Sentiment Analysis of English-Punjabi Code-Mixed Social Media Content

Mukhtiar Singh[1], Vishal Goyal[2], Sahil Raj[3]
[1,2]Department of Computer Science, Punjabi University, Patiala
[3]School of Management and Studies, Punjabi University, Patiala
{mukhtiarrai73,vishal.pup,dr.sahilraj47}@gmail.com

Abstract

Sentiment analysis is a field of study for analyzing people's emotions, such as Nice, Happy, ਦੁਖੀ (sad), changa (Good), etc. towards the entities and attributes expressed in written text. It noticed that, on microblogging websites (Facebook, YouTube, Twitter), most people used more than one language to express their emotions. The change of one language to another language within the same written text is called code-mixing. In this research, we gathered the English-Punjabi code-mixed corpus from micro-blogging websites. We have performed language identification of code-mix text, which includes Phonetic Typing, Abbreviation, Wordplay, Intentionally misspelled words and Slang words. Then we performed tokenization of English and Punjabi language words consisting of different spellings. Then we performed sentiment analysis based on the above text based on the lexicon approach. The dictionary created for English Punjabi code mixed consists of opinionated words. The opinionated words are then categorized into three categories i.e. positive words list, negative words list, and neutral words list. The rest of the words are being stored in an unsorted word list. By using the N-gram approach, a statistical technique is applied at sentence level sentiment polarity of the English-Punjabi code-mixed dataset. Our results show an accuracy of 83% with an F-1 measure of 77%.

1 Introduction

In the last decade, the social media platform has been become the medium of communication such as Facebook, Twitter, LinkedIn, etc. (Yang, Chao et al., 2013; Fazil, Mohd et al., 2018). On social media platform, everybody has a short time and the information to be analyzed is huge. Sentiment analysis helps us to whether the message or sentence follows positive or negative. Sentiment analysis is also known as opinion mining or opinion analysis. By using, the web forums there are so many sources to express their views to track and analyze opinions and attitudes about and product. In India, there are 22 official languages, and many more regions languages used for communication (W. Medhat et al., 2014).

There are lots of social media communication, which people use more than one languages to convey their opinion or sentiments (Kalpana et al., 2014; Sharma, S et al., 2015). So, necessary to analyze the data to find appropriate sentiments. N-grams are one of the most commonly used features (G. Rodrigues Barbosa et al., 2012; Kaur, A., & Gupta, V. 2014). We used the n-gram approach up to fivegram and found that the results of fivegram are similar to trigram approach for English-Punjabi code mixed text. The type of ngram also depends on the type of domain used as some domains are more popular in phrases to express the sentiment. Accordingly, our tool gives the power to the users to choose one of two approaches: trigrams and fivegram.

2 Methodology

The main target of current research is sentiment analysis of English-Punjabi code mixed language at sentence level. The foremost task for developing the system is collection of Social Media Code-Mixed text using API twitter threads for **Twitter**, selecting some prolific users comments for **Facebook** as data and some student community prolific users chat for **Whatsapp** followed by cleaning of extracted data.

The dataset used in the current research consists of 10 Lakh sentences (tafter preprocessing) which have been tagged as en

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 24–25
Patna, India, December 18 - 21, 2020.

(English), pb (Punjabi), univ (Universal) and both (mixing of two languages inside a word), The features used are contextual features, capitalization features, special character features, character N Gram features and lexicon features.

In social media text people use creativity in spellings rather than traditional words. The deviation of text can be categorized as acronyms, slangs, misspellings, use of phonetic spellings etc. Contractions like hasn't- has not, ma'am-madam etc. which are handled by mapping. Plenty of common English words e.g. nyt – night, jan- January, gm- gud morning have changed their existence on social media. A dictionary of such out of vocabulary has been maintained in order to normalize them.

3 Results

Generally, the performance of sentiment classification is evaluated by using four indexes: Accuracy with Precision plus Recall and F1-score. A random sample of 200 sentences is picked up for testing and firstly manual testing identified and then tested by a statistical tool. This comparison also discusses the challenges and solutions. We faced and devised on evaluating sentiment analysis. Table 1 represents an accuracy of 83 % with F1-score 77 % on the English-Punjabi code mixed data set the statistical approach.

The accuracy represents the rate at which the method predicts results correctly. The precision also called the positive predictive rate, calculates how close the measured values are to each other. A F- measures that combines precision and recall is the harmonic mean of precision and recall. This score takes both false positives and false negatives into account.

Metrics	**Fivegram Approach**	**Trigram Approach**
Accuracy	82%	83%
Precision	0.83	0.83
Recall	0.71	0.71
F1-Score	0.76	0.77

Table 1: Accuracy, Precision, Recall and F1-score

In order to compute the accuracy of each technique, by calculating the intersections of the positive or negative proportion given by each technique. Table 1 presents the percentage of accuracy for fivegram approach and trigram approach.

References

Fazil, Mohd,and Muhammad Abulaish. A hybrid approah for detecting automated spammers in twitter. IEEE Transactions on Information Forensics and Security, vol. 13, pages 2707-2719, 2018.

G. Rodrigues Barbosa, I. Silva, M. Zaki, W. Meira, R. Prates and A. Veloso, Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment, Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts- CHIEA, vol.1, pages 1186-1195, 2012.

Gelman, A. & Hill, J. Data analysis using regression and multilevel/ hierarchical models, vol.1, pages 1-6, 2007.

Kalpana, R., Shanthi, N., & Arumugam, S. A survey on data mining techniques in agriculture, International Journal of Advances in Computer Science and Technology, vol. 3, pages. 426-431, 2017.

Kaur, A., & Gupta, V. Proposed algorithm of sentiment analysis for punjabi text.Journal of Emerging Technologies in Web Intelligence, vol. 6, pages 180-183, 2014.

Kaur, H., Mangat, V., & Krail, N. Dictionary based sentiment analysis of hinglishtext, International Journal of Advanced Research in Computer Science, vol. 8, pages 1-6, 2017.

Sharma, S., Srinivas, P. Y. K. L., & Balabantaray, R. C. Text normalization of code mix and sentiment analysis, In 2015 international conference on advances in computing, communications and informatics, vol.1, pages 1468-1473, 2015.

W. Medhat, A. Hassan, and H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Eng. J., no.4, vol. 5, pages 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

Yang, Chao, Robert Harkreader, and Guofei Gu. Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Transactions on Information Forensics and Security, vol. 8, pages 1280-1293, 2013.