International Journal of Applied Research 2015; 1(7): 391-394

ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 5.2
IJAR 2015; 1(7): 391-394
www.allresearchjournal.com
Received: 17-04-2015
Accepted: 20-05-2015

Suman
Student, M. Tech
Deptt. of Computer Science & Engg., Manav Institute of Technology & Management, Jevra, Hisar (Haryana)

Madhurima
Asstt. Professor,
Deptt. of Computer Science & Engg., Manav Institute of Technology & Management, Jevra, Hisar (Haryana).

Dr. Vijay Bhardwaj
Co-Guide, Associate Professor & HOD, Deptt. of Computer Science & Engg., Manav Institute of Technology & Management, Jevra, Hisar (Haryana)

Correspondence:
Suman
Student, M. Tech
Deptt. of Computer Science & Engg., Manav Institute of Technology & Management, Jevra, Hisar (Haryana)

Text mining model to identify criminal activities

Suman, Madhurima, Vijay Bhardwaj

Abstract
Today most of the information communication over the web is in textual form. This communication includes emails, chats, tweets etc. These easy means of communication reduced the sensitivity while communicating with a person. Because of this sometimes the misuse of these kind of medium is done. This misuse can be identified in terms of crime messages, spam messages etc. These message contents can be an advertisement, threat, blackmailing etc. To improve the communication reliability it is required to identify these kinds of messages. In this paper work, a statistical sentiment analysis approach is defined to identify these kinds of spam and invalid messages. The work is divided in two stages. In first stage, the statistical analysis over the textual form is done. This form includes the identification of relevant message including the positive aspect messages and negative aspect messages. The aspect criticality is also considered. To consider this criticality, a weighted approach is defined. In this work, a fuzzy adaptive approach is defined to identify the message weights and identify the message sensitivity. The work is applied on some real time textual messages obtained from web sources. The obtained results show that the work has clearly identified the hidden message sentiment.
Keywords: Web Communication, sentiment analysis approach, weighted approach, Fuzzy adaptive approach, positive aspect, negative aspect.

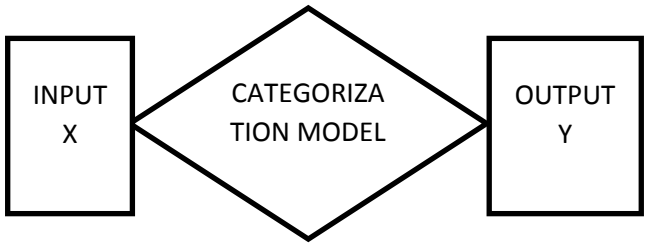
Keywords: Criminal Activities, Text Mining, communication.

Introduction
Sentiment Analysis comes under intelligent textual processing to acquire the effective information over some message and present it in some decision form. This decision can be the answer to some question in terms of acceptance or rejection. Sentiment Analysis comes under document or text information processing and it provides the speaker level flow in the architecture so that the effective information transition will be obtained. The text is here defined respective to some topic based on which the actual information evaluation can be obtained. The sentiment extraction over the text is always a critical and challenging task.
In this paper, the sentiment analysis is performed on product review collected by an organization from its customers. The reviews are here defined under evil character generation and story line specification. The work is here defined to obtain these all consideration about the products and product features and assign them relative based on associated word level sentiment specifications. Once these all sentiments are attached, the next work is to generate the overall sentiment specification so that the overall product sentiment will be obtained from the work.
The categorization is a supervised form of machine learning. Machine learning comprises of supervised, unsupervised, semi- supervised and reinforced learning. In the supervised form of learning the learning is from the training data available.

Supervised Learning: Supervised Learning is effective information processing and learning method to process the data and provide the behavior analysis in complex system and environment. This learning method defined with the predictor specification that provides the function specification to generate the hypothesis to find the eligibility to the data class so that effective object derivation and mapping will be obtained.

Unsupervised Learning: Unsupervised learning is defined as the behavior specification so that the data learning and processing can be obtained to generate the output by analyzing the

input on its own. The following figure shows how unsupervised process is performed.



In the above diagram you can see that in unsupervised learning a data is unknown and needs to be categorized and similar kind of data has to form a separate cluster. By applying the machine learning algorithm one can separate out the clustered data and hence identify the different forms of data.

Categorization is one of the main task and form of data mining techniques. It is we have input data field, denoted by X., then an output data field, denoted by Y., a characterization model is placed in between input and out which define the type of characterization that is taking place.

