International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Sentiment Severity on Location-Based Social Network (LBSN) Data of Natural disasters

K Shyamala, Vijay Kumar Kannan, Sheran Dass. D

Abstract: ***Social media emerged as one of the key components to reach disaster affected people, as they supplement planning and operational coordination. Sentiment analysis was expended to identify, extract or characterize subjective information, such as opinions, expressed in a tweet. The sentiment expressed is analyzed and is classified as positive or negative sentiment, which is not versatile enough to capture the exact sentiment conveyed by the user. Opinion mining is a machine learning process used to extract information conveyed by the user in the form of text. In this paper, the lexical analysis to sentiment analysis of twitter data is employed. Conventionally, the sentiment is conveyed using the polarity of the data but in this paper, sentiment intensity is employed to convey the sentiments. Performing sentiment analysis on tweets gives us the sentiment intensity conveyed by the user, which in turn is used to calculate the severity of the disaster event specified by the user. Further, it is also used to classify the tweets based on their severity. This paper proposes a methodology to extract relevant sentiment information from Location Based Social Network (LBSN) and suggests a unique scale to classify this information to help disaster management authority.***

Keywords: ***Sentiment Analysis, Unsupervised Learning, Natural Language Processing, Lexical Analysis, Disaster Management, Location Based Social Network***

I. INTRODUCTION

Information of people in distress and their location are some of the required parameters in disaster management. The best way to gather these parameters is social media, which is one of the most dynamic and innovative inventions of the 21st century. It is also one of the best forms of broadcasting communication in the scenario of disaster and allows a user to broadcast information from anywhere in the world, which can be accessed by anyone in any part of the world, instantly. In case of disaster, the analysis of the data in the social media can act as a vital source of information, which can alert officials to identify the needy people and arrange for immediate aid.

Manuscript published on January 30, 2020.
* Correspondence Author
Dr. K Shyamala*, Associate Professor, Department of Computer Science, Dr.Ambedkar Govt. Arts College, Vyasarpadi, Chennai, India. Email: shyamalakannan2000@gmail.com
Vijay Kumar Kannan, Research Scholar, Department of computer science school of computing sciences, VELS Institute of Science Technology and Advanced Studies, Pallavaram, Chennai, India. Email: vijaydharshan@gmail.com
Sheran Dass. D, M.S, Ira A. Fulton Schools of Engineering, Arizona State University, Tempe, Arizona, USA. Email: sherand96@gmail.com

Some social media platforms store the geospatial data of the post with the data posted by the user. These state-of-the-art platforms, which give us the geospatial information of the posts, are called Location Based Social Network (LBSN). This is one of the primary reasons to perform sentiment analysis on a LBSN like twitter for disaster management as it will give us the specific location of the disaster event [3]. LBSN includes any social media network, which handles geospatial data like twitter, facebook, Snapchat, Instagram and other similar platforms. The data can be filtered to remove irrelevant tweets based on the geospatial information of the tweet as the sentiment conveyed by a person away from the actual disaster location may not convey the exact scenario of the disaster at the affected location.

The purpose of this paper is to propose a specific methodology for scaling the severity of sentiment on the disaster management. In this method, the tweets were extracted from twitter application and the sentiment analysis was performed. The main objective of this paper was to develop a scale or metric, which could indicate the severity of the disaster event based on the information derived from each tweet. Conventional methodology dictates the sentiment polarity be used to convey the sentiment of the tweet [3][2] but disaster management demands a much more complex classification of sentiment [7]. Even though the intensity of the sentiment is a feature of sentiment analysis, which on its own cannot convey more sense than the sentiment polarity, it can become a powerful metric when handled with skill and further, can classify the data with finesse. Furthermore, the sentiment intensity value can reveal the severity of tweets. Differentiation of the severity index based on rating scale of 10 ordinal values gives the officials of disaster management the liberty to assign priorities and allocate their resources to different disaster events in a better manner. Moreover, the sentiment analysis mainly focuses on two methodologies, supervised and unsupervised learning. The supervised learning method, though effective and more accurate, does not fulfill our objective and is not particularly suitable for social media texts [6]. The unsupervised learning demands substantially less resources when compared to the supervised learning method [15] and fulfills our objective.

