Sentiment Analysis for Code-Switched Languages

A Project Report

Submitted in partial fulfillment of requirements for the course CS626 of

Master of Technology

By

Ankita Singh
Roll No.: 19305R002
Deepti Mittal
Roll No.: 193050025

under the guidance of

Prof. Pushpak Bhattacharya

Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

December, 2020

Acknowledgment

We would like to thank **Prof. Pushpak Bhattacharyya** for always helping us and pointing us in the right direction whenever we were stuck. We can't thank enough my parents for their constant encouragement and support. We am also very grateful to my peers and my seniors for all the guidance and discussions held through out the semester. we would also like to thank everyone in the CFILT LAB, IIT Bombay for their help and guidance and CSE department for providing such high-quality education.

Abstract

Sentiment Analysis deals with the automatic detection of opinion orientation in text.Mixing languages, also known as code- mixing, is a norm in multilingual societies. Multilingual people, who are non-native English speakers, tend to code-mix using English-based phonetic typing and the insertion of anglicisms in their main language. In addition to mixing languages at the sentence level, it is fairly common to find the code-mixing behavior at the word level. This linguistic phenomenon poses a great challenge to conventional NLP systems, which currently rely on monolingual resources to handle the combination of multiple languages. In this we tried to perform the task of sentiment analysis for code switched language i.e., Hindi and English.

Contents

List of Figures

Chapter 1

Introduction

1.1 Problem Statement

Input: Given tweets in Code Switched Language,i.e., Hindi and English. Output the sentiment expressed by the given tweet. Sentiment can be positive, negative or neutral.
Output: Output the Sentiment bit of given data.
The task is to figure out the sentiment expressed by the data.
Example 1: *"bholy bhayaa. Ufffff dil jeet liya ap ne. Love you imran bhai. Mind blowing ap ki acting hai."*
The sentiment expressed by this statement is positive since it involve praising of the actor in some series.

1.2 Motivation

With the rapid penetration of the Internet into people's daily life, online consumption has grown in popularity. Its emergence and pervasiveness provides opportunities for new types of communication between customers and service providers. People tend to share their consumption experiences and express their opinions with regard to any product or service via e-commerce websites, following the consumption behavior.

Naturally, code-mixing is more common in geographical regions with a high percentage of bi- or multilingual speakers, such as in Texas and California in the US, Hong Kong and Macao in China, many European and African countries, and the countries in South-East Asia. Multilingualism and code-mixing are also widespread in India, which has more than 400 languages with about 30 languages having more than 1 million speakers. Language diversity and dialect changes trigger Indians to frequently change and mix languages, particularly in speech and social media contexts. So study of the sentiment analysis for code mixed language is also quite important.

1.3 Sentiment Analysis

As per the book[Liu et al., 2010],Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes.

Although NLP and linguistics have a large history but this field was least explored field before 2010. But, with the growth of social media, large growth occurred in field of sentiment analysis. The major reason was expressing of views, opinions and emotions on blogs, twitter,product forum and on other social media websites by the users.Many organization took this as an opportunity to know what people feel about their product and services.

Therefore, they thought of using data of people's opinions available on the web for the improvements.Also, Opinion-gathering from public and consumers has been a long business itself for marketing and political- party campaign companies. In fact, we can say sentiment analysis is now right at the center of social media research.

1.4 Road Map

The report discusses various resources and techniques required for the sentiment and emotion analysis. Chapter 2 describes the basics of sentiment analysis Chapter 3 discuss about the dataset used in the implementation.Chapter 4 discusses about the extraction of features for the purpose and the experiments and findings of the work. we also talk about the error analysis done on the work. Finally chapter 5 has the conclusion and future work for the report.

Chapter 2

Introduction to Sentiment Analysis

In this chapter, we define abstraction of the sentiment analysis or opinion mining problem. this abstraction gives us a statement of the problem which helps us in understanding the inter-related sub problems in the task of sentiment analysis.

