International Journal of Research in Engineering, Science and Management
Volume-3, Issue-6, June-2020
www.ijresm.com | ISSN (Online): 2581-5792

Information Extraction and Analysis from Text Messages Using Predictive Analysis

Nikhil Yadav[1*], Utkarsh[2], Chandra Prabha[3]
[1,2]*Stduent, Department of Computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida, India*
[3]*Assistant Professor, Department of Computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida, India*
*Corresponding author: nikhilyadav50586@gmail.com

***Abstract*: We tend to reside in an associate degree era of exaggerated pressure and mental disorders. The exaggerated level of stress and pressure leads to the inclination of range the amount |the quantity} of individuals showing unsafe tendencies and so a bigger number of individuals are committing suicide. Stress is often caused because of the family dispute, job discontentedness, health problems, etc. within the world of recent computing, individuals are happy to share their views and feelings over social media with peers and members of the family via services like electronic communication. because of the reserved nature and busy schedules of individuals, it's very troublesome to move with peers and members of the family, thus social media platforms are thought-about because of the most used platform for private conversations. This paper aims to estimate the unsafe tendencies of an individual by applying data processing techniques to the text messages an individual sends to the associated individuals. By analysing the parts of the text messages (keywords and emoticons) we will estimate the unsafe tendencies of an individual so necessary steps are often taken to avoid wasting the life of the subject.**

***Keywords*: Text mining, Information discovery, Sentiment analysis, Opinion mining.**

1. Introduction

The need for applying data mining techniques to text massages arrives from the ever-increasing suicide rates in numerous components of the planet. Saving the life of humans is the task of prime importance for a nation to avoid wasting the lifetime of individuals their sentiments should be illustrious and inferred so the desired steps are often taken on time. The most effective thanks to understanding the sentiment of an individual are by applying data processing techniques to the text messages an individual sends. If an individual shows symbols of hyper stress then informing the individuals on the brink of that person will facilitate in saving the lifetime of the topic. The text process is applied to the text obtained from the user [12].

Text pre-processing contains as tokenization, stop-word-removal, stemming and some other techniques. Tokenization splits the text in the form of words called chunks or tokens.

Tokenization is used to identify keywords in the texts. Stop-word-removal is the process of removal of words that do not convey a special meaning in the document like the, and, this ... etc. Stemming is a process where we erase suffixes like -ing, -ion, etc., to get the root word of the data.

This research focuses on sentiment analysis for predicting the stress level of a person. The prediction model comprises SVM and K-NN algorithms. This is done by feeding the system with a data set for training the system. This approach can be used to predict the results of elections when applied at a larger scale and for multiple subjects. It is highly effective in predicting the results regarding different opinions of people. It can be used to get prior knowledge about terror attacks or unorganized violent protests or riots [14].

Emoticons are a very important and most expressive part of any textual conversation over the internet as they convey the real essence of the conversation between the two people. Hence, it is of prime importance to analyze the emoticons used in any text message so that the real sentiment of the text is known [15].

Negative emotions reflect the sad or disturbed sentiments of the people and are thus replaced by negative words. Processing negative emotions is important because this is what affects more human's life.

2. Literature Survey

There have been several types of research done in the area of text mining. Now-a-days data mining is becoming one of the emerging technologies. Based on the study of some papers we have compiled our own literature survey as given in table 1.

3. Proposed Methodology

The proposed methodology helps to save the lives of those people who are facing issues of hyper stress or the other issue that can prove fatal to them.

The aim is to extract information from the text messages of the user and use it for various functions like sentiments analysis. The model additionally includes the analysis of emoticons to utterly analyze the statements. To implement our model we've used python artificial language. it's several libraries that facilitate to unravel the matter simply.

