INTERNATIONAL JOURNAL OF SOFT COMPUTING

MEDWELL PUBLICATIONS

Sentimental Analysis on Blog Data

G. Pradeepini, V. Sahithi, N. Siva Sree and P. Yamini Padmaja
Department of Computer Science, KL University, Andhra Pradesh, India

Key words: Social media, education, data mining, web text analysis, sentimental analysis, conducted

Corresponding Author:
G. Pradeepini
Department of Computer Science, K L University, Andhra Pradesh, India

Page No.: 50-56
Volume: 16, Issue 3, 2021
ISSN: 1816-9503
International Journal of Soft Computing

Abstract: Indian history is a composition of distinct knowledge and automation. It all depends on the progress along the period. In this study we conducted sentimental analysis on various data that is collected. This helps in finding the behavior of the people who are active or inactive. Now-a-days, social media is taking a crucial role in social networking and sharing of data. These sites are favored by many users since it is available to everyone without any restriction to share their opinions, educational learning experiences and concerns via their status. Student's posts on social network offers us a stronger concern to take decisions with reference to the particular education system's learning method of the system. Evaluating such knowledge in social media is sort of a difficult method. Within the proposed system, there will be a work flow to mine the knowledge that integrates each study and big scale data mining techniques supporting the various distinguished themes in which behaviors are categorized into different teams. Data mining techniques are used to enforce the deep-mined knowledge for analysis purpose to urge the deeper understanding of the information. Word based measures are mostly taken to research the results and comparing them with the prevailing sentiment analysis technique.

INTRODUCTION

Now-a-days web applications and social media have grown, there is a rapid growth in the amount of information and this as became the major challenge to reduce. A personal review or survey which describes about the once behavior or the opinion which is playing a major role in every aspect. Since, social network is flexible for everyone and became a immense part of their life. Since it is available all over the world, social network sites had become popular now a day's such as Instagram, Facebook, YouTube, linked-in etc. Figure1 describes the

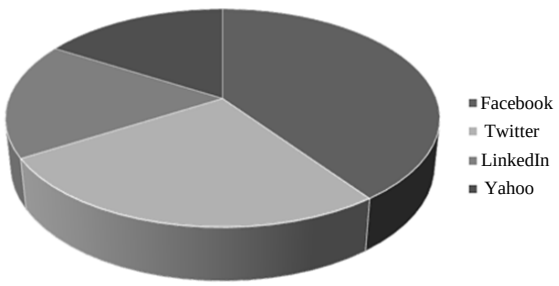

Fig. 1: Impact of social media on society

impact of social network on the current society. This affords a wonderful platform for college students and others to express their anger, fear, pleasure, joy and emotions. Everyday user's discuss and share their encounters in formal and casual way on different social media websites. User's tweets or remarks and a personal survey on specific posts offer big amount of implicit expertise and an entire new attitude for researchers and practitioners to understand the user's conduct outdoor and the controlled surroundings. This understandings are useful for taking the selection at higher level in taking the consideration of scholar's point of view for their development. Even although social media facts gives a lots of possibilities to understand scholar's conduct, however nevertheless there are some methodological problems in making experience of the social media statistics for educational purposes. These social networks generate a huge amount of data in day to day life. This is enable to store the data in the traditional database. In order to overcome this disadvantage a platform was created to store a huge amount of data which was popularly known as Big Data.

Big Data contains extremely large data sets that are analyzed to find the different hidden patterns, fashion and associations, especially relating to the once behavior and the actions. The different patterns are classified by depending on the density of the word in the word cloud or Data cloud. This is a text mining method which highlights the most frequently used words. It is also known as text cloud or tag cloud. Data cloud is formed based on the density of the word or item in the sentence or message, depending on the density of the word, it is highlighted. This data cloud is to classify different behaviours of the user by performing some techniques namely sentimental analysis.

