POLICY RESEARCH JOURNAL
ISSN (E): 3006-7030 (P) 3006-7022
Volume 3, Issue 1, 2025
Received: 02 December 2024
Accepted: 02 January 2025
Published: 09 January 2025

SARAIKI LANGUAGE CORPUS FOR SENTIMENT ANALYSIS

Nisha Rafique[*1], Saher Liaqat[2]

[*1,2]*Department of Computer Science Institute of Southern Punjab Multan, Pakistan*

[*1]nisharafique2@gmail.com, [2]saherliaqat2018@gmail.com

ABSTRACT

This study uses a large corpus created especially for this purpose to investigate sentiment analysis of the Saraiki language. The corpus represents the linguistic and cultural diversity of the Saraiki-speaking community by incorporating a wide range of Saraiki text data that was gathered from different sources. Using this dataset, we investigate the effectiveness of many machine learning algorithms for sentiment analysis tasks, such as Support Vector Classification (SVC), Naïve Bayes, Random Forest, Logistic Regression, and CNN. The Saraiki language corpus is used to train and assess each algorithm, and the results are compared based on performance criteria like accuracy, recall, and precision. The experimental findings indicate that the algorithms are not all equally successful; CNN performs better than Naive Bayes, Logistic Regression, Random Forest, and SVC. In addition, we provide a thorough examination of each algorithm's advantages and disadvantages when it comes to performing sentiment analysis tasks in the Saraiki language context. Our results highlight how crucial it is to use domain-specific corpora and cutting-edge machine learning approaches to accurately analyze sentiment in languages with limited resources, such as Saraiki. By developing sentiment analysis techniques specific to the Saraiki language, this research advances our knowledge of the sentiment dynamics within the Saraiki-speaking population.

Keywords: *Sentiment Analysis, Linguistic, Corpus, Support Vector, Approaches.*

INTRODUCTION

An important part of NLP is opinion analysis, usually known as opinion mining. That concentrates on understanding, analyzing human sentiments, opinions, attitudes expressed in text. With the growing dominance of online platforms and social networking sites, in recent times, sentiment analysis has emerged as an increasingly current field of study. The use of sentiment analysis has increased significant value in several areas, such as social-media monitoring, customer response analysis, and advertising.

1.1 Background

Sentiment analysis is major job in broader NLP field (Mazoochi, 2023). When textual sources are analyzed automatically for sentiment, the system attempts to annotate each text according to sentiment behind it, which can be either negative, neutral, or positive (Krister and Jauhiainen, 2023). Sentiment analysis is especially helpful in determining what the public thinks about certain services, and interesting subjects. Sentiment analysis tools are more advanced for English than for other languages, such Kurdish. This is because languages with lower resources have fewer NLP tools available, such as annotated datasets (Badawi, 2024).

A small language in India and Southern Punjab of Pakistan, including Multan, Muzaffargarh, and Dera Ghazi Khan, the language Saraiki is Indo-Aryan. There are over 26 million native language speakers in our country. Despite having its own set

of 45-letter alphabets, it is written in Perso-Arabic characters. Of the 45 letters, 39 are the same as those used in the Urdu language, and the other six are new. The Main Saraiki, Thalli, Rajanpur, the Southern Saraiki, and the Cholestane Desert are depicted by Multan and surrounding regions, and Thar are among the numerous dialects of this language (author= {Saleemi, 2021).

An annotation process is used with UPOS, morphological tokens, emotional analysis, and Document Term Matrix. Provide an explanation of the Sindhi text corpus's sentence structure, along with the methods and sources used for the data collection. The process culminates in the construction of a text corpus after a full understanding of the issue (Talpur, 2023).

Most of the recent research has been directed at creating English sentiment corpora. As a result, there is a huge demand and room for study to create opinion corpora for non-English languages (Mazoochi, 2023). The corpus is suitable for several jobs, including retrieving information and machine translation, pattern identification and text-to-speech, text to tokenization, dictionary, thesaurus, word-to-vector analysis, feature extraction and analysis, and text recognition of words, classification, and more (Dootio, 2021).

To create corpus for Finnish language, each of our annotators received one of the nine work packages containing three thousand sentences. The polarity distribution of the sentences from Finnish preselection categories was the same in every bundle. (Krister and Jauhiainen, 2023).

The MAC corpus contributes to advancements in sentiment analysis methods in the Moroccan perspective and enables a deeper identification of public opinion, and sentiment of social media in the Arabic language of Moroccan (Garouani, 2022).

In 2023, a study on Analysis of Sentiment for Hausa and Igbo, Low-Resource languages of Africa was presented. In which sentiments analysis was performed for the two African languages Hausa, and Igbo. For this author applied the method Berta on African languages called AFRI BERTa. The resulted F1score was for Hausa:0.809 and Igbo:0.806 (Raychawdhary, 2023).

A Brazilian testing with automated corpus tagging for sentiment analysis was conducted in 2020. A specific location of Recife saw the collection of a dataset as a collection of tweets. Sentiment analysis was carried out using Logistic Regression, Linear SVM, and Multinomial Naive Bayes. Accuracy for the Logistic Regression was 0.7405, 0.6527 for Multinomial Nave Bayes, and the accuracy for Linear SVM was 0.7379 (de Carvalho, 2020).

The document's inverse and term frequency in the data can be found using the TF-IDF, which can also be used to convert the vectors. Different kinds of methods are used, and every makes use of a unique set of categorization algorithms. These are employed for text classification and, consequently, for analysis. The classifiers that were employed were KN- N, RF, and DT. Half the information is employed in training and the other half is employed in testing the train and test split (Patel A., 2023).

1.2 Motivation

There is already much work done on high resource languages like English. There is less work on local languages like Saraiki, Pashto, and languages of other countries. As some authors have tried work on local languages but it is not so enough to perform machine learning tasks like classification, Information retrieval and sentiment analysis. To perform processing like translation, sentiment analysis and other ML algorithms on any language, corpus is needed.

Because of the absence of resources in Saraiki, it is challenging to analyze the sentiment in Saraiki text. But there are not many works on the resource-low native languages for sentiment analysis. In addition, corpus of Saraiki language is not developed for sentiment analysis. So, this research will solve this issue by creating a trustworthy corpus for sentiment analysis.

POLICY RESEARCH JOURNAL
ISSN (E): 3006-7030 (P) 3006-7022

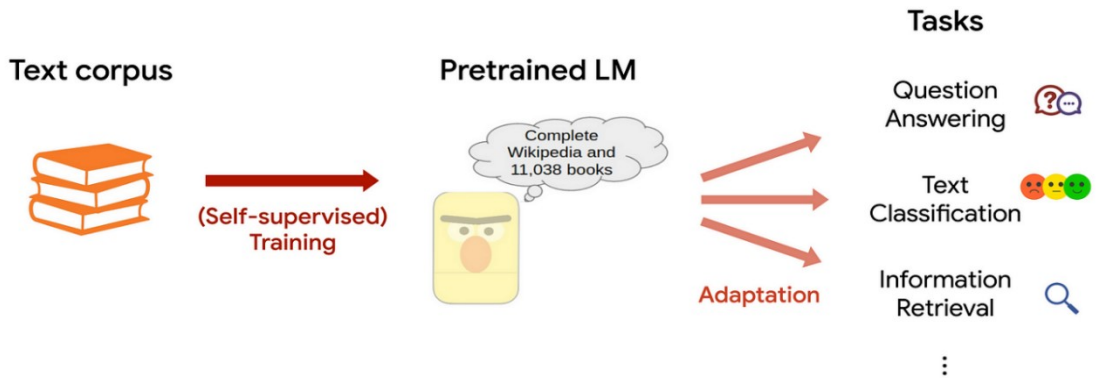

Figure 1.1: General Text corpus Creation.

In this research, a corpus of 7000 sentences of Saraiki language is created by collecting data from different sources like newspapers, dictionaries and from other social networks. After the collection of sentences the corpus is labelled as positive, negative, or neutral by assigning them o or 1. The important task was to make proper corpus of Saraiki. When the data is properly collected and labelled then preprocessing is performed. Preprocessing includes tokenization of sentences, removal of white spaces, removal of stop words and special characters. The feature extraction techniques N-gram and TF/IDF are performed on dataset. Then perform sentiment analysis on Saraiki dataset the ML methods CNN, RF, Logistic Regression, and SVC. Results are obtained, and accuracy is checked. After that the results are compared of these methods. A comparison is performed of Saraiki sentiment model with existing approaches. This study will promote the Saraiki language. This will help to make the Saraiki language specific models which will be used for more work on language.

1.3 Problem Statement

The lack of resources and approaches for sentiment analysis in the Saraiki language, which makes it difficult to accurately analyze the sentiment and emotions portrayed in Saraiki text, is the research problem. The creation of trustworthy sentiment analysis tools for Saraiki text is hampered by the absence of a specialized sentiment analysis corpus, annotated data, and language-specific models. Consequently, the study attempts to address this problem by developing a sentiment analysis framework that includes an annotated data set, a sentiment analysis corpus, techniques for feature extraction, and an evaluation of the Saraiki sentiment evaluation model's implementation in relation to alternative approaches.

1.4 Research Objective

Research objectives of study are:

1. To establish a corpus of text for sentiment analysis using Saraiki language.
2. To assess the Saraiki language corpus and provide annotations for sentiment analysis.
3. To Explore feature extraction techniques for Saraiki corpus sentiment analysis
4. To contrast the Saraiki sentiment analysis model with the methods already in use.

1.5 Research Questions

The research questions of the study are:

1. How can text in the Saraiki language be used to create a corpus for sentiment analysis?
2. How should the Saraiki language corpus be assessed and annotated for sentiment analysis?
3. What are some ways to investigate feature extraction techniques for Saraiki corpus sentiment analysis?
4. How to Compare Saraiki sentiment analysis with methods already in use?

1.6 Significance of study

1. Preservation of linguistic heritage: For linguistic and cultural reasons, it is essential to develop a framework for the analysis of sentiments in the Saraiki language. It contributes to the preservation and advancement of the Saraiki language by providing the necessary tools for understanding

and assessing the sentiment conveyed in Saraiki literature.

1 Deeper understanding of speaker:
By accurately assessing sentiment in Saraiki literature, the research contributes to our understanding of the emotional states, opinions, and attitudes of Saraiki speakers. This data can be applied in a range of industries, including political attitude tracking, shared opinion research, social media analysis, and market research.

2 Deeper comprehensions of speakers:
The research advances our knowledge of the emotional states, worldviews, and attitudes of the speakers by precisely assessing sentiment in Saraiki text. Numerous industries can make use of this data, including social media analysis, market research, public opinion research, political attitude tracking, and more.

3.Improved Decision Making:
Accurate sentiment analysis in the Saraiki language gives decision makers pertinent data. Making well-informed decisions about product development, customer satisfaction, Public opinion and policy formulation is crucial in assisting businesses, governments, and organizations in attaining better outcomes.

4. Research on Sentiment Analysis Progress:
The study is helpful for sentiment analysis in general, especially for languages with little funding for research. For the Saraiki exam, a sentiment analysis corpus, annotation rules, and language-specific models have been established to advance sentiment analysis methodologies and procedures. These resources will be highly beneficial for research in the future.

5.Overcoming language hurdles: The sentiment analysis framework for the Saraiki language aids in overcoming language barriers in sentiment analysis. It illustrates how important it is to consider a variety of languages and cultural contexts while doing research on sentiment analysis, since it will eventually hint to more accurate and inclusive sentimental analysis models and tools.

