Using Twitter Data and Sentiment Analysis to Analyze Users Comment

Abhishek Kori, Dr. Jigyasu Dubey
Department of Information Technology
Shri Vaishnav Vidyapeeth Vihwavidyalaya, Indore, India
abhishekkori.rgpv@gmail.com

Abstract. As to predict real world outcomes Twitter has been analyzed recently, and this is also true for analyzing users comments. In this work, we extract information about users comment from Twitter in Hindi-English mix language. This paper investigates the users comment, using sentiment analysis to assess to what extent Hindi-English mixed data is used by persons. This paper shows the results for a monitoring tool that allow to study the users comments.

Keywords: Hindi-English mix tweets, · Twitter · Natural Language processing, Sentiment Analysis

1. Introduction

Sentiment Analysis (commonly known as Opinion Mining) refers to the problem of identifying the dominant sentiment in a given piece of text. The sentiment modelled [1] are broadly classified as positive, negative, and neutral. With the expansion of social media data such as blogs, news articles and comments on them, YouTube comments, Amazon product reviews and Yelp reviews, online forum discussions, tweets, Facebook posts, and emails, a huge requirement to process this information and its sentimental evaluation become a big challenge. That may be converted into an opportunity if sentiment present in these pieces of text could be identified and analyse the minds of the people. The sentiments are human thoughts, opinions, or ideas which are based on the certain situation, event or past experiences. The sentiments are developed as the temporal effect or may extend with the time span [2]. The sentiments are inherent entities that depend upon certain dependent or independent thought process and share within the society or community.

The Sentimental Analysis is a contextual mining process to identifying opinions in text using computational approaches and categorizing them, in order to determine impact on future

requirements. This may include particular intellectual property or topics, product etc. The mode of analysis are online and/or offline, where reactions forms are collect as textual inputs from the intended individual and monitor the impact as output.

Public Opinions and preferences expressed in social networks are prime pre-requisites for sentiment analysis and data mining. This help to identify problem or customer satisfaction for particular services or products [3]. For example, estimating stock market prices to even predict the results of the elections before it takes place.

In [4] several lexical methods are elaborated and processing emotions in order to teach computers to extract knowledge about how human emotions changes in case of context. Language affects decisions and chat bots may get polarized with the influence. Text belongs to different languages such as English (EN) and Hindi (HI), mixed languages such as EN written in HI or HI written in EN, and some new mixed local words affect context significantly. Therefore, written utterances confirm for considering of emotions in compound sentences and processing to improve the accuracy.

Therefore, it is found that very few research work had done in the area of finding sentiment of language in Hindi-English (Hi-En) mix text. A lot of work still can be done like finding accuracy, recall, precision of this text and correctly analyzing the users sentiment in Hindi-English mixed text on social media. Further methodology will conclude the procedure and necessity of research work proposed.

2. Methodology

The overall methodology of our proposal is depicted to extract twitter data. This data or comments of users mainly concern with Hindi-English mixed comments. As introduced the first step is the extraction of tweets; to fulfill this purpose various tools are available. This paper uses the R tool for extraction of user's comments. For the purpose download latest version of R studio, an open access tool easily available. Then download the R script for writing the various scripts. Further for extract comments one have to create a twitter application as well by creating a twitter account on twitter developer's website. Credentials which are required by R studio from this developer's website are: Consumer Key, Consumer Secret, Access Secret Key, and Access Token apart from that access URL is also required. Two important library required for authentication purpose are twitteR and ROAuth. After assigning all these variables a certificate is

required to be downloaded i.e. "cacert.pem". Also assign the essential URL given in the application on developer's website for the purpose of handshaking. For authorizing the application run overall steps. After authorizing application a pin is generated, this is required to complete the authentication process. Now any number of tweets of any related word can be searched or extracted .Higher the number of tweets higher time is required to extract the tweets. Second step is to filter the user's comments restricted only for Hindi-English mixed text. Also emoticons can be identified and removed from the tweets. For filtering first convert the number of tweets in list or data frame format. Then statement having emoticons can easily be identified and removed. In this filtering process, there is another stage of filtering where positive and negative words can also be identified. In this paper only Hindi-English mixed text is identified. These positive and negative words are user defined. That means one can make its own list of positive and negative words. Using function in the R studio given, list of positive and negative words can be created and modified as well.

