American Journal of Electronics & Communication

Subhodeep Banerjee et. al., American Journal of Electronics & Communication, Vol. II (1), 16-24

A Novel Approach for Emotion Detection from Text Data using Natural Language Processing and Machine Learning

Authors: Subhodeep Banerjee*, Shrivasta Goswami*, Arnab Das*, Neeloy Saha *, Soumyashree Seth*, Sagnik Bhattacharya*, Dr. Sandip Mandal*, UEM Kolkata
{Email:sandy06.gcect@gmail.com}

Abstract *- Emotion can be expressed in many ways that can be seen such as facial expression and gestures, speech and by written text. Emotion Detection in text documents is essentially a content– based classification problem involving concepts from the domains of Natural Language Processing as well as Machine Learning. In this paper we are proposing a solution for emotion recognition based on textual data. The emotion expressed in a blog, review or any kind of textual content remains unused until the text is analyzed and the emotion is retrieved from the data. It is impossible to analyze the huge amount of data manually and gain information from it. So, we are proposing a Model that will analyze the data and make a prediction of the emotion embedded into the textual data.*

We are using algorithms from the domain of Natural Language Processing and Machine Learning. Initially we will take some text as input and in next step we perform tokenization to the input text. Words related to emotions will be identified in the next step afterwards analysis of the intensity of emotion words will be performed. Sentence is checked whether negation is involved in it or not, then finally an emotion class will be found as the required output. In short Emotion Detection is the most important field of research in human-computer interaction.

Keywords*: NLP, Emotion, Textual Data, sentiment analysis.*

INTRODUCTION- Emotion plays a very important role in day-to-day human life. According to human psychology we tend on repeating actions which make us happy and avoid the ones which make us feel sad. Theoretically if we explain, then emotion is a very complex psychological state which primarily involves three basic components: a subjective experience, a physical response and an expressive response.

Complexity- Emotions differ from person to person. If at some point a situation seems funny to someone or makes someone happy, the same thing could make someone sad. Complexity also refers to the phenomenon when we cannot feel ourselves, cannot detect our emotions, let alone knowing how someone else feels.

Response- Whenever we feel emotion, there is a physical response associated like sweating, palpitation, etc. There is also change in gestures and postures like crossing your arm or leg, frowning or smiling, etc.

As Robert Plutchik had proposed his famous drawing on the wheel of emotions to express his proposal in a graphical representation. It mainly consists of the basic eight bipolar emotions: sadness vs joy, disgust vs trust, fear vs anger and anticipation vs surprise. [1].