There are many range of techniques utilized by researchers together with surveys,group to gather facts associated with the user's conduct. These methods are typically very time-ingesting, manual and no longer very frequent. Considering these drawbacks of existing system a new system was proposed. In proposed system a sentiment analysis is used instead of the qualitative analysis because sentiment analysis considers the opinion of the user about a system and categorizes it into many different levels namely anger, disgust, joy, neutral, negative or positive mood.One of the hardest task is to search keywords which leads us to a confusion whether the user's feelings are positive, negative or neutral. By exploring more advanced information retrieval methods there are two ways of extracting data. One among them is semantic based information retrieval in which it uses semantic information to understand the documents and queries. The other method is machine learning based method which is used to reorganize Web documents.

In this study, we mainly discusses on the Users behavior by performing the personal review or survey based on machine learning. So, that we perform a sentiment analysis which is implemented to analyze the behavior of the user as per the category and the results will be reported to a decision makers which helps the person to get the overview of their problems.Sentiment analysis is performed using R-language which is dynamic language for statistical computing designed in 1993 as a successor to S. It combines lazy-functional features and object-oriented programming which was not proposed by computer scientists but became popular surprisingly. In this project we also generated word cloud for the collected users data.So that they can take necessary measures to overcome existing problems in their social life.

Literature review: Hu and Liu[1] collected the data about 6,800 words of customer reviews and then performed sentiment analysis manually and summarized their behaviour.

Das and Chen[2] had developed a methodology to extract sentiment from the stock message boards. They used phase-lag analysis, pattern recognition and statistical methods to improve the quality of sentiment index, which is done manually.

Jindal and Liu[3] has discovered that there are two types of opinions. They are namely regular opinion and comparative opinion. In which regular opinion is used to express the sentiment of the specific entity or its characteristic but comparative opinion comprises multiple entities based on some characteristics.

In order to detect the polarity of words semi-supervised and automatic methods are proposed by Hatzivarsiloglou[4]. He proposed an alogorithm which determines polarity of adjectives.

Turney and Littman[5] proposed an algorithm to calculate the polarity of words by finding the type of most co-occured words .Mohammad, Dunne and Dorr in 2009 found an algorithm which generates sentiment lexicons for >60,000 words which are taken from glossary automatically. They used two sentiment lexicons from tweets using emotions and hashtags.

Since, the manual explanation of data is very costly so they used reserved supervision techniques on short informal texts. According to Go, Bhayani and Huang uses tweets with emoticons for supervised training. According to him emoticons like) are taken as positive tweet label and: (are taken as negative tweet label. In 2012 Mohammad had developed a classifier in order to detect the emoticons using tweets having hashtags as labeled data.

According to work done by Jia,Yu and Meng found that negation plays an important role to determine the sentiment where it involves identification of negation

words such as not, never, doesn't and etc determines the scope of negation.But according to Polani and Zaenen-2004, Kennedy and Inkpen, the scope of negations are assumed beginning from the word which follows negation word till the next punctuation mark or end of sentence is found.

One of the common way for capturing the impact of negation is by reversing the polarity of sentiment words in their scope. In order to do this Taboada *et al*.[6] had proposed a algorithm which shifts the score of the sentiment term in context of their negation toward the opposite polarity with fixed amount. But this was not accepted by many researches in many cases.

With the past knowledge, all the work done is manual and time-consuming in order to overcome that we choose one of the machine learning technique i.e, sentiment analysis. User's behavior is categorized depending on the comments and survey they had given on a particular post are used. Sentiment analysis pursuits to decide the mindset of a speaker or a writer with respect to a few subject matter or the general contextual polarity of a file and impact the user wants to have on the reader. Sentiment analysis has emerge as famous in judging the opinion of purchasers in the direction of numerous brands. The manner where users specify their opinions on social networking web sites allows to choose this opinion. The primary problem is to apprehend this sentiment and being capable of classify it appropriately with less time consuming.

MATERIALS AND METHODS

System architecture: The advancement of social media data sense-making for different purposes, coordinating analysis and applying data mining techniques as illustrated in Fig. 2 and to know the user's formal and informal conversations which are separated into the opinions like negative, positive or neutral in order to know problems confronted in the personal life of the users. The system involves several steps like Data Extraction, Data collection, Data pre-processing, Term Frequency, Word cloud formation, Performing sentimental Analysis and finally analyzing output.

