INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Annotated corpus creation for sentiment analysis in code-mixed Hindi-English (Hinglish) social network data

Neha Garg[1]*, **Kamlesh Sharma**[2]

1 Assistant Professor, Department of Computer Science & Engineering, Manav Rachna International Institute of Research & Studies, Faridabad, Haryana, India. Tel.: +91-8882879289

2 Associate Professor, Department of Computer Science & Engineering, Manav Rachna International Institute of Research & Studies, Faridabad, Haryana, India

OPEN ACCESS

Received: 26.09.2020

Accepted: 08.11.2020

Published: 20.11.2020

Editor: Dr. Natarajan Gajendran

Citation: Garg N, Sharma K (2020) Annotated corpus creation for sentiment analysis in code-mixed Hindi-English (Hinglish) social network data. Indian Journal of Science and Technology 13(40): 4216-4224. https://doi.org/10.17485/IJST/v13i40.1451

* **Corresponding author**.

Tel: +91- 8882879289
gargsneha99@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: Evaluating the sentiments of tweets, blogs, comments and posts have become a crucial part of many applications. Sentiment analysis of social network data is very helpful for decision-making in application areas like movie reviews, product feedback and impact of the speech of a politician etc. The users often comment in their native languages or in slang languages or more often they use abbreviations and do not even stick to grammatical rules of the language. The bilingual and multilingual community often mixes two or more than two languages in their comments. Unavailability of annotated code-mixed data for native language adds to the difficulty in performing sentiment analysis. **Objectives** The motive of this article is to present the process of creating annotated corpus for code mixed social media text in Hindi, English and Hinglish which is collected from twitter. Ambiguous meaning words and inconsistent spellings of both the languages have also been included in the study to provide wide spread canvas. **Method:** This study will provide significant elements that should be considered while developing the annotated corpus of Hindi, Hinglish & English dataset. The annotation is calculated on the basis of polarity of words in three categories as positive, negative and neutral. There are words which have mixed feeling i.e. these words have positive as well as negative sentiments. To consider these words, inner agreement among the polarities has been considered. The words used for sarcasm or slangs have also been taken into account. The study has included ambiguous meaning and inconsistent spelling words of both languages as well. **Findings:** The proposed work provides a standard annotated corpus for code switched social media text in Hindi-English (Hinglish). The process of developing the corpus and calculating the polarity has been shown. It is found that if one considers the code-mixed text, the accuracy can be enhanced. **Application:** The proposed corpus can be utilized in the area of market analysis, customer behavior, polling analysis, brand monitoring, etc. The corpus serves as dataset which can further be extended according to the problem definition.

Keywords: Machine learning; sentiment analysis; data preprocessing; data cleaning; code-switch; linguistic-switching; multilingual

1 Introduction

With the up-growing usage of digital world where users can access the data on large screen terminals to small screen terminals [1], [2]. The use of micro-blogging sites like Twitter, Instagram etc. are becoming more and more popular among all types of users. The users can freely share their thoughts and their views on real-world activities like any popular political event such as the elections, corruptions or about a celebrity performing for a charitable event, supporting a political party or anything which may attract their attention. In other words, the real-world communities are being replicated by these social networking communities. Similarly, the real world events are being replicated by these social network sites events. The increasing use of social networking sites has flourished the following important aspects of research-

1. With the support of WWW to multilingual expressions, more often the users use their native language to express themselves. On many occasions, they use code-switching or linguistic switching [3]. The concept of using multiple languages in a single sentence is very popular among the multilingual community [4] and to perform the sentiment analysis of these statements for more accuracy, the sentence is translated into one language and then polarity is calculated.
2. Secondly, the rapid creations of various virtual communities on social network sites, real-world events are being discussed among these virtual communities. The reflection of these real-world events in virtual events are also dragging the attention of research communities [5].
3. The same real-world events may appear with different names or with the same name among the virtual communities and analyzing these parallel events or sub-events may add great values to the accuracy of results.

Discovering the future aspects of all the events, prediction and taking the precautionary measures, has always fascinated the researchers. Predictive analysis is an area where users can predict the future based on the analysis made on historical as well as on current data. The increasing amount of online data has provided an opportunity to analyze the social and behavioral aspects of real-world communities through their behavior, posts, tweets, etc on social networking and microblogging sites. The new avenues have created for research in the area of predictive analysis.

In the competitive world where sustainability and progress have become so crucial that organizations are going extra mile to analyze the data more preciously and make decisions with better efficiency [6]. In earlier stages, however, the researchers have worked on sentiment analysis of movie reviews dataset [7–9]. The authors [10,11] have found that with the increasing attention on microblogging sites as compared to the traditional systems, social media has become popular among the users, organizations and researchers depending on their needs. The social blogging sites comprise communities like groups of friends, relatives, colleagues or people sharing common interests which replicate the real world's communities [5].

Sentiment analysis of multilingual text, blogs, reviews and posts in English, Dutch and French language has been done [3]. People [12] have performed sentiment analysis on German language by using SentiWordNet score. The sentiment score system considering Spanish and English languages for tweets was presented in SemEval 2013 and TASS 2013 [13]. Ensemble learning technique has been used for English & Vietaname language [14] and is reported with increased precision in comparison to the monolingual dataset. The researchers have considered formal as well as informal language and the local slang language. For example, mixture of English and Tamil (Tamilish) is analyzed by linguistic mixing. The techniques and challenges faced in formal and informal combination of code-mixing and linguistic switching for various languages such as Arabic, German, Romanian, French, English, Singaporean etc. are compiled by [15].

