AUTOMATIC DETECTION OF OFFENSIVE LANGUAGE

FOR ROMAN PUNJABI

By

Muhammad Zeeshan Kahoot

Supervisor

Asst Prof Dr. Shibli Nisar, PhD

A thesis submitted to the faculty of Computer Software Engineering Department, 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

November 2022

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS Thesis written by **Muhammad Zeeshan Kahoot** Registration No. **00000359409**, of Military College of Signals 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 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: _____

Supervisor: Asst Prof. Dr. Shibli Nisar

Date: _____

Signature (HoD): _____

Date: _____

Signature (Dean): _____

Date: _____

ABSTRACT

Cyberbullying has emerged as a deadly nuisance for social norms and ethics, posing one of the most critical challenges for researchers to detect and monitor cyberbullying on social media platforms through the use of offensive and abusive wordings to express their annoyance or frustration. Much of the advancements have already been carried out by the linguistic community but in resource-rich languages and very little effort has been put in for resource-poor languages to develop an automatic monitoring and detection system for abhorrent remarks. The main reason is the non-availability of datasets for native/ local languages. The objective of this research work is to detect abhorrent and abusive remarks automatically for resource-poor language i.e. "Punjabi", through the Zero-Shot Learning technique, wherein the model classifies the samples without ever seeing or training examples. The seen and unseen categories are combined using Zero-Shot approaches by using auxiliary information to indicate observable differentiating/ distinguishing properties of objects. Dataset creation is the foremost task and is done manually due to its non-availability online. We collected two datasets of 0.1 Mn comments/ feedback and 1000 comments/ feedback separately from different social media platforms for "Punjabi". Manual labeling of 1000 comments as 'Offensive' and 'Non-Offensive' was done for comparison with the prediction of our proposed model based on feature extraction, to calculate the accuracy. The proposed classification by the model is done by taking the Euclidean distance from the centroid of the offensive words and the document. The zero-shot learning model gave an accuracy of 84.6% against the manually labeled dataset as 'Offensive' and 'Non-Offensive'. Moreover, automatic labeling of the 0.1 Mn dataset is also done by the proposed model. The corpus created in this work is made available for the researcher working in this domain

Keywords — Natural Language Processing, Text Mining, Automation, Deep Learning

DECLARATION

*I, Maj Muhammad Zeeshan Kahoot declare that this thesis titled "**Automatic Detection of Offensive Language for Roman Punjabi"** has not been submitted before for any degree application at NUST or any other educational Institutes. This synopsis is presented as a result of my original research.*

_____

Maj Muhammad Zeeshan Kahoot
(00000359409 / MSSE27)

DEDICATION

This thesis is dedicated to

MY FAMILY AND TEACHERS

for their love, endless support and encouragement

ACKNOWLEDGEMENTS

All praises to Allah and thanking HIM in gratitude with humility for the strengths and HIS blessings in the completion of this thesis.

I'm honored to express my sincere gratitude to my supervisor Dr. Shibli Nisar for his abutment, and counsel throughout my studies. His guidance and prolific exhortations were the beacons to complete my dissertation in a timely fashion. I'm equally honored to express my hat tip to Dr. Naima Iltaf. Without her sincere and bounteous contribution, I would not be able to achieve the milestone. I would like to pay special thanks to Assoc Prof Dr. Muhammad Waseem Iqbal and Assoc Prof Dr. Ihtesham Ul Islam, for their tremendous support and cooperation. I am also thankful to Dr. Adnan Ahmed Khan for his guidance and uplifting motivation.

Last but not the least, I am pleased to express my gratitude to all the individuals who have duly encouraged me to MS studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Motivation

Man, being a social animal, needs to communicate among the masses and express his views of love, hatred, admiration, annoyance, etc, which were initially governed by ethical and social norms. The parturition of social media and subsequent technological advancements have, on one hand, effectively provided a ready reckoner for awareness, knowledge acquisition, and communication, but conversely, influenced the methods and intent of mass communication [1]. Now, with such easily accessible free platforms of social media, it has laid a foundation and flexibility for its users to opine and comment on various happenings in society, videos, articles, products, and even people's personal lives. Subsequently, unethical behavior in society has also evolved to new heights, which cannot be overlooked [2].

Social media is serving as a podium for its users to express themselves in their local/ native languages, thus people utilize this feature more often, as they feel more comfortable and habituated, and express their natural flare for communication in the native language on social media. Another reason can be that some people are only familiar with the native language in which they speak or communicate.

Communication being undertaken on social media may include verbal, non-verbal, pictorial, emojis, etc, but the non-verbal type of communication is generally preferred by the people. Comments/ content published on social media platforms can broadly be classified into two types i.e. non-offensive or offensive. Non-offensive comments include appreciation, motivational, and general comments regarding various happenings coming from the positive pillar of society. However, offensive comments include abusive, insolent, and harsh words to express their rage/ frustration in terms of their views/ feedback about any aspect of life coming from negative people in society, thus giving birth to the phenomenon of cyberbullying [3]. An enormous rise in cyberbullying in society is due to the availability of so many easily accessible and freely available social media platforms like Twitter, YouTube, Facebook, etc. Such ever-increasing trends in cyberbullying need to be addressed sternly and as a priority as this will ruin the ethical fiber of society. Therefore, we opted to take the initiative to deal with cyberbullying being carried out in the traditional local language "Punjabi".

1.2 Contribution

The utmost need of the hour is to monitor and block offensive content on social media platforms through automatic detection systems. Resource-rich languages have traversed extensively forward, but resource-poor languages lag far behind due to the non-availability of datasets. A lot of work has already been conducted to address the cyberbullying issue for resource-rich languages like English, French, German and Arabic, etc. [4–6] because of the bulk availability of datasets online for these languages, however, resource-poor languages even lack language specific vocabularies available on the web.

Now in this study, we will be focusing on Punjabi, the resource-poor language, and discuss the work already done for resource-rich languages as a model. During the past few years, the trend of cyberbullying has significantly increased in resource-poor languages like Pashto, Urdu, Punjabi, and other local languages due to the provision of features to communicate in their native languages and exceptionally low progress achieved in resource-poor languages so far. In this research, we have proposed a model for the resource-poor language "Punjabi", which automatically detects offensive language/ words/ phrases from the feedback and an unlabeled dataset of 0.1 Mn comments/feedback, in Punjabi, is created from different social media platforms with automatic labeling done by the proposed model and will be made available online.

1.3 Goals

The main goal of our research work is to address the cyberbullying challenge through a focused and dedicated approach in a resource-poor language like Punjabi. Like other language-speaking communities, the Punjabi community is also taking the leverage of using social media platforms for expressing their views/anger/frustration related to any event/happening. The availability of the Punjabi dataset online was the major limitation during the research. To overcome this limitation, a dataset of 0.1 Mn comments/feedback in the Punjabi language, and 1000 comments/feedback have been collected from different social media platforms. 1000 comments dataset was labeled manually as 'Offensive' and 'Non-Offensive' for comparison with the proposed model prediction to calculate the proposed model's accuracy. To counter this underlying evil of cyberbullying in Punjabi, we proposed a 'Zero-Shot Learning' model for the classification of the text into 'Offensive' and 'Non-Offensive'. The accuracy of the proposed model is 84.6% against the labeling done manually. The corpus created in this work is made available online for subsequent usage by researchers in this field.

1.4 Observations

Previously, ethical and social norms of society initially governed mass communication for awareness, dissemination of information, and cultivation of knowledge. As explained above, it is inferred that social media has revolutionized the methods available and intent of mass communication [7]. Currently, social media users globally connect and communicate their views/ perceptions/ thoughts via easily accessible and freely available social media platforms such as Facebook, Twitter, YouTube, TikTok, etc. [6] with no Face-to-Face contact among its users, which empowers individuals to share their opinion without any fear and ethical norms. Such leverage so provided has made people less tolerant of their emotions/ conduct and imparts aggression to their behavior and content [8]. Thus, people use language that antagonizes the feelings of others. Hence comes the parturition of cyberbullying, which is a big challenge for researchers these days.

Social media platforms have been utilized for evil as well as good, thus requiring imposing potentially intrusive controls to encounter the violations. But the same is quite difficult to address manually due to the huge tally of data on social media platforms [9, 10]. The availability of such easily accessible and diverse social media platforms delivers the privilege to social media users for expressing their feelings and opinions in native and resource-poor languages, thus making it complicated to detect the violations. Therefore, an automated system for the detection of hate speech and offensive/ abusive language on social media has become an active area under research based on the need of the hour [11].