There were two data sets with data on major international disasters that took place during the month of April 2019. The first data set was on Iran floods which occurred during the months of March 2019 and April 2019. The second dataset was on the cyclone Fani which originated on 26th April and by the start of May 2019 had developed to an equivalent of Category - 4 hurricanes.

It is noted that most of the tweets from Iran did not have their location tagged, and the reach of internet was not huge in Iran. In the second case, in India, each government departments had their own official verified twitter page where they actively communicated the real-time events to the world.

Moreover, during this research, it is observed that the probability of authentic tweets with geospatial information is more in developed countries than in underdeveloped countries.

Though information in real-time is a really powerful resource for disaster management, there are two major hurdles, when it comes to using data from twitter for disaster management. The first hurdle is that the format of social media text is ever evolving and when this important text form is lost during pre-processing it is impossible to get the exact context conveyed by the user. This means that though we could get relevant information it sometimes would not be the entire context conveyed by the user. The second hurdle is that the sentiment polarity does not give relevant information related to disaster management. Though we have enormous amounts of data, it is important to analyze the data efficiently and to process this unstructured data to get relevant structured information to aid in disaster relief. The methodology proposed here aims to tackle these problems and help make efficient relief efforts.

II. REVIEW OF LITERATURE

As a consequence of over exploitation of natural resources and climatic changes, we have resulted in the extinction of many human lives in the past century. With concern on increase in climatic changes, there are continuous natural disasters that might unfavorably affect both the long-term and short-term health consequences in emerging nations.

Previous researchers had mainly concentrated on the influence of distinct, significant events of disaster that are more characteristically affect the health of people. It is not only the immediate effect as observed in [4] and experience towards one disaster in the past rises the acute illnesses like fever, diarrhea, and severe illness in respiration for children within five years. The household socioeconomic status affected by the nature and magnitude of effects. Methods to proficiently collect, discover, search, organize, and distribute real disaster data have developed the main national concern for effective management of crisis and recovery tasks in disaster.

It means that general social media usage is getting more and more important to share information. For handling the vast data, social media information has to be streamed with low volume with a model of valuable information. As possibly massive useful data produced on social networks variety, direct social media usage is unviable to extract relevant information. Henceforth, progressive filter methods are essential [11].

Huge messages on social media platforms like Facebook, Twitter, Ushahidi etc. are immediately produced during and after large disasters for real-time information exchange about condition improvements, by persons over the affected areas. The primary added value of this content is that: a) It is nearly real-time, and typically faster than mainstream news. b) It potentially contains fine-grained, factoid information on the situation on the field, far more densely distributed geographically than official channels from the management of crisis organizations. However, numerous problems deter social media from explaining its full possible for applications in real-world. Initially, the content is enormous, comprising a high information duplication rate, because of several people reporting content about the same fact, partly due to platform-specific content-linking practices, such as re-tweeting. Secondly, as messages from social media can be created with some metadata that is platform-specific, an essential content part is still encoded in natural language text [16].

Event detection and tracking in social media have gained increased attention in last few years, and many approaches have been proposed, to overcome the information overload in emergency management for significant scale events, and to improve the sensitivity of small case incident monitoring. As information and communication technology (ICT) becomes more pervasive, we can access information about the world in ways and with speed never before possible. Micro-blogging was one social media which is quickly implemented. It offers ways to retrieve produce and spread information; the nature of that sharing has a lifecycle of information production and consumption that is rapid and repetitive. Gradually, micro-blogging is reflected as an emergency communications source because of its increasing ubiquity, communications speediness, and accessibility through cross-platform [17].

