International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

A Design on Recognization of Sentiment Analysis of Marathi Tweets using Natural Language Processing

Renuka Ashokrao Naukarkar[1], Dr. A. N. Thakare[2]

[1]M.Tech, Computer Science and Engineering , Department of Computer Engineering Bapurao Deshmukh College of Engineering, Sevagram Wardha, Maharashtra, India

[2]Assistant Professor, Department of Computer Engineering, Department of Computer Engineering Bapurao Deshmukh College of Engineering, Sevagram, Wardha, Maharashtra, India

Article Info
Volume 8, Issue 2
Page Number : 451-455

Publication Issue
March-April-2021

Article History
Accepted : 02 April 2021
Published : 10 April 2021

ABSTRACT

Sentiment Analysis of Marathi Tweet using Machine learning Concept is done in this paper. The tweets are classified into Positive, Negative and Neutral by using different concepts. It is difficult to predict Marathi tweet results from tweets in Marathi language. So, we used different tool to get tweets in Marathi tweet. Sentiment Analysis also shows the higher accuracy of Marathi Tweet data. The proposed work explain Sentimental Analysis of Marathi tweets, which have been classified into positive, negative and neutral using machine learning algorithms like NLP. and shows the higher accuracy of text data.

Keywords :- Machine Learning Algorithm, Sentiment Analysis, Marathi Tweets.

I. INTRODUCTION

Sentimental Analysis is the process of decide whether a piece of writing is positive, negative or neutral. Sentimental analysis is the technique to analyzes the text that imperfection polarity within text, whether a entire documents, passage, sentence or clause. Nowadays, due to the huge number of daily posts on social network people opinion is very critical for decision making millions of people gives their opinion in their other tongue through social media like Twitter, Blogs, Facebook, Instagram, etc. Sentimental Analysis plays essential role in the field of business, politics and Public Action. Text limitation of tweet massages with 280 characters per each. So, Sentiment level analysis is one of the main directions in sentiment analysis.

Maharashtra's mother tongue is Marathi is most commonly used language to express their opinion through twitter. The Sentimental Analysis of Marathi twitter message is unavoidable since there exists on automatic sentiment analysis in this language. Marathi language is very expressive language and the language is write in the form of text. Text contain word as well as hyperlink, special character, punctuation, number, symbols, etc. to removing such type of expression is the major task. The proposed work explain Sentimental Analysis of Marathi tweets, which have been classified into positive, negative and

neutral using machine learning algorithms such as NLP. and shows the higher accuracy of text data.

Bag-of-words (BOW) is the most popular technique to model text in statistical machine learning approaches in sentiment analysis. However, the performance of BOW sometimes stand short due to some fundamental deficit in handling the polarity shift issue. For that we collect the Positive words, Negative words and stopwords in Marathi language. And make a dictionary of that words.

Objectives

Sentiment Analysis for the Marathi language is the new trending work research field as number of system is available for many other languages but for Marathi language not much research work has been done. Classifying and identifying sentiment in the form of text. Nowadays, social media a huge form of sentiment oriented rich data in the form of blog post, Facebook, twitter, etc. This user generated web oriented data may contain very useful information that helps for finding the sentiments of the crowed data or gating useful information from unstructured data. Sentiment Analysis is to predict the emotion. Emotions are the representation of different facial expression. For analyzing this expression we use different methods and algorithms. For that following are the defined objectives for SA on Marathi data.

1. To classify tweet data by using different methods.
2. To identify tweet data by using different algorithms, whether it is positive negative or neutral.
3. To shows accuracy of tweet data.

II. BACKGROUND AND RELATED WORK

SA has been studied and employed widely for the last two decades. Most of the works in SA are specific for the English language.

Pang and Lee proposed three different machine learning algorithms such as NB, Maximum Entropy, and SVM with unigram and bigram features for SA of movie reviews in English. They showed that SVM outperforms other two classifiers [1].

SA has been done in different Indian languages like Bengali, Hindi, Punjabi, Manipuri, Kannada, Tamil and Malayalam. Soumya S., Pramod K.V proposed Sentiment analysis of Malayalam tweets using machine learning techniques classified into positive and negative using different machine learning algorithms such as NB, SVM, RF [2]. Sentiment Analysis of Malayalam Tweets using Machine Learning Technique is done in this paper. By using different machine learning technique tweets are classified into positive and negative. And also shows the higher accuracy. Marathi language is very expressive language and the language is write in the form of text. Text contain word as well as hyperlink, special character, punctuation, number, symbols, etc. to removing such type of expression is the major task. In the Marathi language, there is multiple meaning of the one word when we are talking. In Marathi Language there are multiple pronunciation of one word buts its meaning is different. For that word is difficult calculate accuracy of Marathi word. For that we refer Hindi language paer. Charu Nanda, Mohit Dua, Garima Nanda proposed Sentimental Analysis for Movie Reviews in Hindi Language using Machine Learning. In this paper an approach to sentiment Analysis on movie review in hindi language is discussed for social websites like facebook, twitter are widely posting the user review about different thing such as movie, food, fashion etc. Review and opinion play a role in identifying the level of satisfaction of user [3].