In this paper, filtering process consists of removing the punctuation and then converting tweets into lowercase, removing decimal by using various function available in R studio. First convert all tweets into data frames apply functions to filter and then apply function to merge all the data-frame. Use score function which cleans the tweet and merged data-frames. The library used for cleaning and merging data is reshape. Applying filtering and reshaping procedure generate score for both positive words and negative words. After filtering lexical analysis would be performed to combine and calculate the score of positive and negative words.

3. Lexical Analysis:

To perform lexical analysis three parameters are required: First is sentences or tweets available, second and third are list of positive and negative words. Before performing it various filtering should perform like removing blank spaces, removing punctuations, control statements and new lines by using various filtering functions. After that convert sentences to lowercase and split the data, for that stringr function is required. After that convert list to a character vector. These vectors took one word at a time and compare it with positive word and negative words. After matches use sum function to add all positive words similarly add it for negative words. The spitted data returns a list which could be unlisted by unlist function. Then each word is compared by match function. To check it compare each word with the list of positive words and

similarly compare each words with negative word. If it return true it would be added to the list of positive words or if it returns true for negative list of words it would be added to list of negative words. Then calculate overall score of positive words and negative words separately. Now from the data take each line paragraph calculate number of positive word score and negative word score and merge these overall data and convert it into a single merged data frame. Finally return these merged data frame. Now clean the tweets and returned the merged database, as cleaning the tweets is already explained. Create the copy of result data frame by using reshape library. To combine all the score generated melt function is used, which could be executed by storing each tweet in a separate variable. Use different tables to contain results of different data frames. Then merge results of each table to get single score as given in fig.1 below. To analyze the result perform information analysis to compare resultant.

$core	Positive	Negative
$ -1	1	2
$ 2	3	1
$ 0	1	1
$ 0	2	2
$ 1	1	0
$ -2	1	3
$ 3	3	0
$ 1	1	0
$ 1	1	0
$ 1	1	0
$ 0	2	2
$ 1	1	0
$ 2	2	0
$ -1	0	1
$ 1	1	0
$ 5	5	0
$ -1	1	2
$ -2	1	3
$ 2	2	0

Fig.1

4. Information Analysis:

In previous section lexical analysis is performed where score of positive sentiment and score of negative sentiment is calculated. Now here one more field is required to add to depict the information. This field is required to calculate the percentage of positive and negative score. For that each score is to be divided by overall score. For example positive score is divided by overall score to get positive percentage similarly to get negative percentage negative score is to be divided by overall score. Now one problem is raised here when the case is zero which in turn return result as NaN which denotes not applicable result. To remove this problem one have to

replace each NaN field with zero itself for both positive and negative score. The analysis could be done on the basis of generating histogram. The function used is hist in which rainbow can be used as color field which provides a nice colorful view of histogram given in figure below Fig.2. To analyze differently pie chart can be generated. The library used to generate pie chart is plotrix. The advanced version of pie chart can be generated using three dimensional view. In this process of extracting features hash tag can also be identified and can determine the frequency of any particular hash tag on the basis of number of times it is used. As increasing number of using hash tags in tweets increase its requirement. These frequencies can be ordered increasingly or decreasingly. Now these frequencies can be plotted using histogram plot. The library used is ggplot. These ways can be used to determine and visualize easily various dominating factors of positivity and negativity. The various trending tweets all over the world and famous celebrity hash tags as well.

