A NOVEL APPROACH TO DETECT SENTIMENTS BASED ON SOCIAL MEDIA USING MACHINE LEARNING

Abstract

It might be helpful to estimate a product or service's future scope by doing a sentiment analysis of it. However, it might be tedious and difficult to manually analyse a lot of papers in a short amount of time. As a result, several attempts have been made to address this issue in the literature, and various sentiment analysis approaches have been put forth. Various machine learning methods that use supervised or semi-supervised techniques are currently available. These algorithms can use a hybrid, unigram, bigram, n-gram, or other suitable approach. This study makes advantage of semi-supervised learning. This study combines many approaches to create a unique model that employs a number of approaches and yields results. This fictional method produced greater result. People frequently post a lot of stuff on social media with the goal of discovering memes that capture their emotions. Because people are more inclined to convey their feelings through pictures and written descriptions, image and textual sentiment analysis (SA) is advancing at a rapid rate. Social media users are increasingly expressing themselves and sharing their experiences through photos and videos.

***Key expressions*:** Sentiment-Analysis**,** Emoticons, Social Media, OCR, Image, Natural language Processing, Machine Learning

Authors

Neetesh kumar Nema
Research Scholar Dr. C. V. Raman University Bilaspur, C.G.

Bharat Choudhary
Department of Computer Science and Engineering, LCIT Bilaspur.
hi.bharat2002@gmail.com

Dileshwar Patel
Department of Computer Science and Engineering, LCIT Bilaspur.

I. INTRODUCTION

Nowadays in field of computer science and technology the hottest research area is sentiment analysis. Sentiment analysis can be defined as technique/method of identifying the view of people, given in the form of text regarding to a specific object (event, individual, decision, change etc.). Other synonyms of sentiment analysis are opinion mining, confidence analysis, people attitude towards an object, deriving opinion etc. The main reason behind the popularity of sentiment analysis it gives us overview of wide spread public opinion/thinking related to a topic.

Sentiment associated to a particular object is categorized in one of the following category:

- Positive
- Negative
- Neutral

Sentiment analysis is used at multilevel, it can be used to identify sentiment hidden in a document, to be more précised the analysis can be used to calculate the sentiment associated with each paragraph or may be each line.

The basic method adopted to identify the overall sentiment score is tokenization of each sentence in document into expressions, further each expression is categorized into positive, negative and neutral expressions. In next step further, these expressions are categorized according to associated impact (for example: extremely happy is having more impact than happy), finally summation of numbers of positive and negative expression is done to find out overall sentiment score.

The major challenges in sentiment analysis are:

- Multilingual
- Sarcasm
- Emoticons handling
- Natural language processing overheads

To increase the accuracy and overcome challenges of sentiment analysis, there are advance sentiment analysis mechanism which incorporates the sarcasm and emoticons handling technique. Moreover, nowadays natural language processing software's (such as Open NLP by Apache) are also used in the process of semantic analysis.

Example: I am very glad with India's succeed in hockey match.

If we tokenize the given expression there are two positive expressions: glad and succeed, and there is no negative expression, so it will produce overall positive sentiment score

Natural Language Processing (NLP) is involved in sentiment analysis, which ensures with the computational operations of belief, emotion, objectivity, and indifference inside the provided text. Sentiment encompasses thoughts, judgements, actions, feelings, and other things. Social Network Analysis (SNA) has grown to be a vital tool for experts and analysts in social computing as the number of users and usage of social media is rapidly growing.

Sentiment categorization is the process of categorising texts according to the polarity of their expressed sentiments. A product or service's sentiment analysis might be useful in estimating its future market potential. However, it may be tedious and taxing to manually analyse a lot of papers in a short amount of time. The topic of memes on social media platforms like Facebook, Instagram, and Twitter has gained a lot of attention in recent years as the Internet has gained full respect. Over the past 10 years, memes have proliferated on the internet, frequently via social media platforms and mainly for amusing purposes. This trend has made it easier to analyse people's thoughts, feelings, and views using social media images using machine learning (ML), sentiment analysis (SA), emotion analysis, or opinion mining.