1.5 Social Media as a key Influencer in Cyberbullying

Easily accessible social media platforms, interactive technologies, and globally collaborative means are the main influencer in the parturition of cyberbullying worldwide as they provide a central access point to communicate for people around the globe. Individuals belonging to several geographic locations, religions, colors, creeds, cultures, and political affiliations troll each other by using offensive/ violent language [2]. To present their views/ feedback, judgment, response/ remarks about online products, videos, articles, and various happenings, people feel comfortable and favor using their native language [12]. Comments that form the basis of confrontation and conflict need to be detected, analyzed, and immediately deleted [13]. The global community is facing a big challenge to impose restrictions/ precautionary measures to control pungent cyberbullying in local/ native/ resource-poor languages as no formal mechanism is yet implemented to encounter such comments/ feedback from negative users on social media.

1.6 National Interests and Benefits

Besides considering the international and global effects of social media, the utmost need of the hour is to consider the effectiveness of social media along with its side effects nationally and on the Pakistani community in particular. Cybercrimes have recorded a momentous 83% increase in Pakistan in the last three years as claimed by "The News" dated 28th August 2021 [14]. There has been a significant increase of 24% in social media subscribers in Pakistan which has a total stack of around 46 million social media users according to a report published in February 2021 [15]. Within almost 1 year, around 55% increase has been observed in January 2022 with 71.70 million social media users according to the latest report published on 16th February 2022 [16]. The Punjabi-speaking community, being the 1st largest speaking language in Pakistan, must have contributed significantly to the above facts and figures regarding social media, the 3rd most-spoken native language in the Indian Subcontinent and the 9th largest speaking language in the world [17, 18]. Punjabi speakers are much more popular on TikTok and other social media platforms these days and must have significantly contributed to the above-mentioned 83% increase in cybercrimes prejudiced by cyberbullying on social media platforms.

1.7 Research Objective

As per the above-mentioned facts, social media aspects in our country need stern regulations related to NLP (Natural Language Processing) for local languages in Pakistan.

Punjabi is one of the most popular and widely spoken languages in Pakistan with about 80.5 million speakers which is approximately 39% of the population and about 125 million speakers around the world [19]. Nowadays the Punjabi-speaking community is facing great challenges on cyber grounds and facing the consequences of the easily accessible uncontrolled and unsupervised launch of social media platforms. Thus, making it obligatory to propose such a system that can automatically detect, alter or prohibit offensive/ aggressive comments to be published online. In this research work, we propose the Zero-Shot Learning model for the classification of the text into 'Offensive' and 'Non-Offensive' for Punjabi comments/ feedback by the Punjabi-speaking community on social media platforms and form the basis to counter cyberbullying effectively and efficiently in the Punjabi language.

1.8 Literature Review

The linguistic community has articulated the issue to detect any hate speech/ abusive or offensive language for resource-rich languages from various social media applications online [13, 20, 21] like Twitter, YouTube, Facebook, and blogs because of the focused intention and availability of resources

in the resource-rich languages. Researchers have achieved noticeable success in addressing cyberbullying issues in several resource-rich languages i.e. English, Arabic, German and Indonesian, etc. [4–6].

In the past few years, researchers have successfully employed various machine learning techniques for Natural Language Processing (NLP) for offensive language and abusive comments from social media users [22, 23].

In 2019, Gamal et al. used machine learning models with special emphasis on n-gram feature extraction techniques in Arabic to detect/ deter aggressive/ offensive comments for YouTube [24]. Similarly, the same technique and model have also been used in the Indonesian language to recognize belligerent comments from social media [5]. Schneider et al. addressed the issue of cyberbullying in the German language by using the convolutional networks technique to detect antagonistic comments on Twitter [6].

In 2019, G. I. Sigurbergsson et al. detected unscrupulous and offensive comments in the Danish and English languages using Long Short Term Memory (LSTM) and Logistic regression techniques [10, 21]. Pelicon et al. (2021) proposed a detection system for hate speech classification for the English language using zero-shot cross-lingual offensive language [25]. Akhter et al. (2020) proposed the model to utilize specific and combined n-gram techniques for the extraction of features at the character as well as word levels by applying various classifiers techniques to detect offensive/ abusive language for Urdu and Roman Urdu text comments from social media [26]. Hammad et al. (2020) presented a vocabulary of hateful words in Roman Urdu (RU) and developed a dataset of 10,012 tweets in Roman Urdu called RUHSOLD. They proposed the CNN-gram technique and novel deep learning architecture for the detection of hate speech and offensive comments from Twitter [27]. All aspects discussed above in the literature review have great significance for Natural Language Processing (NLP) and foresee much more progress to peruse the same for local languages. We performed the research work for Punjabi, an extremely popular, historical, local but resource-poor language, and focused on the underlying challenge of cyberbullying. YouTube is the most frequently used video website with millions of users around the world [28, 29], and the 2nd most visited website after Google. YouTube has over 2 billion users and billions of hours of content. People watch videos on YouTube and share their views in the comment section [30]. Likewise, the Punjabi-speaking community gives a fair amount to the viewer's stack and appeared to be highly active on YouTube with traces of cyberbullying detected in the contents/ comments on the YouTube platform. Other social media platforms, such as Facebook, offer a popular platform for individuals all over the world to communicate and share their material/ posts/ comments. TikTok is another such application that is

contributing tremendously to the uplifting graph of cyberbullying, especially, in Pakistan and other countries like India nowadays.

Punjabi, being the 9th most frequently spoken language across the world, has approx. 125 million speakers [17, 18], and the 3rd most widely spoken language in the Indo-Pak subcontinent. Nearly 39% of Pakistanis, especially in Punjab, with about 80.5 million speakers, use it as their native language. The rise in cyberbullying in the Punjabi-speaking community is also high as is done for other languages on social media platforms. Cyberbullying, in resource-poor languages like Punjabi, gets more pronounced as people express their views in the form of Unicode as well as Roman scripts. In such a case, remedial measures demand more strenuous and determined effort.

The social calamity of cyberbullying has deeply rooted among the people of Pakistan and needs to be handled at preference. Therefore, we took the responsibility to combat the social evil of cyberbullying for the Punjabi-speaking community. We do our best to formulate an instinctive (automatic) model to control the virus of cyberbullying on social media platforms in Punjabi. We attained optimistic results for each suggested model in this study and encountered the aggressive/ offensive comments in our dataset very efficiently.

1.9 Thesis Structure

This work comprises mainly 4 chapters. In chapter 1, we briefly introduce our research topic, motivation, contributions, goals, observations, national interests and benefits, and literature review. The chapter is summarized by discussing the thesis structure. In chapter 2, we present the methodology and deliberate about the proposed model and technique adopted to achieve the milestone. This chapter includes the process of data collection, flow diagrams, and a brief discussion about the data classification and purification techniques.

In chapter 3, we carried out a detailed analysis of the proposed models and discuss the outcomes along with the comparison of the outcomes of each model. Tables and graphs of comparison of the results of each model are also presented in this chapter. 4th, the very last chapter, consists of a detailed conclusion to this spectacular achievement and recommendations for future contributions in this regard.

Chapter 2

METHODOLOGY

Once decided to combat the social evil of cyberbullying in the Punjabi language, we encountered various issues, that need to be addressed as a pre-requisite to carrying out the research. Firstly, we confronted with the issue of the non-availability of Punjabi dataset online for any purpose or any automated application programming interfaces (API) to extract Punjabi language comments from any social media platform, not specific to cyberbullying, as Punjabi is a resource-poor language and there is no/ very little work done online. Therefore, the initial focus was on the compilation and collection of the Punjabi language dataset and the basic objective of our thesis is finding offensive language detection through the zero-shot learning method by associating seen and unseen classes based on the auxiliary information (feature extraction) for offensive text classification of raw comments. The complete pipeline of the model proposed in the research work is shown in Figure 2.1.

2.1 Dataset Collection and Preprocessing

The collection of datasets is the foundation for processing any machine-learning task. The first challenge that we faced while doing the research work was the collection of the Punjabi dataset online as there was no such dataset available and no automated tool/ application available to extract commented data of Punjabi language from social media platforms as per our knowledge.

