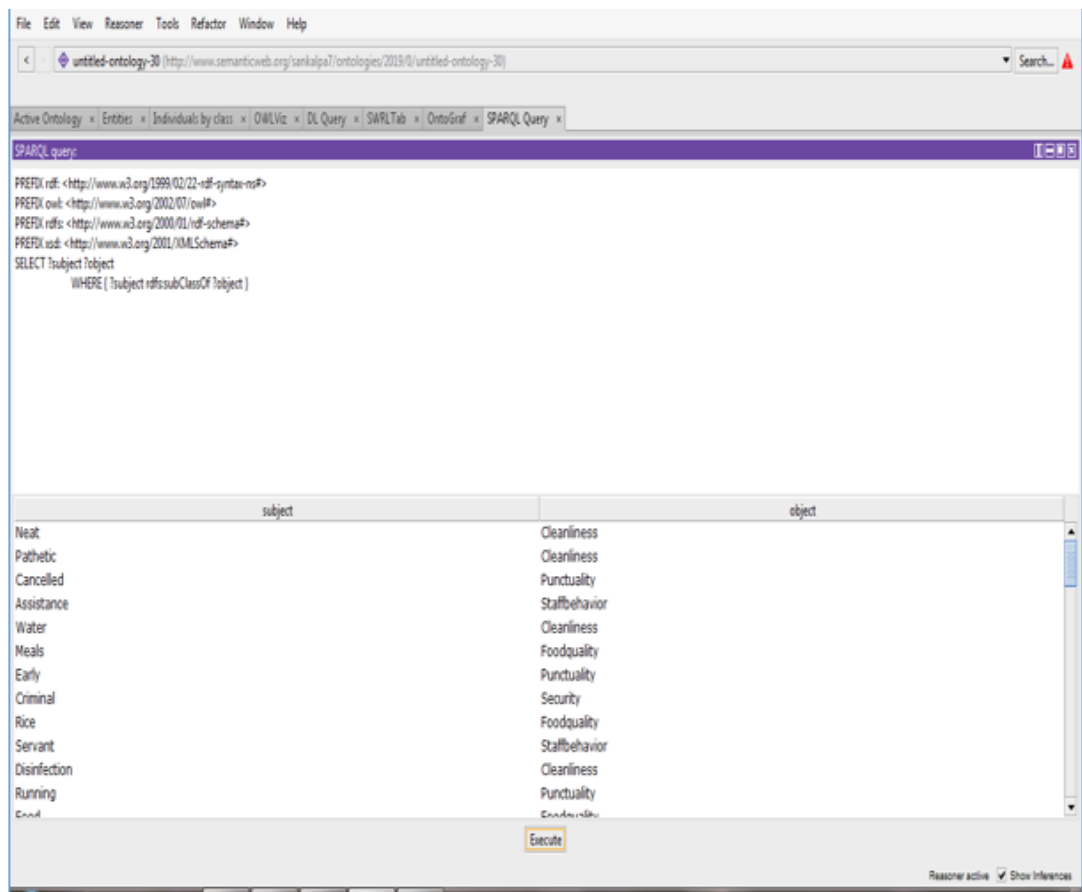


Figure 6.6: Sample SPARQL query made

6.2 summarizes the chapter

Knowledge based intelligence system is made by implementing RDF schema to OWL having machine readable and interoperable access to a knowledge base. The problem of deep web access in the contemporary era where knowledge needs to be represented and shared in a quite natural and intuitive way among organizations and individuals is investigated. A framework is proposed to have a systematic approach to convert traditional RDB model to semantic data model. The rationale behind this is that Web Ontology Language (OWL) and RDF schema provide machine readable knowledge base that not only represents knowledge in the form of concepts but also provide means of data exchange across web applications. Such data can be accessed with either direct SPARQL query or query made through a program in M2M environment. An algorithm is proposed which has multiple procedures to map different objects of relational database to the equivalent objects in RDF schema. The empirical study revealed that the proposed framework is useful for transforming RDB behind deep web into readily accessible semantic data model which is proved to be ideal for modern business use cases. The proposed framework needs to be standardized with different database dialects (RDB) available in the real world so as to make it robust and work for all databases. This is one direction for future work. Another direction is to work on simplifying query interface to end users by building query browser to facilitate interactive means of accessing knowledge to make well informed decisions.