Data extraction and collection: Data extraction is a place where information is broken down and crept through extracting important data from information sources (like database) in a particular example. In future information handling is done in which metadata is included along with other information in-corporation; another procedure in information work process. The data extraction is used to extract the data from various web sources like Twitter, Facebook, blogs and from any personal product reviews. If there is any unstructured data,data preprocessing must be done to integrate, normalize and to get the sturctured data.The major share of information extraction originates from unstructured information sources and diverse information positions.This unstructured information can be in any form, like tables, records, tuples and analytics.

The data is collected from social network sites and other personal review or survey.The data collection is the toughest problem because it is very secure and encrypted end to end so in order to collect the data we used some tools namely,scraper world and face pager etc.Face pager was made for bringing open accessible information from Facebook, Twitter, Instagram and other JSON-based APIs. All the information is stored in a SQLite databases and exported to csv or xls files and are saved in the form of .csv or .xls spreadsheets.

Data storage: The data may be stored in the form of structural or unstructural and this may be stored somewhere for few purposes. The purposes are like

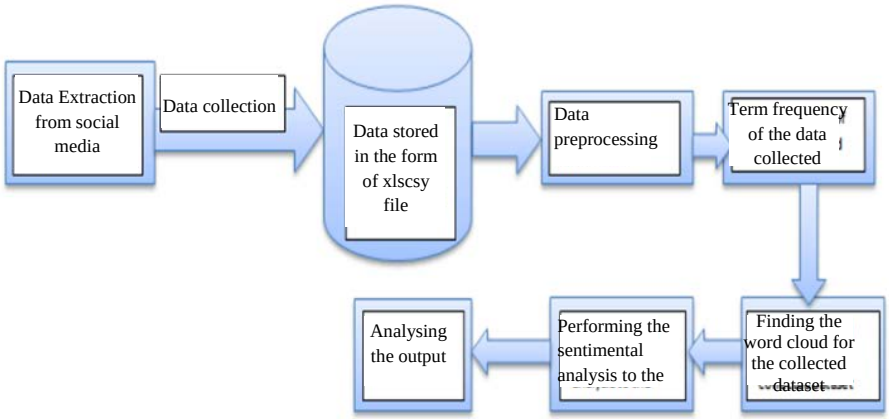

Fig. 2: Workflow that we expand to make sense of social network data

knowing the behaviour or knowing the opinions of the social network users. This data is stored in tables or spread sheets like .csv or .xls format. The stored data is now processed and data mining techniques are applied in order to perform data pre-processing, data cleaning and data reduction. The refined data is used for performing sentiment analysis easily.

CSV file,a comma separated values files in which the data is stored into the tables. An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

Data pre-processing: Data pre-processing is an information mining strategy that includes changing crude information into a justifiable format. Real world information is regularly deficient, conflicting as well as containing many errors. Data pre-processing is a proven strategy for settling such issues. Data pre-processing plans crude information for further handling. Data pre-processing is utilized database-driven applications, for example, client relationship administration and lead based applications (like neural systems). There are series of steps involved in Data pre-processing:

Data cleaning: Cleaning of data can be done through procedures, for example, filling in missing qualities, smoothing the boisterous information, or settling the irregularities in the information. Here we clean data by replacing the empty spaces in excel sheet with zeros.

Data integration: Data with various representations are assembled and clashes inside the information are settled.

Data transformation: Data is standardized, collected and summed up.

Data reduction: This step expects to show a diminished representation of the information in an information distribution centre.

Data discretization: Involves the lessening of various estimations of a consistent characteristic by isolating the scope of property interims.

Term frequency: In order find the term frequency first we have to find term-document matrix which is used to find the relationship between the terms and the documents, in which every row represents a term and every column represents document and also an entry is the number of existence of the terms in document. Now we have to find the frequency of most popular words in the document. This is called term frequency where most popular words are visualized largely. Here, same frequency words are visualized in same color.