2.1.1 Data collection

At first instance, an unlabeled dataset (UDS) of approx. 0.1 Mn comments in Punjabi were extracted/ collected through social media platforms like YouTube, Facebook, Instagram, Twitter, etc. Moreover, another dataset of 1000 Punjabi comments (LDS) was also collected, which was labeled manually for comparison with the predicted labels of our proposed system. Similarly, around 150 offensive words/ comments for the Punjabi language were also extracted from various social media platforms and maintained as an offensive words/ comments dictionary for the Punjabi language.

2.1.2 Data preprocessing

During the data preprocessing phase, we carried out Tokenization along with the removal of stop words and then manual labeling of the dataset of 1000 words, as the already labeled dataset for the Punjabi language was not available online.

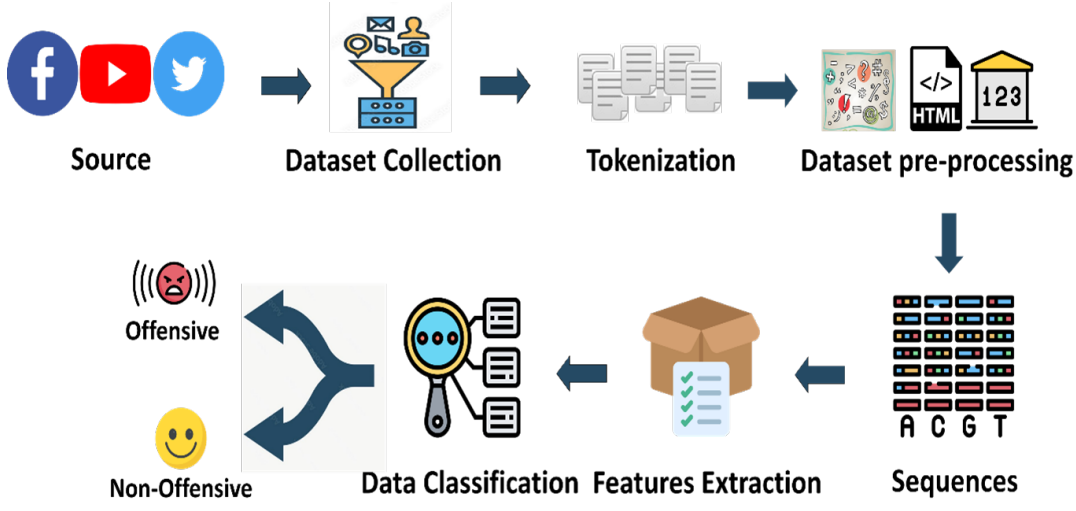

Figure 2.1: The pipeline of the proposed model

2.2 Tokenization

Tokenization is a process for any Natural Language Processing (NLP), which breaks the natural language text and unstructured data into information chunks, called 'Tokens'. Tokens are discrete elements for any structure data and form the basis for any Natural Language Processing (NLP). Then the datasets i.e. LDS, UDS, and Offensive dictionary were preprocessed and converted into the Tokens.

2.2.1 Removal of Stop Words

In preprocessing, the datasets (UDS and LDS) in raw form were refined to remove unwanted contents including punctuations, HTML codes/ commands, stop words like "is", "he", and "we" and digits, HTTP links and emoji, etc by applying preprocessing techniques to make ready our datasets for labeling after successful cleaning as they do not contribute to the classification of offensive/ abusive text.

2.3 Labeling of Datasets

The labeling of datasets for resource-poor languages like Punjabi is itself a challenging task as there is no online data of Punjabi language available which is labeled as 'Offensive' or 'Non-Offensive'. As this is a prerequisite task for our research, so we had to label the dataset manually as 'Offensive' and 'Non-Offensive'. For the accomplishment of the task, we labeled the dataset (LDS) in terms of 'TRUE' for offensive and 'FALSE' for non-offensive comments, by analyzing each comment critically and semantically. Our analysis is so deep and precise that we also considered the contexts of each comment. We are confident about the labeling of comments because it was done with human effort and understanding, based on keen observations.

Table 2.1: Examples of designing n-grams model

N-grams	**Roman Punjabi**
Sentence	‘mei Pakistan wich renda a’
Unigram	‘mei’, ‘Pakistan’, ‘wich’, ‘renda’, ‘a’
Bigram	‘mei Pakistan’, ‘Pakistan wich’, ‘wich renda’, ‘renda a’
Trigram	‘mei Pakistan wich’, ‘Pakistan wich renda’, ‘wich renda a’

2.3.1 Comments and Labels splitting

Then we carried out the splitting of comments based on the new line ‘\n’ for each dataset i.e. UDS (Unlabeled), LDS (Labeled), and Offensive Dictionary of comments/ words. Moreover, we also carried out the splitting of labels for the LDS dataset based on ‘\t’ and saved it separately in different arrays for subsequent usage. Then the datasets (UDS, LDS, and Offensive Dictionary) are structurally ready to be passed through the feature extraction phase.

2.4 Feature Extraction

The formulation of the desired model for any machine learning problem is based on the features of the dataset. Therefore, feature extraction becomes the most important aspect of machine learning and requires a thorough overview for the selection of feature extraction techniques. There are multiple feature extraction techniques available in machine learning. Better choice and efficient utilization of the technique require a detailed overview of respective machine learning techniques, covered in subsequent paras.

2.4.1 N-gram technique

An n-gram sequence is a sequence incorporating the neighboring sequences of words/ symbols/ characters/ tokens in a given document. It is the continuous sequence of the items, based on contextual and semantic analysis. It deals with the sequence of words for sentence tokenization known as ‘Word N-Gram’ or tokenizing a word in a sequence of characters known as ‘Character N-Gram’. It’s a commonly used model in NLP (Natural Language Processing) tasks, and one of its applications is sentence completion. In natural language processing (NLP), sequenced words are employed as features. It assigns probability values to the sequenced words or characters utilized in the categorization process. For text or audio categorization, classifiers employ token probabilities to predict the next item in a series and complete sentences automatically. Here is an example of designing n-grams from Punjabi sentences as shown in Table 2.1.

2.4.2 Bag of Words

'Bag of Words' is a feature extraction technique from the text, that is used for modeling in Natural Language Processing (NLP) tasks to analyze text/ documents based on the word count for word occurrence in a document. In this technique, the ordering of the words is not accounted for in a document e.g. Aslam is smarter than Akram or Akram is smarter than Aslam, will have the same Bag of Words representation.

2.4.3 TF – IDF (Term Frequency – Inverse Document Frequency)

Term Frequency – Inverse Document Frequency (TF-IDF) is a feature extraction method that incorporates the concerned text relevancy to a specific document as well as the collection of documents. 'TF' is the concerned term frequency in a document whereas 'IDF' incorporates the term relevancy factor inside the documents collection and their multiplication gives us the TF-IDF factor.

TF-IDF ('term', 'doc, 'doc collection') = TF('term', 'doc) x IDF('term', 'doc collection') TF-IDF caters to both TF, as it defines the frequency of the term 't' in a document 'd', which is directly proportional to the number of occurrences within a document, whereas IDF, defines the relevancy factor of the term 't' in a collection of documents 'D', which is inversely proportional to the frequency of the term 't' in the collection of documents 'D' (thus rarity is given the higher weightage). Hence, most common words or stop words will rank low besides appearing many times.

2.4.4 Word Index Dictionary

The textual data is converted into numerical data and considered a prerequisite to be used for machine learning algorithms, Artificial Intelligence (AI), and Natural Language Processing (NLP) algorithms. This converted numerical data will then be inputted into the system for further processing. Therefore, the datasets (UDS, LDS, and Offensive dictionary) are then converted into a sequence of numbers. This sequenced information of words thus forms the word-indexed dictionary.

2.4.5 Sequence Padding

All documents in the datasets will not be of equal lengths (size and shape), there arises a requirement to make all the documents in the corpus of equal lengths (size and shape) as vector processing is best done on the same vector dimensions for optimized results. The word-indexed dictionary contains vectors with variable lengths. Hence, we calculated the maximum length of sequenced documents for datasets and padded the sequence document to the size of the maximum length. Padding is of two types i.e. 'pre-padding' and 'post-padding'. For 'pre-padding', vectors having shorter lengths from the maximum length of the sequenced document are padded with 0's at the beginning and vice versa

for ‘post’ type padding.

2.5 GloVe Method

The gloVe is an unsupervised learning technique that is used for the conversion of words to vector representations, which exhibit the linear substructures of words to vector representations. Based on these linear structures, the learning model is trained and is dependent upon the co-occurrences of the words across the corpus. The vector similarity is then calculated based on finding the nearest neighbor in which distance formula or cosine similarity can be used.