According to the research, the involvement of Indian users has increased greatly these days and so is the presence of Indian languages [16]. Now, more users express their feelings in Hindi, Bengali, Punjabi, Telugu, Tamil and Malayalam etc. [17,18] and sentiment analysis is also done for these languages. More research work is going on in the field of bilingual sentiment analysis and the researchers are analyzing of every possible pair of languages. Hindi and Marathi has been analyzed with 72% accuracy [19]. The authors considered blogs and Sunday Travel editorial with Hindi corpus of around 11k words and Marathi that of 12k words. People have performed sentiment analysis by considering spelling mistakes during writing in Hindi and Telugu [20]. The authors had created corpus of mis-spell words by considering most frequent words for both languages; 15,000 for Hindi and 10,00 for Telgu and expand this corpus to approx 90,000 words.

Important techniques for sentiment analysis of Hinglish text has been aptly discussed along with their limitations [21]. The process of sentiment analysis has been compartmentalized in three phases. NLP comprising tokenization, stop word removal, punctuation removal, stemming and lemmatization is defined as first phase. Feature extraction has been put for the second phase and in third phase, classifier is chosen with Naïve Bayes being prominent choice. Emotion detection has been performed in three categories as happy, sad and anger on some 12000 code-mixed sentences of Hindi-English which were mainly collected from social media (FB & Twitter) [22]. With using CNN-BiLSTM model of sentiment analysis, accuracy of 83.21% has been achieved. The authors applied two systems for Hinglish code-mixed tweets for around 15,000 instances; a supervised classification incorporating cross-lingual embeddings for English transliterated Hindi data and a transfer learning approach trained on English sentiment data and cross-lingual embedding and applied to code-mixed data [23]. The system obtained f1 score of 0.556.

The increment of 4-5% in accuracy was shown when Hindi-English code mix was used for misspelling, morpheme and application of Sub-word LSTM architecture instead of traditional approaches [24]. The linguistic code-switching is considered by taking into account the grammatical transition among languages and the recursive neural tensor network (RNTN) [16]. Further research has been done for Hindi-English code mixed data to identify the polarity of a speech as normal or hate [25]. The authors have proposed two-way method for sentiment analysis of Hinglish Data: one by using Lexicon based approach with an accuracy of 86% and secondly by using machine learning techniques with an accuracy of 76% [26].

The social media (SM) data is increasing at such a rapid pace that one can always enhance the accuracy of decision making and can take better preemptive or precautionary measures in advance to avoid the unwanted conditions[27,28].

People have worked on sentiment analysis in English language. Some of the important work has been done in Hindi and Hindi-Hinglish combination as well. There are many such combinations used in other pair of languages. However, combination of more than two languages has not been worked upon extensively. Coupling of Hindi-Hinglish-English languages have been missing from sentiment analysis due to the difficulties encountered for various reasons for example ambiguous words, inconsistent spelling, part of speech and bag of words etc.

Present paper is an attempt to highlight dataset creation for sentiment words of Hindi-Hinglish text, three level annotation schemes for positive, negative and neutral words. This paper also focuses on conflicting words through inner annotation agreement and on the properties and statistics of dataset.

2 Material and Methods

Most of the research work is targeted in the direction of event detection and performing the sentiment analysis of events. The researchers have considered the text written in one language as English, Hindi, Chinese, Romanian, French and German etc. and bilingual like English & Chinese, English & Romanian etc along with different multilingual communities. However Indian researchers have focused on unilingual and bilingual communities. People have worked on code mixing and linguistic switching in Hindi-English bilingual combination. The presence of ambiguous words and inconsistent spellings has rendered the analysis rather incomplete and less effective. Words written in roman can have many different appearances. For example words such as mei, mein, main, etc. refer to one word 'मैं' in devanagri script. There are many words which are present in English and in Hindi with same phonetics but different meaning like bus, suffer, Holi etc. In the proposed strategy, the aim to provide more accurate results for social feeds in Hindi-English (Hinglish) and bridge the mentioned research gaps. To cover gaps, a large amount of data set is required. For English language, python is well-equipped with rich library of sentiment words and natural language processing functions. On the other hand, dealing with Hindi and Hinglish, utf encode and decode functions are not very helpful in translating each and every word that is used by users in SM. In this study, Hindi and Hinglish language processing and word base generation techniques have been proposed. [Figure 1] depicts the complete process of generating an annotated corpus of tweets written in Hindi & Hinglish language in CSV format.

