An intelligent Framework for effective Sentimental Analysis in Urdu Language

By

Maria Masood

FALL 2018-MS-18(CSE) 00000275893

Supervisor

Dr. Farooque Azam

Department of Computer Engineering

College Of Electrical & Mechanical Engineering (CEME)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

APRIL, 2022

An intelligent Framework for effective Sentimental Analysis in Urdu Language

By

Maria Masood (0000000275893)

Supervisor

Dr. Farooque Azam

Co-supervisor

Dr. Wasi Haider Butt

A thesis submitted in conformity with the requirements for the degree of *Master of Science* in Software Engineering

Department of Computer Engineering
College Of Electrical & Mechanical Engineering (CEME)
National University of Sciences and Technology (NUST)
Islamabad, Pakistan

APRIL, 2022

Declaration

I, *Maria Masood* certify that this research work titled *"An intelligent Framework for effective Sentimental Analysis in Urdu Language"* is my own work under the supervision of **Dr. Farooque Azam**. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged/referred.
I confirm that:

1. This work was done wholly or mainly while in candidature for a Master of Science degree at NUST
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at NUST or any other institution, this has been clearly stated
3. Where I have consulted the published work of others, this is always clearly attributed
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work
5. I have acknowledged all main sources of help
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself

---

Maria Masood,
FALL 2018-MS-18(CSE) 00000275893

Copyright Notice

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of SMME, NUST. Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.

- The ownership of any intellectual property rights which may be described in this thesis is vested in SMME, NUST, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of SMME, which will prescribe the terms and conditions of any such agreement.

- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of SMME, NUST, Islamabad.

Dedication

Dedicated to my beloved parents,sisters and husband whose tremendous support and cooperation led me to this wonderful accomplishment.

Abstract

In recent times, Sentiment analysis has become a significant means for framing a successful business and can be very helpful in predicting customer trends to help organizations in their decision-making process. Though many software applications are available in the market for text analysis, one of the major limitations of such applications is that they are developed for rich languages like English, German, Spanish, Arabic, etc. and less popular languages like Urdu, Hindi, Roman Urdu are neglected due to lack of availability of resources. Therefore, this research project will provide an implementation of sentiment analysis in the Urdu language. Firstly, preprocessing is performed and a small-scale manual dictionary of 830 Urdu stem words is introduced for stemming. Then a deep learning-based framework of LSTM is used for Urdu sentiment analysis. Experimental results show a high classification accuracy of 86.03% with LSTM that captures sequence information effectively to analyze sentiments than the conventional supervised machine learning approaches.

Keywords: *Sentiment Analysis, LSTM classifier, Urdu, Preprocessing, Dataset*

Acknowledgments

I am extremely thankful to Allah Almighty for his bountiful blessings throughout this work. Indeed, this would not have been possible without his substantial guidance through each and every step, and for putting me across people who could drive me through this work in a superlative manner. Indeed, none be worthy of praise but the Almighty.

I am profusely grateful to my family for their love, prayers, support, and sacrifices for educating and preparing me for my future. I also thank my siblings who encouraged me and prayed for me throughout the time of my research.

I would also like to express my gratitude to my supervisor **Dr. Farooque Azam** and my co-supervisor, **Dr. Wasi Haider** Butt, for their constant motivation, patience, enthusiasm, and immense knowledge. Their guidance helped me throughout my research and writing of this thesis. I could not have imagined having a better advisor and mentor for my MS study.

I would like to pay special thanks to **Muhammad Waseem Anwar** for his incredible cooperation and for providing help at every phase of this thesis. He has guided me and encouraged me to carry on and has contributed to this thesis with a major impact.

I would also like to thank my Guidance Committee Members **Dr. M. Umar Farooq** and **Dr. Arsalan Shaukat** for being on my thesis guidance and evaluation committee. Their recommendations are very valued for improvement of the work.

Last but not the least, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Thanks for all your encouragement!

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

ULP	Urdu Language Processing
NLP	Natural Language Processing
SA	Sentiment Analysis
UTSA	Urdu Text Sentiment Analysis
DL	Deep Learning
ML	Machine Learning
RNN	Recurrent Neural Network
RNTN	Recursive Neural Tensor Network
CNN	Convolutional Neural Network
GRU	Gated Recurrent Unit
LSTM	Long Short Term Memory
ANN	Artificial Neural Network
SVM	Support Vector Machine
KNN	K-Nearest Neighbor
NB	Naïve Bayes

DT	Decision Tree
NER	Named Entity Recognition
RMSE	Root Mean Squared Error
LR	Logistic Regression

Chapter 1

Introduction

Definition of the sentiment word defines various terms, including awareness of the anonymous performance regarding emotions, opinions, views, and feelings. Different authors have introduced it as "opinion mining" [2]. In those days, research in sentiment analysis played a significant role; people could examine emotions and show their feelings in their own words. Both categories, such as thoughts and opinions, are emotional. They communicate somebody's feelings or emotions which may be written or spoken. Classified the emotion can be as positive or negative [3]. Social media have become a significant and vital part of our life. It unites individuals with the world. Therefore, it is expected that individuals should encourage online entertainment for the right and accurate data. However, tragically, while investigating the posts on online entertainment, then, at that point, it was seen that individuals had incorrect data by posting figures references.

1.1 Background Study

In the most recent times, quite a considerable number of researchers have been attracted to blogging websites, social media networks, and online discussion forums. The exchange of information on various social networking platforms has resulted in advanced applications to facilitate firms and individuals in their decision-making process [4]. Big enterprises bring about a significant advancement in their planning and

decision making to structured and unstructured data by applying analysis on data and sentiments. Sentimental analysis has extended over multiple fields such as marketing, social analysis, and customer information. In addition, sentiment analysis is utilised to improve product quality or to better understand public opinion on various topics. Sentiment analysis, also known as opinion mining, is the study of people's feelings, opinions, emotions, ratings, and behaviour in relation to objects such as products, services, firms, people, events, subjects, and traits [5]. Sentiment analysis has recently been applied to a variety of sectors, including marketing, customer information, and sociological analysis. Many business domain applications of sentiment analysis may be found, including brand reputation, recommender systems, e-commerce, and social media monitoring.

Sentiment mining of the English Language holds diverse literature. Currently, a small number of scholars are focusing on the sentiment classification of other languages such as Italian, Arabic, Urdu, and Hindi. Pakistan's national language is Urdu. For written and oral communication, there are around 200 million individuals who speak Urdu [6]. The Urdu language is hard to understand and process due to challenging script, complex morphological structure, vague word boundaries, and difficulty in stemming [7]. In comparison to other languages, the Urdu language has a dearth of language processing resources such as stemming and lemmatization tools, as well as stop words lists. In addition, the use of deep learning approaches for Urdu sentiment analysis has received little attention [8]. Sufficient literature exists on the difficulties, utilization, and policies of SA for the English language. However, researchers have done very little research on the SA of the Urdu language. As a result, a comprehensive framework for text processing in Urdu is required to meet preprocessing needs such as stop words removal, stemming, and normalising. Our research intends to improve the quality of classifiers in terms of accuracy, precision, f-measure, and recall using a superior deep learning approach in sentiment analysis, and to develop a framework that is particularly suited for dealing with the scarce resources language Urdu. The proposed methodology is done by preprocessing followed by a deep learning approach of LSTM.

1.2 Motivation

Social media platforms spread negativity and positivity. Negativity severely affects the younger generation, prompts social conflict, and spreads false impressions. This should be featured, halted, and eliminated from social media to keep away from social conflict. Nowadays, the presence of different websites, including Urdu, has altogether expanded. The majority of Pakistan's official language is Urdu, and also India's specific language is used by a colossal number of people overall for conveying. In Pakistan, for the most part, visited sites presenting their substance in Urdu [9]. Urdu shows a couple of challenges in the handling of language. For instance, Urdu uses the formal and easygoing type of action word, and each thing has two prospects, i.e., either masculine or feminine.

Research is being conducted on sentiment analysis and gaining interest in academics, industry, and academia. It is the natural ability of humans to analyze sentiments by using expressions or tone of voice. Still, it is challenging to train the machine and produce accurate output. The fundamental methodologies for performing text analysis in English (lexicon-based, machine learning, and hybrid method) can also be applied for Urdu. Despite the considerable variations in grammar, orthography, and morphology between English and Urdu, research and development are still required. Furthermore, English morphological analyzers and POS taggers cannot be rigorously employed due to Urdu's complicated morphology and unstructured format.

The suggested thesis' major purpose and motivation is to use a deep learning approach to facilitate sentiment analysis for the Urdu language, as there has not been much research done on the subject. As a Pakistani, our research will also support the Urdu language in sentiment analysis to help it gain traction in digital marketing.

1.3 Problem statement

Dissimilar to English, the Urdu language is made from right to left side. The limit between the words isn't perceivable with the end goal that all of the time, what lies

there is sensible despite the fact that it contains no additional areas between different words. By applying the NLP dealing with strategies, sentiment analysis can remove emotional data. It is used to describe the text as neutral as O, positive as P, and negative as N regarding different scenarios like thing, organization, subject, and event. This direction is to investigate to help the researchers, which is rotated to help plan strategy. Nevertheless, minimal research work has been done in sentiment analysis of Urdu [10]. Sentiment analysis planned for English is not adequate for the Urdu language because of the different substance and structure of morphological [11].

Urdu has linguistic components that are different and similarly conflicting with the English language. Like a few one or two dialects, the online asset of Urdu is as well as getting popular as individuals should share impressions and pass on sentiments in their nearby dialect. The Urdu language has different varieties of words that have the same meaning. For example, the word "KAY" in the Urdu language can be confused for "KEH" and correspondingly "PEH" with the "PAY." It implies that similar rhyming words might have different meanings. In Urdu Compositions, space utilization isn't contagious, and when a rule, space addition or prompts exclusion of space issue in words. Another example of the word "INKA" has an issue with space exclusion; cardinal words are a combination of hybrid processes as an individual word. Also, the addition of space issue, for instance, "AQALMAND," meaning as Intelligent, however when the split into the tokens in all single actuality word, for example, "AQAL" and "MAND," then the machine will consider this word as a compound word, which is tended to by a 2-phase framework [11].

In comparison to English, sentiment analysis for the Urdu language is underdeveloped. For sentiment analysis of the English language, a number of approaches have been developed. However, because scholars have not focused on Urdu, the number of strategies and methodologies for Urdu is limited. Furthermore, English morphological analyzers and POS taggers cannot be rigorously employed due to Urdu's complicated morphology and unstructured format. As a result, our research intends to improve the accuracy, f-measure, recall, and precision of deep learning approaches based on sentence polarity.

1.4 Generic Solution Steps

Literature represents various sentiment analysis procedures that have a few basic steps. These steps are primarily used, which are shown in Figure 1.1.

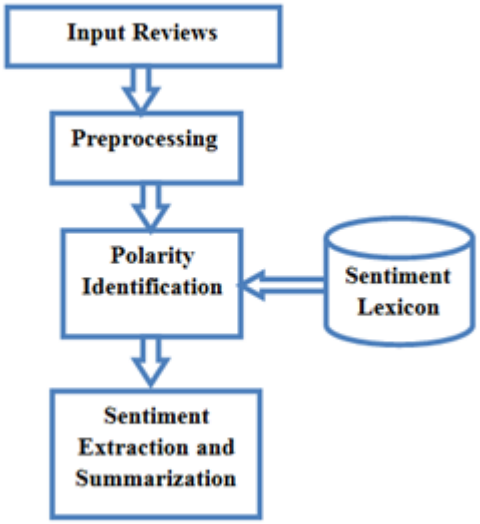

Figure 1.1: Sentiment Analysis' Basic Steps

The first step is Preprocessing. A progression of all basic steps is completed in this stage, for example, Part of Speech Tagging, Noise Removal, Detection of Sentence Boundary, and sometimes Tokenization of Words, Detection of Sentence Boundary. After that, related issues with preprocessing steps are more needed to handle. The next step is to use the label and its enduring popularity. Then, a word is considered, sentence and text or instant messages to discover whether it is neutral, positive, or negative. It is fundamental to give accurate data training to models that plan different algorithms effectively. Besides, data should be appropriately big to set up the trained model effectively. The next step of sentiment analysis of Urdu or any other language is Tokenization, which is performed. The "Tokenization" method separates or breaks the given density into the units called "Tokens." These tokens can be words, punctuation marks, and digits. Tokenization does this work by searching the boundaries of the word. At this stage, the exact meaning of words when different tokens are produced

utilizing the different sentences, and then it proceeds to the identification of polarity confirmation stage. Polarities of each word are ultimately settled as Negative as -1, Neutral as 0, and Positive as 1. Similarly, when the singular polarities are dispensed to each word, at this stage, combine the polarity of a sentence is determined—for example, assuming a particular sentence which has two words of the positive one and negative word by and huge calculated the polarity and utilizing the summation of polarities. For instance, +1 (+2 - 1) is considered a sure sentence because of its positive worth. The result shows that the reporter's remark had a negative-positive and neutral end towards the news and things.

1.5 Methods for Sentiment Analysis

There are different techniques used for sentiment analysis that incorporate dictionary-based strategies, techniques based on machine learning and deep learning, and a combination of these methods as hybrid techniques. More discussion on these methods is followed.

