THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

Sentiment Extraction: A Technique For Punjabi Tweets Using Maxent Classifier

Amarinder Kaur

(amarinder.21482@lpu.co.in)

Assistant Professor, School of Computer Science & Engineering
Lovely Professional University, Jalandhar, Punjab

Sudha Shanker Prasad

(sudha.21826@lpu.co.in)

Assistant Professor, School of Computer Science &
EngineeringLovely Professional University, Jalandhar,
Punjab

***Abstract*— This research paper describes the work done towards sentiment extraction from social media posts. We focused on twitter for our data source that helps the users to describe their thoughts and views through short message called as Tweets. The tweets have only 140 character, hence very short to carry views and to analyse user sentiment. We have a twitter training dataset in Punjabi Language. We tried different data mining approach including MaxEnt as classifiers for extracting sentiments. We used one on most popular Data Mining Tool Weka to automatically extract the sentiment of Punjabi tweets into negative, positive or neutral.**

***Keywords*—** Sentiment Extraction, Data Mining, Machine Learning, Supervised Learning, Information Retrieval

I. INTRODUCTION

A Computer Science domain that deals with study of human languages(linguistics), that show interaction or interface of computer and natural(human) language is generally known as Natural Language Processing [1].In Natural Language Processing(NLP), one of the most prominent research area is Sentiment Extraction. It tries to extract the human perception or sentiments form the day to day topics, discussion, social media interaction. In simple words, it tries to extract user opinion form a text with two major opposite sentiments i.e. Negative or Positive. As we know Internet, Chats, Microblogging, Social Blogging, etc. has created it large presence in the users. Hence, web contents generated from users are increasing rapidly. This actually create a new research paradigm for Sentiment Extraction in contents generated from Social Media. One of most proper microblogging site, Twitter that is majorly used by users to express their opinion and views. Tweets are limited to 140 characters. Due this restriction, users generally uses variation in spelling, different punctuation marks & emoticons to emphasize in their tweets. Using the mentioned variations actually createSentiment Extraction a new scope for research & task more challenging. Sentiment Extraction in foreign languages such as English, French, etc has achieved a significant amount of result already. It also helps to create the better corpus, dictionary and various resources and tool for foreign languages. Now a day's web contents in Indian Languages has created a vast space in internet and also growing day by day. Hence, creating an opportunity for research in Sentiment Extraction for Indian Language. We also want to point out that limitation of various resources and tools for Indian Language, still make this task very challenging. We know that Punjabi Language has large nos. of audience on the basis of substantially large number of speaker. Thisresearch paper shows the work we carried out for achieving sentiment extraction forPunjabi Language Tweets into classes based on sentiments i.e. neutral, negative or positive. In order to achieve the result, we carried our work through MaxEnt algorithm.

Our paper has been arranged as follows. Section 2 describe past work done in this research area. Section 3 focuses on system information. Section 4 focuses on the way we solved the problem. Section 5 describes the Experiment results. Section 6 describeanalysis & discussion. Section 7 concludes the paper by dealing with conclusion.

II. RELATED WORKS

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

In research community, Sentiment Extraction has attracted many researchers to work upon. In past, enormous amount of research has been carried out in this paradigm. Some of prominent work in this paradigm include research done by Pang [3] &Turney [2] for classification of polarity for reviews of various products. A polarity bases classification is tried by Synder[5] and Pang[4] for document classification in multiway approach. They worked upon star ratings of movie reviews for classification into negative or positive. Synder focused on reviews of restaurant on different aspects such as ambience, taste and quality of food.In past, several task has been carried out for feature extraction, entities identification, feature opinion in multiway classification namely neutral, negative or positive. In Liu's NLP Handbook chapter "Sentiment Analysis and Subjectivity" [6] has discusses sentiment extraction in much depth.In last decade, we can notice a prominent amount of research has been carried out by researchers in web contents such as social blogging, microblogging. These social tools has gain vast user base and prominently used for expressing public sentiments, stock market information, results of elections [5, 12, 14] and also for management & preparedness of disaster [13]. But, Sentiment Extraction in Indian language is still a new paradigm of research and research community has shown a good interest. Some of past works include Indian Language is Bengali [7]. A sentiment detection reported for Hindi in [8]. A wordnet for Bengali using various dictionary that supports Bengali-English and lexicons available in English has been discussed by Bandyopdhya and Das [9]. Detecting emotional expressions and labelling them for Bengali corpus is carried out by Bandyopdhya and Das [10]. Sentiment Analysis on movie review data in Malayalam on rule based technique carried out by Jisha&Deepu [11]. Sentiment Classification for Manipuri on basis of verb as lexicon is carried out by Kumar and Kishorjit [12]. Balmurali [13] discussed a feature generated though WordNet on basis of cross lingual sentiment classification. It internally uses SVM(Support Vector Machine).

