Abstract

Now a days, Emotions and opinions are the most influential & solely the driving factor of all social media content. It is observed that people can accentuate their sentiments in regional languages and thus the ways of writing with code-mixing and code-switching has unconsciously taken up in the social media as a trend. Users feel that their opinions are more emphasizing when written in code mix language. Hence the recent challenge is understanding the opinions in code-mixed content in social media. Such text is available in Romanized English format in Indian social media, which is the transliteration of Indian regional language in Romanized English, which demands text normalization to get further insights into the text. Thus this project aims to use and process code-mixed language (Hinglish: English + Hindi) input and perform Text Normalization for the same. Further, the project will judge the polarity of the statement as positive or negative using various sentiment resources.

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Overview of the Research

Text normalization is the process of transforming text into a single canonical form that it might not have had before.

Text normalization is a prerequisite for a variety of speech and language processing tasks and hence there is a demand for such systems. Normalizing text before storing or processing it allows for separation of concerns, since input is guaranteed to be consistent before operations are performed on it. T66666ext normalization requires being aware of what type of text is to be normalized and how it is to be processed afterwards; there is no all-purpose normalization procedure.

The rapid expansion of Internet use, electronic communication and user-oriented media such as social networking sites, blogs and micro blogging services has led to a rapid increase in the need to understand casual written English, which often does not conform to rules of spelling, grammar and punctuation. Text Normalization is important to get further insights into such text and for sentiment analysis.

1.2 Background of the Research

Text Normalization is a discipline of Natural Language processing and a necessity for various speech and language processing tasks. In this bilingual speech community, there is a natural tendency of speakers to mix phrases and sentences during conversation, which has led to substantial code switching in Hindi and English language.

We came across systems that use a statistical language independent approach for automatic detection of foreign words in mixed language. For Hindi-English a bilingual syntactic parser has been implemented. Conditional random field method has been used to identify the language of the words in mixed language documents.

As the users on the social media are commonly using Abbreviation (short form) or SMS language

to communicate, just dealing with language identification and transliteration cannot help us in understanding the text in social media sites.

1.3 Motivation of the Research

Social Media is the frontier for opinions. Its anonymity provides the perfect ground for public voice. It also witnesses an active participation of people across the world and we get diverse outlooks over the same topic making it the best source for text mining.

However, extensive usage of net lingo, i.e. use of various acronyms for common phrases and slangs and different forms of short hand words in place of normal words is a limitation. Also, users employ bilingual (or more than 2) languages users for convenience and ease of communication making it difficult to analyze such text.

Hence this system will be designed to understand casual written English and code-mixed text, which often does not conform to rules of spelling, grammar and punctuation.

1.4 Significance of the Research

With the rapid development of social media, the irregularity of language poses a barrier to automated task. Posts are often highly ungrammatical, and filled with spelling errors, and resorted to selecting clusters of spelling variations manually. The interest in content of this type, both from researchers and corporations, shows a pressing need for effective text normalization. Natural Language Processing tasks such as Machine Translation, Information Retrieval and Opinion Mining, require Text Normalization due to the irregularity of the language featured.

1.5 Objective of the Research

Our objective is to use concept of text normalization for code-mixed (Hindi-English) and impure social media text and perform sentiment analysis. Social Media is the most budding platform with a great global outreach and an important source for text mining for social media analytics. Text Normalization can be used to make such text consistent. Hence, this project presents a method for Text Normalization and how it can be used to treat and process social media content which in the above form as mentioned.

1.6 Scope of the Research

The scope of the system would be to normalize code-mixed Hindi-English content in social media which is available in Romanized English format. This project will also deal with impure social English consisting of Abbreviations (Short forms), Word play/ Intentional misspelling for verbal effect and Slang words(acronyms). The system converts impure social English to pure English, followed by tagging English and Hindi words. Hindi words are transliterated to Devanagari script for sentiment analysis based on lexicon approach.

1.7 Beneficiaries

The recipients of the system would be organizations which use social media monitoring such as public opinion, reviews and rating of the product which provide valuable information about emerging trends and what consumers and clients think about specific topics, brands or products.

Chapter 2

Literature Survey

This chapter consists of the literature survey that we have conducted on the various systems employing normalization of code-switched text and sentiment analysis. The following chapter explains the systems that we surveyed.

Shashank Sharma, PYKL Srinivas and Rakesh Chandra Balabantaray's [1] proposed paper proves to be the base paper for further research as it presents various methods to normalize code - mix text available on social media, which is the transliteration of one language into another. In this paper, firstly, the language of code-mix text, which includes Phonetic Typing, Abbreviation (Short forms), Word play, intentionally misspelt words and Slang words is identified. This is followed by transliteration of these Romanized English language words which have a variety of spellings. The words are then judged for the sentiment of the statement to be positive or negative based on lexicon approach. An accuracy of 85% was achieved with the model.

DinkarSitaram, Savitha Murthy's [2] discusses an approach for sentiment analysis of the mixed language arising through the fusion of two languages: Hindi and English. In the mixed language, the resulting grammar usually alternates between the source language and deviates significantly from the grammatical structure of its source languages. This methodology incorporates the determination of the grammatical transition and the application of sentiment combination rules across languages to evolve the overall sentiment of a sentence in the mixed language. The model uses a recursive neural tensor network (RNTN) for sentiment classification. The model is trained with a training set consisting of 345 text samples in mixed language.Accuracy of model is low due to small training set data. Many important words that determine the overall sentiment of a sentence are not known to the classifier and hence assigned a neutral sentiment.

Ben King and Steven Abney's [3] uses a weakly-supervised learning method for identifying the languages of individual words in mixed language documents. A conditional random field model trained with generalized expectation criteria, a hidden Markov model (HMM) trained with

expectation maximization (EM), and a logistic regression model trained with generalized expectation criteria is implemented. The paper concludes that CRF trained with GE is clearly the most accurate option among the methods examined. It also outperforms sentence-level language identification, which is too coarse to capture most of the shifts between languages. Also, named entities are not handled properly.

Heeryon Cho, Jong-Seok Lee and Songkuk Kim's [4] presents a method of improving lexicon-based review classification by merging multiple sentiment dictionaries, and selectively removing and switching the contents of merged dictionaries. First, the system compares the positive/negative book review classification performance of eight individual sentiment dictionaries. Then, selects the seven dictionaries with greater than 50% accuracy and combine their results using (1) averaging, (2) weighted-averaging, and (3) majority voting. It is shown that the combined dictionaries perform only slightly better than the best single dictionary (65.8%) achieving (1) 67.8%, (2) 67.7%, and (3) 68.3% respectively. To improve this, the approach combines seven dictionaries at a deeper level by merging the dictionary entry words and averaging the sentiment scores. Moreover, it leverages the skewed distribution of positive/negative threshold setting data to update the merged dictionary by selectively removing the dictionary entries that do not contribute to classification while switching the polarity of selected sentiment scores that hurts the classification performance. The revised dictionary achieves 80.9% accuracy and outperforms both the individual dictionaries and the shallow dictionary combinations in the book review classification task.

Subhash Chandra, BibekanandaKundu and Sanjay Kumar Choudhury's [5] analyzes the reasons of language mixing and its characteristics. The paper focuses on the mixed language called Benglish and Hinglish which are actually fusion of English with Bangla and Hindi language. A hybrid approach combining rule based and statistical methods has been proposed here. After manual introspection of the sentences of CMIC (Computer Mediated Informal Communication), the system extracts some linguistic patterns for detection of English words in Benglish text. The statistical model has two components viz. (1) Grapheme Language Model (GLM) and (2) Phoneme Language Model (PLM). When tested on 9152 Benglish sentences containing 13795 unique mixed words collected from CMIC, the proposed approach yielded an accuracy of 95.96% comparatively higher than 83.67% and 54.70% achieved by rule based and statistical approach

respectively.

G.Vinodhini and R M Chandrashekaran's [6] covers a survey of various techniques and methods that are employed in performing sentiment analysis. It also covers the challenges that are faced while doing so. The paper covers the Machine Learning algorithms that are based on supervised learning. The other problem that it focuses on is semantic orientation which focuses on unsupervised learning. The role of negation and the various models that have been developed to handle the problem of negation are also discussed. The paper also discusses the feature based learning and the various algorithms. The paper also maps the function of various models based on certain metrics such as precision, recall and F-measure. The paper focuses on various methods that are already in existence and measures their working based on a few metrics.

Eleanor Clarka and Kenji Arakia's [7] discusses problems involved in automatically normalizing social media English. The paper describes an experiment which examines the efficacy of conventional spell checkers on casual English, and to what extent this could be improved with pre-processing with the given system. Results showed that average errors per sentence decreased substantially, from roughly 15% to less than 5% with use of spell checker.
Different spell checker has strengths and weaknesses with different types of errors.

Thomas Gottron and NedimLipka's [8] compares the performance of some typical approaches for language detection on very short, query-style texts. The results show an accuracy of more than 80% and for longer texts, accuracy values close to 100% were achieved. The approach focuses on methods based on character n-grams, for short n-grams. Next method studied was Naive Bayes which is a classical approach. Markov processes are used to determine the language of the text. The last approach is based in vector space of all possible n-grams.

R. Mahesh K. Sinha and Anil Thakur's [9] presents a mechanism for machine translation of Hinglish to pure (standard) Hindi and pure English forms. The approach uses a Hindi and English morphological analyzer.The morphological analyzer yields part of speech information for each of the words and marks the words that are unknown in Hindi and English respectively. The words that remain unknown, are marked for cross morphological analysis for plural noun forms. Complex code-mixed (CCM) Hinglish sentence is segmented into simple code-mixed Hindi

(SCMH) and simple code-mixed English (SCME) parts. Next, the method isolates SCMH and SCME from CCM using heuristics and shallow parsing. Then it translates SCMH/SCME into pure Hindi language/pure English using FSM. The system fails in case of polysemous verbs, due to a very shallow grammatical analysis used in the process, the system is unable to resolve their meaning.

2.3 Literature Summary

Table 2.1: Summary of Literature Survey

Sr. no	Languages	Year	Research paper Details	Remarks
1	Hindi-English	2015	**Text Normalization of Code Mix and Sentiment Analysis** • Tokenization and Language Identification • Transliteration of Romanized Hindi to Devanagari script • Study & development of corpus like Opinion Lexicon, AFINN for English &HindiSentiWordnet for Hindi • Standardization for judging sentiment	• Overall accuracy of more than 85%. • Corpora was collected from FIRE 2013 and FIRE 2014. Accuracy in transliteration. • Model achieved Precision of 0.90.
2		2015	**Sentiment Analysis of mixed language employing hindi-english code switching** • Recursive neural network for sentiment classification • Training set of 345 text samples (includes shorthand)	• Proposed technique focuses on sentiment analysis at both phrase and sub-phrase level • Accuracy low due to small training set data • Many important words that determine the overall sentiment of a sentence are not known to the classifier and hence assigned a neutral sentiment

3		2005	Machine Translation of Bi-lingual Hindi-English (Hinglish) Text Steps: • Hindi and English morphological analyzer. Complex code-mixed (CCM) Hinglish sentence is segmented into simple codemixed Hindi (SCMH) and simple code-mixed English (SCME) parts. • Isolating SCMH and SCME from CCM: heuristics and shallow parsing (steps mentioned) • Translate SCMH/SCME into pure Hindi language/pure english - FSM used.	In case of polysemous verbs, due to a very shallow grammatical analysis used in the process, the system is unable to resolve their meaning.
4		2013	**Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods** • Word language classification using: logistic regression, naive Bayes (best performer), decision tree, and Winnow2 weakly- and semi supervised methods • Implements a conditional random field model trained with generalized expectation criteria, a hidden Markov model (HMM) trained with expectation maximization (EM), and a logistic regression	Named entity errors, when a named entity is given a label that does not match the label it was given in the original annotation, Shared word errors, when a word that could belong to either language is classified incorrectly.

			model trained with generalized expectation criteria	
5	Hindi-English; Hindi-Bengali	2013	**Enhancing lexicon-based review classification by merging and revising sentiment dictionaries** Presents a method of merging multiple dictionaries, and removing and switching the merged dictionary's contents to achieve greater accuracy in the lexicon-based book review classification	The revised dictionary achieves 80.9% accuracy and outperforms both the individual dictionaries and the shallow dictionary combinations in the book review classification task
6		2013	**Hunting Elusive English in Hinglish and Benglish text: Unfolding challenges and Remedies** • A hybrid approach combining rule based and statistical methods has been proposed here. • After manual introspection of the sentences of CMIC (Computer Mediated Informal Communication), the system extracts some linguistic patterns for detection of English words in Benglish text	• The proposed methodology has been evaluated on a corpus of 9152 sentences with 13795 unique words • The proposed approach yielded an accuracy of 95.96% comparatively higher than 83.67% and 54.70% achieved by rule based and statistical approach respectively
7		2012	**Sentiment Analysis and Opinion Mining** • Machine Learning algorithms based on supervised learning • Semantic orientation which focuses on unsupervised learning • Models to handle the problem of negation	The paper maps the function of various models based on certain metrics such as precision, recall and F-measure.

8		2011	**Text normalization in social media: progress, problems and applications for a pre-processing system of casual English** Designed a text classification system with manually compiled and verified database and phrase matching rules	• Results showed that average errors per sentence decreased substantially, from roughly 15% to less than 5% with use of spellchecker. • Different spellchecker has strengths and weaknesses with different types of errors.
9		2010	**A Comparison of Language Identification Approaches on short query style texts** Method based on n-gram, naive bayes, markov processes	The results show that for single words an accuracy of more than 80% can be achieved, for slightly longer texts accuracy values close to 100% was observed.

2.2 INFERENCE OF LITERATURE SURVEY

- The result of 4 papers on Regional Language and 5 others about various algorithms and techniques, it has been concluded that Text Normalization would be the best approach for Hindi-English Code Mix content.
- Recursive neural network for sentiment classification analysis at both phrase and sub-phrase level Accuracy low due to small training set data. Many important words that determine the overall sentiment of a sentence are not known to the classifier and hence assigned a neutral sentiment.
- Due to a very shallow grammatical analysis used in Machine Translation of Bi-lingual Text, the system is unable to resolve their meaning.
- Language classification results in problems like Named Entity errors and Shared Word errors, thereby lowering overall accuracy of result expected.
- Merging and revising sentiment dictionaries will not help when the corpus consists of an Impure Language base and it would only enhance sentiment analysis of English content.
- Other papers show research about employing methods to Extract English words from code mixed and code switched text, using algorithms such as n-gram markov process etc for

language identification by comparison, machine based algorithms for enhancement of precision of Sentiment analysis.

- So, hereby, we conclude that using Text Normalization harbors the best method for Social Media content, as it presents various methods to normalize code - mix text available on social media.