Different approaches vary concerning the level of analysis they perform on crowd-sourced content and the amount of structure they extract. Micro-blogs are gradually attaining attention as a significant information source in the management of the disaster [14].

Tweets are the best recent and comprehensive current events information and commentary stream. However, they are also noisy and fragmented, and encouraging the systems need to extract, categorize and aggregate essential events. Unique characteristics of Twitter present fresh opportunities and challenges for extraction of open-domain event [13].

For the past few years, Twitter is one of the standard media for news and communication. Twitter, a micro blogging service in which registered users can post messages called tweets [9]. Registered users can broadcast tweets, follow other users' tweets etc. Twitter messages are only 280 characters long, and they are called tweets. Tweets can be published from multiple platforms and devices. The advantage of Twitter is that anyone can follow anyone on public Twitter. Tweets are delivered to users in real-time. To connect to a general topic, users can add a hashtag as keywords to their posts. The hashtag is a Meta character which is expressed as #keyword or #hash tag. The hashtag helps people to pursue their interested topics very easily and quickly. The hashtag will provide tweets related to a standard or particular subject. E.g. the keyword #flood will retrieve all the tweets that contain the keyword flood. During natural calamities such as flood, earthquake or hurricane, at places in which the traditional connection has been down universally, these tools of social media seem to be extra beneficial [10].

Twitter shows a leading role in the rapid information propagation at the time of disasters [5, 12]. It allows accessing or dispersing crucial information or breaking news directly from the affected areas.

Social media analysis is the process of collecting information or data from social media sites such as Twitter, Facebook, LinkedIn, YouTube etc. and a study was carried out to get meaningful or useful outcomes from it. Twitter exploration comprises inspecting the tweets or matter [8].

Several machine learning algorithms such as Decision Tree, Support Vector Machine, Random Forests, Naive Bayes, Logistic Regression etc., can be applied to the data for analysis. These algorithms aid in gaining useful results from the data and help in data visualizing in a precise manner.
Nonetheless, a comprehensive standard ontology compassing all those subject areas does not exist, while different sub-ontologies cover critical subject areas (such as Resource, Damage and Disaster). It poses an issue of ontology mapping and integration for a crisis response information system which would like to make use of Linked Open Data for its data sharing.

III. METHODOLOGY

A. Analytical Process

The tweets related to cyclone Fani were collected using twitter API (Application Programming Interface). From the entire data set collected from twitter API only the tweets with geospatial information, which is from India dated from 03rd May 2019 to 06th May 2019 were analyzed. Further, the data obtained after passing through those three filters were analyzed to extract the sentiment conveyed by the users. The pipeline to be followed is shown below:

B. Pre-processing of the data

The data obtained after the filtering still has lot of noise. Thus, the tweet needs to be pre-processed before it is analyzed to get accurate results. There are a lot of established pre-processing methods to process text. However, social media text differs from normal text as social media text has emoticons, URL's, Capitalized words, slang words, punctuations, numbers, and many other forms of text which cannot be analyzed directly. The conventional process is to remove all these forms of text, but since our sentiment severity model can make sense of emoticons and capitalized words, we do not remove them. The standard procedure after removing these forms of text is to remove the stop words like a, the, and, of and to perform lemmatization and stemming. But for this research, only stop-words are removed. Stemming and lemmatization is not performed as the words are visualized later and it is better to view whole words than their base forms.

C. Sentiment Analysis

Natural Language Processing is a machine learning process which allows the machine to understand the natural language spoken by people. Sentiment analysis is a Natural language process that deals with making the machine understand the sentiment conveyed by people in text. Any machine learning process follows supervised learning methodology or unsupervised learning methodology [7] [1] [2]. In this paper, sentiment analysis is performed on tweets to get the sentiment conveyed by the user in each tweet. This implies that the tweets should be analyzed for sentiment without having a training dataset to use for supervised learning methodology. Thus, unsupervised learning methodology is used to analyze the tweets and extract the sentiment conveyed by the user.

