DEMOGRAPHICAL BASED SENTIMENT ANALYSIS FOR DETECTION OF HATE SPEECH TWEETS

By

Kamal Safdar

Supervisor:

Dr. Shibli Nisar

A thesis submitted to the Department of Computer Software Engineering, Military College of Signals, National University of Sciences and Technology, Islamabad, Pakistan, in partial fulfillment of the requirements for the degree of MS in Software Engineering

December 2022

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS Thesis written by **Kamal Safdar,** Registration No **00000359483**, of **Military College of Signals (MCS)** has been vetted by undersigned, found Complete in all respect as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes And is accepted as partial, fulfillment for award of MS/MPhil degree. It is further certified that Necessary amendments as pointed out by GEC members of the scholar have been also incorporated In the said thesis.

Signature: ________________________________

Name of Supervisor: **Dr. Shibli Nisar**

Date: _____________________________________

Signature (HOD): ____________________________

Date: ______________________________________

Signature (Dean/Principal): ____________________

Date: ______________________________________

DECLARATION

I hereby declare that this thesis titled "Demographical Based Sentiment Analysis for Detection of Hate Speech Tweets" is entirely based on my personal efforts under sincere guidance of my supervisor Dr. Shibli Nisar (MCS, NUST). All the sources used in this thesis have been cited and the contents of this thesis have not been plagiarized.

Signature of Student

Kamal Safdar

COPYRIGHT STATEMENT

.

ABSTRACT

As online content continues to grow, so does the spread of hate speech. The use of vast online social communication forums helps the user to express their opinion freely at any time. While the ability to freely express oneself is a human right that should be cherished, inducing and spreading hate towards another group is an abuse of this liberty. The availability of a large amount of data with demographical information such as location, time, and events are helpful to analyze the hidden patterns and understanding spontaneously expressed opinions in the sentiment analysis process which would enable more accurate results. Sentiment Analysis is a technique that is being used abundantly nowadays for customer reviews analysis, popularity analysis of electoral candidates, hate speech detection, and similar applications. This thesis aims to perform a spatiotemporal-based sentiment analysis of hate speech tweets in the Pakistan region. The process starts with the collection of political and religious-oriented demographic tweets through Twitter data API and web scrapper. After necessary text preprocessing the refine dataset will be annotated in three categories (Positive, Neutral, and offensive). The labeled data will be passed to the feature extraction module for relevant feature mining. The extracted feature will be trained on state-of-the-art machine learning and deep learning classifiers to investigate the performance of the proposed model.

DEDICATION

This thesis is dedicated to

MY FAMILY, FRIENDS, AND TEACHERS

For their love, endless support, and encouragement

ACKNOWLEDGEMENT

I am grateful to God Almighty who has bestowed me with the strength and the passion to accomplish this thesis and I am thankful to Him for His mercy and benevolence. Without his consent, I could not have indulged myself in this task.

LIST OF FIGURE

LIST OF TABLES

ACRONYMS

European Union	EU
Prevention of Electronic Crime Act	PECA
National Language Processing	NLP
Bag of Words	BoW
Part of Speech	PoS
Term Frequency Inverse Frequency Document	TF IDF
Support Vector Machine	SVM
Logistic Regression	LR
Long Short Term Memory	LSTM
Convolutional Neural Network	CNN
Sentiment Analysis	SA
Offensive Speech	OS
Hate Speech	HS
Neutral	N
Annotator	A
Data Set	DS
Word 2 Vector	W2V
Temporal Analysis	TA
Spatial Analysis	SA
Precision	P
Recall	R
F Score	F

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

The advancement of technologies and social media platforms like Facebook, Twitter, YouTube, and Instagram allow users to connect and communicate to share their ideas, and thoughts with relatives and friends in no time with the purpose to bring the social media community under one umbrella. In old days, different communication media is used to send messages from source to destination like smoke signals, telegraphs, carrier pigeons, balloon mail, etc. The main problem with using these modes is the delay factor, message received with delay losses its importance. The dramatic rise in technologies like high-speed networks like 5G supports the social media platform to share ideas within no time with the provision of freedom of speech. It permits every individual to extend hateful ideologies among the community. The use of abusive language in social media leads to hate crimes. Given the collaborative nature of social media, the detection of hate speech content and hate speech crime has become effortless. The traditional law enforcement institution somehow established some laws against hate speech content to reduce the spread of hate speech crimes, but the problem is still there. It affects the mental and emotional health of the target group (i.e. Shia, Sunny, Political, Ethnic group, etc.). The life cycle of hate speech content is comprised of four steps defined by Chatty and Alathur 2018, First step, hate speech remains high on social media, then it's gradually reduce after a few days in the second step. After some days the hate speech remains zero and then in the fourth stage, the hate speech again returns subject to content type, location, target class, etc. According to a recent online hate speech report of Pakistan, the most of hate speech promulgated is religiously and culturally motivated. 42 % from religion, 16% from Sex/gender / sexual orientation, 22 % from race/ ethnicity, and 23 % from nationality. The Root cause of hate speech promulgation on social media is the lack of awareness of hate speech and the lack of proper legislation and implementation by law enforcement agencies. Social media platforms like Facebook, Twitter, etc. Somehow formulate and implement AI-based hate speech detection Algorithms that automatically detect and remove the contents from their platform but there is limitation subject to diversity of content, language, and location. Legislation from different countries including the USA, Australia, Denmark, and the UK to protect their people from harassment and hate speech content. EU code of conduct was launched in 2016 and was implemented with the four internet social platforms (Facebook, Twitter, YouTube, and Microsoft) with the purpose to control and stopping hate speech content on the internet. Pakistan has formulated similar policies and laws to encounter hate speech. Pakistan Penal Code [1] states any violation against race, ethnicity, community, religious group and any cast will in result five years of imprisonment. The anti-terrorism act 1997[2], declares the individual guilty if he or she founds with threatening or abusive language or words. Article 19[3], every citizen have a right to freedom of speech with some limitation imposed by the law. The Prevention of Electronic Crimes Act (PECA) 2016, restricts the user to post/sharing hate speech content on social media that leads to interfaith, sectarian, or racial hatred.

Since the increase of online hate speech to social media companies like Twitter and Facebook, they were under public and political pressure from many anti-hate government agencies. Germany has passed a law that could fine Twitter, Facebook, and other social media companies up to 40 million for failing to remove defamation, violence, and hate speech within 24 hours. The European Commission has issued a code of conduct to combat online hate speech. According to a report [4], Facebook removes hate speech content faster than Twitter and YouTube. Facebook accessed 95% of hate speech notifications in less than 24 hours, while Instagram responded 62%, Twitter 44%, and YouTube 9% on hate speech notifications. In 2020 Mark Zukerberg announced a comprehensive policy on hate speech content used in Ads, Facebook will remove all such contents that target a specific group (Race, national origin, gender, sexual orientation, etc.).

1.1 Problem Statement and Objectives

Despite the immense contribution of law enforcement agencies and social media platforms like (Facebook, Twitter, and Youtube) to reduce the promulgation of hate speech content, it largely remains unchecked. The belief in the user to report abusive language can leave the content unreported and unnoticed. On the reporter side, it is very difficult to maintain and track the manual offensive contents to be flagged and removed. It is not only a massive uphill task but it can also be personal bias and subjective views. With the rise of information flow on social media, it is insufficient to manually filter the hate speech contents on social media which implies the automated detection of online hate speech. It is also a complex problem to have automated systems as it involves capturing the context, keywords, sarcasm, irony, analyzing tone, and diversity of language available to communicate ideas over social media platforms. English is the most common language used to express ideas. There is a lot of contribution reported in the detection of hate speech content in the English language. Being a national language of Pakistan, Urdu has grabbed less attention in the detection of hate speech content. A study conducted on Pakistan's cyberspace showed that 51% of the respondents had been the target of online hate speech. Given the history of terrorism in Pakistan and the ongoing efforts of war against terrorism, there is a dire need of developing resources comprising hateful content in the Urdu language and intelligent systems that could detect such content automatically. Objectives of this thesis are:

- To develop vast offensive and hate speech content in the Urdu Language.
- To develop automatic detection of hate speech contents in the Urdu language using Machine / Deep Learning Techniques.
- To perform visualization and analysis of Urdu hate speech data using Techniques like (Heat Maps, choropleth maps, and time series plots).

1.2 Contribution

Being a low resource language Urdu, there is very less amount of work done in hate speech detection. To the best of my knowledge, all available Urdu corpus is insufficient for further analyses as Machine / Deep learning Algorithms require a lot of data to understand the hidden patterns and to perform efficiently. The research is based on three target classes (Neutral/Positive, Offensive, and Hate Speech). The dataset contains useful information like time and location for the purpose to visualize the hidden patterns and better understanding. To the best of 3 our knowledge, there are no existing publicly available resources comprising hate speech in the Urdu language, the primary contributions of this thesis are:

- To make the Urdu textual Corpus include all categories (Ethnic, Religious, Political, national origin, gender, Sex).
- The data is extracted based on location and time, the focus of the research is to target data about Punjab including 36 districts in light of five years from 2018 -2022.

- Classify the Urdu text corpus into three target classes (Neutral, Hate Speech, and Offensive). Subsequently, using data cleaning and feature extraction techniques to extract useful information/features for further analysis and classification.

- To establish the baseline for automatically detecting offensive speech and hate speech.

- To perform visualization and time series analysis on Corpus using techniques like (Heat Map, choropleth map, and time series plot).

1.3 Thesis Outline

This thesis is divided into seven chapters:

- **Chapter 1**: This chapter includes the basic introduction, and establishes the objectives and primary contribution of my research work.
- **Chapter 2**: This chapter describes the previous work on Hate Speech Detection in the Urdu Language.
- **Chapter 3**: This chapter defines offensive and hate speech in Urdu Language Corpus.
- **Chapter 4**: This chapter describes the data collection process of the Urdu language.
- **Chapter 5**: This chapter includes the Comprehensive techniques used for the detection of hate speech text in corpus. It also includes visualization and time series techniques used to describe the data,
- **Chapter 6**: This chapter presents the model evaluation and result.
- **Chapter 7**: This chapter concludes the report and highlights the direction for future work.

Chapter 2

LITERATURE REVIEW

In recent past years, Hate speech detection remained a hot topic area for research as NLP leads and a lot of research has been conducted to understand the hidden patterns and useful information from textual data. There are different approaches used to detect hate speech content in text such as Supervised, unsupervised, and semi-supervised learning. State-of-the-art tackles this problem with supervised techniques as the Dataset is to be labeled and feed into the model for training and testing. Mainly two approaches are used Traditional approach (SVM, Decision Tree, and LR, etc.) and other is deep learning approach (LSTM, CNN, RNN) in which the model learns the pattern with the support of multiple layers of neural network based on input. The first part of the Literature review for each above approach encompasses English including other low resource languages except Urdu and another part is for the Urdu language.