Nearest Neighbor

The nearest neighbor can be found out based on the following:

- Distance Formula
- Cosine similarity

2.5.1 Distance Formula

The formula to find the distance between the vector representation of the words forms the basis for determining the similarity of the words contextually and semantically. But, this formula is not accurate for synonyms or homonyms. Synonyms are words having a different set of alphabets and different pronunciations but possess the same meanings e.g. (big – huge), (end – finish), and (begin – start) are some examples of synonyms but homonyms are words with the same set of alphabets or same pronunciations but different meanings e.g. (bat – bat), (desert – desert), (lie – lie), are examples of same spellings and (right – write), (rain – reign), (cat – kat) are examples of same pronunciation. The distances so calculated through the above formula for synonyms will give outcome as dissimilar words, and for homonyms as similar words, whereas the same is not true. Such statistic may not be identified correctly as it is uncommon and outside the vocabulary of the average human, thus making it impossible to be incorporated them into the proposed model design.

To compute the similarity between data points (feature representation of the words), distance metrics are used. For the Zero-Shot Learning technique of whether for classification or clustering, an efficient distance measure is required to increase the accuracy and applicability of the proposed model. The distance formula for Minkowski distance (a generalized version of Euclidean/ Manhattan Distance) is as follows:

$[\sum_1^n |P_i - Q_i|^p]^{1/p}$

where

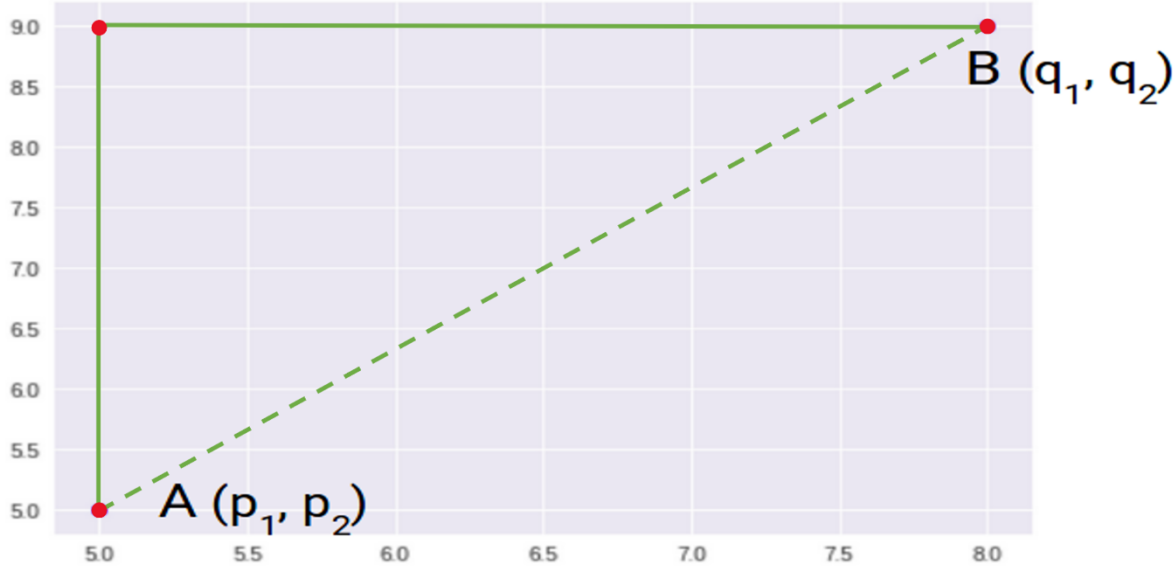

Figure 2.2: Two-dimensional Euclidean Distance between two points

- n = number of dimensions
- pi, qi = data points
- p = no of dimensions

Minkowski distance can be termed as Manhattan Distance for a one-dimensional vector when 'p' is equal to 1, however the same can be termed as Euclidean Distance for the two-dimensional vectors when 'p'= 2. We have used Euclidean distance for finding the similarities between the sequence of words for classifying the comments as 'Offensive'/ 'Non-Offensive'.

Euclidean Distance

For two-dimensional vectors, Euclidean distance is the smallest distance between two points. Numerous machine learning algorithms, including K-Means, use Euclidean distance as a distance metric for finding the similarity between vectors. Assume we have the following two points as shown in Figure 2.2.

Minkowski Distance formula will be modified into Euclidean Distance as follows (p=2):

$$d = ((p_1 - q_1)^2 + (p_2 - q_2)^2)^{1/2} \tag{2.1}$$

Euclidean Distance can be generalized for n-dimensional space as:

$D_e = [\sum_1^n |P_i - Q_i|^2]^{1/2}$

Where, 'n' is the number of dimensions $[P_i, Q_i]$ represents data points

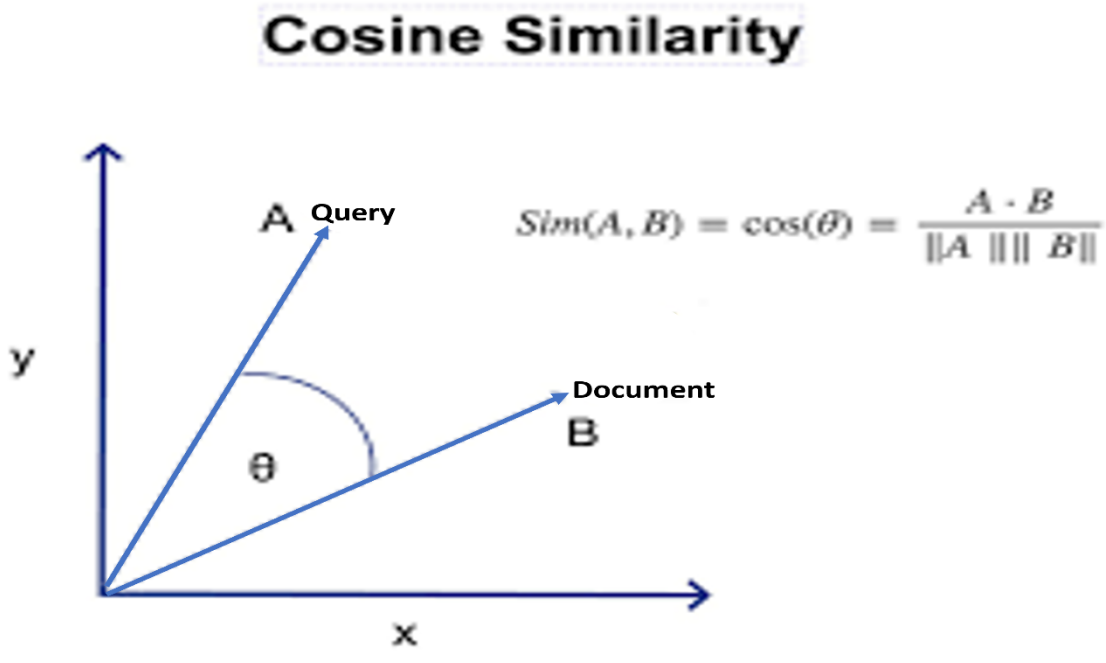

Figure 2.3: Cosine Similarity

2.5.2 Cosine Similarity

Cosine similarity is another method of measuring the similarity between two vector representations. The converted sequences of words are symbolized as vectors in a dot (inner) product. The cosine similarity is calculated by finding the dot product of the vectors (A.B) and then dividing it by the product of the magnitudes of their lengths and is represented as the cosine of the angle between the two vectors as shown in Figure 2.4. Cosine similarity depends on the cosine of the angle between the vectors and not on the magnitudes of the vectors. Vectors are similar and proportional to each other if the cosine of the angle between them is 1. When the cosine of the angle between the vectors is 0, then the vectors are orthogonal, and once the cosine of the angle between the vectors is -1, then the vectors are disjoint and opposite.

