© 2019 IJSRSET | Volume 6 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET196516

An Approach Towards Bilingual Sentiment Analysis

Farhin Falak
ECS Department, BSSITM, Lucknow, Uttar Parades, India

ABSTRACT

In the existing An Approach towards Bilingual Sentiment Analysis, Sentiment analysis is an emerging area of research. Over the past few decades increase in user generated content provides important aspect for researchers, companies and even government organization.
With the increase in online user generated content sentiment becomes a big challenge and brings promising opportunity in field of sentiment analysis.
User content varies tremendously since a user may.

- ✓ Use single language
- ✓ Use mixed language
- ✓ May use mixed languages (Hindi + English)
- ✓ May use abbreviations instead of full words

Keywords : Bilingual Sentiment Analysis, polarity, Addressing Negation, POS, Subjectivity

I. INTRODUCTION

This Thesis will analyze the exactness of the present strategy utilized for checking estimation of statements written Hinglish (Hindi + English) language. Sentiment analysis is the procedure of computationally recognizing and arranging opinion communicated in a piece of text, particularly so as to decide if the writer's attitude towards a specific point, item, and so on., is sure, negative, or impartial.

The speculation is that by utilizing sentiment analysis framework upgraded for the investigation of Hinglish language, a more precise estimation characterization will be accomplished than if Hinglish statements were preprocessed and grouped utilizing a similar procedure as English languages.

Sentiments in Hinglish are very much important for organizations, government officials, and different associations since it gives a medium to a great number of clients to express their opinion for assessments and proclamations.

Pang and Lee published a paper in 2008, Opinion Mining and Sentiment Analysis [3], on opinion mining and sentiment analysis where they talk about procedures and ways to deal with straightforwardly empower supposition arranged data looking for frameworks. In spite of the fact that the applications and datasets talked about are identified with audits, proposals, business and government insight, and backing of legislators or other open figures and these sorts of writings are all the more formally composed, numerous looks into received the procedures examined in the overview paper to tweets.

II. BACKGROUND AND RELATED WORK

In 2008, Pang and Lee published paper [3] on"opinion mining and sentiment analysis" where techniques and approaches were discussed that enable opinion-focused information-searching systems. Though the applications discussed are focused on reviews and recommendations for business and government, these reviews are more formally written than the text discussed in this paper. Pang and Lee's paper took it as the main challenge. They also addressed the following points

(1) Extracting documents focused on relevant topic
(2) Extracting sentiment of the document
(3) Classifying the polarity of the sentiment identified.

They also focused on the key concepts in classification which is also relevant in informally written small text, For example classifying documents based on sentiment polarity, addressing negation, POS, subjectivity, and topic relevance.

III. SENTIMENT ANALYSIS

Sentiment Analysis, sometime also refer as opinion mining, is a part of NLP and text analysis which helps in finding out polarity of the user's sentiment from a written piece of text.

Previously sentiment analysis was focused on formally written text such as movie reviews etc. Since the launch of Twitter in 2006, the number of users has increased exceptionally. In 2015 around 500 million users were registered in twitter, Many times while twitting user express their feelings toward any product, person, or idea. Sentiment Analysis of such data can provide insight over many applications.

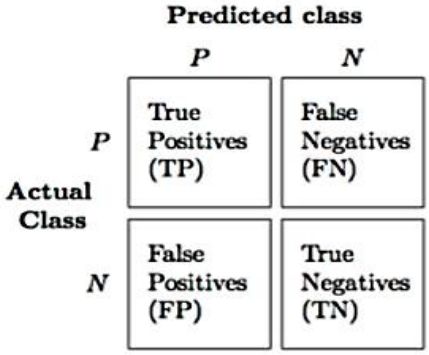

Figure 1. Derive formulas from the confusion matrix

Two programming packages used here for sentiment analysis are Python's NLTK and R's 12 sentiment package. Python's NLTK provides easy interfaces for lexical resources as well as libraries for classification, tokenization, stemming, etc. Also, R's sentiment analysis has become popular now a day because of its integration with data visualization tools.

HINGLISH

Two programming packages used here for sentiment analysis are Python's NLTK and R's 12 sentiment package. Python's NLTK provides easy interfaces for lexical resources as well as libraries for classification, tokenization, stemming, etc. Also, R's sentiment analysis has become popular now a day because of its integration with data visualization tools.

Table 1 : Examples of Hinglish Tweets

1	ENGRAVER SETHI @EngraverSethi @lovudsharma	
2	Ramandeep Kaur ?@ramandkaur89 17	#Samsung galaxy note 5 India ka sabse mehnga aur bakwaas android phone hai.
3	@NeerajDhake	Ab pta chala

	d	@SamsungMobileINsamsung ke sare mobile hi ghatiya hein lelo.
4	Ramandeep Kaur ?@ramandkaur8917	Mujhe yeh samjh main nahi aata ke #Samsung ke sabhi mobiles mein ek hi jaise features kyu hote hain. Rs.45000 and Rs. 25000 are same
5	Navneet Bhullar ?@NavyBhullar	# I think HTC and samsung phones are quite similar bcoz maine dono use kiye hain..mujhe dono phones pasand hain
6	@lovudsharma @EngraverSethi	Naye #Samsung smartphones and #iphone apni lokpriyata sirf removable battery naa hone ki wajha se kho rahe hai.
7	@ap_pune	acha hai na ...means some one doing your marketing ... Its like I don't like apple phone doesn't mean its bad
8	@NaveenxAsad	i hate window phone tbh. Apple acha hai 5s but not like android
9	@OyeSaaann @Mahnoor_Agha	*die hard apple boy* "android sucks! IOS for life!"