Facial expression recognition is one of the most crucial factors for human expression beings to read their feelings and intentions in touch. With their ability to convey a variety of information and emotional meaning, facial expressions have emerged as one of the most significant information networks in interactive communication. The goal of visual SA is to determine the feeling that a picture evokes.

Sentiment analysis is a type of opinion mining that is used to identify text on the web. The only goal is to hear the actual opinions of consumers about a certain product, service, film, news item, or topic. Sentiment analysis may be carried out on many different levels, including the phrase, document, and object or attribute levels. A person's attitude may serve as his or her assessment on a given invention. Because the majority of people buy or sell things online, feedback is crucial for both consumers and producers. Before making a purchase, some shoppers may want to see what other customers have to say about the product.

II. RELATED WORK

Essentially, Sentiment Analysis is used to articulate individual person’s sentiment. According to current state of the art sentiment analysis is used to categorize sentiments into two categories positive and negative. Some works classified them into as positive, negative and also in one more group as neutral.

C. Hauff et al. offers instructions on how to handle negation words like not, no, neither, couldn't, etc. in sentences. Even when there are negative phrases in a statement, the sentence may nevertheless have a good connotation.

A. Neviarouskaya et al uses To undertake fine-grained phrase categorization, there are 10 categories—nine emotional ('angry', 'disgust', 'fear', 'guilt', 'interest', 'joy','sadness' ('distress'),'shame', and'surprise') and one neutral.

Anurag P. Jain et al. Using Twitter API v. 1.1, they were able to retrieve information on political news.The technique utilised to forecast people's general predisposition towards political problems and circumstances is described in this study. After preprocessing the uncooked tales, two information sets—the training information set and the testing information set—are created according to the approach described. Additionally, writers used sentiment analysis to create a model that divides tweets into three categories: good, negative, and neutral.

Antonio Teixeira et al. are likewise developing sentiment analysis using data from Facebook. This paper's authors' primary goal is to describe how Facebook information extraction, information preparation, and sentiment analysis are carried out (using free and open source software).

Rincy Jose et al. provided a method for doing sentiment analysis on tweets using SentiExpressionNet, ExpressionNet, and Expression Sense Disambiguation as lexical resources. They added negation handling to the preprocessing stage for increased accuracy. Information gathering, preprocessing, and sentiment categorization are the three processes they included in their APPROACH presentation.

J.M. Weibe et al. the researcher brings out different algorithms in best identification of sentiment analysis.

M. A. Hearst et al. had come up with adding intelligence to sentiment analysis. Different machine learning methods are being used by researchers.

V. Suresh et al. presented an approach that used stop expressions and gaps between stop expressions as the feature for sentiment analysis.

Murthy G et al. made a comparative study on sentences and web context based sentiments.

Pang et al. suggested with unigram approaches in their research work.

Dave et al. used a tool to synthesize reviews.

III. PROBLEM DEFINITION

To understand people's thoughts and feelings based on their proposed text. It gives an overview of the different sentiments classification approaches and tools used for sentiment analysis. The machine learning approach is used for predicating the polarity of sentiments based on trained data sets. In this study automated analysis of social media is accomplished by building predictive model.

IV. OBJECTIVE OF RESEARCH

The main goal of this study issue is to identify the feelings and views of the clients or users through text. Analyze social media data to look for any patterns that might reproduce depressive symptoms in relevant social users. The best way to spot sadness in social media comments is by using machine learning techniques. Determine the elements to search for in social annotations to diagnose depression. Enhance the performance by removing additional sorts of emotional elements by employing a different expertise. Utilizing efficiency, effectiveness, and authenticity, the suggested approach is improved.

V. PROPOSED METHODOLOGY

R studio (IDE) will be utilized in the suggested technique for information analysis. Four main steps make up the proposed approach for analysis:

a. Gathering data from the target person's social media accounts.
b. Information preparation.
c. Analyzing the data to discover underlying trends and patterns
d. Outline the various Knowledge Patterns

A. Collection of information of target figure from his/her social media account: Basically, this methodology is focusing on extracting information from two platforms-

1. **Facebook:** Using the Facebook APP is the easiest approach to connect to the Facebook API and get the data. These are the stages that go into it:
   a. Install the necessary packages in R Studio.
   b. Build an application on the Facebook developer platform
   c. Create a connection using the app authentication key between R studio and Facebook.
   d. After a connection has been successfully established, obtain the necessary data.
   e. The Facebook Graph API and other third-party tools are only two of the many alternative techniques for information extraction that are accessible.