Mohammed Arshad Ansari, Sharavari Govilkar, proposed Sentiment Analysis of mixed code. In that they transliterated Hindi and Marathi text The designed system is an effort which classifies Hindi as well as Marathi text transliterated documents automatically using KNN, NB and SVM and ontology based classification; and results are compared to in order to decide which methodology is better

suited in handling of these documents [4]. Mohd Sanad Zaki Rizvi Article on 3 Different NLP Library for Indian Language Analytics Vidya Article. In that they discuss about the different types of NLP library and how it works. This Article for indian language[5]. Parul Sharma and Teng-Sheng Moh proposed Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter in that they used SVM, NB for prediction [6]. Binita Verma, Ramjeevan Singh Thakur proposed Sentiment Analysis using Lexicon and Machine Learning Based Approach [7]. Sentiment Analysis of Urdu Tweets is done in this paper. By using different Lexicon based and machine learning technique. Tweets are classified into positive and negative [8].

III. PROPOSED METHODOLOGY

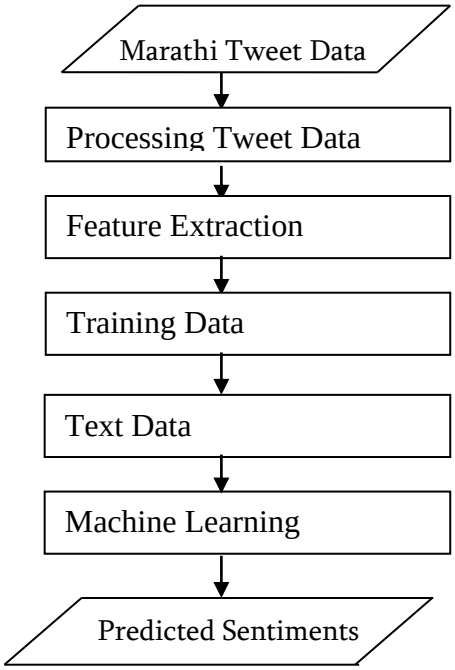

Fig 1. Proposed Architecture for SA

This section explains about the Dataset, Preprocessing Methods, Feature Selection, and Classifiers used in our experimental setup. The architecture of the proposed method is shown in Fig. 1.

3.1. Dataset

Due to the unavailability of the sentiment tagged dataset in Marathi, we have created the dataset in Marathi Language. Then used that data for identifying it is positive, negative and neutral.

3.2. Preprocessing

The retrieved tweets contain hyperlinks, punctuations, special characters, etc., these have been removed using regular expressions in python language. After that manually verified that data and assigned it is positive , negative or neutral. For that first we input data manually. And then doing tokenization process on the input data and then remove stopwords and special symbols of the input data.

3.3. Feature selection

Feature selection is the procedure of reducing the number of input variables when progressing a predictive model. It is used to calculate accuracy of input data set. BOW, TF-IDF, Unigram with sentiwordnet and Unigram with Sentiwordnet with negation word have been considered for feature vector formation of the input data set.

- BOW: In BOW, the text is transformed into a bag of words where each entry corresponds to the number of occurrences of a particular term in the sentence. The feature matrix is created with m * n dimension where m is the number of sentences and n is the number of unique words in the corpus.

3.4. Machine Learning Approach

Machine learning is an application of artificial intelligence (AI) that gives systems the ability to automatically learn and better from experience cut off being direct programmed. NLP is the one of the part of machine learning language. In that we used NLP library for identify sentiments for Marathi language. The selection of hyper parameter is most challenging for accurate prediction of data.

- NLP

Natural Language Processing (NLP) allows machines to break down and interpret human language. NLP is a field in machine learning with the ability of a computer to understand , analyze, manipulate and potentially generate human language.