Existing Work

In this section we briefly present some of the research literature related to minimize criminal activities. This kind of information processing is defined along with the identification associated relation words to the document and to generate the associated polarity based on which estimation and effective class derivation will be done. This learning mechanism is defined to provide the information abstraction so that the early information processing will be done [1] [2] [3]. Qiu [4] has defined a work propagation analysis so that the word sentiment extraction based on polarity analysis will be performed and this kind of analysis is defined at topic and document level. Author has presented a work on machine learning algorithms to processing with sentiment extraction and identification. Whitelaw [5] has defined a work on the sentiment processing based on group generation under theoretical frameworks. Author separate the associated adjectives based on the feature specification so that the word feature based extraction and word review analysis can be done. Author worked on different critics and message for message reviews. Mullen has presented a SVM based approach information processing defined at the topic specification. Author defined a measure specific formation of word partition generation so that the extraction of relative object and subjectivity analysis will be performed and the polarity of word will be identified based on which the word class identification and classification will be done [6]. Hu and Liu [7] and presented a summarized approach customer review generation and processing so that the message formation and opinion generation will be done in an effective way. Author defined the message review processing under the abstraction no positive and negative review or aspect generation. Blair Golesohm [8] has presented a summarization effective work to analyze the customer review collected for the hotel service and provide the analysis based on the algorithmic approach and also generate the aspects related to the service also identify the relevant information and polarity values based on the message strength analysis. This kind of derivation is defined under processing the positive and negative message aspects so that effective information processing and message derivations will be obtained. Yi [9] has defined a work on message separation and formation applied on news articles to identify the impact on the society. This kind of impact analysis is here derived in the form of syntactic parser so that the sentiment lexicon based polarity analysis and classification is defined. Miyoshi [10] has presented the sentiment class identification based on the message level analysis so that the customer review processing can be done in a contextual form and the change in the customer mood can be identified. Zhang [11] has presented a work on content level analysis defined under rule derivation and filtration so that the abstracted and aggregative information formation will be done in an effective and innovative way. This kind of information processing includes the weight adjustment information processing and the contribution derivation based on which the sentimental polarity analysis is done. N N Shaikh [12] defined a work on the derivation of sentence level message formation so that the domain and application specification analysis is defined. Author used the sensenet tool to obtain the derivation and provide the effective visualization on these words and also provided the aspect generation over the messages. This kind of analysis is defined as the score specification to the message.

Problem Defination

Information transmission over the web is one of the safe and effective ways for conversion. But in last few years, there are various side effects are identified from this social media conversion. The social media has increased the crime rate and its having a high participation in crime. This kind of crime includes digital theft, blackmailing etc. By analyzing the conversion of users, some crime can be identified and can be stopped from happing. In this work, a text mining approach is defined for social media to analyze the conversation messages and to identify the abnormal chat pattern. To analyze this, a layered approach is defined in this work. In first layer of this work, the backend dataset is generated in which the social media keywords are defined under the specification of category. This manual categorization is based on the type of chat message and its criticality level. Once the backend dataset is defined in specialized way, the next work is to process the chat message and extract the keywords from it. In second layer of this, the clustering will be performed on these keywords based on topic specification. To perform this categorization an entrophy value and topic similarity analysis approach is defined. Once the chat topic is identified, the next work is to match the words respective to specific topic and identify the hidden criticality in the message. The work will categorize the message as the crime or non crime message as well as the type of crime associated in the message. To identify the criticality of the words and the polarity, the fuzzy system based analysis will be performed. Finally the crime activity of the user will be identified. The work will be implemented in java environment.

Research Methology

In this present work, a layered model is presented to identify the crime involvement of a user who is performing any kind of online conversation on social media. The work will use a two stage model to identify the conversation topic as well as the crime involvement of a web user. The work will also identify the type of crime associated with the message. To

perform this analysis, in first phase, the keyword based text clustering approach is defined. This clustering will be based on the entropy value analysis and similarity analysis. Once the topic of conversation will be identified the next work will be to identify the crime based messages as well as user based on polarity and priority analysis. The work stages of this work are given here under

DB Generation: The first stage of work is to generate the back end db that can represent the social media conversation topics as well as frequently used words. These words will be represented under the topic and criticality based. This topic will also include the crime message along with crime type.