Chapter 3

Text Normalizer for Code-mixed (Hindi + English) Text

In this chapter we would be discussing about the system architecture. The input to the system would be would be code-mixed (Hinglish), impure language text available from different social media domains like Facebook, Twitter and YouTube. The first unit is tokenization. This will be followed by conversion of impure SMS language to English. Language Identification will then be done to tag English and Hindi words. The next part is transliteration of Hindi to Devanagari script. Finally using a lexicon based approach sentiment analysis of the text will be done.

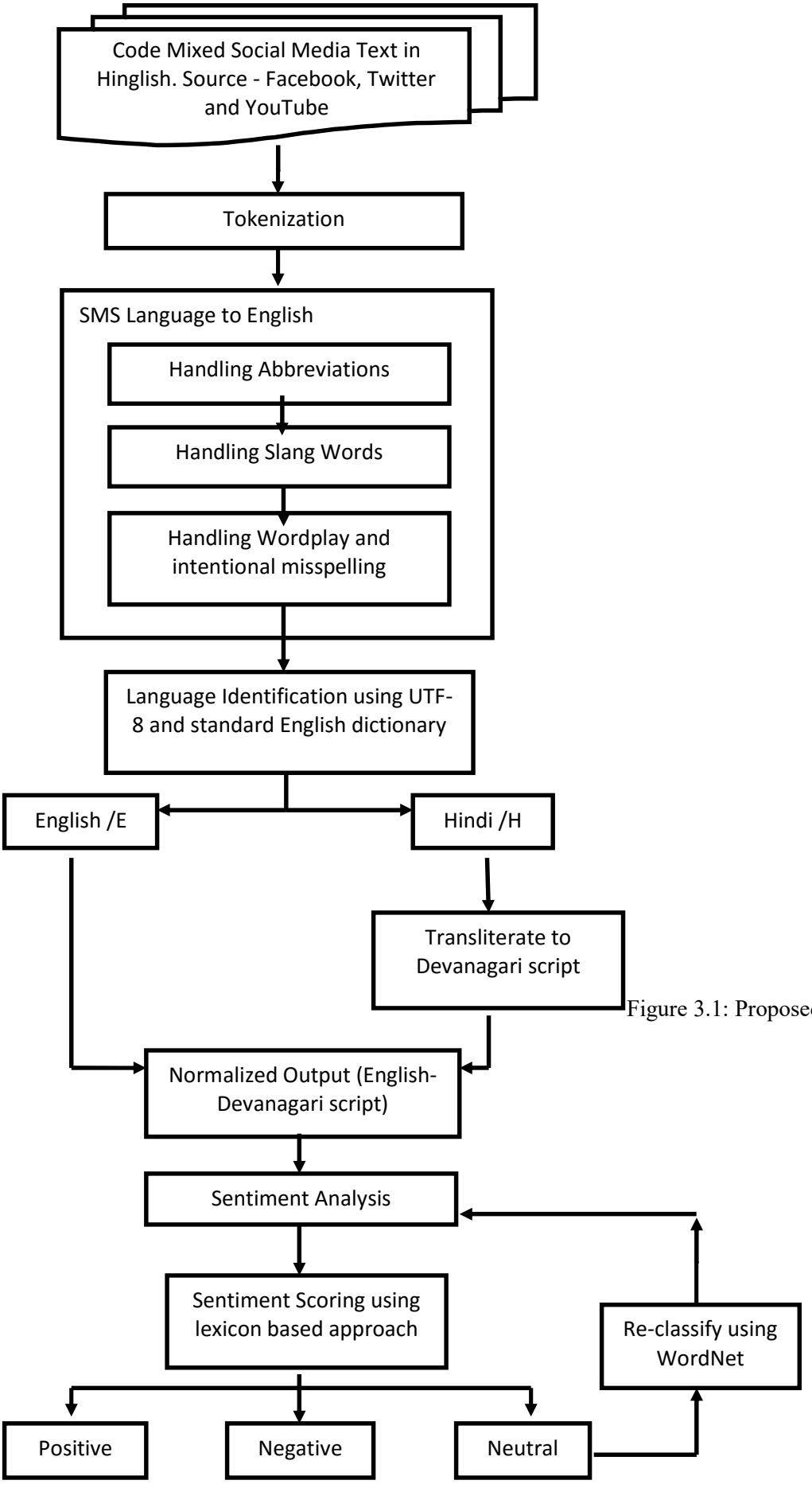

Figure 3.1: Proposed architecture

3.1 Input Documents

The input to the system would be code-mixed (Hinglish), impure language text. This text will be in Romanized English format which is transliteration of Hindi in Romanized English. The text would be from different social media domains like Facebook, Twitter and YouTube.

3.2Tokenization

Process of converting sentence into a chain of words so that processing word by word can be easily performed.Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation. We use white space character for tokenization.

Algorithm for Tokenization:

Input: Code-mixed text

Output: List of token

Steps:

1. START
2. Initialize all pointers (input (for character), output (for token), initially assign tokens = NULL)
3. Scan the input document.
4. Check for not End of file.
    a. Read a character from input file.
    b. if character is a special character do the following
        i) Treat that special character as a token.
        ii) Add the token in token list file.
        iii) Increment the respective pointers.
    c. If character is not special character then do following while space character is not found
        i) Treat all character as a token.
        ii) Add token into token list file.
        iii) Increment respective pointers.
5. Repeat step 4 until end of file.
6. STOP.

3.3 SMS Language to English

Input text contains different net lingo which is use of various acronyms for common phrases and slangs and different forms of short hand words in place of normal words. Before further processing this SMS language is transformed to pure English.

3.3.1 Handling Abbreviations

Now a days, on social media we find a lot of abbreviations(short forms) being used in English. We have mapped differentkinds of shorthand notation spellings into one, for example:atb ("all the best") or lol ("laugh out loud"). A dictionary is used to handle and expand abbreviations.

Aamof	as a matter of fact	gm	good morning
Aimb	as I mentioned before	gtg	got to go
Asap	as soon as possible	hand	have a nice day
Atb	all the best	hhh	hip hip hooray
Atm	at the moment	lol	laugh out loud
Atst	at the same time	ruok	are you ok
atwd	agree that we disagree	tc	take care
awttw	a word to the wise	ttyl	talk to you later
bb	bye bye	tysm	thank you so much
gg	good name	wru	where are you

Table 3.1: Sample of abbreviations dictionary

Algorithm for Handling Abbreviations:

Input: Tokens

Output: Full-form of abbreviations

A dictionary based approach is been utilized to handle abbreviations in the document. A dictionary of 2300 words has been used for the comparison purposes. The algorithm is implemented as below given steps.

Steps:

1. A single token is read from token array.

2. Dictionary of abbreviations ("**abbreviations.xls**") is opened.
3. Initialize a pointer variable to start of dictionary file.
4. Compare token and word from dictionary.
5. If match is found, replace abbreviated token with its full form.
6. If match is not found, increment pointer variable and go to step 4.
7. If end of file is reached, close dictionary file and stop.
8. Repeat the above process for all the tokens.

3.3.2 Handling Slang Words

The next most commonly used phenomenon on social media is the usage of slang words or acronyms, for example: 4get ("forget"). A slang dictionary with a number of words is used to train the system for correct language identification and transform slang words to pure English.

2day	Today	gr8	great
4u	for you	gud	good
awsm	awesome	hv	have
B	Be	hw	How
betn	between	k	okay
bt	But	lyk	Like
D	The	r	Are
der	there	v	We
eg	example	wat	what
fav	favourite	wer	where

Table 3.2: Samples of Slang dictionary

Algorithm for Handling Slang words:

Input: Tokens where abbreviations has been handled.

Output: Slang words converted to pure English.

A dictionary based approach is been utilized to handle slang words in the document. A dictionary of 250 words has been used for the comparison purposes. The algorithm is implemented as below given steps.

Steps:

1. A single token is read from token array.
2. Dictionary of slang words ("**slang words.xls**") is opened.
3. Initialize a pointer variable to start of dictionary file.
4. Compare token and word from dictionary.
5. If match is found, replace slang word with its pure form.
6. If match is not found, increment pointer variable and go to step 4.
7. If end of file is reached, close dictionary file and stop.
8. Repeat the above process for all the tokens.

3.3.3 Handling Wordplay

A lot of users often use creative spellings, which includes phoneticspelling and intentional misspelling for verbal effect e.g. thatwas soooooo big ("that was so big"). In this case, a simple algorithm may be used to identify the flaw and corrected it with the right English word.

Algorithm for Handling wordplay:

We use regular expressions (regex) to identify and remove multiple consecutive occurrences of characters which are typed intentionally.

1. re.sub(r'([a-zA-Z])\1{2,}', r'\1', inputs)
2. re.sub(r'([a-zA-Z])\1{2,}', r'\1\1', inputs)

1) Input :soooooo

Output :

so

soo

Search in English dictionary, 'soo' is not found, hence it is discarded. Hence 'soooooo' is translated to 'so'.

2) Input :aweeeeeesomeeeee

Output :

awesome //match found in English dictionary

aweesomee

3) Input :yummmmmy
Output :
yumy
yummy //match found in English dictionary

4) Input :greaaaaaattttt
Output :
great //match found in English dictionary
greaatt

3.4LANGUAGE IDENTIFICATION

Language Identification is an important unit and used to tag language of text as English(/E) or Hindi(/H). This is important for further processing of Romanized Hindi and eventually sentiment analysis. The process may start with identifying English followed by Hindi.

3.4.1 English Language Identification

The tokens are fed into an English Language Identifier. The tokens belongingto English language are tagged as /E at the end of each token. To identify the English tokens we use UTF-8 and a standard and vastEnglish dictionary.

Unicode (or Universal Coded Character Set) Transformation Format means it uses 8-bitblocks to represent a character. UTF-8 is a variable-length byte encoding of Unicode, thecharacter numbering system for all languages defined by Unicode. UTF-8 is the dominantcharacter encoding for the World Wide Web and can support many languages and canaccommodate pages and forms in any mixture of those languages. Its use also eliminates theneed for server-side logic to individually determine the character encoding for each page servedor each incoming form submission. This significantly reduces the complexity of dealing with amultilingual site or application and also allows many more languages to be mixed on a singlepage than any other choice of encoding.

Unicode has a totalof 1,114,112 code points in the range 0(hex) to 10FFFF(hex).UTF-8 has the capacity ofencoding all the code points defined in Unicode. Code points with lower numerical values (i.e.,earlier code positions in the Unicode character set, which tend to occur more frequently) areencoded using fewer bytes. The encoding in UTF-8 has certain variations that are thoroughly followed. The first 128 characters of Unicode are encoded using a single octet with the samebinary value as ASCII, making valid ASCII text valid UTF-8- encoded Unicode as well. AndASCII bytes do not occur when encoding non-ASCII code points into UTF-8, making UTF-8 safeto use within most programming and document languages that interpret certain ASCIIcharacters in a special way.

A	do	heritage	not	spread
Adopt	excited	hurt	our	tell
All	feel	immortal	past	thank
anybody	field	in	pride	the
awesome	finally	like	richest	to
best	finals	match	rivalry	today
between	follow	match	science	was
cricket	from	miss	see	were
culture	good	much	sentiment	you
day	have	nice	so	yummy

Table 3.3: Samples of English dictionary

Algorithm for Language Identification:

Input: Tokens in code-mix (Hinglish).

Output: English words are tagged as /E.

To identify the English tokens we use UTF-8 and a standard and vastEnglish dictionary.

Steps:

1. Use the character set as UTF-8
2. Scan the token character by character.
3. Compare each character from scanned input with UTF-8.
4. If character is present in the UTF-8, then it is valid to English script otherwise not.

5. Ignore all the special characters.
6. Identifying English words.
    6.1.Open Standard English dictionary.
    6.2. Initialize a pointer variable to start of dictionary file.
    6.3. Compare token and word from dictionary.
    6.4. If match is found, tag token as /E.
    6.5. If match is not found, increment pointer variable and go to step 6.3.
    6.6. If end of file is reached, close dictionary file and stop.
7. Repeat the above process for all tokens.

3.4.1 Hindi Language Identification and Transliteration to Devanagari

The tokens which are not labeled as /E are considered to be in Hindi language which is written in Roman script and is tagged as /H. Romanized Hindi words are transliterated to Devanagari script.

Machine transliteration refers to the process of automatic conversion of a word from one language to another without losing its phonological characteristics. Machine transliteration of English-Hindi is done using rule based approach. Some rules are constructed for syllabification. Syllabification is the process to extract or separate the syllable from the words.

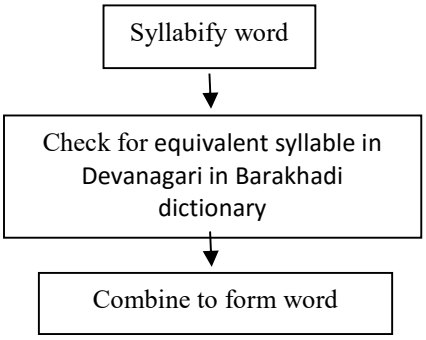

Figure 3.2: Transliteration Process

For constructing our rules, we are using syllabification approach. The syllable is acombination of vowel and consonant pairs such as V, VC, CV, VCC, CVC, CCVC and CVCC.Almost all the languages have VC, CV or CVC structures so we are using vowel and consonantsas the basic phonification unit.

Some possible rules for syllabification are shown in table:

Syllable Structure	Example	Syllabified form (English)
V	Eka	[e][ka]
CV	Tarun	[ta][run]
VC	angela	[an][ge][la]
CVC	sidhima	[si][dhi][ma]
CCVC	odisha	[o][di][sha]
VCC	obhika	[o][bhi][ka]

V: Vowels, C: Consonants

Table 3.4: Possible combinations of syllables

If we have a word P then it will be syllabified in the form of {p1,p2,p3….pn} where p1,p2..pn are the individual syllables. Syllabification of English input is done using rule based approach for which the following algorithm is used.

Algorithm for syllable extraction

1. Enter input string in English.
2. Identify Vowels and Consonants.
3. Identify Vowel-Consonant combination and consider it as one syllable.
4. Identify Consonants followed by Vowels and consider them as separate syllables.
5. Identify Vowels followed by two continuous Consonants as separate syllable.
6. Consider Vowel surrounded by two Consonants as separate syllable.
7. Transliterate each syllable into Hindi.

a	अ	Ee	ई	ki	कि
aa	आ	ei	ऐ	me	मे
cha	च	Ga	ग	o	ओ
chha	छ	Ha	ह	oo	ऊ

e	ए	i	इ	ra	र
ea	ए	Ja	ज	sa	स

Table 3.5: Samples of English-To-Hindi-Conversion list.

3.5 Example Sentences

Example 1:

2day's match was soooooawsm. Hameshakitarah sab were excited to see match betn gr8 rivalry in cricket. SachinaurSehwagkejodine ne to kamalkardiya. Finally, v r in finals. ATB Indian Team ko. Feeling very proud 2 b Indian.