1.5.1 Lexicon-based methods

The lexicon-based approach is an approach that is not supervised or unsupervised learning by creating dictionaries for various domains. This technique relies upon emotional research, labeled with the sentiment and data about every section's subjectivity. These all approaches depend upon lexicons of words and phrases. Also, lexicon-based can be created physically and consequently, using different machine learning techniques [12, 13]. Many researchers have utilized "WordNet" for sentiment analysis [14, 15]. The learning technique is not required as outrageous term inclusion for marked and methodology. Because of the vocabulary, the techniques typically use a rundown of various words connected with the sentiment word. Two automated techniques are used to gather a list of words. First is a Lexicon-based strategy that works with the acknowledged heading of words and employs glancing through the same importance and inverse significance words, i.e., antonyms and synonyms, accessible in straightfor-

wardly open word combinations to improve the variety of the first overview. Until the never new word is searched, this work of this process is going repetition. In sentiment analysis, the score is calculated using the summary of the lexicons' polarity. In sentiment, analysis score is made the subtraction then this situation of lexicon-based with the polarity of negative. The improvement is possible of lexicon dictionary reference as manual assessment like Wordnet, semi-naturally, and VADER like SenticNet, and complete automation based. Regularly utilized, mostly word references are n-grams and unigrams. Additionally, just using unigrams isn't sufficient for Sentiment analysis. For instance, an identical text expression might have inverse familiarization in different conditions. The essential detriment of this stage is the inaccessibility of area-specific dictionaries [16]. This arrangement relies on a methodology based on a corpus of words. A summary of seed words, for example, by using a corpus having the similar domain adjectives are extended. The words that are used together are expected to have similar polarity.

1.5.2 Machine learning methods

Machine learning (ML) techniques are utilized for automation. These ML strategies are divided into two fundamental sorts: i.e., i) supervised (classification) and ii) unsupervised (clustering). These approaches work by training an algorithm before applying it to the testing dataset. Such techniques train an algorithm with known outputs of a particular input [17]. Also, these techniques proposed by [18] the issues of sentiment analysis can be classified into two distinct groups: i.e., i) traditional models and ii) deep learning models. Traditional models represent classical models for example the Support Vector Machine (SVM) [19]and Naïve Bayes (NB) classifier [20] as depicted in Figure 1.2.

The input sources are given to such classifiers incorporate lexical features, lexicon-based features, various adverbs, and different adjectives and parts of speech. These frameworks' accuracy depends on which all features of elements are selected. Deep learning models can give an efficient outcome instead of traditional models. Different Deep Learning models, such as more complex CNNs, ANNs, and time series RNNs,

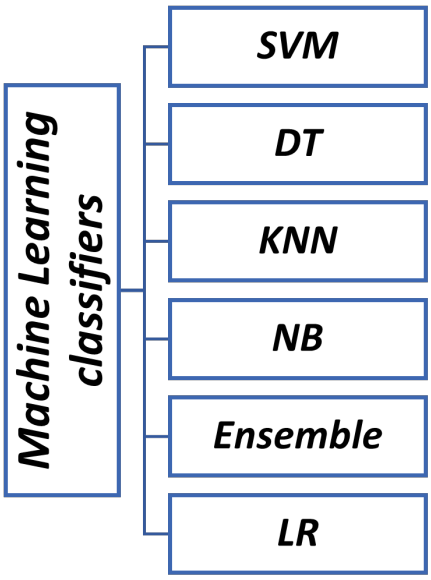

Figure 1.2: Machine learning classifiers

can be used for sentiment analysis. Such techniques at various levels include the classification issues: i) document level, ii) at the aspect level, and iii) sentence level. Machine learning-based elements of these techniques of sentiment for the classification are [15]:

- Frequency of term presence incorporates different n-grams and unigrams alongside their frequencies.
- Information of grammatical form as part of speech was utilized for a feeling of vulnerability, which is used in the choice of element direction.
- Negations have feeling examination words, which can communicate good or pessimistic opinions.

In machine learning techniques, there is no need for a dictionary. They show the highest accuracy of classification. In many scenarios, single-domain training for working is not appropriate with other trained classifier domains. It is essential to make sure the availability of data is marked, which could be costly as a result of the modest relevance of new data.

1.5.3 Hybrid approaches for sentiment analysis

A hybrid approach combines both different Machine learning methods and a lexicon-based approach [21]. This study proposed [22] the primary hybrid framework to suggestion, which takes benefits content composing and joint effort both techniques are filtering. For instance, a content-based strategy utilized might incorporate the recognition of client profiles for the proposal. It relies upon the user's particular interest in various points brought from different web connection or pages while the coordinated effort strategy relies upon the reaction of various clients. Although sentiment analysis is not acted in this work, it could be understanding the precursor of other studies that are merging approaches that acquire particular user reactions for sentiment analysis. In the most recent review, [23] related movies have utilized the merging of coordinated effort strategy and content-based techniques to develop a different initial list of the proposal obtained which presents a hybrid methodology. In a similar domain, offer the various purposes of sentiment analysis classifiers present from the feedback of film as one more strategy after collaboration technique [24].

1.5.4 Deep learning approach

In recent trends, deep learning has set an interest in machine learning. In a traditional neural network (NN), the hypothetical establishments of deep learning are established in a wide range. It additionally allows assessment models, which are made from many layers of processing, to learn defining of data with various levels of reflection [25]. Those approaches have an existing improvement in visual object recognition, object detection, text analysis, speech recognition, and other different domains, for example, genomics and drug [26]. In massive datasets, deep learning develops complex designs by using the algorithm of backpropagation which defines how a machine should convert its inside parameters of hidden layers to calculate error rates which are used at the output layer [27]. Different classifiers are used to detect different sentiments from textual data. These classifiers are defined in Figure 1.3.

A Deep Neural Network (DNN) known as Convolutional Neural Network (CNN) Now,

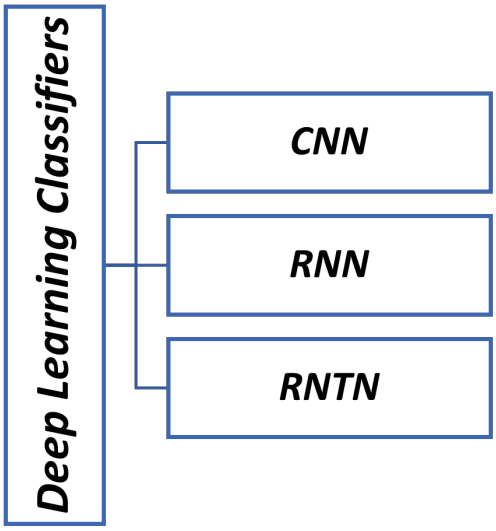

Figure 1.3: Deep learning classifiers

they have recently been utilized for multiple NLP domains [28]. Recently, CNNs have been utilized to analyze textual data and get outstanding results. A standard CNN comprises a fully connected layer, pooling layer, outcome layer, input layer, and convolutional layer [29]. The input is processed, and a vector matrix is utilized for the input layer. Convolutional layer significant functionality that it gets the meaningful information from the features, spatial wise, and then calculates the outputs result of these features, which is known as "feature maps" [30]. Different kernels are used to extract the different information from input features for this purpose. With a reduced size, the convolutional layer holds the previous layer's information and then moves to the next layer. While keeping informative data to decrease the dimension of feature plots, the pooling layer is utilized. Normally, consisting of the end of the network and then converts the output of the pooling layer, which refers to the assemble layer. One-dimensional vector the output layer gives, for example, input and to classify logistic regression used soft-max. Recurrent neural network (RNN) is a widespread technique used in NLP issues. For most literature, RNN is appropriate when its recurrent structure is to process the content size of a sentence. Present in the input data RNN reproduces the sequential data, for example, while creating predictions dependency between all words in the text [31]. It is present in the concept that the input layer saves and takes care of this previous to the contribution to request and

anticipate the output layer. It is good by a methodology that a grouping of varying size recursively has to do with a progress capacity to its inner secret part vector of the input order [32]. In the field of deep learning (DL), Long short-term memory (LSTM) is used as an artificial neural network (RNN) model [33]. LSTM is typically allocated in time and space. Its estimation time complexity per phase and weight is also denoted as O (1). It is a model which extends the whole memory of RNNs. But RNNs have short-term memory; in that case, they utilize previous persistent data to be used in the current neural network layer. RNNs are also utilized for the NLP domain to show better results. Recursive Neural Tensor Networks (RNTN) is a unique neural network that is helpful for the natural language processing (NLP) domain. They have a tree design, and each node of the tree has a neural network. RNTNs are used for colossal partitioning to decide which word categories are negative or positive [34]. In addition, RNTN provides external components such as Word2vec. Word 2 vectors are used as a basis for sequential classification and features. They are merged into the sub-phrases, and the sub expresses are merged into strings or sentences that can be classified by sentiment analysis and other metrics [35]. A neural network word showed such as parameters continuous vector can be shown to determine the utilized text. These words contain not only vectors information about the related word but also the information related to the nearby words, which contains usage, context, and semantic information of each word.

1.6 Research Objectives

In order to perform sentiment analysis of Urdu text, the following research objectives are met. The research objectives are illustrated in Figure 1.4.

- Explored different topics on sentiment analysis of Urdu text specifically and reviewed the approaches used in this area.
- A literature review is performed to deeply explore the existing methodologies by critically analyzing their pros and cons.

Table 1.1: RNN working in different domain

Domain	**Detail Description**
Text generation and language modeling	Models doing work by forecasting, the most preferable the next character and the next sentences or word in the generated text. A model detects the flow of a complete body of the given text. It is understood to write some extra text in a similar way.
Machine translation	A given sentence is embedded in the network. For example, a group from beginning to end because the translation is normally not word to word. In the language, the input is a stream of continuous words in any given language of source and while the desired output is same meaning words in target language.
Speech recognition	The input is in the form of recording analyzed into acoustic or heard signals and model outcomes the very suitable syllable section matching equally every part of the recording.
Time series anomaly detection	The input is serial or sequential data for example the response of a user on network throughout month. Such models can detect which information in the series shows an anomaly behavior as compared to normal events.
Image generating description	The input is an image or picture and the model recognizes major features in the image and generate a text that sketch the image.
Video-tagging	The input is a series of video frames and then generates a model a textual explanation of what is occurring in the video from frame to frame.

- Design and implementation of the proposed work are performed.
- Urdu dataset is obtained to perform classification.
- Data preprocessing is performed for the normalization of data.
- Training, validation, and testing of the data are performed using deep learning approaches.
- Investigation on the results obtained on the classification and error of misclassification are observed.
- Comparison of the proposed approach is conducted with the previous research studies to show the improved results.
- In the end, research is analyzed with limitations, and improvements are recommended on Urdu sentiment analysis in the conclusion and future work section.

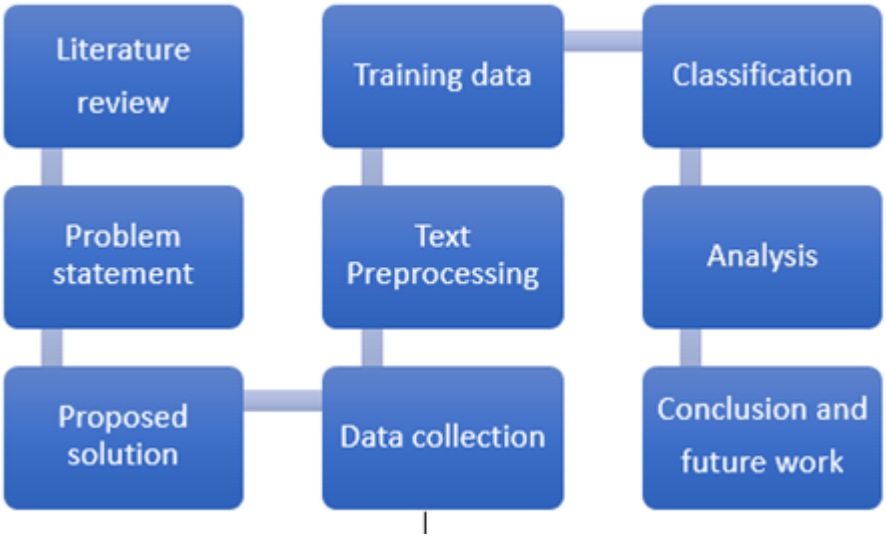

Figure 1.4: Research Objectives

1.7 Research Contribution

Following are the objectives of the research as shown below:

- Preprocessing step is executed by removing stop words, digits, and spaces to eliminate noise from the dataset. Since very few preprocessing resources and tools for the Urdu language are available, a supporting stop word list is made.
- A small-scale Urdu stemming dictionary of 830 words is created for effectively performing stemming steps in preprocessing.
- Framework for Urdu text sentiment analysis by implementing a deep learning classifier of LSTM is proposed.
- Lastly, a comparative analysis of the proposed framework with a related approach is done.

1.8 Scope of the Research

- Sentimental classification is used in multiple domains like business, politics, and communication domains. For example, websites of customer reviews, different search engines, language identification, political debate and discussion forums, spam filter, and detection systems of emails and messages.
- In the business domain context, it can be used in recommender systems, brand reputation, e-commerce, and social media monitoring.