Algorithm 6 RDB to RDF Mapping (RRM)

INPUT: All relational tables **T**

Mapping Database

Procedure MapDatabase(S)

Begin

MapTables(S)

MapColumns(S)

MapConstraints(S)

MapRelationships(S)

End

Mapping Tables

Procedure MapDatabase(S)

Begin

for each the table **t** in **T** **do**

 Create the Class C in RDF repository C with RDF

 <owl:Classrdf:ID="Ci" />

end for

End

Mapping Coloumns

Procedure MapColumns(S)

Begin

for each the table **t** in **T** **do**

for each the table **c** in **C** **do**

 Use class c of t

Set c as property of hast

getType(t)

<owl:DatatypePropertyrdf:Identity="hasA">

<rdfs:domainrdf:resource framework="#C" />

<rdfs:rangerdf:resource framework="& xsd;type_ equivalent" />

</owl:DatatypeProperty>

end for

end for

End

Chapter 7

Conclusion and Future scope

Information gathering considers different aspects. However, finding what is the opinion of other people is an important aspect in information gathering. In the capacity of customers, people do think about what other people are thinking about a service or product. This aspect is even more important to organizations as they can make use of opinions of others to improve product or services. Moreover, organizations can influence opinions of people as well by improving Quality of Service (QoS). Organizations want to know the events and the feedback of the events towards product management and marketing. They wanted to have both traditional customer feedbacks in conventional channels and the social feedback through web based social networks. Thus a comprehensive business intelligence (BI) is expected by the organizations to ensure that their brands or services will give positive influences.

The technologies associated with Web 2.0 paved way for different aspects of higher importance in the contemporary era. They are known as social networking, tagging, podcasting, social bookmarking, reviewing and blogging. With these there is increase in the opinionated resources that bring about challenges and opportunities to organizations. This has led to significant attraction towards sentiment analysis or opinion mining. Therefore, the textual content with subjectivity, opinions or sentiments is give importance with computational treatment. Many tools came into existence for text analytics. With these tools and tools specific to sentiment analysis such as WordNet and SentiWordNet organizations are striving to gain timely social feedback that reflects opinions of customers on specific product or service [1]. By this BI, companies can take necessary steps and see that the customers' opinions will be positive and in turn influence other

people as well.

In this thesis, we have performed the Sentiment Analysis on Indian Railways(IR) Tweets on respective features such as Punctuality, Staff Behaviour, Cleanliness, Security, Food quality. This analysis will help the railway organisation to improve their quality of services on above set features. As we know that Indian Railways is Asia's largest and the world's second largest rail network of India operated by the Ministry of Railways. It has more than 11,000 locomotives and over 70,000 passenger coaches. It transports around 2.5 crore passengers daily. IR carried around 8.26 billion passengers and 1.16 billion tonnes of freight in the fiscal year ending March 2018. As IR is the preferred transport to most of the Indians, it is observed that the Online Social Network (OSN) carry social feedback on IR. In addition to the direct feedback given by passengers in traditional approaches, passengers are equipped with social platforms like Twitter to provide their feedback in the form of reviews, micro-blogging and exchange of opinions in social media. Nevertheless, the tweets from Twitter carry the essence of social feedback given by passengers of IR. In this context, it is not wise to ignore social feedback. In fact, it is indispensable for any organization to consider opinions of public available in social media. IR is no exception to this. Swacch Bharat Abhiyan Prime Minister Organization (PMO) considers the whole nation including IR to improve in various parameters of healthy approaches. Having understood about the importance of social feedback for IR, Sentiment analysis is made on the tweets pertaining to Indian Railways so as to help it to get benefited for improving services and gain highly positive opinions on its services. We aimed to perform sentiment analysis on different features of IR and further to develop knowledge based sentiment analysis system that provides essential of social feedback on different attributes of IR such as Staff Behavior, Punctuality, Cleanliness, Food Quality and Security. The research carried out in this thesis leads to an out of the box solution that can be adapted as part of Decision Support System (DSS) for IR. It also has impact on other stakeholders of IR.