Example 2:

GM all Ab India ke cultural heritage kebaare mekyabatavu. India has d richest culture in d world. Pooreduniyamai spread huahai ye. Aamof, MohiniAttam ne India ko world level pe represent kiyahai. A.R. Rehmanekaurzindaudaharanhai. Iirc he has won oscars 4 India in music. Not just art lekin science ke field me bhibohotinsaanhai who hv brought pride to our nation. Hamare culture kopooreduniya ne adopt karnekikoshishkihai. Tysm HAND.

Example 3:

Aajmaa ne terifavkheerbanayithind it was so yummmmy. Humne aapkobahut miss kiya. Cum soon Didi

Chapter 4

Result and Discussion

4.1 Input Datasets and GUI

For the evaluation of the working of our system, we generated many test sets. One test setwas generalized which catered to all the different inputs that we are considering, i.e Hinglish words , Slang words, Abbreviated words, Wordplay and intentionally misspelled words as well. A total of 70 such sentences were considered in that particular test set. In order to check the individual working of all the different modules of the system, we also considered 10 sentences for each module. Thus we used four more test sets for calculation of individual module's efficiency. The in depth analysis of the outputs and their performance evaluation has been described in detail in the following sections.One such example is as shown below:

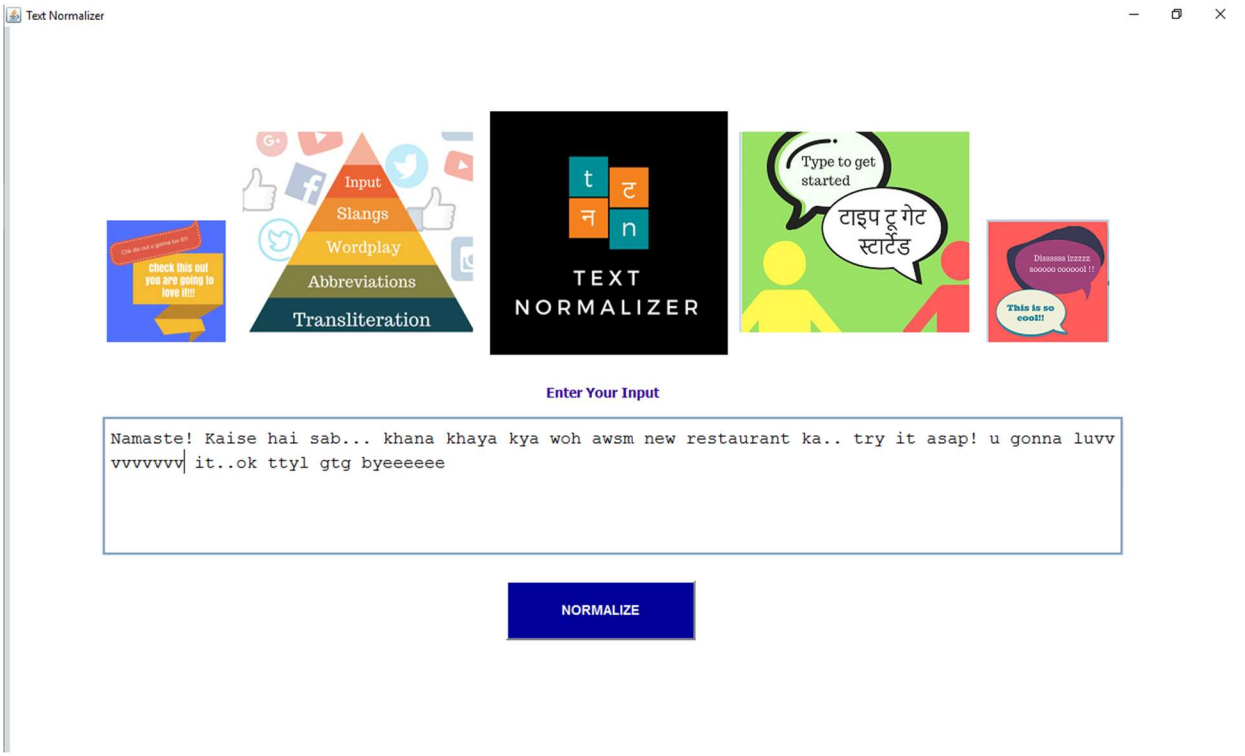

Figure 4.1.1 Homepage of the Text Normalizer system

The above image is the first page of the GUI. The user can enter the input sentence in the text area.

Figure 4.1.2 Output of the Text Normalizer system.

This is the second screen of the system. It has various options, where in the user can select the stage whose output the user wants to view.

Figure 4.1.3 Handling Abbreviations stage of the system.

The figure 4.3 is what the user would see as the output when the user selects the Handling

Abbreviations as the option. The output of the input sentence at this stage may also be seen in the right output screen.

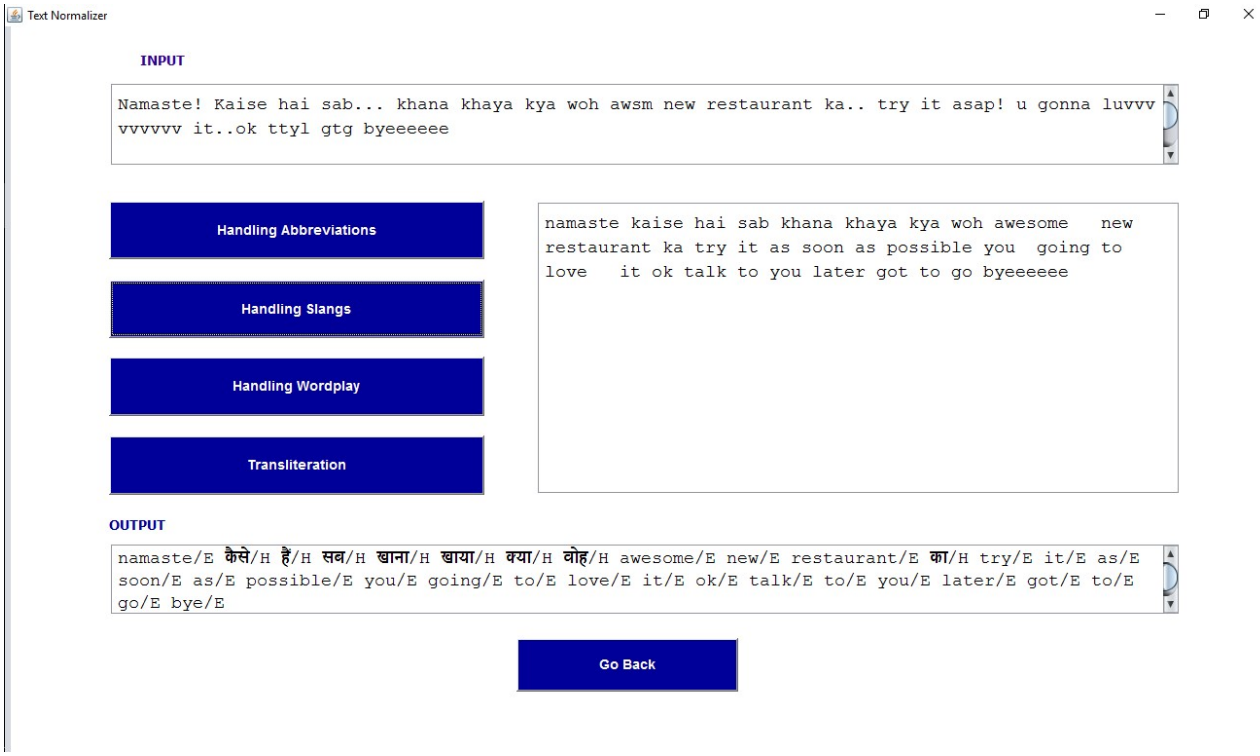

Figure 4.1.4 Handling Slangs stage of the system.

The above page is what the user would see as the output when the user selects the Handling Slangs as the option. The output of the input sentence at this stage may also be seen in the right output screen.

The figure 4.5 represents another stage of the block diagram which is represented in the system. This stage handles the Wordplay and intentionally misspelled words. The figure shows what the system would give as output for the given input from the user.

Figure 4.1.5 Handling Wordplay stage of the system.

Figure 4.1.6 Handling the Transliteration stage of the system.

The figure 4.6 represents the final stage of the system. Here the transliterated words are handled. The input words considered here are those that are in Hinglish(Hindi + English)

4.2 Performance Evaluation and Result Analysis

The quality of our system can be obtained by calculating the efficiency of the system. There is no particular method to calculate such an efficiency. Thus we have calculated the percentage of correct outputs for all the different datasets considered for testing.

$$\textbf{Efficiency} = \frac{\textbf{Total No of Correct words in the output}}{\textbf{Total No of words in input}} * \textbf{100}$$

4.2.1 Analysis of Handling Abbreviations

Sr no.	Input Sentence	Output	Incorrect Words	Total Words
1	Dmi I am glad to help you df	don't/E mention/E it/E i/E am/E glad/E to/E help/E you/E dear/E friend/E	0	11
2	Dwbits an automatic reply	don't/E write/E back/E it/E is/E an/E automatic/E reply/E	0	8
3	After watching friends icl its roflnd lol funny man!	after/E watching/E friends/E **ड्रnull/H** it/E is/E rolling/E on/E the/E floor/E laughing/E and/E laugh/E out/E loud/E funny/E man/E	1	17
4	Abd work needs to be add by the owner of the works, kindly comply	already/E been/E done/E work/E needs/E to/E be/E add/E **bye/E** the/E owner/E of/E the/E works/E kindly/E comply/E	1	16
5	Bb , hand and call me when you reach at your destination	bye/E bye/E have/E a/E nice/E day/E and/E call/E me/E when/E you/E reach/E at/E your/E destination/E	0	15
6	Cpg is highly overrated, one should use organic items instead	consumer/E packaged/E goods/E is/E highly/E overrated/E one/E should/E use/E organic/E items/E instead/E	0	12

7	Cot is the secret for every successful relationship in this world	circle/E of/E trust/E is/E the/E secret/E for/E every/E successful/E relationship/E in/E this/E world/E	0	13
8	Dmi , it was my pleasure	don't/E mention/E it/E it/E was/E my/E pleasure/E	0	7
9	Fb is my fav social media / social networking site man!	**फnullnullबूक** is/E my/E favorite/E social/E media/E social/E networking/E site/E man/E	1	10
10	Gmab , is what ross was intending to tell Rachel, but it came out all wrong.	give/E me/E a/E break/E is/E what/E ross/E was/E intending/E to/E tell/E **राचेल /H** but/E it/E came/E out/E all/E wrong/E	1	19

Table 4.1 Test case for Abbreviations

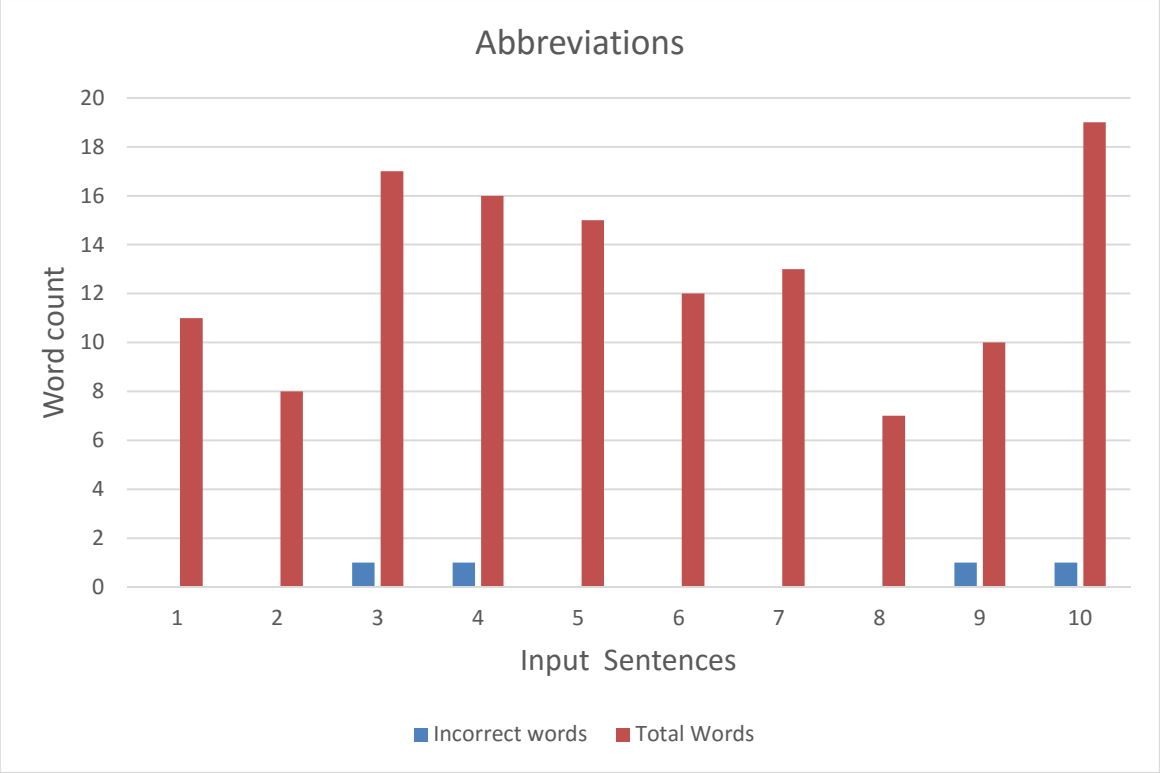

Figure 4.2.1 Simulation results for abbreviated words as input

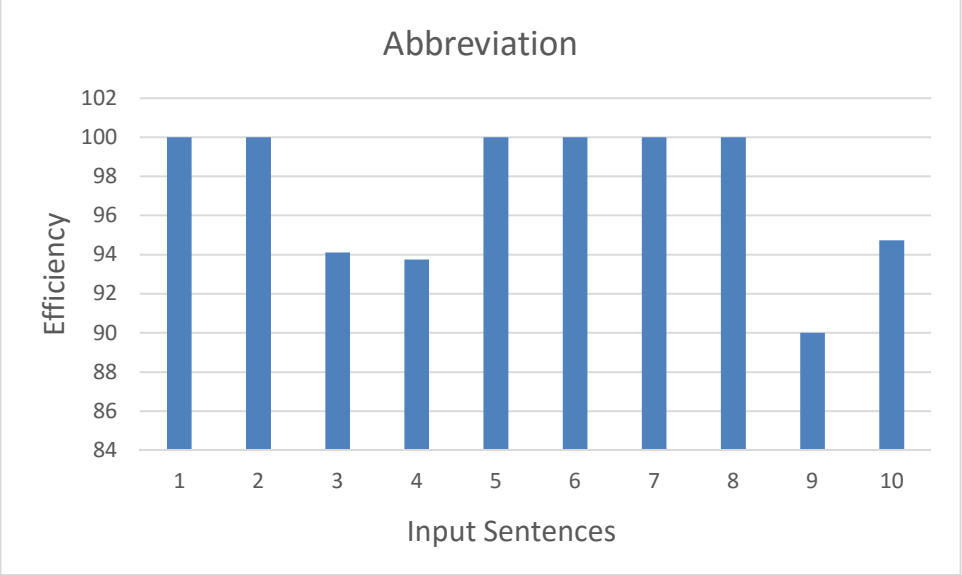

Figure 4.2.2 Efficiency evaluation of Abbreviated words

In order to evaluate the proposed system, we conducted tests on the different test cases that we developed. Firstly we considered a test case with 10 sentences as inputs to check how the system worked with only abbreviated words as inputs. The figure 4.2.1 graph shows the output that the system generates; the correct and incorrect word count for every input sentence can be viewed in the graph. The accuracy of the system for evaluation of the abbreviated words came around 96.8%. The correct words predicted were 128 and the incorrect ones were 4. Thus for this test case, the system is working at a very high efficiency.