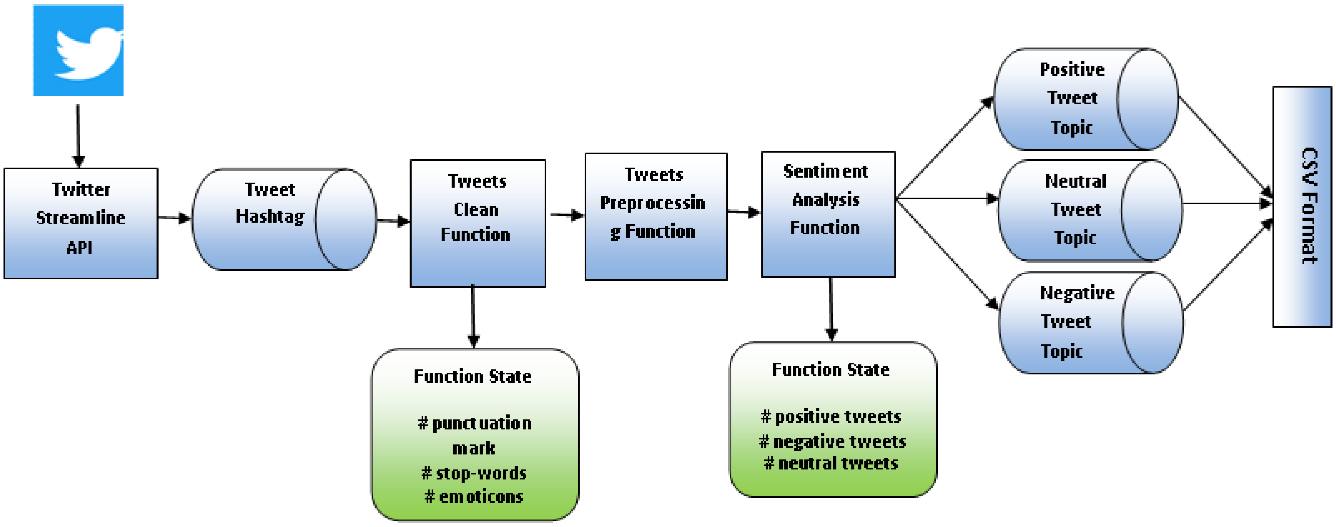


Fig 1. Proposed method for annotated corpus of Hindi & Hinglish text

3 Data Collection

3.1 Source

With the advancement of technologies and increasing involvement of users with digital media, a person with a certain type of thinking and likings will have friends with similar thinking in his or her personal life and will be grouped with people of the same set on SM sphere. Through the real-life events can be reflected in virtual space and people use monolingual, bilingual, and multilingual expressions to express themselves. To limit the area of interest, this research work has been carried out for tweets which are analyzed on parameters like user mention, hashtags, user information and location information and user text. The work done for fetching the required data is elaborated in the preprocessing section.

3.2 Preprocessing

To fetch the required details from Twitter various APIs are used in python version no: 3.7 32-bits. Libraries such as NLTK, TEXT BLOB, CSV, TWEEPY, PANDA, STRING and OS are used for the purpose. With the existing libraries, many challenges were faced. Unicode, used for encoding and decoding tweets, often does not encode a text properly and may fetch it as noise, thereby missing some important emotional content. NLTK has rich dictionary for English language text but provides no support for Hindi or other such languages.

4 Data set creation

4.1. Data selection

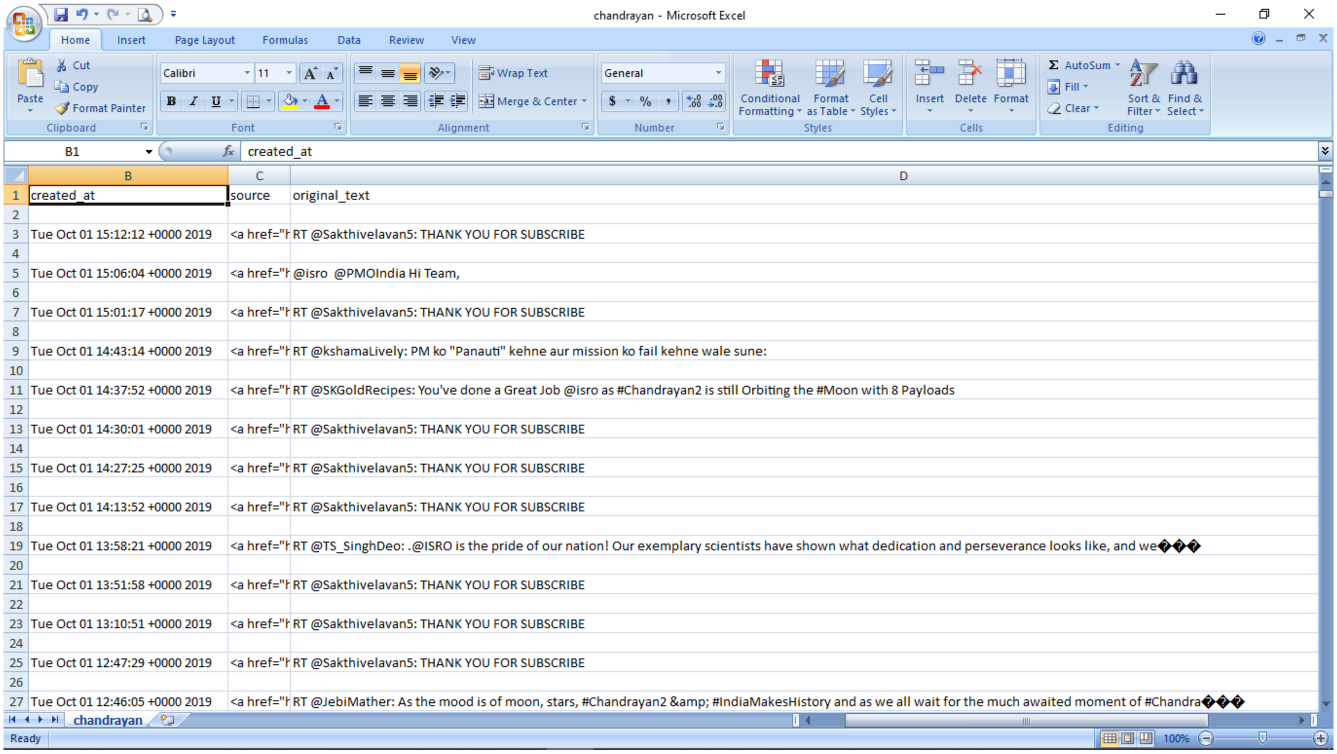
As the social networking sites, online shopping, blogging etc are used by a vast range of users. These users react to the events which occur in their surroundings according to their interest and routine. Various combinations of languages are used in bilingual or multilingual modes. Not only the variation, users also post with linguistic code switching and code mixing such as Hinglish, Tamilish etc. The data selection has been fixed for a set of tasks that are closely related to the goal of proposed research work. In the proposed work, the data selection task comprises selection of language-specific tweets i.e. English, Hindi & Hinglish language and selection of events.

