**Spam Detection in Social Networks: An Algorithm and Data Mining Approach**

Balogun Abiodun Kamoru1 , Azmi Bin Jaafar2, Masrah Azrifah Azmi Murad3, Marzanah A Jabar4, AbduMajid Babangida Umar5

1,2,3,4,5 Department of Software Engr and Information Systems, Faculty of Computer Science and Information Technology, Univerisiti Putra Malaysia, Serdang 43300 Selangor Malaysia.

E-mail: [Balogun@consultant.com](mailto:Balogun@consultant.com), [Azmij@upm.edu.my](mailto:Azmij@upm.edu.my), [masrah@upm.edu.my](mailto:masrah@upm.edu.my), marzanah@upm.edu.my,bumar05@gmail.com

**Abstract**

The problem with Social networks has been a major issues globally. In recent years, Spam Detection on social networks has been focused. However, Spammer has seen that social networks are vulnerable to attack in order to perpetrate their evil. Influx of spam has been a great threat to individual, organization, government, institution if left unchecked, spam threatens to undermine resource sharing, interactivity, and openness.

Due to the ubiquitous use of social networks it has generated huge amount of social data which gives the spammer the leverage to performance various forms of malicious attack and spam activities.

This paper survey three computational categorization issues on social networks like Size, Noise and Dynamism, due to the issue often experience on social networks data that are complex to analyse. The paper talk on the Various Data mining techniques used in mining diverse aspects of the social networks site analysis over decades going from the inception to the up to date models, include use of novel algorithms likes Porter Stemmer algorithm, TF-IDF algorithms that are proposed.

General Terms

Porter Stemmer Algorithm (PSA), TF-IDF Algorithm,

Keywords

Spam, Social network sites, TF-IDF( Terms Frequency –Inverse ,GAD (General Activity Detection)

1. INTRODUCTION

In the recent years, spam activities has tremendously increase on social networks sites. Balogun et al 2017, Define Social networks as a term used to describe web-based services that allow individuals to create a public/semi-public profile within a domain such that they can communicatively connect with other users within the network. Social networks has improved on the concept and technology of web 2.0, by enabling the formation and exchange of user-Generated content Kaplan et al,(2010).

Xin Jin et al,2015 Define Data Mining as a data or knowledge discovery, it allows users to analyze data from various dimensions/angles, categorize it, and summarize the relationships that have been identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Social networks sites has become very popular in the last decade, it is more affordable to access social networks sites like Facebook, Twitter,Sina weibo, LinkedIn , Tagged and Badoo,e.t.c. through the internet. people rely on social networks sites for news, information and opinions of other users on various subjects.

Social network is a term used to describe web-based services that allow individuals to create a public/semi-public profile within the domain such that they communicatively connect with others users within the network [3]. Social network have gone from being a small niche of the web to one of its important components. Social network has really upgrade on the concept and technology of web 2.0, by giving enabling environment for the formation and exchange of user-Generated content [9], social network is a graph consisting of nodes and links used to represent social relations on social network sites.

Spam Detection on social network will be highly focus, Spam can be in the form of images, text, videos etc. Social network are prone to malicious attack. Spam may contain virus links which could lead to personal or business loss[1]. There have been extensive research on email spam detection[5,6]. Spam messages[15], spam images [10], spam video [11], web spam [12], spammers[13][7][2] etc.

Social networks are important sources of online interaction and communication sharing[], subjectivity [1]. assessments [5], approaches [37] , evaluation [24], influences [27], observation [30], opinion and expressions []. Social network platforms enable rapid growth of information exchange between users regardless of the location. Social network enables big organizations, celebrities, government official and government institution to obtain knowledge on how their audience reacts to postings that concerns them out of the enormous data generated on social networks, however, it is possible for the spam to prey on any individuals.

Data mining techniques have been capable of handling the three computational dominant disputes with social network data like ; **size, noise and dynamism.** The voluminous nature of social network datasets require automated information processing for analysing it within a targeted time. Basically, data mining techniques also require huge data sets to mine remarkable patterns from data; social network appear to be perfect sites to mine with data mining tools [10]. Data mining tools surveyed in this paper ranges from supervised, un-supervised and semi-supervised machine learning. A table itemizing the previous techniques is covered in this paper.

***The rest of the survey is organized as follows. Section 2 examines the problem with spam detection on social networks. Section 3 discuss the algorithms and flow chart for spam detection on textual feature with the proposed clustering algorithm. Section 4 present supervised and un-supervised classification techniques employed in social networks analysis. Section 5 discuss data mining tools used to analyse social network spam. While the section 6 is conclusion work and direction.***

**2. PROBLEM AND ISSUES WITH SPAM DETECTION ON SOCIAL NETWORKS**

Quite number of research works have talk about spam issues and challenges. One approach is to view the spam **problem as a classification problem**: given a list of objects, partitions it into two lists, one list of spam content and one list of non-spam content the traditional concepts of precision and recall:

Precision: of the good (or bad) content that we acted on (identified, changed the rank of, or prevented) as good or bad what percentage actually good?