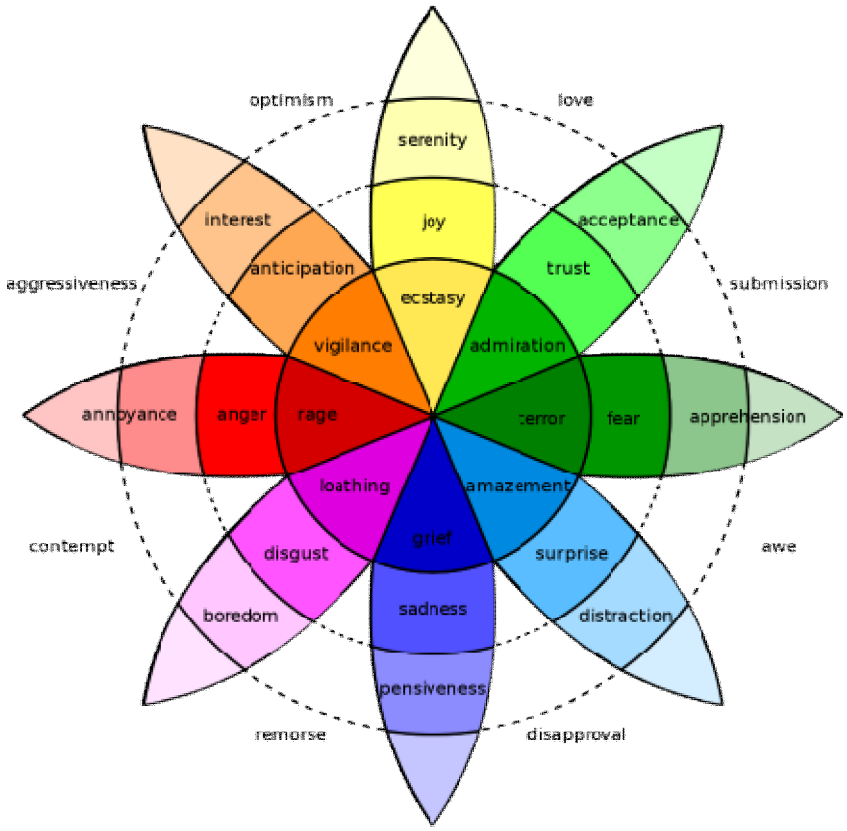


Fig. 1.

With the help of NATURAL LANGUAGE PROCESSING information on the subjective basis can be extracted from various resources such as recommendations, reviews, publications, etc. which allows us to understand the emotion the author of the text tried to portray. With the help of NLP, we can extract huge amount of useful information from unstructured textual data such as Facebook posts, twitter posts, blog posts, product reviews and recommendations, etc.

The impact of Natural Language Processing in our daily life could be huge depending on how we use it to extract the data. For eg, if you are unsatisfied with the previous product that you had ordered from an e-commerce site, then the same search history keeps appearing. With the help of NLP your textual emotion of your review could be detected and the product would not appear quite often. Another impact of NLP could be in saving social mishaps. For eg, if a person wrote a depressing post on social media, many a times human sensation cannot detect the in-depth emotion of a particular text. In those scenario NLP can be life saving in many cases, perhaps can save someone from committing suicide by recognizing the emotion and thereby if some tags referring to the emotion gets attached to the post then their relatives or friends could do something about it.

In short Emotion Detection is the most important field of research in human-computer interaction.

LITERTAURE REVIEW-

I. Natural Language Processing is a branch of artificial intelligence is concerned with making computers understand human language. The task of exploring natural language processing to process Arabic texts efficiently is never easy. Building NLP systems using the NLP services via web API created by several companies, made it easier. The paper concerned explored the available NLP web API's that helps in supporting Arabic Language. [2].

II. Mobile based data can also be retrieved using NLP. The IT industry not only has moved their own industry towards advancement but also other sectors as well. The tentacles of IT industry have spread over the agriculture industry as well. Well, now in this upcoming digital world, the farmers are using smart phones to convey their knowledge about agriculture. To document the knowledge from the farmers the SME (subject matter experts) of agriculture keep it in a database. To provide further solution the IT people use efficient approach through NLP and RDF (Resource Description Framework) which is in the form of several mobile applications using SPARQL queries. [3].

III. An application of NLP could be also in text summarizing which could be used to summarize product reviews. With growing use of smartphones and Internet, there is a speedy growth in online shopping. How do you decide which one to buy out of several similar items? How do you know which one is genuine? So here is where reviews help. Each and every user goes through the review section before ordering something. So, what if the review section is brought down in some few words, instead of those long reviews? In this context text summarization is of pretty much use. There are various types of text summarization techniques both basic and advanced which summarizes the text to a few lines or some words as per your own benefit. Increased accuracy could be gained by working with seq2seq model along with LSTM and attention mechanism. [4].

IV. Identification of topic and extraction of key phrase using synonymous term grouping and term frequency is also a very important application of NLP. In this case we mainly utilize linguistic patterns and morphology syntax. This model can be considered as the probability of word sequencing and frequency or words. The method approaches on the process of semantic based topic clustering and key phrase selection.[5].

V. There are also models on handwritten recognition based on artificial intelligence. There is a proposed

general algorithm for efficient handwritten recognition. Tis algorithm is mainly used to reduce the workload. The time to convert documents into letters can be hugely reduced. The proposed algorithm is designed with several artificial intelligence technique in such a way that it can work with multiscript. The recognition accuracy is high as well as the need for such algorithms.[6].