Table 1
Literature survey

No.	Research Paper Title	Name of Authors	Advantages	Disadvantages
1.	Information Extraction Through text Messages using data mining.[1]	Rishab Verma, Sartaj Ahmad.	• Easy to implement and understand. • The model implements the SVM algorithm. • The model implements the KNN algorithm.	• The model is very specified. • It violates user privacy to a certain extent.
2.	Enhancing Predictive Power of Cluster-Boosted Regression With Text-Based Indexing.[2]	Mark Chignell, Nippon Charoenkitkarn, Jonathan H. Chan, Wutthipong Kongburan.	• This model analyzes the Electronic Health Record to improve efficiency. • The model also examines whether textual features can be used to improve the accuracy of ICU mortality prediction.	• Very complex to implement because even a single mistake could harm someone's health.
3.	Financial Latent Dirichlet Allocation (FinLDA): Feature Extraction in Text and Data Mining for Financial Time Series Prediction [3]	Nout Kaunungsukkasem, Teerapong Leelanupab.	• In this model basically, technical and fundamental analyses are used by investors to predict financial time evolution, such as stock prices.	• This model is not fast and takes much time to predict.
4.	An Artificial Intelligence Driven Multi-Feature Extraction Scheme for Big Data Detection [4]	Jian Wan, Piaopiao Zheng, Neal N. Xiong.	• The model categorizes new articles according to the user's demand. • It has a vast application area.	• The main obstacle is poor portability because templates are designed for a specific purpose and they are difficult to reuse.
5.	Multistage Gene Normalization and SVM-Based Ranking for Protein Interactor Extraction in Full-Text Articles.[5]	Hong-Jie Dai, Po-Ting Lai, and Richard Tzong-Han Tsai.	• Using the multistage GN algorithm, we have been able to improve system performance by 1.719 percent compared to a one-stage GN algorithm. • Our experimental results also show that with full text, versus abstract only, INT AUC performance was 22.6 percent higher.	• This model can only be used for multistage gene normalization for protein interactor extraction.
6.	Bilevel Feature Extraction-Based Text Mining for Fault Diagnosis of Railway Systems.[6]	Feng Wang, Tianhua Xu, Tao Tang, MengChu Zhou, Fellow, IEEE, and Haifeng Wang.	• Impactful text mining of such impotant data plays an important role to find anomalies and making better fault diagnosis efficiency. • Perform an improved $\chi 2$ statistics-based feature selection at the syntax level to overcome the learning difficulty caused by an imbalanced data set.	• unstructured data and high-dimensional data, and unnecessary fault class distribution put challenges for feature selections.
7.	Comments Mining With TF-IDF: The Inherent Bias and Its Removal.[7]	Inbal Yahav, Onn Shehory, and David Schwartz.	• The model highlights one such shortcoming–the popular use of standard text analysis techniques (specifically, tf-idf), applied to comment classification, resulting in a biased analysis. • Comments of the social media platforms could be analysed to get a better insight into the business.	• We find that content extracted from the discourse is often highly correlated, resulting in dependency structures between observations in the study, thus introducing a statistical bias. Ignoring this scenario can brought in a non-robust analysis at best and can lead to an completely wrong conclusion.

A. Data set description

The data is obtained by extracting all the text messages sent by the topic. this may be achieved from multiple sources like Facebook, Whatsapp, etc. All the messages sent through these electronic communication services are saved and we will apply our model and analyze the feelings. {the knowledge|the info|the information} set can contain a text sort of data and emoticons. No alternative sort of knowledge like pictures is going to be analyzed through the model [8], [9].

B. Model components

Sentiment analysis:

In the component, data is labeled as positive and negative, the extent of it by performing data pre-processing using the SVM algorithm.

In this component, the data is assigned a sentiment such as positive or negative, and the extent of it by performing data pre-processing using the SVM algorithm [10].

Text Pre-processing:

Text Pre-processing is a major part of data analysis and data modeling. More than 75% of work id-data pre-processing and rest is analysis and training. Data preprocessing is the very first step in data analysis. There are many steps involved in preprocessing like Tokenization, Data standardization, emoji conversion, stop word removal, abbreviation analysis. All steps have been described below-

Tokenization:

Tokenization is the first step in data pre-processing. each new message is split into significant words known as tokens. Example –

"Have a decent day" is regenerate to "Have", "a", "good", "day".

Data standardization:

It involves changing all words within the message in commonplace kind, changing all words in small letters.