2.1 Definition

Sentiment Analysis is the task of detection of sentiment of the user about products, services and other entities. Here, the major focus is sentiment analysis from textual data. There are basically 3 types of sentiment:

1. **Positive**
   Example: *Coke tastes very good.*

2. **Negative**
   Example: *This camera sucks.*

3. **Neutral**
   Example: *I purchased this mobile last week.*

2.2 Different Dimensions of Sentiment Analysis

Sentiment analysis can be viewed as a five dimension problem on the basis of literature. It can be shown as follows:

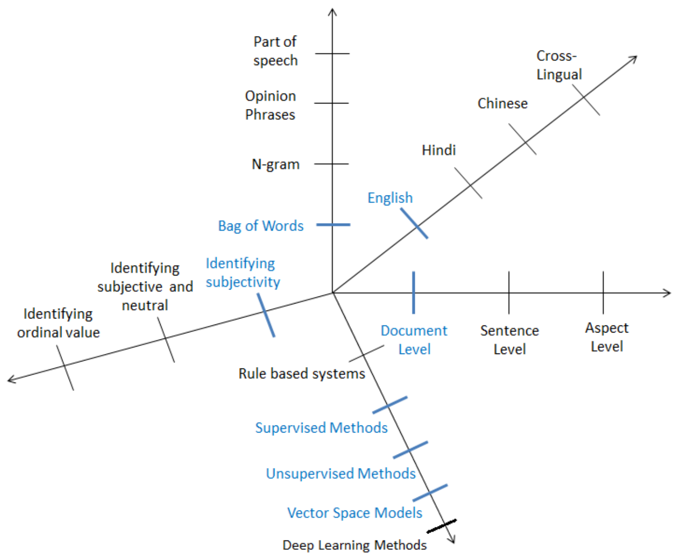

Figure 2.1: Different dimensions of sentiment analysis based on literature[1]

2.3 Different Levels of Analysis

In general, sentiment analysis has been investigated at three levels on the basis of granularity of research.

2.3.1 Document Level

In this level, the task is to extract the opinion for the whole document. This level of analysis assumes that document contains a single entity for expressing

[1]https://www.researchgate.net/figure/The-figure-showing-different-dimension-of-sentiment-analysis-based-on-literature-The_fig1_320471023

opinions.
e.g for the product review given by a customer, The task is to classify whether the whole document expresses positive, negative or neutral sentiment about the product.

2.3.2 Sentence Level

In this level, analysis is done at sentence level to extract the sentiment. Each sentence expresses a positive, negative or neutral sentiment. This level of analysis is also known as *Subjectivity classification*, where subjective statements(statements which expresses sentiment and opinions) are distinguished from objective statements (also known as factual statements).
e.g. *Apple is doing well in this lousy economy.* The sentence expresses *positive* sentiment about the growth of Apple company.

2.3.3 Entity and Aspect Level

It is possible that a single sentence can express more than one sentiment on various features(also known as aspect) of a single product. It assumes that an opinion contains both sentiment(positive, negative or neutral) and target(of opinion).
Aspect level performs finer-grained analysis.It ignores the language constructs like documents, paragraphs,sentences etc. The goal of this level of analysis is to discover sentiments on entities and/or their aspects.
E.g. *"The picture quality of the camera is too good but the lens is heavy."*
The sentence expresses *positive* sentiment about the 'picture quality' aspect of the entity camera but *negative* sentiment towards the 'weight of the lens'.

2.4 Opinion & its representation

As per the book[Liu et al., 2010],It is a view, statement or judgement of a person towards a product, service or institution. The opinion is subjective in nature.

Opinion is represented/defined in two ways.

1. **Opinion as a quadruple** (g, s, h, t)**:**
   where, **g** is the opinion or sentiment target, **s** is the sentiment about the target, **h** is the opinion holder and **t** is the time when opinion was expressed.
   Example: *"The picture quality of canon G12 camera is actually good"*
   Here, the opinion-target **g** is canon G12, and **s** is positive.