VI. Almost every social networking site twitter, Facebook, Instagram, etc. are a huge source of information which probably could help a business firm with their product marketing. Twitter is one of the important sites where people tend to give their product reviews. The model used NLP to filter tweets from pre-processed data. Secondly Bag of Words (BoW) model is incorporated and the TF-IDF model is used to analyze the part where sentiment is concerned. Together with BoW and TF-IDF model the positive and negative tweets can be clarified. TF-IDF vectorizer further increases the accuracy of the sentiment analyzer.[7].

VII. Again, the social networking sites could be a very important platform to discuss social issues and causes. Sentiment analysis through NLP could avoid as well as make you aware of a social cause going on. Analyzing sentiments from the data collected from social media could help predict public mood on any particular event or news or any social cause. Accuracy of the model is increased by using semantics and Word Sense Disambiguation. The filtered text is further subjected to Ensemble classification to analyze the sentiment. This type of classification involves the combination of various independent classifiers on any particular classification related problem.[8].

VIII. Business Process Management can be automated by a key technique known as Natural Language Processing (NLP). The performance of Parts of Speech (PoS) tagging depends on the respective annotated data. The trained PoS tagger could nearly reduce the overall tagging error by 12%. In BPD the verbs being keywords could increase the error reduction by 27%. The ambiguity caused by the OOV words can be easily solved by extracting local contextual knowledge from images which are attached to help the users in better understanding of the process. [9].

IX. Sentiments of software developers could also influence the product quality and productivity. First the dataset of issue comments are selected and then an entity level sentiment analyzing tool consisting of several sentiments classification and entity recognition is used. It receives an overall precision of 77.19%, which is quite high. [10].

X. Sentiment analysis takes raw data from comments or posts and figures out the emotion behind it. It classifies the categories as either positive or negative or neutral. The process of this classification is known as polarity classification.
Computational linguistics and text analysis helps to perform sentiment analysis. Machine learning algorithms performs sentiment analysis on social media data. It is seen that in this way logistic regression has achieved as far the greatest accuracy when used with n-gram and bigram models. [11].

PROPOSED WORK- We will be using algorithms from the domain of Natural Language Processing and Machine Learning.

To continue with the process, we need to follow the steps ahead:

I. **Tokenization**:

The key aspect of working with textual data is tokenization. It is a process of breaking a string into small pieces of words called tokens. These

tokens can be used for understanding the context of the sentence and for further development of the model. The meaning of the sentence is embedded into these tokens. The order in which the tokens appear in the sentence is also important. So that is also taken care of.

The first and foremost step in the process of modelling text data is tokenization. Tokenization can be broadly classified into three categories:

i. **Word Tokenization**: The most commonly used tokenization algorithm is word tokenization, which splits a text into individual words based on certain parameters.

One of the drawbacks of word tokenization is that it cannot deal with out of vocabulary words. It does not recognize new words encountered during testing.

ii. **Character Tokenization**: It splits a textual data into a set of characters. It overcomes the drawback caused by word tokenization.

Although it overcomes the out of vocabulary problem, the length of the sentence becomes too long to handle it as a sequence of characters.

Therefore, we go to a next level tokenization which is in between word and character tokenization.

iii. **Sub word Tokenization**: This type of tokenization splits the piece of textual data into sub words.

Thus, tokenization is one of the most powerful ways to deal with pieces of textual data.

II. **Removing Stop words**:

Stop words are those words which are frequently used in many articles but carry a very less amount of significant meaning. Such words are 'and', 'the', 'if', 'of' and many more. These words may carry meanings when they are used in human conversation but are useless when it comes to natural language processing. Removing stop words may make the dataset cleaner and easier to use for training as there will be a smaller number of errors caused by those stop words.

When we remove the stop words, the size of the dataset as well as the time in training the model decreases. Besides this, removing stop words increases classification accuracy.

III. **Normalization**:

Normalization is little more complex than tokenization. It is easier for us to understand that words of similar kind carry similar meaning, but that is exactly not the case with machines. It converges all kind of similar words into one single root word or one single token. For example, "watched", "watching", "watches" all can be normalized into a single token "watch".