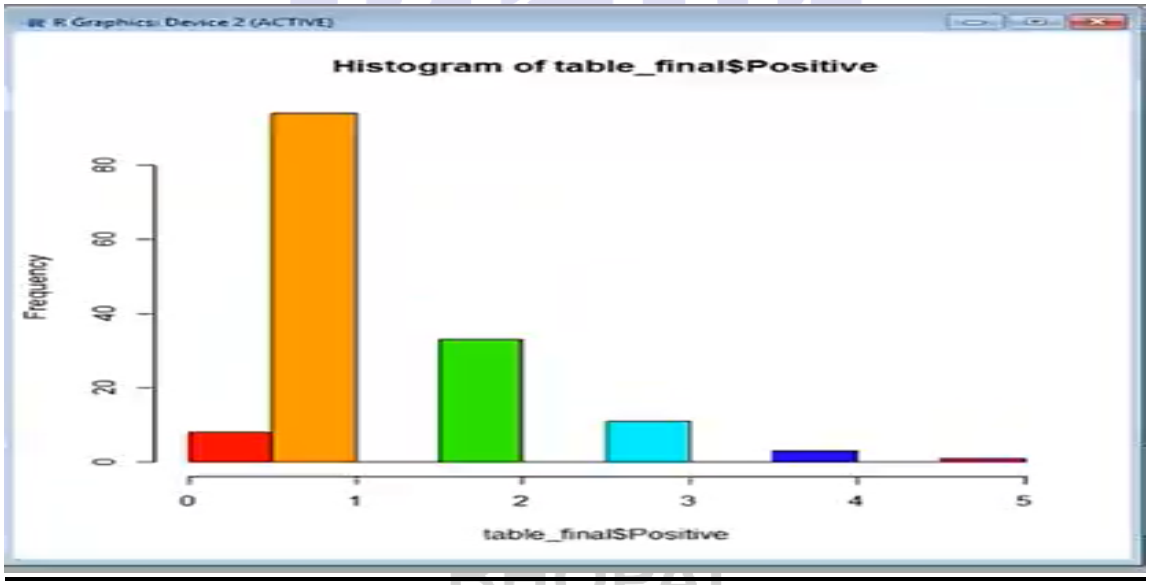

Fig.2

5. Result:

This section shows how procedure is applied, discussed in previous section, to get the desired result. The R platform is made use of the various Twitter libraries and Twitter Stream APIs to extracts the stream of tweets in particular region and analyzed the number of tweets containing Hindi English Text used as a mode of communication.

Twitter Data and Sentiment Analysis is used to analyze Hindi English Text. The total number of tweets collected was about 300 generated by about 100 unique users. Tweets have then been processed with the R Studio based platform to perform all text-processing operations described in the previous section. The number of Hash tags used is also analyzed using Twitter library available for it. Then on the basis of text collected the positive and negative sentiments are identified. Filtering of data is performed by removing various annotations. In fig.3 Histogram shows the frequency of negative and fig.4 shows positive sentiments expressed by the user using various text in twitter.