2. **Opinion as quintuple** $(e_i, a_{i,j}, s_{i,j,k,l}, h_k, t_l)$**:**
   where e_i is the name of an entity, $a_{i,j}$ is the aspect of entity, h_k is the opinion holder and t_l is the time when opinion is expressed by h_k. The sentiment $s_{i,j,k,l}$ is positive,negative or neutral.
   Example: *"The picture quality of canon G12 camera is actually good but it is too heavy".*
   Here, the entity is e_i is canon G12, and $a_{i,j}$is 'picture quality' and sentiment $s_{i,j,k,l}$ with the aspect is 'positive', whereas, another aspect $a_{i,z}$ is 'weight of camera' and sentiment $s_{i,z,k,l}$ with the aspect is 'negative'.

With the definition of opinion, we can move towards defining various types of opinion. Next section discuss about various types of opinion:

2.5 Summary

In this manner, task of sentiment analysis is done on the textual data. In this chapter, we covered a brief overview of sentiment analysis of textual data.

Chapter 3

DataSets

3.1 Abstract

In this chapter, we will discuss the data set and its statistics used for experimentation.

3.2 Code Switched dataset with Hindi and English Languages

We conducted our experiments on sentiment analysis on a benchmark dataset made available by Codalab for sentimix competition [Patwa et al., 2020] in 2019.
The datasets consist of tweets labeled into one of the three classes:

- **Positive (Pos)**: Tweets which express happiness, praise a person, group, country or a product, or applaud something. Hinglish example: "bholy bhayaa. Ufffff dil jeet liya ap ne. Love you imran bhai. Mind blowing ap ki acting hai." (bholy bhayaa, you won hearts. love you imran bhai your acting is mind blowing).
- **Negative (Neg)**:Tweets which attack a person, group, product or country, express disgust or unhappiness towards something, or criticize

something. Hinglish example: “You efficiency of anchoring a program is continuously deteriorating. Ab to dekhne ki himmat hi nahi” (Your efficiency of anchoring is continuously deteriorating. Now can’t even dare to watch it)

- **Neutral (Neu):** Tweets which state facts, give news or are advertisements. In general those which don’t fall into the above 2 categories. Hinglish example: “Nahi wo is news ko defend kerne ki koshesh ker rhe hain h” (No, they are trying to defend this news).

3.3 Statistics of data

- The statistics for the dataset can be given as follows:

Language	Split	Total	Positive	Neutral	Negative
Hinglish	Train	14,000	4,634 (33.10%)	5,264 (37.60%)	4102 (29.30%)
	Validation	3,000	982 (32.73%)	1,128 (37.60%)	890 (29.67%)
	Test	3,000	1,000 (33.33%)	1,100 (36.67%)	900 (30%)
	Total	20,000	6,616 (33.08%)	7,492 (37.46%)	5892 (29.46%)

Figure 3.1: Dataset Snapshot[Patra et al., 2018]

- These are most frequently occurred negative and positive words in the dataset.

- A pictorial representation of hindi words present in the dataset.

This chapter covers the brief description of the dataset and its statistics. In next chapter we will discuss about related work done in the field of code-switched languages.