6.Practical Applications: There are several sectors in which the sentiment analysis framework can be used practically, such as social media monitoring, brand reputation management, customer comment analysis, political sentiment analysis, and mental health monitoring. It makes sentiment analysis useful for a variety of reasons for Saraiki-speaking researchers and organizations. 8. In general, the framework for creating a sentiment analysis for the Saraiki language recognizes the significance of language, culture, and practicality. It makes Saraiki speakers possible, improves comprehension, encourages dedication, and advances the field of sentiment analysis research for underserved languages.

1.7 Sequence of thesis
Chapter 2: The information in this chapter provides the background of Saraiki language corpus for sentiment analysis. It provides previous work done on the Saraiki language and for the development of the corpus that how the data was collected for the corpus creation and from which sources text was collected. Related to the sentiment analysis, all the previous work done is discussed in this chapter. In sentiment analysis the data is labelled as positive, negative, and neutral classes. Related to sentiment analysis work like data collection, labelling, preprocessing, feature extraction, ML methods on the dataset of different local languages is presented. The results after applying different methods are also given. Related to the corpus creation all the requirements and the sources from which data is collected, and sentiment analysis is presented.

Chapter 3: This chapter presents the complete methodology of Saraiki language corpus for sentiment analysis. The methodology of this research with Model diagram is properly discussed. Moreover, the methods which are applied to extract features like TF/IDF, N-gram are presented with the results. Regarding the categorization of sentiments as negative, positive, or neutral ML methods are given with their proper working and behavior on Saraiki Corpus. The

POLICY RESEARCH JOURNAL
ISSN (E): 3006-7030 (P) 3006-7022

overall execution of the model with charts is properly presented in this chapter.

C**hapter 4:** This chapter gives the overall results and evaluation obtained after applying the methods CNN, Naïve Bayes, Logistic Regression, Random Forest, SVC on Saraiki dataset. Moreover, proper discussion about the results and the comparison between these methods is which one gives less accuracy, and which one is performing better than the other.

Chapter 5: In this chapter the conclusion of the research is discussed. The evaluation of classifiers on Saraiki dataset and overall performance of model results are presented. Moreover, the future related to this is also given with some suggestions.

2. Literature Review.

Sr	Year	Language	Text Resources	Method	Results	Ref
1	2023	Finish	Scraped from online web stores	CNN	79.0%	(Krister and Jauhiainen, 2023)
2	2018	Persian	Customer Reviews, social media	Fleiss ‘s kappa measure	Target Word Annotation 62.60	(Hosseini, 2018)
3	2021	Arabic	Extraction of social media posts	Long short-term memory (LSTM)	MSA POS tagging + Arabic POS tagging +85.7% 9, -1.24%, neutral 88.65%	(Nerabie, 2021)
4	2022	Nigerian	Annotated tweets	NLP and many machine learning models	Better by 41.8%, 20.8%, 4%, and 19.1% respectively.	(Muhammad S. H., 2022)
5	2022	Moroccan Arabic	Tweets, Facebook comments	SVM, Logistic Regression, LSTM and CNN,	91.27%, obtained by using the LSTM classifier	(Garouani, 2022)
6	2020	Urdu	Urdu social media content, comments	Statistical data-driven methods	93.8% precision, 92.9% recall	(Baig, 2020)
7	2020	Tamil	comment posts from YouTube	Logistic regression and random forest.	positive is 2,075, negative is 424, neutral is 173	(Chakravarthi B. R., 2020)
8	2019	Sindhi	Social media, newspaper, Sindhi websites	Matrix and TF-IDF models	Preprocessed and normalized	(Dootio, 2021)

9	2018	Saraiki	Essays, newspapers, social media	Named Entity Recognition Resolution,	Not much work is done on Saraiki language	(Saini, 2018)
10	2020	Regional Languages	Majhi dialect and then it has been translated into Saraiki.	Levenshtein algorithm	orthographical similarity of 80.9% between Punjabi and Saraiki	(Khaled, 2020)
11	2020	Saudi Tweets	Arabic tweets	Support Vector Machine (SVM)	91% accuracy	(Almuqren, 2021)
12	2020		dataset gathered from Twitter	hybrid approach of corpus based and dictionary-based techniques	More accuracy in the word list	(Yekrangi, 2021)
13	2021	Saraiki	Newspaper Jhoke	POS tagging NLP tools	Nouns about 1500 extracted from 2 million corpora of Saraiki	(Zamir N. a., 2021)
14	2023	Pashto	News, books, Wikipedia	Gradient descent approach	F1-measure of 93% and 98% accuracy	(Haq, 2023)
15	2020	Brazilian	Set of tweets	Linear SVM, Logistic Regression and Multinomial Naïve Bayes,	0.6527 for Naïve Bayes, 0.7405 for Logistic Regression, and 0.7379 for Linear SVM.	(de Carvalho, 2020)
16	2020	Saraiki	Newspaper, poetry, social media	quantitative corpus-based approach	49.38% with a frequency of 82 verbs.	(Zamir N. a., 2020)
17	2021	Saraiki	Newspaper, dictionary	POS tagging	A well collection	(author= {Saleemi, 2021)

					of hierarchical words	
18	2023	Saraiki	Saraiki dictionary, documents	Rule-based stemming and LSTM model	68.53% accuracy was achieved	(Malik, 2022)
19	2022	Rojak	Bahasa Rojak Crawled Corpus (BRCC).	BERT and XLM as baseline	Best performance in all domains	(Romadhona, 2022)
20	2021	Saraiki	newspapers, stories, essays, and poetry	Expansion approach	Developed wordnet	(Gul, 2021)
21	2017	Telugu	Twitter, news websites and Facebook	SVM	12.3% error rate	(Mukku, 2017)
22	2023	Urdu	Websites Nawa-e-Waqt, BBC Urdu etc.	LSTM	85% on sentence-based, 50% on paragraph-based corpus	(Bashir, 2023)
23	2023	Ethiopian	Facebook and twitter	SVM	86%	(Astarkie, 2023)
24	2023	Arabic	OCA corpus, movies Reviews	Binary Equilibrium Optimization Algorithm	84%	(Rahab, 2023)
25	2023	Arabic-Egyptian	Tweets	Ensemble SVM with LR	92.6%	(Kora, 2023)
26	2023	English	TripAdvisor website	SVM and CNN	0.86 and 0.82 respectively	(Namee, 2023)
27	2019	Punjabi	Newspapers, News items, Novels, Published Books, Poetry	POS tagging	Corpus of 2 million words	(Hashmi, 2019)
28	2023	Telugu	Newspapers	Naive Bayes and Decision Tree	F1 score 0.87 and 0.93 respectively	(kumari Bygani, 2023)
29	2023	Urdu	news, social media, Wikipedia, and historical text	Bayesian Classifiers Chain and Nearest Set Replacement	0.50 and 0.53	(Shafi, 2023)

30	2023	English	online movie reviews and Twitter	Boost, SVM, RF and LR with embedding IWVS	0.65, 0.61, 0.51 and 0.61 respectively	(Samih, 2023)
31	2023	Urdu	Social media	LSTM	87.00%	(Akhtar, 2023)
32	2023	Pashto	text documents	Multinomial Naïve Bayes, SVM and Logistic Regression	0.81, 0.84 and 0.85	(Baktash, 2023)
33	2023	Perso-Arabic languages	Wikipedia, and Newspaper	Hierarchical Modeling with fast Text	F1 score 0.90	(Ahmadi, 2023)
34	2023	Uzbek	News websites, articles, and press releases	CNN and BERT	F1 80.8 and 83.4 respectively	(Kuriyozov, 2023)
35	2023	African	Twitter	AfroXLMR	F1 71.2	(Muhammad S. H., 2023)
36	2023	Uzbek	Twitter and Newspaper	POS-tagger tool	89.78%	(Sharipov, 2023)
37	2020	Tamil-English	Social media comments	Random Forest, Naive Bayes	F1 score 0.65, 0.63	(Raja Chakravarthi, 2020)
38	2020	English and Urdu	NEWS sites, social media	Logistic Regression	F1 0.98	(Asad, 2020)
39	2023	Sindhi	Textbooks of different subjects	n-gram and TF/IDF	Complexities in language	(Talpur, 2023)
40	2022	Tigalari	Social media	Multi-Layer Perceptron, SVM	F1 0.62, 0.60	(Hegde, 2022)
41	2023	Bengali	Dataset available	CLSTM	F1 0.86	(Haque, 2023)
42	2023	English	Airline reviews from Kaggle	BERT, Random Forest	F1 0.82, 0.74	(Patel A. a., 2023)
43	2023	Spanish	Tweets	Neighbor-sentiment algorithm metrics	F1 64.06	(Pilar, 2023)
44	2020	Malayalam-English	Social Media Comments	BERT	F 1 0.75	(Chakravarthi B. R., 2020)
45	2023	Romanized Sindhi	Different online sources	Online Python tool	86% neutral, 9% negative, and 5% positive	(Sodhar, 2023)
46	2023	Thai	Reviews form travel websites	ridge regression, SVM, LR	F1 0.891, 0.896 and 0.896	(Khamphakdee, 2023)

					respectively	
47	2023	Sindhi	Articles on different topics	TPTS approach	F score 89%	(Nawaz, 2023)
48	2023	English	magazine, newspaper, and academic journal	TF/IDF	Recall 88.900%	(Xu, 2023)
49	2023	Urdu	blogs, websites, and tweets	LSTM+GRU, RNN+CNN	84.5% and 85.8% respectively	(Muhammad K. B., 2023)
50	2023	Arabic	LABR and HARD from books and reviews	Word2Vec embedding with LSTM	Dataset1: 94.58% Daataset2: 85.85%	(Elhassan, 2023)
51	2023	Bengali	Comments from YouTube	Bangla BERT	F1 score 0.82	(Hasan, 2023)
52	2023	Persian	social microblogs twitter, Instagram	CNN with fast text	0.72	(Mazoochi, 2023)
53	2023	Kurdish	Twitter	SVM and LR	0.61 and 0.57 respectively	(Hameed, 2023)
54	2023	Persian	Stock market tweets	SVM	F1 score 0.57	(Ahangari, 2023)
55	2022	Croatian	Movies and news articles, reviews	BERT	84.71 accuracy	(Thakkar, 2023).
56	2023	Danish and Norwegian	Novels, Dictionaries	BERTa	F1 score 0.72	(Allaith, 2023
57	2023	Bangladeshi	Reviews from fintech apps	CNN, BiLSTM	F1 score 0.870, 0.970	(Hasan, 2023)
58	2023	Arabic	Arabic stock platform Tadawul	BERT	F1 score 0.87	(Ahmadi, 2023)
59	2023	Hausa, and Igbo	Twitter	AfriBERTa	F1 for Hausa:0.809 and Igbo:0.806	(Muhammad S. H., 2023)
60	2023	Spanish	newspaper articles	BERT	F1 for +ve 75.3, -ve 74.2 and neutral 68.5	(P{\'e} rez, 2023)

Table21:Literature Review

2.1: Summary of Literature Review

In the year 2023, an annotated social media corpus for Finland for sentiments was conducted. In which Finn data was collected by scraping from online web stores. The method used for it was CNN. The paper presents about 27,000-sentence data set that has been neutrally annotated about sentiment polarity by three native annotators. And cross-validation runs were 0.54 for the given dataset on the CNN model. (Krister and Jauhiainen, 2023).

Integrating corpus of Pashto language with POS tagging was developed and determination of integrating Machine Learning was intended to increase precision of computerized POS assignment, speeding up the tagging activity and reducing instruction manual labor. In the recent past, the corpus development pertinent to sentiment analysis also advanced (Haq, 2023).