Recall: of the good or bad content in the system, what percentage did we act upon as good or bad. **(Paul Heymann et al, 2007)**

Another approach is to view the **spam problem as a ranking problem:** given a list of objects, compute a rank for each object. This makes the most sense for evaluating demotion-based strategies. In this case we could take into consideration spam objects and their list position. the issue of spam factors sets in, it measures the spam for a set of ranked results**.(Paul Heymann et al,2007)**

Another approach **is degree of accuracy problem** among the concurrent numerous spam content and non content, limiting user actions or difficult to change user interfaces. tagging system in which users annotate objects like photo and likes.(://paulgraham.com/antispam.html)

Analysis of the linkage behaviour of the social network so as to ascertain relevant nodes, links, communities and imminent areas of the network i.e linkage based and structural analysis problem. Dynamic and static analysis which involves presumed networks, while the social networks change with time and over time. **Papadopoulous et al; 2012.** There are some major problem in the spam detection of fake review is the unavailability of ground truth datasets for model training (Ma and Li,2012), and the situation is worse in the case of an imbalance in the data of fake and truthful reviews [29]**.** Spam is in different forms like images, text, videos etc. So Social networks websites need to be untainted for long term success **[16]**

**3. PROPOSED CLUSTERING ALGORITHM**

**3.1 Data Mining used**

* **NLP Parser (Natural language Processing)**
* **TF-IDF Algorithm ( Terms Frequency-Inverse Document Frequency)**
* **PSA algorithm (Porter Stemmer Algorithm)**

**3.1.1 NLP Parser**

it stands for Natural language Processing, It is an Arena in Artificial intelligence *(AI)* and linguistics. It Pertains to interactions between computers and human i.e. Natural language. It has its own vocabulary dataset against which it checks for the semantics of words present in a given document.

A natural language parser is a program that figures out the grammatical structure of sentences for instance, which groups of words go together ( as "Phrases") and which words are the Subject or Object of a verb. The knowledge of language obtained from hand-parsed sentences is used by the probalistic parsers which try to yield the most likely analysis of new sentences. The development of NLP parsers was one of the biggest advancements in Natural Language Processing in 1990s.

**3.1.2 TF-IDF**

TF-IDF stands for Term Frequency-Inverse Document Frequency. TF-IDF weight is a weight used in data text mining and information retrieval. It is a statistical measure which is used to ascertain how important a word is to a document in a collection or corpus, The Importance of a word is directly proportional to the number of times it appears in the document. There are certain variations of TF-IDF weighting scheme which are often used by search engines as a central tool to score and rank a document's relevance, when given a user query. Summing the TF-IDF for each query term in a query is one of the simplest ranking functions. TF-IDF can be used successfully for filtering of stop-words.

**3.1.3 PORTER STEMMER ALGORITHM**

Natural language texts usually contain many different variants of a basic word.

Morphological Variants for examples are COMPUTATIONAL, COMPUTER, COMPUTERS, COMPUTINGS etc are generally the most common, with other sources including valid alternative spellings, mis-spelling, abbrevation, etc,

In a typical IR environment, one has a collection of documents, each described by the words in the document title and possibly by words in the document abstract. Ignoring the issue of precisely where the words originate, it can be said that a document is represented by a vector of words, or terms. Terms with a common stem will usually have similar neanings.

for example:

CONNECT

CONNECTED

CONNECTING

CONNECTION

CONNECTIONS

frequently, the performance of an IR system will be improved if term groups such as this are conflated into single term. This may be done by removal of the various suffixes -ED, -ING, -ION, IONS to leave the single term CONNECT. the suffix stripping process will reduce the total number of terms in the IR system, and hence reduces the size and complexity of data system, which is always advantageous [18]

**3.1.4. GAD(General Activity Detection) CLUSTERING ALGORITHM**

GAD clustering is a data mining techniques widely used in numerous applications. It has also been studied in research areas such as biology, pattern recognition, statistics, machine learning, information retrieval, market research and multi media research [2]. Many papers have been published form fast clustering on large data. Some develop fast core clustering algorithms; some develop pre-processing methods, such as sampling, sub- space and compression in order to reduce the data to smaller size to achieve speedup.

Within the GAD framework a set of algorithms have been designed for different scenarios:

1. Exact GAD algorithm E-GAD, which is much faster than K-means and get the clustering result [6]
2. Approximate GAD algorithms with different assumptions, which are faster than E-GAD while achieving different degree of approximation.
3. GAD based algorithms to handle the "large clusters" problem which appears in many large scale clustering applications.