III. SYSTEM INFORMATION

We mention the insights of our research below:

A. Dataset

The dataset available for training contains 1282 Punjabi tweets with either of the three sentiment classes namely Positive, Neutral and Negative. The Test Data contains 3265 without annotations. We used one on most popular Data Mining Tool WEKA in our research to achieve the sentiment extraction.

B. Classifier Used

The idea behind Maximum Entropy (MaxEnt) model is to use a set of user-defined features and accordingly learn appropriate weights. The model internally works on search-based optimization technique to find weights for the features that maximize the likelihood of the training data.

Assuming only the word level features, we define a joint feature $f(x, y) = S$, for each word x and class $y \in C$ where, S is count of x occurs in query in class C. With iterative optimization technique, we assigned a weight to each joint feature to increase the overall likelihood of the training data. The probability of class z given a query x and weight λ.

$$P(z|x, \lambda) = \frac{exp \sum_i \lambda_i f_i(z,x)}{\sum_{y' \in C} exp \sum_i \lambda_i f_i(z',x)}$$

These parameters are used to maximize the entropy of distribution.

IV. PROPOSED APPROACH

Sentiment extraction is a technique that is used to classify a Punjabitweet into their sentiment annotation namely either positive, neutral or negative. we noticed our research work as a classification problem that has three classes i.e. Positive, Negative & Neutral. It can be considered as a Supervised Machine Learning Problem.

A flowchart demonstrates a step by step solution of the proposed approach for sentiment extraction:

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

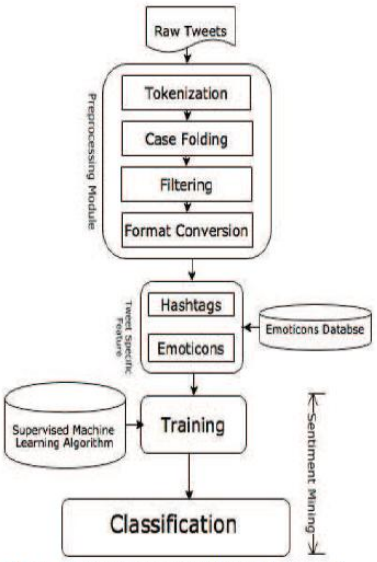

Fig. 1. System Description for Proposed Approach

Given below steps describe the steps we considered following steps to accomplish our research work.

A. Pre-Processing Track

Given Steps are carried out for considering pre-processing of thetwitter data from test data & training data.

1) Token Conversion: As twitter data are created through users,generally user mistakenly add multiple words without considering any whitespace in between them.

2) Filter Step: Tweets generally have Web(URL) links, user information &punctuation that doesn't have sentiment information and can be removed safely.

Conversion of Format: .arff conversion is considered.

B. Feature Findings

This steps handles of finding features that participate in sentiment of the Punjabi tweets.

1) Tweet features: Hashtags and Emoticons are two tweet specific features.

• Hashtags: Hashtags are keywords that start with hash sign (for eg. #happy_birthday). It is considerable element for twitter. We tried consideringthe keywordsavailable in the hash-tag to look into the affectiontowards a peculiarsentiment class after ignoring the hash sign.

• Emoticons Symbol: Generally, weuse different emoticonsto show our sentimental Views. Wecaptured this as apeculiar feature and we tried polarizing emoticons on basis of sentiments.

C. Sentiment Extraction

The above considered resources & features are taken into consideration with MaxEntsupervised algorithm to help our classifier system to learn from test data. The trained system(classifier)is further used to classify Punjabitweets from the testing data into their given sentiment classes.

V. EXPERIMENTS RESULTS

We choosed three parameter namely Precision, Recall & F-Measure to measure the achievement of our system. Table 1 describes the training result of the system created throughMaxEnt classifier.