2.1 Traditional Approaches

2.1.1 English & Low resource language except for Urdu

Burnap & Williams [5] used a Bag of words with n-grams (n=1-5) and Algorithms ruled-based and spatial-based classifiers and achieved a 98% accuracy. Waseem & Hovy [6] used extra-linguistic features and n-grams (n=1-4) and achieved 64.58% efficiency. Linguistic features are used to identify the sense of a word. Davidson et al. [7] in his research and implements the part of speech tag (POS), bigrams, unigrams, trigrams, and tf-idf using machine learning algorithms SVM, Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT) and linear SVM and achieved efficiency 90% on English tweets. Gamback & Sikdar [8] used a char n-gram and word2vec model and achieved 78.3% efficiency. The word2vec is a word embedding technique used to learn word association from a large dataset. Malmasi & Zampieri [9] focuses on hate speech profanity and anti-social behavior with char n-gram, n skip-gram and uses a linear SVM model and achieved 78% efficiency. Garima Koushik & Mr. Suresh Kannan Muthusamy [10] used BOW and TF-IDF approaches to train machine learning models, after conducting exhaustive experiments on the Twitter dataset the logistic regression outperforms with the accuracy of 94.11 % on detecting binary classes either hate or not hate. Kelvin, George, Richard, and Kennedy[11] develops an approach for detecting hate speech content by the self-identified hateful community, Naive bayse classifier gives better results with precision, recall, and accuracy values of 58%, 62%, and 68% respectively. HAJIME & MONDHER [12] used a unigram approach on a small dataset of 2010 tweets. The experiments are conducted based on binary and ternary classification. Results show that accuracy achieved 87% on binary classification and 78.4% on ternary classification. Trisna & Arif [13], worked on hate speech and cyber pulling detection Indonesian language on the data promulgated during the election 2019. The paper comprehensively describes the process of developing a dataset with more than 1 Million tweets using Twitter developer API. In the basic preprocessing and implementing machine learning algorithms, the Latent Dirichlet Allocation LDA is used to extract the topic from collected tweets and detail sentiment analysis on each category applied to generate a polarity score on balance data. The naïve bayse classifier achieved an accuracy level of 78.7%. Yasemin & Rehime[14] works emphasize hate speech on women. The Turkish data is collected from Twitter with the approach to search tweets from the specific hashtag on a choice of clothing of women. In their research, they applied five machine learning algorithms for the detection of hate speech content against women including Support vector machine, J48, Naïve Bayse, Random Forest, and Random tree. Results show that Naïve bayse performed best with an f score of 62 % among all. OLUWAFEMI & EDUAN [15] targeted the development of an English corpus from South African tweets to find the hate of offensive content by implementing different machine learning algorithms. Character n-grams, word n-grams, and negative sentiment are used to extract useful features from the dataset. In machine learning, support vector machines, random forest, logistic regression, and gradient boosting are used. Preliminary results show that support vector machine with n-gram is best in the detection of hate speech with a true positive rate of 89.4% and optimized gradient boosting with word n-gram performs best with a positive rate of 86%. The comprehensive analysis presented that multi-tier learning models could overcome the misclassification error rate by 34%. Purnama & Budhi [16] used a multinomial logistic regression classifier with a tf-idf feature extraction technique that achieved the best average score of precision of 80.02 %, recall of 82%, and accuracy of 87.66%. Sattam & Pablo & Francisco & Alexey used a supervised classifier including a support vector machine, Gaussian naïve bayse, Decision tree, nearest neighbors, and random forest and the target language is English and Spanish. Results show that Naïve bayse, support vector machine and random forest performs wells into account all features with an average f score of 77%. Shervin & Marcos [17] used n-grams, word n-grams, and word skip grams with a supervised learning model on the annotated dataset with hate speech tweets 2399, offensive 4836 and ok tweets with 7247 out of 14509 tweets and it is found that Support vector machine has been outperformed well for native and variety language identification. The SVM achieved an accuracy of 78% with char 4 gram and with word unigrams SVM achieved 77.5% accuracy. Tom De Smedt & Guy De Pauw [18] examine the quantitative and qualitative analysis of Twitter data containing Jihadist hate speech. The data corpus was collected in compliance with the online procedure. The total data collected is 45K tweets from 2014-2016 covering a region Syria, Iraq, France, United States, Israel, Russia, Jorden, Iran, Egypt, Yemen, Damascus, and London. The SVM model trained on the balanced training set of 45K hates speech tweets and the same for safe tweets. The accuracy achieved is 82% (F1 Score) by applying 3-fold cross-validation.

2.1.2 Urdu Language

M. MOIN & Khurram & M. Kamran [19] worked on hate speech detection in roman Urdu tweets, 5000 roman Urdu tweets were collected. Tweets are further classified into three classes' Neutral-Hostile, Simple-Complex, and Offensive-Hate speech. Five different machine learning techniques. The results show that logistic regression outperformed all with an F1 score of 0.756 for offensive hate speech tweets. M Z Ali & Ehsan & Kashif & Sarmad [20] contributed to improving hate speech detection of Urdu tweets using sentiment analysis. The research addressed the challenges and problems including dimensionality, sparsity, and high skewed classes. The data is annotated in five classes (Neutral, positive, highly positive offensive, highly offensive). The target category is national security and religion. The SMOTE, variable global feature selection techniques are used to handle the sparsity, class imbalance problem. The two machine algorithms SVM and naïve bayse are used. Initial baseline results show that SVM performed well with an F Score (0.626), after improving the performance of the classifier, the results improved with an f score of 0.93. M. PERVEZ AKHTER & ZHENG & IRFAN & M. Abdul Majeed & Tariq [21] collected 5000 tweets of roman Urdu and Urdu respectively. The N-grams technique is used on character and word levels. The seven machine learning algorithms are used to detect offensive or non-offensive tweets from a corpus. The experiment shows that regression models perform best with n-grams about process Urdu tweets. Logitboost and simple logistics outperform others with a score of 95%. M Owais & Qaiser & Ghulam [22] used machine learning algorithms including logistic regression, begging, decisions tree, and ANN to detect abusive language within-corpus of 2400 tweets (1187 Abusive and 1213 no abusive). After performing the classifier task, results show that logistic regression performs best with an f score of 83%.

2.2 Deep Learning

Deep learning uses an artificial neural network to learn abstract representations of data using layers (input, hidden, and output). The most famous deep-learning techniques are CNN, RNN, and LSTM. CNN is best for learning spatial patterns in a dataset.

2.2.1 English & Low resource language except for Urdu

Badjatiya & Shashank [23] worked on a 16K annotated dataset 16K with three target class's racist, sexist, and neither. In their research, extensive experiments were conducted with multiple deep learning architectures in contrast with word embedding to handle the complexity. Results on the benchmark dataset showed that deep learning methods outperform the char/word gram method by 18 f points. Aya & Zakaria & Nadia [24] contributed well to hate speech detection from multiple languages that appeared in tweets. The experiments were performed on a Convolutional Neural network (CNN) with character-level representation. The result with the best parameter was 0.889 for the dataset containing five languages and 0.83 for the dataset containing seven languages. Lin & Yoshimi [25] did experiments on two different datasets with different sizes (Dataset A containing 9925 and Dataset B containing 31962 records). Traditional approaches (Logistic regression, SVM) and deep learning (LSTM, Stacking, and GRU) were applied to two different datasets. The result showed that Logistic regression outperformed dataset A with an f score of 43% and LSTM on dataset B with an f score of 67.30%. Gameback &

Utpal [26] introduced a deep learning-based hate speech model. The text was classified into four categories i.e. racism, sexism, both, and not hate speech from the dataset of 9K. The experiments showed that CNN performed well with word2vec with an f score of 78%. M. Umar & Imran & Arif & Saru & Saleem [27] proposed a combination of CNN & LSTM for performing sentiment analysis for the detection of hate speech on three datasets. The model is analyzed with traditional models i.e. SVM, Logistic regression, voting classifier, Random forest, and SGD. This study also investigated two different feature extraction techniques TF-IDF and word2vec to determine their impact on accuracy. The results showed that CNN –LSTM performed well among all. Roy & Kumar & DAS & XIAO [28] developed a deep Convolution neural network for hate speech detection. In this, they used Glove embedding to analyze the semantics of tweets promulgated on Twitter and achieved precision, recall, and f score values of 0.97, 0.88, and 0.92 respectively. Nabila & Nasrun & Setianingsih [29] used an artificial neural network with a back propagation method. The case study identified the hate speech in the sentence. The random accounts were analyzed who involved in hate speech had almost 1235 tweets of which 626 tweets were categorized as hate speech and 583 tweets were classified as non-hate speech. The result was analyzed with hypermeters like Epoch size and learning rate and has been found that results were improved with the tuning of hyperparameters. The overall result obtained an average recall of 90.03%, a precision of 80.6%, and an accuracy of 89.4%.

2.2.2 Urdu Language

Raza & Umar & Umair & Waseem [31] worked on the Urdu tweets dataset of 10K. The different machine learning algorithms are used for hate speech detection and transfer learning to exploit fast text and Bert multi-lingual embedding model. The result shows that Bert's improves the f scores of 0.67, 0.68, and 0.69 respectively. Hammad & Haroon & Asim [32] developed annotated roman Urdu dataset of 10K and proposed a CNN-gram deep learning architecture. The results show that transfer learning is better and more beneficial as compared to training a dataset from scratch. Lal & Ammar & Noman & Hsien [33] worked on multi-class sentiment analysis of Urdu text using Word / Char n-gram, fastText, and BERT. The result shows that BERT pre-trained embedding outperformed Deep learning and achieved an f score of 81%.

In this chapter, previous work done on hate speech detection has been discussed in detail. A lot of work done in the English language and research on low-resource languages is in focus. A minimum of work has been observed especially in the Urdu language. The thesis aims to develop a hate speech detection system on a large Urdu corpus with the support of Machine learning and deep learning techniques. In addition, a comprehensive demographical analysis of time and location features is available in the Collected Urdu Corpus.

CHAPTER 3

OFFENSIVE LANGUAGE AND HATE SPEECH

It is very important to understand both terms offensive and hate speech. There is no unique definition of these terms. In this chapter, we briefly explain and define the terms in general and in our research context.

3.1 Difference between Offensive and Hate Speech

In general, there is no unique definition exists that defines the boundary between offensive and hate speech. Sometimes these terms are used interchangeably but the need is to understand the difference between both terms. The offensive is sort of abusive and insulting language that contains a set of offensive words regardless of category. Hate speech is instantly considered offensive but every offensive language is not necessarily to be considered hate speech. Hate speech is a type of public or free speech that encourages hatred and violence among individuals or groups based on their protected characteristics such as Sex, Gender, Ethnic, religion, politics, etc. Each progressive country and social media platform has defined hate speech terms and developed comprehensive legislation. Twitter defines hate speech as any content that promotes violence against or directly attacks or threatens people based on race, sex, gender, ethnicity, religious, etc. The suspected account has been suspended and the content is to be removed within 24 hours. Facebook defines any content that directly attacks people based on protected characteristics referred to as hate speech. EU has defined a comprehensive code of conduct on hate speech. According to the EU, hate speech is treated as public incitement to violence or hatred based on certain characteristics, including race, color, religion, descent, and national or ethnic origin.

3.2 Research-Oriented Definition of Offensive vs. Hate Speech

It is interesting for us to develop an understanding of offensive and hate speech definitions. The following set of parameters briefly explains and elaborates on the offensive and hate speech terms in our research.

3.2.1 Offensive

In our research, we have developed lexicons/sets of keywords that are commonly used in our dataset that are considered abusive or offensive. For example لعنت, بیغرت, کتا, گشتی, کھوتا, حرامی, خنزیر, کمینے, بھڑوے, غدار

3.2.2 Hate Speech

In our research, the main concern is to establish the boundaries that briefly explain and define hate speech. Thoroughly going through the data, the following important and interesting factors have been established that cover the hate speech term especially for our dataset contains 0.2 M tweets:

- Every tweet that includes any offensive word aims to target an individual or group based on protected characteristics such as religious, ethnic, political, etc.