2. **Instagram:** In order to extract the information from Instagram we can use instagram App. Following are steps involved in this-

a. To support the Instagram connection, install and load the necessary packages in R studio.
b. Create an Instagram app and obtain the Consumer key (API key) and Secret key (API Secret).
c. Using the authentication key specified in the preceding step, connect the Instagram app to R studio.
d. Take the necessary information after successful authentication.

B. Preprocessing of information - Initial phase helps in removal of all other information present in all other languages except the English language. For the next level of preprocessing first we need to load the target file in a user defined object. Further the target information columns are loaded into the information corpus (collection of documents containing (natural language) text. In next step the operations are carried out on the information corpus to clean the information.

Figure shown below presents the complete method of preprocessing-

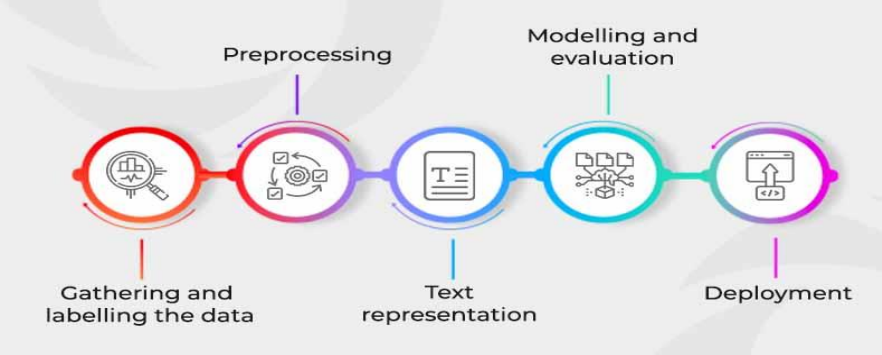

Figure 1: Sentimental Analysis Process

Level 2 of preprocessing of information basically involves:

1. **Removing punctuations:** English language is supported by different punctuation marks such as dot (.), coma (,) etc. However, punctuations are meaningless whenever we have to perform analysis, so it is become important to remove the punctuation marks.

2. **Removing white spaces:** It might be possible that extracted text contain unwanted whitespaces, which may act as noisy information during analysis. For better results, it is advisable to remove the white spaces.

3. **Converting all the text into lower case:** Most of the analysis/mining code treat are case sensitive, so to reduce errors it is advisable to have our all our text in same case.

4. Remove the stop expressions of English-Stop expressions are comprised of general expressions which is to support our sentence such as I, me, my, do, should etc. However, these are not important from the information analysis point of view, so it is advisable to remove such expressions.

5. **Stemming of the expressions:** For the analysis purpose it is important to convert all the expressions to base expressions such as played is converted to play.

6. **Removing other unwanted expressions or symbols:** It might be the case that we want to remove certain targeted expression from the text file for more specific results.

C. Mining the information to find hidden pattern and trends

In our proposed methodology we are basically focusing on analysis for the purpose of feature extraction of target candidate and confidence of people related to the candidate. So for the general analysis we can use the Information Corpus and Document term matrix for presenting the generating the knowledge based on the frequency of expressions such as expression cloud and 'n grams'.

For more depth analysis one need to perform natural language analysis. There are certain open sources libraries are available that we can integrate with R studio such as OpenNLP (provided by Apache). The basic aim of semantic analysis is to generate a plot that represents the score of a candidate on the basis of kind of expression used in his/her posts/stories. Additional we can identify confidence of a candidate among people with the help of semantic analysis of people comment and stories information.