This research which is aimed at developing a Knowledge Based System(KBS) for Sentiment Analysis dataset that act value to IR. The main topic and challenges of this thesis is to develop KBS using ontology of Sentiment Analysis dataset based on different features of IR such as Punctuality, Staff Behavior, Security, Cleanliness and Food quality which will help IR to make a strategic decision to

improve its services as well as it will also help the passengers to book their ticket of respective trains not only on basis of the availability of the reservation but also on the basis of the performance of the said features.

We have investigated the nature of tweets and subjected them to sentiment analysis which has mechanism to process twitter tweets in order to classify sentiments values. We extend this research to explore the framework with more ideal and efficient means of sentiment analysis and evaluate it with different domains. Further, we present a framework with a hybrid approach for sentiment Analysis. It exploits lexicon creation and polarity detection concepts besides a three-point classification of tweets of Indian railways as result of sentiment analysis. The topic based sentiment analysis algorithm is presented to implement the phases in the proposed framework. We propose a framework to evaluate the performance of machine learning algorithm such as Naïve Bayes and SVM and deep learning algorithm such as LSTM in sentiment classification. We have also used ensemble classifier with stacking that ensemble machine-learning algorithm with deep learning algorithm. The experimental results show that the F-Measure of single classifier i.e. Naïve Bayes is (80.5%), Support Vector Machine is (85.6%) and for Long Short Term Memory the F-Measure is (88.1%). Further, the F-Measure of Ensemble classifier i.e. Naïve Bayes with LSTM is (90.1%) while the F-Measure of Support Vector Machine with LSTM is (92.1%). Next, we have compared the F-Measure of different single and ensemble classifier and results explored that Support Vector Machine with LSTM is better. The empirical study revealed the utility of the proposed framework in ascertaining valuable insights besides understanding the additional value added by long short-term memory in Ensemble stacking approach towards better performance of Sentiment Classification. Further, we developed a knowledge-based system in which implementation of Relational Database (RDB) to Resource Description Framework (RDF) schema and to Ontology has been done which have machine readable and interoperable access to a knowledge based. The problem of deep web access in the contemporary era where knowledge needs to be represented and shared in quite natural and intuitive way among organizations and individuals is investigated. A framework is proposed to have a systematic approach to convert traditional RDB model to semantic data model in knowledge based intelligent system. The proposed framework needs to be standardized with different database dialects available in the

real world so as to make it robust and work for all databases.

Despite the possible positive outcomes shown, there are some limitations in applying automatic analysis due to the difficulty to implement it because of the ambiguity of natural language processing and also the characteristics of the posted content. The analysis of social media tweets is an example of this, for they are usually coupled with hashtags, emoticons and links, etc. creating difficulties in determining the expressed sentiment analysis. In addition, there is a need for automatic techniques that require large amount datasets of annotated posts or lexical databases where emotional detection words are associated with sentiment analysis values. Another important aspect of that analyses are suitable for the English language, in which there is a limitation for other languages [4].

In the field of sentiment analysis are some challenges in a range of scenarios, in terms of architecture and application domains with unclear or scarce datasets. Also, there is a lack of labelled machine learning data, which can pose a barrier to the advancements in this sentiment analysis area [5].

This is one direction for future work. Another direction is to work on simplifying query interface to end users by building query browser to facilitate interactive means of accessing knowledge to make well-informed decisions. This knowledge based system is accessible to humans and also programs in heterogeneous Machine-to-Machine (M2M) environments. It can be used in decision support system (DSS) of Indian railways for effective decision-making as it conveys sentiments of Indian Railways Tweets that is helpful for passengers, organization, and administrators to take further steps.

References

- [1] Georgios Paltoglou. Sentiment Analysis in Social Media. pages 3–17.