- Two or more offensive words used within a tweet based on the intensity of offensiveness like (خنزیر غدار).

- Any hurtful term used for individual or group like (ککڑی، پورن سٹار, بھکاری, عمرانڈو, ٹریکٹر ٹرالی, رانا باندری ، فضلو).

- Any offensive tweet by considering the subjectivity of the sentence.

CHAPTER 4

DATA COLLECTION AND DEMOGRAPHICAL ANALYSIS

In this chapter, we briefly discuss the process of collection of tweets and subsequently refining the process to make it prepare for annotation/ labeling.

4.1 Dataset Collection

There are multiple options available to extract data from social media like Twitter, Facebook, Instagram, and YouTube. Facebook and Twitter are the social media platforms being preferably adopted in Pakistan due to their rich features and user-friendly interface. According to Global Statistics, Facebook was used 82% and Twitter 15% in Pakistan in Sep 2022. The hashtag feature in Twitter is used to represent the topic on Twitter. Twitter allows the user to tweet according to their interest freely. Recently political disability in Pakistan enables Social media users in Pakistan to express their political affiliations and opinion freely. The hashtag #امپورٹڈ_حکومت_نامنظور [] trend has more than 106 M tweets within one week which witnessed the exponent growth of hate speech tweets in Pakistan. Being the second widely adopted social media platform in Pakistan, we selected Twitter as our source of data as Twitter allows privileged users to access the information for research purposes. For this Twitter developer team allows the researcher to get access to its content by defining the registration process. Our research initially used Twitter developer API that allows to access the information for only 7 days. The process to extract an ample amount of data from this method takes too much time. To make the process fast many open-source scripts are available to fetch the required information in no time.

We examined and explored the all options and found the "SNscrape" one of the useful open-source scripts to extract the required data for our research

4.1.1 SNscrape: Social Network Scraper

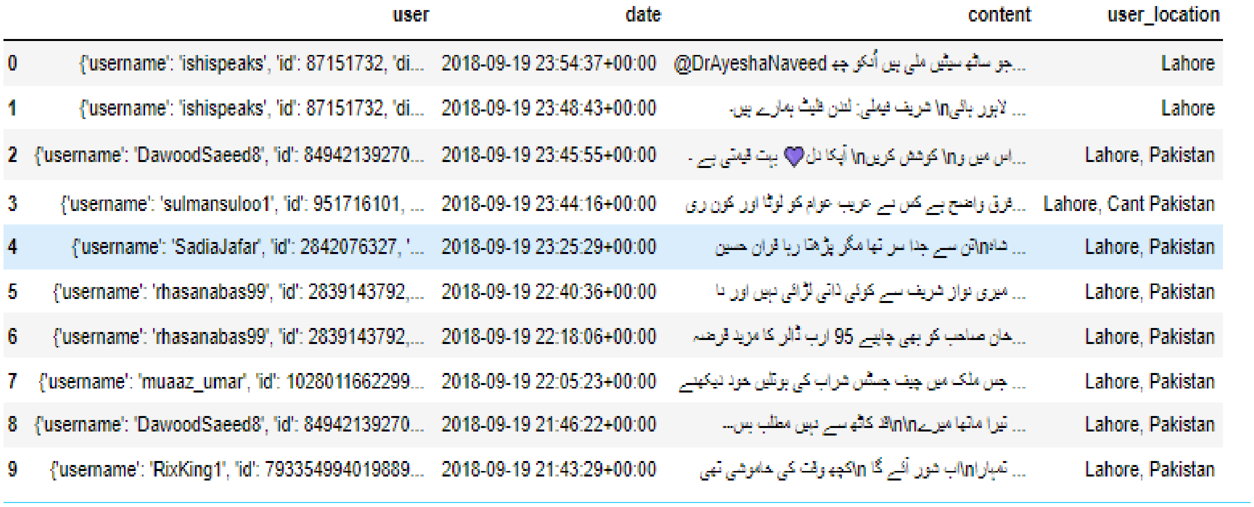
SNscrape [] is an open-source scraper for social networks that enables to extract the data from Social Media platforms like Twitter, Facebook, Instagram, Reddit, and Weibo. The basic functionality is that it provides access to user profiles, groups, hashtags, and trends. As a prerequisite python 3.8 or above is required for the installation of the scraper. Initially, 90 slur terms are shortlisted that are commonly used in the Urdu language. Our slur terms such as غدار (Mutinousness), لعنت (damn/curse), خنزیر (Pig), حرامی (Bloody), کتا (Dog), (donkey) گشتی , دلہ , بےغیرت کھوتا are some examples and other slur terms used to target individual such as ککڑی ، نانی , باندری , فضلو ، , جادوگرنی , بھکاری , بھگوڑی , بھگوڑے, عمرانڈو, ٹریکٹر ٹرالی , پورن سٹار, , کرائم منسٹر, چوکیدار ، ڈبو , پٹواری , are shortlisted. The terms that are used to extract the religious tweets like کافر قادیانی, کافر ، شیعہ خنزیر ، شیعہ فتنہ, گستاخ صحابہ, وہابی, سنی and terms like افغان, بلوچ ، ایران ، پشتونوں also shortlisted to extract maximum tweets against these slur terms. The SNscrape script has many variations to extract tweets. In our research, we have to extract the demographical information such as Location and time with tweets. SNscrape provides a script (shown in fig 4.1) with the following parameters that fulfill our research requirement:

- **Location** (The information is given through Latitude , Longitude parameter with provision of surrounding distance)
- **Time** (with the parameter Since and until)
- **Keyword search** (The slur terms like قادیانی , پشتونوں)
- **Language** (The required language is given in the Lang parameter such as in our case Urdu so we set the parameter as Lang =”ur”)
- **Number of tweets (**Required no of tweets to be fetched**)**
- **User profile**

```
import pandas as pd
import snscrape.modules.twitter as sntwitter
import itertools
loc = '31.523844543701532, 74.35154811757151, 10km'
df_coord = pd.DataFrame(itertools.islice(sntwitter.TwitterSearchScraper('قادیانی  ,lang:ur
since:2018-04-01 until:2018-09-20
geocode:"{}"'.format(loc)).get_items(),20000))[['user','date','content']]
df_coord['user_location'] =  df_coord['user'].apply(lambda x: x['location'])
```

Figure 4.1 SNscrape Script

Being a low resource language, Urdu has very fewer tweets as compared to English, through SNscrape scripts we extract tweets with a timestamp range between 2018- Apr 2022 i.e. five years. This is a very cumbersome job as we have to get Urdu tweets as well our focus is to extract tweets with location (within Pakistan) and time. We restricted our research to only Punjab districts as overall Pakistan has almost 160 districts and covering all locations and extracting tweets against all were a difficult job. Punjab is one of the provinces of Pakistan that has 36 districts such as Lahore, Rawalpindi, Gujarat, Jhelum, D.I Khan, RYK, BWP, etc. . Initially we made a list of all districts of Punjab with their Latitude and Longitude values and then extract tweets against each district with the help of parameters (Time, slur terms, No of tweets). Figure 4.2 shows a brief description of the output SNscrape:



	user	date	content	user_location
0	{'username': 'ishispeaks', 'id': 87151732, 'di...	2018-09-19 23:54:37+00:00	@DrAyeshaNaveed جو ساتھ بیٹھی ملی ہیں اُنکو چ...	Lahore
1	{'username': 'ishispeaks', 'id': 87151732, 'di...	2018-09-19 23:48:43+00:00	... لاہور باقی\n شریف فیملی: لندن فلیٹ ہمارے ہیں.	Lahore
2	{'username': 'DawoodSaeed8', 'id': 84942139270...	2018-09-19 23:45:55+00:00	...اس میں ور\n کوشش کریں\n آپکا دل💜 بہت قیمتی ہے ۔	Lahore, Pakistan
3	{'username': 'sulmansuloo1', 'id': 951716101, ...	2018-09-19 23:44:16+00:00	...فرق واضح ہے کس نے غریب عوام کو لوٹا اور کون ری	Lahore, Cant Pakistan
4	{'username': 'SadiaJafar', 'id': 2842076327, '...	2018-09-19 23:25:29+00:00	... شاہ\nاتن سے جدا سر تھا مگر پڑھتا رہا قران حسین	Lahore, Pakistan
5	{'username': 'rhasanabas99', 'id': 2839143792,...	2018-09-19 22:40:36+00:00	... میری نواز شریف سے کوئی ذاتی لڑائی نہیں اور نا	Lahore, Pakistan
6	{'username': 'rhasanabas99', 'id': 2839143792,...	2018-09-19 22:18:06+00:00	...خان صاحب کو بھی چاہیے 95 ارب ڈالر کا مزید قرضہ	Lahore, Pakistan
7	{'username': 'muaaz_umar', 'id': 1028011662299...	2018-09-19 22:05:23+00:00	... جس ملک میں چیف جسٹس شراب کی بوتلیں خود دیکھے	Lahore, Pakistan
8	{'username': 'DawoodSaeed8', 'id': 84942139270...	2018-09-19 21:46:22+00:00	... تیرا ماتھا میرے\n\nلا کاٹھ سے نہیں مطلب ہیں...	Lahore, Pakistan
9	{'username': 'RixKing1', 'id': 793354994019889...	2018-09-19 21:43:29+00:00	... تمہارا\nاب شور آئے گا \nکچھ وقت کی خاموشی تھی	Lahore, Pakistan

Figure 4.2 SNscrape Script Output

As a result, we extracted almost 0.25 M tweets covering all 36 districts of Punjab, and more than 0.5 M tweets were found ambiguous and duplicated. The duplicate tweets were removed and overall 0.2 M unique tweets were left as a final dataset.

The data collection process starts from the finalization of slur terms/keywords that are most commonly used in the Urdu Language, especially in Pakistan. Subsequently desired tweets are extracted against each district with defined parameters and saved the tweets of each district separately. The tweets of each district merged in one file to have a combined dataset of all districts to start the Labeling and Annotation process. Fig 4.3 shows the complete data collection process.