To find out the confidence score the people opinion information can be broken down into single expressions (tokens). For the purpose of converting the text document into stream of tokens some natural language processing tool is required such as OpenNLP by Apache. Further with the assistance of NLP tool expression can be categorized into a scale (-5 [very negative expression] to +5 [very positive expression]) as shown in Figure

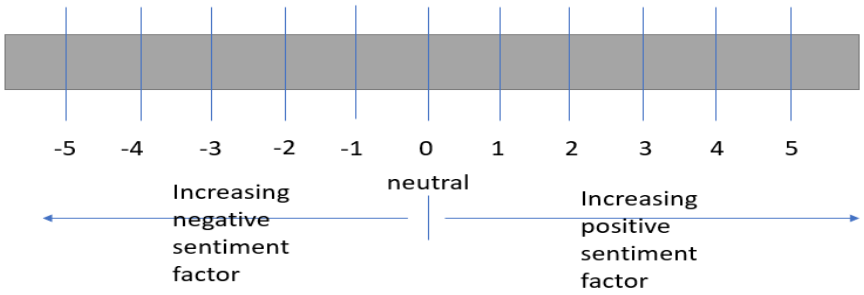

Figure 2: Proposed scale for calculate sentiment information for each expression

D. Presenting the Knowledge: To present the knowledge different graphical structures can be used-

1. Expression Cloud is an image comprised of expressions; in which size of expression depend on the frequency of expression in the document.
2. N- gram in Corpus is to present the group of expressions which are used together.

5.1 Phases of the Methodology

1. **Information collection**: The act of obtaining and analysing data from a variety of sources with an eye towards answering questions, testing hypotheses, and assessing outcomes is known as information collection. The customer reviews will be used to gather the information.

2. **Identification of information**: It identifies the information according to its value and what we are going to use. After information has been identified it will be given as an input to the system.

3. **Pre-processing:** Pre-processing is done to the customer's viewpoint to remove any extraneous or irrelevant language. Our approach simply works with the description of each review's speech component; processing entails dividing reviews into phrases in order to distinguish them.

4. **Part of Speech tagging**: It breaks down each sentence into its component parts of speech, indicating if a given term is a noun, verb, adjective, adverb, etc. It also recognises basic noun and verb groupings. Find the closest noun or noun phrase of the opinion expression as an uncommon feature when a sentence lacks any frequent features but contains one or more opinion expressions. Every statement of feedback is classified as a noun, adjective, or adverb using speech tagging.

5. **Negation detection**: It is also a crucial component in implementing sentiment analysis using term scores because negation would flip the opinion orientation of words like "funny" and "interesting," which are often used to describe things that are amusing or interesting.

6. **Stop expressions removal:** With the use of the Parts of Speech tagging approach, we delete terms like prepositions, numbers, articles, and nouns like "name of product" etc.

because their presence in the system has no functional value. It helps the tagged file's opinion statements and expressions be extracted more effectively.

7. **Rule-Based approach:** Rules must be defined in the Rule-Based method and comprise some specified relation that has an associated originator and outcome. In this process, certain rules must be established before the feelings may be examined or analyzed in light of those rules.