In 2022, Arabic from Morocco corpus was developed for classification of sentiment. In this study researchers provide a complete dataset that gets the distinctive linguistic and cultural expressions of sentiment expression in Moroccan Arabic. By operating the MAC corpus, researchers can address challenges such as dialectal differences and local sentiment tones in sentiment analysis. The accessibility of corpus promotes cooperation and enables formation of a strong feeling analysis methods for Moroccan Arabic. The MAC corpus contributes to advancements in sentiment analysis methods in the Moroccan perspective and enables a deeper identification of public opinion, market trends, and social media sentiment in the Moroccan Arabic language (Garouani, 2022).

In 2022, a study was conducted about a Nigerian Twitter Sentiment Corpus was created. The first freely available Twitter sentiment dataset that is huge-scale was carefully annotated for the four languages that are most spoken of Nigeria (Nigerian-pidgin) was introduced by the authors of this study as Naijerian Senti. They offer methods that make it possible to gather, filter, and annotate language data with such limited resources (Muhammad S. H., 2022).

A study conducted in 2021 for the development of sentiment analysis for Persian in which data collected from different sources, which was sentiments quantified was used (Hosseini, 2018).

A study for Development of a Corpus containing POS Tagged in Urdu Tweets in 2020 was done. In this study all the tweets in Urdu language were collected to make a corpus with POS tagging. And the approach used was statistical data-driven methods. (Baig, 2020).

The study on the development of the Sindhi corpus is also done. The data is collected from different sources like newspapers, social media, dictionaries, and essays. The idea to create the corpus was for the use of language for machine algorithms. A preprocess and normalized dataset were prepared (Dootio, 2021).

A study on the progress of Saraiki WordNet: A Corpus-based Method in 2021 was held, which based the creation of Saraiki WordNet on the Urdu WordNet. UET Lahore developed Urdu WordNet, which is based on Princeton WordNet. The growth of the Saraiki WordNet has had a significant impact on Natural Language Processing (NLP). For the data ideas, dictionaries or lugats, literary sources like fiction and poetry including non-fictional sources just like newspapers in the Saraiki language, are employed. The Urdu word senses are then plotted against the senses of Saraiki word. The charting procedure in this study uses the expand approach and the mapping strategy (Gul, 2021).

Corpus-based research on the Saraiki language was held in 2020 in which, after text annotation and encoding, the 1 million corpora of words from newspaper of Saraiki Jhoke was created, and a file of 160 verbs was produced. With the use of a device user-friendly dictionary, the lexicon-semantic linkages of verbs were established, and the occurrence of each relationship was discovered using Antconc 3.5.7 by Laurence Anthony. The findings showed that four semantic relations such as synonym, antonymy entailment, were required for building a WordNet and varied in their frequency and percentage (Zamir N. a., 2020).

In 2020 a corpus for the sentiments analysis was created in the Tamil-English text. In which originated a 15,744 YouTube comments are included in the basic Tamil-English code-transferred, sentiment-annotated corpus. The procedure for building the corpus and assigning polarity was described in that study. The agreement of Inter-annotator is also displayed, and the SA

findings using dataset as a benchmark are exhibited (Chakravarthi B. R., 2020).

SentiBahasaRojak: The First Bahasa Rojak Corpus for Pertaining will be released in 2022 by BRCC and was conducted. In which data was collected from Bahasa Rojak Crawled Corpus (BRCC). The XLM and BERT were chosen as the baseline were used for dataset annotation and process. Best performance was achieved in all domains (Romadhona, 2022).

In 2023, Rule-Based and LSTM Based Sequence-To-Sequence Model Approach for Saraiki Language Hybrid Stemmer is conducted. In which data was collected from the Documents and a dictionary in Saraiki are accessible on an alternative website. An LSTM sequence to sequence model and a rule-based stemming module hybrid models were used. For that model 68.53% accuracy was achieved (Malik, 2022).

In 2021, An Elementary Parts of Speech (POS) corpus for the Saraiki language's morphological, syntactic, and lexical annotations was conducted. The data was collected by Newspaper, dictionary. POS tagging was used for corpus development. A well collection of hierarchical words was obtained (author= {Saleemi, 2021).

An Automated Corpus Explanation for Sentiment Analysis in Brazil in 2020 was done. Dataset was collected as a collection of tweets was made at a certain Recife location. Linear SVM, Polynomial Naive Bayes and Logical Regression were used for sentiment analysis. Accuracy for respective methods was the values for Logistic Regression are 0.7405, Multinomial Naïve Bayes is 0.6527, and Linear SVM is 0.7379 (de Carvalho, 2020).

In 2023, the study on Automatic POS Tagging Using Pashto Corpus and Machine Learning was conducted. Dataset was collected from News, books, and Wikipedia. For the purpose Gradient descent method by using the L-BFGS was used. Accuracy was 98% and F1-measure of 93% (Haq, 2023).

In 2021, a study on A corpus-based study was conducted to examine the semantic-lexicon associations of the nouns used in the Newspaper of Saraiki. was conducted. For that purpose, the data of Saraiki language was gathered from Newspaper Jhoke. POS tagging NLP tools were used for corpus development. Therefore, 1500 nouns were taken from the 2 million Saraiki corpus. (Zamir N. a., 2021).

In 2020, research on Sentiment analysis of financial markets: creating a specialized Lexicon was done. For the proper development and annotation, the dataset gathered from Twitter. A hybrid approach of corpus based, and based on dictionary techniques were applied. As a result, a better grade of accuracy in the word list was generated (Yekrangi, 2021).

In 2020, AraCust: Sentiment analysis corpus on a Saudi Telecom Tweets was developed. Dataset was collected from Arabic tweets. Support Vector Machine (SVM) was used for the sentiment analysis from the Arabic dataset. 91% accuracy was obtained from the dataset with the AraCust (Almuqren, 2021).

In 2020 will be the year that Punjab province studies the orthographic differences between the Punjabi language and the Saraiki dialect. was conducted. In this study Majhi dialect and then it has been translated into Saraiki. Levenshtein algorithm was used for that procedure. The orthographical similarity was 80.9% between Punjabi and Saraiki (Khaled, 2020).

In 'Saraiki NLP': A Comprehensive Meta Analytical Study of Its History, Development, and Evolution in 2018, there was a proper early work discussed about the Saraiki language. Resolution, NER, and Word-Sense-Disambiguation. Authors concluded that not much work is done on Saraiki language (Saini, 2018).

In NaijaSenti: A Multilingual Sentiment Analysis Sentiment Corpus for Nigerian Twitter, 2022. Annotated tweets were used for the process. There were various machine learning models and NLP employed. The performance by methods was 41.8%, 20.8%, 4%, and 19.1%. (Muhammad S. H., 2022).

To enable Telugu Sentiment Analysis, authors of a publication outline an effort to provide a Telugu annotated corpus of the highest caliber utterances. The ACTSA (Annotated Corpus for Telugu SA) dataset had a set of Telugu phrases that were gathered from various sources, pre-processed, and then manually marked up by native Telugu speakers in accordance with our annotation rules. They have labelled 5410 texts altogether, making dataset the largest source at that time. The corpus

and annotation rules are made available to the general audience (Mukku, 2017).

In 2023, research is concentrated on the collection of data in the pure Urdu language, data preprocessing, feature extraction, and novel sentiment analysis techniques. After reviewing previous research, ML and DL methods were used to data. This research analyzed the outcomes and proposed hybrid techniques, which creates new opportunities for sentiment analysis on Urdu-language data.

(Muhammad K. B., 2023).

A paper examined the writers' opinions on the subject and user comments of Ethiopia in 2023 using Facebook as a platform that was manually marked by Telugu native speakers. This study uses a Facebook application to analyze sentiment analysis and opinion mining of user comments and posts on social media are conducted using NLP and ML derived from a novel sentiment analysis and opinion mining framework. SVM was applied for classification, and the obtained accuracy was 86% (Astarkie, 2023).

This study of 2023 suggested a new binary equilibrium optimization metaheuristic algorithm along using an Arabic sentiment analysis technique based on classification rules as an optimization strategy for generating classes rules from Arabic-language records. The Opinion Corpus for Arabic (OCA) was used to test the suggested method, which produced a class-based technique. The consequences of the comparison with cutting-edge techniques demonstrate that the suggested method exceeds all previous white-box models in terms of classification accuracy. (Rahab, 2023)

Research has been conducted which made the impacts: First, they proposed a meta-collective DL method for improvement the execution of SA. In technique, they trained standard methods with use of levels of meta-learners. Second, they also suggested using the "Arabic-Egyptian Corpus 2" dataset as an expansion of already made corpus. Ten thousand more annotations have been added to the corpus, increasing its size by conversational on many topics. There were conducted numerous tryouts on six benchmarks of SA (Kora, 2023).

A study in 2023 was conducted on Sindhi. An online Python application was utilized in this study to parse text and produce outcomes. The analysis's findings showed that 9% of the findings had negative sentiments, 5% of the phrases had neutral sentiments, and 86% of the sentences had neutral sentiments. Based on scraped values, the RST's precision was calculated and found to be 87.02% accurate. Based on the accuracy discovered using the online confusion matrix calculator, an error ratio of 12.98% was estimated (Sodhar, 2023).

In a paper of 2023, an aspect-based technique on reviews of customer was presented. Studied hotel features included courteous service, spotless rooms, good value, and convenience. Their suggested approach gave a summary of opinions of customer created on score and explained why guests liked or disapproved of various hotel features. A state-of-the-art model was compared with the suggested method's best model (Namee, 2023).

A study on Text pre-processing Techniques for Sindhi Language in 2023 was organized, the Sindhi language text preprocessing model TPTS is introduced. For the language of Sindh, TPTS completes crucial NLP responsibilities such text tokenization, standardization, removal of stop-word, stemming, and POS tagging.

For testing, 1.5k Sindhi text documents from several online news websites were combined to form the Sindhi Text Corpus (STC). To find stop-words with high-frequency in Sindhi language, the TF-IDF technique was used. Additionally, in Sindhi input text, a rule-based algorithm tags words with parts of speech were used. The suggested TPTS approach achieves 89% accuracy on the STC corpus when evaluated using the ROUGE evaluation metric (Nawaz, 2023).

An article of 2023, in which a sentiment analysis model was constructed using natural language processing (NLP) technology. The weighted average technique was employed in that article's suggested enhanced TF-IDF algorithm to assess the emotional significance of each emotive word. To determine the emotional value and tendency of the English corpus, inspirational words were employed. The findings demonstrate that the model is highly operational and classificational accurate when choosing feature words and the recall obtained by TF-IDF was 88.900% (Xu, 2023).

A study of Lexicon-Semantic Relationships of Punjabi Shahmukhi Nouns was conducted. The study is important and helpful in creating Punjabi Shahmukhi WordNet. With growth of WordNet, it is feasible to use digital tools like machine translation, information retrieval, report production, archive querying, speech recognition, data mining, read aloud, robotics, and many other in Punjabi Shahmukhi. On the other hand, WordNet supports Punjabi Shahmukhi's continued international recognition. So, the 2 million words corpus was established (Hashmi, 2019).

In 2023, A study on Telugu News based on Sentence Classification with SA was presented. In which they introduced classification, which was two-phase, method for Telugu news words using Telugu sentiment analysis. It first acknowledges the classification of subjectivity, which labels claims as 0, 1, -l. The second phase categorizes subjective statements into groups that are either good or negative. Using this technique, we get a sentiment analysis categorization accuracy of 81 percent. In which naïve bayes and decision tree were applied and got the F1 score 0.87 and 0.93 respectively (kumari Bygani, 2023).