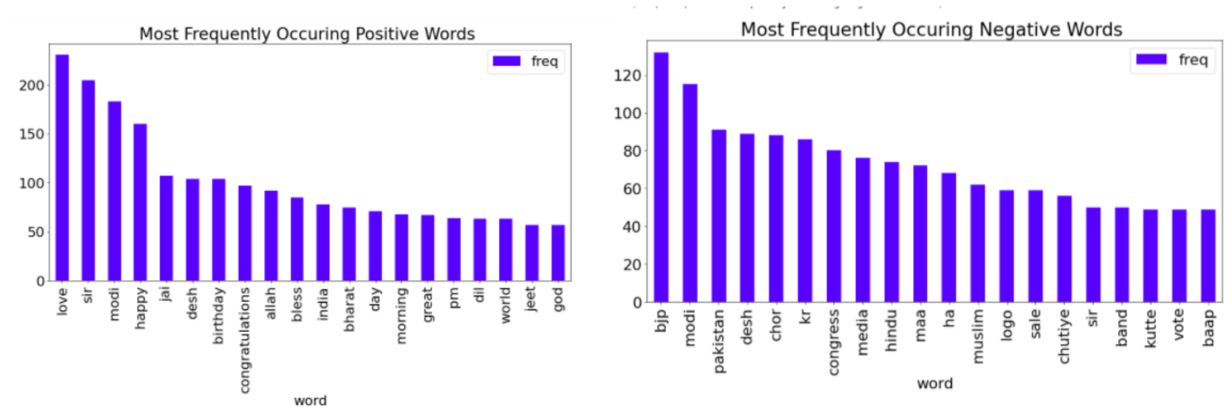

Figure 3.2: Most frequent words in the dataset

Figure 3.3: Word Cloud for the hindi words

Chapter 4

Models and Experimentation

4.1 Abstract

Sentiment Analysis has become one of the Natural Language Processing (NLP) tasks which has gained popularity over the past decade.Sentiment Analysis requires large sentiment labelled corpus for training. However, code-mixing demands new research methods where the focus goes beyond simply combining monolingual resources to address this linguistic phenomenon. Codemixing poses difficulties in a variety of language pairs and on multiple tasks along the NLP stack, such as word-level language identification, part-of-speech tagging, dependency parsing, machine translation, and semantic processing. In this chapter we will discuss different models used for training and results obtained from them [Patra et al., 2018].

4.2 Dataset PreProcessing

- Intially basic Cleaning of dataset has been done by removal of short forms like ”h” to ”hai” etc. Performed preliminary data processing operations like removal of unwanted links,special characters,etc and basic cleaning, expansion of short forms,stemming etc.
- Processed emoticons separately based on **Emoji Sentiment Ranking** as mentioned by [Kralj Novak et al., 2015]

- Processed hindi slang words separately using **Offensive hindi tweet dataset** and finally involved there rating as part of dataset [Mathur et al., 2018].

- Used the English/ Hindi tagged words to **segregate chunks of hindi words** from chunks of english words.

- use of **Google translator** for converting the hindi slangs and chunk of words with english data for better processing.

- **Sentiment Intensity Analyser** : it is a tool of nltk to calculate the intensity of sentiment in a sentence which has been used to calculate the intensity of negative , positive and nuetralty in the sentence [Venugopalan and Gupta, 2015].

- Dataset has been splitted into the ratio of 70:10:20 for traning , validation and testing respectively.

These all steps has been covered to pre-process the data before feeding it to the actual model. The snapshot of data is given as follows which has following features:

1. code mixed sentence
2. review translated into english language
3. whether slang exists in data or not?
4. Negative score of the sentence
5. Positive score of the sentence or review
6. neutral score of the sentence
7. offensive rating of the slang
8. Positive emoticon score
9. Neutral emoticon score
10. Negative emoticon score

	sentence_mixed	sentence_eng	sentiment	slang_existance	slang_rating	positive_emoji	negative_emoji	neutral_emoji	neg_score	pos_score	neu_score
1	haan yaar neha 😞😞 kab karega woh post 😭 usne na sach mein photoshoot karna chahiye phir woh post karega ... URL	yes friend 😞 😞 when will he post 😭 he didn t really photoshoot then he will post ... URL	0	False	0	1896	2412	1218	0.000	0.296	0.704

Figure 4.1: Dataset snapshot after Pre-processing

4.3 Models Trained

These are the classification models which we trained for the task of sentiment analysis:

- **Random Forest Classifier**: Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (see figure below).
  A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

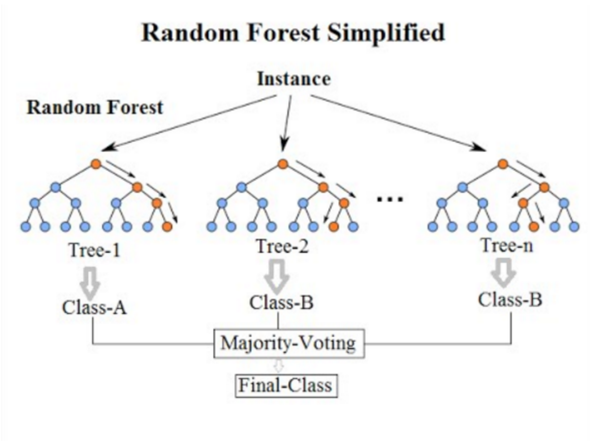

Figure 4.2: Random Forest Classifier[1]

- **Decision Tree Classifier**: The decision tree classifiers organized a series of test questions and conditions in a tree structure. The following figure [1] shows a example decision tree for predictin whether the person cheats. In the decision tree, the root and internal nodes contain attribute test conditions to separate recordes that have different characteristics. All the terminal node is assigned a class lable Yes or No.

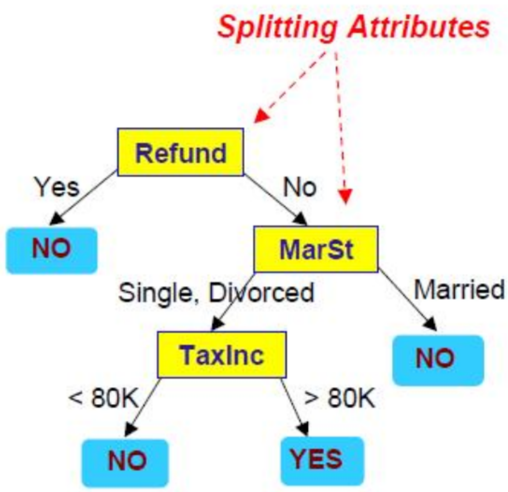

Figure 4.3: Decision Tree Classifier[2]

- **Logistic Regression Classifer**: A contradiction appears when we declare a classifier whose name contains the term 'Regression' is being used for classification, but this is why Logistic Regression is magical: using a linear regression equation to produce discrete binary outputs. And yes, it is also categorized in 'Discriminative Models' subgroup of ML methods like Support Vector Machines and Perceptron where all use linear equations as a building block and attempts to maximize the quality of output on a training set.

- **XGboost Classifer**: Rather than training all of the models in isolation of one another, boosting trains models in succession, with each new model being trained to correct the errors made by the previous ones. Models are added sequentially until no further improvements

can be made.
The advantage of this iterative approach is that the new models being added are focused on correcting the mistakes which were caused by other models. In a standard ensemble method where models are trained in isolation, all of the models might simply end up making the same mistakes.