4.3.1 Analysis of Handling Slangs

Sr no.	Input Sentence	Output	Incorrect Words	Total Words
1	Hey cn u ack the receipt of the product at ur end	hey/E can/E you/E acknowledge/E the/E receipt/E of/E the/E product/E your/E end/E	0	11
2	Reply asap. The mtng is imp nd I want u thr	reply/E as/E soon/E as/E possible/E the/E meeting/E is/E important/E and/E i/E want/E you/E there/E	0	14

3	Cn u b ne more annoying, she said. His reply ws masked in the silence of his sadness nd agony	can/E you/E be/E any/E more/E annoying/E she/E said/E his/E reply/E was/E masked/E in/E the/E silence/E of/E his/E sadness/E and/E agony/E	0	20
4	BRB gotta eat smt M famished r8 nw	be/E right/E back/E gotta/E eat/E something/E i/E am/E famished/E right/E now/E	0	11
5	Cmb urgent wrk here now HRU ndhwzur health	call/E me/E back/E urgent/E work/E here/E now/E how/E are/E you/E and/E how/E is/E your/E health/E	0	15
6	Hey Elon Musk started a new startup! I mgonna apply thr dude. It's a once in a lifetime opportunity	hey/E एलों/H musk/E started/E a/E new/E startup/E i/E **i/E** am/E going/E to/E apply/E there/E dude/E it/E **s/E** a/E once/E in/E a/E lifetime/E	2	22
7	So hws life treating u R u happy wid d way things are proceeding r8 nw	so/E **ह्वस/H** life/E treating/E you/E are/E you/E happy/E with/E the/E way/E things/E are/E proceeding/E right/E now/E	1	16
8	Any1 intererested in buying this d2d use chair from us, its in f9 condition	anyone/E interested/E in/E buying/E this/E day/E to/E day/E use/E chair/E from/E us/E it/E is/E in/E fine/E condition/E	0	17
9	Hrt diseases kills many Indians every year	heart/E diseases/E kills/E many/E indians/E every/E year/E	0	7
10	n/t is allowed during exams	**and/E t/E** is/E allowed/E during/E exams/E	2	6

Table 4.2 Test case for Slang

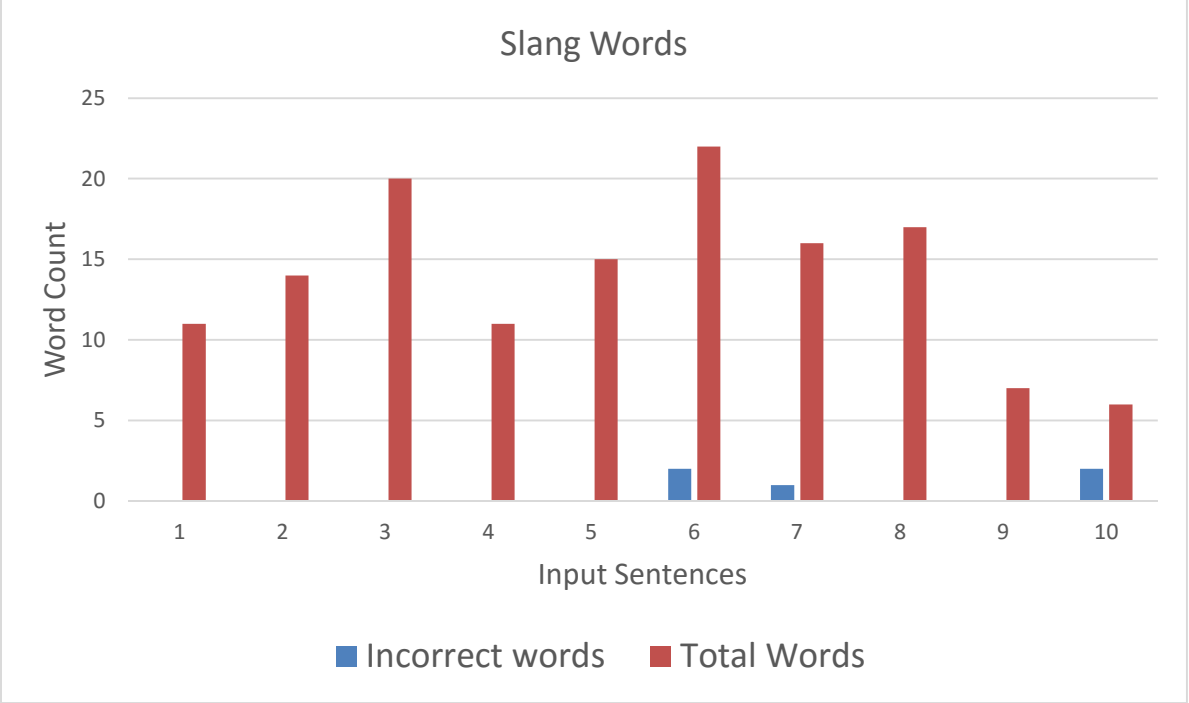

Figure 4.2.3 Simulation results for slang words as input.

The next block in the block diagram is the handling of Slang words. A test case of 10 sentences has been considered. This test case has only the slang words in its sentences; no other types of inputs have been considered. The graph shows the output that the system generates; the correct and incorrect word count for every input sentence can be viewed in the graph.

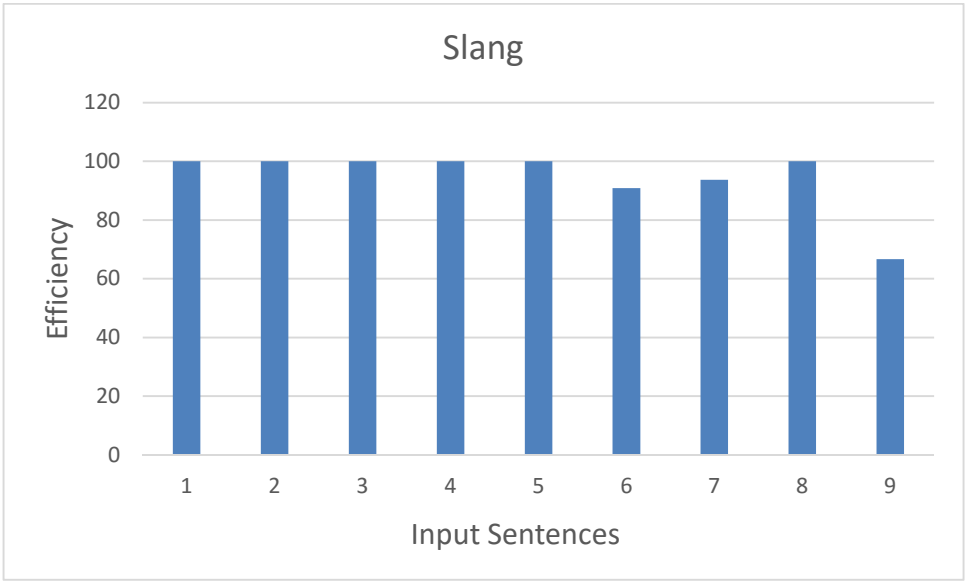

Figure 4.2.4 Efficiency evaluation of Slang words.

The accuracy of the system for evaluation of the slang words came around 96.4%. The correct words predicted were 139 and the incorrect ones were 5.

4.3.1 Analysis of HandlingWordplay

Sr no.	Input Sentence	Output	Incorrect Words	Total Words
1	Heyyyy! I am sooo glad I ran into youuuu! Common lets go outttttttt	hey/E i/E am/E so/E glad/E i/E ran/E into/E you/E common/E lets/E go/E out/E	0	13
2	YOOO! Are you coming to the partyyyyyy! I heard its going to be a blasssst	yo/E are/E you/E coming/E to/E the/E party/E i/E heard/E it/E is/E going/E to/E be/E a/E blast/E	0	16
3	Omgggggggg! Did you listen to that new Ed Sheeransonggggggg!! Its out of thisssssworlddddd.	oh/E my/E god/E did/E you/E listen/E to/E that/E new/E ed/E शीरं/H song/E it/E is/E out/E of/E this/E world/E	0	18
4	The new hippie lifestyle is soooooocooool! I want to try ittttttsooooobadlyyyymannnn	The/E new/E hippie/E lifestyle/E is/E so/E cool/E i/E want/E to/E try/E it/E so/E badly/E man/E	0	15
5	Heyyyy… will you dance at my weddinggggggg!! Pleaseeeeeyaaaarrr… Dancers are neeeeeded dude!	hey/E will/E you/E dance/E at/E my/E wedding/E please/E यार/H dancers/E are/E needed/E dude/E	0	13
6	Like, did you seeeee the new Trump rulingssssss, like what was he thinking likeeee	like/E did/E you/E see/E the/E new/E trump/E rulings/E like/E what/E was/E he/E thinking/E like/E	0	14
7	The lil flea market is soooooocooolll.. I reallllyyyy like the vibes that place gives dudeeeee	the/E little/E flea/E market/E is/E so/E **nullnullल/H**i/E **रेnull/H** like/E the/E vibes/E that/E place/E gives/E dude/E	3	16
8	Did you seeee the Disney movieeee!! It's a book based movie on the life of Belle!!!! Its awsmmmmmm	did/E you/E see/E the/E disney/E movie/E it/E**s/E** a/E book/E based/E movie/E on/E the/E life/E of/E belle/E it/E is/E awesome/E	0	20
9	Watchhhhhandddddlearnnnnnn!!!!! That's howwwwwwitsssssdoneeee…	watch/E and/E learn/E that/E **s/E** how/E it/E is/E done/E	0	9

10	I ammmmmsooooomadddddd at uuuuurtttttnwwwwww.. get outttttttt of hereeeeee	i/E am/E so/E mad/E at/E you/E right/E now/E get/E out/E of/E here/E	0	12

Table 4.3 Test case for Wordplay

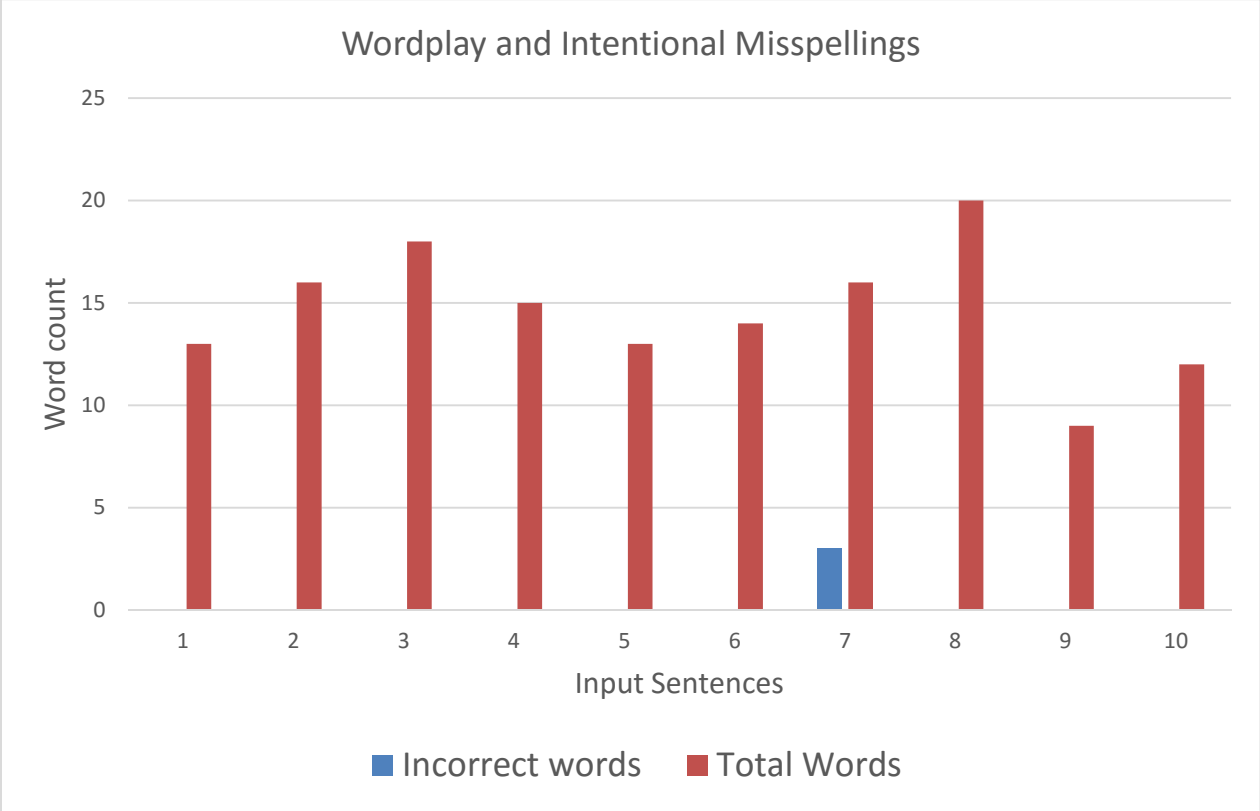

Figure 4.2.5 Simulation results for wordplay and intentionally misspelled words as input.