2.5.3 Linear Substructures

After going through the above discussion, it is deduced that the similarity metrics are generated to quantify the similarity of two words based on the nearest-neighbor evaluations. But words usually have similarities considering one aspect in question and differ on the other, e.g. man and woman are considered similar as both describe humans; however, they are opposite in gender. Similarly, Employee name and employee ID, Country name, and Zip Codes can refer to the same employee or country and are considered similar but the computers store words and texts differently so treat them

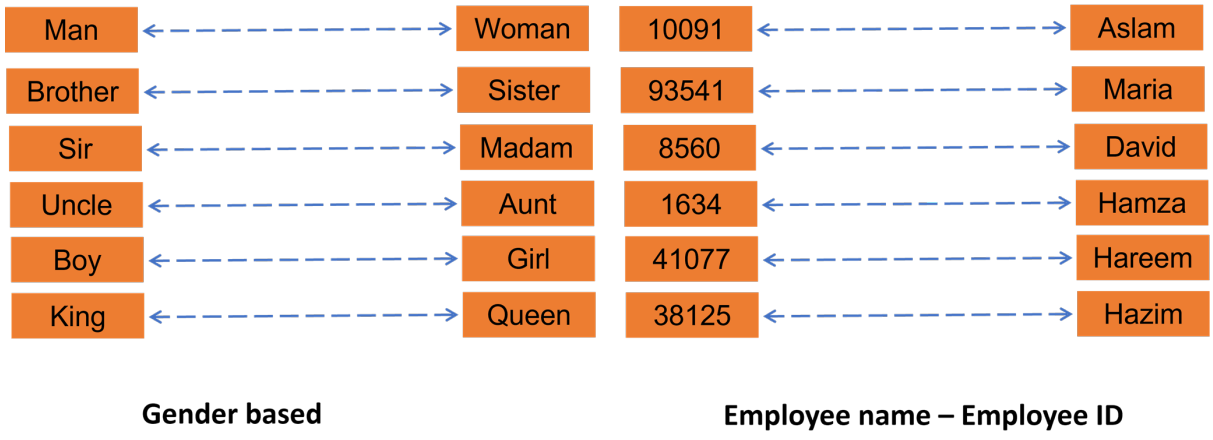

Figure 2.4: Linear Substructures

as different in this purview.

To find the distinguishing features for two vectors in question e.g. a man and a woman, the model needs to incorporate not only the difference in vector representations for the two words but also capture the contextual meaning formed by the association of the words as much as possible. Thus, the GloVe model fits best for comprehending word association. There are two major methods of the GloVe model, i.e. global matrix factorization and local context window.

Global matrix factorization is a method of reducing huge term frequency matrices using linear algebra matrix factorization algorithms in NLP. In most cases, these matrices indicate the presence or lack of words in a document. Latent Semantic Analysis (LSA) is an example of finding associations between global matrix factorizations and term frequency matrices.

The local Context Window is the prediction process for the word context based on the surrounding words by keeping the concerned word at the center and vice versa. Examples of local context windows are CBOW – "Continuous Bag of Words", and "Skip-Gram". The proposed model when envisages the centered word contextually depending upon the surrounding words is called the CBOW model, and when the model forecast the surrounding words contextually depending upon the centered word is known as the Skip-Gram model. For frequently occurring words, CBOW trains several times faster and has greater accuracy, whereas, Skip-gram shows better results in the case of rare words.

2.5.4 Training – GloVe Model

Training for any machine learning algorithm on the trained data is a prerequisite for further processing. The single pass of the corpus will save the non-zero occurrences in a word-word co-occurrence matrix as the GloVe model does not take into account the non-existent words, thus storing the frequent words' co-occurrence with each other in a given corpus. It will be cost-efficient for small

Table 2.2: Designing – GloVe Model based on probabilities

Probability and Ratio	**k=Doc1**	**k=Doc2**	**k=Doc3**
P(k/word1)	P11=P(w1,d1)	P12=P(w1,d2)	P13=P(w1,d3)
P(k/word2)	P21=P(w2,d1)	P22=P(w2,d2)	P23=P(w2,d3)
P(k/word1)| P(k/word2)	P11/ P21	P12 / P22	P13 / P23

corpus, however, for large corpus, the single pass technique will be computationally expensive. After carrying out a single pass, the later training iterations would be much faster as the non-zero matrix in a given corpus would be smaller than the actual number of words.

2.5.5 Overview – GloVe Model

The gloVe model is a weighted model of the least-squares with a log-bilinear objective, based on the ratios of probabilities of word-word co-occurrence to encode some type of meaning. Let us consider the co-occurrence probability of the word1-word2 co-occurrence in a corpus of documents as shown in Table 2.2.

Similarly, the GloVe model can train itself for many related attributes between the vectors such as the employee names to the employee IDs, Countries' names to their Zipcodes, etc, and some of the resulting vector representations illustrate various linear substructures (associations) of the words as shown in Figure 2.4.

2.6 Words Embeddings

Word embedding is the technique used to represent features of words semantically. It is a type of word representation that allows words to have comparable representations. It represents the words as a vector from a corpus for a predetermined fixed-size vocabulary. The position of the words in a vector space is determined by the words themselves and the surrounding words with similar representations. Each word is thus represented as a real-valued vector. Word embedding has a reduced distance between comparable words semantically than words having no semantic link. Another advantage of word embedding for building input to model matrices is the number of columns would be equivalent to one-hot encoding irrespective of the unique words in the text. Word embedding can be done using various techniques such as word2vec, glove, and embedding layer.

2.6.1 Word Embedding of LDS, UDS, and Offensive dictionary

We then carried out the word embedding for our datasets (Labeled dataset - LDS, Unlabeled dataset -UDS, and offensive dictionary). For Labeled Dataset (LDS), the following workflow is used for building a deep learning model from the labeled dataset and was tested for accuracy against the manual labeling of the dataset.

- Data be split into text (X) and labels (Y)
- Preprocessing of text (X)
- Word Embedding Matrix be created from the text (X)
- Tensor input be created from the text (X)
- Deep Learning models be trained using the tensor inputs and labels (Y)
- Predictions shall be made on new data

For Unlabeled Dataset (UDS), the learned model is tested against a variety of benchmark datasets and the dataset is labeled autonomously by the model based on the text classification.

2.7 Text Classification

Text classification is a natural language processing (NLP) task in which the proposed model must predict the classes of text data based on machine learning algorithms and classify the text into predetermined and predefined classes/ categories. Techniques like 'human annotator' (interprets the text substance and classify manually, thus producing excellent results but costly and time-consuming), **'Automatic text categorization'** (employs machine learning algorithms, natural language processing (NLP), and AI-based approaches in the classification of the text quickly with efficient and accurate results) are some of the methods for text classification. These classification techniques have a record of producing specialized and more accurate models but require a lot of effort and deliberation in training the model onto the labeled data and are inconsistent when a class/ classes are added to already known classes and lack the reusability of datasets for similar projects. Hence, building more intelligent generalized models without vast quantities of labeled data is the need of the hour as labeling everything in the world is difficult. Thus **'Zero-Shot Learning'** technique becomes important to be used for text classification.

2.7.1 Zero-Shot Learning

An approach in machine learning wherein the classification of the samples at the test time is done without observing the data during training. This approach is not dependent on specific labeled data but uses any labeled data of any other task/ project. The seen and unseen categories are combined based on the auxiliary information to indicate observable differentiating/ distinguishing properties of objects [1].

Zero-Shot makes pattern recognition with no training data using semantic information. As is the case for a set of images of animals possessing some auxiliary and textual descriptions of animals

Figure 2.5: An untrained system may recognize Zebra as Striped Horse

used for classification. A Zero-Shot Learning (ZSL) model trained to recognize horses if known that zebras look like striped horses (additional information), can still recognize a zebra even though zebra has never been observed during training.

The capacity to accomplish a task without any training examples is known as zero-shot learning. Consider the instance of recognizing a sort of word/ comment without ever seeing or training on such words. Recently, researchers of artificial intelligence have made significant development in building AI systems that can learn from large volumes of labeled data. The supervised learning paradigm has a record of producing specialized models that excel at the task for which they were trained. But requires a lot of effort and deliberation in training the model onto the labeled data. Moreover, the whole effort must be re-done once a class is added to already known classes, and reusability of datasets created for similar projects cannot be done as a complete dataset needs to be re-created with desired classes.

Thus, building more intelligent generalized models to execute various tasks and learn new skills without vast quantities of labeled data is the need of the hour. Practically speaking, labeling everything in the world is difficult. Some jobs, like training translation systems for resource-poor languages, simply do not have enough labeled data. If AI systems are provided with a deep and more nuanced grasp of reality besides the training of the dataset, will prove more beneficial and give results similar to human-level intelligence as shown in Figure 2.6 .

