International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 10 | Oct 2021 www.irjet.net p-ISSN: 2395-0072

A Study on Sentiment Analysis of Twitter Data for Devanagari Languages

Nirmiti Bhoir[1], Aarushi Das[2], Mrunmayee Jakate[3] Snehal Lavangare[4], Deepali Kadam[5]

[1,2,3,4]*Student, Information Technology, Datta Meghe College of Engineering Airoli, India*
[5]*Asst. Professor, Information Technology, Datta Meghe College of Engineering Airoli, India*

Abstract - *Sentiment Analysis is a process of identifying the emotion of a sentence. Sentiment Analysis has various applications like Social Media Monitoring, Customer Feedback, Brand Monitoring, etc. One of these applications of Social Media Monitoring will be implemented in our project. We will be implementing Sentiment Analysis on the social media platform - Twitter. Most of the work in Sentiment Analysis has been done in the English language. As a result, less work has been done in regional languages. Therefore we decided to perform Sentiment Analysis in the Marathi language, which is the official language of Maharashtra. This paper will be focussing on the Lexicon approach to perform Sentiment Analysis.*

***Keywords*: Sentiment Analysis, Devanagari Language, Marathi Language, Lexicon Approach, Twitter, Tweets**

1. INTRODUCTION

The number of social media users is increasing on a daily basis and the use of social platforms has become a common thing among people. Social media is a platform where people can connect, have conversations, share views and information, and create content on the internet. One of the most popular social networking sites is Twitter. Twitter is an American microblogging and social networking service on which users post and interact with messages known as "tweets". These tweets can contain text, GIFs, images, and videos. The posts made from your account are visible to your followers and can be found in the search bar.

Users on Twitter can also retweet and privately chat with their followers. By default, the tweets are public in nature and can be viewed by everyone, but a user can make them private by changing their settings. Users on Twitter can tweet about anything, right from giving updates about the number of covid patients as in the pandemic to posting memes or a joke. It is surprising to know that every single

day around 500 million tweets are posted. This shows us the power of social media and its usage. Social media criticism and negative comments are a norm rather than an exception. With thousands of users connected to social media accounts, negative comments are unavoidable. On average around 25 % of users are subjected to social media trolling. So there is an urgent need to solve this issue and reduce its effectiveness. Twitter has an existing feature of banning/reporting users under its terms and conditions.

There are 121 languages spoken in India and more than 19500 local languages or dialects. This is only about our country. If we go on exploring the entire world then the number will be shocking. Users convey their judgments and feelings in different languages and English isn't everyone's native language so it becomes a challenge to analyze people's opinions on particular issues.

Marathi is a language spoken in India and is the co-official language in Maharashtra and Goa which are the states of Western India. Balshastri Jambhekar is the father of the Marathi language. It belongs to the Indo-Aryan language family. It originated from Maharashtri Prakrit which is said by a lot of people. Marathi has the third-largest number of native speakers in India and ranks 10th in languages spoken with most native speakers in the world.

Twitter has localized the web UI experience in 7 Indian regional languages like Hindi, Gujarati, Marathi, Urdu, Tamil, Bengali, and Kannada. Twitter has 192 million users (as of 2021) out of which many users post in the Marathi language.

Giant e-commerce websites like Amazon ,Myntra, Flipkart, etc rate the products on their websites using customer reviews. There are over 100 million users, but they don't have the time to go through each review. So the study of these reviews is done using Sentiment Analysis.

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative, or neutral. A primary sentiment analysis system draws on a sentiment library to understand. Sentiment libraries are very large collections of adjectives such as (good, wonderful, awful, horrible) and phrases (good game, wonderful story, awful performance, horrible show) that have been listed by human coders. There are two ways to analyze the sentiments of the tweets.

a) Natural Language Processing Approach(NLP)

The three common sentiment labels: positive, neutral, and negative. The mixed Sentiment label exists in tweets that have two different meanings.

b) Machine Learning(ML)

It is a scientific discipline that explores the construction and the study of algorithms that can learn from data.

Some of the applications of Sentiment Analysis are as follows:

Social Media Monitoring – Traversing through all that data in minutes, to analyze individual emotions and overall public sentiment on every social platform.