Research on Based on Word Embeddings and Deep Learning on Sentiment Analysis of Arabic was presented in 2023. In which the Arabic dataset was collected from LABR and HARD from books and reviews. In that study different embeddings were applied with deep learning methods. And Word2Vec embedding with LSTM was applied for Arabic dataset. For two different datasets the acquired accuracy was Dataset1: 94.58% Daataset2: 85.85% (Elhassan, 2023).

A paper in 2023 offered a sizable corpus and approaches for Urdu language's tagging of semantic challenge. The suggested corpus has 8,000 tokens, with 2K tokens in each of the genres or domains of historical text, Wikipedia, social media, and news. The corpus has been manually labelled with 21 major semantic fields and 232 sub-fields using the USAS (UCREL Semantic Analysis System) semantic taxonomy, which provided a full range of semantic fields for coarse-grained annotation. Every word in our suggested corpus has been annotated with a minimum of one and maximum of nine semantic field tags to give a comprehensive semantic analysis of the language data. This has made it possible for us to tackle the problem of semantic tagging as a multi-target classification work under supervision. From the recommended corpus, they collected local, topical, and semantic features and used them as input for seven distinct supervised multi-target classifiers. Findings show that our suggested corpus, which is openly available for download, is 94% accurate. (Shafi, 2023).

In a paper, based on improved word-embeddings and XGboost in 2023, an innovative technique called Improved Vector of Words for Sentiment Analysis was proposed. XGboost to raise the sentiment categorization F1-score. The suggested technique, known as Sentiment2Vec, created sentiment vectors by averaging word embeddings. Additionally, they examined the Split language for categorizing good and negative emotions. Using several machine learning techniques and datasets of sentiments, we compared the F1-score of sentiment categorization using our technique. IWVS outperforms. When we evaluate the quality of our proposal with baseline models, we find that Doc2vec and frequency-inverse document frequency (TF-IDF) on the F1-measure for sentiment classification are the most effective. XGBoost with IWVS features was the best model in our study at the same time (Samih, 2023).

Research in 2023 was done for SA. That study provided a novel framework for categorizing emotions conveyed in Urdu. The study's primary contributions were to draw attention to the importance of this multifaceted research problem. as well as expert components, involving the corpus, and algorithm of parsing. The suggested work was thoroughly compared to the baseline standard approach in the results. For the Urdu dataset LSTM was applied and the accuracy achieved was 87.00% (Akhtar, 2023).

A paper in 2023 for Pashto language for Creating a Pashto automatic text classification system is the goal of that work. They created a Pashto dataset, which was group of Pashto tesxt, Additionally, to determine the most efficient method, study compared several models that incorporate Machine learning methods based on neural networks and statistics, including logistic regression, gaussian nave Bayes, multinomial nave Bayes, decision trees, Multilayer Perceptron (MLP), Support

Vector Machine (SVM), and K Nearest Neighbor (KNN). Additionally, this study assessed two distinct feature extraction techniques, namely unigram and IF/IDF. Consequently, this study's MLP classification algorithm and TFIDF feature extraction method yielded an average testing accuracy rate of 94% in that situation (Baktash, 2023).

A study on language of Uzbek, as part of the text categorization process, author examined the dataset creation procedures and evaluation methods. First, they provide a recently acquired dataset for the classification of Uzbek texts, which was gathered from ten different news and press websites and includes texts in 15 different kinds related to the news, and legislation. Then produced dataset, they also gave a thorough assessment of various models, models to deep learning architectures, that has range from conventional bag-of-words. Their test demonstrated that the rule-based models are outperformed by models built using convolutional and recurrent neural networks (RNN and CNN, respectively). Trained on the Uzbek corpus, the BERT model is a transformer-based BERT model, has the best performance (Kuriyozov, 2023).

A study in 2023 was conducted in which author offered AfriSenti, a sentiment analysis platform for 14 African languages of tweets, including, Native speakers had annotated the following languages: Oromo, Swahili, Tigrinya, Twi, Xitsonga, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, and Swahili, and Yorùbá from families of language. The information was utilized in SemEval 2023 job 12, the initial shared SemEval job. They outline the procedures for gathering data, the annotation process, and any associated difficulties encountered when curating each dataset. The method applied was AfroXLMR and F1 score was 71.2. (Muhammad S. H., 2023).

In 2023, A part-of-speech (POS) annotated dataset and tagger tool were presented for the low-resource Uzbek language, together with the rule-based POS tagger was presented in that research work. A POS-tagger of rule-based tool was developed with use of dataset's 12 tags. To guarantee its representativeness, it must be balanced over 20 distinct fields. The tagger tool demonstrated great accuracy in recognizing and classifying elements of speech in Uzbek text when tested on the annotated dataset. Several natural languages processing applications, including language modelling, machine translation, and text-to-speech synthesis, can be performed using the recently released dataset and tagger tool (Sharipov, 2023).

A Corpus Tamil-English Text was created in 2020. In which corpus of mixed script Tamil-English was created by authors. For which, the dataset was collected from social media. Different NLP methods were applied to the dataset. Good accuracy was achieved on two methods which were Random Forest and Naive Bayes. The F1 score was 0.65 and 0.63 respectively (Raja Chakravarthi, 2020).

A paper on the topic Sorting News Articles using ML Approach in 2020 was published. The dataset used in both English and Urdu text. The data collected from e-newspapers, articles, and social media. For the evaluation Logistic regression was applied to dataset and got F1 score 0.98 (Asad, 2020).

In a study, a technique for categorizing Thai sentiment in the hotel industry using sentiment analysis was proposed in 2023. To construct word embeddings with various vector dimensions, the technique of Word2Vec (using the continuous bag of words (CBOW) and skip-gram approaches) was first implemented. Second, to see how every word vector dimension outcome affected the final model, embedding model of every word was merged with DL methods. Pre-trained models of BERT were used to perform grouping on the dataset. Finally, their research demonstrates that the Wangchan BERTa model marginally enhanced accuracy. With an accuracy of 0.9170, the skip-gram and CNN model combo outperformed other DL models, yielding a value of 0.9225. They discovered from the trials that the performance of sentiment classification was impacted by the word vector dimensions, hyperparameter settings, and DL models' layer count (Khamphakdee, 2023).

In 2023, Linguistic analysis and building a corpus from a primary school Sindhi language textbook was done. In that research a Sindhi corpus was developed. The data was collected from Textbooks of different subjects like English, Urdu, history, and math. On that language n-gram and TF/IDF

was performed. The authors concluded that there are too many complexities in the language and for classification there should be large dataset (Talpur, 2023).

A study was conducted in 2022 in which Tulu text was discussed. The study was done in Tigalari language. For that purpose, the dataset was collected from social media. A 7,171 YouTube comments gold standard trilingual code-mixed Tulu annotated corpus is produced. The methods applied were Multi-Layer Perceptron, SVM and F1 score was F1 0.62, 0.60 respectively (Hegde, 2022).

In 2023, authors conducted a supervised DL classifier. The study intended to give a comparison analysis with the baseline models and maximize accuracy utilizing the proposed model. As baseline models, six machine learning models with two different feature extraction techniques were considered. The authors proposed that CLSTM architecture can greatly outperform SA with 85.8% accuracy and 0.86 F1 scores on a labeled dataset of 42,036 Facebook comments. A web application based on the recommended model and the best baseline model was developed to determine the true sentiment of social media comments (Haque, 2023).

In 2023, A work on conducting SA was put out in a study. Data for databases is sourced from KAGGLE. Several machine learning (ML) methods are used, Logistic Regression. Their backing was to check the behavior of BERT design. The "Random Forest" is used as a benchmark to compare the results of the "BERT Model" because it outperforms all other machine learning models. The better outcomes were got from BERT, Random Forest which were F1 0.82, 0.74 (Patel A. a., 2023).

In 2023, A method for Spanish tweets was conducted. On the reviews, a clustering genetic algorithm extracted feature. The algorithm has been used to categorize a set of 1,899 evaluation reviews from two Spanish-language corpora consisting of 3,413 and more than 63,000 tweets, respectively. Precision was employed to evaluate the results. In comparison to earlier literature efforts, the algo has increased the outcomes for metrics and on both corpora, attaining a M F1 of 0.640 and an accuracy of 0.689. The main qualitative advancement of the classifier has been the flexibility feature in extraction of feature (Pilar, 2023).

A paper introduced a brand-new dataset having Malayalam-English texts that has been annotated voluntarily in 2020. That dataset achieved a dataset of greater than 0.8. The standard for SA in texts with hybrid Malayalam-Eng coding is provided by that new dataset. The dataset was gathered from social media like YouTube, Facebook, and their comments. The BERT was applied on that dataset and acquired F1 was 0.75 on Malayalam-English texts. (Chakravarthi B. R., 2020).

A study on A Language Identification Benchmark for Perso-Arabic Scripts was done in 2023. The research clarified the difficulties in identifying languages written in Perso-Arabic scripts, particularly in multilingual settings when "unconventional" writing is used. To solve this problem, authors classified sentences into their respective languages using a variety of supervised techniques. The Arabic dataset was collected from Wikipedia, Tatoeba and SETimes. Hierarchical Modeling with fast Text was evaluated on dataset. The F1 was 0.90 (Ahmadi, 2023).

In a study of 2023, authors recommended a dataset for sentiment analysis of the recent crisis between Russia and Ukraine that has annotations in Bangla. The dataset was created by compiling comments about Bangla from various videos of three recognized Bangladeshi TV report on channels the present conflict. 10,861 comments in Bangla were totaled. By testing with various transformer-based language models that were all previously trained on unlabeled Bangla corpora, a benchmark classifier was created. The models were fine-tuned using the dataset we purchased. All transformer language models Bangla BERT, XLM-RoBERTa-large, DistilmBERT, and mBERT were subjected To hyperparameter optimization. Several assessment criteria, such as the F1 score, accuracy, and AIC, were used to evaluate and analyze each model. The model Bangla BERT performed the highest had a maximum accuracy of 86% and an F1 score of 0.82 (Hasan, 2023).

A study on Building Persian dataset in 2023 built the ITRC-Opinion user opinion dataset from scratch using a collaborative setting. 60,000 casual and everyday Persian writings from social

microblogs made up dataset. Second, a fresh deep convolutional neural network (CNN) model is put forth in study for a more successful sentiment analysis of informal text. Additionally, several models, including Fasttext, Glove, explored dataset and assessed the outcomes.Results show that dataset and the suggested model are advantageous (72% accuracy) on CNN model with fasttext. (Mazoochi, 2023).

A study on Kurdish language was done in 2023. In study discussion about Kurdish dataset was made. For this purpose, authors investigated a few traditional methods. They also used a learning about transfer strategy to leverage pre-trained. They showed that augmentation of data, despite the task's difficulty, achieves a high F1 score and accuracy. The better results were acquired on SVM and LR methods that were F1 score 0.57 and 0.61 (Hameed, 2023).

A study in 2023 which tells about sentiments analysis wase performed on Persian language, which is spoken in Iran. The dataset was collected from Stock market tweets. The SVM method was applied to the dataset. which acquired the F1 score of 0.57 (Ahangari, 2023).

In 2023, a study was done on the topic A Sentimental labelled Dataset about Reviews of Film. In which a Croatian movie review dataset with sentiment annotations. Over 10,000 sentences make up the dataset, which has been annotated at the sentence level. Authors offered both transformer-based fine-tuning approach and the overall annotating process (Thakkar, 2023).