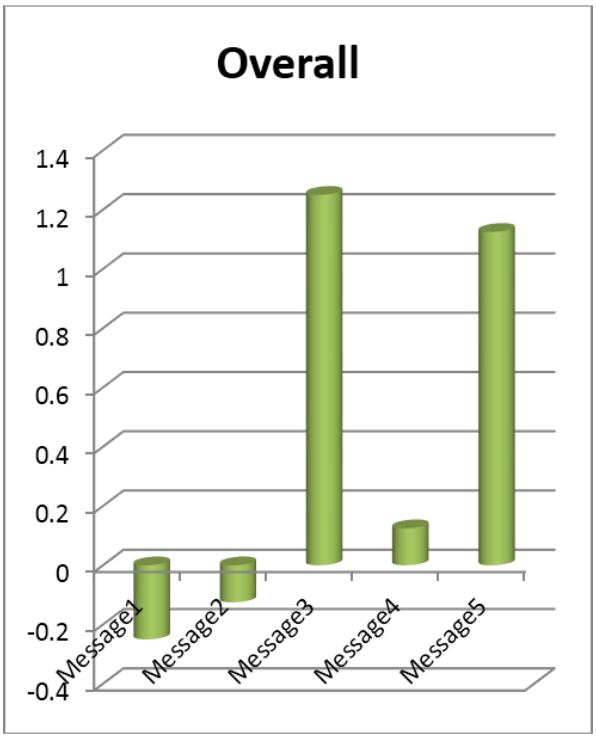
Keyword Analysis: In this work, the social media message will be processed to extract the keywords. This word includes the filtration stage in which, the stop list words will be removed from the message and the keywords will be identified. These keywords will be identified along with the topic specification.

Clustering: The third stage includes the text based clustering to divide the chat message in various related clusters. The clustering will be defined here under the similarity analysis and entropy analysis approach.

Criminal Activity Identification: In the final stage of work, the topics will be filter against the specification of crime messages and the crime related keywords. The particular crime message category will be identified from the message.

Results

The presented work is defined to perform the message driven analysis to identify the associated sentiment in the message. These negative sentiments here define the crime intension or some wrong intension of user. The positive sentiment messages are the proper communicated messages. The analysis of work is here defined in terms of message aspect analysis and user aspect analysis



Conclusion and Future Work Scope

In this environment, different kind of messages can be transmitted using emails, chats, sms, blog post etc. But some of the users can include some irrelevant messages or the crime inclusive contents in these messages. This kind of message includes abusive language, threats, blackmailing etc. Because of this, there is the requirement of some intelligent approach that can real and analyzes this kind of message and identifies the crime intension of a person. In this present work, a statistical weighted measure approach is defined under fuzzy rule to identify such messages. The work is here defined to perform the fuzzy adaptive keyword criticality analysis so that the positive or the negative aspect from message will be identified. The work is implemented in user friendly environment. The work is here applied in different forms with different graphical interfaces. In first model of this work, the individual message analysis is done. In this stage, the message criticality is under different associated aspects and the sentiment of single message is obtained. In second model, the user based analysis is performed. In this form, all the messages submitted by a particular user are analyzed and obtained the hidden sentiment. In this stage, user driven sentiment analysis is obtained. The obtained results show that the presented work has provided an effective solution to identify the crime sentiments from different messages.

References

1. Farkhund Iqbal. Mining Criminal Networks from Chat Log, 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology 978-0-7695-4880-7/12© 2012 IEEE.
2. Dragoljub Pokrajac. Incremental Local Outlier Detection for Data Streams, IEEE Symposium on Computational Intelligence and Data Mining (CIDM), April 2007.
3. Simon Hawkins. Outlier Detection Using Replicator Neural Networks.
4. G. Qiu, B. Liu, J. Bu, C. Chen. Expanding domain sentiment lexicon through double propagation, Proceedings of the 21st International Joint Conference on Artificial Intelligence (Morgan Kaufmann, San Francisco, 2009).
5. C. Whitelaw, N. Garg, S. Argamon. Using appraisal groups for sentiment analysis, Proceedings of the 14th ACM Conference on Information and Knowledge Management.
6. T. Mullen, N. Collier. Sentiment analysis using support vector machines with diverse information sources, Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing, 2004.
7. M. Hu, B. Liu. Mining and summarizing customer reviews, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.
8. S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G.A. Reis, J. Reynar. Building a sentiment summarizer for local service reviews, Proceedings of WWW 2008 Workshop: NLP Challenges in the Information Explosion Era, 2008.
9. E. Yi, J. Wiebe. Learning extraction patterns for subjective expressions, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing 2003
10. J. Wiebe, E. Miyoshi. Creating subjective and objective

sentence classifiers from unannotated texts, Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, 2005.
11. C. Zhang, D. Zeng, J. Li, FY. Wang, W. Zuo. Sentiment analysis of Chinese documents: from sentence to document level, Journal of the American Society for Information Science and Technology, 2009; 60(12).
12. M. Shaikh, B. Liu. Mining opinions in comparative sentences, Proceedings of the International Conference on Computational Linguistics, 2008.