Grenze International Journal of Engineering and Technology, July Issue

Sentiment Classification of Marathi Text using word's N Gram Polarity and Machine Learning Algorithms

Pallavi V. Kulkarni[1] and Deepa S. Deshpande [2]
[1] Government Engineering College , Computer Science & Engineering Department, Aurangabad, India
Email: pallavi.k11@gmail.com
[2]Jawaharlal Nehru Engineering College, Computer Science & Engineering Department, Aurangabad, India
Email: deepadeshpande@jnec.ac.in

***Abstract*—Sentiment Analysis is initial step of Emotion Recognition which is one of the driving force in the area of research in Artificial Intelligence. Availability of large Web text and improved tools in Natural Language Processing are attracting researchers to this field. However there is tremendous research scope in Marathi Language spoken mainly in Maharashtra, India which is very ancient and rich in morphology. This paper presents an attempt to develop a machine learning model for sentiment classification of Marathi Corpus . N Gram feature of word and polarity of each word is considered to calculate document polarity. Both Negative score and Positive score are fed to classifiers and performance is compared. The conclusion doesn't stop on any one of the classifier but SVM and Logistic Regression gives overall good results in all circumstances. Stochastic Gradient Descent gives highest accuracy for both bigram and trigram features when 70:30 split is used.**

***Index Terms*— Sentiment Analysis, Ngram, Natural Language Processing, Machine Learning Algorithms.**

I. INTRODUCTION

Emotion is an integral part of human life. Irrespective of location ,caste,country every human being wants or bound to express his feelings in the form of emotions. Sentiments are more general form of emotions which are described as positive or negative. Amongst various ways to express emotions, language is the most powerful and long lasting medium. Textual representation of language can be done only by human being amongst all living beings. Sentiment analysis is the area of Natural Language Processing, grooming the research in Artificial Intelligence. A large amount of text available on Web in the form of structured or unstructured form motivates the task. The scope and challenges of sentiment classification increases as language changes. A lot of work is done in English and other Europian languages like French, German, Spanish etc[1]. Urdu and Arbaic are also contributing in solving the problem. Japanese and Chinese language sentiment classification work is progressing apart from their unique pictorial scripts. Indians are also putting their efforts for sentiment analysis of different Indian languages like Bengali, Gujarati,, Marathi, Punjabi etc along with Hindi.

Marathi spoken in Maharashtra being an ancient language has its own importance due to its scripting similarity with Sanskrit. The research in sentiment analysis of Marathi text is in its infant. The paper

Grenze ID: 01.GIJET.6.2.6_1

describes implementation of different machine learning models applied on polarity annotated corpus for tourism database created at CFILT IIT Bombay[1]. The seed database from Emolex is used by assuming the universal fact that words express same type of sentiment or emotion irrespective of language it belongs[3].

II. LITERATURE SURVEY

Chitra Choudhari[6] calculated document polarity using polarity of individual sentence. Presence of positive or negative words from wordnet which decide sentence polarity.A java suite of tools GATE (General Architecture of Text Engineering is used to accomplish the purpose . Again the size of dictionary of positive and negative words is the limitation. Sujata Deshmukh[7], implemented sentiment classification of Marathi text using English sentiwordnet. A feasible corpus of Marathi language using English Sentiwordnet is created in first step and second step maps the keywords for sentiment analysis. This attempt has limitation due to small corpus size and unavailability of NLP tools required for preprocessing. Snehal Pawar[8] used collection of positive and negative words and applied support vector machine , naive bayes and maximum entropy for classification of sentences. Only adjectives are extracted as a feature to decide the polarity of the sentence. She found sentiments to be domain dependent and difficulty in handling scarsum .
Mohammed [9] classified Hindi as well as Marathi text transliterated documents using KNN, Naive Bayes and Support vector machine(SVM). Bilingual dictionary was used to translate Marathi word to Hindi and then Hindi Senti WorNet is used to get the sentiment score. After processing 1200 Hindi and 300 Marathi documents they concluded Linear SVM is performing better than lexical analysis based naive bayes.

A. Seed Database

Saif M. Mohammad [2][3] constructed an Emotional Lexicon along with semantic orientation of words through annotation based method using amazon mechanical Turk . The database consist of more than 14000 words along with emotional scores and semantic polarity . The crowdsourcing process was implemented for English later lexicon is extended for 104 different worldwide languages including Marathi using google translator .The subset of Marathi words along with positive and negative polarity is used as the seed sentiment dataset for the purpose.
Balmurali[1] created Marathi Sentiment polarity database for travel destination reviews. A review consists of 4 -5 sentences and 10 – 15 words each. This corpus consists of 75 positive and 75 negative reviews. The corpus was annotated by native speakers and later analyzed using cross lingual sentiment analysis . CSVM machine Learning technique is applied using word as a feature and achieved 92.78% accuracy.The polarity sense travel review dataset is used as a basekine for the implementation purpose.