In a study of 2023, in which a dataset in Danish and Norwegian was collected from dictionaries and novels. To get good results BERTa was applied to the annotated corpus. After that got the F1 score 0.72 on created dataset (Allaith, 2023).

A study on of User Reviews of Bangladeshi Mobile Services, primarily focused on the responses of these apps' users. Based on the users' published feedback, sentiment analysis is being used to elicit their feelings. Examining the perspectives of these application users is the main objective of this article. From the Google Play Store, 5414 different bits of information were gathered and categorized as either negative, neutral, or good. Using the CNN, LSTM, and BiLSTM algorithms, the data model has been assessed. The BiLSTM method created a model with accuracy of 97.07% when compared to CNN and LSTM (Abdullah Al Ryan, 2023).

BERT model and the logistic regression model are the foundations of the Arabic SA technique presented in a recent paper of 2023. To highlight the polarity, the economic justifications for the articles based on semantic units were separated into seven economic aspects. According to SA, the supervised BERT model classified articles with an accuracy of 88%, and the unsupervised mean Word2Vec encoder clustered economic aspects with an accuracy of 80% (Alasmari, 2023).

In a study, authors focused on the problem of targeted sentiment analysis for news headlines on the 2019 Argentinean Presidential Elections that were released by significant news outlets. Authors offered a polarity dataset with 1,976 headlines that mention candidates at the target level to aid study in this area. Their tests showed the value of target information for this task when utilizing cutting-edge classification algorithms built on previously trained language models. To encourage additional research, they also made data and models accessible to the public (P{\'e} rez, 2023).

3. Proposed Methodology

In the proposed methodology the first task is to create a Saraiki dataset. This purpose, the Saraiki sentences are collected from different sources like Saraiki website named “Saraiki News.com”, Facebook and Books. When collection of data is done then sentiment labeling is done as positive, negative, and neutral. This annotation of dataset is done properly by experts of this field. Then the methods of preprocessing, feature extraction and classification are applied. After that the evaluation of methods are compared.

POLICY RESEARCH JOURNAL
ISSN (E): 3006-7030 (P) 3006-7022

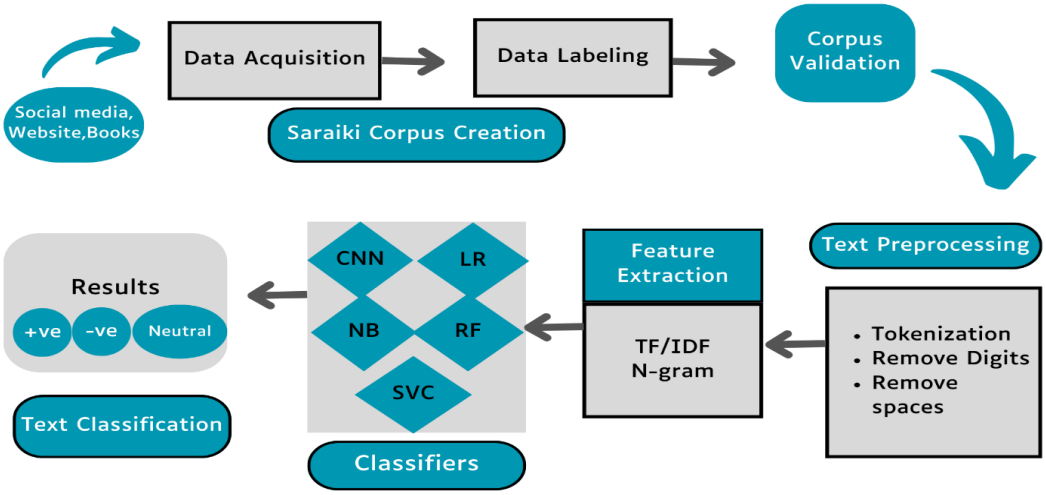

Figure 3.1: Proposed Methodology

3.1 Dataset Acquisition

For Saraiki language the first step is to collect data. As there is no corpus of Saraiki language existed before, the data is collected by different platforms. In the data acquisition, the sentence base data of Saraiki language is collected from different sources. Different sources are used for data collection like Facebook posts, books, and Saraiki website. First, the random Saraiki data in paragraphs is collected and then sentences are formed from them.

3.2 Annotation of Sentences

A sentence with no positive and negative sentiments is then labelled as neutral. Then these sentences are manually annotated. Python has a variety of labeling methods; however, labeling can also be completed by hand. When compared to Python libraries, manual labeling turns out to be more accurate because human annotation techniques incorporate deeper terms. The sentences are properly labelled as positive, negative, and neutral sentences based on sentiments. If a sentence has positive sentiment, then labelled as positive by giving 1 value. If a sentence has negative sentiment, then labelled as negative by giving it 1 value. The dataset is properly validated by the experts of the field.

3.3 Preprocessing

After the acquisition of text, the further step is to preprocess the collected dataset. So, for this purpose first, Saraiki sentences are preprocessed. Pre-processing starts with converting text documents into a word format that can be read and then it is performed which includes the following:

3.3.1 Tokenization

Tokenization is applied to the data. The sentences are tokenized into words by using spaces.

Sentence	میں اجڑ گیا تاں کوئی گل نئیں						
Tokens	میں	اجڑ	گیا	تاں	کوئی	گل	نئیں

Table 3.1: Saraiki Tokenization Example

Saraiki Stop Words Example		
تے	ہک	اچ
اے	کوں	انہاں
تاں	نے	کر

Table 3.2: Saraiki Stop Words Examples

3.3.2 Stop Words Removal

There are some stop words in every language that are commonly used so also in Saraiki language. So, they are removed. There were some extra spaces also which were then removed. Unimportant terms must go because stop words are frequently used. Basically, stop words are those words whose removal does not affect your actual dataset and are not necessary.

3.3.3 Elimination of Special Characters

If there are some special symbols, simile, or any other marks like #, @ etc. in the sentences then they are removed.

3.4 Feature extraction

Lately, embedding techniques have proven more effective than conventional text-feature extraction techniques like bag of words. New words that are absent from training texts can be categorized using keywords that are like each other. Numerous words embedding methods exist, including Word2Vec, Glove, and BERT (Elhassan, 2023). A document is first considered a string, and then it is processed into a list of tokens.

3.4.1 Term frequency-inverse document frequency (TF-IDF)

It is a technique of rating the significance of words in a text according to how often they show up in different documents (Kashif, 2019). In Saraiki language corpus for sentiment analysis, TF/IDF is used for feature extraction.

3.4.2 N-Gram Features

The fundamental elements frequently employed in sequence-based malicious code detection techniques in computer virology research are called n-grams. However, as an n-gram gets longer, the feature space expands exponentially, requiring a large amount of storage and processing power (Hamid Parvin, 2020). In N-gram there are unigram, bigram, and trigram. So, in unigram there is single words are taken. In bigram there are combination of two words taken and in trigram combination of three words are taken

N-Gram	Generated Sentences	Number of N-gram features
Unigram(1-Gram)	"کمپنی", "شدید", "مالی", "بحران", "توں", "ڈوچار", "ھِ"	7
Bigram(2-Gram)	"کمپنی شدید", "شدید مالی", "مالی بحران", "بحران توں", "توں ڈو چار", "ڈو چارھِ"	6
Trigram(3-Gram)	"کمپنی شدید مالی", "شدید مالی بحران", "مالی بحران توں", "بحران توں ڈو چار", "توں ڈو چارھِ۔"	5

Table 3.3: N-gram Examples

3.5 Classifiers

There are five classifiers which are applied on created Saraiki Dataset.

3.5.1 CNN

Among the applications of the CNN, a kind of neural architecture, are face detection, object detection, picture segmentation, and image processing. The CNN model outperforms conventional ML algorithms in sentiment categorization when used. It can shorten execution times while detecting intricate elements in the data (Khamphakdee, 2023). One of the best techniques for grouping is the convolutional neural network (CNN), which includes a convolutional layer for extracting information from longer text. The convolution layer system based on byte n-grams are the three main layers that make up the CNN

POLICY RESEARCH JOURNAL
ISSN (E): 3006-7030 (P) 3006-7022

model. Text is one dimensional, in contrast to visuals. Therefore, one-dimensional convolutions are used to the word embeddings rather than two-dimensional convolutions. To add non-linearity, a non-linear activation function is added after convolutions. One or more totally connected layers function as the classifier to translate the features to the desired output classes (e.g., positive, negative, and neutral) once the vector has been flattened.

C_a = The a^{th} feature map.
F = An activation function like ReLU
W = The filter
b= The bias term.

The following can be used to carry out convolution to classify text:

$$Ca = f(W \cdot Xa{:}\, a + h - 1 + b) \qquad (3.1)$$

$X_{a:\, a+h-1}$ = The word embeddings' subsequence from location from a to a+h-1

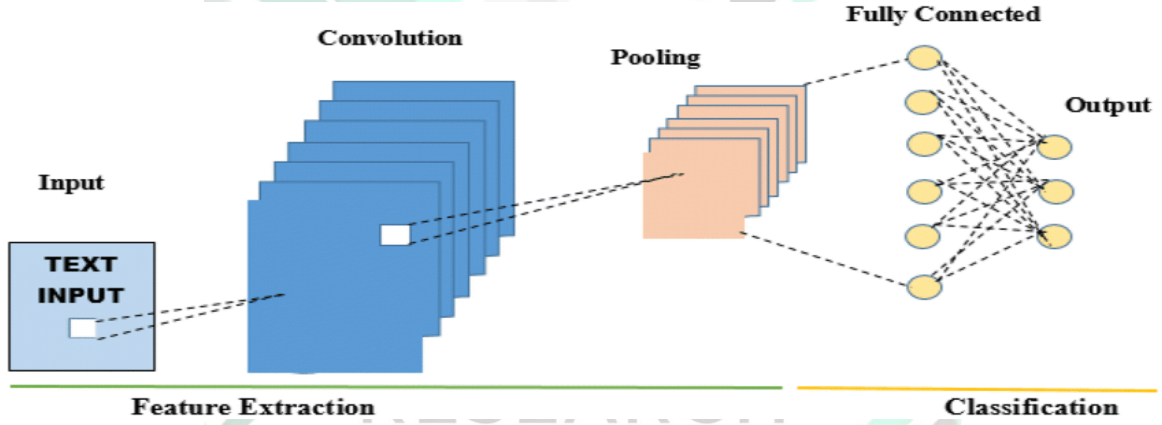

Figure 3.1: CNN working.

3.5.2 Logistic Regression

One type of supervised learning problem is logistic regression. This is a binary classification algorithm. This classifier is a statistical technique for dataset analysis. To determine the best fitting line and forecast a binary result, use logistic regression (Sandeep Nigam, 2018). Sentiment analysis and other text categorization tasks use the flexible statistical model known as logistic regression. To classify text features into sentiment classes like positive, negative, or neutral, it converts the features into numerical representations and fits them to a logistic function. Even though it is straightforward, Logistic Regression is highly appreciated for its efficiency and readability, which makes it especially helpful for smaller datasets or situations with controlled processing sources. It could, however, not be able to represent complicated relationships in data as well as more sophisticated models like neural networks.

3.5.3 Random Forest

Tree-based models are the foundation of Random Forest. According to a certain criterion, a tree-based model divides the provided dataset into two separate groups iteratively until the predetermined stopping point is attained. Leaf nodes, also referred to as leaves, are the terminal locations where these decision trees culminate. These leaves stand for the results or forecasts that result from the characteristics and circumstances that are met during the recursive partitioning procedure. Regression and classification issues are both addressed by it (Poulami Basu, 2024).