Customer support - Using NLP for reading regular human language and understanding their meaning, emotion, tone related to any customer request.

Customer Feedback - Gaining insights through customer feedback that are available online and also work on specific user issues related to the same.

Brand Monitoring and Reputation Management - Bad or negative brand reviews or mentions will get notified easily, also one can keep track of their brand's image and reputation by transforming the data into a statistical one.

Market and Competitor research - Analyze one's competitors and find out who's trending or drifting amongst them. Also, understand the strong and weak points of your venture.

Our proposed system will be analyzing the sentiment of the Marathi tweets posted on our twitter-clone. To accomplish sentiment analysis of Marathi tweets we are going to use a lexicon-based approach.

2. LITERATURE SURVEY

2.1 A Study on Sentiment Analysis Techniques of Twitter Data [1]

In 2019 Abdullah Alsaeedi, Mohammad Zubair Khan presented a research paper on A Study on Sentiment Analysis Techniques of Twitter Data [1]. Their approach focuses on using diverse techniques for Twitter sentiment analysis like Supervised Machine Learning Approaches, Ensemble Approaches, Lexicon-based Approaches (Unsupervised Methods), Hybrid Methods. In this paper, they had difficulty in deciding the best approach for detecting sentiment.

2.2 Lexicon-Based Approach to Sentiment Analysis of Tweets Using R Language [2]

This research paper was published by Nitika Nigam and Divakar Yadav 2018 on a Lexicon-based approach to Sentiment Analysis of tweets using R language [2]. They made dictionaries that consist of adjective words and then the input sentence tokens are compared with that dictionary.

They faced some issues like

i) The context-based dependent words,

ii) The combination of multiple opinion words in one sentence in this approach.

The second approach was a holistic lexicon-based method within which they made a dictionary consisting of words with their polarity depending on the strength of the word. Their third approach focused on identifying the cynicism on Twitter in which they used a pattern-based method. They also overcame the problem of polarity shift detection by using a model called dual sentiment analysis.

2.3 Sentiment Analysis in Marathi Language [3]

In 2017 Snehal V Pawar and Prof Swati Mali presented a paper on Sentiment Analysis in Marathi Language [3]. This paper emphasized the methods of analyzing the sentiments of Marathi sentences. There are 2 methods to find the polarity of a sentence- using Machine learning and the Lexicon approach. Machine learning works with different algorithms and in Lexicon Approach we make thesaurus of words having prebuilt sentiments(negative and positive). The lexicon approach is further divided into corpus and dictionary forms. The performance of this model is good.

2.4 Sentiment Analysis in Marathi using Marathi WordNet[4]

In 2017 Chitra V Chaudhari, Ashwini V. Khaire, Rashmi R. Murtadak, Komal S. Sirsulla presented a paper on Sentiment

Analysis in Marathi using Marathi wordnet [4]. Their approach focussed on using the Gate processor (Natural Language Processor)to find out the sentiment of the overall document which can consist of multiple sentences using GATE(General Architecture for Text Engineering) which also performs Natural Language Processing(NLP) operations.The NLP approach used in this model can work less efficiently if used with grammatically incorrect sentences. Therefore the efficiency of this model is great.

2.5 A Literature Review on Twitter Data Analysis [5]

This Manuscript was published by Hana Anber, Akram Salah, A.A.Abd El-Aziz on December 31, 2015, which was accepted on June 8, 2016. In this paper, the authors review different information analysis techniques; starting with the analysis of different hashtags, twitter's network topology, events spread over the network, identification of influence, and finally analysis of sentiment. Literature Review was done where Datasets problem was solved and data retrieval was done Twitter-API. Repeatability was studied by deploying two different features: the Content (URL and hashtags), and the Contextual feature from 74 million tweets. There are two ways to analyze the sentiments of the tweets Natural Language Processing Approach(NLP) . And Machine Learning approach where the study of algorithms that can learn from data.