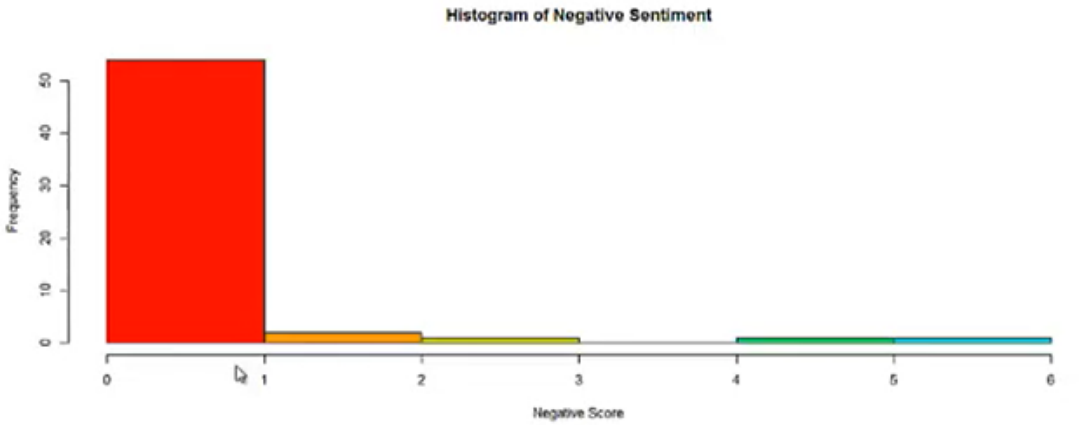

Fig.3

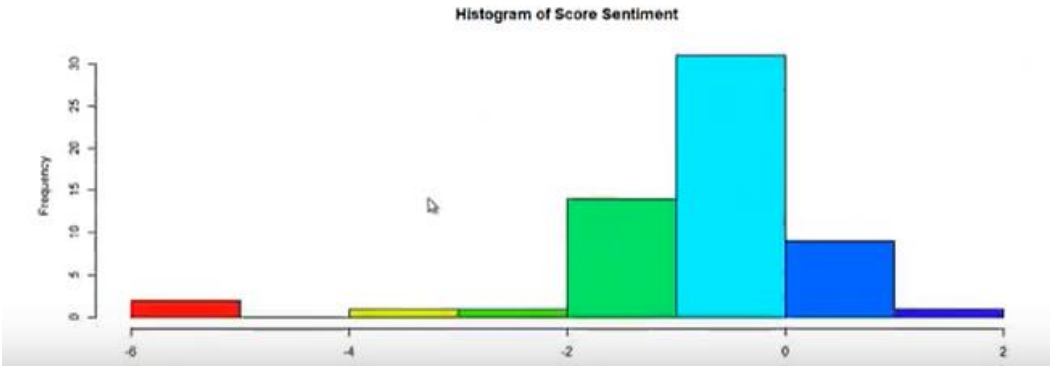

Fig.4

6. Conclusion:

The result introduced an approach to Tweeter data processing aiming at extracting sentiment of users frequently using Hindi English mix text to express their sentiments. This is achieved by also exploring the various inbuilt Twitter libraries and sentiment analysis technique used. The

final goal is to get data for studying the polarity of the sentiments in the area being considered, and first results are encouraging. The considering of other future questions as:

– the comparison with other existing proposal/tools, e.g.

– the contribution that following and followers can provide to improve the accuracy and the meaning of collected data

– how profiling users (according to age, gender, residence area, device type...)leads to better (targeted) analysis.

– how to explore other sentiment analysis methods, for instance combining lexical- and machine learning- based methods, in order to improve the effectiveness of the proposed approach Using Twitter Data and Sentiment Analysis to Study Usage of Hindi English Text

– to gather a larger number of tweets even in different geographical areas, to validate our proposal.

References:

[1] V. Jha, M. N, P. D. Shenoy and V. K. R, "Sentiment Analysis in a Resource Scarce Language:Hindi," *International Journal of Scientific & Engineering Research*, vol. 7, no. 9, pp. 968-980, September-2016.

[2] G. Abercrombie and D. Hovy, "Putting Sarcasm Detection into Context: The Effects of Class Imbalance and Manual Labelling on Supervised Machine Classification of Twitter Conversations," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop, pages 107–113*, Berlin, Germany, August 7-12, 2016.

[3] M. O. Shiha and S. Ayvaz, "The Effects of Emoji in Sentiment Analysis," *International Journal of Computer Electrical Engineering*, vol. 9, no. 1, pp. 360-369, 2017.

[4] R. Rzepka, M. Takizawa, J. VallverduI, M. Ptaszynski, P. Dybala and K. Araki, "From Words to Emoticons: Deep Emotion Recognition in Text and Its Wider Implications," *International Journal of Computational Linguistics Research,* vol. 9, no. 1, pp. 10-26, 2018.

[5] D. B. a. N. A. Smith, "Contextualized Sarcasm Detection on Twitter," in *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 2015.

[6] R. W. Gibbs, "Irony in Talk Among Friends," *Metaphor and Symbol,* vol. 15, no. 1&2, pp. 5-27, 2000.

[7] A. Joshi, P. Bhattacharyya and M. J. Carman, "Automatic Sarcasm Detection: A Survey," *ACM Computing Surveys,* vol. X, no. Y, p. 22, 2017.

[8] F. Barbieri, F. Ronzano and H. Saggion, "Relying on intrinsic word features to characterise subjectivity, polarity and irony of Tweets," *EVALITA 2014,* pp. 104-107, 9-11 December 2014.

[9] S. Phani, S. Lahiri and A. Biswas, "Sentiment Analysis of Tweets in Three Indian Languages," in *93 Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing*, Osaka, Japan, December 11-17 2016.

[10] A. Ghosh and I. Dutta, "Real-time Sentiment Analysis of Hindi Tweets," Conference Paper · November 2014.

[11] B. G. Patra, D. Das, A. Das and R. Prasath, "Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets - An Overview, pages 650–655. Springer International Publishing, Cham.," in *Mining Intelligence and Knowledge Exploration*, Lecture Notes in Computer Science book series (LNCS, volume 9468), 2015, p. 650–655.

[12] A. Joshi, A. Prabhu, M. Shrivastava and V. Varma, "Towards Sub-Word Level Compositions for Sentiment Analysis of Hindi-English Code Mixed Text," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, December 11-17 2016.

[13] F. Barbieri, H. Saggion and F. Ronzano, "Modelling Sarcasm in Twitter a Novel Approach," in *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Baltimore, Maryland, USA, 2014.

[14] M. Khodak, N. Saunshi and K. Vodrahalli, "A Large Self-Annotated Corpus for Sarcasm," arXiv:1704.05579v4 [cs.CL] 22 Mar 2018, 2018.