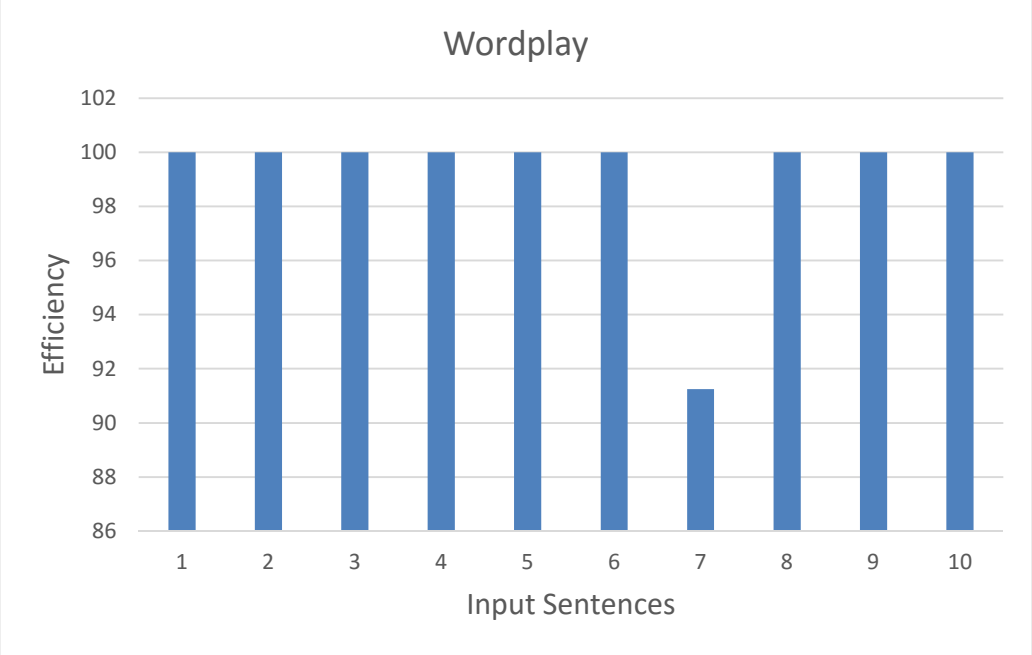

Figure 4.2.6 Efficiency evaluation of Wordplay and intentionally misspelled words.

Now a test case for checking the Wordplay and Intentional Misspelling block has been checked and demonstrated. A test case of 10 sentences has been considered here as well. The graph shows the output that the system generates; the correct and incorrect word count for every input sentence can be viewed in the graph.

The accuracy of the system for evaluation of the wordplay and misspelled words came around 97.9%. The correct words predicted were 143 and the incorrect ones were 3.

4.3.1 Analysis of Transliteration

Sr no.	Input Sentence	Output	Incorrect Words	Total Words
1	Heyyyy! I am sooo glad I ran into youuuu! Common lets go outttttttt	hey/E i/E am/E so/E glad/E i/E ran/E into/E you/E common/E lets/E go/E out/E	0	13
2	YOOO! Are you coming to the partyyyyyy! I heard its going to be a blasssst	yo/E are/E you/E coming/E to/E the/E party/E i/E heard/E it/E is/E going/E to/E be/E a/E blast/E	0	16

3	Omgggggggg! Did you listen to that new Ed Sheeransongggggg!! Its out of thisssssworlddddd.	oh/E my/E god/E did/E you/E listen/E to/E that/E new/E ed/E शीरं/H song/E it/E is/E out/E of/E this/E world/E	0	18
4	The new hippie lifestyle is sooooocooool! I want to try ittttttsooooobadlyyyymannnn	The/E new/E hippie/E lifestyle/E is/E so/E cool/E i/E want/E to/E try/E it/E so/E badly/E man/E	0	15
5	Heyyyy… will you dance at my weddinggggggg!! Pleaseeeeeyaaaarrr… Dancers are neeeeeded dude!	hey/E will/E you/E dance/E at/E my/E wedding/E please/E यार/H dancers/E are/E needed/E dude/E	0	13
6	Like, did you seeeee the new Trump rulingssssss, like what was he thinking likeeee	like/E did/E you/E see/E the/E new/E trump/E rulings/E like/E what/E was/E he/E thinking/E like/E	0	14
7	The lil flea market is sooooocooolll.. I reallllyyyy like the vibes that place gives dudeeeee	the/E little/E flea/E market/E is/E so/E **nullnullल/H**i/E **रेnull/H** like/E the/E vibes/E that/E place/E gives/E dude/E	3	16
8	Did you seeee the Disney movieeee!! It's a book based movie on the life of Belle!!!! Its awsmmmmmm	did/E you/E see/E the/E disney/E movie/E it/E **s/E** a/E book/E based/E movie/E on/E the/E life/E of/E belle/E it/E is/E awesome/E	0	20
9	Watchhhhhandddddlearnnnnnn!!!!! That's howwwwwwitsssssdoneeee…	watch/E and/E learn/E that/E **s/E** how/E it/E is/E done/E	0	9
10	I ammmmmsoooooomaddddddd at uuuuurtttttnwwwwww.. get outttttttt of hereeeeee	i/E am/E so/E mad/E at/E you/E right/E now/E get/E out/E of/E here/E	0	12

Table 4.4 Test case for Transliteration

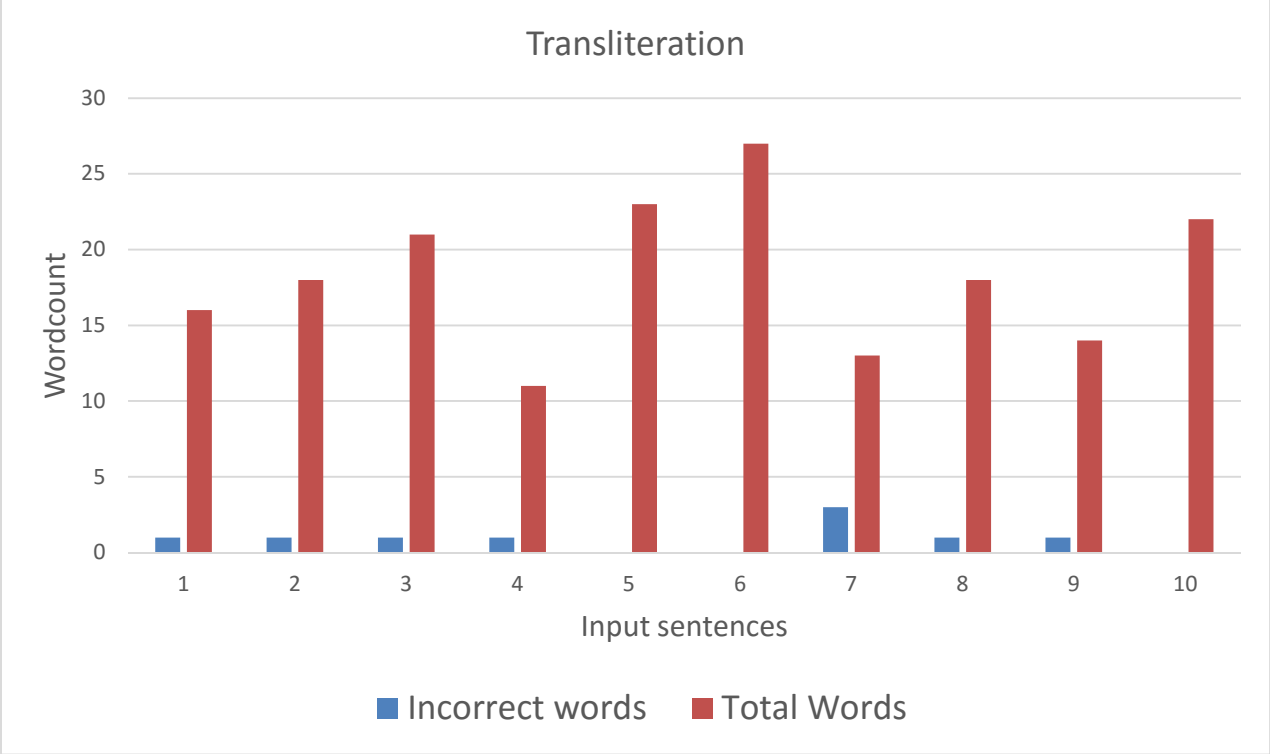

Figure 4.2.7 Simulation results for transliterated words as input.

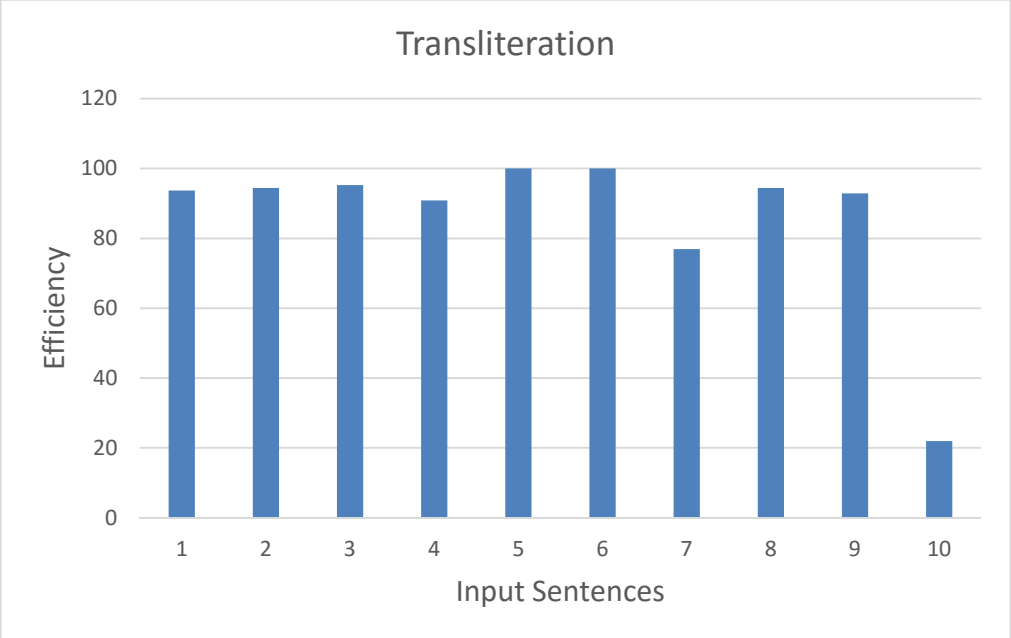

Figure 4.2.8Efficiency evaluation of Transliterated words.

Now a test case for checking the Transliteration block has been checked and demonstrated. Even for evaluation of this block, 10 input sentences have been considered.

The accuracy of the system for evaluation of the transliterated words came around 95% .The correct words predicted were 187 and the incorrect ones were 9.

4.3.1 Analysis of mixed inputs

Sr no.	Input Sentence	Output	Incorrect Words	Total Words
1	2day's match was soooooawsm. Hameshakitarah sab were excited to see match betn gr8 rivalry in cricket. SachinaurSehwagkejodine ne to kamalkardiya	today/E s/E match/E was/E so/E awesome/E हमेशा/H की/H तरह/H सब/H were/E excited/E to/E see/E match/E between/E great/E rivalry/E in/E cricket/E साचिन/H और/H सेह्वाग/Hके/H **जोदीने /H any/E** to/E **कमल/H**कर/H दीया/H	3	31
2	Finally v r in finals. ATB Indian Team ko. Feeling very proud 2 b Indian.	finally/E we/E are/E in/E finals/E all/E the/E best/E indian/E team/E को/H feeling/E very/E proud/E to/E be/E indian/E	0	17
3	GM all Ab India ke cultural heritage kebaare me kyabatavu. India has d richest culture in d world. Pooreduniyamai spread huahai ye.	good/E morning/E all/E ab/E india/E के/H cultural/E heritage/E के/H बारे/H me/E क्या/H बतावु/H india/E has/E the/E richest/E culture/E in/E the/E world/E पूरे/H दुनियाँ/H में/H spread/E हुआ/H हैं/H ये/H	0	28
4	Aamof, MohiniAttam ne India ko world level pe represent kiyahai. A.R. Rehmanekaurzindaudaharanhai.	as/E a/E matter/E of/E fact/E मोहिनी/H अत्ताम/H **any/E**india/E को/H world/E level/E पे/H represent/E किया/H हैं/H a/E **are/E**रहमान/H एक/H और/H ज़िन्दा/H **उदहारं/H**हैं/H	3	24
5	Iirc he has won oscars 4 India in music. Not just art lekin science ke field me bhibohotinsaanhai who hv brought pride to our nation.	if/E i/E remember/E correctly/E he/E has/E won/E oscars/E for/E india/E in/E music/E not/E just/E art/E लेकिन/H science/E के/H field/E me/E भी/H बहुत/H इन्सान/H हैं/H who/E have/E brought/E pride/E to/E our/E nation/E	0	31
6	Hamare culture kopooreduniya ne adopt karnekikoshishkihai. Tysm HAND.	हमारे/H culture/E को/H पूरे/H दुनियाँ/H **any/E** adopt/E करने/H की/H कोशिश/H की/H हैं/H thank/E you/E so/E much/E have/E a/E nice/E day/E	1	20

7	Heyyyyyyy……I am soooooohappyyyyy. I gt d jobbb! D intrvwwssooooooogooood! D off s soooooohugggeee!	hey/E i/E am/E so/E happy/E i/E got/E the/E job/E the/E interview/E was/E so/E **god/E** the/E **off/E s/E** so/E huge/E	3	19
8	ND d ppll! OMG! Dey r d nicessstppl on earthhhh! U Free? Wanna celeb 2nigt!!! Letsgtdnnr!	and/E the/E **nullलल/H** oh/E my/E god/E they/E are/E the/E nicest/E people/E on/E earth/E you/E free/E want/E to/E celebrate/E **nullि null/H** lets/E **got/E** dinner/E	4	22
9	My Treat! Soooooo! D buk I read wssooooogooooddd☹ I wannnacryyyyndlaughhhndsmillle at d sime time!!!!!!	my/E treat/E so/E the/E book/E i/E read/E was/E so/E **god/E**i/E wanna/E cry/E and/E laugh/E and/E smile/E at/E the/E **sime/E** time/E	2	22
10	Omgommmmmmggggg! U hve to see dis pic! D place is in Russia! Btitsssssoooooobeautifulll! Nd the treessssndanimals r soooocuuuttttteeeeeee!	oh/E my/E god/E oh/E my/E god/E oh/E my/E god/E you/E have/E to/E see/E this/E picture/E the/E place/E is/E in/E russia/E but/E its/E so/E beautiful/E and/E the/E trees/E and/E animals/E are/E so/E cute/E	0	32

Table 4.5 Test case sample for Mixed inputs

A bigger test case has been demonstrated. This test case has a total of 70 sentences. The diagram shows the graph with the correct outputs and the incorrect outputs for the mixed test case. The total correct output words generated was 1076 and the incorrect outputs was 56.

The efficiency of the systemcomes around 95%.