Word cloud formation: After finding the term frequency, to provide effective visual overview of data we generate word clouds alternatively. Here, we can adjust the frequency also. To this we can include some colour in-order to display. We can also change the ranges of font size used for the plot.

Performing the sentimental analysis: Sentimental analysis follows after the word cloud formation in order to find the different behaviors of the user. Sentiment analysis considers the opinion of the user about a system and categorizes it into many different levels namely anger, disgust, joy, neutral, negative or positive mood.

Output analysis: In order to suggest the solutions for the user we need to analyse the results of sentimental analysis performed. By analyzing we can suggest the user about their behavioral changes. So, that users get benefitted.

Implementation: We collected data from one of the blog (http://www.klusn.com) in which our university is running. The blog contains all the information related to student activities. The data set is collected from university admistartion. The data set contains post_id (where each and every post has its unique id), permlink, post_message (messages the admin post on the blog), users_feedback etc. This data is stored in the form of .xls or .csv file format

Initially the data set is cleaned by replacing the missed columns with zeros. Now the data set is loaded in R-studio for the further process to continue. Initially to perform the analysis in R-studio packages are to be installed. R-Studio downloads the package from CRAN, so we need to be connected to internet. Here we use some of the packages like Table 1.

After packages are being installed we load them into R-studio. Now we load data set from our hard disk and perform data pre-processing activities under ‘tm’ package by using ‘corpus’.

Corpus () is a general function in R-environment which represents collection of documents. It involves basically the loading of files which are created in text mining folder into the object of the corpus, this is enabled by ‘tm’ package.

Another important step in text analysis is data cleansing. For example we see, very clumsy data will play confusion with results. Besides, we also see that data cleaning is always a continuous process since there are problems that are ignored for the first time about. The ‘tm’ package also offers a number of changes that comfort the dullness of cleaning data. This can be done using get Transformations () function. Still there are few primary clean-up steps needed to done before using these powerful transformations. These clean-up activities include removing punctuations, removing numbers, some special

Table 1: R-Studio downloads the package from CRAN

Package	Description
Lazyeval	It is an alternative approach for non-standard evaluation using formulas. It also provides a full implementation a full implementation of LISP style for the easier generation of code with other code ggplot2
Winston Chang	It is a system used for creating graphics, based on the data we provide and tell 'ggplot2' how to map variables to primitives to use, and it takes care of the details tm
Feinerer	It provides a clear way of preparing documented data for the statistical study and also offers an easy extensibility by well documented interfaces
wordcloud in Fellows	It is used to plot the cloud of words that are shared across documents
syuzhet in Matthew Jockers	It is used to extract sentiment and sentiment driven plots from text
NLP in Kurt Hornik	It is used to compute explanations by continuously calling the given explanations with the given text and current explanations, and then integrate the newly formed explanations with current ones

Table 2: Then term document matrix for this corpus would be like

Parameters	She	is	a	good	girl	-
	She	is	a	dancer	-	-
Doc 1	1	1	1	1	1	0
Doc 2	1	1	1	0	0	1

Eg: Doc1: She is a good girl. Doc 2: She is a dancer

symbols, unnecessary white spaces, stop words and converting any upper case letters to lower case letters. The next step of the process is creating a Term Document Matrix (TDM) which lists all the words that occur in corpus by the documents where documents are represented in rows and words by the columns. In a particular document the word occurrence can be represented as an entry of matrix corresponding to that row and column is 1, else 0.If any word occurs twice it is recorded as "2" in the respective matrix (Table 2).

Before proceeding further there are two points about the DTM. Initially DTM's are very huge i.e. measurements of the matrix would be number of documents X number of the words in the corpus. Secondly, DTM in invariably sparse matrix since majority of its entries are 0. It builds relation between terms and documents[7-10].