1.9 Thesis Organization

The thesis organization is depicted in Figure 1.5. In Chapter 1, an introduction of the research work is provided, and the rest of the document is organized as follows.
Chapter 2: a literature review is presented, which highlights the detailed review of research being conducted on Urdu sentiment analysis.
Chapter 3: design and implementation of the research work are provided, and it also provides the experimental implementation of the research work.
Chapter 4: provides the validation of the research. The discussion about results achieved by this research is discussed in the chapter. For better understanding, desired

tables and figures are provided in this thesis.

Chapter 5: conclusion is drawn with future directions.

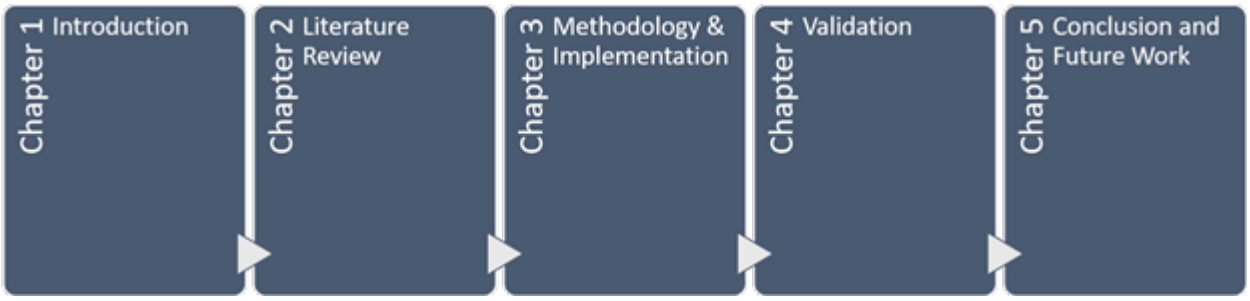

Figure 1.5: Thesis Organization

Chapter 2

Literature Review

One of the most active areas of computer science is Sentiment analysis (SA). Generally, SA is defined as opinion mining. Opinion mining of SA is a text-oriented technique that addresses extracting, detecting, and analyzing the text, determining the positive and negative opinions, and determining how an entity (people, product, organization, etc.) is considered positive or negative. Sentiment analysis examines the feelings of people, their attitude, opinions, attitudes, appraisals, evaluations, and emotions regarding anything like products, services, individuals, organisations, issues, and themes, as well as their characteristics [36]. Microblogging websites and social networks have seen significant growth. Micro-blogging websites are one of the most critical web objectives for the users, according to the different contexts that are very helpful for the user's point of view, which includes opinions and attitudes [37, 38].

The social networking services primarily utilized micro-blogging, Twitter, which gives enormous data. Researchers recently employed social media data for sentiment analysis of people's perceptions on the event context. Furthermore, "opinion mining" is another term for sentiment analysis, which is important in the NLP sector. Introduce and explain the text-related sentiment analysis, such as neutral, positive, and negative [39, 40]. However, computational linguistics represents the sentiment analysis for recognizing data and categorizing the user's thinking through NLP and text analytics implementations. Usually, according to the similar context of the specific contextual document's file polarity, the primary purpose of the sentiment analysis is to define the

researcher's opinion. An assessment or judgment of the user's point of view can be, deliberated communication of emotion and affective condition. Recently, sentiment analysis is becoming the main research domain in demand under Natural Language Processing (NLP). English language resources are filled with sentiment analysis. It also includes a large number of NLP, part-of-speech, and lexicons [41].

Sentiment analysis in an individual language improves the probability of necessary information in the form of textual data of different languages being left [42]. The national language of Pakistan is Urdu which is usually spoken. The Urdu language is also used in India and its different areas. The Urdu language faces different problems in Sentiment analysis. Resources are not lack acknowledged of Urdu lexical [43–45]. Due to this shortage, for the most part, includes Urdu sentiment analysis, the converting of data from the English language overflowing with the resources to the Urdu language that is needed in these resources [46, 47]. Similarly, Urdu websites were organized other than an appropriate encoding of the text plan in an exemplified format.

The current situation faces hurdles when trying to design a structure of corpus that is definitely readable for a machine. Sentiment analysis lexicon in any language is necessary for making the framework of an opinion examination. The English language is a complete resource along with a considerable vocabularies size of opinion (like Senti WordNet) that is well organized. However, in sentiment lexicons, the Urdu language has insufficient resources. As a result, many researchers focusing on Urdu sentiment analysis have been rare. The lack of linguistic resources and language developer elements can be set down to the absence of awareness.Previous research has focused on the Urdu language, which emphasises several aspects of language processing [41, 47]. This includes Urdu language morphology, along with Named Entity Recognition (NER), stop words identification, and datasets concept stemming and searching. Moreover, the Urdu language has extraordinary features, and therefore a few basic steps are needed for the process of sentiment analysis. Such as, from right-to-left, Urdu language script is composed.

Table 2.1: Stages of Sentiment Analysis

S.N	Stages of Analysis	Opinion Discovery Type	Type of data	Characteristic of the data
1	Document level	Positive, Negative	Document as a single entity	Everything
2	Sentence level	Positive, Negative, Neutral	Sentences	Views or Opinion
3	Aspect level	Opinion itself	Documents or sentences	Opinions

2.1 Different levels of analysis

Remarkable research is done on a word or phrase level of sentiment mining with different levels such as document level, user-level sentiment analysis, and sentence level. First, the sentence level is inspected to find the direction of words to determine the opinion mining. Then, the sentences are analyzed to determine sentence-level opinion mining. The result shows sentiment classification as negative, neutral, and positive. While at the document level, the document is analyzed to find out the opinion of the whole document. In user-level sentiment, the comments of users on social media are analyzed to find out the user has the same opinion on the topics [48]. These stages are defined in Table 2.1.

2.1.1 Sentiment Analysis at documents level

Whole document polarity is figured out by analyzing that whole document. Reviews of one product can be held in one text file. The system estimates the score of negative and positive comments for the product. Hence determine the opinion about the product. The opinion about a single product is determined using the document-level sentiment. The purpose and advantage of opinion mining at document are to find that overall opinion is getting about one entity. On the other hand, the main drawback is that we cannot find out opinions about different entities [49].

2.1.2 Sentiment Analysis at Sentence level

In sentence level of sentiment analysis, the sentence is analyzed to find whether the sentence shows a positive opinion, neutral or negative. The Neutral opinion is determined as having no meaning. Sentiment analysis at the sentence level is closely related to the classification of subject and entity in the sentence. Subjective sentences express information related to the subjective point of view. Therefore, the primary mission of sentence-level opinion mining is to determine the subjective and objective classification of the sentence [50].

2.1.3 Sentiment analysis at entity and aspect Level

The object-level sentient mining discovers the actual opinion of the people about the product. The entity-level had done a perfect Analysis of the people's points of view. However, the sentence and document level analyses do not analyze what people actually say about a product. The aspect level analyzes the opinion rather than the construction of the sentence. The aspect level consists of the target opinion. Hence, it is considered that the opinion target helps to understand the problem related to the sentiment. Consider as an example, "Despite the poor service, I still enjoy this restaurant". has positive sentiment, but this is not entirely positive. The sentence is negative about the restaurant's service but gives positive vibes about the restaurant. The central theme of this analysis is to find out the entity-level sentiments on the products. Take one more example "The phone's quality is good, but its battery life is short." The analysis of sentiment explains the negative and positive points of the phone entity. Two aspects of the entity are evaluated: quality of call and battery life. Call quality shows positive sentiment, whereas battery life shows negative sentiment. Thus, entity-based sentiment converts unstructured text into structured ones. The obtained information is used to analyze quantitative and qualitative attributes [51].

2.2 Sentiment analysis Approaches

The way of finding out the classification and categorization of texts and opinions as negative or positive is known as sentiment analysis and opinion mining. However, labeling sentiment words manually is considered a time-consuming process. Therefore, two popular techniques are used to obtain the sentiment analysis automatically.

- Weight words are used as a lexicon.
- Machine learning approach to automate sentiment lexicon.

In the lexical-based approaches, a list of words or corpus is used to determine the polarity of the words. On the other hand, the machine learning approach determines the sentiment by training the data set using an algorithm and done classification using features, and done testing to check accuracy [52].

2.3 Classification of Sentiment Analysis Approaches

Similarly, two fundamental classifications techniques are used in sentiment analysis: automatic detection and manual detection. A detailed diagram of the sentiment analysis approaches is depicted in Figure 2.1.

Traditional machine learning (TML) methods and Deep Learning (DL) methods are two types of machine learning approaches. Many TML algorithms use either supervised or unsupervised methodologies. DL, on the other hand, is rapidly expanding. It's also a popular topic of Machine Learning (ML), which demonstrates these approaches for encoding learnable characteristics in a progressive system in a supervised or unsupervised manner. For example, in Deep learning-based sentiment analysis techniques, DL displayed outstanding performance in the multi-layers representation of automatic features [53]. Moreover, DL performs the best result in the understood semantic removals of components that would help across various domains.

An application for deep learning algorithms in the sentiment analysis area has decreased the need for human intervention, feature engineering, and time process cal-

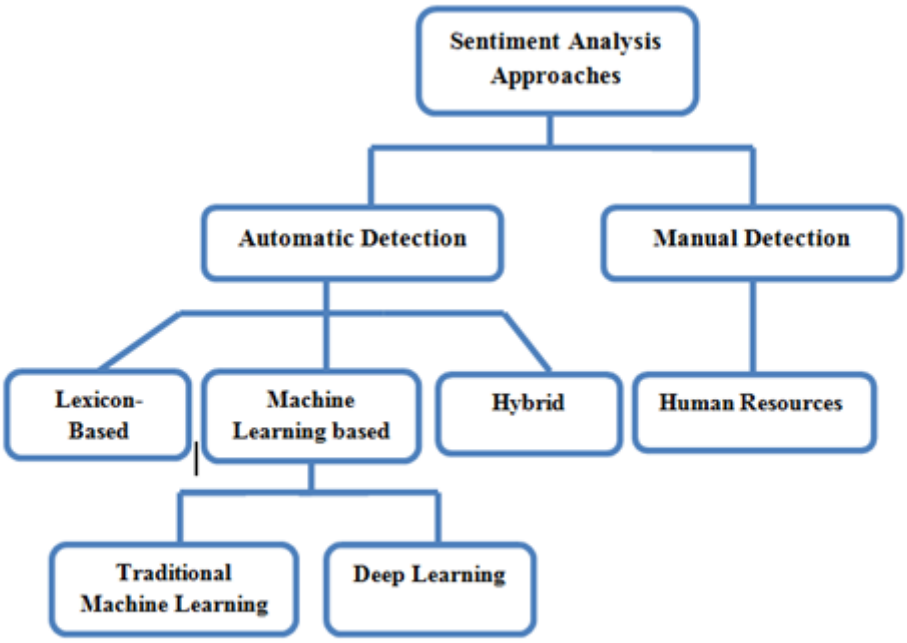

Figure 2.1: Classification of sentiment analysis approaches

culations [54]. DL techniques that are most widely recognized are based on simple ANNs, such as the time-dependent Recurrent Neural Networks (RNNs) and the spatial-based complex Convolutional Neural Networks (CNNs). As the text-based information is coming promptly, the well-known variants of RNNs are Gated Recurrent Unit (GRU), Bidirectional long short-term memory (BiLSTM), and long short-term memory (LSTM). Each DL technique is used, and model RNN is clearly the most often used procedure in sentiment analysis. Its sequential technique is best suited to the concept of text data and the capacity to manage varied input and output sizes.

2.3.1 Manual Inspection

Sentiment analysis from text-based data on social networks can be done utilizing manual inspections, but it is a challenging task. The reason is that many resources to inspect humans are required, and it is not possible to capture each individual opinion of a million users. However, a human can capture different emotions from the text,

but not machines do well; it is also a time-consuming task, and the results may be based on the nature of the inspector.

2.3.2 Automatic Detection

There are several automatic or machine-based detection of sentiment analysis methods, including Lexicon-based methods and machine/deep learning-based methods.

Lexicon level detection

The lexicon-based approach utilizes a dictionary of different words with opinion words associated. This method matches words with the dictionary to find out the scoring value. The lexicon technique does not require any dataset to be trained. While in the machine learning method, data is preprocessed first and then trained. The sentiment of the sentence and document is determined by combining the polarities of the single word present in the document. Combining the score delivers the sentiment of the whole sentence and document. This approach utilizes a predefined list of words, and each word has a sentiment in the list. Different Lexicon-based approaches are depicted in Figure 2.2.

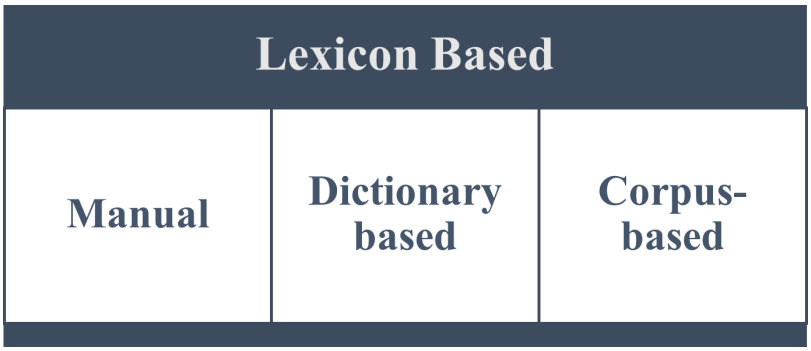

Figure 2.2: Three categories for lexicon based

1. **Dictionary Based methodology**

   Positive opinion words and negative opinion words are identified by utilizing lexicon-based dictionaries such as Wordnet dictionaries. Dictionary of sentiment analysis can be divided into 3 ways: manual approach, dictionary-based

approach, and corpus-based. However, the work is hard, so that's why the manual approach [55, 56]is solid. "General Inquirer," which is developed manually [57] for the English language, is a significant resource for sentiment analysis. In dictionary-based approach which is automatically extended sentiment dictionary from a seed of words [58–62]. Generally, for the development of words, two methods are used, which are Wordnet [63], and Senti WordNet [64].