4.2 Refining Process

The process to prepared the data for Annotation/labeling comprises two phases, one is to collect the data against each district and combined them to have a combined dataset and the other is the refining process which is required to remove ambiguous and duplicate data and to make it in readable form as the Urdu language needs to be encoded for clear visibility and readability of text. The following steps were taken to prepare the extracted tweets for labeling / Annotation:

- Create a New Excel workbook, from the option Data -> from text ->open Combined dataset CSV File
- Select Option Delimited and from **file origin** select 65001: Unicode (UTF-8)
- Delete Empty rows
- Remove duplicate records
- Filter the dataset subject to requirements
- Wrap the text

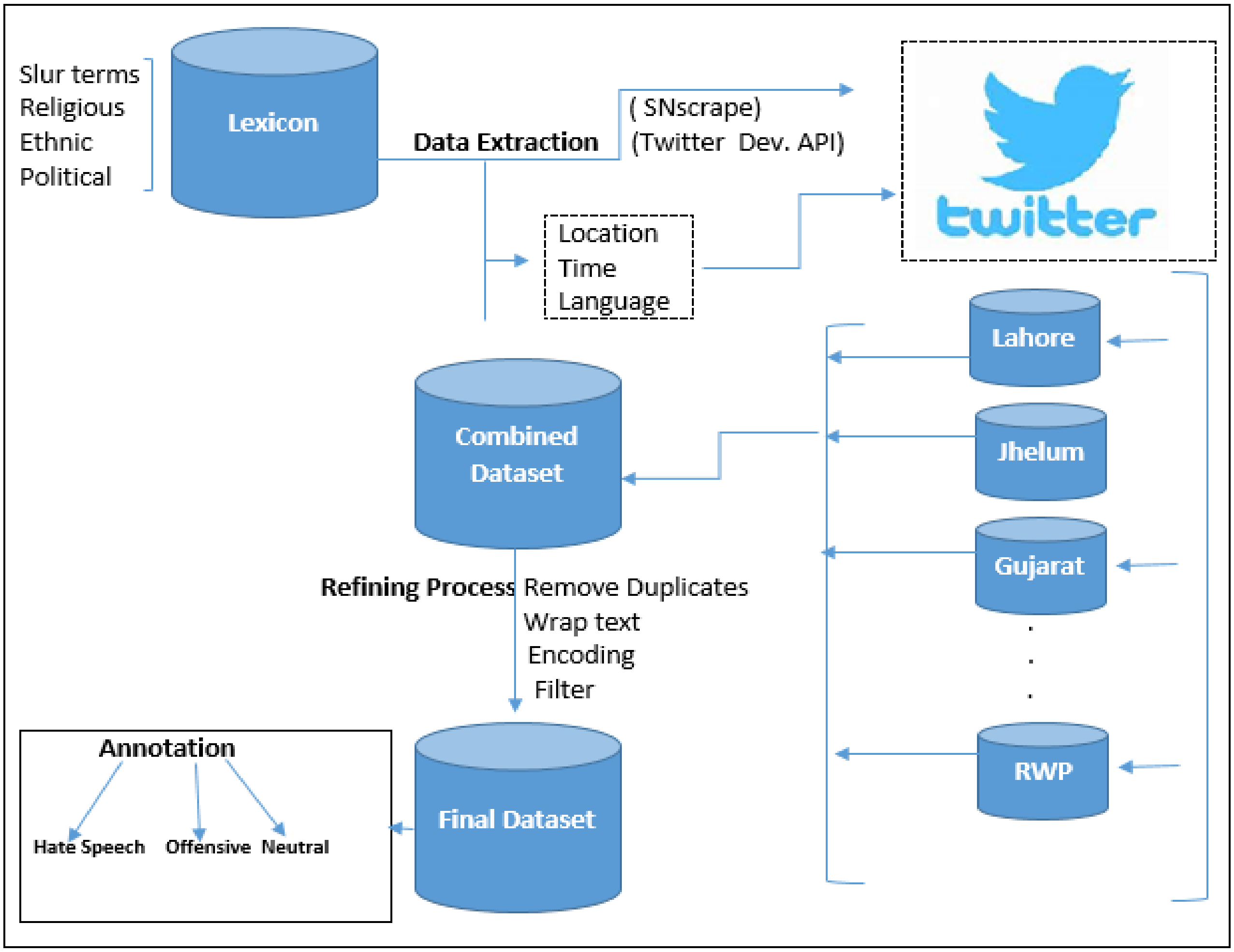


Figure 4.3 Data Collection and Refining Process

We randomly select 70% of our dataset to form the training set while the remaining 30% of the dataset comprises the testing set**.** The distribution of tweets between training and testing is shown in table 4.1. The Combined Dataset is highly imbalanced as Neutral tweets dominate the other classes hate speech and Offensive. The percentage ratio of neutral tweets in the combined dataset.

Table 4.1 Train Test Distribution of Dataset

Data set	**Hate Speech**		**Offensive**		**Neutral**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Combined DS (0.2 M)	**4981**	2069	15744	6321	127,342	52,143

Are 84 % and offensive 11% and hate speech with 4% tweets. Hate speech exists very rarely so to keep the originality of data we do not need to balance data. In our research, we did experiments on both balanced and imbalanced data.

4.3 Annotation Process

The annotation process was a very cumbersome job as we had to annotate the dataset containing 0.2 M tweets. Three annotators including one domain expert started annotation on the combined dataset. The process was started by dividing the dataset into 3 parts each annotator got 67K tweets to be annotated. Annotation guidelines were already formulated in Chapter 3. To annotate such huge data we mutually decided to complete the annotation task within 3 months timestamp. It was decided to label a dataset in three classes Hate speech labeled as -1, Offensive as 1, and Neutral / Positive as 0. As the initial annotation by each annotator, the file of Annotator A handed over to Annotator C and vice versa for cross-verification of the annotation process and omission of any human mistake (if any). The voting system was maintained while finalizing the labeling. For example, to finalize the tweet label, there should be a minimum of 2 annotators who agreed on the same label either Hate speech or Offensive. Table 4.2 shows the annotation process:

Table 4.2 Annotation Process

Tweet	Annotator			Final Label
	1	2	3	
چینی چور کرپٹ حماد اظہر کو گرفتار کر لیا گیا	-1	1	-1	**-1**
گانڈ میں لو اپنا ووٹ	1	-1	1	**1**
توں دلال ہے پتا ہے اے آر وائے کا	-1	-1	1	**-1**
قادیانی کائنات کا بدترین کافر ہے	-1	-1	-1	**-1**
جاھل بکاو ٹٹو صحافی	1	-1	-1	**-1**
نیازی رنڈی کا بچہ	-1	-1	1	**-1**
رنڈی اپنے کنجر باپ کو بلا اپنے کنجر بیٹے کو بلا اپنے یار قطری کو بلا گشتی	-1	1	-1	**-1**

There are very variation in tweets that makes it difficult to label either in the offensive or hate speech category. Negative tagging referred to targeting individuals or groups such as Rana Sanaullah being a killer or Qadainies as being the worst creatures in the universe.

یہودی اور قادیانی کبھی سامنے سے وار نہیں کرتے

The above tweet represents the specific group but no abusive language is used in this tweet that makes it hate speech, so we consider it as offensive tweet

مختلف جماعتوں اور علماء کرام کے احتجاج پر حکومت کا ایک اور یو ٹرن
لیتے ہوئے قادیانیوں کو اقلیتی کمیشن میں شامل کرنے کا فیصلہ واپس لے لیا

Table 4.3 shows the group-wise hate speech tweets in the dataset such as Ethnic – Hate Speech, Political - Hate Speech, Religious - Hate Speech

Table 4.3 Group wise Hate Speech Tweets

Group	Tweets
Religious Hate Speech	قادیانیوں پر لعنت بے شمار غدار ختم نبوت لعنت ہو تم پر قادیانی اس ملک کی جڑوں کو کاٹ رہیں ہیں قادیانیوں لعنتیوں اور کافروں یہودیوں کے یاروں اور حمایتیوں پر بے شمار لعنتیں
Ethnic Hate Speech	تم پاکستان کے دشمن ہی نہی بلکے غدار ہو اس میراثی کو راجپوت کہہ کر ہماری توہین مت کرو۔ یہ افغانی کتا ھے پنحاب اپنی تقسیم کرنے والوں پر لعنت بھیجتے ہوئے
Political Hate Speech	میرے پیارے بھڑوے صحافی کل پارلیمان آپ کو کھسرا ڈیکلئر کر دے تو کہاں جاو گے لکھ لعنت نواز شریف تجھ پر حامد میر تم جیسا غدار وطن اور ن لیگ کا دلال میں نے اپنی زندگی میں آج تک نہیں دیکھا لعنت تم جیسے صحافی پر لعنت ہو اس شخص پر جوتجھ جیسے بیغیرت کا لیڈر ہے تم جیسے پی ٹی آئی والوں کو دیکھ کر عمران نیازی سے نفرت بڑھ جاتی ہے

4.4 Data Distribution

The Dataset collected through SNscrape for the period of five years comprises from 2018- Apr 2022 was labeled into three classes (Hate Speech, Offensive, and Neutral). It has been observed that data is highly skewed as the majority of rows were associated with one class. The following statistics show the distribution of classes in the overall dataset containing 0.2 M.

Hate Speech	**7101**
Offensive	**22229**
Neutral	**172491**

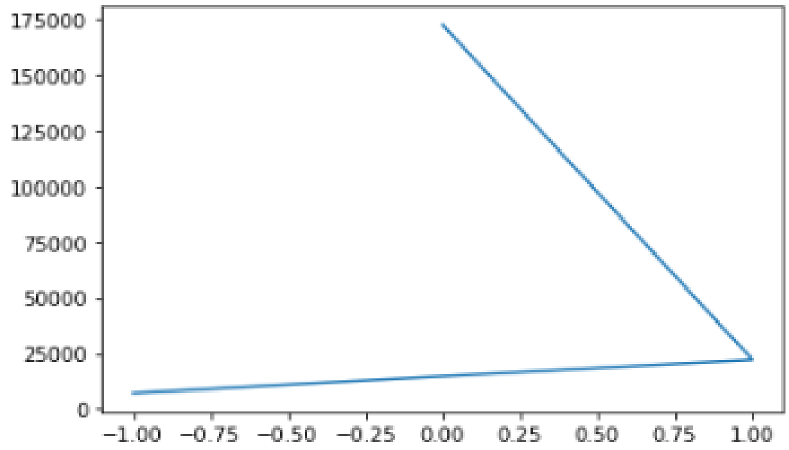


Figure 4.4 Class Distribution

4.5 Demographical Analysis

Spatial-temporal-based analysis of big data provides the opportunity to understand interesting patterns and trends such as event detection. We did the same Spatio-temporal analysis for our data to find the interesting facts that are helpful for decision and policy-making for stakeholders to formulate a comprehensive roadmap on hate speech, especially in Pakistan..

4.5.1 Temporal Analysis

This section is about the temporal analysis of our research dataset. The overall dataset contains the time in which the tweet has been recorded from the duration Jan 2018 – Apr 2022. The table shows the overall statistics of data w.r.t time:

Table 4.4 Temporal Analysis of Target Classes

Year	HS	Offensive	Neutral	Total
2018	**768**	2483	15727	18978
2019	**1130**	2948	16845	20923
2020	**471**	1590	16863	18924
2021	**833**	2786	20199	23818
2022	**3903**	12258	100831	116937
Total	**7101**	22229	172491	201821

Fig 4.5 depicts the class distribution w.r.t time (2018-2022). It has been observed that hate speech and offensive tweets were found more in 2019, and 2022 compared with the year 2018, 20, and 2021. The hate speech (3903 in no) tweets in the first 4 months of year 22 stand high with offensive containing 12258 tweets. The trend shows that the growth in hate speech tweets rapidly increases as we moved from 2018 -22.