The use of algorithms to determine properties of natural, human language so that computers can

understand what humans have written or said. NLP includes teaching computer systems how to extract data from bodies of written text, translate from one language to another, and recognize printed or handwritten words. Notably, NLP is the field that allows for our everyday use of virtual assistants.
The best known library for NLP is,

- iNLTK.

iNLTK stands for Natural Language Toolkit for Indic Language iNLTK, an open-source NLP library consisting of pre-trained language models and out-of-the-box support for Data Augmentation, Textual Similarity, Sentence embeddings, Word embeddings, Tokenization and Text Generation in 13 Indic Languages. This identification is done based on this library. For identification this language they give the various types of code for every language for example Marathi “mr”, hindi “hi”, etc. by using this toolkit we doing process on input data.

3.5. Fornulation

After this all process we calculate the polarity of that data it means it is positive, Negative or Neutral. For that we used percentage formula. And then we show it in the form of graph. For calculating polarity we used percentage formula. For that first we calculate total number of word for that used

total no. words = (no. positive words + no. negative words +no. neutral words)

Then calculate polarity,

$$\text{Percent positive} = \frac{\text{No. positive word}}{\text{Total no. word}} \times 100$$

$$\text{Percent negative} = \frac{\text{No. negative word}}{\text{Total no. word}} \times 100$$

$$\text{Percent neutral} = \frac{\text{No. neutral word}}{\text{Total no. word}} \times 100$$

IV. RESULT AND DESCUSSION

If we input Marathi data,

Fig. 2 screenshot of input data

In that we input data in Marathi language.
Then we get this type of output,

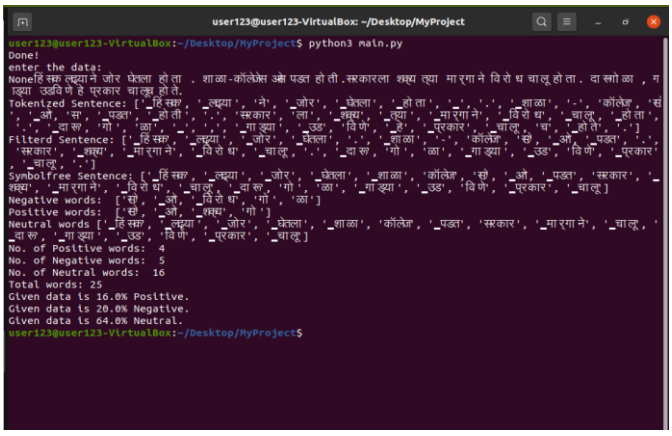

Fig. 3 screenshot of output data

In this screenshot it shows the output of that Marathi input data. It shows the input data is positive, negative or neutral. And also shows the polarity of input data.

V. CONCLUSION

Sentiment Analysis of Marathi tweets using machine learning algorithm such as NLP are proposed in this work. Different feature selection method are considered for feature vector formation in the input data. And shows the better higher accuracy of Marathi tweet data.

VI. ACKNOWLEDGEMENT

We would like to thank many people for designing on identification of sentiments by using machine learning algorithms for different language. A special thanks to Dr. A. N. Thakare for help with

coordinating across different time zone and discussion with topic.

VII. REFERENCES

[1]. Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, Thumbs up?: senti ment classification using machine learning techniques, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10, Association for Computational Linguistics, 2002. A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in Workshop on Languages in Social Media, Stroudsburg, PA, USA, 23 June 2011, pp. 30–38.

[2]. Soumya S., Pramod K.V "Sentiment analysis of Malayalam tweets using machine learning techniques" classified into positive and negative using different machine learning algorithms, IEEE, April 2020.

[3]. Charu Nanda, Mohit Dua, Garima Nanda, Sentimental Analysis pf Movie Reviews in Hindi Language using Machine Learning, 2018 International Conference on Communication and Signal Processing (ICCSP), 1069-1072, 2018.

[4]. Mohammed Arshad Ansari, Sharavari Govilkar, Sentiment Analysis of mixed code for the transliterated Hindi and Marathi text, international journal on Natural Language computing (IJNLC) Vol7, 2018.

[5]. Mohd Sanad Zaki Rizvi, "3 Different NLP Library for Indian Language", Jan 23, 2020, Analytics Vidya Article.

[6]. Parul Sharma and Teng-Sheng Moh "Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter", 2016 IEEE International Conference on Big Data.

[7]. Binita Verma, Ramjeevan Singh Thakur, "Sentiment Analysis using Lexicon and Machine Learning Based Approach", Springer, 2020.

[8]. Zarmeen Nasim, Sayeed Ghani, "Sentiment Analysis on Urdu Tweets using Markov Chain", Springer, 2020.

Cite this article as :

Renuka Ashokrao Naukarkar, Dr. A. N. Thakare, "A Design on Recognization of Sentiment Analysis of Marathi Tweets using Natural Language Processing", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 8 Issue 2, pp. 451-455, March-April 2021.

Journal URL : https://ijsrst.com/IJSRST218289