4.2. Data fetching

Once the domain of the problem has been selected, the data is collected using API access credentials for social networking sites. Online available data sets are also used. Later on, the data can be stored in the required format[29]. To make the data in the required format, following steps are followed-

1. Data Format- Changing the data type and syntactic data structure of attributes and values (if needed)
2. Data Construction- Identifying, collecting and creating new attributes (if required)
3. Data integration- Identifying, correcting of integration conflicts, establishing data relations, and selecting the data integration schema.

These tasks are restricted to be language-specific i.e. Hindi, English & Hinglish language and social site-specific i.e. tweets and could well be used for other applications. In [Figure 2], a sample data set of around 3, 000 tweets have been produced which is fetched using Twitter APIs with keyword “Chandrayan” “Chandrayan2”, “IndiaFails”, “Pulwama”, ”TeleMedicine” etc. Using these keywords, tweets with hashtag or user mention are stored in the table.



B1 f_x created_at

	B	C	D
1	created_at	source	original_text
2			
3	Tue Oct 01 15:12:12 +0000 2019	<a href="h	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE
4			
5	Tue Oct 01 15:06:04 +0000 2019	<a href="h	@isro @PMOIndia Hi Team,
6			
7	Tue Oct 01 15:01:17 +0000 2019	<a href="h	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE
8			
9	Tue Oct 01 14:43:14 +0000 2019	<a href="h	RT @kshamaLively: PM ko "Panauti" kehne aur mission ko fail kehne wale sune:
10			
11	Tue Oct 01 14:37:52 +0000 2019	<a href="h	RT @SKGoldRecipes: You've done a Great Job @isro as #Chandrayan2 is still Orbiting the #Moon with 8 Payloads
12			
13	Tue Oct 01 14:30:01 +0000 2019	<a href="h	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE
14			
15	Tue Oct 01 14:27:25 +0000 2019	<a href="h	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE
16			
17	Tue Oct 01 14:13:52 +0000 2019	<a href="h	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE
18			
19	Tue Oct 01 13:58:21 +0000 2019	<a href="h	RT @TS_SinghDeo: .@ISRO is the pride of our nation! Our exemplary scientists have shown what dedication and perseverance looks like, and we���
20			
21	Tue Oct 01 13:51:58 +0000 2019	<a href="h	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE
22			
23	Tue Oct 01 13:10:51 +0000 2019	<a href="h	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE
24			
25	Tue Oct 01 12:47:29 +0000 2019	<a href="h	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE
26			
27	Tue Oct 01 12:46:05 +0000 2019	<a href="h	RT @JebiMather: As the mood is of moon, stars, #Chandrayan2 &amp; #IndiaMakesHistory and as we all wait for the much awaited moment of #Chandra���

Fig 2. Original text data fetched from twitter

4.3 Data cleaning

The process of data cleaning comprises following steps for the languages-

Noise Removal- In the very first of data cleaning, the cleaning of text is performed for the normalized representation of text. The Unicode, which is not supported by Python, has been replaced by special symbols. Similarly, email-id, phone numbers and URLs are replaced by special tokens. For cleaning English text, Python has an in-built library “clean-text”. As for the text in Hindi & Hinglish, a new library is developed.

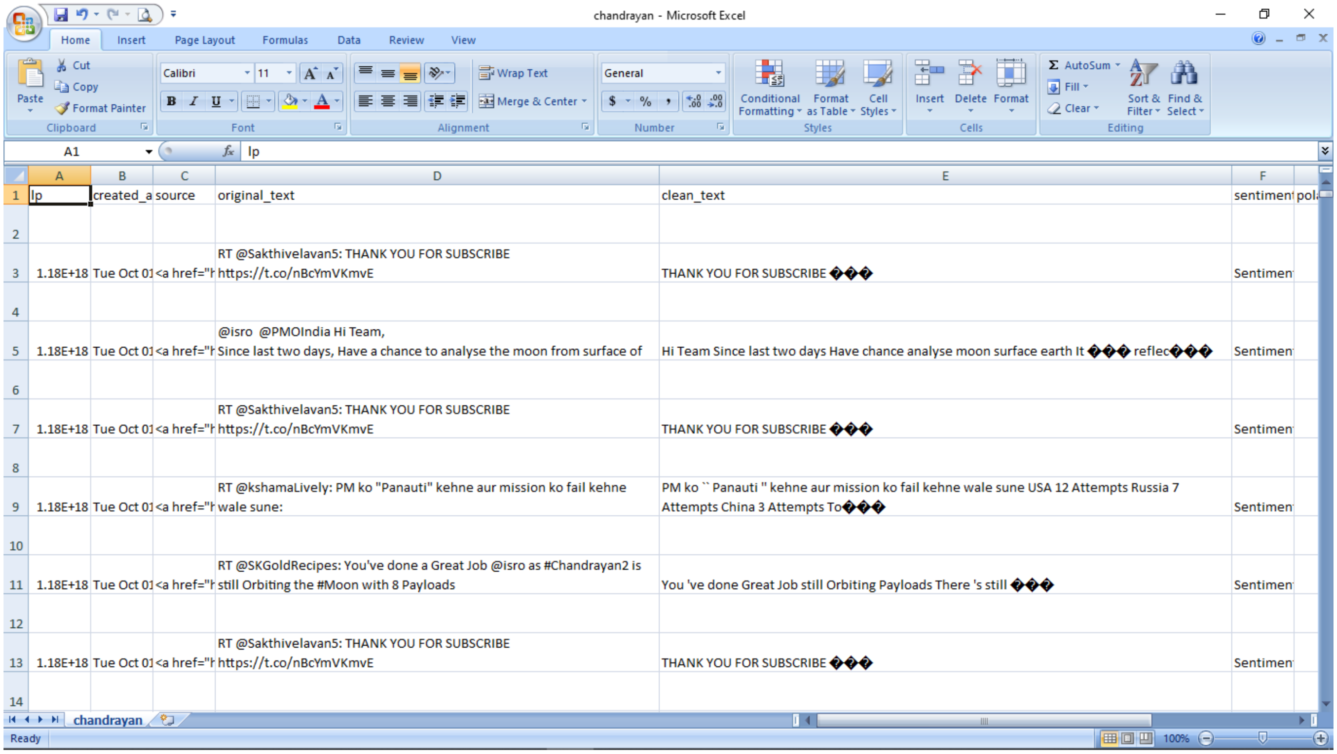
Stopword Removal- In the step of cleaning, words which are not important to the sentiment analysis, are listed and removed from the data. NLTK is provided in python library as in-built function for stopword removal. NLTK supports approximately 217 stop words for English language. But for Hindi and Hinglish, there are no stop words supports by NLTK so a list of 240 stop words has been developed. There are words which are left out or removed in English dictionary because they do not bear significant meaning. On the other hand, In Hindi and Hinglish dictionary, due to the fact that these words carry impact on sentiment analysis, we have included such words in analysis. Words such as “या”, “परंतु” are kept after stopword removal.