2.6 Sentiment Analysis on Hindi Content : A Survey [6]

In December 2015 Mukesh Yadav, Varunakshi Bhojane presented a paper on Sentiment Analysis on Hindi Content: A Survey [6]. Their approach focuses on In-language Sentiment Analysis, Machine Translation(MT)-based Sentiment Analysis, Resource-based Sentiment Analysis. They stored a 300 sentences dataset in XML and sentence-level sentiment is adopted. In the Data preprocessing stage, Deep belief architecture is divided into 2 stages: the pre-training model & fine-tuning step. The first restriction of their research is that the sentences are marked incorrectly as negative or positive even though they are not. Negation handling is missing and the second restriction is that a subjective lexicon can be developed for the unexplored languages which do not have wordnet. According to them, work in the lexicon approach can be extended to incorporate word sense disambiguation (WSD)

NAME	AUTHOR	YEAR	SELECTED FEATURE
Sentiment Analysis Techniques of Twitter Data	Abdullah Alsaeedi, Mohammad Zubair Khan	2019	Study of Twitter sentiment Analysis
Lexicon-Based Approach to Sentiment Analysis of Tweets Using R Language	Nitika Nigam, Divakar Yadav	2018	Lexicon Approach and Calculation of sentiment score
Sentiment Analysis in Marathi Language	Snehal V Pawar , Prof Swati Mali	2017	Find Polarity of the sentence
Sentiment Analysis in Marathi using Marathi WordNet	Chitra V Chaudhari, Ashwini V. Khaire,Rashmi R. Murtadak, Komal S. Sirsulla	2017	Extraction functionality
A Literature Review on Twitter Data Analysis	Hana Anber, Akram Salah, A.A.Abd El-Aziz	2015	NLP Approach, Machine Learning Approach
Sentiment Analysis on Hindi Content	Mukesh Yadav, Varunakshi Bhojane	2015	Language Sentiment Analysis

Table 1: Literature Survey in Tabular format

3. PROPOSED SYSTEM

There are two techniques to find the sentiment of sentences - The Machine Learning approach and the Lexicon-based approach. In our proposed system we will be using the lexicon-based approach.

We will be building a Twitter-clone to implement Sentiment Analysis. The vocabulary of a group of people, language, or field is called a lexicon. Items within a lexicon are called lexemes, and groups of lexemes are called lemmas, which is a unit used to represent the size of a lexicon. Our methodology involves the collection of the negative and positive words in the Marathi language and storing them in a list/array. The main aim of our project is to analyze the tweets posted by the users on our twitter-clone. According to the sentiment of a tweet, the appropriate action will be taken by the admin.

The Twitter clone will analyze the sentiment of every tweet posted by each user and calculate its score. The admin of the Twitter clone can check the sentiment scores of all the users.

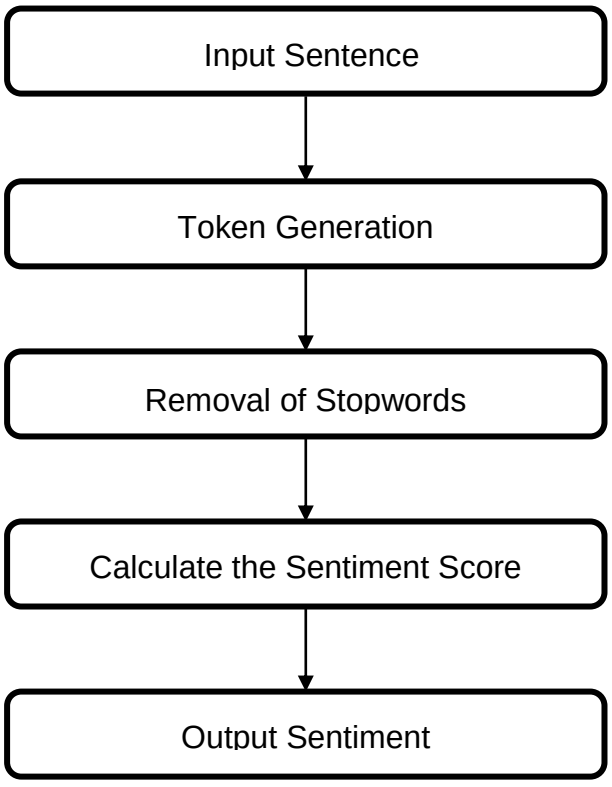

Figure 1: Flowchart of Proposed System

3.1 Collection of Dataset:

Tweets will be collected from Twitter API to evaluate the performance of our model.