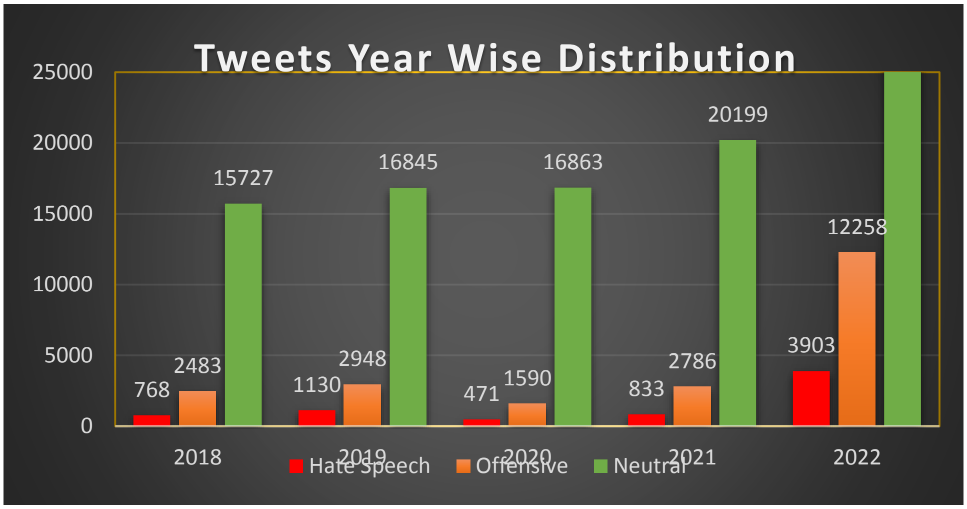


Figure 4.5 Year-Wise Class Distribution

Fig 4.6 depicts the year-wise tweets recorded. It has been observed that trend to adopt twitter in Pakistan has witnessed exponential growth in 2022.

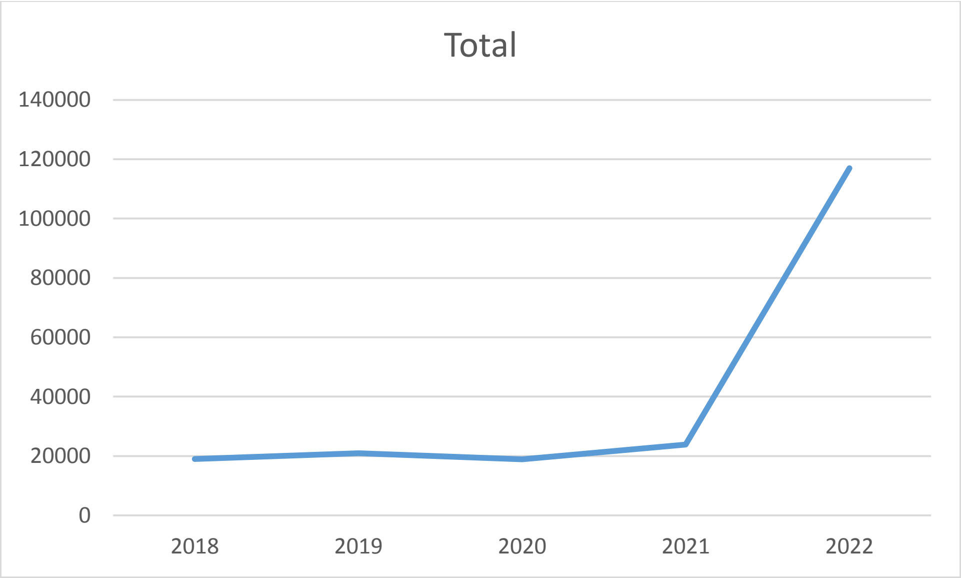


Figure4.6 Tweets Year-Wise Distribution

4.5.2 Spatial Analysis

This section represents the spatial analysis of our dataset. We extracted the Latitude and longitude of each Punjab district against the tweet. Table 4.5 shows the spatial data statistics:

Table 4.5 Spatial Data Statistics

District	Total	Hate Speech	Offensive	Neutral
Attock	**619**	14	76	529
RYK	**2319**	142	318	1859
RajanPur	**530**	21	47	462
BWP	**5386**	231	718	4435
Lodhran	**1509**	41	82	1386
Bahwalnagur	**1264**	69	225	970
Chakwal	**1513**	72	132	1309
Vehari	**4421**	80	215	4126
Chinot	**2261**	59	203	1999
DGK	**3644**	137	654	2853
FSB	**12454**	593	1722	10136
GJW	**17954**	784	2183	14987
Gujrat	**8557**	202	601	7754
Hafizabad	**1004**	72	135	797
Jhang	**1231**	36	72	1123
Jhelum	**3802**	86	266	3448
Kasur	**2619**	116	340	2163
Khanewal	**3796**	81	289	3426
Khushab	**914**	29	115	769
Lahore	**29103**	962	3341	24800
Layyah	**210**	8	14	188
Mandi	**5822**	134	449	5239
Mianwali	**2209**	89	250	1869
Multan	**22855**	596	2274	19985
Muzafargar	**1699**	97	335	1267
Nanka	**425**	13	43	369
Norwal	**507**	31	114	362
Okara	**3780**	181	593	3006
Pakpattan	**934**	83	133	718
RWP	**26433**	784	2811	22838
Sahiwal	**4174**	123	431	3620
Bakhar	**3595**	114	339	3142

Toba	**454**	26	85	343
Shiekhpura	**2631**	83	218	2330
Sargodha	**8446**	341	863	7242
Sialkot	**12769**	575	1543	10651
Total	**201821**	**7105**	**22214**	**172265**
Ratio		**3.52%**	**11.01%**	**84.50%**

A Choropleth map is a statistical map used to provide the visualization of the variable varies across the geographical location. We used the same map (**as shown in fig 4.7**) for visualization of hate speech data promulgated in Pakistan from 2018-22. It has been observed that hate speech remains high in Lahore (962), Gujranwala (784), Faisalabad(593), Rawalpindi(784), Multan(596), Sialkot(575), and Sargodha (575) districts. **Fig 4.8** shows the overall offensive tweets recorded in the Punjab district. It has been recorded that offensive tweets in Punjab districts such as Lahore (**24800**), Rawalpindi (**22838**), Multan (**19985**) Gujranwala (**14987**), Faisalabad (**10136**), and Sialkot (**10651**) stayed high, especially in 2022 the overall offensive tweets were found 12258 overall.

4.5.3 Map Generation Process

Python Plotly library is used to generate the choropleth map. Plotly is used for data analysis and visualization tools. To generate a Choropleth map, we first installed a plotly library in python. The JSON file is required to represent the geolocation on the map. Pakistan geojson file is freely available on the internet. The data to be further analyzed is saved in CSV format with the following features:

- Coordinates (Information regarding latitude and longitude)
- Geometry (Defined as Multi polygon)
- Provinces (Set as Punjab)
- District (36 districts of Punjab)
- Shape area
- Status
- Offensive and Hate Speech data information (Number of offensive/hate speech tweets observed in a labeled dataset)

The Choropleth map box function is used to draw final a choropleth map. The following parameters are set to generate the district-wise Pakistan choropleth map

- Location
- Geojson
- Color
- Hover name
- Title
- Map box style
- Center
- Zoom and opacity