Therefore, text normalization transforms a word into its canonical form and there are two ways of doing it.

i. **Stemming**: it is the process of reducing words into simpler forms by removing prefixes and suffixes. It reduces vocabulary and increases the efficiency in the process of retrieving information. Stemming is usually one rule-based approach. It is faster than Lemmatization but when simpler cases are concerned, both give the same effect.

ii. **Lemmatization**: Different forms of the same word can be transformed into its canonical form or the base form. If spoken technically then the canonical form is known as a lemma. We can thereby ignore morphological variations on a single word.

For lemmatization to be done right both the lemma and the original data

have to stored to continue with the further process.

IV. **POS tagging**:

In a sentence, each and every word is co-related to each other. We call every word as parts of speech. These parts of speech are important to determine the emotion of the sentences. So, each word needs to be tagged with correct type. After tokenization the tokens should be tagged with correct type of parts of speech.
Universal POS tags: [12].
ADJ: adjective
ADP: apposition
ADV: adverb
AUX: auxiliary
CCONJ: coordinating conjunction
DET: determiner
INTJ: interjection
NOUN: noun
NUM: numeral
PART: particle
PRON: pronoun
PROPN: proper noun
PUNCT: punctuation
SCONJ: subordinating conjunction
SYM: symbol
VERB: verb
X: other

V. **Vectorizing the Tokens**:

Vectorization is a process that transforms a token into a vector, or a numeric array. As machine cannot process any kind of raw text, so that tokens must be converted into numeric values which can be processed further for training the model.

There are three broad ways of vectorization:

i. **Bag of Words Vectorization**: Certain terms should be understood like:

Documents- it is a single piece of text data.
Corpus- it is a collection of all of the documents given.
Feature- each word in the corpus is a feature.

Bag of Words vectorization takes a document from the corpus set and helps to converts into a numeric vector. It maps each document word to its feature vector.
All values inside the array will be zero except the position representing the address inside the feature vector.

ii. **Tf-idf Vectorization**: Tf-idf vectorization solves the problem of very common or rare words. Term frequency- inverse document frequency (Tf-idf) takes a word into consideration as to how many times it occurs in a document and a corpus.

tf ("word") = Number of times "word" appear in a document / total number of words in a document

idf ("word") = log (Number of total documents / numbers of a document with "word" in it)

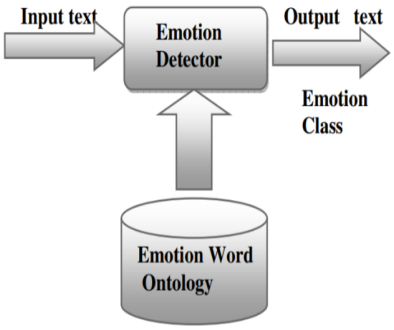
Both the above vectorization is unable to state the relation between two related words.

iii. **Word Embedding**: Word embedding is a vectorization technique which captures both the syntactic and semantic relationship from large corpus of text.

ALGORITHM AND TRAINING MODEL-

Here the SVM machine learning model is used as a classifier to classify the emotions from a text data. The preprocessed data is used as the training and test data to train and test the model. SVM is a classic model used for mainly classification. Here, the main goal of our project is to classify the emotions into multiple classes such as happy, sad, worried,

feared, panicked etc. The SVM model is given a labeled dataset to train itself. The SVM takes the datapoints from the vectorized dataset and outputs a hyperplane that best separates the emotion classes. This hyperplane is called as decision boundary.



RESULT AND ACCURACY-

Following are the results of the trained SVM when tested with 10000 tweets.