3.2 Collection of Positive, Negative, and stopwords:

As there are no predefined packages for lexicons in the Marathi Language, we have manually collected around 100 words from each category of positive, negative, and stopwords. Some adverbs are also collected which can change the sentiment score of the sentence. For example- **खूप वाईट** will have a greater negative score than **वाईट** when used in a sentence. These words have been collected from various sites and some from public Github repositories.

Examples

- Positive Words: चांगले,उत्कृष्ट,अनुकूल, अद्वितीय, प्रेमळ
- Negative Words: कृतघ्न,मूर्ख,हरामखोर,हलकट,बेशरम
- Stopwords: परवा,लगेच,दिवसभर,वर, खाली

3.3 Removal of stopwords:

Stopwords are words in a stop list that are removed before calculating the sentiment score of a sentence. These stopwords can also contain punctuation marks, URL or hashtags, special characters, or digits which do not contribute to any sentiment and hence should be removed.

3.4 Calculation of Sentiment Score:

The Score of each tweet will be calculated in this way:

Step 1) Input sentence is broken down into tokens

Step 2) Stopwords from each sentence are removed

Step 3) Each word in the sentence is mapped with the positive and negative words list.

Step 4) Every positive and negative word is associated with a numeric value that represents its score. Positive words have positive values and negative words have negative values.

Step 5) According to the following formula

$$\textbf{Sentiment Score(avg)}=\frac{1}{n}\sum_{i=1}^{i=n} S_i$$

Step 6) This formula calculates the average sentiment score of the sentence giving the result in the range of -1 to +1

Step 7) -1 indicates a strong negative sentence,+1 indicates a positive sentence, and 0 specifies a neutral sentiment

3.5 Classification of Tweets:

The score will be presented in a percentage format indicating how positive or negative a tweet is. There are 3 main sentiments to be analyzed - positive, negative, and neutral. The negative sentiment will be categorized as depressing and abusive tweets.

Sometimes users express their feelings of loneliness, or sadness online through social media platforms, but rarely do they get any support from it, meaning followers of that user might help by communicating with that person but a strong helping resource like any mental health organizations contact point is not provided by the social media platform. Abusive tweets are the ones that have more cuss words or hateful words used in the sentence.

The web-based twitter-clone will allow the admin to disable/report the users posting abusive tweets frequently. The admin can also provide valuable resources to the user posting gloomy tweets by providing links in the user's search bar. So that the user can communicate with the help if they wish to.