2. **Corpus Based methodology**

   A large number of words are collected in a corpus. Synthetic pattern-based words are collected, and words related to other opinions are found within the context. Corpus-based a seed of words from sentiment lexicon by utilizing a corpus then it is increased [65–67]. The methodology of Lexicon-based sentiment analysis is shown in Figure 2.3.

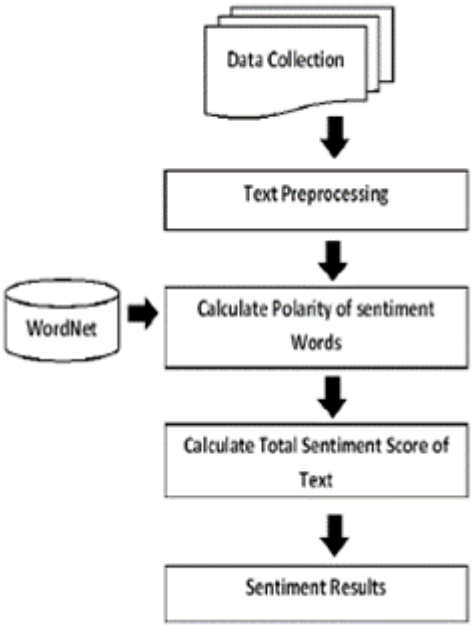

Figure 2.3: Lexicon based sentiment analysis

These techniques rely on lexicons expressions or phrases, and these are also pre-labeled with sentiment analysis, information according to each entry based on subjectivity or polarity. By utilizing machine learning (ML) techniques lexicon can be created manually. For example, data is generated as automatically. However, few researchers have

utilized (WordNet) for the analysis of sentiment [21]. Each word (to be examined) in the text is compared with a lexicon entry. Therefore, as a result, it is related to positive or negative scores. In order to complete the sentence score, it submitted the result as scores of several words. Let us consider an example of a sentence, "A quantity range camera." In this sentence, "quantity range" is the only consisting (semantic unit) of sentiment analysis. Therefore, the excess words in the given sentence are unbiased; henceforth, this sentence is considered a not negative one. Many researchers have developed an Urdu lexicon. Lexicon-based Roman-Urdu Adjective is proposed by [68]. In this research, they proposed architecture to search the polarity of the sentiment using the Natural Language Processing technique (NLP). Instead of that, they developed a lexicon feature that includes adjectives as a negative or positive one. This study also used three types of a dataset that performed 1620 comments, and also performs; as a result, comments were wrong classified 21.1%, with an accuracy of 78.9%. A Lexicon-based work proposed by [69] has a total number of 7335 sentences which includes 4728 as a positive and 2607 as a negative entry. The absolute polarity is the total amount of all the related weight terms of a sentence. In addition, the dataset includes Urdu-related comments from various Urdu websites to check the framework's effectiveness which the author proposes. The result of this study is that the framework performs an accuracy at 66%, and the dataset depends on text from many news websites. It can be attention to handling the several linguistic glitches and linguistic, with the Urdu electronic text, which can expand the lexicon. One more research work proposed by [24] is an Urdu language Sentiment analysis Lexicon for adjectives, nouns, and verbs along with the negative intensifiers and context-dependent 9578 positive words and 11,739 negative words. As a result, it gained an accuracy of 89.03%, including a dataset of 6025 sentences with 151 blogs. The different authors have made a vast range of Urdu lexicon sentiment analyses, including adjectives, verbs, and nouns. The term negations in Urdu language, intensifiers, and conjunctions are most necessary for the lexicon. The lexicon-based (for example, sentiment analysis in Urdu) for the Urdu websites developed rather than a supervised machine learning equivalent. Accuracy of this framework is 89.03%, precision 0.86, recall 0.90 and 0.88 F-measure. It is huge because the whole literature work is done in Urdu language sentiment analysis.

The biggest accuracy increased by 73.8-89.03% according to handling with the context, negations, and intensifiers, which is totally dependent on the words effectively and which express those 3 types are utilized in the Urdu text/message often and then should not be ignored while acting sentiment analysis text of Urdu language. The Lexicon-based approach functions admirably rather than the SML approach because of an extreme inclusion of the dictionary. It is a time-consuming process, but if data is enough to be gathered in each domain, then a productively SML approach can be managed with the information from different areas [22].

Machine Learning based Detection

This method involves training an algorithm with a training data set before applying it to real-world data for testing and outcomes. The algorithm can be better trained for the testing data set if the inputs and outputs of training data are known before the results. Support vector machines, N-gram sentiment analysis, Nave Bayes, and K-means are examples of popular machine learning approaches. Unsupervised (without labels) and supervised (with ground labels or information) ML techniques are two types of techniques. The more details are given below.

1. **Supervised Learning approach**

   Supervised classification is the most common machine learning strategy used in sentiment analysis. Training, testing, and validation are all phases of supervised categorization. We can get correct results when annotated data is used for training, unannotated data is used for testing, and a classifier is trained on the training data. With aspect-oriented explanation and justification, it delivers more structured and comprehensible outputs than pure learning-based systems. A sample supervised machine learning model is depicted in Figure 2.4.

2. **Unsupervised Learning approach**

   Unsupervised models are more sophisticated than supervised Learning for unannotated and unlabeled datasets. Unsupervised Learning aims to uncover hidden patterns in data that have not been labeled. As a result, no prior training is

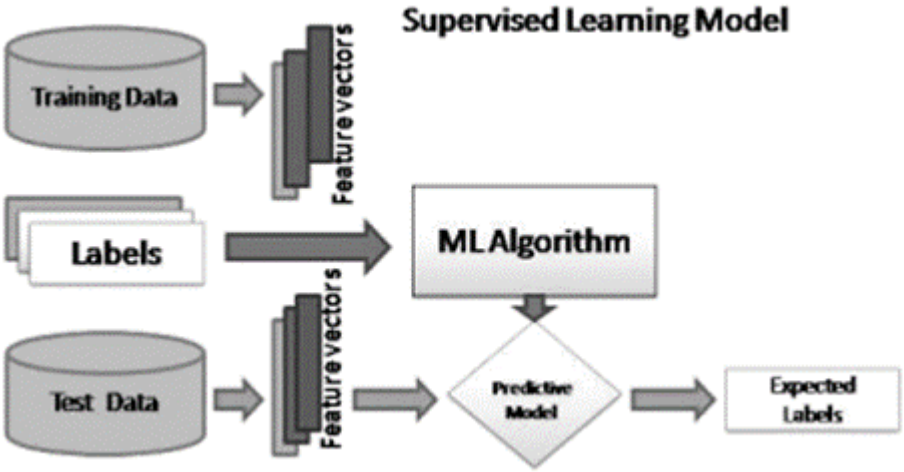

Figure 2.4: Supervised machine learning model [1]

required to examine the data. It was once common to group data into categories solely based on their statistical features. Common examples include clustering algorithms, k-means, matrix factorization, principal component analysis, and many others.

Machine Learning techniques depend on vectors feature that is selected by domain. A labeled corpus usually trains a single classifier in ML. In this technique, feature selection is a vital issue that extraordinarily influences the main results. For example, Urdu blogs of sentiment analysis, in this research work [40], that is based and focused on the sentence level. Regarding this, different steps are selected. The information is collected from many internet-based online blogs in the first step. In the second step, every sentence from text in online blogs is defined as negative, positive, and neutral through the annotators. In the last third step, the order of sentences (as negative or positive) is completed by using different classifiers, which consist of Decision Tree (DT) of J48 type, SVM, and KNN [22]. The result declared in this research work has gained a precision of 76%, an accuracy of 67%, a recall of 73%, and the f-measure reached 73%. For this scenario, the dataset is collected and extended from both quality type data and quantity type. Dataset is taken from various 151 sites utilizing 6025 sentences with 14 different sorts is examined. Therefore, there is a big deal of variety in the messages, and each type is enough to address. The primary focus of another research is machine

learning and differentiating between the main classifier in Sentence level, which Urdu sites utilize for sentiment [70]. The achievement of this stage is made possible through different steps which are expressed previously. Additionally, the annotator clarified sentences 2753 as a negative result,1388 sentences as a neutral outcome, and sentences 1876 as a positive. Because of contention between the annotators, eight sentences are disposed of 3 classes, which characterize all sentences, including neutral, negative, and positive. It is completed by utilizing the Weka device, which comprises machine learning algorithms. For example, five different classifiers are utilized for this purpose Naïve Bayes (NB), DT (J48), KNN, Lib SVM, and PART. Compiling the results after 10-fold cross, a rundown of the top three classifiers is chosen because of better execution regarding the accuracy, recall, precision, and f-measure. There are three classifiers selected, including KNN, J48, and SVM. Additionally, the KNN performs better than J48 and KNN in all the performance metrics. To accept the result of predictions (sentences will be marked as negative, neutral, and positive) is done by these 3 classifiers with two different training datasets, each dataset consisting of 1800 sentences and six different test datasets with 180 sentences. By utilizing, 3 evaluation measures and also these predictions are calculated. This whole evaluation measures McNamar's test, and Root Mean Squared Error (RMSE), and Kappa measurement. This study's accuracy was calculated at 87%. Urdu Text-based Word Sense Disambiguation (WSD) utilizes 3 different classifiers proposed by [71] those classifiers known as SVM, NB, and DT. This study has used the dataset from international and national news websites and performed the result f-measure of 0.71%. The system's outcome could be enhanced by utilizing an adjustable window size for unclear Urdu words. To obtain the result of learning about Urdu and English language tweets, different sentiment analyzers with the algorithms of machine learning [72] focus on the best accuracy. To calculate in a text, a lexicon-based in sentiment analysis, semantic orientation is focused on words, phrases, or sentences. In this research, the main contribution is to sentiment analysis in the Twitter dataset using the RNN-LSTM model and other machine algorithms, including KNN, DT, and SVM. Instead of all other algorithms, RNN-LSTM conducts efficient results and has the highest accuracy. For example, the result produces 92% for the English twitter data set 87 and for the Urdu Twitter data set. [1] performed

sentiment analysis on the Urdu language. They extracted their Urdu dataset from 151 Urdu blogs. The researchers have annotated their dataset into positive, negative, and neutral classes and performed preprocessing step by stop words removal. Then, they performed supervised classification using three classifiers of KNN, SVM, and decision tree in WEKA using 10-fold cross-validation. They achieved 67.01% accuracy by KNN. [73] proposed a sentiment analysis system for Roman Urdu on the dataset of 779 reviews. For this task, they used unigram, bigram, and uni-bigram features and used five diverse classification algorithms to determine accuracy. They concluded that Naïve Bayes (NB) and Logistic Regression (LR) gave better results than others after 36 experiments. Results were enhanced after feature reduction. [74] gave discourse-based sentiment analysis on Roman Urdu datasets. They collected a larger Roman Urdu data corpus from social media websites. Then performed normalization, POS tagging, and tokenization to clearly identify the discourse element. After taking discourse into consideration, the system is ready to use for neural network-based sentiment analysis for future work. Sana et al. [75] developed Emotion Detection System in the Urdu language from online business tweets using supervised machine learning approaches. They applied different classifiers of ML methods, such as a basic Naïve Bayes (NB), an old Random Forest (RF), a simple K-Nearest Neighbors (KNN), and a complex Support Vector Classifier (SVC), to classify tweets of Urdu emotions. They showed that SVC gave efficient results, giving 81.09 for the sports dataset and 80.05 for the smartphone dataset. [76] paper focused on Urdu Roman reviews of Daraz.pk website. Different experts have annotated 20.286 K total reviews into 3 classes. The bag of word model is used for feature extraction and passed to SVM for sentiment classification. MATLAB Linux server is used for experimentation, and the dataset is public.

Deep Learning Models for the Detection

Deep Learning is a comparatively novel approach that has been used to perform Urdu sentiment analysis. This is the advanced stage of machine learning, in which neural networks are primarily used for data learning and prediction. Deep Learning has been found to be better than machine learning, or lexical-based approaches, especially for

a considerable amount of training data [77]. After the potential of deep learning was identified, investigations revealed that when employed in sequential data, linguistic analysis, word recognition, and word prediction, Recurrent Neural Networks (RNN) produced satisfactory results [78]. As a result, RNNs have been extensively used for the retrieval of subsequent data.However, because to the well-known gradient exploding and vanishing difficulties, as well as difficult-to-read long-term patterns, these RNNs are typically difficult to train. To address these issues, the long short-term memory (LSTM) was created. The LSTM is the neural network's recurrent architecture, which displays the most recent results for sequential data. LSTM is mostly used to calculate the longer-term dependencies between text data. Figure 2.5 shows a sample deep learning network architecture with a two-unit output layer, three-unit input layer, and a couple five-unit hidden layers as an example.