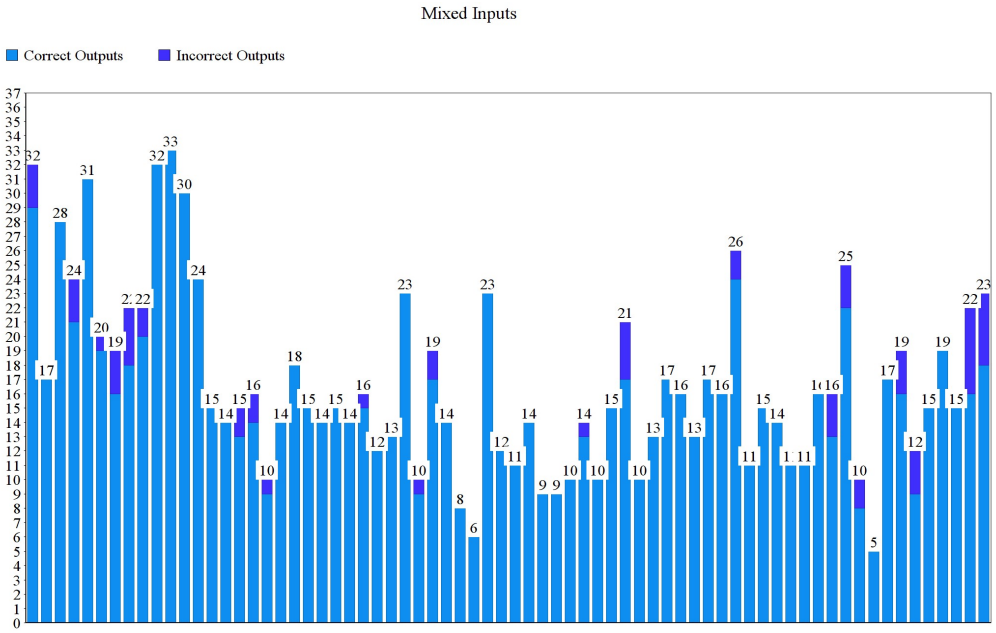

Figure 4.2.9 Simulation results for mixed input sentences.

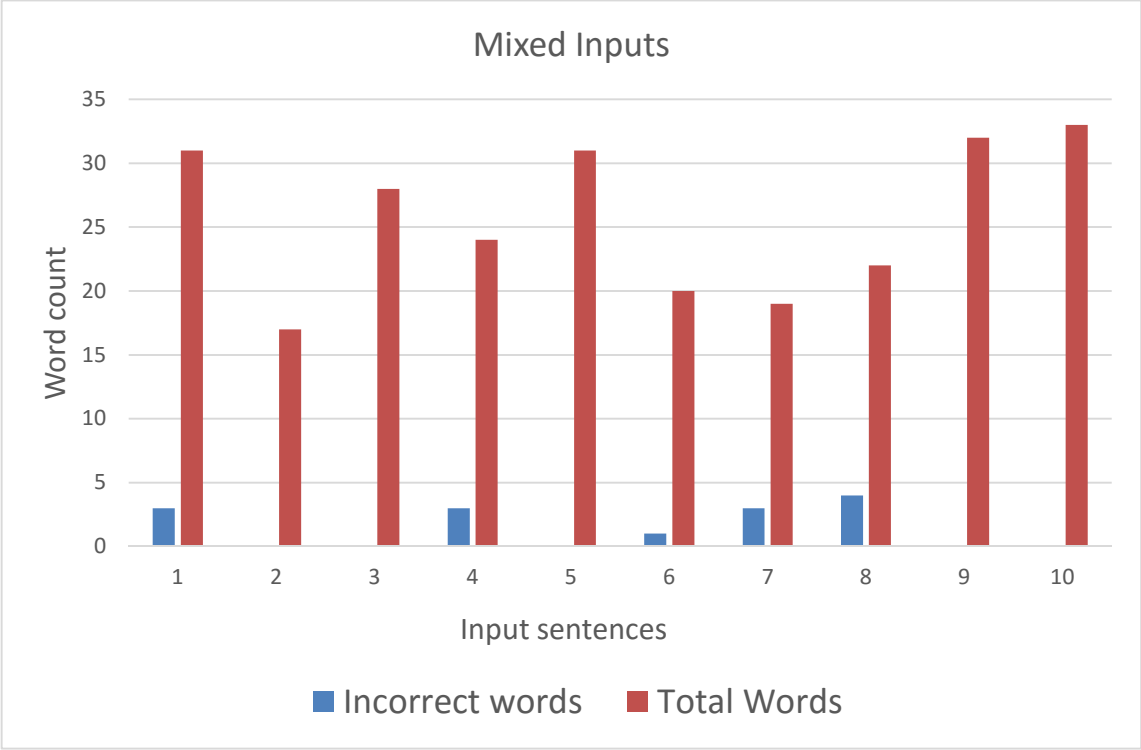

Figure 4.2.10 Simulation results for sample mixed input sentences (10 inputs).

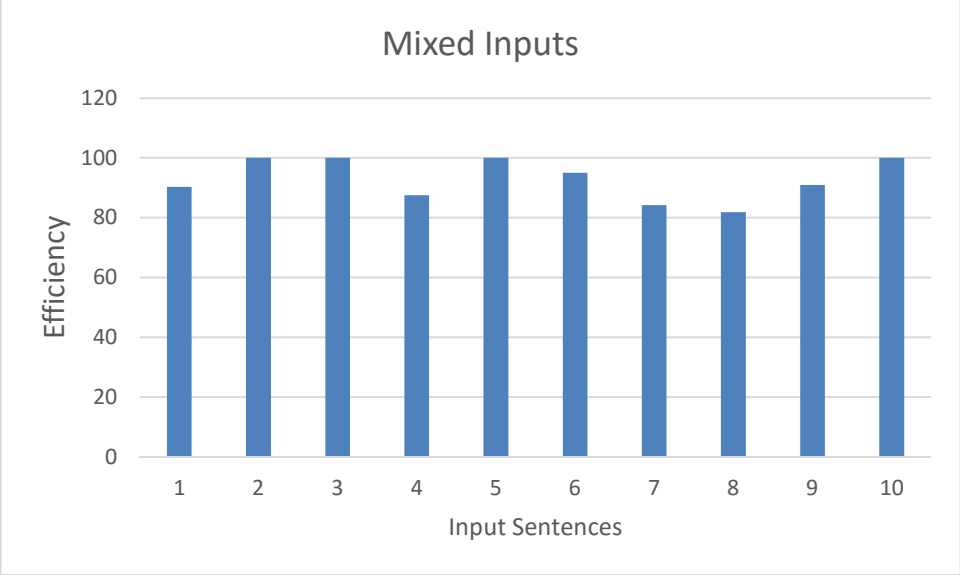

Figure 4.2.11Efficiency evaluation of Mixed inputs.

Chapter 5

Applications

5.1 Sentiment Analysis

Sentiment Analysis, which is also known as opinion mining, is used to obtain the subjective information in any text. The opinion hence obtained can be processed and different outcomes based on the text can be achieved. There have been many successful attempts made in the past to perform Sentiment Analysis of pure English and pure Hindi text. In being one of a kind, this project would be helpful in bridging the gap between both the languages in their pure form and hence allowing to process the languages in code-mixed format.

5.2 Identifying Detractors and Promoters

Customer reviews and opinions have great importance and provide impetus in the company's development. Categorizing the reviews, especially when they are in code-mixed format can be conundrum. Hence, the text normalization of code-mixed data allows this particular problem to be solved. The reviews can then be separated and ordered into Detractors and Promoters and thus allowing a deeper understanding of the position the customer is offering. Also it allows to identify the clients with negative sentiment in social media or news and to increase the margin for transactions with them.

5.3 To Forecast Market Movement Based On News, Blogs and Social Media Sentiment

Social Media is a very strong and important tool for expression of one's ideas, opinions, etc. Because of the ease-ness in uploading text and data on the Internet and social media, there is an abundance of input material available. The input which is in code-mixed language was not processed earlier. Now, because of text normalization the processing of code-mixed data will be possible. And this processing may lead to outputs such as trend prediction, market movement, forecasting any scenario etc.

5.4 Business Analytics

The text analysis tools when deployed in the business sector, it allows to provide reliable and accurate analysis of unstructured and code-mixed text in this case. The business can then use predictive modelling techniques which may then lead to uncovering an abundance of information. Improvement of customer experiences, discovering new market leads, predicting market trends, etc. are some of the further applications that are possible in this area.

Chapter 6
Conclusion and Future Scope

Text Normalizer analyses code-mixed (Hindi+English) text and normalizes it. The different ways in which the sentences are written in code-mixed format i.e. Abbreviations, Slang words, intentional misspelling and wordplay are taken into consideration. Impure text in input is normalized to English text. Input in Romanized Hindi is transliterated to Devanagari. The system has been tested for the above kinds of inputs, individually. Abbreviations are handled with an accuracy of 96.8%. Slang words are handled with an accuracy of 96.4%. Intentionally misspelt words are handled with an accuracy of 97.9%. Transliteration gives an accuracy of 95%. The system has been tested against a sample input data set of 100 inputs. The total accuracy achieved by the system is approximately 95%. The system has certain drawbacks but considering a holistic point-of-view, the output generated is very useful.

The system may be extended to perform sentiment analysis of code-mixed (Hindi+English) text. To handle code mixed input of other regional languages and English, The dictionary used for transliteration may be altered accordingly. To achieve a greater accuracy for abbreviations and slang words instead of rule –based approach machine learning can be used to auto-update the respective dictionaries. Wordplay handling cannot handle words such as ‘gooood’ or ‘toooo’ as they may either be ‘good/god’ or ‘to/too’. Also , words which are common to both hindi and English dictionaries such as ‘sun’ or ‘the’ which may have different meanings in hindi and English are handled incorrectly by the system. The system may be modified to interpret the input sentences and normalize the same by taking the meaning of the sentences into consideration.

Chapter 7

References

[1] Shashank Sharma, PYKL Srinivas,Rakesh Chandra Balabantaray, 2015, Text Normalization of Code mix and Sentoment analysis.

[2] DinkarSitaram, Savitha Murthy, Debrajrai, 2015, Sentiment analysis of mixed language employing Hindi-English code switching.

[3] R. Mahesh K. Sinha, Anil Thakur, 2005, Machine Translation of Bi-lingual Hindi-English (Hinglish) Text.

[4] Ben King, Steven Abney, 2013 Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods

[5] Heeryon Cho, Jong-Seok Lee, Songkuk Kim, 2013 Enhancing lexicon-based review classification by merging and revising sentiment dictionaries

[6] Subhash Chandra, BibekanandaKundu and Sanjay Kumar Choudhury,2013, Hunting Elusive English in Hinglish and Benglish text: Unfolding challenges and Remedies.

[7] Bing Liu ,2012, Sentiment Analysis and Opinion Mining.

[8] Eleanor Clark, Kenji Araki, 2011, Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English

[9] Thomas Gottron, Nedim Lipka2, 2011, A Comparison of Language Identification Approaches on Short, Query-Style Texts

Chapter 8

Acknowledgement

We remain immensely obliged to Dr. SharvariGovilkar for providing us with the idea of this project topic, for her invaluable support in garnering resources for us either by way of information or computers and also her guidance and supervision which made this project happen.

We would like to thank the H.O.D. of Computer Department Dr. Madhumita Chatterjee for her invaluable support.

We are thankful to our principal Dr. R.I.K.Moorthyfor the immense support he gave us while participating in competitions and representing out college at various platforms.

We would also like to thank our project coordinators,Prof. RupaliNikhareand Prof. VarunaakshiBhojane. Also, we would like to say that it has indeed been a fulfilling experience for working out this project topic.

Ms. Neha Pai
Ms. Swetha Ramaswamy
Ms. Manali Vedak

Chapter 9

Annexure

ANNEXURE A: Sample of Abbreviations

aaf	always and forever	cil	check in later
aab	average at best	clm	career limiting move
aaf	as a friend -or- always and forever	cm	call me
aak	alive and kicking	cmb	call me back
aamof	as a matter of fact	cmf	count my fingers
aamoi	as a matter of interest	cmiw	correct me if i'm wrong
aap	always a pleasure	cmu	crack me up
aar	at any rate	cob	close of business
aas	alive and smiling	col	chuckle out loud
abd	already been done	coo	short for cool
abh	actual bodily harm	cot	circle of trust
abk	always be knolling	cpg	consumer packaged goods
abt	absolutely, about	crb	come right back
acd	alt control delete	csa	cool sweet awesome
ace	access control entry	cto	check this out
bak	back at keyboard	dba	doing business as
bau	business as usual	dbd	don't be dumb
bb	bye bye,	ddg	drop dead gorgeous
bb4n	bye bye for now	df	dear friend
bbbg	bye bye be good	dfik	darn if i know
bbc :	bring beer and chips	dftba	don't forget to be awesome
bbiab	be back in a bit	dfwly	don't forget who loves you
bbn	bye bye now	dga	don't go anywhere
bbq	Barbecue	dgt	don't go there
bbr	burnt beyond repair	dhyb	don't hold your breath
bbs	be back soon	diku	do i know you
bbsd	be back soon darling	diy	do it yourself
bbsl	be back sooner or later	djm	don't judge me
bbt	be back tomorrow	dk	don't know
bcbs	big company, big school	dkdc	don't know don't care
bcnu	be seein' you	dmi	don't mention it
bd	big deal	dnbl8	do not be late
bdn	big damn number	dnc	does not compute
bdoa	brain dead on arrival	dnd	donotdisturb

beos	Nudge	doa	dead on arrival
bf	boy friend	doc	drug of choice
bfd	big frickin deal	doe	depends on experience
bff	best friend forever	doei	goodbye (in dutch)

ANNEXURE B: Sample of Slang words

2mrw	Tomorrow	h/o	hold on
2d4	to die for	h/p	hold please
2day	Today	h2cus	hope to see you soon
2dloo	toodle oo	h2s	here to stay
2g2b4g	too good to be forgotten	h4u	hot for you
2g2bt	too good to be true	h4xx0r	hacker -or- to be hacked
2g4u	too good for you	h8	hate
cul	Cool	j/c	just checking
d2d	day-to-day	j/j	just joking
dt	Date	j/k	just kidding
da	There	j/p	just playing
every1	Everyone	j/w	just wondering
evre1	every one	j2lyk	just to let you know
e2eg	each to his/her own	j4f	just for fun
g2g	got to go	j4g	just for grins
g2glys	got to go love ya so	jft	just for today
g4i	go for it	l@u	laughing at you
g4n	good for nothing	l8	late
g98t	good night	lng	long
g9	Genius	lst	last
g8	Great	ly4e	love you forever
gud	Good	m/f	male or female
gd	Good	m2ny	me too, not yet
h&k	hug and kiss	m4c	meet for coffee
2l8	too late	nc	nice
2u2	to you too	no1	no one
2mor	Tomorrow	ntng	nothing
2qt	too cute	nw	new
2n8	Tonight	nxt	next
3sum	threesome	o3	out of office
4col	for crying out loud	sm1	someone
4e	Forever	w4u	waiting for you
4eae	forever and ever	w8	wait
4f?	for friends?	w8 4 me	wait for me
4nr	foreigner	w8am	wait a minute

a2d	agree to disagree	w8n	waiting
a3	anywhere, anytime, anyplace	w9	wife in room
abt	about	wan2	want to
abl	Able	wan2tlk	want to talk?
abt2	about to	wht	what
awsm	awesome	wid	with
any1	anyone	spk	speak
b2a	business-to-anyone	wrst	worst