D. Supervised Learning

Supervised learning is a machine learning process which uses a training dataset to find the probability for each event and to use this to predict the outcome of the test dataset [7] [1] [2]. There are many algorithms to implement supervised learning like Naive Bayes algorithm, SVM algorithm and random forest algorithm [6]. In this research, there is no training dataset and even if training dataset was prepared it would not be applicable for all cases. The objective of this research is to find a metric to represent the disaster severity expressed by users of a LBSN platform. Hence, supervised learning methodology is not used in this research.

E. Unsupervised Learning

Unsupervised learning makes use of a predefined data set with assigned values to be mapped with the given dataset for the problem. This implies that this methodology does not require any training dataset and is able to predict the outcome of the given dataset by mapping it to the dataset of the model. There are a large number of algorithms to implement unsupervised learning of which this research uses lexical algorithm which maps the dataset to its predefined dictionary to find the outcome of the given dataset [1] [2]. The lexical approach in this research is implemented using VADER (Valence Aware Dictionary and sEntimentReasoner) and TextBlob. However, it is observed that VADER is more suited to the disaster management scenario [6].

F. VADER

The Valence Aware Dictionary and sEntimentReasoner is a rule-based model for sentiment analysis of social media text, which calculates the sentiment of any given text at the entity level and takes the mean of all the sentiment values of entities in the given text [6]. The dictionary VADER uses was created by using Amazon Mechanical Turk, where the dictionary is built by human raters and all their ratings are averaged to give a standardized value. The sentimental value from the dictionary ranges from -4 to 4 but, after getting the mean it is normalized using the logistic function where x is the sentimental score of the words in the sentence and a is the normalization constant and is assumed to have the value of 15 [44].

$$logistic\ function = \frac{x}{\frac{((x)2\ +\ a)1}{2}} \ldots\ldots\ldots\ldots\ldots (1)$$

As the sample size increases the value of sentiment grows closer to -1 to 1. Thus, VADER is most efficient in analyzing small samples like tweets and other social media posts [6].

Since humans' rater give value in VADER to punctuations, capital letters, degree modifiers and negators like but and not good are all taken into account unlike other analyzers. VADER gives four outputs after performing sentiment analysis,

Negative, neutral, positive and compound values of sentiment [6]. Conventionally sentiment polarity values like neutral, negative and positive are used to classify the tweets but this research uses the compound value output, which gives the Sentiment intensity.

```
def sentiment_score(sentence):
    snt = analyser.polarity_scores(sentence)
    return snt
```

G. Text Blob

The Text blob model is also a rule-based model for sentiment analysis of text. It calculates the sentiment polarity and subjectivity of the given text. The Text blob is a simple mathematical machine learning model that is able to give very high accuracy. The text is tokenized first and the tokens are assigned a polarity value. The polarity value of all the tokens in the given text are averaged which gives the sentiment value of the text. The special functionality of text blob is its modifier detection. When text blob encounters a modifier, it multiplies, or inverse multiplies the intensity of the modifier to the polarity value. The mathematical formula text blob uses is:

Polarity value of text = averaged polarity value * intensity of the modifier (inverse or direct)

This keeps the sentiment polarity of the given text as accurate as possible. Text blob is also able to do Part of Speech (POS) tagging, Language generation and many other natural language processing techniques, which VADER cannot perform.

```
tet = TextBlob(tweet['text'])
print(tet.sentiment)
```

H. Severity Classification

The sentiment intensity value given by VADER shows us the intensity conveyed by the user. But it does not make sense on its own. Thus, we apply our severity conversion to make this data more versatile and usable. First, we observed that all the tweets with the positive sentiment had no relevance to severe distress scenarios and the neutral sentiment only provided information and not the severity of the effect of disaster. Thus, any sentiment value equal to or greater than 0 is converted to 0 in the severity scale. Now, only considering the negative sentiment intensity value we normalize it to a scale of 0 to 10. First, the mean of the sentiment intensity values is calculated. Then the disaster factor, severity value and the severity score were calculated using the formula given below and is shown in Table 1.