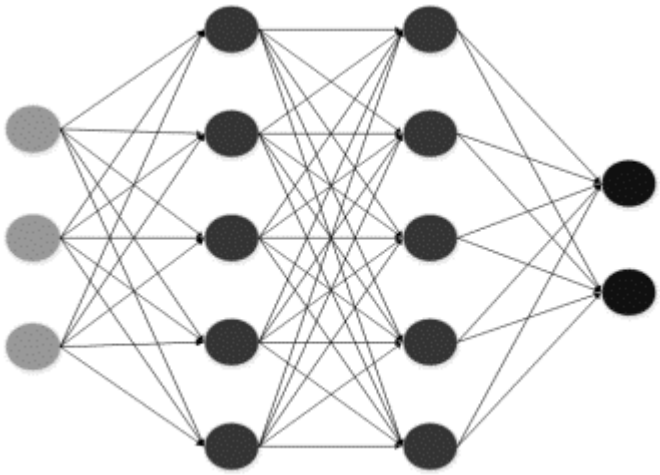

Figure 2.5: Deep Network architecture [1]

The Deep Learning (DL) model is developed for sentiments analysis used for under-resourced language. On these topics, they developed an open-source corpus of about 10,008 reviews and 566 internet-based strings, for example, entertainment, food, sports, politics, and software. The main contribution of this research [79] is defined, firstly for the research work of Urdu sentiment analysis to create a human-defined corpus and secondly, utilizing a measurement of corpus and model performance up-to-date. Finally, for evaluation results, they performed their work in the classification of binary (two) and ternary (three), which used other various models, that are: N-grams,

tiem dependent LSTMs, Rule-based LSTM, complex ML based SVMs, spatial dependent CNNs, hybrid RCNN. RCNN model shows outstanding performance instead of other models for the binary classification with the accuracy of 84.98% and 68.56% accuracy for the ternary classification. To work with many analysts working in a similar domain, we have publicly displayed the corpus and code created for research. Urdu Text Sentiment Analysis (UTSA) proposed by [9] they investigate deep learning (DL) models which gather different representations of word to vector. The resulting outcome of deep learning (DL) models such as CNN-LSTM, Convolutional Neural Networks (CNN), attention-based Bi-LSTM, and Long Short-Term Memory (LSTM) is working for sentiment analysis. Stacked layers in the sequential model are applied as, C-LSTM, BiLSTM-ATT, and LSTM. Various filters are used with the individual convolution layer in the CNN model. In addition, the pre-trained and self-trained embedding models of unsupervised Learning are explored in the sentiment analysis domain for classification. The complete results show the BiLSTM-ATT outperformance other than deep learning (DL) models by obtaining an accuracy of 77.9% and an F1 score of 72.7%. Manzoor et al. [80] proposed a novel approach of "Self-attention Bidirectional LSTM (SA-BiLSTM)" to deal with varying patterns of text representation. It addressed the unidirectional nature of traditional architecture. In SABiLSTM, Self-Attention deals with the sophisticated formation by comparing the complete sentence. BiLSTM handles the lexical variation of appearing embedding in prior and posterior directions by extracting context representations. For improved performance of the proposed model, they performed preprocessing and normalized the Roman Urdu dataset and gave accuracies of 68.4% and 69.3% with the deep learning approach of SABiLSTM. The fundamental objective of [81] is twofold: (1) to create a standardized dataset for Urdu and (2) to utilize and assess different machine learning and deep learning classifiers. They compared two techniques of text representation: fastText pre-trained word embeddings for Urdu and count-based using n-gram feature vector. They used the following two models: i) DL models, including LSTM and 1D-CNN, and ii) supervised machine learning (NB, RF, SVM, LR, and MLP) for Urdu sentiment analysis. They concluded that the combination of n-gram features with LR surpassed others with an accuracy of 81.94%. [82]research has two aims a) the formation of a

human-annotated corpus (b) assessment of performance using a corpus. For this purpose, they executed binary and ternary classification using LSTM, Rule-based CNN, SVM, and recurrent convolutional neural network (RCNN). The RCNN model outperformed others, giving 68.56% accuracy for ternary classification and 84.98% accuracy for binary classification. Their code and corpus are publicly available.

Hybrid approaches

Machine learning and lexicon-based approaches are combined in Hybrid Techniques. This combination increases classification performance, according to researchers [83].They introduce pSenti, a concept-level sentiment analysis system that combines lexicon- and learning-based techniques. The fundamental benefit of their hybrid technique, which is built on a lexicon/learning symbiosis, is that it offers the best of both worlds: stability and readability from a well-designed lexicon and high accuracy from a strong supervised learning algorithm.

Table 2.2: Summary of existing Literature on Urdu sentiment analysis

Author	Year	Task	Model	Polarity	Data	Acc
[10]	2010	SA	Classification	Pos/Neg	Urdu	72%
[84]	2011	SA	Classification	Pos/Neg/ Neutral	Urdu	86.80%
[85]	2011	Opinion Mining	SVM	N/A	Urdu	69.81%
[86]	2012	SA	SVM	Pos/Neg	Urdu	62.12%
[24]	2014	Expressions Detection	Classification	Pos/Neg	Urdu	82.50%
[87]	2016	NEWS Corpus for Salience Analysis	Heuristic Approach	Pos/Neg	Urdu	84.5%.
[88]	2017	Opinion Mining	Classification	Pos/Neg	Urdu	50-52%
[46]	2018	Data Cleaning	Lib SVM, DT, KNN	Pos/Neg/ Neutral	Urdu	67%
[73].	2018	NA	NB, SVM, KNN, DT	Pos/Neg	Roman Urdu	67.58%
[74]	2018	Tokenization, POS Tagging	N/A	Pos/Neg	Roman Urdu	80%
[75]	2019	NA	SVC, RF, NB, KNN	NA	Urdu	81.09%
[76]	2019	NA	SVM	Pos/Neg/ Neutral	Roman Urdu	60%
[89]	2019	SA	LSTM	Pos/Neg	Roman Urdu	90%
[90]	2020	SA	RCNN	Pos/Neg/ Neutral	Roman Urdu	71.30%
[80]	2020	Normalization	CNN, LSTM	Pos/Neg	Roman Urdu	69.30%
[81]	2021	SAs	ML, DL	Pos/Neg	Urdu	81.94%.

Chapter 3

Proposed Methodology

In this section, the proposed methodology has been discussed in detail. The abstract level of the proposed methodology is shown in Figure 3.1.

Our framework has two stages. The first stage is preprocessing of text and the second step is a classification and categorization of dataset. In the initial step, the dataset is preprocessed. In the second step, the processed data is given to the classification algo-

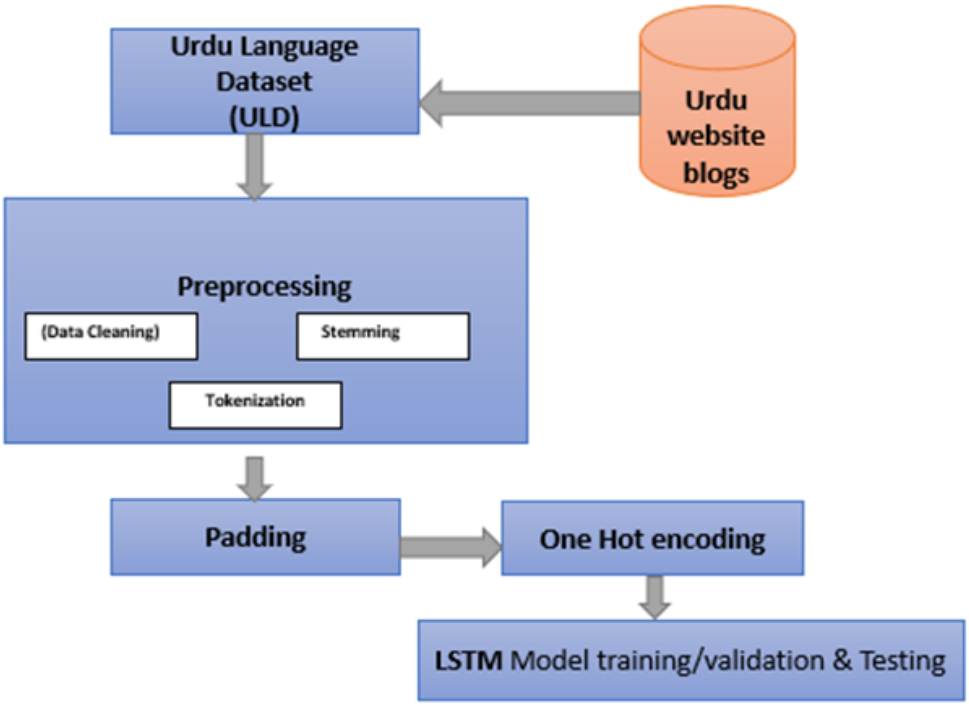

Figure 3.1: Proposed methodology

rithm for sentiment classification. The preprocessing of tweets includes data cleaning, stemming, and tokenization. In data cleaning, special characters, digits, stop words, white spaces are removed. In the second step, stemming is done and for that purpose, a stemming dictionary is manually created. Thirdly, tokenization is performed in preprocessing. For the classification of preprocessed data, deep learning is performed by LSTM. Firstly, padding is done to equalize the length of the vector and then one hot encoding is performed. Then classification is performed by LSTM and results are obtained.

3.1 Dataset Analysis

We have used a dataset comprising of 3995 sentences in the Urdu language. Dataset is split into 70:15:15 between training, testing, and validation. Dataset distribution is depicted in Figure 3.2.

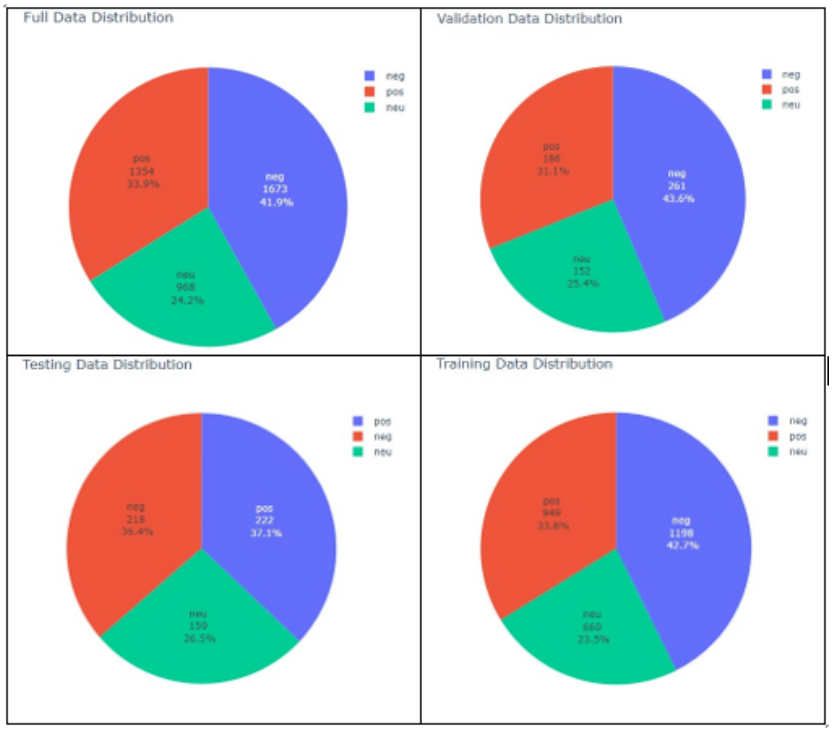

Figure 3.2: Dataset Distribution

3.2 Dataset Preprocessing

3.2.1 Data Cleaning

The dataset used in this research is in the Urdu Language. This dataset has around 4000 rows and three labels, i.e., Positive, Negative, and Neutral. This dataset has noise in different forms. These noises are removed by following steps.

1. **Removal of special characters** The special characters like brackets, quotations, commas, and different signs usually have no information and are considered noise in natural language processing. These special characters are removed from the dataset by finding each special character in all the words of all sentences.

2. **Removal of digits** Like special characters, digits also have no meaning in sentiment analysis. Digits are also removed from the dataset by finding out digits.

3. **Removal of stop words from the manually generated list** Stop words like "The, To, For," etc., contain zero information in natural language processing, and these stop words should be removed from the data. The issue with the Urdu Language is no stop words list is available. In this research, an Urdu stop words list is created first, and then all the stop words in the data are removed.

4. **Removal of white spaces** After removing irrelevant information from the dataset, some white spaces were left, which were considered words in some cases. Therefore, these white spaces are also removed from the dataset.

3.2.2 Stemming

1. **Manual generation of stemming dictionary**

   Stemming is one of the most critical steps in natural language processing. There are many forms of a single word like different forms of word 'Play' are; 'Playing, Played, Plays, Playable.' All these forms are considered a new word for a computer, but all the words have the same meaning. An issue with the Urdu

Language is that there are many forms of a single word and no stemming dictionary is available that maps the different forms of a word to its root or stem word. In this research, a small-scale stemming dictionary of 830 words is manually created.

2. **Mapping of all relevant words to their stem word**

   After creating the stemming dictionary for the Urdu Language, an algorithm is designed for searching, comparing, and mapping all the different forms of words to their stem or root word based on the stemming dictionary. An overview of stemming dictionary is shown in Figure 3.3.