ANNEXURE C: Sample of Standard English Dictionary

abandon	break in	cabinet	day	ugly	wonderful
abandoned	break into	cable	dead	ultimate	wood
ability	break off	cake	deaf	ultimately	wooden
able	break out	calculate	deal	umbrella	wool
about	break up	calculation	deal in	unable	word
above	breast	call	deal with	unacceptable	work
abroad	breath	call back	dear	uncertain	worker
absence	breathe	called	death	uncle	working
absent	breathe in	call for	debate	uncomfortable	work out
absolute	breathe out	call off	debt	unconscious	world
absolutely	breathing	call up	decade	uncontrolled	worried
absorb	Breed	calm	decay	under	worry
abuse	Brick	calm down	December	underground	worrying
abuse	bridge	calmly	decide	underneath	worse
academic	Brief	camera	decide on	understand	worship
accent	briefly	camp	decision	understanding	worst
accept	bright	campaign	declare	underwater	worth
acceptable	brightly	camping	decline	underwear	would
access	brilliant	can 1	decorate	undo	wound 1
accident	Bring	can 2	decoration	unemployed	wounded
accidental	bring back	cancel	decorative	unemployment	wrap
accidentally	bring down	cancer	decrease	unexpected	wrapping
according to	bring out	candy	deeply	unfair	write
account	bring up	cannot	defeat	unfairly	write back
account for	Broad	cap	defense	unfortunate	write down
accurate	broadcast	capable	defend	unfortunately	writer
accurately	broadly	capacity	define	unfriendly	writing
accuse	broken	capital	definite	unhappy	written
achieve	brother	captain	definitely	uniform	wrong
achievement	brown	capture	definition	unimportant	wrongly
acid	Brush	car	degree	union	yard
acknowledge	bubble	card	delay	unique	yawn

a couple	budget	cardboard	deliberate	unit	yeah
acquire	Build	care	deliberately	unite	year
across	building	career	delicate	united	yellow
act	build up	care for	delight	universe	yes
action	Bullet	careful	delighted	university	yesterday
active	Bunch	carefully	deliver	unkind	yet
actively	Burn	careless	delivery	unknown	you
activity	burn down	carelessly	demand	unless	young
actor	Burnt	carpet	demonstrate	unlike	your
actress	Burst	carrot	dentist	unlikely	yours
actual	burst into	carry	deny	unload	yourself

ANNEXURE D: Hindi to English conversion list (Barakhadi for Transliteration process)

a,अ	gah,गः	zee,झी	do,डो	da,द	fu,फु	yo,यो	shah,शः
a,अ	gh,घ	zu,झु	dau,डौ	daa,दा	foo,फू	yau,यौ	shna,ष्न
aa,आ	gha,घा	zoo,झू	dan,डं	di,दि	fe,फे	yan,यं	shnaa,ष्ना
i,इ	ghaa,घा	ze,झे	dam,डं	dee,दी	fai,फै	yam,यं	s,स
ee,ई	ghi,घि	zai,झै	dah,डः	du,दु	fo,फो	yah,यः	sa,सा
u,उ	ghee,घी	zo,झो	dh,ढ	doo,दू	fau,फौ	r,र	saa,सा
oo,ऊ	ghu,घु	zau,झौ	dha,ढ	de,दे	fan,फं	ra,रा	si,सि
e,ए	ghoo,घू	zan,झं	dhaa,ढा	dai,दै	fam,फं	raa,रा	see,सी
ea,ए	ghe,घे	zam,झं	dhi,ढि	do,दो	fah,फः	ri,रि	su,सु
ai,ऐ	ghai,घै	zah,झः	dhee,ढी	dau,दौ	b,ब	ree,री	soo,सू
ei,ऐ	gho,घो	tra,त्र	dhu,ढु	dan,दं	ba,ब	ru,रु	se,से
o,ओ	ghau,घौ	traa,त्रा	dhoo,ढू	dam,दं	baa,बा	roo,रू	sai,सै
ou,औ	ghan,घं	tri,त्रि	dhe,ढे	dah,दः	bi,बि	re,रे	so,सो
au,औ	gham,घं	tree,त्री	dhai,ढै	dh,ध	bee,बी	rai,रै	sau,सौ
an,अं	ghaah,घः	tru,त्रु	dho,ढो	dha,ध	bu,बु	ro,रो	sam,सं
am,अं	ch,च	troo,त्रू	dhau,ढौ	dhaa,धा	boo,बू	rau,रौ	san,सं
ah,अः	cha,च	tre,त्रे	dhan,ढं	dhi,धि	be,बे	ran,रं	san,सां
aha,अः	chaa,चा	trai,त्रै	dham,ढं	dhee,धी	bai,बै	ram,रं	sah,सः
ru,ऋ	chi,चि	tro,त्रो	dhah,ढः	dhu,धु	bo,बो	rah,रः	h,ह
k,क	chee,ची	trau,त्रौ	na,ण	dhoo,धू	bau,बौ	l,ल	ha,हा
ka,क	chu,चु	tran,त्रं	naa,णा	dhe,धे	ban,बं	la,ल	haa,हा
kaa,का	choo,चू	tram,त्रं	ni,णि	dhai,धै	bam,बं	laa,ला	hi,हि
ki,कि	che,चे	trah,त्रः	nee,णी	dho,धो	bah,बः	li,लि	hee,ही
kee,की	chai,चै	t,ट	nu,णु	dhau,धौ	bh,भ	lee,ली	hu,हु
ku,कु	cho,चो	ta,ट	noo,णू	dhan,धं	bha,भ	lu,लु	hoo,हू
koo,कू	chau,चौ	taa,टा	ne,णे	dham,धं	bhaa,भा	loo,लू	he,हे
ke,के	chan,चं	ti,टि	nai,णै	dhah,धः	bhi,भि	le,ले	hai,है
kai,कै	cham,चं	tee,टी	no,णो	n,न	bhee,भी	lai,लै	ho,हो

ko,को	chah,चः	tu,टु	nau,णौ	na,न	bhu,भु	lo,लो	hau,हौ
kau,कौ	chha,छ	too,टू	nan,णं	naa,ना	bhoo,भू	lau,लौ	han,हं
kan,कं	chhaa,छा	te,टे	nam,णं	ni,नि	bhe,भे	lan,लं	ham,हं
kam,कं	chhi,छि	tai,टै	nah,णः	nee,नी	bhai,भै	lam,लं	hah,हः
kah,कः	chhee,छी	to,टो	t,त	nu,नु	bho,भो	lah,लः	ksh,क्ष
kh,ख	chhu,छु	tau,टौ	ta,ता	noo,नू	bhau,भौ	v,व	ksha,क्ष
kha,ख	chhoo,छू	tan,टं	taa,ता	ne,ने	bhan,भं	va,व	kshaa,क्षा
khaa,खा	chhe,छे	tam,टं	ti,ति	nai,नै	bham,भं	vaa,वा	ksha,क्षा
khi,खि	chhai,छै	tah,टः	tee,ती	no,नो	bhah,भः	vi,वि	kshi,क्षि
khee,खी	chho,छो	th,ठ	tu,तु	nau,नौ	m,म	vee,वी	kshee,क्षी
khu,खु	chhau,छौ	tha,ठा	too,तू	nan,नं	ma,म	vu,वु	kshu,क्षु
khoo,खू	chhan,छं	thaa,ठा	te,ते	nam,नं	maa,मा	voo,वू	kshoo,क्षू
khe,खे	chham,छं	thi,ठि	tai,तै	nah,नः	mi,मि	ve,वे	kshe,क्षे

mba,म्ब	dm,द्म	phi,फ़ि	jam,जं	p,प	mee,मी	gnaa,ज्ञा
rsh,र्श	dr,ड्र	phee,फ़ी	jah,जः	pa,प	mu,मु	Gni,ञ्जि
pra,प्र	dr,द्र	phu,फ़ु	z,झ	paa,पा	moo,मू	gnee,ञ्जी
mru,मृ	dhr,ध्र	phoo,फ़ू	za,झ	pi,पि	me,मे	gnu,ञ्जु
rva,र्व	dv,द्व	phe,फ़े	zaa,झा	pee,पी	mai,मै	gnoo,ञ्जू
kk,क्क	dy,द्य	phai,फ़ै	zi,झि	pu,पु	mo,मो	gne,ञ्जे
kka,क्का	ft,फ़्त	pho,फ़ो	thee,ठी	poo,पू	mau,मौ	gnai,ञ्जै
kke,क्के	gr,ग्र	phau,फ़ौ	thu,ठु	pe,पे	man,मं	gno,ञ्जो
kki,क्की	gy,ग्य	my,मी	thoo,ठू	pai,पै	mam,मं	gnau,ञ्जौ
kkai,क्कै	khm,ख्म	khai,खै	the,ठे	po,पो	mah,मः	gnam,ञ्जं
kku,क्कु	khy,ख्य	kho,खो	thai,ठै	pau,पौ	y,य	gnah,ञ्जः
kru,कृ	kl,क्ल	khau,खौ	tho,ठो	pan,पं	ya,य	bb,ब्ब
gru,गृ	kr,क्र	khan,खं	thau,ठौ	pam,पं	yaa,या	Bhr,भ्र
ghru,घृ	ks,क्स	kham,खं	than,ठं	pah,पः	yi,यि	br,ब्र
nru,नृ	ky,क्य	khah,खः	tham,ठं	f,फ	yee,यी	ddh,द्ध
vru,वृ	ld,ल्द	g,ग	thah,ठः	fa,फ	yu,यु	dhr,ढ्र
shtra,ष्ट्र	lm,ल्म	ga,गा	d,ड	faa,फा	yoo,यू	tn,त्न
shree,श्री	mb,म्ब	gaa,गा	da,ड	fhi,फ़ि	ye,ये	tren,ट्र
shri,श्री	mh,म्ह	gi,गि	daa,डा	fee,फ़ी	yai,यै	tt,त्त
kra,क्र	nd,ण्ड	gee,गी	di,डि	vai,वै	kshai,क्षै	vr,व्र
kraa,क्रा	nn,न्न	gu,गु	dee,डी	vo,वो	ksho,क्षो	hw,ह्व
gra,ग्रा	ns,न्स	goo,गू	du,डु	vau,वौ	kshau,क्षौ	dhy,ध्य
rgi,र्गी	nt,न्त	ge,गे	doo,डू	van,वं	kshan,क्षं	dhya,ध्या
swa,स्व	pr,प्र	gai,गै	de,डे	vam,वं	ksham,क्षं	dhyaa,ध्या
sva,स्व	rc,र्च	go,गो	dai,डै	vah,वः	kshah,क्षः	w,व
gve,ग्वे	rf,र्फ़	gau,गौ	to,तो	sh,श	dny,ज्ञ	wa,व
bda,ब्द	rs,र्स	gan,गं	tau,तौ	sha,श	dnya,ज्ञा	waa,वा
hin,हिं	rt,र्त	gam,गं	tan,तं	shaa,शा	dnyi,ज्ञि	shw,श्व
di,दी	rth,र्थ	chhah,छः	tam,तं	shi,शि	dnyee,ज्ञी	shwa,श्वा

nt,त	ry,र्य	j,ज	tah,तः	shee,शी	dnyu,ज्ञु	shwaa,श्वा
nd,द	shr,ष्र	ja,ज	tha,थ	shu,शु	dnyoo,ज्ञू	pt,प्त
shan,शां	shr,श्र	jaa,जा	thaa,था	shoo,शू	dnye,ज्ञे	lk,ल्क
lin,लिं	shv,श्व	ji,जि	thi,थि	she,शे	dnyai,ज्ञै	ph,फ
ad,अद	sj,स्ज	jee,जी	thee,थी	shai,शै	dnyo,ज्ञो	pha,फ
dwai,द्वै	sk,स्क	ju,जु	thu,थु	sho,शो	dnyau,ज्ञौ	phaa,फा
dve,द्वे	sn,स्न	joo,जू	thoo,थू	shau,शौ	dnyan,ज्ञं	
wai,वै	st,स्ट	je,जे	the,थे	shan,शं	dnyam,ज्ञं	
bhon,भों	st,स्त	jai,जै	thai,थै	sham,शं	dnyah,ज्ञः	
sw,श्व	sth,स्थ	jo,जो	tho,थो	tham,थं	chch,च्च	
xn,ङ	sv,स्व	jau,जौ	thau,थौ	thah,थः	chchh,च्छ	
gna,ज्ञ	thr,थ्र	jan,जं	than,थं	d,द	chhu,ट्टु	

ANNEXURE E: Hindi to English conversion list (Maatra for Transliteration process)

a,ा
a,ा
i,ि
ii,ी
u,ु
uu,ू
r,ृ
e,े
ai,ै
o,ो
au,ौ
n,ं
m,ं

ANNEXURE F: Sample of Hindi to English dictionary of common words

aah आह	aanevaala आनेवाला	baadha बाधा	chalte चलते
aaha आहा	aanevaalaa आनेवाला	baadi बाड़ी	chaltey चल्ते
aahaa आहा	aanevaale आनेवाले	baadlon बादलों	chalte छलते
aahaahaa आहाहा	aanevaalii आनेवाली	baadshaah बादशाह	chalthi चलती
aahat आहत	aanewaala आनेवाल	baadshaah बाद्शाह	chalti चलती
aahaten आहटें	aanewaala आनेवाला	baadshaahon बाद्शाहों	chalu चालू
aahaton आहटों	aanewaale आनेवाले	baadshah बादशाह	chalun चलूँ
aahe आहे	aanewaali आनेवालि	baadshahon बादशाहों	chalun चलूं
aahee आहे	aanewala आनेवाला	baaen बाएं	chalungee चलूँगी
aahen आहें	aanewale आनेवाले	baag बाग	chalungi चलूँगी
aahista आहिस्ता	aanewaley आनेवाले	baag बाग़	chaluun चलूँ
aahistaa आहिस्ता	aang अंग	baag भाग	chaluun चलूं
aaho आहो	aangan आँगन	baaga बागा	chaluunga चलूंगा
aahon आहों	aanganaa आँगना	baagad बागड़	chaluungaa चलूंगा
aaida आईदा	aanganon आँगनों	baageshvari बागेश्वरी	chaluungi चलूँगी
aaika आईका	aangdayi अंगड़ाई	baaghi बागी	chalye चले
aaina आइना	aangna अंगना	baaghon बागों	cham चम
aaina आईना	aangna आँगना	baaghon बाग़ों	cham छम
aainaa आइना	aanhe आँहे	baago बागों	chama चम
aainaa आईना	aanhen आँहें	baagon बागों	chama छमा
aaine आइने	aanhon आँहों	baagon बाग़ों	chamaacham चमाचम
aainey आईने	aankade आँकड़े	baagon बाहों	chamacha चमचा
aais आइस	aankana आँकना	baah बाँह	chamacham चमचम
aaitabaar आइतबार	aankein आँखें	baahaake बाहाके	chamache चमचे
aaitbaar आइतबार	aankh आँख	baahaane बाहाने	chamak चमक
aaiya आइया	aankhe आँखे	baahaar बाहार	chamak छमक
aaiye आइये	aankhein आँखें	baahar बहार	chamaka चमका
aaiyey आईये	aankhen आँखें	baahar बाहर	chamakaa चमका
aaiyega आइयेगा	aankhen आँखेन	baaharakii बाहरकी	chamakaae चमकाए
aaiyo आइयो	aankhey आँखें	baahe बहे	chamakaakar चमकाकर
aaj आज	aankhiyaan आँखियाँ	baahe बाँहें	chamakaana चमकाना
aaja आजा	aankhiyaan अंखियाँ	baahe बाहे	chamakaao चमकाओ
aajaa आजा	aankho आँखो	baahe बाहें	chamakaati चमकाती
aajaae आजाए	aankho आँखों	baahein बाँहें	chamakaauun चमकाऊं
aajaaegaa आजाएगा	aankhoen आँखोएन	baahein बाहें	chamakaaye चमकाये
aajaaein आजाएं	aankhon आँखों	baahen बाँहें	chamakaayen चमकायें
aajaao आजाओ	aankkhon आँखों	baahen बाहें	chamakaayi चमकाई
aajaate आजाते	aankon आँखों	baahen बाहों	chamakaayuचमकाऊँ
aajaaye आजाये	aanso आँसू	baahir बाहिर	chamakan चमकन
aajaayen आजाएं	aansoo आँसू	baaho बाँहों	chamakana चमकना
aajama आजमा	aansu आँसु	baaho बाहों	chamakane चमकने