Punctuation Sign- Punctuation signs are also insignificant for sentiment analysis. Hence, such words (“.”, “!”, “?”) are removed.

Emoticons consideration- The annotation list for emoticons has been employed from SentiStrength and the content of tweets is matched against the annotated list. The matched emoticon has been replaced by their pre-specified polarity.

Affected word polarity- The words are mapped on the scale from -1 to +1 according to their polarity where -1 signifies highly negative and +1 is for highly positive and 0 refers to neutral sentiment.

Once the preprocessing for the English, Hindi, and Hinglish words is complete, the text is now forwarded for feature extraction phase. In [Figure 3], the sample data set is presented after removing the noise, Emoticons, stop words and punctuation. The erroneous code which is shown in column ‘clean_text’ represents the code that is not treated by Python utf-8 code.



A1 = lp

	A	B	C	D	E	F	
1	lp	created_a	source	original_text	clean_text	sentiment	pol
2							
3	1.18E+18	Tue Oct 01	<a href="h	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE https://t.co/nBcYmVKmvE	THANK YOU FOR SUBSCRIBE ���	Sentimen	
4							
5	1.18E+18	Tue Oct 01	<a href="h	@isro @PMOIndia Hi Team, Since last two days, Have a chance to analyse the moon from surface of	Hi Team Since last two days Have chance analyse moon surface earth It ��� reflec���	Sentimen	
6							
7	1.18E+18	Tue Oct 01	<a href="h	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE https://t.co/nBcYmVKmvE	THANK YOU FOR SUBSCRIBE ���	Sentimen	
8							
9	1.18E+18	Tue Oct 01	<a href="h	RT @kshamaLively: PM ko "Panauti" kehne aur mission ko fail kehne wale sune:	PM ko `` Panauti '' kehne aur mission ko fail kehne wale sune USA 12 Attempts Russia 7 Attempts China 3 Attempts To���	Sentimen	
10							
11	1.18E+18	Tue Oct 01	<a href="h	RT @SKGoldRecipes: You've done a Great Job @isro as #Chandrayan2 is still Orbiting the #Moon with 8 Payloads	You 've done Great Job still Orbiting Payloads There 's still ���	Sentimen	
12							
13	1.18E+18	Tue Oct 01	<a href="h	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE https://t.co/nBcYmVKmvE	THANK YOU FOR SUBSCRIBE ���	Sentimen	
14							

Fig 3. Cleaned text after removal of stop words

4.4 Feature extraction

As per the identified gaps above, the words are mapped based on the following features

1. Ambiguous words- Hinglish and English have certain ambiguous words like “bus”, “holy” etc. Such ambiguous words can be handled by context handling technique. Taking example of the word “evil weather”, here, the word ‘evil’ has a negative meaning but in combination with word ‘weather’, phrase keeps positive meaning [30].

2. Inconsistent feature check- The inconsistent feature has two important aspects, inconsistent spelling of words and inconsistent meaning words. Here inconsistent spelt words like “mei”, “mein” etc. are handled by lemmatization method whereas inconsistent meaning words like “date”, “lying” , “kal”, “idhar”, “udhar” are taken care of by context handling.
3. Negative Polarity Shift Feature- The group of words or phrases representing the negative polarity are mapped to match the words with required sentiment polarity for example- ‘do not like’ are mapped to dislike, ‘nahi acha’ are mapped to “bura”,

4.5 Annotation

A scheme is proposed to label the words of Hindi & Hinglish language to represent the sentiment polarity and to scale them accordingly. Certain words represent facts or state and do not have positive or negative opinion; hence such words are labeled as neutral words. For assigning the values to English and Hindi language words, the WorldNet dictionaries of both the languages have been referred and the Hinglish words are mapped accordingly.

1. **Positive words**- These words represent the positive expression, emotion, feeling or opinion to support, motivate, admire, inspire or praise any person, object, or situation. Such words of positive impact are assigned values from “0” to “1”.
   Example- ‘सौभाग्यशाली, खुशकिस्मत’, ‘saubhagyashaaly, khushkismat’ is given the highest positive score..
2. **Negative words**- Some words have negative expression, emotions, feelings, or opinion to show disagreement, criticism, failure, sadness, negative attitude. Such words are called negative words and scaled from “-1” to “0”.
   Example- ‘अपमानजनक’, ‘apmanjanak’ are given the highest negative score.
3. **Neutral words**- These words have neither positive nor negative sentiment. These words are not counted for evaluating any sentence in analysis. Such words are placed with value “0”.
   Example- words ‘फिलहाल, filhal’ have been assigned value ‘0’ for their neutrality toward sentiment.