2.7.2 What is Zero-Shot Text classification?

Text classification is a natural language processing (NLP) task in which the model must predict the classes of text data. For traditional procedures such as supervised learning, a massive amount of labeled data is required for model training to predict unseen data. Zero-Shot Text classification has pushed natural language processing to new heights. The basic goal of any zero-shot text classification

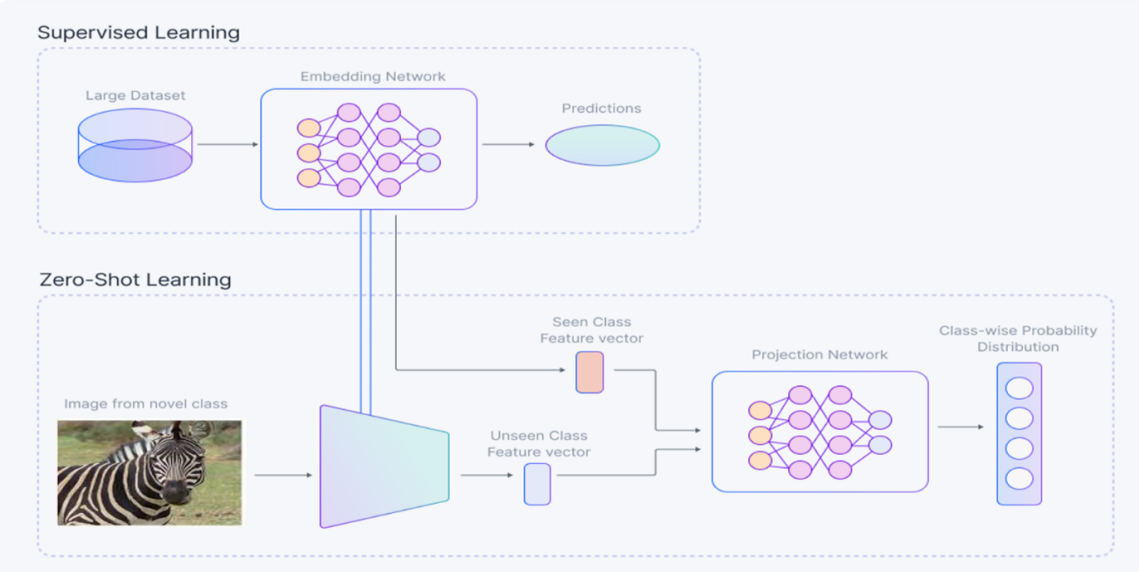

Figure 2.6: Zero-Shot Learning vs Supervised Learning

model is to classify text documents without utilizing any labeled data or having seen any tagged text. Thus, Zero-Shot Text classification is termed a combination of text classification and zero-shot learning techniques.

2.7.3 How does Zero-Shot Learning work?

The data, which is being used in Zero-Shot Learning, is as follows:

- **Seen Classes:** These are the data classes used for the training of the deep learning model.
- **Unseen Classes:** The data classes which are not used in training the deep learning model, and incorporate the generalization of the data are known as Unseen classes.
- **Auxiliary Information:** In the absence of the labeled data for the unseen classes, the Zero-Shot learning approach requires auxiliary information comprising descriptions, semantic information, and word embedding for all unseen classes.

2.8 How to choose a Zero-Shot Learning method?

To select the appropriate zero-shot learning method for choosing the appropriate technique, an overview of the following techniques is elaborated below:

2.8.1 Classifier-Based Methods

These methods include a one-versus-rest solution for training the model on a zero-shot classifier. Classifier-based methods train a binary one-versus-rest classifier for unseen classes. We divide classifier-based approaches into three subcategories based on the approach used to build classifiers.

Correspondence Method

Building the classifier for unseen classes by comparing the binary one-versus-rest classifier with its corresponding class prototype is the major goal of the correspondence method. Each class has only one associated prototype in the semantic space. As a result, this prototype might be considered the class's "representation" in the feature space, which represents a binary one-versus-rest classifier.

Relationship Method

This method involves building a classifier vis-à-vis inter-class and intra-class associations for unseen classes. The model can be trained for seen classes as per the available data in the feature space, based on binary one-versus-rest classifiers that can be learned. Meanwhile, computing these associations among related prototypes can be used to determine the relationships between seen and unseen classes. Based on the class associations and the learned binary observed class classifiers, relationship approaches tend to classify the unseen classes. The computation of finding the associations within related prototypes can be used to determine the relationships between the seen and unseen classes.

Combination Method

Combination methods are used to build a classifier for unknown classes by combining classifiers that make up the classes. Combination methods consider the most important factors which form the basis of the classification. Each of the seen and unseen classes incorporates these important factors.

2.8.2 Instance-Based Methods

Labeled instances are firstly obtained in Instance-based approaches for the unseen classes, and then train the zero-shot classifier with these instances. The existing instance-based approaches can be of the following types:

Projection Method

Projection methods generate labeled examples by projecting the feature and the semantic space instances, into a shared space for unseen classes. In the feature space, there exists labeled training instances of visible classes whereas the semantic space contains prototypes of the seen and unseen classes. These instances and prototypes represent real number spaces in the form of vectors in the feature and semantic spaces. The prototypes are labeled instances, that form the basis to classify instances in two different areas i.e. the feature space and the semantic space.

Synthesizing Methods

The labeled instances for unseen classes can be created/ synthesized using the synthesizing pseudo-instances based on the approximation of some unknown classes' distribution parameters. These in-

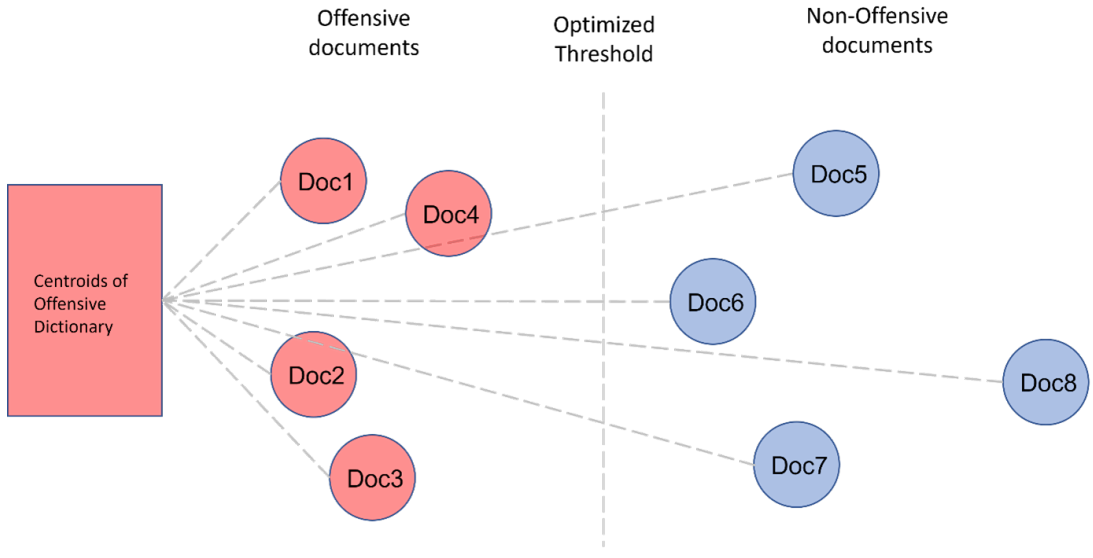

Figure 2.7: Distance between offensive dictionary centroid and each document

stances are presumed to follow some distribution parameters to synthesize the pseudo-instances.

2.9 Model Prediction

The prediction by the model used during the research is based mainly on the Adam optimizer which is one of the TensorFlow Keras optimizers. It involves a combination of two gradient descent, that exponentially descent the gradient based on the weighted average of the gradients. It converges towards the minima at a faster pace as it is computationally efficient and easy to implement [31].

2.10 Calculations

We calculated the centroids of all documents of the Labeled dataset (LDS), the Unlabeled dataset (UDS), and the centroids of the comments/ words of the offensive dictionary. Euclidean distance between the centroids of the offensive dictionary and the Labeled dataset (LDS) was calculated and then normalized on a scale of 0 and 1. Documents with smaller distances are classified as 'Offensive' while 'non-offensive' documents have greater distance values.

An optimum threshold distance value is calculated for segregating the documents into two clusters. We also captured the labels of all documents of Labeled Dataset (LDS) in a separate array for comparison with the prediction of the proposed model. Based on the comparison between the proposed system model predictions and labels of the labeled dataset, the accuracy of the model is calculated as shown in Figure 2.7.