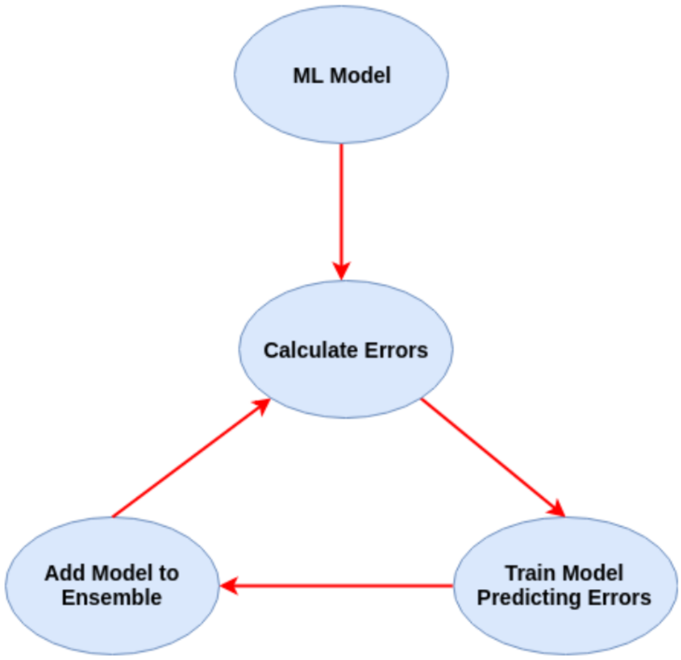

Figure 4.4: Xgboost Classifier[3]

- **Support Vector Machine**: "Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges.
  However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well
  .

[3] `https://www.google.images`

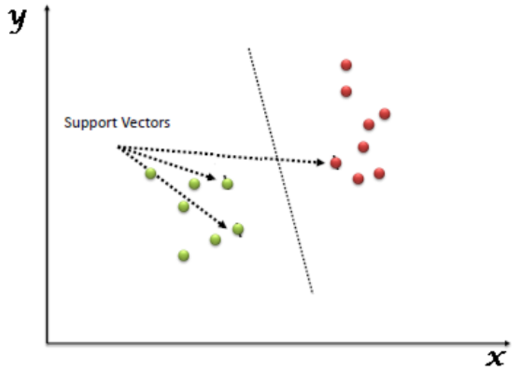

Figure 4.5: Support Vector Machine[4]

- **Bi-LSTM Model**: Using Bidirectional LSTMs, you feed the learning algorithm with the original data once from beginning to the end and once from end to beginning. There are debates here but it usually learns faster than one-directional approach although it depends on the task.

- **BERT Embeddings**: BERT, which stands for Bidirectional Encoder Representations from Transformers and its application to text classification. BERT is a text representation technique like Word Embeddings.
  Like word embeddings, BERT is also a text representation technique which is a fusion of variety of state-of-the-art deep learning algorithms, such as bidirectional encoder LSTM and Transformers. BERT was developed by researchers at Google in 2018 and has been proven to be state-of-the-art for a variety of natural language processing tasks such text classification, text summarization, text generation, etc. Just recently, Google announced that BERT is being used as a core part of their search algorithm to better understand queries.

4.4 Experimentation

These are the parameters and settings taken care-of while performing the experimentation of the different Models:

1. We have used keras and sklearn library for the processing.
2. Length of the embedding taken is 50.
3. We have experimented on multiple type of embeddings like google embedding, hindi sentinet embedding, BERT embedding, Glove twitter embeddings, bag of words, and td-idf embeddings.
4. For some model, we experimented by stacking the numerical and handcrafted features along with the sentences.
5. In case of BERT Model, emoticons are pre-processed by converting them to labels positive negative and neutral.

4.5 Results

- **Impact of the Bag of words features along with numerical features:** The table shown below gives us the accuracy for the dataset for positive sentiment, negative sentiment and neutral case when bag of words features has been used.

Model	Positive			Negative			Neutral		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
SVC	**.62**	**.55**	**.59**	.53	.43	.48	.46	.57	.52
Random Forest	.58	.57	.57	.49	.55	.52	.48	.54	.50
XGBoost	**.58**	**.58**	**.59**	.56	.53	.54	.49	.54	.52
Logistic Regression	.51	.57	.54	.46	.48	.47	.45	.40	.42
Decision Tree	.42	.44	.43	.41	.38	.39	.41	.42	.41
Bi-LSTM	0.54	0.50	0.52	**0.62**	**0.67**	**0.64**	0.49	0.48	0.48

Figure 4.6: Accuracy obtained when model is being trained using bag of words features

- **Impact of Glove textual embeddings & BERT Embeddings**: The table shown below gives us the accuracy for the dataset for positive sentiment, negative sentiment and neutral case when Glove textual embedding for twitter dataset has been used. Also it contains the case when BERT Embeddings were being used.