ANNEXURE G: Transliteration in example sentences with output

1	Kya aapne humare lia khana laya hai? Aj meri maa ne dabba nai dia hai humko	क्या/H आपने/H हमारे/H लिए/H खाना/H लाया/H हैं/H अज/H मेरि/H मां/H **any/E**दब्बा/H नै/H दिया/H हैं/H हमको/H
2	Aaj kal garmi kitni badh gai hai. Ghar se bahar nikalne ke lia bohot taklif hoti hai.	आज/H कल/H गरमी/H कित्नी/H बढ़/H गयी/H हैं/H घर/H से/H बाहर/H निकलने/H के/H लिए/H बहुत/H **ताक्लिफ/H**होतीं/H हैं/H
3	Baki sabko bhi bata dena, aaj ka khana mere ghar pe hai. Mere naya ghar dekhne zaroor ana tum sab log.	बाकी/H सबको/H भी/H बता/H देना/H आज/H का/H खाना/H मेरे/H घर/H पे/H हैं/H मेरा/H नया/H घर/H देखने/H ज़रुर/H आना/H तूम/H सब/H **log/E**
4	Kya hum aaj ghumne jaa sakte hai. Aaj meri chhutti hai.	क्या/H **hum/E**आज/H घुमने/H जा/H सकते/H हैं/H आज/H मेरि/H छुट्टि/H हैं/H
5	Vo kya hai na, aaj meri dost ki shaadi hai, toh isiliye mai itne saj daj ke nikli hu ghar se.	वो/H क्या/H हैं/H ना/H आज/H मेरि/H dost/E कि/H शादी/H हैं/H तोह/H इसीलिए/H में/H इतने/H सज/H दज/H के/H निकली/H हु/H घर/H से/H
6	Mere ghar walon ne mujhe aj raat tere ghar pe rukhne k lia kaha hai. Mere mata aur pita bahar jaa rahe hai na, isiliye.	मेरे/H घर/H वालों/H **any/E**मुझे/H आज /H रात/H तेरे/H घर/H पे/H रुखने/H **ok/E**लिए/H कहा/H हैं/H मेरे/H माता/H और/H पिता/H बाहर/H जा/H राहें/H हैं/H ना/H इसीलिए/H
7	Mujhe kab milaoge apne doston se, me sirf unki kahaniyan sun rahi hu.	मुझे/H **kab/E**मिलाओगे/H अपने/H दोस्तों/H से/H **me/E**सिर्फ़/H उनकी/H कहानियाँ/H **sun/E**रही/H हु/H
8	Mujhe ek din ki chhutti chahiye, jab mai poori din kitaab lekar ghar pe baithke padh sakti hu.	मुझे/H एक/H दिन/H कि/H छुट्टि/H चहिये/H जब/H में/H पूरी/H दिन/H **ख़िताब/H**लेकर/H घर/H पे/H बैठके/H पढ़/H सकती/H हु/H
9	Aj bohot thak gai hu yaar, kuch paani ya kuch de dena .	आज /H बहुत/H थक/H गयी/H हु/H यार/H कुछ/H पाणी/H **ya/E** कुछ/H दे/H देना/H
10	Mujhe tum paise kab tak vapis dogi, tangi padhi hai yaar . Thoda jaldi se de dena, mere bhi halat ho raha hai.	मुझे/H तूम/H paise/E कब/H तक/H वपिस/H दोगी/H तंगि/H पढ़ी/H हैं/H यार/H थोड़ा/H जल्दी/H से/H दे/H देना/H मेरे/H भी/H हालत/H हो/H रहा/H हैं/H

ANNEXURE H: Wordplay in example sentences with output

1	Heyyyy! I am sooo glad I ran into youuuu! Common lets go outttttttt	hey/E i/E am/E so/E glad/E i/E ran/E into/E you/E common/E lets/E go/E out/E
2	YOOO! Are you coming to the partyyyyyy! I heard its going to be a blasssst	yo/E are/E you/E coming/E to/E the/E party/E i/E heard/E it/E is/E going/E to/E be/E a/E blast/E
3	Omgggggggg! Did you listen to that new Ed Sheeran songgggggg!! Its out of thisssss worlddddd.	oh/E my/E god/E did/E you/E listen/E to/E that/E new/E ed/E शीरं/H song/E it/E is/E out/E of/E this/E world/E
4	The new hippie lifestyle is sooooo cooool! I want to try ittttttt sooooo badlyyyy mannnn	The/E new/E hippie/E lifestyle/E is/E so/E cool/E i/E want/E to/E try/E it/E so/E badly/E man/E
5	Heyyyy… will you dance at my weddingggggg!! Pleaseeeee yaaaarrr… Dancers are neeeeeded dude!	hey/E will/E you/E dance/E at/E my/E wedding/E please/E यार/H dancers/E are/E needed/E dude/E
6	Like, did you seeeee the new Trump rulingssssss, like what was he thinking likeeee	like/E did/E you/E see/E the/E new/E trump/E rulings/E like/E what/E was/E he/E thinking/E like/E
7	The lil flea market is sooooo cooolll.. I realllllyyyy like the vibes that place gives dudeeeee	the/E little/E flea/E market/E is/E so/E **nullnullल/H** i/E **रेnull/H** like/E the/E vibes/E that/E place/E gives/E dude/E
8	Did you seeee the Disney movieeee!! It's a book based movie on the life of Belle!!!! Its awsmmmmmm	did/E you/E see/E the/E disney/E movie/E it/**E s/E** a/E book/E based/E movie/E on/E the/E life/E of/E belle/E it/E is/E awesome/E
9	Watchhhhh andddddd learnnnnnn!!!!! That's howwwwww itsssss doneeee…	watch/E and/E learn/E that/E **s/E** how/E it/E is/E done/E
10	I ammmmm sooooo madddddd at uuuuu rtttttt nwwwwww.. get outttttttt of hereeeeee	i/E am/E so/E mad/E at/E you/E right/E now/E get/E out/E of/E here/E

ANNEXURE I: Slang example sentences with output

1	Hey cn u ack the receipt of the product at ur end	hey/E can/E you/E acknowledge/E the/E receipt/E of/E the/E product/E your/E end/E
2	Reply asap. The mtng is imp nd I want u thr	reply/E as/E soon/E as/E possible/E the/E meeting/E is/E important/E and/E i/E want/E you/E there/E
3	Cn u b ne more annoying, she said. His reply ws masked in the silence of his sadness nd agony	can/E you/E be/E any/E more/E annoying/E she/E said/E his/E reply/E was/E masked/E in/E the/E silence/E of/E his/E sadness/E and/E agony/E
4	BRB gotta eat smt M famished r8 nw	be/E right/E back/E gotta/E eat/E something/E i/E am/E famished/E right/E now/E
5	Cmb urgent wrk here now HRU nd hwz ur health	call/E me/E back/E urgent/E work/E here/E now/E how/E are/E you/E and/E how/E is/E your/E health/E
6	Hey Elon Musk started a new startup! I m gonna apply thr dude. It's a once in a lifetime opportunity	hey/E एलों/H musk/E started/E a/E new/E startup/E i/E **i/E** am/E going/E to/E apply/E there/E dude/E it/E **s/E** a/E once/E in/E a/E lifetime/E
7	So hws life treating u R u happy wid d way things are proceeding r8 nw	so/E **ह्वस/H** life/E treating/E you/E are/E you/E happy/E with/E the/E way/E things/E are/E proceeding/E right/E now/E
8	Any1 intererested in buying this d2d use chair from us, its in f9 condition	anyone/E interested/E in/E buying/E this/E day/E to/E day/E use/E chair/E from/E us/E it/E is/E in/E fine/E condition/E
9	Hrt diseases kills many Indians every year	heart/E diseases/E kills/E many/E indians/E every/E year/E
10	n/t is allowed during exams	**and/E t/E** is/E allowed/E during/E exams/E

ANNEXURE J: Abbreviations in example sentences with output

1	Dmi I am glad to help you df	don't/E mention/E it/E i/E am/E glad/E to/E help/E you/E dear/E friend/E
2	Dwb its an automatic reply	don't/E write/E back/E it/E is/E an/E automatic/E reply/E
3	After watching friends icl its rofl nd lol funny man!	after/E watching/E friends/E **इnull/H** it/E is/E rolling/E on/E the/E floor/E laughing/E and/E laugh/E out/E loud/E funny/E man/E
4	Abd work needs to be add by the owner of the works, kindly comply	already/E been/E done/E work/E needs/E to/E be/E add/E **bye/E** the/E owner/E of/E the/E works/E kindly/E comply/E
5	Bb , hand and call me when you reach at your destination	bye/E bye/E have/E a/E nice/E day/E and/E call/E me/E when/E you/E reach/E at/E your/E destination/E
6	Cpg is highly overrated, one should use organic items instead	consumer/E packaged/E goods/E is/E highly/E overrated/E one/E should/E use/E organic/E items/E instead/E
7	Cot is the secret for every successful relationship in this world	circle/E of/E trust/E is/E the/E secret/E for/E every/E successful/E relationship/E in/E this/E world/E
8	Dmi , it was my pleasure	don't/E mention/E it/E it/E was/E my/E pleasure/E
9	Fb is my fav social media / social networking site man!	**फnullnullबूक** is/E my/E favorite/E social/E media/E social/E networking/E site/E man/E
10	Gmab , is what ross was intending to tell Rachel, but it came out all wrong.	give/E me/E a/E break/E is/E what/E ross/E was/E intending/E to/E tell/E **राचेल/H** but/E it/E came/E out/E all/E wrong/E

Chapter 10

Publications and Achievements

- We were selected to represent Pillai College of Engineering in the Research Project Convention Avishkar , wherein we presented our Project's poster. The following are the certificates that was awarded to us for presenting in the event.

University of Mumbai

Department of Students' Welfare

Organises

AVISHKAR

Research Convention 2016-2017

DISTRICT LEVEL ROUND

Venue:

Ramrao Adik Institute of Technology

Nerul, Navi Mumbai - 400 706.

Certificate

This is to certify that Mr. / Ms. SWETHA RAMASWAMY *student of* PILLAI COLLEGE OF ENGINEERING *has participated in District Level Round of Avishkar Research Convention held on 19th October 2016 and presented his/her project titled.* TEXT NORMALIZATION OF CODE-MIXED TEXT & SENTIMENT ANALYSIS

Dr. Mukesh Patil	Dr. Ramesh Vasappanavara	Dr. R. C. Patil	Dr. Sunil Patil
Co-ordinator	Principal	Overall Co-ordinator	Director, DSW, Convener

University of Mumbai

Department of Students' Welfare

Organises

AVISHKAR

Research Convention 2016-2017

DISTRICT LEVEL ROUND

Venue:

Ramrao Adik Institute of Technology

Nerul, Navi Mumbai - 400 706.

Certificate

This is to certify that Mr. / Ms. NEHA PAL *student of* PILLAI COLLEGE OF ENGINEERING *has participated in District Level Round of Avishkar Research Convention held on 19th October 2016 and presented his/her project titled.* TEXT NORMALIZATION OF CODE-MIXED TEXT & SENTIMENT ANALYSIS

Dr. Mukesh Patil	Dr. Ramesh Vasappanavara	Dr. R. C. Patil	Dr. Sunil Patil
Co-ordinator	*Principal*	*Overall Co-ordinator*	*Director, DSW, Convener*

University of Mumbai

Department of Students' Welfare

Organises

AVISHKAR

Research Convention 2016-2017

DISTRICT LEVEL ROUND

Venue:

Ramrao Adik Institute of Technology

Nerul, Navi Mumbai - 400 706.

Certificate

This is to certify that Mr. / Ms. MANALI VEDAK *student of* PILLAI COLLEGE OF ENGINEERING *has participated in District Level Round of Avishkar Research Convention held on 19th October 2016 and presented his/her project titled.* TEXT NORMALIZATION OF CODE-MIXED TEXT & SENTIMENT ANALYSIS

Dr. Mukesh Patil	Dr. Ramesh Vasappanavara	Dr. R. C. Patil	Dr. Sunil Patil
Co-ordinator	*Principal*	*Overall Co-ordinator*	*Director, DSW, Convener*

- We also participated in the Technical Paper Paper presentation competition held during Alegria 2017. We were awarded first prize for the same. The following certificates were awarded to us for the win.

Mahatma Education Society's

Pillai Group of Institutions

CERTIFICATE OF ACHIEVEMENT

This is to certify that Mr./Ms. Neha Pai
has been declared the Winner of the TPP
competition conducted as part of Alegria-The Festival of Joy 2017,
organized by Mahatma Education Society's Pillai Group of Institutions
at Dr.K.M Vasudevan Pillai Campus,New Panvel.

alegria
THE KINGDOM OF GAMES

Dr. Priam Pillai
Chief Operating Officer
Mahatma Education Society

Dr. Daphne Pillai
Chairperson, Management Board,
Mahatma Education Society