[15] D. Ghosh, W. Guo and S. Muresan, "Sarcastic or Not: Word Embeddings to Predict the Literal or Sarcastic Meaning of Words," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 17-21 September 2015.

[16] A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden and A. Reyes, "Sentiment Analysis of Figurative Language in Twitter," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, 2015.

[17] A. Mensikova and C. A. Mattmann, "Ensemble Sentiment Analysis to Identify Human Trafficking in Web Data," in *Proceedings of ACM Workshop on Graph Techniques for Adversarial Activity Analytics (GTA3 2018)*, New York, NY, USA, 2018.

[18] R. J. Kreuz and K. E. Link, "Asymmetries in the Use of Verbal Irony," *Journa of Language and Social Psychology*, vol. 21, no. 2, pp. 127-143, June 2002.

[19] R. Giora, "On irony and negation," *Discourse Processes*, vol. 19, no. 2, pp. 239-264, 1995.

[20] A. Ghosh and T. Veale, "Fracking Sarcasm using Neural Network," in *Proceedings of NAACL-HLT 2016*, San Diego, California, 2016.

[21] A. Reyes, P. Rosso and T. Veale, "A multidimensional approach for detecting irony in Twitter," *Lang Resources & Evaluation*, vol. 47, p. 239–268, 2013.

[22] M. S. Kaiser, K. T. Lwin, M. Mahmud, D. Hajializadeh, T. Chaipimonplin, A. Sarhan and M. A. Hossain, "Advances in Crowd Analysis for Urban Applications Through Urban Event Detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, no. 99, pp. 1-21, 2017.

[23] W. Wang, K. Zhu, H. Wang and Y.-C. J. Wu, "The Impact of Sentiment Orientations on Successful Crowdfunding Campaigns through Text Analytics," vol. 11, no. 5, 2017.

[24] Z. Hai, G. Cong, K. Chang, P. Cheng and C. Miao, "Analyzing Sentiments in One Go: A Supervised Joint Topic Modeling Approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 6, pp. 1-14, 2017.

[25] A. U and S. M. Thampi, "Linguistic Feature Based Filtering Mechanism for Recommending Posts in a Social Networking Group," *IEEE Access*, vol. 6, pp. 4470-4484, 2018.

[26] Z. Chen, F. Lu, X. Yuan and F. Zhong, "TCMHG: Topic-Based Cross-Modal Hypergraph Learning for Online Service Recommendations," *IEEE Access*, vol. 6, pp. 24856 - 24865, 2017.

[27] X. Lei, X. Qian and G. Zhao, "Rating Prediction Based on Social Sentiment From Textual Reviews," *IEEE Transactions on Multimedia* , vol. 18, no. 9, pp. 1910 - 1921, 2016.

[28] D. Davidov, O. Tsur and A. Rappoport, "Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon," in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, Uppsala, Sweden, 2010.

[29] F. Barbieri and H. Saggion, "Modelling Irony in Twitter: Feature Analysis and Evaluation," in *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, April 26-30, 2014.

[30] B. Liu, "Sentiment Analysis and Subjectivity," in Handbook of Natural Language Processing, 2010, pp. 1-38.

[31] Akshat Bakliwal, Piyush Arora, and Vasudeva Varma. 2012. Hindi subjective lexicon: A lexical resource for hindi polarity classification. In Proceedings of International Conference on Language Resources and Evaluation.

[32] Utsab Barman, Amitava Das, JoachimWagner, and Jennifer Foster. 2014. Code-mixing: A challenge for language identification in the language of social media. In In Proceedings of the First Workshop on Computational Approaches to Code-Switching.

[33] Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In Proceedings of the Forum for Information Retrieval Evaluation, FIRE '14, pages 48–53, New York, NY, USA. ACM.

[34] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boomboxes and blenders:Domain adaptation for sentiment classification. In In ACL, pages 187–205.

[35] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. CoRR, abs/1607.04606.

[36] Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In AAAI, pages 2153–2159.

[37] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement,20(1):37–46, April.

[38] Amitava Das and Sivaji Bandyopadhyay. 2010. Sentiwordnet for indian languages. In Proceedings of the EighthWorkshop on Asian Language Resouces.

[39] C´ıcero Nogueira dos Santos and Victor Guimar˜aes. 2015. Boosting named entity recognition with neural character embeddings. CoRR, abs/1505.05008.

[40] Ronen Feldman. 2013. Techniques and applications for sentiment analysis. Commun. ACM, 56(4):82–89, April.