	Content	Emotion_predicted	Emotion_actual
0	people be so rude to you isaac they should get...	worry	sadness
1	totally screw up my ability to talk to a parti...	neutral	hate
2	n o t e x t	neutral	sadness
3	to take days off of work or have the money to ...	worry	worry
4	text	neutral	neutral
5	Got ta go to bed soon	neutral	neutral
6	n o t e x t	neutral	love
7	n o t e x t	neutral	neutral
8	that suck B be you ok	worry	neutral
9	my ipod i cant fall asleep	worry	sadness
10	i ve just be wrap up in day to day stuff so i ...	neutral	neutral
11	to summer strawberry be not availble in the Ch...	sadness	worry
12	ae get marry and it wasn t to alex	worry	worry
13	the hill in london u will realise what tourtur...	sadness	sadness
14	text	neutral	worry
15	when she start type on her computer in the mid...	worry	hate
16	talk at the Balisage Markup Conference 2009 Pr...	sadness	neutral
17	think i m definitely go to get an ear infectio...	worry	sadness
18	want to go out	sadness	neutral
19	I miss you	worry	sadness

Fig. 2.

CONCLUSION-

Natural language Processing has improvised the machine-user interaction to a great extent out of which emotion detection plays a very important role. Emotion detection can be used in social media content analyzing, product review, and also in post production market survey. Emotion Detection in future can further enhance user experience to a great extent.

The main aspect of sentiment analysis is to understand a body of text and the opinion it tries to express. With the help of emotion detection, suicides could be avoided to a great extent. Thus, emotion detection from text will be very fruitful in the near future.

REFERENCES-

[1] meaningcloud, introduction to emotion recognition in text, by Janine Garcia available at https://www.meaningcloud.com/blog/introduction-to-emotion-recognition-in-text accessed on 11th March,2021.

[2] S. A. Al-Ghamdi, J. Khabti and H. S. Al-Khalifa, "Exploring NLP web APIs for building Arabic systems," *2017 Twelfth International Conference on Digital Information Management (ICDIM),* Fukuoka, Japan, 2017, pp. 175-178, doi: 10.1109/ICDIM.2017.8244649.

[3] N. Srinivasan and A. Selvaraj, "Mobile based data retrieval using RDF and NLP in an efficient approach," *2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM),* Chennai, India, 2017, pp. 427-428, doi: 10.1109/ICONSTEM.2017.8261416.

[4] R. Boorugu and G. Ramesh, "A Survey on NLP based Text Summarization for Summarizing Product Reviews," *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India, 2020, pp. 352-356, doi: 10.1109/ICIRCA48905.2020.9183355.

[5] K. Nokkaew and R. Kongkachandra, "Keyphrase Extraction as Topic Identification Using Term Frequency and Synonymous Term Grouping," *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, Pattaya, Thailand, 2018, pp. 1-6, doi: 10.1109/iSAI-NLP.2018.8693001.

[6] N. Chumuang and M. Ketcham, "Model for Handwritten Recognition Based on Artificial Intelligence," *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, Pattaya, Thailand, 2018, pp. 1-5, doi: 10.1109/iSAI-NLP.2018.8692958.

[7] M. R. Hasan, M. Maliha and M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data," *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2),* Rajshahi, Bangladesh, 2019, pp. 1-4, doi: 10.1109/IC4ME247184.2019.9036670.

[8] M. Kanakaraj and R. M. R. Guddeti, "Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques," *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, Anaheim, CA, USA, 2015, pp. 169-170, doi: 10.1109/ICOSC.2015.7050801.

[9] X. Han, Y. Dang, L. Mei, Y. Wang, S. Li and X. Zhou, "A Novel Part of Speech Tagging Framework for NLP Based Business Process Management," *2019 IEEE*

International Conference on Web Services (ICWS), Milan, Italy, 2019, pp. 383-387, doi: 10.1109/ICWS.2019.00068.

[10] J. Ding, H. Sun, X. Wang and X. Liu, "Entity-Level Sentiment Analysis of Issue Comments," *2018 IEEE/ACM 3rd International Workshop on Emotion Awareness in Software Engineering (SEmotion)*, Gothenburg, Sweden, 2018, pp. 7-13.

[11] Samarth Garg, Divyansh Singh Panwar, Aakansha Gupta, Rahul Katarya, "A Literature Review on Sentiment Analysis Techniques Involving Social Media Platforms", *Parallel Distributed and Grid Computing (PDGC) 2020 Sixth International Conference on*, pp. 254-259, 2020.

[12] universal dependencies, universal POS tags, available at https://universaldependencies.org/u/pos/ accessed on 11th March,2021.