Chapter 3

RESULTS AND DISCUSSION

The research work carried out aimed to address the issue of cyberbullying by detecting and classifying the comments as 'Offensive' or 'Non-Offensive' automatically on social media platforms. Our target was to control cyberbullying in resource-poor language i.e. Punjabi, which is spoken as the first language for almost 39% of the population (mainly in Punjab province), 3rd most spoken language in the subcontinent, and 9th most widely spoken language of the world. Much of the research work has already been carried out for resource-rich languages like English, Arabic, French, German, etc, and is publicly available for enhancing further research in these resource-rich languages. However, research for resource-poor languages still lags far behind due to the non-availability of the dataset and automated tools/ APIs to extract Punjabi language comments from social media platforms.

Therefore, the collection of data for the Punjabi language was a major challenge faced while carrying out the research. Initially, an unlabeled dataset (UDS) of Punjabi was extracted/ collected through various social media platforms. This dataset consists of a corpus of approx. 0.1 Mn documents. Another dataset of approx. 1000 documents, called labeled dataset (LDS), were also collected and manually marked after carrying out critical and semantic analysis with 'Offensive' comments as True while 'non-Offensive' as False respectively to find the proposed model's accuracy. Moreover, an offensive dictionary of approx. 150 offensive comments (a set of highly offensive words) have also been collected.

We used the Zero-Shot Learning classifier for the Punjabi language to classify offensive comments. The major advantage of using the Zero-Shot Learning classifier was its non-dependence on a very large dataset. In contrast with the supervised learning method which requires a lot of effort and deliberation in the training of the model onto the labeled data, Zero-Shot learning model training is applicable for all classes of the language, one-time effort and can easily be reused and is amongst the best techniques for resource-poor languages. For supervised learning, the whole effort has to be redone once a new class is added for already trained data and a new dataset has to be created again for any project change, besides similarities. Zero-Shot learning classification, an unsupervised learning method, takes the distance of every document in the dataset against the offensive dictionary (a set of highly offensive words). This offensive dictionary serves as a target for distance measurement

from the documents of the dataset. Zero-Shot learning classifier then classifies the documents into offensive or non-offensive clusters based on the nearest distance of each document and offensive dictionary. Offensive documents have smaller distance values and are closer to the target, while non-offensive documents have greater distance values.

Algorithm 1 Proposed Model Algorithm

Require:
0.1M Unlabeled Dataset (UDS), 1000 Labeled Dataset (LDS), Offensive Dictionary `algorithmic`.

Preprocessing
Features Extraction using Embedding model
Finding centroids of UDS, LDS, Offensive dictionary (Off Dict)
Finding averages of UDS, LDS, Off Dict ‘
Finding Euclidean distance between each document of LDS and Off Dict average
Normalizing the distances of LDS on a scale of 0 and 1
Initialization of threshold $T_{hr} = 0.4$
for `<i = 1 to 1000 >` **do**
    if Dist I > T_{hr} **then** Assign 0 to label
        if
            thenAssign 1
        end if

    Diff = label – LB(manually assigned label)

Ensure: Accuracy

We carried out the comparisons of accuracies for datasets. For optimum results and clear segregation of labels (offensive/ non-offensive) based on the distances, a threshold distance value is selected to decide the cluster for each document. A random threshold value is initially specified for the evaluation of the classifier model. The new categorization of the LDS documents into offensive and non-offensive labels is performed based on these clusters. The comparison of these labels is carried out against already assigned labels which were done manually based on critical semantic and contextual analysis of each LDS document and the accuracy of our proposed model is calculated. As for threshold values equal to 0.4, the accuracy of our proposed system is 84.6% whereas, for values equal to 0.7, accuracy is less than 60% (59.7%). It is quite evident from Figure 8 below, that by increasing the threshold value, the performance of the proposed system is decreasing.

During the process of research, we carried out a dedicated and focused analysis of the findings of the research and opines as per the following paras. Availability of labeled datasets for any language for NLP-related tasks is a must and is the most challenging task for resource-poor languages like Punjabi. This deficiency gets further worsened due to the non-availability of applications to automat-

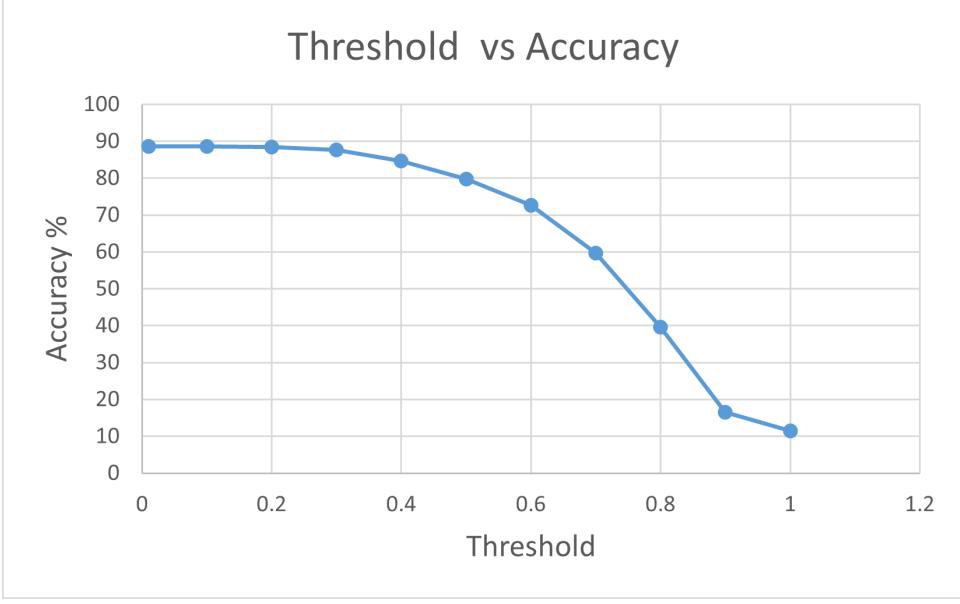

Figure 3.1: Threshold Vs Accuracy comparison

ically extract comments from social media platforms. Such an issue can only be overcome with the increase of focus by the researchers on resource-poor languages.

The availability of an increased Punjabi language dataset and offensive words dictionary for resource-poor languages is another important issue, for which compilation of the same needs to be done as a priority and made available online for further research in the field as well as the evaluation of the proposed model's performance based on enhanced parameters.

Selection of the most optimum text classification technique based on accuracy and effort is another important aspect. As for traditional text classification, supervised learning is the best technique to be used to get better accuracy but requires a lot of effort, time, and deliberation in the training of the model onto the labeled data, and the whole effort has to be redone once a new class is added for already trained data and a new dataset has to be created again for any project change, besides similarities. But the latest trending technique of Zero-Shot Learning gives optimum results as regards accuracy and effort. Zero-Shot learning model training is applicable for all classes of the language, one-time effort and can easily be reused, and is one of the best techniques for low-resource languages. The major advantage of using the Zero-Shot Learning classifier was its non-dependence on a very large dataset. Last but not the least, we opine that an increase in the offensive dictionary size and the Unlabeled Dataset (UDS) will cause an increase in the accuracy of the model. Moreover, the issue of dialects in resource-poor languages like Punjabi, which has more than 25 dialects, is another major limitation in improving the proposed model's accuracy.

Furthermore, the incorporation of synonyms and homonyms for any NLP task in finding the sim-

ilarity amongst the words, for any language, to achieve maximum accuracy and bringing NLP as nearer to human intelligence as possible can be another important aspect for future study.

Chapter 4

FUTURE WORK AND CONCLUSION

In this research work, we propose an automatic monitoring and detection system for abhorrent remarks in the Punjabi language. It is found that no dataset for the Punjabi language is available online and no direct Application Programming Interface (API) is available to extract Punjabi comments automatically from social media platforms as per our knowledge. One dataset of about 0.1 Mn and another dataset of 1000 comments of Punjabi from Facebook, Twitter, and YouTube were collected. A major contribution of this work is the collection and compilation of the Punjabi dataset. A model for the detection of abusive and unethical comments in Punjabi, a resource-poor language has been proposed in the research work. 1000 comments dataset was labeled manually as 'Offensive' and 'Non-Offensive' for comparison with the proposed model prediction to calculate the proposed model's accuracy.