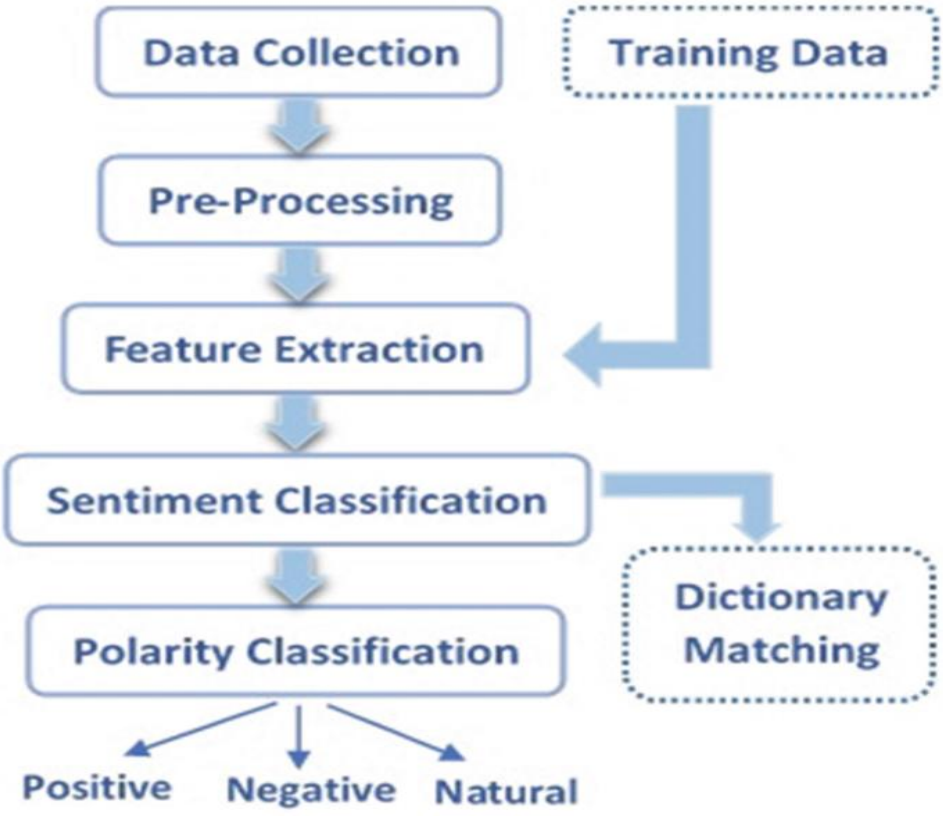

Figure 3: Sentiment analysis Process

VI. CONCLUSION AND FUTURE WORKS

The main motivation behind this research is to define the, we determine if the sentiment is good or negative in this study. In addition to its advantages, it aids in social media monitoring and provides public opinion on particular subjects. The prime outcome will be detection of depression in sentiments caused due to Social media. We will improve the performance of the system in sentiment analysis to detect depression. We will build model for finding the sadness, happiness, customer behavior in sentiment analysis. The result will be obtained in lesser time using proposed method. Proposed work can be beneficial for E-Commerce websites, Social media etc, enhancing the customer satisfaction. Since it can deliver more trustworthy signals and information for a number of data analytics activities using digital platforms for prediction, we believe sentiment categorization on sizable amounts of online user-generated content is advantageous. Because there are many negative memes

centered on racism, religion, politics, and terrorism and because a high text score does not always suggest a positive meme, no single model can analyze all genres in a trend.

REFERENCE

[1] Wiebe Janyce "Identifying subjective characters in narrative, Proceedings of the International Conference on Computational Linguistics (COLING-1990)."1990.
[2] Hearst M., 1992, Direction-based text interpretation as an information access refinement in TextBased Intelligent Systems, P. Jacobs, Editor 1992, Lawrence Erlbaum Associates, 257-274.
[3] V. Suresh 2011, A Non-syntactic Approach for Text Sentiment Classification with Stopwords, WWW 2011,March 28–April 1, 2011,Hyderabad, India.
[4] Murthy G. and Bing Liu, 2008, Mining opinions in comparative sentences, Proceedings of the 22nd international conference on computational linguistics (Coling 2008), Manchester, August 2008, 241248.
[5] B. Pang L. Lee, and S. Vaithyanathan., "sentiment classification using machine learning techniques" 2002.
[6] Matsumoto and Takamura, "syntax-based features construct parse trees", 2005.
[7] Dave, Lawrence &Pennock, "Opinion extraction and semantic classification of product reviews", 2003.
[8] Alena Neviarouskaya, Helmut Prendinger, Mitsuru Ishizuka, Affect Analysis Model: novel rule-based approach to affect sensing from text, Natural Language Engineering, Cambridge University, Vol. 17, pp. 95- 135, September 2010.
[9] C. Hauff, Dadvar, Maral and Jong de, Franciska, Scope of negation detection in sentiment analysis, Dutch-Belgian Information Retrieval Workshop, Netherlands, February 2011.
[10] Hao, F., Park, D., Pei, Z.: When social computing meets soft computing: opportunities and insights. Hum. Centric Comput. Inf. Sci. 8, 8 (2018).
[11] A. P. Jain and V. D. Katkar, "Sentiments analysis of Twitter data using data mining," 2015 Int. Conf. Inf. Process., pp. 807–810, 2015.
[12] A. Teixeira, "Data extraction and preparation to perform a The example of a Facebook fashion brand page."
[13] R. Jose, "Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation," no. November, pp. 638–641, 2015.