812	کالونی	کالونیوں	کالونیاں		
813	نشان	نشانیاں	نشانیوں	نشانی	
814	سرگرمی	سرگرمیاں	سرگرمیوں		
815	گرمی	گرمیاں	گرمیوں	گرم	
816	کاروائی	کاروائیاں	کاروائیوں		
817	چاہت	چاہتیں	چاہتوں		
818	مل	ملائے	ملنا	ملوں	
819	اٹھا	اٹھائیں	اٹھانا		
820	کمپنی	کمپنیاں	کمپنیوں		
821	جل	جلے	جلنا		
822	دھو	دھونا	دھونے	دھوتے	دھوئے
823	کھل	کھلنے	کھلنا		
824	کلی	کلیاں	کلیوں		
825	لی	لیئے	لینا		
826	پھیل	پھیلنا	پھیلنے		
827	بوڑھا	بوڑھوں	بوڑھے		
828	گر	گرے	گرنا		
829	کاٹ	کاٹنے	کاٹے	کاٹنا	
830	اڑ	اڑنے	اڑنا		

Figure 3.3: Stemming dictionary overview

3.2.3 Tokenization

Tokenization of words is a procedure of breaking strings into tokens which in turn are small units that can be used for tokenization. In this research, our data has gone through the process of tokenization. The text has been assigned to different tokens. Firstly, indexing is performed and then replacement is done with the respective indices'

numbers. For example, the word "Khelna" has given an index number of 37 so it has been replaced with 37.

Figure 3.4: Tokenization

3.3 Padding

Padding is used to equalize the lengths of vectors. In our case, some sentences have many words while some have fewer words, leading to the unequal size of vectors. Zeros are padded to equalize the length of all the sentences. Also, input to a machine learning model should be of the same lengths.

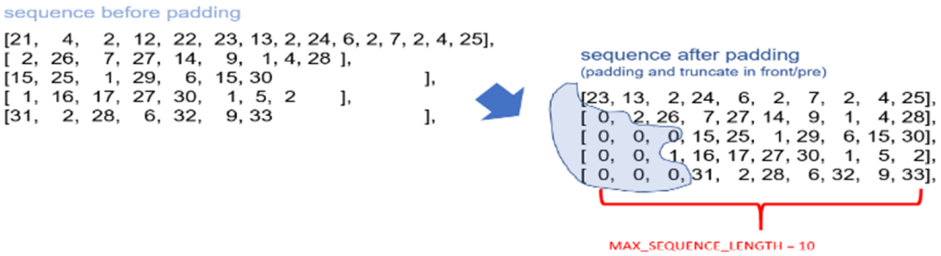

Figure 3.5: Padding

3.4 One-Hot Encoding

As mentioned before, a computer understands digits and binary numbers only. We have converted our labels to one-hot form. It means the label "Positive" is converted to 001, "Negative" is converted to 010, and "Neutral" is converted to 100.

Index	Label
0	positive
1	negative
2	neutral

index	Positive	Negative	Neutral
0	1	0	0
1	0	1	0
2	0	0	1

Figure 3.6: One-hot encoding

3.5 GloVe Embedding

GloVe Embedding has been used for vector representation of words. These are the pretrained weights that are trained on a well-known English dataset containing different English words that are repeatedly used in English vocabulary. The gloVe is used, since it does not depend just on local statistics but incorporates statistics on a global level. 200D glove embedding is selected because the vector size of input data is around 200.

3.6 Classification

The dataset used in this thesis contain textual data. The available classification deep learning models for textual or sequential data is Recurrent Neural Network (RNN). These models are useful for processing of time-series and sequential data. However, RNNs are quite difficult to train and fit because of vanishing gradient along with the exploding issues. This issue is overcome by another advanced version of RNN, which is known as Long Short-Term Memory (LSTM). The advanced RNN version is also a recurrent model which contain memory cell and forget gate along with the input and output gates. The functionalities of these four gates are:

- Forget Gate: The functionality of this gate is to forget less useful or infrequent information inside Long Term Memory (LTM).
- Learn Gate: A current input (an event) and Short Term Memory (STM) are merged together for remembering the recently learned information from STM

and applied to that event.

- Remember Gate: As the name suggests, the main functionality of this gate is to remember the previous information up to certain limits. LTM information which are not forgotten. An event and STM which are merged together in this gate works as updated version of LTM.
- Use Gate: Functionality of Use Gate is to predict the current event's output by involving the usage of STM, LTM and an Event.

The detailed working of an LSTM unit/model is further divided into three main steps.

3.6.1 Step 1

Initially, an LSTM unit identifies some irrelevant information used in the data and disappear such irrelevant information from the cell. A well-known activation function, *sigmoid*, is used for the identification followed by elimination. This is done by getting the final LSTM unit, $h_t - 1$, output at $t - 1$ time for the available X_t input at t time. The functionality of *sigmoid* is to simplify the part which needs to be removed from the old output. Another thing that needs to be noted down that the output should be between 0 and 1 for any $C_t - 1$ cell stage, that is stored in f_t vector. Final decision is takend by the *sigmoid* function that which data or information should be discarded or remains kept, depending on the output. The formula explains the step 1 operation is given by;

$$f_t = \sigma(Weight_f[h_t - 1; X_t] + bias_f) \tag{3.6.1}$$

σ is used for the functionality of *sigmoid* functionality, while $Weight_f$ represents weighted matrices and $bias_f$ the respective biases, corresponding to forget state.

3.6.2 Step 2

After step 1, X_t, the new input, is stored as well as the cell state is updated. Two actions are executed: one on sigmoid layer while other one on *tanh* layer. *Sigmoid*

layer takes a decision which information needs to be updated or discarded while *tanh* layer assigns weights to the passing values. Then to update the cell state, these values are multiplied and then new memory Y_t is added to old memory $Y_t - 1$.

$$i_t = \sigma(Wi[h_t - 1; X_t] + b_i) \tag{3.6.2}$$

$$N_t = tanh(Wn[ht - 1; X_t] + b_n) \tag{3.6.3}$$

$$Y_t = Y_t - 1 * ft + N_t * i_t \tag{3.6.4}$$

where Y_t and $Y_t - 1$ are showing the cell states at time t and $t - 1$. While W shows weight matrices and b represents bias to the cell sate.

3.6.3 Step 3

The last step consists of output values h_t. These values depend upon output cell state Y_t ; however, in a processed form. In ordered to create output, *sigmoid* layer selects the part of cell state. After that *sigmoid* gate Y_t output is multiplied by the new values that are produced by *tanh* layer from the cell state Y_t.

$$Y_t = (W_o[h_t - 1; X_t] + b_o) \tag{3.6.5}$$

$$h_t = Y_t * tanh(C_t) \tag{3.6.6}$$

W_o and b_o shows the weight matrices and bias.

3.7 Deep Learning Model Training

During deep learning use glove embedding generation as LSTM training data. Different activation functions have been applied to discover the best readings.

Chapter 4

Results and Discussion

In this chapter, results are generated and validated to demonstrate the applicability of the proposed approach. After applying deep learning classifiers of LSTM on our dataset, we have used the standard definition of precision, recall and F-measure and confusion metrics for further evaluation of the results. For performance and efficiency evaluation of the model, precision, recall and F-measure are used.

4.1 Evaluation measures

F1 Score calculation and accuracy were used to assess the accuracy of the results. In binary classification statistical analysis, the F1 score (F-score/F-measure) is a measurement of how accurate a test is. It looks at the test's recall r and precision p while computing the score: p is the number of accurate positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that should have been returned.

The F1 score can be thought of as a weighted average of recall and precision, with 1 being the highest and 0 being the lowest. The number of accurately anticipated data points out of all the data points is known as accuracy. It's calculated by dividing the number of true positives and true negatives by the total number of true positives, true negatives, false positives, and false negatives.

$$Precision = TP/(TP + FP) \tag{4.1.1}$$

$$Recall = TP/(TP + FN) \tag{4.1.2}$$

$$F1Score = 2 * (Recall * Precision)/(Recall + Precision) \tag{4.1.3}$$

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \tag{4.1.4}$$

4.2 Experimental Setup

Before digging into the results compiled and achieved by the proposed methodology, let us discuss the experimental setup. For compiling the results, this thesis has used Python Programming Language in the Jupyter Notebook IDE. Implementation of the proposed methodology is done using Python 3.0. LSTM model is trained with the help of TensorFlow library. Different Python libraries used in the implementation of this code are explained in below sections.

1. Data Exploration
   (a) Pandas
   (b) Numpy
2. Preprocessing:
   (a) NLTK
   (b) Sci-Kit Learn
   (c) RE
3. Model Training/Testing
   (a) Sci-Kit Learn

   (b) Keras/TensorFlow

4. Data Visualization

   (a) MatPlotLib

   (b) Plotly

   (c) Seaborn

4.2.1 Hyper- and Parameters Configuration

Configuration of parameters and hypterparameters is required for the fitting of any DL model. In this thesis, experiments on various combinations of parameters and hyperparameters are performed and the best values are found out. For setting up parameteres and hyperparameters of the LSTM model, Table 4.1 provides all the configuration of the model. The initial adaptive learning rate was set to 0.001 with the reduce factor of 0.1 after patience of one epoch upto the minimum value of 0.00001. The reduction is monitored with the validation loss. The initial weights are embedded from Glove embedding 200. These weights were trainable for the Urdu data. The early stopping patience was set to 3. It means that when the validation loss does not reduce after three epochs, the training of the model stops to avoid overfitting. The optimizer used in this research is $ADAM$ and loss function used is $CategoricalCrossEntropy$. Different values for batch size is tried and the best value found is 512. The number of epochs for training of the model is infinity.

4.2.2 Model Architecture

A simple LSTM model is used in this research as the given data is already preprocessed intelligently and was neat and cleaned. The input layer has a shape of 200 tokens followed by an embedding layer where trained weights from Glove Embedding are provided as initial weights. Total number of parameters at this layer were 7150200. Then an LSTM layer with an output shape of 256 hidden neurons with a total of 467968 trainable parameters was implemented in the model followed by two

Table 4.1: Parameters Configuration

Hyper-\Parameter	**Configuration**	
Learning Rate	Initial Value	0.001
	Nature	Adaptive
	Reduce factor	0.1
	Patience	1
	Minimum	0.00001
	Monitor	Val Loss
Weights	Initial input	Glove 200
	Trainable	True
Stopping Criteria	Early Stopping	True
	Monitor	Val Loss
	Patience	3
Training-Validation	Optimizer	ADAM
	Loss	Categorical Cross Entropy
	Max Epochs	Inf
	Batch Size	512
	Val Split	20%
	Performance Metrics	Accuracy
		Loss
		Val Accuracy
		Val Loss

dense layers having 32896 trainable parameters. The LSTM layer outputs are initially dropped out to avoid overfitting of the model and then normalized using the Batch Normalization Layer. Final output layers have a total of 3 labels as our problem is of three classes. The activation function after first dense layer is $SeLU$ where final layer has $SoftMax$ activation function. The schematic diagram of the architecture used in this research is shown in Fig.4.1.

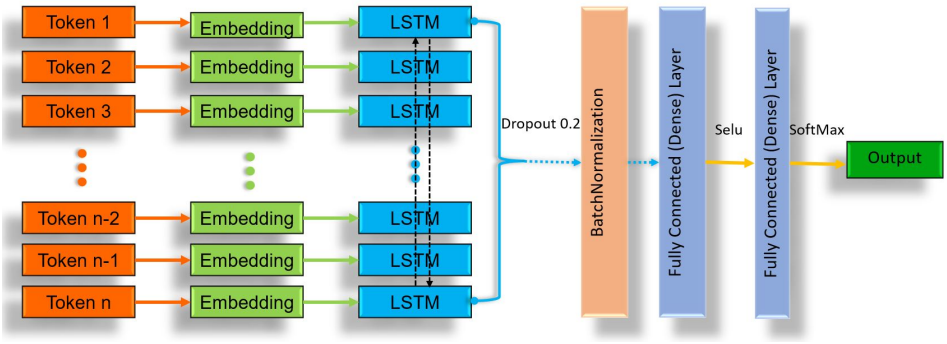

Figure 4.1: Model Architecture

4.3 Results

This section highlights an overview of best results that are achieved for the parameteres and hyperparameters given in Table 4.1. First data is splitted into train test with a ratio of 80:20. From training data, before passing it to the proposed model, we have splitted it further into 80:20 ratio between train and validation set. Moreover, callbacks are also applied on the basis of validation loss while training the model to save the best weights. Below figures and graphs represent overall results of training, validation, testing and comparison. The experimental results using LSTM for Urdu sentiment analysis and shows the effectiveness of LSTM approach by achieving 86.81% accuracy and 0.86 F1-score. Standard metrices i.e., precision, recall and F-measure have been used for evaluation of our deep learning LSTM based framework. Classification results have been shown in Table 4.2. The average precision of the proposed methodology for the cleaned Urdu data reaches to 87.21%, while the recall is slightly lower, reaches 86.81%. The F1-score for the proposed model touches 87% value for the same data. This shows that proposed model is neither underfit nor overfit on the preprocessed and cleaned Urdu data.