B. Machine Learning techniques

Logistic regression
It is a linear classification model where probabilities describe the possible outcomes using logistic function. It is discriminative model which directly estimates p(y/x) in a single trial.
Consider each example is represented by a feature vector x and it is classified into class y ,where y have only two values y= -1,1. First x is mapped to a real number, if x is having very positive value means x is likely to be positive (y = 1). If x is equal to very negative value indicates x is negative (y = −1). The parameter vector θ is calculated using inner product where θ: $\theta > x$. The range of inner product is between $(-\infty,\infty)$. It is worth emphasizing that $\theta \in R^d$, where d is the dimension of input features x.[12].

Stochastic Gradient Descent
It is an iterative model which optimizes an objective function using suitable smoothing properties. It can be regarded as a stochastic approximation of gradient descent optimization as it replaces the actual gradient by an estimate. It is simple and very efficient approach to fit linear models . Particularity for large size input Stochastic Gradient Descent algorithm is used. SGD classifier fits a logistic regression model with loss=”hinge”]12].

Multinomial Naive Bayes:
It implements naive bayes for multinomially distributed data. Two or more variables are involved and they decide output class. Probability distribution of these input variables is used used to calculate the outcomes of experiments .
A word vector is represented as

$$\theta y = (\theta y_1, \ldots . \theta yn) \quad (1)$$

for each class θy_i is the probability $p(x_i/y)$ of features appearing in a sample belonging to class y[12].

Bernoulli Naive Bayes

Implements the naive Bayes training and classification algorithm for data that is distributed according to multivariat Bernoulli distributions[12].

Bernoulli distribution is the discrete probability distribution of random variables which either 1 or 0.

$$\text{pmf} = q = p-1 \quad \text{if} \quad k=0$$
$$= p \quad \text{if } k=1$$

Linear SVC Classifier

It is implementation of support vector classification implicitly for case of linear kernel. Support vector classifiers try to find best hyperplane to separate different classes by maximizing the distance between sample points and the hyperplane[12].

III. System Development

The system is developed in three steps: Preprocessing, Feature Construction and Classification. Figure 1 describes the overall system implemented for sentiment classification.

A. Preprocessing

The input baseline database is 150 files of sentiment annotated reviews. As creation of this corpus is based on Synaset ids, the words are present with Synaset ids. These ids are removed and tokenization is done for all elements of the database.

B. Feature Construction

Bi-gram or Tri-gram sequence of words is created for each file as per the requirement. For each n gram a triplet is constructed by extracting positive and negative polarity of each word from Emolex sentiment polarity dataset of Marathi text using the following formula.

Polarity score of n gram= Σ polarity score of each token present in the documents.

C. Classification

The n grams of positive reviews and negative reviews is calculated separately and then mixed and shuffled in a single file. Average of positive polarity and Negative Polarity of each triplet per review file is calculated to construct the featureset . The train -Test split function from Scikit Learn library of python is used to divide training and testing featuresets. This input is fed to Machine Learning Algorithms for constructing Machine Learning Model. Scikit Learn library for machine learning is used from python as it has simple and effective tools for predictive data analysis. It is open source and reusable in various context. Logistic Regression,Stochastic Gradient Descent, Bernoulli Naive Bayes, Multinomial Naive Bayes, SVC Classifier are selected from the domain. These classifiers are useful for classification of text with multiple feature vectors with large size.

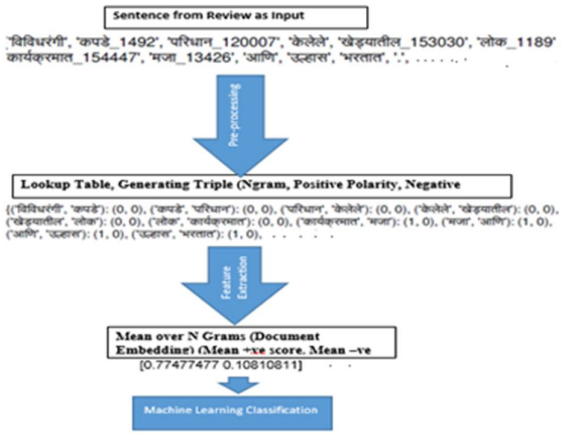

Figure 1. The overall structure of the proposed system

IV. Results and Discussion

TABLE I: ACCURACY OF MACHINE LEARNING CLASSIFIERS FOR N GRAM WORD SENSE BASED SENTIMENT ANALYSIS

Model	Train :Test Ratio	Bigram Accuracy(%)	Trigram Accuracy (%)
Logistic Regression	70:30	86.667	75.556
	60:40	76	83.333
Bernoulli Naive Bayes	70:30	44.44	35.556
	60:40	51.667	50.00
Multinomial Naive Bayes	70:30	80.00	57.778
	60:40	75	80
SVC	70:30	82.22	75.556
	60:40	75	83.333
Stochastic Gradient Descent	**70:30**	**88.889**	**84.44**
	60:40	78	60.00

In the experiment every polarized file is scanned iteratively. Sentence from review is extracted using inbuilt function TokenizeSentence('marathi') from NLTK. The Synaset ids associated with words are removed with the help of regular expression .Now these tokens are fed to find-ngrams(input_list,n) user defined function to generate bigrams and trigrams.