Based on collaborative learning, random forest is kind of managed machine learning technique. Distinctive algorithm types or iterations of algorithm which are same, are joined to create an extra potent model of prediction. The name 'Random Forest' comes from the algorithm of random forest, which links many algorithms which are of the similar type, i.e. multiple DT, for creation of forest of trees.

POLICY RESEARCH JOURNAL
ISSN (E): 3006-7030 (P) 3006-7022

3.5.4 SVC

Because Support Vector Classification (SVC) can efficiently categorize data points in high-dimensional spaces, it is a reliable technique for sentiment analysis of text corpora. SVC is used in sentiment analysis to choose the best hyperplane to divide various sentiment classes by first converting textual information into numerical representations like Bag of Words or TF-IDF vectors. The regularization parameter and kernel selection are two examples of the hyperparameters that affect the model's efficacy and need to be carefully adjusted.

Naive Bayes executes calculations quickly, the SVM approach takes longer. Compared to SVM, naive bayes is less accurate because it only needs a minimal quantity of data. This demonstrates that data models pertaining to presidential elections and text classification based on social media may benefit more from the SVM approach (Asno Azzawagama Firdaus, 2023).

3.4.5 Naive Bayes

Assuming, the Naive Bayes classifier that features are independent, which makes calculating probabilities easier. Standard metrics like F1-score and accuracy are used by the classifier to assess its performance on a validation set after training. By defining the next probability of each sentiment group by choosing the highest probability class, the trained model infers the sentiment of fresh text samples during inference. The simplicity, efficacy, and efficiency of naive Bayes classifiers are highly appreciated, specifically in text classification tasks; yet their accuracy in capturing intricate feature interactions may be inferior to that of more improved models.

The Naive Bayes classifier, founded on the Bayes theorem, is a basic probabilistic classifier. Multinomial Naive Bayes works well in situations when the classification problem contains many word occurrences. Bernoulli Naive Bayes can be used when a specific word is not present (Sandeep Nigam, 2018)

$$p\left(\frac{A}{B}\right) = \frac{p\left(\frac{B}{A}\right) * p(A)}{p(B)} \quad (3.4)$$

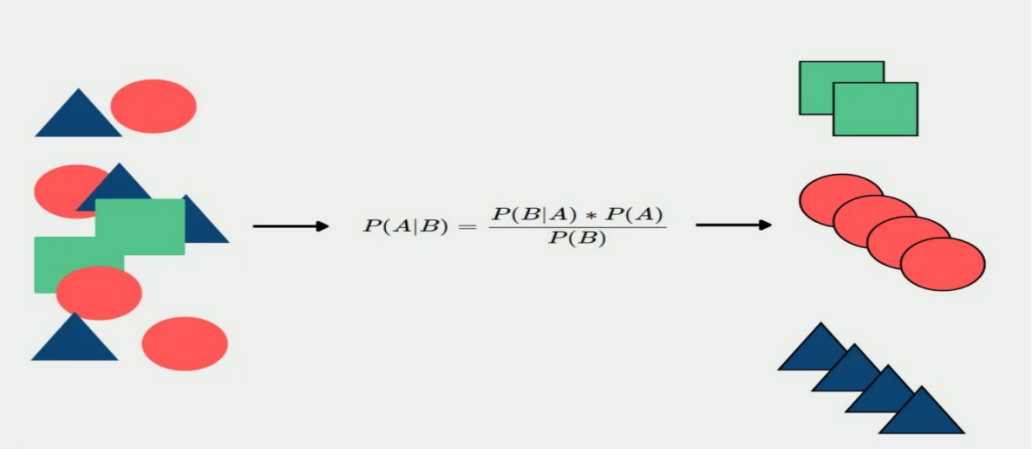

Figure 3.2: Naive Bayes Working

4.Results and Discussion

In the chapter the scores and discussion are posed related to evaluation of classifiers applied to Saraiki corpus.

4.1 Saraiki Language Corpus

The first step was to create a proper dataset of Saraiki. As there is not much work done on this language, so this was important to collect the sentences from different sources. So, we have collected the Saraiki data from Saraiki website, Facebook, and different Saraiki books. First data was randomly collected. Then we form them into a meaningful form by arranging the data in the proper sentences. When the data was fully collected then we complete the labeling as positive, negative, and neutral sentences. The sentence that seems to be happy or good is labelled as positive. The sentences that look bad or angry are then labeled as negative. Others are neutral considered.

The dataset validation is done by the people who are experts in this field. After that Saraiki corpus fully created for the next processing.

Sr	Saraiki Sentences
1	چینی حکام نے پاکستانی سفارتخانے توں خصوصی ٹاسک فورس کوں ووہان ونجڻ دی اجازت ڈے ڈیتی
2	زندگی دے تنگ حالات دے وچ بندے چنگے چنگے پھسدے
3	تمام بھیراواں دی محبتاں دا مشکور ہاں
4	عید تے قرآن مجید دی بے حرمتی مسلم امہ کیتے کہیں صورت اچ قابل قبول کینی،
5	ساکوں رول کے گلیاں وچ میکوں ڈس تاں سہی کہڑا بخت لگئ

Table4.1: Saraiki Dataset

4.2 Preprocessing

When the dataset acquisition is completed then the next step was to preprocess the Saraiki dataset. which includes cleaning, tokenization, removing white spaces, removal of special characters and removal of digits. The file of 7000 Saraiki sentences is tokenized. After tokenization the total 151829 token were extracted. In tokenization a sentence is divided into many tokens or words. Data cleaning is performed. All the extra spaces, digits, stop words are removed from dataset to reduce ambiguity from corpus. So, these tasks are performed correctly.

4.3 Feature Extraction

Feature Extraction on Saraiki dataset is performed to extract specific features from data. For this purpose, the technique TF/IDF is applied on our Saraiki dataset. In feature extraction TF/IDF Vectorizer is applied to dataset and 5 rows ×1000 columns are extracted. Moreover, N-gram also applied on Saraiki dataset that showed required results for method implementation.

1.4 Results

When the data is collected then the dataset is properly formed for the experiments. After that, different methods of preprocessing are applied to dataset for initial work. When dataset is in machine readable form then methods are applied for sentiment analysis from extracted dataset after preprocessing. Then the sentences are classified as positive, negative or neutral sentiment.

Here are the results after applying SVC, Logistic Regression, Random Forest, CNN, and Naïve Bayes to Saraiki Language Corpus for sentiment Analysis.

4.4.1 SVC

The support vector classifier is applied to the Saraiki language corpus. And here are the results of this classifier given in a chart. In which appears the actual and predicted positive, negative, and neutral text. On positive there is higher accuracy.

POLICY RESEARCH JOURNAL
ISSN (E): 3006-7030 (P) 3006-7022

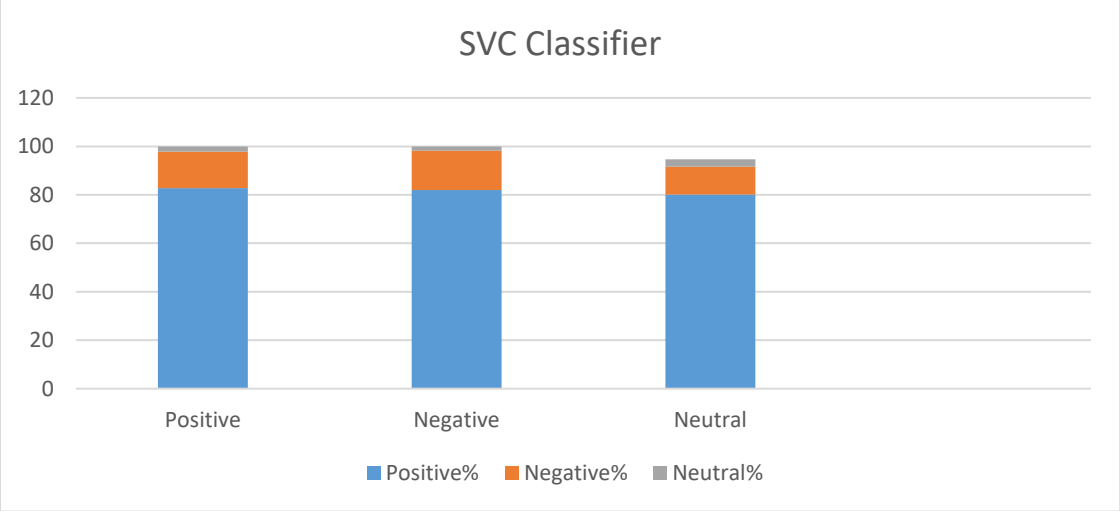

Figure 4.1: Bar chart of Support Vector Classifier

4.4.2 Logistic Regression

The Logistic Regression classifier is applied to the Saraiki language corpus. And here are the results of this classifier given in the bar chart. Which indicates the actual and predicted positive, negative, and neutral text. On positive there is higher accuracy. Also, the negative and neutral are predicted but results are rather good on positive.

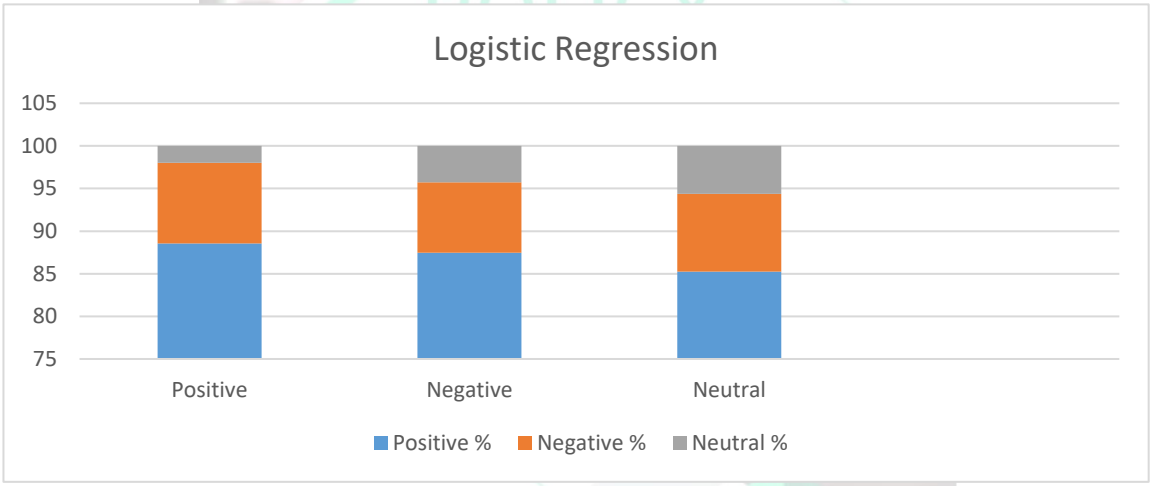

Figure 4.2: Bar chart of Logistic Regression

4.4.3 Random Forest

The Random Forest classifier is applied to the Saraiki language corpus. And here are the results of this classifier given in the form of a bar chart.