Now we calculate the collective frequency of in the document and sort them in advance. In the next step word cloud for the words is created. Here in order to get the same view every time we set seed value. We can also add colours to the word cloud for the words having the same frequency as one colour.

In next step we perform sentimental analysis were the behaviour of the user is classified based on there comments, feedbacks etc. The sentiment score is classified into 10 emotions like anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise and trust. Internal working of the sentimental analysis work as in the following Algorithm 1.

Algorithm 1: Algorithm for each level of sentence

```
1: for each word wi in sentence Senti do
2:        initialize the sentiment score (wi)<- 0
3: end for
4: for each word wi in sentence Senti do
5:        if the word K(wi) in sentiment words list SList then
6:                set K(wi)<- (positive: 1 and negative: -1)
7: end if
8: if the word wi in negation words list NList then
9:        for wj in the same clause with wi do
10:               set K(wj) <- -1* K(wj) do
11:       end for
12:       if all words wj in the same clause with wi and wj not in
SList then
13:               set K(w|Clause| )<-  -1
14:       end if
15: end if
16: if the word wi in transition words list TList then
17:       for wj in the same clause with wj and wi not in SList  do
18:               set K(w|Clause| )<-reverse value of the former clause
19:    end for
20: end if
21: if the word wi in adverbs of degree list DList then
22:       for wj in the same clause with wj and wi in SList do
23:               set K(wj)<- degree(wi) *K(wj)
24:    end for
25:    if all words wj in the same clause with wj and wi not in SList
then
26:       set K(w|Clause| ) <-former clause is positive: 1, negative: -1
27:  end if
28: end if
29: end for
30: for each word wi in sentence Senti do
31:       score <-Error! Reference source not found
32: end for
```

The above algorithm 1 is general code for sentimental analysis for knowing the positive, negative and neutral emotions of the user. The algorithm can be modified for the emotions like anger, anticipation; disgust, fear etc. can also be added. By implementing as per above said manner we identify the behaviour of the based on their reviews written[11].

RESULTS AND DISCUSSION

Mining of web data is very complex, it give us lot of information what we require. In this project the data from the blog is collected and above steps are done accordingly to find the behaviour of the user by performing the sentimental analysis, in their life. The total score of the sentiment of different users for their reviews is as shown in Fig. 3 which represents different behaviours of the user[12].

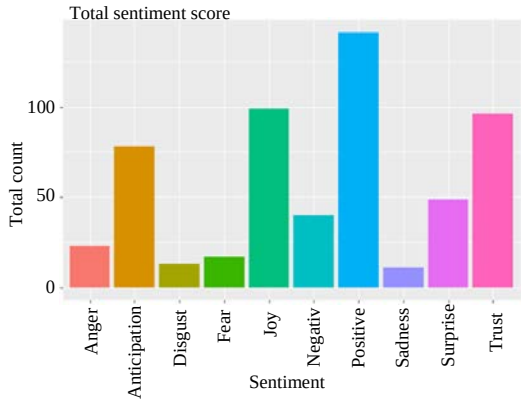

Fig. 3: Sentimental analysis

Fig. 4: Word cloud

We also create word clouds(since they are alternative graphical representation and are also easy to understand) for the data collected from the blogs Fig. 4 represents the word cloud for the blog data such that each colour represents the words having the same frequency. The process is conducted for different number of records and they are compared among them. The same process is conducted for 50,000 records; 20,000 records; 10,000 records and compared among different behaviours of the user (Fig. 5).

Input: Input Dataset file of user's data or comments on the blog.

Output: Displaying categorization of user problems and issues with 10 dissimilar groups.