Example. "Wish you a great day" is regenerate to "wish", " you", " a", " great", "day"

Emoji conversion:

The emoticons gift within the text messages are tagged a keyword-based on the means of feeling categorical. These days emojis are fashionable to use in messages in the order that we are categorically clear which means and feeling of our expression. Typically, these emojis are higher than writing a talk as a result of they're straightforward and quick to know and additionally straightforward in knowledge analysis [11].

The emoticons are classified into the following 2 categories:

1. Positive emotions: these are the emoticons that convey positive sentiment and are replaced by positive words supporting the image.
2. Negative emotions: these are the emoticons that categorical negative sentiment and are replaced by negative words supported symbols.

Stop-word-removal:

All the words within the message that don't convey a special which means ar removed sort of a, the, then, etc.

Stemming:

It involves getting the foundation word connected to each word by having suffixes ling -ing, -ion, etc. Example- enjoying to play, gambling to game, electronic communication to message, etc. as well as improves the efficiency of our system. Less but meaningful data is important not only data.

Abbreviation analysis:

Altering the abbreviations having in the messages by their complete forms. Example ASAP by as soon as possible, MSG by message, GN by Good night etc.

Flow chart of every step is given below to describe it visually.

N-gram:

The next step in data pre-processing is N-gram options extraction. N-gram could be a series of n tokens. N-gram is a model very widely used in NLP tasks The model creates N-grams from the messages within the data set to extract keyword options from the data set.

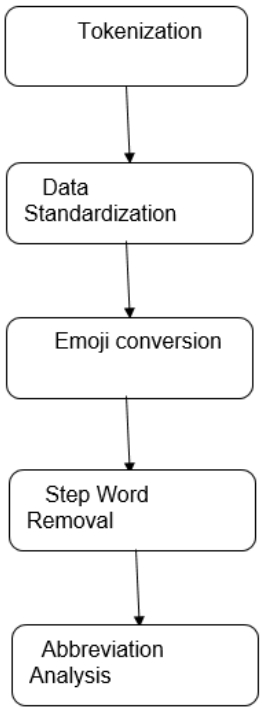

Fig. 1. Steps involved in data pre-processing

For n = 4 a sequence of four-words for every message is generated. The method of N-gram will increase the potency and accuracy of the classification step as a result of the feature extracted from 4 sequences of token combinations.

Example: "Hard work pays off" is analysed as "hard work pays" "work pays off".

Term Frequency:

The number of times a token occurs in each data sample is called the term frequency of that particular token. Tokens having high frequency have a better relationship with the sample.

Support Vector Machines:

The stream of words when the text pre-processing step is processed by the SVM algorithmic program to classify the messages as "normal" or "critical" sentiment. the method is applied to each message within the information set to classify the chat united among "normal" and "critical" sentiment. so we are going to get a sentiment related to the messages related to the user. SVM's are supervised learning models that are used for classification and multivariate analysis of information used. An SVM model represents examples as points in house, totally different categories of examples are divided by a particular gap that should be as wide as potential. New examples once mapped into the house are expected to belong to a category of examples supporting that aspect of the gap they fall.

KNN Algorithm:

The output gain from Support Vector Machines Algorithm (SVM Algorithm) are group of two sentiments with class labels

"normal" and "critical". Based on the output KNN algorithm is applied in order to deduce the overall sentiments of the subject. The input for the KNN algorithm is the sentiments associated with all the chats that the subject is involved in. The last task is to predict the sentiment of a person according to gathered feature set. Data is divided into training and testing sets, and the KNN algorithm is used to predict the sentiment. KNN algorithm is a method for classifying data based on the nearest training sets in the feature space. The class label is assigned the same class as the nearest K instances in the training set. KNN is a type of slow learner algorithm. KNN algorithm is considered a flexible and simple classification technique based on machine learning concepts.