**3.1.5 J48 DECISION TREE ALGORITHM**

The data mining model using the decision tree J48 is created using WEKA. WEKA stands for ***Waikato Environment for Knowledge Analysis*** and has been developed by the University of Waikato, New Zealand.

WEKA is a toll which comprises 60 machine learning algorithms. They developed a prototype system which detected spam on the Facebook. The application runs on the server to where a Facebook request is rerouted for carrying out the process of spam checking. The decision tree model J48 was used for classification. The selected attributes were the number of keywords, the number of links, the length of the post, and the average number of words in a post [3]. The relationships of the attributes were further to be scrutinized in the future. The goal of this work was to only demonstrate the use of data mining model in detecting spams in the Facebook application. The model trained 150 sample posts and tested 75 posts. The features used included the number of words ,the length of the post, and the number of the links nodes with the recall rate of 66%.

* + 1. **K-MEANS ALGORITHM**

K-Means algorithm is a simple iterative method for partitioning a given dataset into a user-specified number of clusters, say K, It was discovered by many researchers belonging to different disciplines, especially, Lloyd (1957, 1982) Forgey (1965), Friedman and Rubin (1967). In database management, clustering of data is done. It is the process of dividing the data elements (input data) into groups so that the items in the same group are as similar as possible and items in different groups are as dissimilar as possible [5].

This method utilizes the K-means clustering algorithm to group the messages or emails based on the similarity of their attributes or features into K disjoint groups to improve the accuracy of spam detection. In classification, the objects are assigned to predefined classes; whereas in clustering, the classes are formed, in which, a data point can belong to only one cluster.

* + 1. **FLOW CHART FOR SPAM DETECTION ON CONTENT TEXTTUAL FEATURE**

Messages, comments and reviews are extracted from Social Networks Sites database. Two parallel activities are carried out.

1. The extracted data is given to Keyword parser,
2. The extracted data is given to the NLP parser as well,

**Social Networks Sites, Reviews, Comments, Data sets**

**Message**

**Keyword Parser**

**Tokens**

**Filtering**

**NLP Parser**

**TF-IDF(Terms Frequency Inverse Document Frequency**

**Grammar/Construct**

**Score Ranking**

**Spam Detection**

**Filtered Messages**

**Final Features**

**Messages Category**

**Identification**

**Filtered Keywords**

**Fig1 : Flow Chart Diagram for Spam Detection in Content Textual Feature**

**3.1.8 *EXPLANATION AND ACTIVITY***

Messages, comments and reviews are extracted from the social networks sites database, Two parallel activities are carried out:

1. The extracted data is given to Keyword parser and
2. The extracted data is given to the NLP parser as well.

***Activity 1***

Java’s in-built function **string tokenizer** will take data in the sentential form and break it into tokens (words. Phrases, symbols and other meaningful elements).

* Filtering function will separate out the regular English like “is”, “an”, “ the” etc. from the given input tokens and produces output as **“filtered tokens”.**
* Filtered tokens act as “terms” for ***TF-IDF algorithm.***
* TF-IDF algorithm produces a score rank for each and every query term.
* The ***score ranking*** will again apply filtering function which will eliminate all the words (terms) which never occurred or appear the least number of times.

**Activity II**

* **NLP Parser** will check for the semantics of the queried words.
* It has its own vocabulary dataset against which it will check for the semantics of words present in a given document.
* ***From the results of activities 1 & 2, Spam detection is done on the given input***.
* Finally, a dataset of filtered messages is obtained from the result.
* The conclusion part and technical details of the system will be presented, including spam features, algorithms and efficient implementation. The Hypothesis behind the design will be analysed, especially on the scalability and accuracy issues, in order to show how this system can handle a huge number of posts and monitor real-time social activities in social networks to identify spam. This method currently detects spam only in text but can further accommodate other features like images, videos and social networks features as well, The complexity of this approach is low and it can be used in reality easily.

1. **SUPERVISED CLASSIFICATION OF SOCIAL NETWORKS**

Clustering techniques are used where basis of data is established but data pattern is unknown [4], classification techniques are supervised learning techniques used where the data organisation is already identified. It is worthy of mention that understanding the problem to be solved and opting for the right data mining tool is very essential when using data mining techniques to solve social network issues. Pre-processing and considering privacy rights of individual. Nonetheless, since social network site is a dynamic platform, impact of time can only be rational in the issue of topic recognition, but not substantial in the case of network enlargement, group behaviour/influence or marketing. This is because this attributes are bound to change from time to time. Information updates in some social network such as Twitters, Facebook, Sina Weibo present APPLICATION PROGRAMMERS INTERFACES (API’s) that makes it possible for crawler, which gather new information in the site, to store the information for later usage and update.