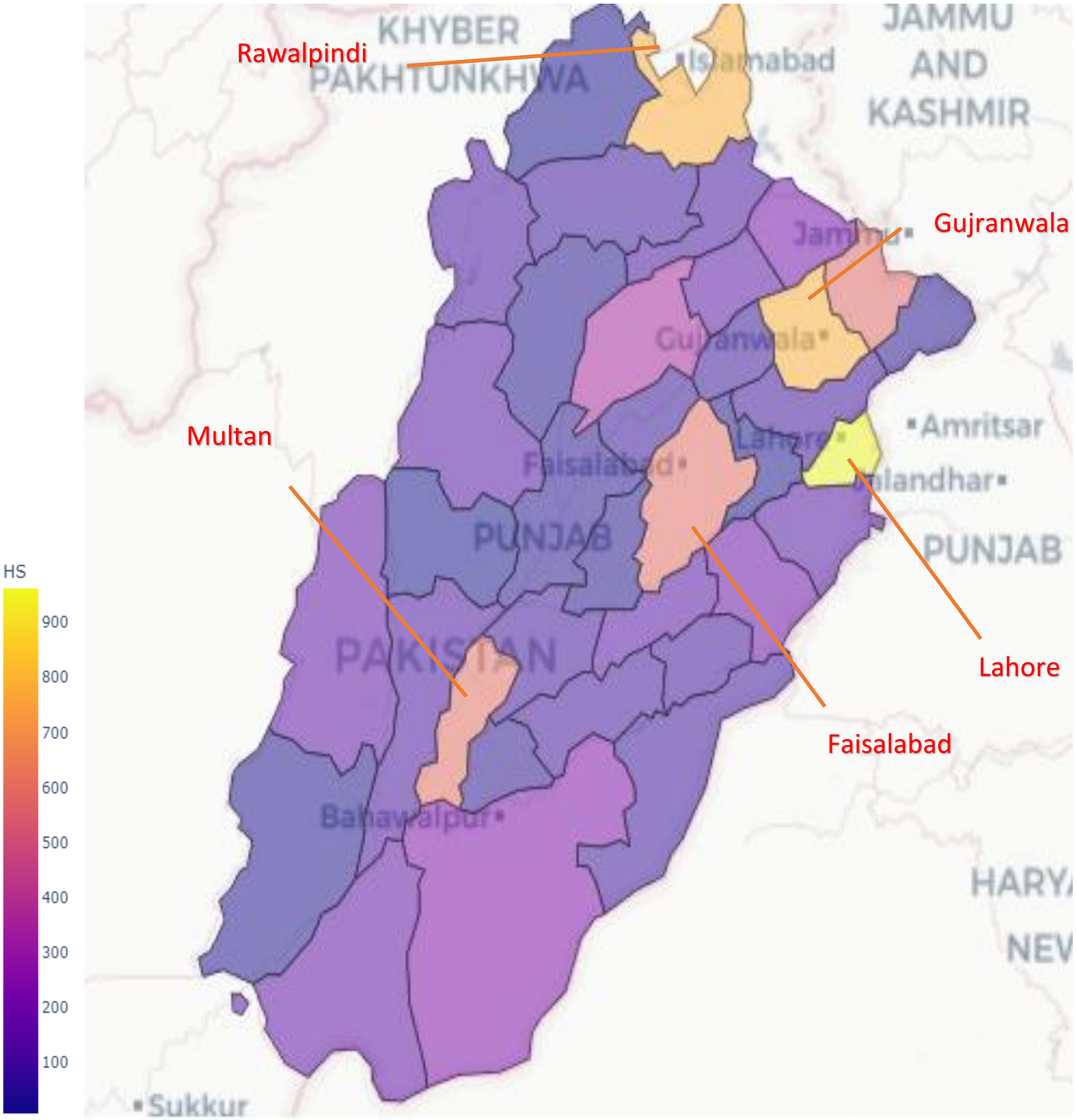


Figure 4.7 Choropleth Map for visualizing Hate Speech Data

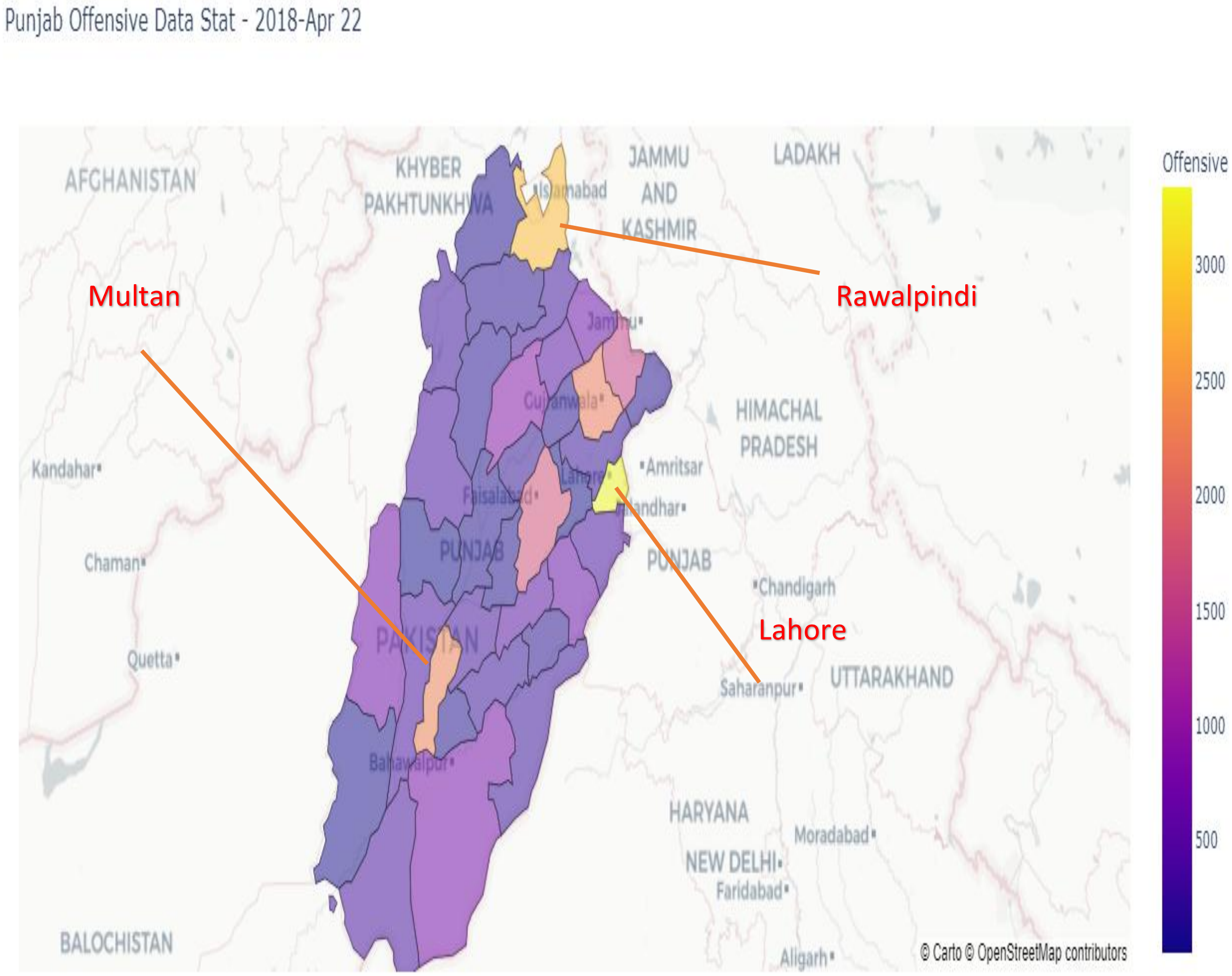


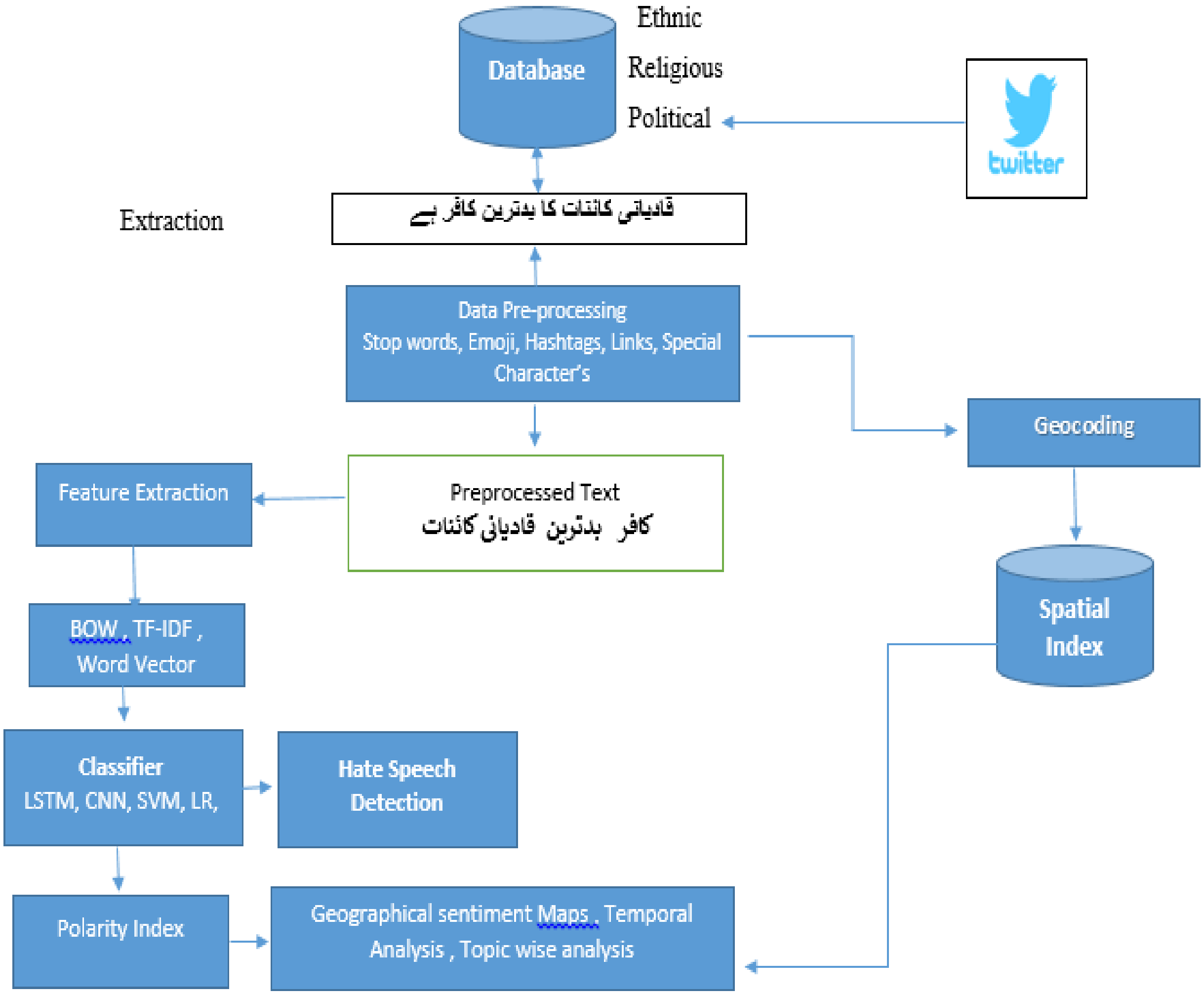
Figure 4.8 Choropleth Map for visualizing Offensive Speech Data

CHAPTER 5

Offensive Language and Hate Speech Detection System

Our research establishes a baseline model for the detection of offensive language and hates speech detection. This chapter includes detailed Methodology including data extraction (already discussed in Chapter 4), Data preprocessing, feature extraction, and classifiers, and the most important part of this methodology is spatiotemporal analysis. Fig 5.1 demonstrates the overall methodology of our research.

`



Database
Ethnic
Religious
Political
twitter
Extraction
قادیانی کائنات کا بدترین کافر ہے
Data Pre-processing
Stop words, Emoji, Hashtags, Links, Special Character's
Geocoding
Feature Extraction
Preprocessed Text
کافر بدترین قادیانی کائنات
Spatial Index
BOW , TF-IDF , Word Vector
Classifier
LSTM, CNN, SVM, LR,
Hate Speech Detection
Polarity Index
Geographical sentiment Maps , Temporal Analysis , Topic wise analysis

Figure 5.1 Proposed Methodology

5.1 Data Pre-processing

In the Urdu language, many variants are used to express the Urdu lexicon. Suffixes and prefixes are very commonly used in the Urdu language. For example, بے غیرت and بیغرت are two different words as one is used with white spaces but has the same meanings. So we have to remove white spaces to overcome the said problem. The list of tasks that were performed during the data pre-processing phase:

- We remove all white spaces from raw tweets.

- In raw tweets, we have URL links that are not contributing so we remove all hyperlinks associated with tweets for example (پیلی ٹیکسی https://t.co/ysiOJysUxW).

- Hashtags are commonly used in tweets with the purpose to identify the topic of tweets. Hashtags are very important to decide the intensity of tweets like #کلیجہ_چبانے_والی_عورت. Hashtags in Urdu are mostly used with underscore _. We initially filter the text that contains the hashtags then examined the intensity of the tweet with a hashtag and labeled the data accordingly and then remove hashtags from tweets.

- We clean the tweets by removing emojis, RT, special characters like (), $ ", user mentions, and punctuations because they do not have linguistic significance.

- We removed stop words from tweets by using the Frozen set library that contains more than 450 predefined Urdu stop words.

- We tokenized the text by separating it by a comma.

- We used an Urdu segmentation tool and stemmer to reduce the term in the base form.

- The English words are mostly used in our dataset like Imported Government Namanzoor. It was necessary to remove such English text to have a pure Urdu corpus. We used Regular expressions to remove English terms from the dataset.

5.2 Feature Extraction

Feature extraction is a technique that is used to convert raw data into numerical or vector representation by preserving meaningful information. We used the following feature extraction techniques in our research:

5.2.1 Word n – Gram

Word n-grams are used to capture consecutive perspectives. We use the word n-grams with 'n' ranging from 1 to 3 in our research. Let m represent a word in a sentence. The set M word grams can be represented as**:**

$$M = \{m1, m1\ m2, m1\ m2\ m3, m2, m2\ m3, m2\ m3, m4, \ldots.\ mt.\} \quad \textbf{(5.1)}$$

Or can be represented in form of equation 5.2

$$F1 = Mi(tf\ idf) \quad \textbf{(5.2)}$$

5.2.2 Char n – Gram

Character n-gram is used to capture the sequential context. We practice char n-grams weighted by their TF-IDF scores with 'n' from 3 to 6. Let c denote a character in a sentence. The feature set representing char (3-6) grams C can be represented as

$$\mathbf{C} = \{c1\ c2\ c3, c1\ c2\ c3\ c4, c1\ c2\ c3\ c4\ c5, c1\ c2\ c3\ c4\ c5\ c6, c2\ c3\ c4... \; ct-2\ ct-1 \quad ct\} \quad \textbf{(5.3)}$$

Or can be represent as equation 5.4

$$F2 = Ci(tf\ idf) \quad \textbf{(5.4)}$$

5.2.3 K Skip Gram

K skip grams are used to represent a context that has a long distance. We used in our Research 3-2 skip grams which results in forming a bigram of (3, 2, 1, 0 skips). The S represents the feature as shown in equation 5.5

$$S = \{w1\ w2,\ w1\ w3,\ w1\ w4,\ w2\ w3,\ ......,\ wt-1\ wt\} \qquad \textbf{5.5}$$

5.2.4 Embedding Features

Embedding is used to reduce the complexity of data by translating the data into vectors. It is very challenging to do experiments on non-numeric data. The embedding converts the high-dimensional data into low-dimensional data by preserving its meaningful information. One more benefit of embedding is that it captures the semantics from the input. The data is converted into numeric or vector form based on the distance. We used the Word2vec model in our research. We trained our large dataset containing 0.2 M tweets with the dimensions m=128. The feature vector word embedding F can be represented as shown in equation 5.6

$$F5 = \{w1e,\ w2e,\ w2e,\ ......,\ w_{ne}\}^{\ n\times m} \qquad \textbf{5.6}$$

5.3 Experiments

In our research, we labeled the data into three different classes (Hate Speech, Offensive, and Neutral). To have experiments on multiclass problems we explored different algorithms like SVM and LR with BOW and TF-IDF feature extraction techniques. Support vector machine and Logistic regression have been witnessed as useful algorithms in identifying the multi-class problem. We used SciKit learn python library for the implementation of SVM and LR. We performed 7-fold cross-validation with the combination of different random splits of data into training and testing with each feature (BOW, TF-IDF).