Severity value= (sentiment intensity * scale constant (10)) / Model range (1)

Disaster factor = severity value – calculated mean

Severity score = Disaster factor * Severity value

Model range - each sentiment analysis model has different range of values for sentiment intensity like -4 to 4 and -1 to 1. Now we have calculated the severity value for all the tweets, which can be communicated to the disaster management officials so that they can send aid to the locations with high severity value.

```
calculated_mean = ( Sum of all sentiment scores / number of sentiment scores )
scale_constant = 10
Model_range = 1
def severity_score(sentiment_intensity):
    severity_value = (sentiment_intensity * scale_constant) / Model_range
    disaster_factor = actual_mean - calculated_mean
    severity_score = disaster_factor * severity_value
        x = severity_score / 10
    y = round(x,1)
    return y
```

IV. RESULTS & DISCUSSIONS

Wordcloud is a new data visualization technique, which has, risen to prominence recently. It is able to convey large number of information to the viewer just by different styles of visualization. The words with the highest frequency are the largest in size. From this word-cloud, it can be observed that cyclone Fani is the word with the highest frequency. Within this word-cloud, it is also observed that Bhubaneshwar, Odisha and Puri are the most frequently named places which indicate that these are the most affected areas. It is also seen that electricity and water issues are present and that has been a devastating disaster event. Thus, wordcloud is able to convey a lot of information with just one graph as shown in (Fig 1).

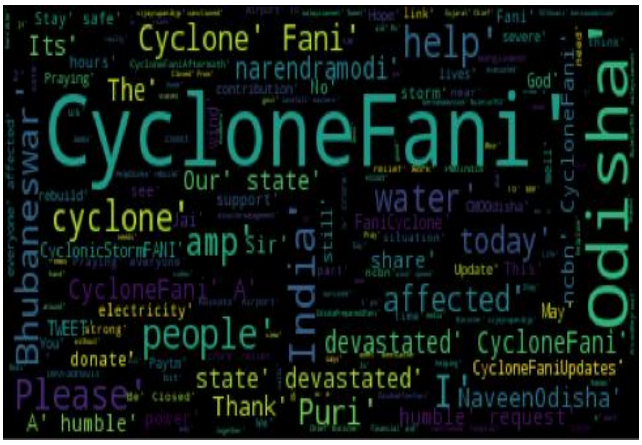

Fig. 1.Wordcloud that visualizes the word and its frequency

A. VADER for Sentiment Analysis

The unsupervised learning method for sentiment analysis can be implemented using many frameworks of which VADER and Text Blob are the most widely used models [12] [5].

This paper has performed sentiment severity using both models, thus the performance of both the models on cyclone Fani data can be measured. We use a simple mathematical accuracy calculation in-order to analyze the performance of the models. The data is classified based on sentiment and the classified data is saved in separate files. The models used at the token level for their sentiment polarity check the data in these files again.

Table- I: Dataset containing sentiment and severity scores

ID	Follower Count	Sentiment	Severity
alpharay63	336	0	0
spriyadarshin10	983	-0.4215	1.6
dkinwild	354	0.5859	0
dharitri	938	0	0
satishscorpion	884	0	0
AKT_BJP	947	0	0
stevereports	3501	0.4767	0
latestlyHindi	69	0	0