4. Result Analysis

Our model predict the result in numerical value between -1 and 1. According the numerical value we mapped emotion so that end user could understand it very clearly. Numbers below 0 shows negative sentiment while numbers greater than 0 shows positive sentiment otherwise it is 0 then it means neutral sentiment. The following table shows experimental result based on value,

Table 2

No.	Numerical value	Corresponding sentiment
1.	1	Ecstatic
2.	.8	Blissful
3.	.6	Very happy
4.	.4	Happy
5.	.2	Joyful
6.	.1	fine
7.	0	Neutral
8.	-.1	Not good
9.	-.3	sad
10.	-.6	Stressful
11.	-.8	Miserable
12,	-1	Depressed

5. Future Scope

The proposed model can be used in situations where sentiment analysis is required to achieve the desired result and use it for various different purposes such as critical reviews for hotels, movies, videos, etc. Sentiment analysis methods till now have been used to find degree of thoughts and opinions of all the users towards a point that access social media. Businesses are fascinated to analyse the thinking of people and how they are giving response to all the products and services provided.

Companies use sentiment analysis to pitch their advertisement campaigns and to make better their products. Companies targeting use of sentiment analysis tools in the field of customer feedback, e-commerce, marketing and CRM.

6. Conclusion

The proposed model takes input from the data set created by accumulating all the text messages sent by the subject. All the messages may be from different social media platforms such as facebook, whatsapp, etc. The messages are then pre-processed to get the meaningful words from the data sets. After pre-processing we use probabilistic language models like n gram. The next step is to use the classifying algorithms to classify the conversations as "normal" or "critical". First a supervised algorithm is used which is SVM as it proves to be highly efficient for such computations and then an unsupervised algorithm is used which in turn increases the efficiency drastically, in our case we use the KNN algorithm. Thus we own to give a highly effective method of finding the sentiment of the person by analysing the text messages and also analysing emoticons. Emoticons are very common- type of tokens in any kind of text message.

References

[1] Muhammad Inaam Ul Haq, Qianmu Li, and Shoaib Hassan, "Text Mining Techniques to Capture Facts for Cloud Computing Adoption and Big Data Processing."
[2] Lihui Zhang, Gancheng Zhu, Shujie Zhang, Xiangping Zhan, Jun Wang1, Weixuan Meng, Xin Fang, and Peng Wang, "Assessment of Career Adoptability: Combining Text Mining and Item Response Theory Method."
[3] Jian Wan, Piaopiao Zheng, Huayou Si, Neal N. Xiong, Wei Zhang, And Athanasios V. Vasilakos, "An Artificial Intelligence Driven Multi-Feature Extraction Scheme for Big Data Detection."
[4] Nont Kanungsukkasem and Teerapong Leelanupab, "Financial Latent Dirichlet Allocation (Fin LDA): Feature Extraction in Text and Data Mining for Financial Time Series Prediction."
[5] Wutthipong Kongburan, Mark Chignell, Nipon Charoenkitkarn, And Jonathan H. Chan, "Enhancing Predictive Power of Cluster-Boosted Regression with Text-Based Indexing, Data Science and Engineering Laboratory, School.
[6] Muhammad Aman, Abas Bin Md. Said, Said Jadid Abdul Kadir, And Israr Ullah, "Key Concept Identification: A Sentence Parse Tree-Based Technique for Candidate Feature Extraction from Unstructured Texts."
[7] Li Shi, Chen Jianping, And Xiang Jie, Prospecting Information Extraction by Text Mining Based on Convolutional Neural Networks—A Case Study of the Lala Copper Deposit, China.
[8] Inbal Yadav, Onn Shehory, and David Schwartz, "Comments Mining with TF-IDF: The Inherent Bias and its Removal."
[9] Feng Wang, Tianhua Xu, Tao Tang, MengChu Zhou, and Haifeng Wang, "Bilevel Feature Extraction-Based Text Mining for Fault Diagnosis of Railway Systems."
[10] Hong-Jie Dai, Po-Ting Lai, and Richard Tzong-Han Tsaim, "Multistage Gene Normalization and SVM Full-Text Articles."
[11] Sartaj Ahmad and Rishabh Varma, "Information extraction from text messages using data mining techniques."
[12] https://en.wikipedia.org/wiki/Sentiment_analysis
[13] https://www.javatpoint.com/text-data-mining
[14] M. Kantardzic, Data Mining: Concepts, Models, Methods and Algorithms, John Wiley & Sons, 2011.
[15] List of Text Emoticons, Retrieved 20 July 2012.