We used two deep learning algorithms, long short-term memory (LSTM) and Convolutional Neural Network (CNN) on our dataset. We randomly used a combination of training and split data and found the best results on training Data (90%) and Test Data (10%). We trained the data and validate the model over 10 models by considering the validation loss factor important to detect the overfitting and underfitting in the model. We used a batch size of 16 in our research.

The detailed experiments with each model are explained below:

5.3.1 SVM

Support vector machine is very efficient and useful in multi-class problems, memory efficient, and very effective in high dimensional data. The SVM takes the data points as input and output hyperplanes that best separate the points. The Hyper plane equation is represented as:

$W^t X=0$ **5.7**

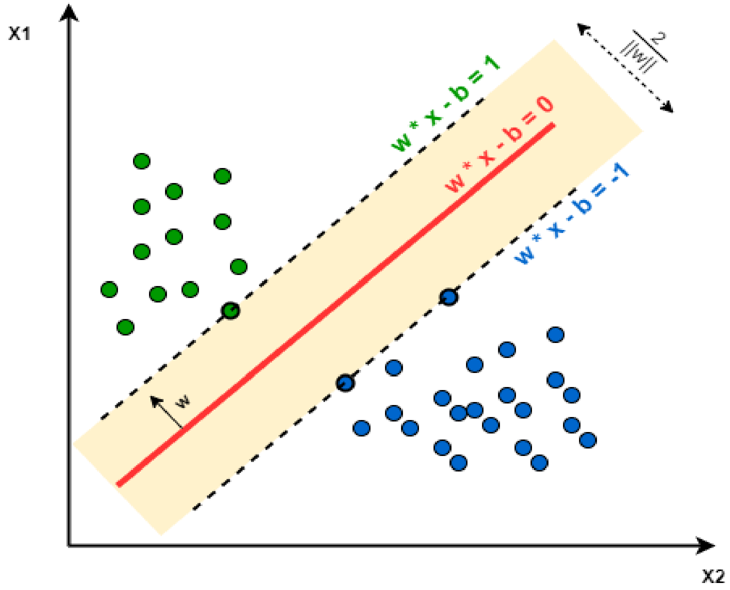
W represents the normal to hyperplanes. The Kernel function is used to calculate the data p point's separations. Given n feature vector f for three classes [1, 0,-1] the hyper plane can be defined in three equation

$w.fn + b = -1$ **5.8**

$w.fp + b = 1$ **5.9**

$w.fp + b = 1$ **5.10**

The distance between positive and negative hyperplanes is 2/||W||and the margin size is 1/|W|.



SVM Multi-Class Problems

5.3.2 LR

Logistic regression works well on independent variables. The outcome of logistic regression is the basic probability so the dependent variable remains bounded in a range between 0 and 1. For the input vector Fi, weighted matrix S, and bias values b, the probability that Fi relates to class ‘K’ is the value of the variable y which can be mathematically represented by the equation:

$$h\theta(Fi) = P(y = K|Fi\ , s, b) \qquad \mathbf{5.11}$$

Where h is the hypothesis and θ represents parameters s and b. The probabilities for the input vectors can be determined by the softmax function as represented by equation 5.12:

$$P(y = K|Fi, s, b) = \text{softmax}\ (s.Fi + b) \qquad \mathbf{5.12}$$

$$P(y = j|\mathbf{F}_i, w, b) = \frac{e^{w_j.\mathbf{F}+b_j}}{\sum_{k=1}^{k} w_k.\mathbf{F} + b_k}$$

To have a minimum loss function during training, we used stochastic average gradient descent (SAG) solver.

5.3.3 LSTM

Long short term memory consists of four layers, the Embedding layer also known as the Input layer, the LSTM layer, the dense layer, and the Output layers. The embedding layer has some predefined parameters like Input dimensions we assigned a vocab size that is 68671 for our dataset, output dimensions assigned as 64 and a maximum input length is 108 for our dataset. We used a hidden layer to have stable and effective results. Rectified linear unit (ReLu) is used in the dense layer and on the output layer Softmax function is used for prediction, the number of neurons used in these layers is equal to the number of target classes. Sparse categorical entropy is used to calculate the cost of learning algorithms. We use a callback to monitor the overfitting. We set the threshold as 3 which means if the validation loss did not change for 3 consecutive iterations the iterations automatically stops.

5.4.4 CNN

CNN has four layers. Convolutional, pooling, fully connected, and an output layer. Input layer that extracts useful information from the input for our case we set the parameter with the size of vocabulary i.e 68671 with embedding dimensions 64. We set max pooling value 2 to keep salient features. Convolutional layer that is used for useful feature extraction. We Used ReLu activation function in this layer. We use dense layers with units 1024 and 512. All extracted features are concatenated to form a feature vector and passed as input to the output layer using the Softmax activation function to classify the sentence. We set the dropout value to 0.02, the learning rate to 0.000055, and 10 epochs.

CHAPTER 6

Model Evaluation

In this chapter, we briefly discussed the metrics used to evaluate the models' efficiency and potential causes of misclassification. In research, we implemented the algorithms on both balanced and imbalanced data.

6.1 Evaluation Metrics

As our dataset contains the 0.2 M tweets that cause the imbalance problem. We have to choose such evaluation metrics that will evaluate the model correctly. In practice, Accuracy is used to evaluate the model that was tested with balanced data. For Imbalanced data, the data is highly skewed towards one class or other words biased having a majority of samples of one type that makes accuracy higher. The models perform well on majority classes but not well in detecting minority classes. We evaluate the model by individually calculating Precision, recall, and F scores against each class.

Precisions are used to measure how much results are relevant. The ratio of True positive and the sum of True positive and False positive.

$$\textbf{Precision} = \frac{\text{True Positive (TP)}}{\text{True positives + False positives}} \qquad \textbf{6.1}$$

The recall represents how many returned results are relevant. It estimates how many actual samples belonging to a certain class were correctly predicted by the model.

$$\textbf{Recall} = \frac{\text{True Positive (TP)}}{\text{True positives + False Negatives}} \qquad \textbf{6.2}$$

F score is used to evaluate the model having tested with imbalanced data. F Score is the harmonic mean of Precision and recall.

$$\textbf{F Score} = \frac{\text{2. (Precision. Recall)}}{\text{Precision + Recall}} \qquad \textbf{6.3}$$

The receiver Operating Curve (ROC) plots the false positive rate (FPR) on the x-axis and the true positive rate (TPR) on the y-axis for values between 0 and 1.

6.2 Results and Discussion - Imbalanced Data

The experiments have been performed on data having 0.2 M tweets including religious, political, and ethnic groups. Two different approaches Support Vector Machine (SVM), Logistic Regression (LR) from Machine learning and Long Short Term Memory (LSTM), Convolutional Neural Networks were used for the detection of offensive and hate speech content in a dataset. Bag of Words (Bow), TF-IDF, and word2vec are used for features engineering. The Precision, Recall, and F Scores are obtained against each and compared to the result. The result highlighted in bold represents the Highest F Score achieved against respective algorithms. **Table 6.1** shows below the results of all 3 target classes against each algorithm:

Table 6 Result of All classifier - Imbalanced Data

Classifiers	**Features**	**Neutral**			**Offensive**			**Hate Speech**		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
SVM	BOW	97	95	96	57	64	60	45	59	51
	TF-IDF	95	94	94	55	54	54	44	51	47
LR	BOW	98	94	96	57	69	**62**	48	68	**56**
	TF-IDF	96	95	96	51	61	59	51	57	54
LSTM	-	90	89	89	70	80	**75**	75	54	**64**
	Word2Vec	82	91	87	73	74	**74**	72	52	61
CNN	Word2Vec	96	96	96	60	63	62	64	47	56

The above results show that F score for the target class Neutral labeled as 0 achieved a maximum F score of 96 with all three algorithms Support vector Machine, Logistic Regression, and LSTM. LSTM Sequential model performs outclass in detecting offensive and hate speech contents in imbalanced data containing 0.2 M Tweets. The Highest F score achieved against the offensive type through LSTM sequential model is **75** and for hate speech is **64**. It has been observed during experiments all deep learning algorithms Perform well on large data because they need more data to learn, and train. The experiments through Deep learning algorithms remained outstanding as compared to traditional approaches. Fig 6.1 and 6.2 shows the boxplot depicting F score against all four classifiers.

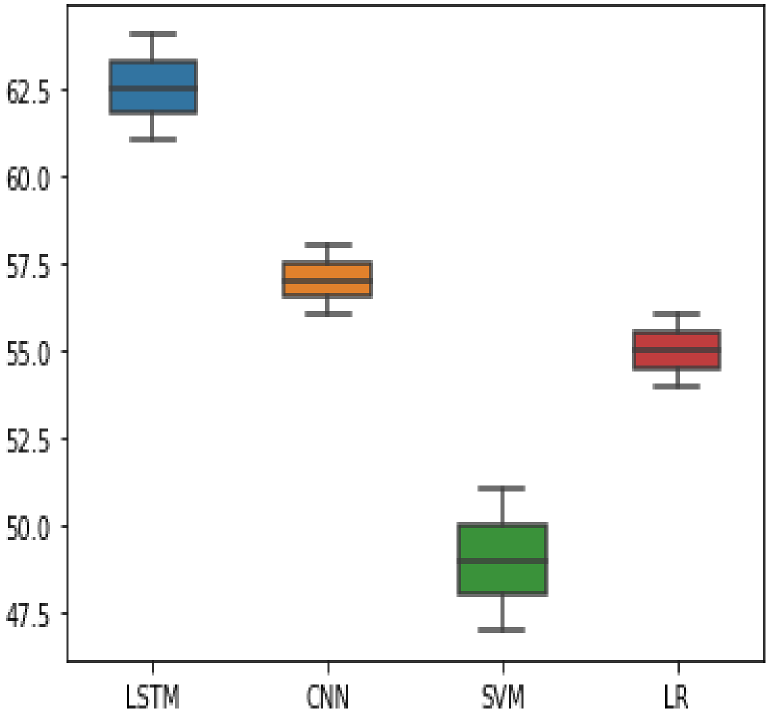


Figure 6.1 Boxplot Yield F Score against Hate Speech

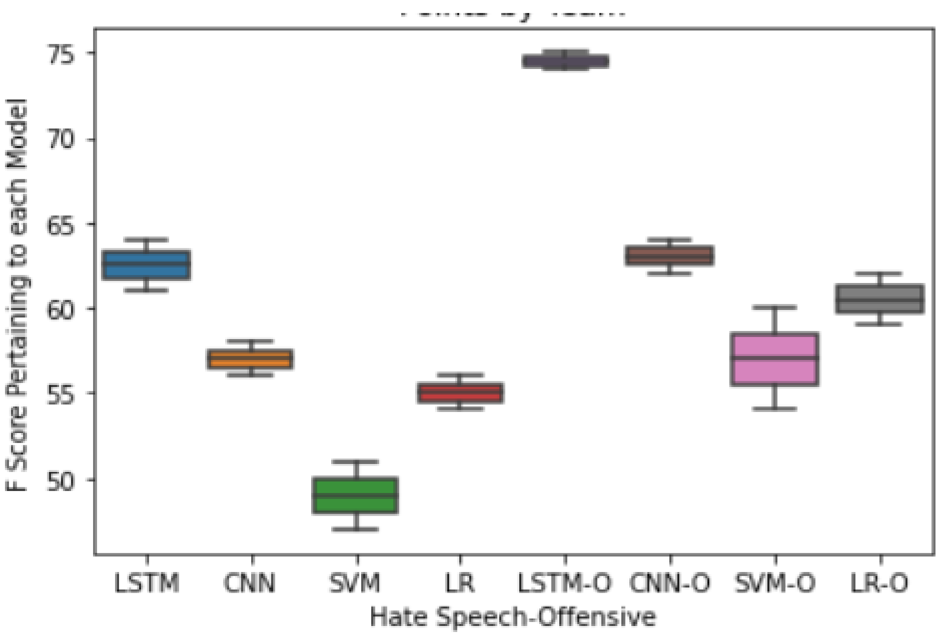


Figure 6.2 Boxplot Yield F Score against Offensive and Hate Speech

6.3 Results and Discussion - Balanced Data

The experiments were performed on balanced data for comparison of results. Oversampling and under-sampling of data have been implemented. The distribution of class in overall data is shown in Table 6.3