Model	Positive			Negative			Neutral		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
SVC	**.60**	**.57**	**.58**	.56	.53	.54	.50	.54	.52
Random Forest	.58	.61	.59	.53	.58	.55	.50	.52	.51
XGBoost	.59	.59	.59	.57	.53	.55	.51	.55	.53
Logistic Regression	**.6**	**.57**	**.58**	.56	.56	.56	.5	.52	.51
Decision Tree	.50	.50	.50	.47	.44	.45	.46	.48	.47
Bi-LSTM	0.54	0.50	0.52	**0.62**	**0.67**	**0.64**	0.49	0.48	0.48
BERT	**0.63**	**0.69**	**0.66**	**0.63**	**0.62**	**0.62**	**0.53**	**0.51**	**0.52**

Figure 4.7: Accuracy obtained when model is being trained using pre-trained Embeddings

- **Training and Validation Accuracy Graph**: This Graph shows the accuracies obtained by model for traning and validation dataset for the case of BERT Model.

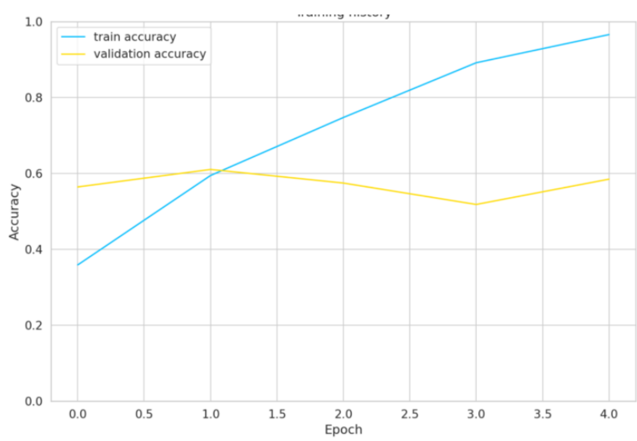

Figure 4.8: Accuracy Plot for the training and validation dataset

- **Confusion Matrix** : Since our model achieved the best accuracy in the case of BERT Model when trained on BERT embeddings which was 61.54%. Here is the attached snapshot for the confusion matrix.

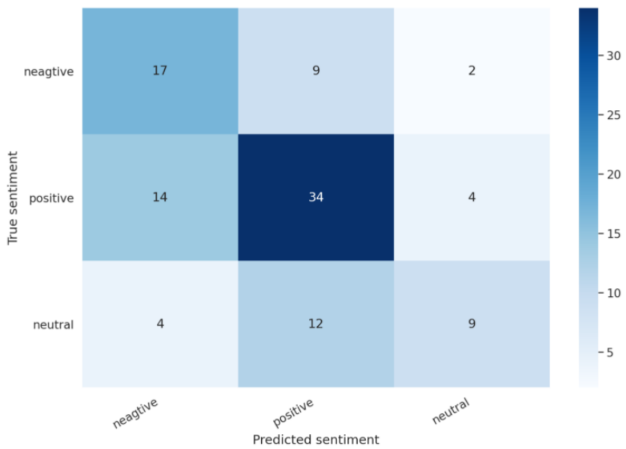

Figure 4.9: Confusion Matrix for the BERT Model

4.6 Observations and Error Analysis

1. It was observed that dataset itself was not correctly annotated. It has very much ambiguity and incorrect annotations.