4. RESULTS

In the following diagram a example is explained using the above formula

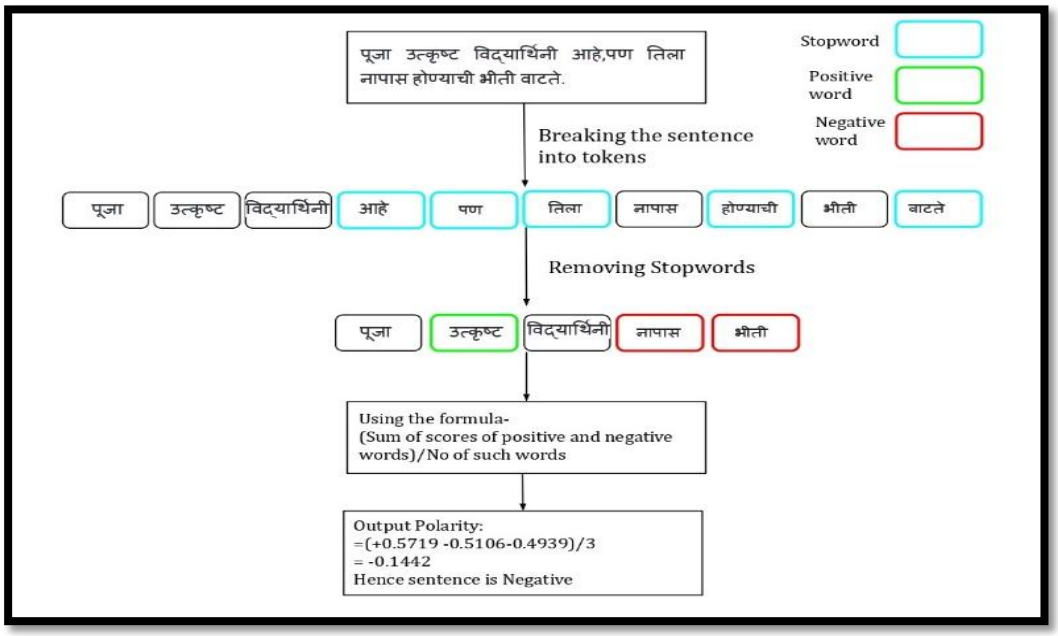

Figure 2: Example

5. DIFFICULTIES FACED IN SENTIMENT ANALYSIS

Sarcasm is difficult to detect in a sentence. In sentiment analysis, sarcasm plays a vital role. It may consist of positive or negative words but the meaning of the sentence differs a lot as it doesn't seem to be it. To avoid the misunderstanding of sentences that indicate sarcasm it is important to Detect sarcasm sentences to extract the exact meaning of the sentence. But it comes as a challenge from a Technical perspective. Example: तू वेडी आहेस का? Meaning- Are you mad? This sentence can be used as a sarcastic remark in some situations. Hence it is difficult to understand the actual sentiment of the sentence.

Anaphora can be a problem while analyzing the sentiment of the tweet. The repetition of words can

create confusion about whether the words are nouns or pronouns. Example: आम्ही नाटक बघायला गेलो आणि नंतर जेवण केले, ते खूप खराब होते meaning-We went to watch a movie and then had dinner, it was awful. Here खराब(Awful) refers to which part of the sentence?

Hybrid Language can be used by some of the users, for example, "तू आज स्कूल ला जाणार आहेस का? " meaning "Are you going to school today?" Here"स्कूल" (School) is an English word written in Marathi Script and hence this sentence is a combination of Marathi and English dialects.

The hybrid language is not considered by us for sentiment analysis. As it will consume more time in preparation for a hybrid dictionary for hybrid language. Our set of dictionaries does not contain such hybrid words. Thus, the hybrid language used in tweets won't be analyzed.

Sentiment lexicon does not contain words that are expressed via emoticons. There are many words available in the dictionary, for the English language. But many times tweets could be expressed in the form of emoticons, words, slang words, etc. and all these words in the lexicon could be difficult. Examples: अबे,खालीफुकट, राव,मोट्या

Polarity of a particular word or sentence is neutral or positive or negative. But how positive or negative the sentence is is another question. "सर्वात वाईट'' and "दुष्ट" both are negative but the second one has a stronger sentiment than the first one.

6. CONCLUSION

Different research papers were studied and understood how various techniques work and how they affect sentiment analysis under different conditions.

An enormous amount of work in sentiment analysis of Twitter has been done in English language. That's why we tried to fulfill the need for sentiment analysis in the Marathi language. This concludes that not only the techniques but the resources are also more important for better results. Also, other algorithms and techniques will be implemented to increase efficiency. According to our research, the words which are not available in our database are considered as neutral though it is a sentiment word. Therefore the database should be as rich as possible for better results.

This project is still under implementation.

REFERENCES

[1] Abdullah Alsaeedi, Mohammad Zubair Khan, "A Study on Sentiment Analysis Techniques of Twitter Data". International Journal of Advanced Computer Science and Application. Vol.10,No.2,2019.

[2] Nitika Nigam,Divakar Yadav, "Lexicon-Based Approach to Sentiment Analysis of Tweets Using R Language". Advances in Computing and Data Sciences (pp.154-164),2018

[3] Snehal V. Pawar, Prof. Swati Mali, "Sentiment Analysis in Marathi Language". International Journal on Recent and Innovation Trends in Computing and Communication. Volume: 5 Issue: 8.

[4] 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.

[5] Hana Anber, Akram Salah, A. A. Abd El-Aziz, "A Literature Review on Twitter Data Analysis". Volume 6 , Number 3, June 2016

[6] Mukesh Yadav, Varunakshi Bhojane, "Sentiment Analysis on Hindi Content : A Survey". International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 4, Issue 12 December 2015.