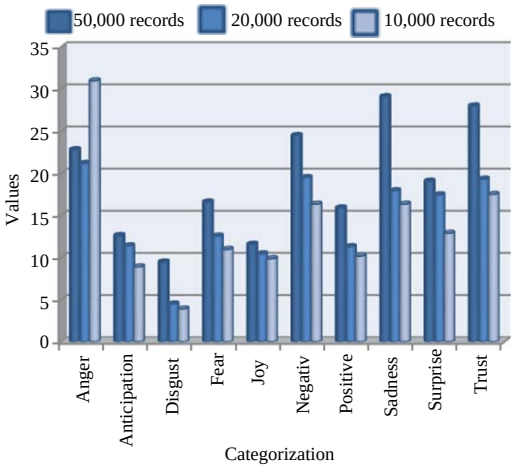

Fig. 5: Categorization of user's data

CONCLUSION

Sentimental analysis is always referred as a opinion mining where results are analyzed by the graphs we encountered during the process. The utmost elements of the graph usually represent positive, negative or neutral feelings. This study describes the behaviour of the user using sentimental analysis and is compared to the general manual process. This study says about how the user reacts to posts or messages in the blog by analyzing there feedback or comments. And we also compare the results for the different categories of the record. According to the data we collected from the blog, we can say that the most of the users react positively to the posts in the blog.

We can enhance the work by processing the images, videos etc. This system is very helpful for manufacturing companies, government sectors, banking sectors etc. in future, for identifying the employees actions, product review, their behaviour, banking reviews in their related services.

LIMITATIONS

Even though there are many advantages like security of data, easy analysis etc but there are some limitations. It limits the mining on huge and significant data when we experiment to calculate sentiment. A particular opinion on a product or any post may influence other indirect causes like someone may have bad day comments a negative remark about something with positive emoji sign which leads us to confusion. With the huge enough sample, outliers bare combined and there may be any alterations. And also sentimental are very feasible to vary over the time

according to user's mood. So, it is necessary to look data from users view point at specific time.

REFERENCES

01. Hu, M. and B. Liu, 2004. Mining and summarizing customer reviews. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data, August 22-25, 2004, ACM Press, Washington, USA., pp: 168-177.
02. Das, S. and M. Chen, 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. Proceedings of the 8th Asia Pacific Finance Association Annual Conference, July 22-25, 2001, Bangkok, Thailand..
03. Jindal, N. and B. Liu, 2006. Mining comparative sentences and relations. Proceedings of the 21st International Conference on Artificial Intelligence Vol. 2, July 16-20, 2006, ACM, Boston, Massachusetts, ISBN:978-1-57735-281-5, pp: 1331-1336.
04. Hatzivassiloglou, V. and K.R. McKeown, 1997. Predicting the semantic orientation of adjectives. Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics, July 7-12, 1997, Madrid, pp: 174-181.
05. Turney, P.D. and M.L. Littman, 2003. Measuring praise and criticism: Inference of semantic orientation from association. Proc. ACM Trans. Inform. Syst., 21: 315-346.
06. Taboada M., J. Brooke, M. Tofiloski, K. Voll and M. Stede, 2011. Lexicon-based methods for sentiment analysis. Comput. Ling., 37: 267-307.
07. Boubel, N., T. Francois and H. Naets, 2013. Automatic extraction of contextual valence shifters. Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, September 7-13, 2013, Hissar, Bulgaria, pp: 98-104.
08. Chen, X., M. Vorvoreanu and K. Madhavan, 2014. Mining social media data for understanding students learning experiences. IEEE. Trans. Learn. Technol., 7: 246-259.
09. Morandat, F., B. Hill, L. Osvald and J. Vitek, 2012. Evaluating the design of the R language. Proceedings of the European Conference on Object-Oriented Programming, June 11-16, 2012, Springer, Beijing, China, pp: 104-131.
10. Switzer, J., L. Khan and F.B. Muhaya, 2011. Subjectivity classification and analysis of the ASRS corpus. Proceedings of the 2011 IEEE International Conference on Information Reuse & Integration, August 3-5, 2011, IEEE, Las Vegas, Nevada, pp: 160-165.
11. Zhang, W., H. Xu and W. Wan, 2012. Weakness finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. Expert Syst. Appl., 39: 10283-10291.
12. Theubl, S., I. Feinerer and K. Hornik, 2012. A tm plug-in for distributed text mining in R. J. Stat. Software, Vol. 51, No. 5.