4.6 Inner annotation agreement

It is used to take into account the degree of disagreement among the predicted annotated classes. For instance, if the annotator vacillates between positive and a negative sentiment class then it difficult to decide for individual cases. To ease the situation, averaging of these sentiments has been taken into account.

Example- word ‘जैसे_तैसे’, ‘jaise_taise’ has positive score of ‘0.125’ as well as negative score of ‘-0.25’ then case average score ‘-0.0625’ has been considered for the word.

5 Dataset evaluation

To annotate the tweets, around seven thousand words of Hindi and Hinglish dataset have been considered and mapped onto positive, negative and neutral annotation scale. Some of the words have been considered with inner annotation agreement due to the conflicting polarity presence. Words with similar meaning have been mapped with the same polarities. For example: words ‘acha’ and ‘badhiya’ in Table 1 have similar meaning and are mapped with the same polarity. To take care of the cases of miss-spelt words or inconsistent spellings, an attempt have been made to include all possible spellings as shown in [Table 1].

Table 1. Hindi & Hinglish dataset containing similar words & 1-gram

Hindi Word set	Hinglish Word set	Polarity
अच्छा,बढ़िया	Achcha, acha, achchha, badiya, badhiya, badia,	0.25
रिक्त,ख़ाली,खाली,शून्य	rikt, riqt, rikat, riqat, khali, shoonya, shunya	0
सोचना,दुखी,सोंचना,अनमनाना	sochna, dukh, sonchna, anmanana	-0.375

Two gram word model has also been included in the analysis to cover more possible combinations of words. Their polarity score is calculated with better accuracy for the words such as “khush” which is used with various combinations like “khush_hua”, “khush_hona”, “khush_lga” to mention a few and same has been shown in [Table 2].

Table 2. Hindi, Hinglish dataset with 2-gram word base

Hindi Word set	Hinglish Word set
उड़ना,उड़ान_भरना	urna, udna, udaan_bharna, uraan_bharna
डूबना,ढलना,अस्त_होना,अस्तगत_होना	doobna, dubna, dhalna, ast_hona, astagat_hona
भुगतान_कृत,अदा,चुकाया_हुआ,चुकता	Bhugtan_krat,bhugtaan_krut,ada,chukaya_hua, chukta, chukata

After mapping the dataset based on similar meaning and 1-gram & 2- gram word base as discussed above, the complete set of approximately 7000 words, arranged in more than 3000 rows, is used for expressions, feelings, opinion etc. As a result of the polarity calculation according to annotation, 385 rows have been considered with positive polarity i.e. with positive impact and 483 were found to be with negative polarity or negative feelings and 2145 rows showed neutral polarity. [Table 3] depicts the complete statistics for dataset evaluated for the Hindi; English & Hinglish words sentiment score.

Table 3. Statistics for data Set

Data Type	#
Total words are taken into consideration	7025
Total rows (mapped according to similar meaning, 2-gram)	3014
Rows Having Positive Polarity	385
Rows Having Negative Polarity	483
Rows Having Neutral Polarity	2145

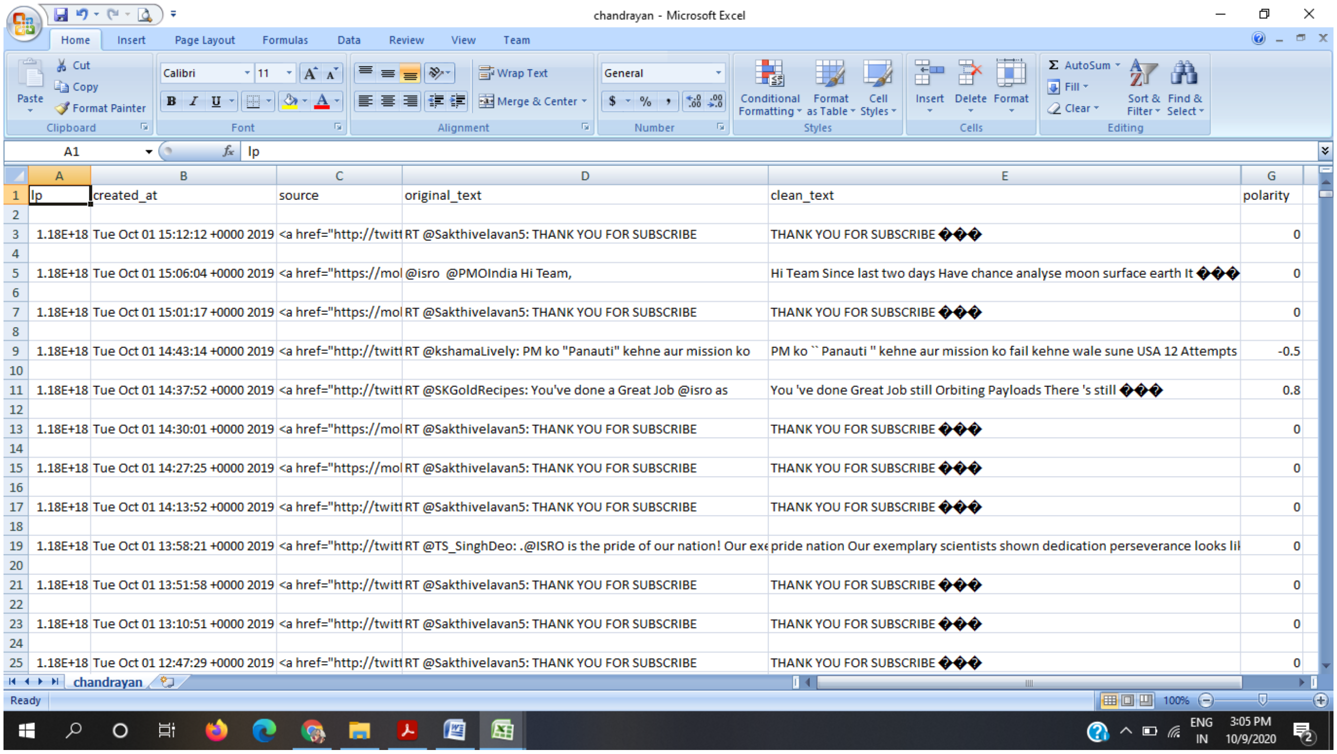
Average of sentiment score for all tokens has been considered to find the sentiment score of a sentence. For a sentence having multiple words with multiple polarities, following method has been proposed and employed.