POLICY RESEARCH JOURNAL
ISSN (E): 3006-7030 (P) 3006-7022

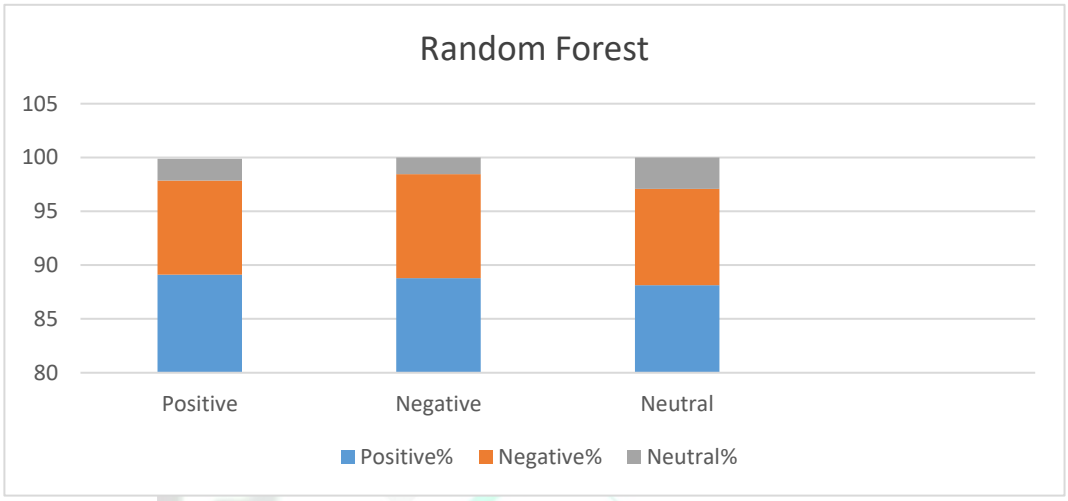

Figure 4.3: Bar chart of Random Forest

4.4.4 CNN

The CNN classifier is applied to the Saraiki language corpus. And here are the results of this classifier given in a bar chart. Which show up the actual and predicted positive, negative, and neutral text. On positive there is higher accuracy.

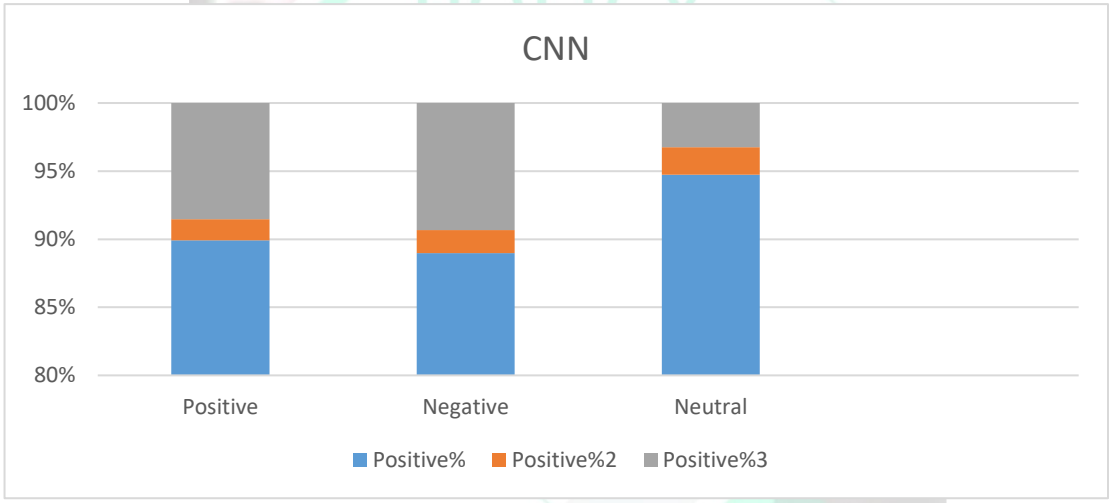

Figure 4.4: Bar chart of CNN.

4.4.5 Naïve Bayes

The naïve bayes Method is employed to the Saraiki language corpus. And here are the results of this classifier given in the shape of a bar chart. Which displays the actual and predicted positive, negative, and neutral text. On positive there is higher accuracy.

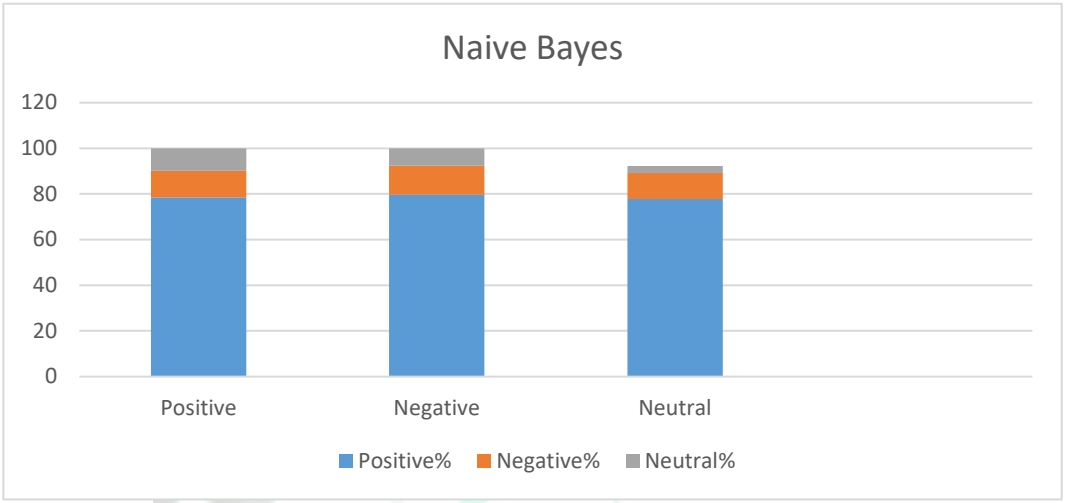

Figure 4.5: Bar chart of Naive Baye

4.4.6 Comparison of All Classifiers

After applying all the classifiers SVC, CNN, Naïve Bayes, Logistic Regression and Random Forest at the end all these classifiers are compared that which one is performing well on Saraiki Language for sentiment analysis. Just looking at the percentages, we can see that the CNN classifier outperforms the Naïve Bayes classifier in terms of percentages for each emotion class. Thus, the CNN classifier seems to be outperforming the Naïve Bayes classifier according to this criterion. As the results of positive, negative, and neutral of CNN are higher than other methods that's why CNN is performing well on the Saraiki language corpus for sentiment analysis. But if we look in depth, we see that CNN is performing well on positive sentences. SVC is doing well on negative class as compared to other one. And Naïve bayes and CNN performing well on the neutral classes. As CNN is doing well in two classes positive and neutral then the overall performance of CNN is better than other method on the Saraiki dataset for sentiment analysis.

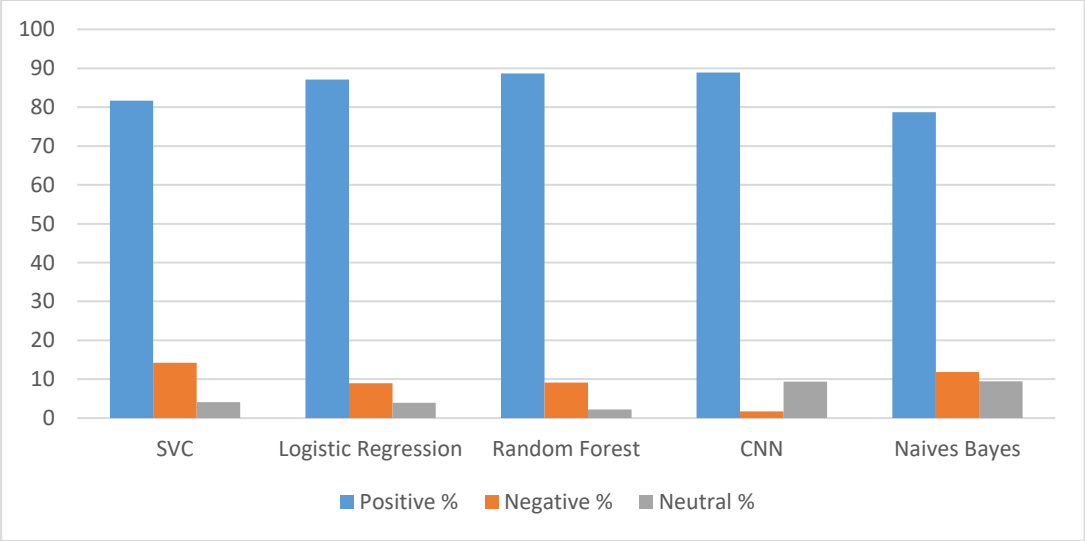

Figure 4.6: Comparison of Classifiers

POLICY RESEARCH JOURNAL
ISSN (E): 3006-7030 (P) 3006-7022

5. Conclusion

In this research we have done the creation of valid corpus of Saraiki for sentiment analysis. As there is no dataset available for Saraiki language. Then after performing labelling on dataset, it is converted into machine readable form file. Then preprocessing was performed on the dataset. feature extraction methods TF/IDF also performed on dataset to extract some features. Then CNN, Naïve bayes, Logistic Regression, Random Forest are applied on Saraiki dataset. By using CNN, captured complex patterns of text, especially understanding the context of word embeddings. As Naïve Bayes performs well for limited data. But it offers an effective approach for Saraiki language and provides fruitful results. Logistic Regression performed well for the straightforward task of sentiment analysis. In Random Forest, by using Saraiki dataset acquired well defined outcome. A comparison is made between results of all, and CNN is performing better than other methods on Saraiki Corpus for sentiment analysis.

By creating a corpus specifically tailored for the language, researchers can examine the emotions, attitudes, and cultural peculiarities expressed in Saraiki text data. The emotional landscape of Saraiki speakers will be clarified by this study, which will help with community involvement, traditional practice preservation, and communication strategy improvement.

5.1 Limitations

• There was not much data available online of Saraiki. So, it is not much as it should be for analysis. This limitation led to future work by increasing the size of dataset.

• As there is much work done on international languages, so the tools are available for them. Regional languages like Saraiki there are not many tools available as there is limited work done on it.

• For the languages like English, there are stop words and difficult words available. In this case there is no such list available which acts as a hurdle for **this.**

6. Future work

Future studies could take the Saraiki language corpus in a few ways to increase its value for sentiment analysis. First off, adding more Saraiki text data from various sources and contexts would expand the corpus's size and variety, improving its representativeness and generalizability. In future by increasing Corpus size of Saraiki, improved results can be obtained. Moreover, adding sentiment labels to the corpus would enable the construction of supervised machine learning models for sentiment categorization. Furthermore, incorporating advanced techniques for natural language processing, such as deep learning models, could better the accuracy and difficulty of which most researchers face in SA in Saraiki text. Analysis of Sentiment will remain relevant and useful if linguists, stakeholders, and Saraiki-speaking individuals are encouraged to collaborate and communicate with one another

ACKNOWLEDGMENTS

I'd like to thank several people and organizations for their help and encouragement throughout my master's program. First and foremost, I would like to express my heartfelt gratitude to my supervisor, Dr. Mubasher Malik, for his enthusiasm, patience, insightful comments, helpful information, practical advice, and never-ending ideas, which have always been invaluable in my project work and development. His vast knowledge, vast experience, and professional expertise in research project guidance enabled me to successfully complete this research. This project would not have been possible without his help and guidance. I could not have asked for a better supervisor during my studies.

Furthermore, I'd like to thank my friends for their moral support and assistance in better understanding the situation.

DEDICATION

This Research project is dedicated to Allah Almighty, my creator, my strongest leader, and my source of inspiration, knowledge, and comprehension. Throughout this program, he has been the foundation of my strength. My studies are also dedicated to my family and teachers. A special thank you to my loving parents and siblings, whose words of encouragement and motivation continue to ring in my ears. I also dedicate this documentation to my friends for their unwavering

support in all my endeavors and each teacher has helped me throughout the process. I will always be grateful for everything they have done, especially my supervisor, Dr. Mubasher Malik, who has spent many hours proofreading.