IV. EXPERIMENT

The data set used here consists of Hinglish tweets and will be analyzed based on following cases using

Python's NLTK package:
1Case 1: Language: English
2Case 2: Language: Hindi
3Case 3: Language: Hinglish

The F-score of classification will be measured and also considering confusion matrix for every case. We hypothesizes that if Hinglish language Sentiment Analysis system is built, it will gain a more accurate classification, as compared to the traditional system where the corpora, dictionary, and other NLP resources.

PYTHON NLTK Package

Pythons NLTK are a package containing libraries and programs used for statistical natural language processing. Here we assumed that NLTK is the most commonly used library for natural language processing. NLTK has most of the functionality required to for a simple sentiment analysis system such as classification, tokenization, etc. NLTK package contain many of its modules having same functionality for non-English languages as well. This package also provides over 50 corpora and lexical resources. In the preprocessing we used the NLTK stop words for English language along with Stemmer provided by NLTK. NLTK SentiWordNet 14 and POS-Tagger were used in building the feature set.

DATASETS

The Hinglish Tweet algorithm checks that the tweet contains English words and Hindi words.
If not (english_dictionary.check (word) and hindi_dictionary.check (word)):

```
If (english_dictionary.check (word)):
        English_word=True
        #hinglish_tweets_file.write ("English Word!" + word +'\n')
        If(hindi_dictionary.check(word)):
                hindi_word=True
                #hinglish_tweets_file.write ("Hindi Word! " + word +'\n')
```

Figure 7. Hinglish Tweet Algorithm snippet

V. RESULT

The Hinglish Sentiment Analysis needed to more attention in the preprocessing so it can adapt and extract features. By close observation we can say that many Hinglish words are constructed and with the help of NLTK Snowball Stemmer, and the Bing Translator, it was observed that translating all words to English was important so that they can be easily mapped to a synsets in SentiWordNet. However, if the word to be translated into English language is not found in the dictionary then it will be unchanged. Many Hinglish words are constructed from English verbs with the Hindi suffix appended to it. So, in this case after translating words to English, and stemming with the Snowball Stemmer, we also apply the Hindi Stemmer available in NLTK.

Results Statistics for Hinglish Sentiment Analysis of Hinglish Tweets

Table 2. Comparison of results statistics for all three cases

	Accuracy	F-Score	Sensitivity	Specificity
Case 1: English	70.54%	57.91%	92.07%	24.68%
Case 2: Hindi	72.20%	63.03%	89.63%	35.06%
Case 3: Hinglish	73.03%	63.32%	91.46%	33.77%

VI. CONCLUSION

This work provided insight into the challenges, resources available, and possible approaches for building a bilingual sentiment analysis tweets. It was observed that around 5% of all tweets in the India were written using English Dictionary and Hindi Dictionary. While 5% is a very small figure but since there are more than millions of tweets every day, this lead to a significant data.

The preprocessing was done according to the language we depend upon for sentiment analysis system. After experimenting on the Hinglish tweets, it was observed that the Hinglish sentiment analysis gave us the highest accuracy of 73.03%, and the best F-Score of 63.32%. It was also observed that in case of misclassification for the Hinglish, this was caused by erroneous Hindi stemming on English words. Further improvements would invoke further preprocessing to better identification of Hinglish words so that the Hindi stemmer would only be applied to these words.

VII. REFERENCES

[1]. Pew Research. “Statistical Portrait of Hispanics in the United States.” Internet: http://www.pewhispanic.org/2016/04/19/statistical-portrait-of-hispanics-in-the-united-states/, April 2016 February 2017].

[2]. Chris Hutchins. “Companies Engaging Hispanics Win Big in the U.S. – and Beyond.”Internet: https://www.motionpoint.com/blog/companies-engaging-hispanics-win-big-in-the-u.s.-and-beyond/, May 2015February 2017].

[3]. B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1-135, 2008.

[4]. Ameeta Asiaee T., Mariano Tepper, Arindam Banerjee, and Guillermo Sapiro. 2012. If you are happy and you know it... tweet. In

Proceedings of the 21 st ACM International Conference on Information and Knowledge Management, CIKM '12, pages 1602-1606, New York, NY, USA. ACM.

[5]. Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, pages 1-6.

[6]. Brand Watch. "Understanding Sentiment Analysis." Internet: https://www.brandwatch.com/2015/01/understanding-sentiment-analysis/, 2015 2017].

[7]. Semantria LLC. Semantria Out-of-the-Box Reliability. (2015). https://semantria.com/case-studies

[8]. Julian Brooke, Milan Tofiloski, and Maite Taboada. "Cross-Linguistic Sentiment Analysis: From English to Spanish." Simon Fraser University.

[9]. Oscar Araque, Ignacio Corcuera, Constantino Roman, Carlos Iglesias, and J. Fernando Sanchez-Rada. "Aspect based Sentiment Analysis of Spanish Tweets." TASS 2015, September 2015, pp 29-34.

[10]. Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. "Learning Sentiment Lexicons in Spanish." Internet: http://lit.eecs.umich.edu/~banea/publications/perez.lrec.2012.slides.pdf, 2012.

[11]. Pew Research. "With fewer new arrivals, Census lowers Hispanic population projections." Internet: http://www.pewresearch.org/fact-tank/2014/12/16/with-fewer-new-arrivals-census-lowers-hispanic-population-projections-2/, December 2014 February 2017].

Cite this article as :

Farhin Falak , "An Approach Towards Bilingual Sentiment Analysis", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 6 Issue 5, pp. 120-124, September-October 2019. Available at doi : https://doi.org/10.32628/IJSRSET196516
Journal URL : http://ijsrset.com/IJSRSET196516