For each word from gram , sentence polarity is extracted from sentiEmolex database and some of positive and negative score is stored with respective gram in tuple form. These tuples along with their class labels of positive and negative sentiment files are mixed and shuffled for feature extraction.Mean of every sentiment score for each file is fed as feature vector to the classifiers. Experimentaion is done for different train test size of both bi gram and tri gram features. The 70:30 ratio indicates 70% training data and 30% testing data, while 60:40 ratio indicates 60% training and 40% testing data.

The result are mentioned in Table 1. It indicates Logistic Regression and Support Vectors machine are giving better performance. Stochastic Gradient Descent is best for 70:30 ratio for both bigram and trigram. Multinomial Naive Bayes is also giving satisfactory results but Bernoulli Naive Bayes is consistently poor compared with all other classifiers.

This attampt for sentiment classification using n gram polarity of words is unique for Marathi text. The sentiment Emolex used is the output of google translation from English words. There are limitations as some of the words are not having equivalent word. Also the view of local speaker is important . Building a strong Sentiment Lexicon through crodsoursing is essential.

V. Conclusion

Performance of five common classifiers from sklearn is observed by changing train-test size for bigram and trigram as feature vectors. Each of one behaves differently in corresponding problem space. Overall Logistic Regression and SVC perform better . Bernoulli Naive Bayes disappoints in all respect compared to other classifiers. Stochastic Gradient Descent gives highest accuracy for both bigram and trigram features when 70:30 split is used. Finally concluding to the point that word polarity playes an important role to calculate sentence polarity and the overall document polarity. A strong senti lexicon and rule based feature extraction is necessary to increase the depth of sentiment classification.

Future Work

The rich domain in vocabulary, grammar, literature of language is challenge for Natural Language Programming researchers. There is a scarcity of NLP tools required for preprocessing the text at every stage of NLP process. Construction of a grammatically structured sentiwordnet is required the most. The future focus will be construction of lexicon with sense annotation for grammatical constructs like verb, adjectives,

adverbs etc. for Marathi Language. Rule Based sentiment analysis attracts for increasing the depth of sentiment classification.

REFERENCES

[1] Balamurali A R, Aditya Joshi, Pushpak Bhattacharyya "Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets", Proceedings of COLING 2012: Posters, pages 73–82, COLING 2012, Mumbai, December 2012.

[2] Saif M. Mohammad and Peter D. Turney, "Crowdsourcing a Word–Emotion Association Lexicon",2013 National Research Council Canada arXiv:1308.6297v1 [cs.CL] 28 Aug 2013

[3] Saif M. Mohammad and Peter D. Turney, "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon"

[4] Saif M. Mohammad, "Challenges in Sentiment Analysis", A Practical Guide to Sentiment Analysis 2015 National Research Council Canada.

[5] Pooja Bolaj, Sharvari Govilkar, "Text Classification for Marathi Documents using Supervised Learning Methods" International Journal of Computer Applications · December 2016 DOI: 10.5120/ijca2016912374

[6] Chitra V. Chaudhari , Ashwini V. Khaire ,Rashmi R. Murtadak , Komal S. Sirsulla , " Sentiment Analysis in Marathi using Marathi WordNet", Imperial Journal of Interdisciplinary Research (IJIR) Vol-3, Issue-4, 2017 ISSN: 2454-1362, http://www.onlinejournal.in

[7] Sujata Deshmukh, Nileema Patil, Surabhi Rotiwar, Jason Nunes, " Sentiment Analysis of Marathi Language", Novateur Publications, International Journal of Research Publications in Engineering and Technology [IJRPET] ISSN;2454-7875 VOLUME 3, ISSUE 6, Jun-2017

[8] Snehal V. Pawar, Swati Mali, " Sentiment Analysis in Marathi Language", International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8, ISSN: 2321-8169 21-25

[9] Mohammad Arshad Ansari, Sharvari Govilkar , " Sentiment Analysis of Mixed Code for the Transliterated Hindi and Marathi Texts", International Journal on Natural Language Computing (IJNLC) Vol 7. No. 2. April2018

[10] Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, "Classification of sentiment reviews using n-gram machine learning approach", Elsevier Expert Systems With Applications 57 (2016) 117–126

[11] Dmitriy Bespalov, Bing Bai, Yanjun Qi, Ali Shokoufandeh, "Sentiment Classification Based on Supervised Latent n-gram Analysis" CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK. Copyright 2011 ACM 978-1-4503-0717-8/11/10

[12] http://scikit-learn.org/stable