Class	Precision	Recall	F-Measure
Neutral	0.715	0.839	0.770
Negative	0.814	0.841	0.829
Positive	0.770	0.373	0.503

Table 1: DETAILED ACCURACY OF PUNJABI TWEETS USING MAXENT

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

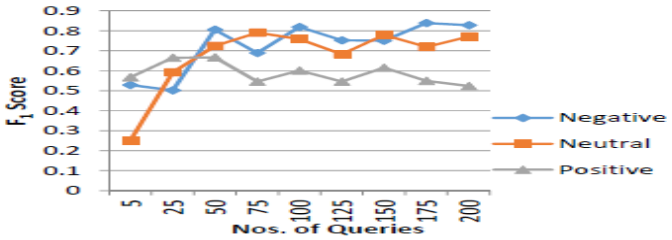

Fig 2: PUNJABI TWEETS USING MAXENT

VI. ANALYSIS & DISCUSSION

We noticed that, Extraction of sentiments in Punjabi Languages is very typical due to its lack in resources. Since the size of training data is very limited, hence we can't build a robust system. In respect of Foreign languages such as English, French, etc., Our accuracy is comparably very low. The availability of limited resources for Punjabi Language is to taken into consideration for poor performance. Tweets generally doesn't have proper sentence construction resulting many variation of spellings, incorrect punctuation marks & incorrect usage of emoticons. In tweets generally multiple language are used to emphasize the context resulting more difficult to extract sentiment.

VII. CONCLUSION

In our research work, we discussed our analysis on Sentiment Extraction in Punjabi Language tweets. The major goal is describing the polarity of tweets for Punjabi Languages into their sentiment namely neutral, positive or negative. To achieve this, we created systems based onprobability algorithms for sentiment extraction for Punjabi tweets. We have suggested a way to create a model using MaxEnt for sentiment extraction.

REFERENCES

[1] A. Trivedi, A. Srivastava, I. Singh, K. Singh, and S. K. Gupta, "Literature Survey on Design and Implementation of Processing Model for Polarity Identification on Textual Data of English." *IJCSI*, 2011.
[2] P. D. Turney. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." *In Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417-424, Association for Computational Linguistics, 2002.
[3] B. Pang, L. Lillian, and S. Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In *Proceedings of the ACL-02 International Conference on Empirical methods in natural language processing-Vol. 10, Association for Computational Linguistics*, pp. 79-86, 2002.
[4] B. Pang, and L. Lillian, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales." In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics*, pp. 115-124, 2005.
[5] B. Snyder, and R. Barzilay. "Multiple Aspect Ranking Using the Good Grief Algorithm." In *HLT-NAACL*, pp. 300-307, 2007.
[6] B. Liu. "Sentiment Analysis and Subjectivity." *Handbook of natural language processing*, pp. 627-666, 2010.
[7] A. Das, and S. Bandyopadhyay. "Subjectivity detection in English and bengali: A crf-based approach." In *Proceeding of ICON*, 2009.
[8] S. S. Prasad, J. Kumar, D. K. Prabhakar, and S. Pal. "Sentiment Classification: An Approach for Indian Language Tweets Using Decision Tree." In *International Conference on Mining Intelligence and Knowledge Exploration*, pp. 656-663. Springer International Publishing, 2015.
[9] A. Das, and S. Bandyopadhyay. "SentiWordNet for Bangla.", Knowledge Sharing Event-4: Task 2, 2010.
[10] D. Das, and Sivaji Bandyopadhyay. "Labeling emotion in Bengali blog corpus–a fine grained tagging at sentence level." *In Proceedings of the 8th Workshop on Asian Language Resources, 2010.*
[11] Nair, Deepu S., Jisha P. Jayan, R. R. Rajeev, and Elizabeth Sherly. "SentiMa-sentiment extraction for Malayalam." In *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1719-1723, IEEE, 2014.
[12] K. Nongmeikapam, D. Khangembam, W. Hemkumar, S. Khuraijam and S. Bandyopadhyay, "Verb Based Manipuri Sentiment Analysis", *IJNLC*, vol. 3, no. 3, pp. 113-119, 2014.
[13] A. R. Balamurali, "Cross-lingual sentiment analysis for Indian languages using linked wordnets.", 2012.