$$sentence\ score = \frac{Total\ token\ score}{Total\ number\ of\ token\ works}$$

6 Results & Discussion

After applying the suggested annotation scheme, a total of 830 sentences were marked as positive. 455 sentences were reported to have negative polarity. A good number of tweets, 1537 to be extract, were found neutral from given dataset. In [Figure 4], annotated tweets have been shown. The collected 2825 code-mixed tweets have been considered for experiment. Each tweet has been cleaned, tokenized and the preprocessing steps have been applied[31]. The tweets have been annotated according the annotation scheme discussed in the above section.

Comparison of this method with other existing sentiments calculation schemes is somewhat unreasonable because of the fact that most of such schemes are based on English language and lack the variety which has been included in the above presented novel method.



	A	B	C	D	E	G
1	Ip	created_at	source	original_text	clean_text	polarity
2						
3	1.18E+18	Tue Oct 01 15:12:12 +0000 2019	<a href="http://twitt	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE	THANK YOU FOR SUBSCRIBE ���	0
4						
5	1.18E+18	Tue Oct 01 15:06:04 +0000 2019	<a href="https://mol	@isro @PMOIndia Hi Team,	Hi Team Since last two days Have chance analyse moon surface earth It ���	0
6						
7	1.18E+18	Tue Oct 01 15:01:17 +0000 2019	<a href="https://mol	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE	THANK YOU FOR SUBSCRIBE ���	0
8						
9	1.18E+18	Tue Oct 01 14:43:14 +0000 2019	<a href="http://twitt	RT @kshamaLively: PM ko "Panauti" kehne aur mission ko	PM ko `` Panauti '' kehne aur mission ko fail kehne wale sune USA 12 Attempts	-0.5
10						
11	1.18E+18	Tue Oct 01 14:37:52 +0000 2019	<a href="http://twitt	RT @SKGoldRecipes: You've done a Great Job @isro as	You 've done Great Job still Orbiting Payloads There 's still ���	0.8
12						
13	1.18E+18	Tue Oct 01 14:30:01 +0000 2019	<a href="https://mol	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE	THANK YOU FOR SUBSCRIBE ���	0
14						
15	1.18E+18	Tue Oct 01 14:27:25 +0000 2019	<a href="https://mol	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE	THANK YOU FOR SUBSCRIBE ���	0
16						
17	1.18E+18	Tue Oct 01 14:13:52 +0000 2019	<a href="http://twitt	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE	THANK YOU FOR SUBSCRIBE ���	0
18						
19	1.18E+18	Tue Oct 01 13:58:21 +0000 2019	<a href="http://twitt	RT @TS_SinghDeo: .@ISRO is the pride of our nation! Our exe	pride nation Our exemplary scientists shown dedication perseverance looks lik	0
20						
21	1.18E+18	Tue Oct 01 13:51:58 +0000 2019	<a href="http://twitt	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE	THANK YOU FOR SUBSCRIBE ���	0
22						
23	1.18E+18	Tue Oct 01 13:10:51 +0000 2019	<a href="http://twitt	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE	THANK YOU FOR SUBSCRIBE ���	0
24						
25	1.18E+18	Tue Oct 01 12:47:29 +0000 2019	<a href="http://twitt	RT @Sakthivelavan5: THANK YOU FOR SUBSCRIBE	THANK YOU FOR SUBSCRIBE ���	0

Fig 4. Annotated dataset

7 Conclusion

This study presented the method of developing a corpus for Hindi & Hinglish sentiment words. The method used to annotate this corpus into categories i.e. highly positive, positive, negative, highly negative and neutral has also been discussed. For the words which belong to both, negative and positive sentiment, inner agreement calculation has also been shown. This data set handles inconsistent spelling and miss-spelt words are also included in the study. This research may help the researchers to address new and exciting problems in Hindi & Hinglish code mixed research. In this article, we have attempted to present a novel scheme to calculate sentiments of sentences on SM and elsewhere which are in native languages along with English. The method can be extended to the combination of various languages such as Gujrati-English, Punjabi-English Bhojpuri-English and Marwari-English.