4.3.1 Model training and validation accuracy plot

The training and validation graph is captured during the model training is shown in the Figure 4.2. Blue line indicates the training curve and orange line is the validation curve while training. Y-axis indicates how a model is trained (in accuracy) and x-axis

Table 4.2: Classification Report

Class	Precision %	Recall	F1-score
Negative	83.19	92.52	87.61
Positive	89.36	75.45	81.82
Neutral	89.09	89.91	89.50
Macro Average	87.21	85.96	86.31
Weighted Average	87.06	86.81	86.69

indicates the number of epochs. As number of epochs proceed, the training accuracy is increased. A comparison of the accuracy between validation, training and testing is shown in Figure 4.3.

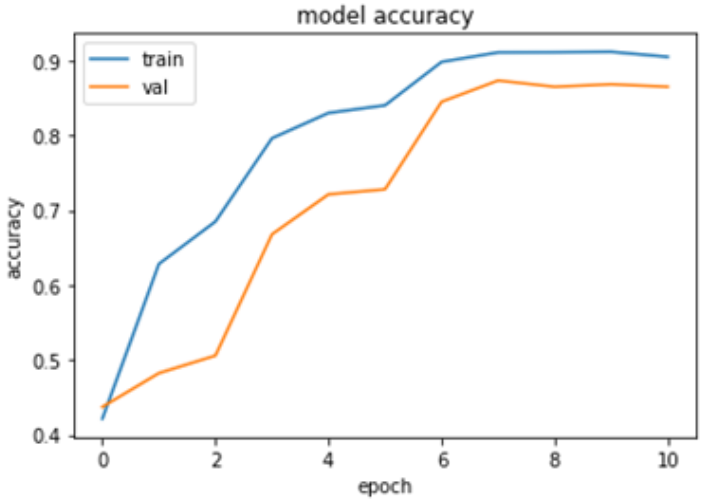

Figure 4.2: Model Training and Validation Accuracy Plot

4.3.2 Confusion Matrix

Confusion matrix is a tabular form of results obtained after applying certain classifier on the data set. It contains values of actual classified objects and predicted values. Confusion matrix gives visualization of the performance of the classification technique. Figure 4.4 shows the confusion matrix. N in confusion matrix represents total number of sentences in data set.

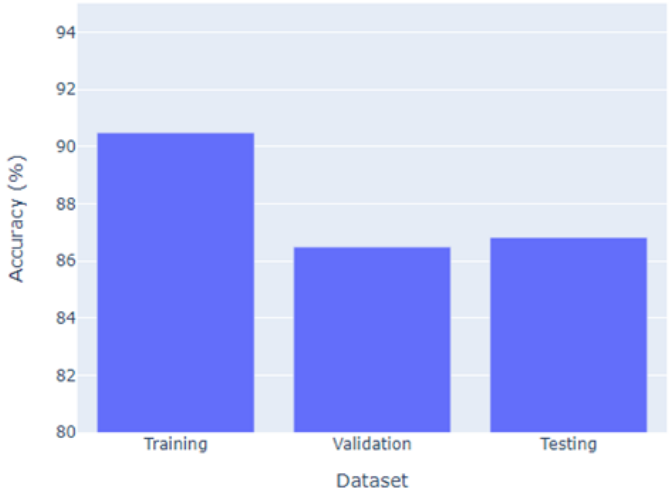

Figure 4.3: Accuracy Comparison

Neg	198	9	7
Pos	24	126	17
Neu	15	7	196
	Neg	Pos	Neu

Figure 4.4: Confusion Matrix

4.4 Performance of each label

On the basis of different evaluation measures discussed in the section 4.1, the performance values of all the three labels (Positive, Negative and Neutral) in terms of precision, accuracy and F1-score is depicted in Figure 4.5. Precision for the Positive and Neutral tweets are more than that of Negative tweets due to the slight imbalanceness of the data. As the number of Negative tweets are comparatively more than Positive and Neutral tweets, that is the reason that Recall for Negative tweets is 92.52%, which is higher than the rest of two performance measures. All the three per-

formance measures, i.e., Precision, Recall and F1-score of Neutral tweets are almost the same which depicts that the model is best fit for Neutral tweets. On the other hand, the Positive tweets have better precision value while comparatively lower recall. Unlike Positive tweets, Negative tweets has better recall value and lower precision. The proposed model performance better for Negative and Neutral tweets but slightly overfits on Positive tweets.

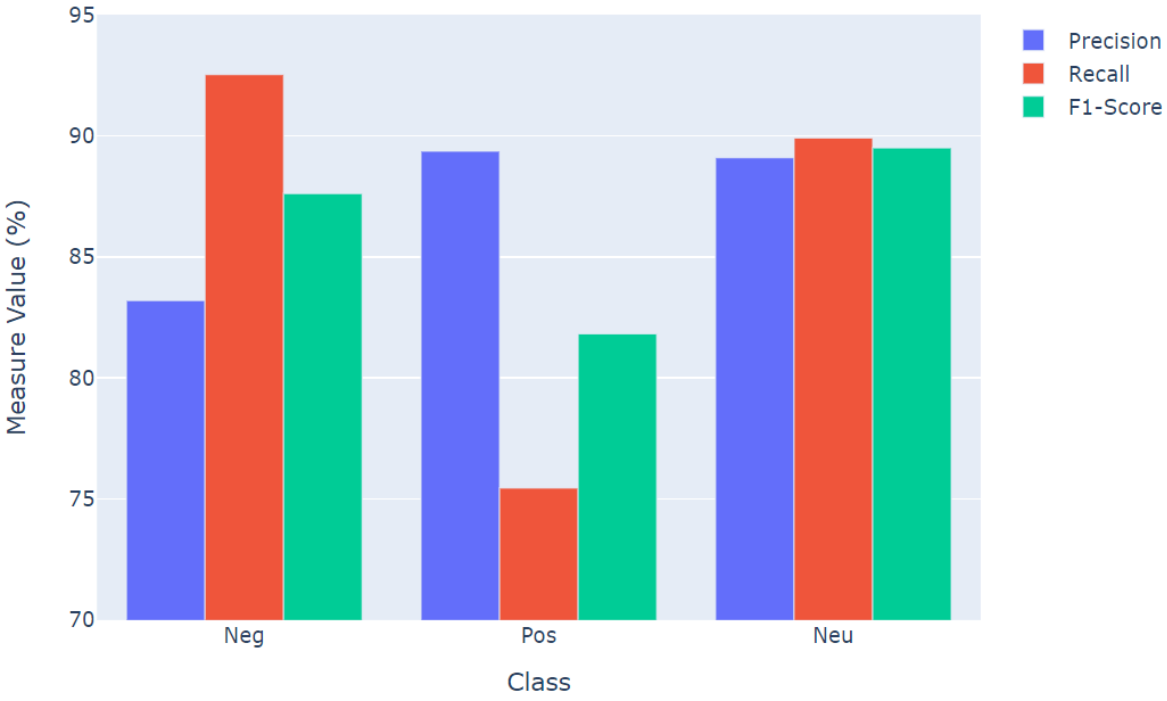

Figure 4.5: Performance of each label

4.5 Comparative analysis

4.5.1 Comparison on the basis of same dataset

In this section, a fair comparison is performed between our proposed method and a paper from literature, which has used the same dataset but different preprocessing steps and classifiers. Mukhtar et al. [46] used the supervised machine learning technique to devise a system that identifies positive, negative and neutral sentences from the data set. In their research, the authors use supervised learning technique to achieve desired

results and very less focus has been given to preprocessing. They have focused on the removal of stop words from the data, only, which results is lower accuracy due to noise in the data. Unlike [46], in our proposed methodology, we have mainly focused on the preprocessing of the data and have used deep learning technique to improve the results. We have performed many data cleaning and preprocessing steps, discussed in chapter 3, including removal of special character, digits, white spaces, stop words, and last but not the least, the stemming of Urdu words generated by this research. Table 4.3 shows when comparison is done to the existing paper using same dataset, our proposed methodology surpassed in terms of utilizing deep learning classifier of LSTM.

Table 4.3: Comparison on the same dataset

Author	Classifier	Preprocessing	Accuracy
N. Mukhtar et al. [46]	NB, Lib SVM, DT, KNN	Stop Word Removal	67%
Proposed Methodology	LSTM	Removal of special characters, digits, stop words, white spaces, and an Urdu Stemming dictionary of 830 words	86.8%

4.5.2 Comparison on the basis of different techniques and different datasets

As discussed earlier in chapter number 1 and 2 that very less research has been done on sentiment analysis of Urdu tweets. Different authors have used different datasets and classifiers. Table 4.4 presents a comparison with existing techniques of supervised and deep leaning techniques and our proposed methodology outperforms all. The comparison has been made on the basis of different classifiers and different datasets. The more details of their datasets is given in Chapter 2 and Table 2.2. This section discusses the performance of the proposed methodology with the recent literature. In 2018, the accuracy achieved for Urdu data was 67.58%, using simple ML classifiers,

Table 4.4: Comparative analysis with different techniques having different datasets

Author	Year	Method	Preprocessing	Accuracy
K. Mehmood et al. [73]	2018	KNN, SVM, DT, LR, NB	Improper preprocessing. Eliminating reviews written in English language and Arabic script.	67.58%
M. A. Manzoor et al. [80]	2020	CNN, LSTM, BiLSTM	Normalization	69.3%
I. Safder et al. [82]	2021	RCNN, SVM, CNN, LSTM	Removal of junk characters	84%
Proposed Methodology	2022	LSTM	Removal of special characters, digits, stop words, white spaces, and Urdu Stemming dictionary of 830 words	86.8%

which is quite a low [73]. In 2020, M. A. Mazoor et al. [80] introduced DL techniques for the Urdu Language but they have not focused on the preprocessing methods which results in lower accuracy of 69.3% only. In 2021, I. Safder et al. [82], have focused on the DL methods along with data cleaning steps by removing noise form the Urdu data. This resulted in a much better accuracy of 84%. The datasets used by these researchers are quite a small for DL methods. In our proposed method, we have proposed a list of stop words and a stemming dictionary of 830 stems. The DL model used in our proposed methodology is a simple LSTM with Glove Embedding but the preprocessing data cleaning was done intelligently. This results in an accuracy of 86.81%, which outperformed more advanced DL methods.

Chapter 5

Conclusions and Future Work

Sentiment analysis is mostly done in English language from last 3 decades. English corpus is available online and many preprocessing tools are also available. Urdu language is ignored from the researcher because of the morphological issues and challenges. Novel techniques that have been applied on other languages are not applicable for Urdu language due to its complex structure. The basic purpose of this research is to apply state of the art deep learning technique of LSTM to perform Urdu sentiment analysis. An efficient technique is achieved with significant increase in accuracy of difficult word identification. 86.8% accuracy and 0.89 F1 Score is attained using labeled dataset of 3995 sentences. This research shows improved result in terms of accuracy, precision, recall and f-measure as compared to the previous research done on Urdu text data and provides a framework that is particularly applicable for dealing with scarce resource language Urdu. In future, further research is required to discover more detailed techniques, methods, architectures and its appropriate training algorithms. Based on the research work done on this topic, some work can be done in future to improve the results and quality of the research. Some of the directions are as follows:

1. In this research work, we have used dataset comprising of 3995 sentences. In future, we would like to extend the span of data set by including different dialects and accents.

2. In future, the system can be upgraded by incorporating different neural networks techniques which are becoming popular in different fields The results produced by

textual sentimental analysis can be further improved by linguistic and rule-based approach.

3. Also, proposed framework can be used for multilingual purpose using multiple languages other than Urdu.

4. Stemming dictionary can be extended further to achieve desired accuracy.

References

[1] Iqbal Muhammad and Zhu Yan. Supervised machine learning approaches: A survey. *ICTACT Journal on Soft Computing*, 5(3), 2015.

[2] Muhammad Abid, Asad Habib, Jawad Ashraf, and Abdul Shahid. Urdu word sense disambiguation using machine learning approach. *Cluster Computing*, 21 (1):515–522, 2018.

[3] Ali Al-Badi, Ali Tarhini, and Nabeel Al-Qirim. Risks in adopting cloud computing: a proposed conceptual framework. In *International Conference for Emerging Technologies in Computing*, pages 16–37. Springer, 2018.

[4] Nazish Azam, Bilal Tahir, and Muhammad Amir Mehmood. Sentiment and emotion analysis of text: a survey on approaches and resources. In *7th International Conference on Language and Technology (CLT)*, pages 87–94, 2020.

[5] Mike Thelwall. Sentiment analysis tasks and methods, 2015.

[6] Cach N Dang, María N Moreno-García, and Fernando De la Prieta. An approach to integrating sentiment analysis into recommender systems. *Sensors*, 21(16): 5666, 2021.

[7] Mubashir Ali, Shehzad Khalid, and Muhammad Haseeb Aslam. Pattern based comprehensive urdu stemmer and short text classification. *IEEE Access*, 6:7374–7389, 2017.

[8] Uzma Naqvi, Abdul Majid, and Syed Ali Abbas. Utsa: Urdu text sentiment analysis using deep learning methods. *IEEE Access*, 9:114085–114094, 2021.

[9] Uzma Naqvi, Abdul Majid, and Syed Ali Abbas. Utsa: Urdu text sentiment analysis using deep learning methods. *IEEE Access*, 9:114085–114094, 2021.

[10] Afraz Z Syed, Muhammad Aslam, and Ana Maria Martinez-Enriquez. Lexicon based sentiment analysis of urdu text using sentiunits. In *Mexican international conference on artificial intelligence*, pages 32–43. Springer, 2010.