2. As per the error analysis done for the BERT model, it was seen that most of the miss-classification occurred for the neutral sentiment.
   Example:rt moeshaaa a week ago we didn ' t win the championship but we came in 5th overall and got all american ohhhh and don ' t forget we made
   Classified as "neutral" in dataset but "positive" as per our BERT implementation

3. Ambiguity in positive-neutral, negative-neutral data-set features:
   Example1:"ye kya bakwas hai kaam dhanda or nahi hai kya practice kro yarrrr there is no space science"
   Classified as "neutral" in dataset but "negative" as per our BERT implementation

Example2: “me to saaare dialogues saath me bolti hu really love that movie ”
Classified as “neutral” in dataset but “positive” as per our BERT implementation

4. Sarcasm recognition:
   Example: ”sahi kaha puja ji aapne aise murkh har desh me badne chahiye jo deshdrohi ko sirf sikhaye hee nahi”
   Classified as “negative” in dataset but “neutral” as per our BERT implementation

5. Language invariance was causing issue since correct translations cannot be obtained. *There is no proper translation for the words like ”kehwa”, ”batawa”, ”janat” in google translate*

6. Positive use of negative smileys, and vice versa and slang words, highly negative/positive words also used in contrasting sense

7. Tried out senti-net for hindi- results not good because scores not given , Wiki-news pre-trained embeddings and glove embeddings in our work.Glove-twitter embeddings gave the best accuracy for Bi-lstm so we proceeded to use that.

8. Only 12.56% of total miss-classifications were positive-negative or negative-positive .

9. Use of hand-crafted features like negative score etc increased accuracy for 6%(approx.) in case of SVM.

10. Smiley and slang sentiment ratings helped a lot in improving accuracies of all the models

4.7 Summary

In this chapter, we discussed all the experimentation done along with the results and observations. BERT seems to be the promising model in case of code-switched languages as well. In next chapter we will discuss brief summary , conclusion and future scope for the work.

Chapter 5

Conclusion and Future Scope

5.1 Summary

Sentiment data present in reviews, tweets, feedback plays a very important role in knowing about different products, services and organizations. It has been an interested research area from quite a long time in NLP tasks. However, reviews are present in the multiple languages even mixed type. In this work we try to explore the existing models to figure out the sentiment of the data when review are given in mixed language, Hindi and English in our case.

5.2 Conclusion

Various Conclusion were Obtained from the implementation of model which can be summarized as follows:

- BERT seems to be the most prominent model for our dataset.
- Hand-crafted features generated by keeping emoticon rating slang rating helps in improvement of accuracy with multiple models.
- Majority of the miss-classification occurred for the case of neutral sentiment..

- Only 12.56% of total miss-classifications were positive-negative or negative-positive .

These are some observation obtained by running multiple experiments while performing sentiment analysis on code-switched dataset.

5.3 Future Work

- Review the data-set again and re-train the models based on revised data-set.
- Fast-text transliterated cross-lingual embeddings can be used
- We observed that even the data-set had incorrectly tagged hindi-english words , a separate project can be taken up to identify highly accurate word tagging. For eg. "are" can be both hindi and english words based on context.
- Other languages for code switching can be also be taken up.

References

[Kralj Novak et al., 2015] Kralj Novak, P., Smailović, J., Sluban, B., and Mozetič, I. (2015). Sentiment of emojis. *PloS one*, 10(12):e0144296.

[Liu et al., 2010] Liu, B., Morgan, C., and San, R. (2010). *Sentiment Analysis and Opinion Mining.*

[Mathur et al., 2018] Mathur, P., Sawhney, R., Ayyar, M., and Shah, R. (2018). Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148.

[Patra et al., 2018] Patra, B. G., Das, D., and Das, A. (2018). Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745.*

[Patwa et al., 2020] Patwa, P., Aguilar, G., Kar, S., Pandey, S., PYKL, S., Gambäck, B., Chakraborty, T., Solorio, T., and Das, A. (2020). Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain. Association for Computational Linguistics.

[Venugopalan and Gupta, 2015] Venugopalan, M. and Gupta, D. (2015). Sentiment classification for hindi tweets in a constrained environment augmented using tweet specific features. In *International conference on mining intelligence and knowledge exploration*, pages 664–670. Springer.