The model is using Zero-Shot Learning approach, wherein the classification of the samples at the test time is done without observing the data during training. The seen and unseen categories are combined based on the auxiliary information to indicate observable differentiating/ distinguishing properties of objects. The basic aim of using the Zero-shot learning technique for our model is to classify text documents without observing labeled data which is not available so far online as per our knowledge. Moreover, the traditional classification techniques require a lot of effort and deliberation in training the model onto the labeled data and are inconsistent when a class/ classes are added to already known classes and lack the reusability of datasets for similar projects. Hence, building more intelligent generalized models without vast quantities of labeled data is the need of the hour as labeling everything in the world is difficult.

After preprocessing the datasets, features were extracted using the word embedding technique. The proposed classification by the model is done by taking the Euclidean distance from the centroid of the offensive words and the document. The zero-shot learning model gave an accuracy of 84.6% against the manually labeled dataset as 'Offensive' and 'Non-Offensive'. Moreover, the 0.1 Mn dataset of Punjabi language has automatically been labeled as 'Offensive' and 'Non-Offensive'. The corpus created in this work is made available online for subsequent usage by researchers in this field.

During the research, we faced many limitations which need to be addressed for optimum results in

combatting cyberbullying in resource-poor languages like “Punjabi”. The suggested future work can be as follows: Punjabi is comprised of various dialects. The association of words of various dialects needs special attention to evaluate the proposed model’s performance using the Zero-Shot Learning approach.

Efforts shall be done to increase the Punjabi language dataset and the offensive dictionary of the Punjabi language to improve the proposed model’s performance against an increased Punjabi language dataset and the offensive dictionary. Another important future work can be to incorporate synonyms and homonyms for finding similarity amongst the words for any NLP task, for any language, which is one of the major hindrances in getting the maximum accuracy and being as nearer to human intelligence as possible.

Bibliography

[1] S. Ray, "Understanding support vector machine (svm) algorithm from examples (along with code), 2017," 2021.

[2] S. Lewandowsky, M. Jetter, and U. K. Ecker, "Using the president's tweets to understand political diversion in the age of social media," *Nature communications*, vol. 11, no. 1, pp. 1–12, 2020.

[3] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 3, pp. 1–30, 2012.

[4] A. Alakrot, L. Murray, and N. S. Nikolov, "Towards accurate detection of offensive language in online communication in arabic," *Procedia computer science*, vol. 142, pp. 315–320, 2018.

[5] M. O. Ibrohim and I. Budi, "A dataset and preliminaries study for abusive language detection in indonesian social media," *Procedia Computer Science*, vol. 135, pp. 222–229, 2018.

[6] J. M. Schneider, R. Roller, P. Bourgonje, S. Hegele, and G. Rehm, "Towards the automatic classification of offensive language and related phenomena in german tweets," 2018.

[7] A. Kumar, S. Saumya, and J. P. Singh, "Nitp-ai-nlp@ hasoc-dravidian-codemix-fire2020: A machine learning approach to identify offensive languages from dravidian code-mixed text." in *FIRE (Working Notes)*, 2020, pp. 384–390.

[8] S. Kiritchenko, I. Nejadgholi, and K. C. Fraser, "Confronting abusive language online: A survey from the ethical and human rights perspective," *Journal of Artificial Intelligence Research*, vol. 71, pp. 431–478, 2021.

[9] F. Husain and O. Uzuner, "A survey of offensive language detection for the arabic language," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 20, no. 1, pp. 1–44, 2021.

[10] G. I. Sigurbergsson and L. Derczynski, "Offensive language and hate speech detection for danish," *arXiv preprint arXiv:1908.04531*, 2019.

[11] D. Ali and L. Xiaoying, "The influence of content and non-content cues of tourism information quality on the creation of destination image in social media: A study of khyber pakhtunkhwa, pakistan," *Liberal Arts and Social Sciences International Journal (LASSIJ)*, vol. 5, no. 1, pp. 245–265, 2021.

[12] Z. Torwali, "Language documentation and description."

[13] F. A. Vargas, I. Carvalho, F. R. de Góes, F. Benevenuto, and T. A. S. Pardo, "Building an expert annotated corpus of brazilian instagram comments for hate speech and offensive language detection," *arXiv preprint arXiv:2103.14972*, 2021.

[14] "Cybercrime increases by 83pc in three years — thenews.com.pk," https://www.thenews.com.pk/print/884453-cybercrime-increases-by-83pc-in-three-years, [Accessed 26-Oct-2022].

[15] S. Kemp, "Digital in Pakistan: All the Statistics You Need in 2021 — DataReportal – Global Digital Insights — datareportal.com," https://datareportal.com/reports/digital-2021-pakistan#:~:text=Social%20media%20statistics%20for%20Pakistan,total%20population%20in%20January%202021., [Accessed 26-Oct-2022].

[16] "Digital 2022 Pakistan (February 2022) v01 — slideshare.net," https://www.slideshare.net/DataReportal/digital-2022-pakistan-february-2022-v01-251182073, [Accessed 26-Oct-2022].

[17] "Punjabi language - Wikipedia — en.wikipedia.org," https://en.wikipedia.org/wiki/Punjabi_language#:~:text=Punjabi%20is%20the%20most%20widely%20spoken%20language%20in%20Pakistan%2C%20being,39%25%20of%20the%20country's%20population, [Accessed 26-Oct-2022].

[18] U. L. Group, "6 Surprising Facts About Punjabi Translation — unitedlanguagegroup.com," https://www.unitedlanguagegroup.com/blog/six-surprising-facts-about-punjabi-translation#:~:text=Punjabi%20is%20the%2010th%20most,than%20German%2C%20Korean%20or%20French, [Accessed 26-Oct-2022].

[19] F. Abbas, M. N. Chohan, M. Ahmed, and M. Kaleem, "Punjabi language in pakistan: Past, present and future," *Hamdard Islamicus*, vol. 39, no. 3&4, pp. 1–14, 2016.

[20] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE access*, vol. 6, pp. 13 825–13 835, 2018.

[21] A. Bisht, A. Singh, H. Bhadauria, J. Virmani *et al.*, "Detection of hate speech and offensive language in twitter data using lstm model," in *Recent trends in image and signal processing in computer vision*. Springer, 2020, pp. 243–264.

[22] D. Bhimani, R. Bheda, F. Dharamshi, D. Nikumbh, and P. Abhyankar, "Identification of hate speech using natural language processing and machine learning," in *2021 2nd Global Conference for Advancement in Technology (GCAT)*. IEEE, 2021, pp. 1–4.

[23] M. H. U. Rahman, M. Divya, B. R. Reddy, K. S. Kumar, and P. R. Vani, "Cyberbullying detection using natural language processing."

[24] D. Gamal, M. Alfonse, E.-S. M. El-Horbaty, and A.-B. M. Salem, "Implementation of machine learning algorithms in arabic sentiment analysis using n-gram features," *Procedia Computer Science*, vol. 154, pp. 332–340, 2019.

[25] A. Pelicon, R. Shekhar, M. Martinc, B. Škrlj, M. Purver, S. Pollak *et al.*, "Zero-shot cross-lingual content filtering: Offensive language and hate speech detection," 2021.

[26] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, and M. T. Sadiq, "Automatic detection of offensive language for urdu and roman urdu," *IEEE Access*, vol. 8, pp. 91 213–91 226, 2020.

[27] H. Rizwan, M. H. Shakeel, and A. Karim, "Hate-speech and offensive language detection in roman urdu," in *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2020, pp. 2512–2522.

[28] C. S. Park, Q. Liu, and B. K. Kaye, "Analysis of ageism, sexism, and ableism in user comments on youtube videos about climate activist greta thunberg," *Social Media+ Society*, vol. 7, no. 3, p. 20563051211036059, 2021.

[29] C. E. Basch, C. H. Basch, G. C. Hillyer, and C. Jaime, "The role of youtube and the entertainment industry in saving lives by educating and mobilizing the public to adopt behaviors for community mitigation of covid-19: successive sampling design study," *JMIR public health and surveillance*, vol. 6, no. 2, p. e19145, 2020.

[30] P. Snickars and P. Vonderau, *The youtube reader*. Kungliga biblioteket, 2009.

[31] A. G. Programmer, "Logistic regression and Keras for classification » AI Geek Programmer — aigeekprogrammer.com," https://aigeekprogrammer.com/binary-classification-using-logistic-regression-and-keras/, [Accessed 26-Oct-2022].