References

1) Singhal S, Garg N. Hybrid web-page segmentation and block extraction for small screen terminals. *International Journal of Computer Applications*. 2013;975.
2) Singhal S, Garg N. Web Page Representation Using Backtracking with Multidimensional Database for Small Screen Terminals. In: Innovations in Computational Intelligence 2018. Singapore. Springer. 2018;p. 299–307. Available from: https://doi.org/10.1007/978-981-10-4555-4_21.
3) Boiy E, Moens MF. A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval*. 2009;12:526–558. Available from: https://dx.doi.org/10.1007/s10791-008-9070-z.
4) Chandu K, Loginova E, Gupta V, Genabith JV, Neumann G, Chinnakotla M, et al. Code-mixed question answering challenge: Crowd-sourcing data and techniques. In: and others, editor. Third Workshop on Computational Approaches to Linguistic Code-Switching. ACL. 2019;p. 29–38.
5) Srivastava R, Bhatia MP, Tayal V, Verma JK. Framework for real-world event detection through online social networking sites. In: Data and Communication Networks. Singapore. Springer. 2019;p. 195–203. Available from: https://doi.org/10.1007/978-981-13-2254-9_17.
6) Tripathy A, Agrawal A, Rath SK. Classification of Sentimental Reviews Using Machine Learning Techniques. *Procedia Computer Science*. 2015;57:821–829. Available from: https://dx.doi.org/10.1016/j.procs.2015.07.523.
7) Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. 2004.
8) Pang B, Lee L, Vaithyanathan S. Sentiment classification using machine learning techniques. 2002.
9) Turney PD. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. 2002.
10) Atefeh F, Khreich W. A survey of techniques for event detection in twitter. *Computational Intelligence*. 2015;31(1):132–164. Available from: https://doi.org/10.1111/coin.12017.
11) Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media. In: and others, editor. Proceedings of the 19th international conference on World wide web. 2010;p. 591–600. Available from: https://doi.org/10.1145/1772690.1772751.
12) Denecke K. Using sentiwordnet for multilingual sentiment analysis. *2008 IEEE 24th international conference on data engineering workshop*. 2008;p. 507–512. Available from: https://doi.org/10.1109/ICDEW.2008.4498370.
13) Balahur A, Perea-Ortega JM. Sentiment analysis system adaptation for multilingual processing: The case of tweets. *Information Processing & Management*. 2015;51:547–556. Available from: https://dx.doi.org/10.1016/j.ipm.2014.10.004.
14) Oborník. Endosymbiotic evolution of Algae, secondary heterotrophy and parasitism. *Biomolecules*. 2019;9(7):266–266. Available from: https://dx.doi.org/10.3390/biom9070266.
15) Lo SL, Cambria E, Chiong R, Cornforth D. Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*. 2017;48(4):499–527. Available from: https://dx.doi.org/10.1007/s10462-016-9508-4.
16) Sitaram D, Murthy S, Ray D, Sharma D, Dhar K. Sentiment analysis of mixed language employing Hindi-English code switching. In: and others, editor. 2015 International Conference on Machine Learning and Cybernetics (ICMLC);vol. 1. 2015;p. 271–276. Available from: https://doi.org/10.1109/ICMLC.2015.7340934.
17) Sharma P, Moh TS. Prediction of Indian election using sentiment analysis on Hindi Twitter. In: and others, editor. 2016 IEEE international conference on big data (big data. 2016;p. 1966–1971. doi:https://doi.org/10.1109/BigData.2016.7840818.
18) Kaur A, Gupta V. A survey on sentiment analysis and opinion mining techniques. *Journal of Emerging Technologies in Web Intelligence*. 2013;5(4):367–371. Available from: https://doi.org/10.4304/jetwi.5.4.367-371.
19) Balamurali AR, Joshi A, Bhattacharyya P. Cross-lingual sentiment analysis for Indian languages using linked wordnets. In: and others, editor. Proceedings of COLING. 2012;p. 73–82.
20) Etoori P, Chinnakotla M, Mamidi R. Automatic spelling correction for resource-scarce languages using deep learning. In: and others, editor. Proceedings of ACL 2018, Student Research Workshop. 2018;p. 146–152.
21) Thakur V, Sahu R, Omer S. Current State of Hinglish Text Sentiment Analysis. In: and others, editor. Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020, Available at SSRN. 2020. Available from: http://dx.doi.org/10.2139/ssrn.3614442.
22) Sasidhar TT, B P, P SK. Emotion Detection in Hinglish(Hindi+English) Code-Mixed Social Media Text. *Procedia Computer Science*. 2020;171:1346–1352. Available from: https://dx.doi.org/10.1016/j.procs.2020.04.144.
23) Singh P, Lefever E. Sentiment Analysis for Hinglish Code-mixed Tweets by means of Cross-lingual Word Embeddings. In: and others, editor. Proceedings of the The 4th Workshop on Computational Approaches to Code Switching. 2020;p. 45–51.
24) Joshi A, Prabhu A, Shrivastava M, Varma V. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In: and others, editor. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016;p. 2482–2491.
25) Bohra A, Vijay D, Singh V, Akhtar SS, Shrivastava M. A dataset of Hindi-English code-mixed social media text for hate speech detection. In: and others, editor. Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media. 2018;p. 36–41.
26) Pravalika A, Oza V, Meghana NP, Kamath SS. Domain-specific sentiment analysis approaches for code-mixed social network data. In: and others, editor. 8th international conference on computing, communication and networking technologies (ICCCNT). IEEE. 2017;p. 1–6. Available from: https://doi.org/10.1109/ICCCNT.2017.8204074.
27) Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*. 2015;35(2):137–144. Available from: https://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007.
28) Lesser V, Horling B, Klassner F, Raja A, Wagner T, Zhang SX. BIG: An agent for resource-bounded information gathering and decision making. *Artificial Intelligence*. 2000;118(1-2):197–244. Available from: https://dx.doi.org/10.1016/s0004-3702(00)00005-9.
29) Chen P, Zhang CL, Y C. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf Sci*. 2014;275:314–347. Available from: http://www.sciencedirect.com/science/article/pii/S0020025514000346.
30) Sinha RMK, Thakur A. Machine Translation Of Bi-Lingual Hindi-English (Hinglish) text. In: 10th Machine Translation summit (MT Summit X). Phuket, Thailand. 2005;p. 149–156.
31) Garg N, Sharma K. Machine Learning in Text Analysis. In: and others, editor. Handbook of Research on Emerging Trends and Applications of Machine Learning. IGI Global. ;p. 383–402.