Table 6.3 Balanced Data Distribution

Class	**Balanced Data**	**Original**	**Operation**
Neutral	23000	172450	**Under Sampling**
Offensive	22225	22225	**-**
Hate Speech	21303	7101	**Oversampling**

Table 6.4 Result of Classifiers - Balanced Data

Classifiers	**Features**	**Neutral**			**Offensive**			**Hate Speech**		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
SVM	BOW	91	87	89	68	82	74	92	81	86
	TF-IDF	86	85	85	67	76	71	90	80	85
LR	BOW	93	87	90	72	82	**77**	88	84	86
	TF-IDF	89	87	88	71	82	**77**	93	83	88
LSTM	-	87	91	89	79	85	**82**	97	88	**92**
	Word2Vec	82	91	87	78	85	**82**	96	88	92
CNN	Word2Vec	90	85	87	84	79	81	88	99	**93**

The result depicts that improvement in achieving a High F Score for offensive and hate speech class on balance data. CNN performs well on balance data with a yielded F score **of 93.** Logistic regression and Long Short term memory models performed well in detecting offensive contents with F scores **of 77** and **82** respectively. The overall accuracy achieved against balanced data and imbalance data is shown in Table 6.5

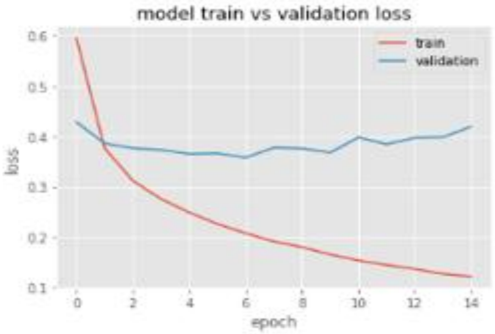
Table 6.5 Result of all classifiers for hate speech detection

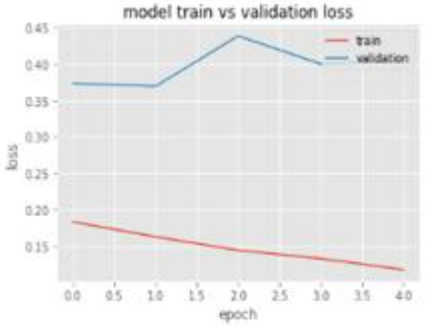
Classifiers	**Features**	**Balanced Data**	**Imbalanced**
Data		**Accuracy (Aggregated %)**	
SVM	BOW	88	**91**
	TF-IDF	87	88
LR	BOW	**89**	**91**
	TF-IDF	88	90

LSTM	-	**91**	89
LSTM	**Word2Vec**	88	87
CNN	**Word2Vec**	87	**92**

6.4 Error Analysis

We recorded the Loss during model training and validation, especially for deep learning algorithms. It has been observed that the model underwent overfitting after 7 epochs. The imbalanced factor of data caused the overfitting problems.





Chapter 7

FUTURE WORK AND CONCLUSION

Hate Speech becomes a global problem on social media nowadays. A variety of languages are used for expressing and sharing ideas on social media which makes the detection of hate speech content a challenge. Machine learning and deep learning algorithms have witnessed effective countermeasures in detecting and removal of such abusive content on social media. Several studies have been carried out on this problem, especially in the English language is the most spoken language in the world. Urdu, being a low-resource language very less amount of work has been carried out either with the small dataset or in roman Urdu. To our best knowledge, there is no work carried out on Urdu's large dataset and demographical parameters in Pakistan. To Our best knowledge, we developed a large corpus having 0.2 M tweets. The corpus is collected against 36 districts of Punjab for the period 2018- Apr 2022. The other contribution to our research is to annotate such a large dataset that takes an ample amount of time. We introduced a new definition of hate speech for our data and annotate the data accordingly. We explored the useful features of Urdu and implement the machine and deep learning algorithms. We observed that deep learning algorithms are most effective and efficient on a large dataset. Embedding features perform well in detecting infrequent patterns of hate speech. The traditional model outperforms deep learning models. It may be due to class imbalance problems, Data Sparsity, and high dimensionality and it is a challenging task to reduce and overcome the problems before moving further in the detection process. That is why we think that deep learning algorithms contribute well in this case. We carried out an error analysis of these algorithms and found challenging to make the process more effective and efficient as we encountered the overfitting problem for our dataset. Our research establishes a baseline for the detection of hate speech in the Urdu language. Future work should address the challenges identified in our research like data sparsity, High Skew, and high dimensionality problems. Another aspect is to incorporate advanced techniques to distinguish between different degrees of language such as sarcasm, implicit hate speech, word sense, and target of abuse. To annotate the large dataset it is necessary to develop a comprehensive sentiment dictionary for the Urdu language. Secondly, the focus should be on minority classes by analyzing every hidden pattern. The language sense is also important, especially for Low resource language like the implementation of word Segmentation, etc. Advanced embedding features should be applied to more data to have more effective and accurate results.

BIBLIOGRAPHY

[1] Pakistan and S. Mahmood, The Pakistan Penal Code (XLV of 1860). Legal Research Centre, 1981.

[2] https://nacta.gov.pk/wp-content/uploads/2017/08/Anti-Terrorism-Act-1997.pdf

[3] https://about.fb.com/news/2020/06/progress-fighting-hate-speech/

[4] https://www.article19.org/

[5] Burnap, P., & Williams, M.L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. Epj Data Science, 5.

[6] Waseem, Z., & Hovy, D. (2016, June). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of the NAACL student research workshop (pp. 88-93).

[7] Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516.

[8] Gambäck, B., & Sikdar, U. K. (2017, August). Using convolutional neural networks to classify hate speech. In Proceedings of the first workshop on abusive language online (pp. 85-90).

[9] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666.

[10] Koushik, G., Rajeswari, K., & Muthusamy, S. K. (2019, September). Automated hate speech detection on Twitter. In 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA) (pp. 1-4). IEEE.

[11] Kiilu, K. K., Okeyo, G., Rimiru, R., & Ogada, K. (2018). Using Naïve Bayes algorithm in the detection of hate tweets. International Journal of Scientific and Research Publications, 8(3), 99-107.

[12] Watanabe, Hajime, Mondher Bouazizi, and Tomoaki Ohtsuki. "Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection." IEEE Access 6 (2018): 13825-13835.

[13] Herwanto, G. B., Ningtyas, A. M., Nugraha, K. E., & Trisna, I. N. P. (2019, December). Hate speech and abusive language classification using fastText. In 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 69-72). IEEE.

[14] Şahi, H., Kılıç, Y., & Sağlam, R. B. (2018, September). Automated detection of hate speech towards a woman on Twitter. In 2018 3rd international conference on computer science and engineering (UBMK) (pp. 533-536). IEEE.

[15] Oriola, O., & Kotzé, E. (2020). Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. IEEE Access, 8, 21496-21509.

[16] Ginting, P. S. B., Irawan, B., & Setianingsih, C. (2019, November). Hate speech detection on Twitter using multinomial logistic regression classification method. In 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS) (pp. 105-111). IEEE.

[17] Malmasi, S., & Zampieri, M. (2017). Detecting hate speech in social media. arXiv preprint arXiv:1712.06427.

[18] De Smedt, T., De Pauw, G., & Van Ostaeyen, P. (2018). Automatic detection of online jihadist hate speech. arXiv preprint arXiv:1803.04596.

[19] Khan, M. M., Shahzad, K., & Malik, M. K. (2021). Hate speech detection in roman Urdu. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 20(1), 1-19.

[20] Ali, M. Z., Rauf, S., Javed, K., & Hussain, S. (2021). Improving hate speech detection of Urdu tweets using sentiment analysis. IEEE Access, 9, 84296-84305.

[21] Akhter, M. P., Jiangbin, Z., Naqvi, I. R., Abdelmajeed, M., & Sadiq, M. T. (2020). Automatic detection of offensive language for Urdu and roman Urdu. IEEE Access, 8, 91213-91226.

[22] Raza, M. O., Khan, Q., & Soomro, G. M. (2021). Urdu Abusive Language Detection using Machine Learning.

[23] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In Proceedings of the 26th international conference on World Wide Web companion (pp. 759-760).

[24] Elouali, A., Elberrichi, Z., & Elouali, N. (2020). Hate Speech Detection on Multilingual Twitter Using Convolutional Neural Networks. Rev. d'Intelligence Artif., 34(1), 81-88.

[25] Jiang, L., & Suzuki, Y. (2019, November). Detecting hate speech from tweets for sentiment analysis. In 2019 6th International Conference on Systems and Informatics (ICSAI) (pp. 671-676). IEEE.

[26] Gambäck, Björn, and Utpal Kumar Sikdar. "Using convolutional neural networks to classify hate-speech." Proceedings of the first workshop on abusive language online. 2017.

[27] Naseem, U., Razzak, I., & Hameed, I. A. (2019). Deep Context-Aware Embedding for Abusive and Hate Speech detection on Twitter. Aust. J. Intell. Inf. Process. Syst., 15(3), 69-76.

[28] Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X. Z. (2020). A framework for hate speech detection using deep convolutional neural network. IEEE Access, 8, 204951-204962.

[29] Setyadi, N. A., Nasrun, M., & Setianingsih, C. (2018, December). Text analysis for hate speech detection using backpropagation neural network. In 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC) (pp. 159-165). IEEE.

[30] Velankar, A., Patil, H., Gore, A., Salunke, S., & Joshi, R. (2021). Hate and offensive speech detection in Hindi and Marathi. arXiv preprint arXiv:2110.12200.

[31] Ali, R., Farooq, U., Arshad, U., Shahzad, W., & Beg, M. O. (2022). Hate speech detection on Twitter using transfer learning. Computer Speech & Language, 74, 101365.

[32] Rizwan, H., Shakeel, M. H., & Karim, A. (2020, November). Hate-speech and offensive language detection in roman Urdu. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 2512-2522).

[33] Khan, L., Amjad, A., Ashraf, N., & Chang, H. T. (2022). Multi-class sentiment analysis of Urdu text using multilingual BERT. Scientific Reports, 12(1), 1-17.