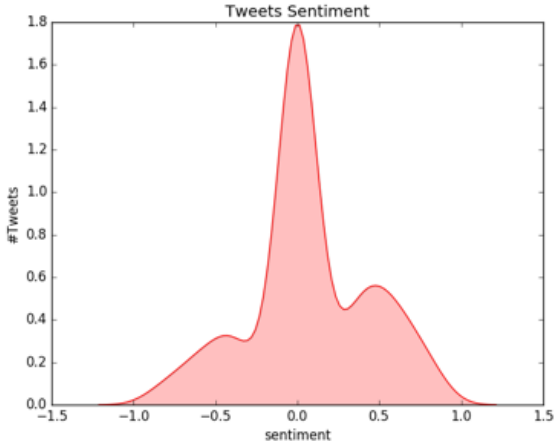

Fig. 2.Graph that visualizes sentiment intensity of tweets

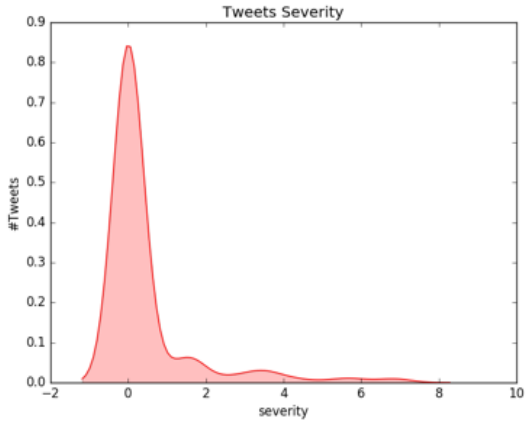

Fig. 3.Graph that visualizes the severity of tweets

The mathematical formula used is
(Correct sentiment / sentiment classified) * 100

Using this formula, the accuracy of the models is calculated [18].

Here, it is observed that Text blob gives 100 percent accuracy while VADER does not [18]. It is also observed that Text blob throws a lot of data while VADER does not. This implies that TextBlob does not analyze text it doesn't know while VADER does and since social media data has most of its text content in slang, VADER is the better choice in Social media analysis than Text blob as it is able to classify a significantly larger data-set when compared to Text blob. It is also observed that VADER is able to accurately classify a negative sentiment when compared to a positive sentiment. Considering the disaster management case where the negativity is the critical feature, which is to be analyzed, VADER appears more suitable to disaster management than any other model in NLP for sentiment analysis [18]

V. CONCLUSION

The research methodology proposed in this paper has various advantages as discussed above but it also has some limitations. The pre-processing of a social media text is a huge hurdle as the text forms is constantly changing and evolving with new acronyms and slang words with different meanings being used. It is also noted that other than VADER most of the unsupervised severity sentiment models are unable to identify and correctly classify a double negative as a positive. The other limitation is that the sentiment expressed by a user not directly related to the event of query may not necessarily be the sentiment felt by the people directly related to the event of query. The gap in this research is the lack of semantic understanding of the text. Future research can fill this gap by semantic analysis on sentimentally classified data related to disasters to achieve even more functionality to the machine learning method of language processing of social media content for efficient disaster management.

REFERENCES

1. Bing Liu, "Sentiment Analysis and Subjectivity", from Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.
2. Bo Pang and Lillian Lee (2008), "Opinion Mining and Sentiment Analysis", Foundations and Trends® in Information Retrieval: Vol. 2: No. 1–2, pp 1-135.
3. Darcy Reynard and Manish Shirgaokar, Harnessing the power of machine learning: Can Twitter data be useful in guiding resource allocation decisions during a natural disaster?, Transportation Research Part D: Transport and EnvironmentVolume 77, December 2019, Pages 449-463
4. DatarAshlesha, Liu Jenny, Linnemayr Sebastian, Stecher Chad. The impact of natural disasters on child health and investments in rural India. SocSci Med 2013;76:83–91.
5. Feng Yang. andYouquan Chen. 2010. Ontology based application framework for Network Education Resources Library. In Proceedings of 2 nd International Workshop on Education Technology and Computer Science,423 – 426.