[11] Zia Ul Rehman and Imran Sarwar Bajwa. Lexicon-based sentiment analysis for urdu language. In *2016 sixth international conference on innovative computing technology (INTECH)*, pages 497–501. IEEE, 2016.

[12] André Luiz Firmino Alves, Claudio De Souza Baptista, Anderson Almeida Firmino, Maxwell Guimarães de Oliveira, and Anselmo Cardoso de Paiva. A comparison of svm versus naive-bayes techniques for sentiment analysis in tweets: A case study with the 2013 fifa confederations cup. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, pages 123–130, 2014.

[13] Marko Balabanović and Yoav Shoham. Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.

[14] Kenneth Bloom and Shlomo Argamon. Unsupervised extraction of appraisal expressions. In *Canadian Conference on Artificial Intelligence*, pages 290–294. Springer, 2010.

[15] Avinash Chandra Pandey, Dharmveer Singh Rajpoot, and Mukesh Saraswat. Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management*, 53(4):764–779, 2017.

[16] Thomas Bos and Flavius Frasincar. Automatically building financial sentiment lexicons while accounting for negation. *Cognitive Computation*, 14(1):442–460, 2022.

[17] D Alessia, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3), 2015.

[18] Misbah Daud, Rafiullah Khan, Aitazaz Daud, et al. Roman urdu opinion mining system (ruomis). *arXiv preprint arXiv:1501.01386*, 2015.

[19] Yulan He. Incorporating sentiment prior knowledge for weakly supervised sentiment analysis. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(2):1–19, 2012.

[20] Yulan He. Incorporating sentiment prior knowledge for weakly supervised sentiment analysis. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(2):1–19, 2012.

[21] Wojciech Gryc Melville. Sentiment analysis of blogs by combining lexical knowledge with text classification, kdd 09, paris, france, 2009.

[22] Hua Pang. Microblogging, friendship maintenance, and life satisfaction among university students: The mediatory role of online self-disclosure. *Telematics and Informatics*, 35(8):2232–2241, 2018.

[23] Gule Zulf Nargis and Noreen Jamil. Generating an emotion ontology for roman urdu text. *International Journal of Computational Linguistics Research*, 7(2016): 83–91, 2016.

[24] Afraz Z Syed, Muhammad Aslam, and Ana Maria Martinez-Enriquez. Associating targets with sentiunits: a step forward in sentiment analysis of urdu text. *Artificial intelligence review*, 41(4):535–561, 2014.

[25] William H Dietz. Health consequences of obesity in youth: childhood predictors of adult disease. *Pediatrics*, 101(Supplement_2):518–525, 1998.

[26] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[27] I Goodfellow, Y Bengio, A Courville, and Y Bengio. Deep learning, 1 mit press, 2016.

[28] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*, 2015.

[29] Xiaoli Zhao, Shaofu Lin, and Zhisheng Huang. Text classification of micro-blog's" tree hole" based on convolutional neural network. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, pages 1–5, 2018.

[30] Yann LeCun, Leon Bottou, Y Bengio, and P Haffner. A b7cedgf hib7prqtsudgqicwvyx hib edcdsisixvg5r cdqtw xvefcds. In *proc. IEEE*, 1998.

[31] Min Yen Wu, Chih-Ya Shen, En Tzu Wang, and Arbee LP Chen. A deep architecture for depression detection using posting, behavior, and living environment data. *Journal of Intelligent Information Systems*, 54(2):225–244, 2020.

[32] Min Yen Wu, Chih-Ya Shen, En Tzu Wang, and Arbee LP Chen. A deep architecture for depression detection using posting, behavior, and living environment data. *Journal of Intelligent Information Systems*, 54(2):225–244, 2020.

[33] Akshaya Ranganathan, A Haritha, D Thenmozhi, and Chandrabose Aravindan. Early detection of anorexia using rnn-lstm and svm classifiers. In *CLEF (Working Notes)*, 2019.

[34] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[35] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[36] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135, 2008.

[37] Surendra Kumar and Suryakant Pathak. Sentiment analysis methods using lexicon approach. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, 14(01):86–92, 2022.

[38] Mark Anthony Cabanlit and Kurt Junshean Espinosa. Optimizing n-gram based text feature selection in sentiment analysis for commercial products in twitter through polarity lexicons. In *IISA 2014, The 5th International Conference on Information, Intelligence, Systems and Applications*, pages 94–97. IEEE, 2014.

[39] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373, 2004.

[40] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631, 2005.

[41] Waqas Anwar, Xuan Wang, and Xiao-long Wang. A survey of automatic urdu language processing. In *2006 International Conference on Machine Learning and Cybernetics*, pages 4489–4494. IEEE, 2006.

[42] K Dashtipour, S Poria, A Hussain, E Cambria, AY Hawalah, A Gelbukh, and Q Zhou. Multilingual sentiment analysis: state of the art and independent comparison of techniques. cogn. comput. 8 (4), 757–771 (2016).

[43] Muhammad Zubair Asghar, Anum Sattar, Aurangzeb Khan, Amjad Ali, Fazal Masud Kundi, and Shakeel Ahmad. Creating sentiment lexicon for sentiment analysis in urdu: The case of a resource-poor language. *Expert Systems*, 36(3): e12397, 2019.

[44] Uzma Naqvi, Abdul Majid, and Syed Ali Abbas. Utsa: Urdu text sentiment analysis using deep learning methods. *IEEE Access*, 9:114085–114094, 2021.

[45] Madiha Ijaz and Sarmad Hussain. Corpus based urdu lexicon development. In *the Proceedings of Conference on Language Technology (CLT07), University of Peshawar, Pakistan*, volume 73, 2007.

[46] Neelam Mukhtar and Mohammad Abid Khan. Urdu sentiment analysis using supervised machine learning approach. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(02):1851001, 2018.

[47] Ali Daud, Wahab Khan, and Dunren Che. Urdu language processing: a survey. *Artificial Intelligence Review*, 47(3):279–311, 2017.

[48] D Alessia, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3), 2015.

[49] Ainur Yessenalina, Yisong Yue, and Claire Cardie. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1046–1056, 2010.

[50] Karine Barzilai-Nahon. Toward a theory of network gatekeeping: A framework for exploring information control. *Journal of the American society for information science and technology*, 59(9):1493–1512, 2008.

[51] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[52] W Medhat, A Hassan, and H Korashy. Sentiment analysis algorithms and applications: a survey. ain shams eng. j. 5 (4), 1093–1113 (2014).

[53] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[54] Peerapon Vateekul and Thanabhat Koomsubha. A study of sentiment analysis using deep learning techniques on thai twitter data. In *2016 13th international joint conference on computer science and software engineering (JCSSE)*, pages 1–6. IEEE, 2016.

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

[56] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language*

technology conference and conference on empirical methods in natural language processing, pages 347–354, 2005.

[57] Philip J Stone and Earl B Hunt. A computer approach to content analysis: studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference*, pages 241–256, 1963.

[58] Yanqing Chen and Steven Skiena. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, 2014.

[59] Yoonjung Choi and Janyce Wiebe. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191, 2014.

[60] Nobuhiro Kaji and Masaru Kitsuregawa. Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 1075–1083, 2007.

[61] Hiroshi Kanayama and Tetsuya Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 355–363, 2006.

[62] Peter D Turney and Michael L Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[63] GA Miller, R Beckwith, C Fellbaum, D Gross, and K Miller. Introduction to wordnet: on-line. *Distributed with the WordNet Software*, 1993.

[64] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[65] Xiaowen Ding, Bing Liu, and Philip S Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240, 2008.

[66] Beata Beigman Klebanov, Nitin Madnani, and Jill Burstein. Using pivot-based paraphrasing and sentiment profiles to improve a subjectivity lexicon for essay data. *Transactions of the Association for Computational Linguistics*, 1:99–110, 2013.

[67] Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356, 2011.

[68] Iqra Safder, Zainab Mahmood, Raheem Sarwar, Saeed-Ul Hassan, Farooq Zaman, Rao Muhammad Adeel Nawab, Faisal Bukhari, Rabeeh Ayaz Abbasi, Salem Alelyani, Naif Radi Aljohani, et al. Sentiment analysis for urdu online reviews using deep learning models. *Expert Systems*, page e12751, 2021.

[69] Vivek Kumar Singh, Mousumi Mukherjee, and Ghanshyam Kumar Mehta. Combining collaborative filtering and sentiment classification for improved movie recommendations. In *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, pages 38–50. Springer, 2011.

[70] Neelam Mukhtar, Mohammad Abid Khan, and Nadia Chiragh. Effective use of evaluation measures for the validation of best classifier in urdu sentiment analysis. *Cognitive Computation*, 9(4):446–456, 2017.

[71] Nadia Tabassum, Tahir Alyas, Muhammad Hamid, Muhammad Saleem, Saadia Malik, Zain Ali, and Umer Farooq. Semantic analysis of urdu english tweets empowered by machine learning. *Intelligent Automation and Soft Computing*, 30 (1):175–186, 2021.

[72] Yibo Wang, Mingming Wang, and Wei Xu. A sentiment-enhanced hybrid rec-

ommender system for movie recommendation: a big data analytics framework. *Wireless Communications and Mobile Computing*, 2018, 2018.

[73] Khawar Mehmood, Daryl Essam, and Kamran Shafi. Sentiment analysis system for roman urdu. In *Science and Information Conference*, pages 29–42. Springer, 2018.

[74] Zareen Sharf and Saif Ur Rahman. Performing natural language processing on roman urdu datasets. *International Journal of Computer Science and Network Security*, 18(1):141–148, 2018.

[75] Lqra Sana, Khushboo Nasir, Amara Urooj, Zain Ishaq, and Ibrahim A Hameed. Bers: Bussiness-related emotion recognition system in urdu language using machine learning. In *2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*, pages 238–242. IEEE, 2018.

[76] Faiza Noor, Maheen Bakhtyar, and Junaid Baber. Sentiment analysis in e-commerce using svm on roman urdu text. In *International Conference for Emerging Technologies in Computing*, pages 213–222. Springer, 2019.

[77] I Goodfellow, Y Bengio, A Courville, and Y Bengio. Deep learning, 1 mit press, 2016.

[78] Ayush Agarwal, Ashima Yadav, and Dinesh Kumar Vishwakarma. Multimodal sentiment analysis via rnn variants. In *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD)*, pages 19–23. IEEE, 2019.

[79] Xueying Zhang and Xianghan Zheng. Comparison of text sentiment analysis based on machine learning. In *2016 15th international symposium on parallel and distributed computing (ISPDC)*, pages 230–233. IEEE, 2016.

[80] Muhammad Arslan Manzoor, Saqib Mamoon, Song Kei Tao, Ali Zakir, Muhammad Adil, and Jianfeng Lu. Lexical variation and sentiment analysis of roman urdu sentences with deep neural networks. *Int. J. Adv. Comput. Sci. Appl*, 11: 719–726, 2020.

[81] Lal Khan, Ammar Amjad, Noman Ashraf, Hsien-Tsung Chang, and Alexander Gelbukh. Urdu sentiment analysis with deep learning methods. *IEEE Access*, 9: 97803–97812, 2021.

[82] Iqra Safder, Zainab Mahmood, Raheem Sarwar, Saeed-Ul Hassan, Farooq Zaman, Rao Muhammad Adeel Nawab, Faisal Bukhari, Rabeeh Ayaz Abbasi, Salem Alelyani, Naif Radi Aljohani, et al. Sentiment analysis for urdu online reviews using deep learning models. *Expert Systems*, page e12751, 2021.

[83] Ganesh K Shinde, Vaibhav N Lokhande, Rasika T Kalyane, Vikas B Gore, and Umesh M Raut. Sentiment analysis using hybrid approach.

[84] Faiza Hashim and M Khan. Sentence level sentiment analysis using urdu nouns. *Department of Computer Science, University of Peshawar, Pakistan*, pages 101–108, 2016.

[85] Smruthi Mukund, Debanjan Ghosh, and Rohini K Srihari. Using sequence kernels to identify opinion entities in urdu. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 58–67, 2011.

[86] Smruthi Mukund and Rohini K Srihari. Analyzing urdu social media for sentiments using transfer learning with controlled translations. In *Proceedings of the second workshop on language in social media*, pages 1–8, 2012.

[87] S Abbas Ali, M Daniyal Noor, Munir Ahmed Javed, M Mohsin Aslam, Omer Ahmed Khan, et al. Salience analysis of news corpus using heuristic approach in urdu language. *International Journal of Computer Science and Network Security (IJCSNS)*, 16(4):28, 2016.

[88] Afraz Z Syed, AM Martinez-Enriquez, Akhzar Nazir, Muhammad Aslam, and Rida Hijab Basit. Mining the urdu language-based web content for opinion extraction. In *Mexican Conference on Pattern Recognition*, pages 244–253. Springer, 2017.

[89] Hussain Ghulam, Feng Zeng, Wenjia Li, and Yutong Xiao. Deep learning-based

sentiment analysis for roman urdu text. *Procedia computer science*, 147:131–135, 2019.

[90] Zainab Mahmood, Iqra Safder, Rao Muhammad Adeel Nawab, Faisal Bukhari, Raheel Nawaz, Ahmed S Alfakeeh, Naif Radi Aljohani, and Saeed-Ul Hassan. Deep sentiments in roman urdu text using recurrent convolutional neural network model. *Information Processing & Management*, 57(4):102233, 2020.