REFERENCES

Abdullah Al Ryan, M. S. (2023). FinTech: Deep Learning-based Sentiment Classification of User Reviews from Various Bangladeshi Mobile Financial Services.

Ahangari, M. a. (2023). A Hybrid Approach to Sentiment Analysis of Iranian Stock Market User's Opinions. International Journal of Engineering, 573--584.

Ahmadi, S. a. (2023). PALI: A Language Identification Benchmark for Perso-Arabic Scripts. arXiv preprint arXiv:2304.01322.

Akhtar, M. a. (2023). A machine learning approach for Urdu text sentiment analysis. Mehran University Research Journal Of Engineering \& Technology, 75--87.

Alasmari, E. a. (2023). Arabic Stock-News Sentiments and Economic Aspects using BERT Model. International Journal of Advanced Computer Science and Applications.

Allaith, A. a.-H. (2023). Sentiment Classification of Historical Danish and Norwegian Literary Texts. In The 24th Nordic Conference on Computational Linguistics.

Almuqren, L. a. (2021). AraCust: a Saudi Telecom Tweets corpus for sentiment analysis. PeerJ Computer Science, e510.

Asad, M. I. (2020). Classification of News Articles using Supervised Machine Learning Approach. Pakistan Journal of Engineering and Technology, 26--30.

Asno Azzawagama Firdaus, A. Y. (2023). Indonesian presidential election sentiment:Dataset of response public before 2024. Elsevier.

Astarkie, M. G. (2023). A Novel Approach to Sentiment In Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 2 (pp. 143-151). Springer.

author= {Saleemi, F. J. (2021). A Basic Parts of Speech (POS) Tagset for morphological, syntactic and lexical annotations of Saraiki language.

Badawi, S. a. (2024). KurdiSent: a corpus for kurdish sentiment analysis. Springer.

Baig, A. a. (2020). Developing a pos tagged corpus of urdu tweets. Computers, 90.

Baktash, J. A. (2023). Tuning Traditional Language Processing Approaches for Pashto Text Classification. arXiv preprint arXiv:2305.03737.

Bashir, M. F. (2023). Context-aware Emotion Detection from Low-resource Urdu Language Using Deep Neural Network. ACM Transactions on Asian and Low-Resource Language Information Processing, 1--30.

Chakravarthi, B. R. (2020). A Sentiment Analysis Dataset for Code-Mixed Malayalam-English. arXiv preprint arXiv:2006.00210.

Chakravarthi, B. R. (2020). Corpus creation for sentiment analysis in code-mixed Tamil-English text. arXiv preprint arXiv:2006.00206.

de Carvalho, V. D. (2020). An automated corpus annotation experiment in Brazilian Portuguese for sentiment analysis in public security. In Decision Support Systems X: Cognitive Decision Support Systems and Technologies: 6th International Conference on Decision Support System Technology, ICDSST 2020, Zaragoza, Spain, May 27--29, 2020, Proceedings 6} (pp. 99--111).

Dootio, M. A. (2021). Development of Sindhi text corpus. Journal of King Saud University-Computer and Information Sciences, 468-475.

D'Orazio, M. a. (2022). Automatic detection of maintenance requests: Comparison of Human Manual Annotation and Sentiment Analysis techniques (Vol. 134). Elsevier.

Elhassan, N. a.-A.-T. (2023). Arabic Sentiment Analysis Based on Word Embeddings and Deep Learning. Computers, 126.

Garouani, M. a. (2022). MAC: an open and free Moroccan Arabic Corpus for sentiment analysis. In Innovations in Smart Cities Applications Volume 5: The Proceedings of the 6th International Conference on Smart City Applications (pp. 849-858). Springer.

Gul, S. a. (2021). Development of saraiki wordnet by mapping of word senses: A corpus-based approach. Linguistics and Literature Review, 46-66.

Hameed, R. a. (2023). Transfer Learning for Low-Resource Sentiment Analysis}. arXiv preprint arXiv:2304.04703.

Hamid Parvin, B. M. (2020). A New N-gram Feature Extraction-Selection Method for Malicious Code. Springer.

Haq, I. a. (2023). The Pashto Corpus and Machine Learning Model for Automatic POS Tagging.

Haque, R. a. (2023). Multi-class sentiment classification on Bengali social media comments using machine learning. International Journal of Cognitive Computing in Engineering, 21-35.

Hasan, M. a. (2023). Natural Language Processing and Sentiment Analysis on Bangla Social Media Comments on Russia--Ukraine War Using Transformers. Vietnam Journal of Computer Science, 1--28.

Hashmi, M. A. (2019). Analysis of lecxico-semantic relations of Punjabi Shahmukhi nouns: A corpus-based study. International Journal of English Linguistics.

Hegde, A. a. (2022). Corpus Creation for Sentiment Analysis in Code-Mixed Tulu Text. In Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (pp. 33-40).

Hosseini, P. a. (2018). SentiPers: a sentiment analysis corpus for Persian. arXiv preprint arXiv:1801.07737.

Kashif, L. &. (2019). An intelligent multi-agent-based voice- enabled virtual.

Khaled, S. a. (2020). THE STUDY OF ORTHOGRAPHICAL DIFFERENCE BETWEEN PUNJABI LANGUAGE AND SIRAIKI DIALECT IN PUNJAB PROVINCE. Hamdard Islamicus, 175-193.

Khamphakdee, N. a. (2023). An Efficient Deep Learning for Thai Sentiment Analysis. Data, 90.

Kora, R. a. (2023). An enhanced approach for sentiment analysis based on meta-ensemble deep learning. Social Network Analysis and Mining, 38.

Krister and Jauhiainen, T. a. (2023). FinnSentiment: a Finnish social media corpus for sentiment polarity annotation. Language Resources and Evaluation, 581-609.

kumari Bygani, J. a. (2023). A Sentence Level Classification of Telugu News Document using Sentiment Analysis. In E3S Web of Conferences (p. 01037). EDP Sciences.

Kuriyozov, E. a. (2023). Text classification dataset and analysis for Uzbek language. arXiv preprint arXiv:2302.14494.

Lei Zhang, S. W. (2018). Deep Learning for Sentiment Analysis: A Survey.

Malik, M. H. (2022). Saraiki Language Hybrid Stemmer Using Rule-Based and LSTM-Based Sequence-To-Sequence Model Approach. Innovative Computing Review.

Mazoochi, M. a. (2023). Constructing Colloquial Dataset for Persian Sentiment Analysis of Social Microblogs. arXiv preprint arXiv:2306.12679.

Muhammad Yasir, 1. L. (2021). Mixed Script Identification Using Automated DNN. Computational Intelligence and Neuroscience.

Muhammad, K. B. (2023). Innovations in Urdu Sentiment Analysis Using Machine and Deep Learning Techniques for Two-Class Classification of Symmetric Datasets. Symmetry, 1027.

Muhammad, S. H. (2022). Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. arXiv preprint arXiv:2201.08277.

Muhammad, S. H. (2023). AfriSenti: A Twitter Sentiment Analysis Benchmark for African. arXiv preprint arXiv:2302.08956.

Mukku, S. S. (2017). Actsa: Annotated corpus for telugu sentiment analysis. In Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems (pp. 54-58).

Namee, K. a. (2023). A Hybrid Approach for Aspect-based Sentiment Analysis: A Case Study of Hotel Reviews. CURRENT

APPLIED SCIENCE AND TECHNOLOGY, 10-55003.

Nawaz, A. N. (2023). TPTS: Text pre-processing Techniques for Sindhi Language. Pakistan Journal of Emerging Science and Technologies (PJEST).

Nerabie, A. M. (2021). The impact of Arabic part of speech tagging on sentiment analysis: A new corpus and deep learning approach. Procedia Computer Science, 148--155.

P{\'e}rez, J. M. (2023). A Spanish dataset for Targeted Sentiment Analysis of political headlines. Electronic Journal of SADIO (EJS), 53--66.

Patel, A. (2023). Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model.

Patel, A. a. (2023). Sentiment Analysis of Customer Feedback and Reviews for Airline. Procedia Computer Science, 2459--2467.

Pilar, G.-D. a.-B.-M.-A. (2023). A novel flexible feature extraction algorithm for Spanish tweet sentiment. Expert Systems with Applications, 118817.

Poulami Basu, D. D. (2024). Using Machine Learning Algorithms With TF-IDF. International Journal of Engineering Research & Technology (IJERT).

Rahab, H. a. (2023). Rule-based Arabic sentiment analysis using binary equilibrium optimization algorithm. Arabian Journal for Science and Engineering, 2359--2374.

Raja Chakravarthi, B. a. (2020). Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text. arXiv e-prints, arXiv--2006.

Raychawdhary, N. a. (2023). Seals_Lab at SemEval-2023 Task 12: Sentiment Analysis for Low-resource African Languages, Hausa and Igbo. In Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023) (pp. 1508--1517).

Romadhona, N. P.-E.-H.-H. (2022). BRCC and SentiBahasaRojak: The First Bahasa Rojak Corpus for Pretraining and Sentiment Analysis Dataset. In Proceedings of the 29th International Conference on Computational Linguistics (pp. 4418--4428}).

Saini, J. R. (2018). An Exhaustive Meta-analytical Study of the History, Evolution and Development of'Saraiki NLP. INFOCOMP: Journal of Computer Science}.

Samih, A. a. (2023). Enhanced sentiment analysis based on improved word embeddings and XGboost. International Journal of Electrical \& Computer Engineering (2088-8708).

Sandeep Nigam, A. K. (2018). Machine Learning Based Approach To Sentiment. Communication Control and Networking.

Shafi, J. a. (2023). Semantic Tagging for the Urdu Language: Annotated Corpus and Multi-Target Classification Methods. ACM Transactions on Asian and Low-Resource Language Information Processing, 1--32.

Sharipov, M. a. (2023). UzbekTagger: The rule-based POS tagger for Uzbek language. arXiv preprint arXiv:2301.12711.

Sodhar, I. N. (2023). Hybrid Approach Used to Analyze the Sentiments of . International Journal of Advanced Computer Science and Applications.

Sodhar, I. N. (2023). Hybrid Approach Used to Analyze the Sentiments of Romanized Text . International Journal of Advanced Computer Science and Applications.

Talpur, N. a. (2023). Researching on Analysis and creating Corpus from Primary level Sindhi language Book . Repertus: Journal of Linguistics, Language Planning and Policy, 37--48.

Thakkar, G. a. (2023). Croatian Film Review Dataset (Cro-FiReDa): A Sentiment Annotated Dataset of Film Reviews. arXiv preprint arXiv:2305.0817.

Xu, J. (2023). A natural language processing based technique for sentiment analysis of college english corpus. PeerJ Computer Science, e1235.

Yekrangi, M. a. (2021). Financial markets sentiment analysis: Developing a specialized Lexicon. Journal of Intelligent Information Systems, 127--146.

Zamir, N. a. (2020). A Corpus-Based Analysis of The Lexico-Semantic Relationships of

Verbs Used in Saraiki language Newspaper. Ilkogretim Online, 4038--4047.

Zamir, N. a. (2021). Analyzing the Lexico-Semantic Relationships of Nouns Used in The Saraiki Newspaper: A Corpus Based Study. Ilkogretim Online.

List of Acronyms

CNN	Convolutional Neural Network
SVC	Support Vector Classifier
TF/IDF	Term Frequency/ Inverse Document Frequency
KNN	K nearest Neighbor
Bert	Bidirectional Encoder Representations from Transformers
LSTM	Long Short-Term Memory
POS	Part of Speech
NLP	Natural Language Processing