6. Gupta, K. (2007). Urban flood resilience planning and management and lessons for the future: a case study of Mumbai, India. Urban Water Journal, 4(3), 183-194.
7. Hutto, C.J. & Gilbert, E (2014)VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, In: Proceedings of the 8th International Conference on Weblogs and Social Media, pp 216-225.
8. Iqbal, M., Karim, A., &Kamiran, F. (2019). Balancing Prediction Errors for Robust Sentiment Classification. ACM Transactions on Knowledge Discovery from Data (TKDD), 13(3), 33.
9. Kandasamy,K., P. Koroth. An integrated approach to spam classification on twitter using url analysis, natural language processing and machine learning techniques. In: Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students Conference on 2014; 1–5.
10. Nona Nader. and Rene Witte. 2010. Ontology-Based Extraction and Summarization of Protein Mutation Impact Information. In Proceedings of Workshop on Biomedical Natural Language Processing.
11. PancerasTalita., Alvin W. Yeo., and Narayanan Kulathuramaiyer. 2010. Challenges in Building Domain Ontology for Minority Languages. In Proceedings of International Conference on Computer Applications and Industrial Electronics, 574 – 578.
12. Poli, R., &Obrst, L. (2010). The interplay between ontology as categorial analysis and ontology as technology. In Theory and applications of ontology: Computer applications (pp. 1-26). Springer, Dordrecht.
13. Srikanth,A., Social media can solve many problems during natural disasters, http://infworm.com/social-media-can-solve-many-problems-during-natural-disasters/.
14. Schulz, A. Ristoski, P., Paulheim, H. (2013) 'I See a Car Crash: Real-Time Detection of Small Scale Incidents in Microblogs' Lecture Notes in Computer Science Volume 7955, 2013, pp 22-33
15. Scott Deerwester and Susan T. Dumais and George W. Furnas and Thomas K. Landauer and Richard Harshman, Indexing by latent semantic analysis, JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIONSCIENCE, year 1990, volume 41, number 6, pages 391—407
16. Shuangyan Liu and Duncan Shaw and Christopher Brewster (2013) 'Ontologies for Crisis Management: A Review of State of the Art in Ontology Design and Usability', Proceedings of the Information Systems for Crisis Response and Management conference (ISCRAM2013 12-15 May, 2013)
17. Tanghuidao Assessment on Social Network Analysis method. Academia, 2009.3 205-208
18. Wang Hong., Gao Siting., and Wang Jing. 2011. The Application Research of Description Logic in Civil Aviation Domain Ontology. In Proceedings of International Conference on Management and Service Science, 1 – 4.
19. VADER, IBM Watson or TextBlob: Which is Better for Unsupervised Sentiment Analysis?", n.d https://medium.com/@Intellica.AI/vader-ibm-watson-or-textblob-which-is-better-for-unsupervised-sentiment-analysis-db4143a39445

AUTHORS PROFILE

Dr. K. Shyamala is working as an Associate Professor in the PG and Research Department of Computer Science, Dr. Ambedkar Govt. Arts College, Vyasarpadi, Chennai, Tamilnadu, India. She has her Masters degree, M.Phil and Ph.D. in Computer Science. She has 29 years of teaching and research experience. Six candidates have completed Ph.D. under her guidance. She has authored numerous books, published 62 research articles and conducted several conferences. She has also chaired sessions in International conferences. She has served as program committee member and chairman for Board of Studies in various colleges and universities. Her area of specialization includes Data Mining, WBAN, Agent Based Computing and Advanced Computer Networks.

Vijay Kumar Kannan, Research Scholar, Department of computer science school of computing sciences, VELS Institute of Science Technology and Advanced Studies, Pallavaram, Chennai, India. Area of specialization is on Big Data Analytics, Data Mining and Text mining on Social Media Applications.

Sheran Dass. D, B.Tech in Electronics and Communications Engineering from Vellore Institute of Technology, Vellore, India, currently pursuing M.S in Software Engineering from Arizona State University, USA. Research contributions in the field of NLP was made while working as a Research Intern at C-DAC, Chennai under the guidance and supervision of Vijay Kumar Kannan. Won the best machine learning application award from Amazon at Sunhacks 2019. Presently conducting research in the field of machine learning with a focus in the field of medicine. Working as a researcher of the Geometry Systems Laboratory at ASU under the guidance of Prof. Yalin Wang.