In [17] a supervised learning algorithm used the combination of multiple base facts to label couple of adjectives having similar or dissimilar semantic orientation. The algorithm resulted in a graph with nodes, and links which represents adjectives and similarity (or dissimilarity of semantic orientation respectively.

**4.1 UNSUPERVISED CLASSIFICATION OF SOCIAL NETWORK**

A straightforward unsupervised learning algorithm can be used to rate a review as ‘thumbs up’ or ‘thumbs down’ [85]. This can be by way of digging out phrases that include adjective or adverbs (part of speech tagging )[33] . The semantic orientation of every phrase can be approximated using PMI-IR [34] and then classify the review using the average Semantic Orientation of the phrase. Cogency of title, body and comments generated from blog post has also been used in clustering similar blogs into significant groups. In this case keywords played very important role which may be multifaceted and bare [3]. EM-Based and Constrained-LDA were utilized to cluster aspect phrases into aspect categories [29].

Other Unsupervised learning used in sentiment analysis in products rating and reviews include POS (Part of Speech) tagging. In POS adjectives are tagged to display positive and negatives ones. Sentiment polarity is the binary classification of an opinionated document into a largely positive and negative opinion [7] . In review this is commonly termed with the ‘thumps up’ and ‘thumps down’ expressions as mentions earlier. The polarity of positive against negative I weighed to give an overall analysis of sentiment expressed on issue under review.

Bootstrapping also forms part of the unsupervised approaches. It utilizes obtainable primary classifier to make labelled data which a supervised process can build upon [20] [17]. Semantic orientation is also unsupervised approaches currently used for sentiment analysis is on social network sites. It attaches different meaning to single word – synonym . This could either be positive or negative. Direction and intensity of words used can determine the semantic orientation of the opinion expressed. Semi-supervised classification is discuss in section 4.2.

**4.2. SEMI- SUPERVISED CLASSIFICATION OF SOCIAL NETWORKS**

Semi-supervised learning is a goal-targeted activity but unlike unsupervised ; it can be specifically evaluated. Authors of [3] worked on a mini training set of seed in positive and negative expressions selected for training a term classifier. Synonym and antonym comparatives were added to the seeds sets in an online dictionary. The approach was meant to produce the extended sets ***P’ and N’*** that makes up the training sets. Other learners were employed and a binary classifier was built using every glosses in the dictionary for both term in P ‘U N’ and translating them to a vector. Their approach discovers the origin of information which they reported was missing in earlier techniques used fpr the task.. Semi-Supervised lexical classification proposed [35] integrated lexical knowledge into supervised learning and spread the approach to comprise unlabelled data.

In [36] semi-supervised learning uses polarity detection as semi-supervised label propagation problem in graphs. Each node representing words whose polarity is to be discovered. The results shows label propagation progresses outstandingly above the baseline and the semi-supervised techniques ***Mincuts*** *and* ***Randomized Mincuts*** The work of [38] ***compared graph-based semi-supervised learning*** with regression and [38] proposed metric labelling which runs SVM regression as the original label preference function comparable to similarity measure. Their result shows that the graph-based semi-supervised learning (SSL) algorithm as per PSP (*positive-sentence-percentage*) comparison (SSL + PSP) proved to perform well.

1. ***DATA MINING TOOLS USED TO ANALYZE SOCIAL NETWORK SITES SPAM***

* 1. ***GRAPH THEORETIC***

Graph theory is probably is the main data mining tools used to analyze social network sites spam concept. The approach is applied to social network analysis in order to determine important features of the network such as the nodes and links ( for example influencers and the followers). Influencers on social network sites have been identified as users that have impacts on the activities or opinion of other users on the network. Fig 1.

Graph theory has proved to be very effective on large-scale datasets (such as social network sites data) This is because it is capable of bye-passing the building of an actual visual representation of the data to run directly on data matrix [32] . In [19] ***centrality measure*** was used to inspect the representation of power and influence that forms clusters and cohesiveness [16] on social network sites. The authors of [34] employed parameterized centrality metric approach to study the network structure and the rank nodes connectivity. Their work formed an extension of a-centrality approach which measures the number of alleviated paths that exists among nodes.

** **

**Influencer Followers**

**Fig” 1 Graph Theoretic on Social Network Site**

* 1. **. SPAM DETECTION USING HIERARCHICAL CLUSTERING**

A community is a smaller compressed group within a larger network . Community formation is known to be one of important characteristics of social network sites. Users with similar interest form communities on social network thereby displaying strong sectional structure. Communities on social network sites, like any other communities in the real world, are very complex in nature and difficult to detect. Applying the appropriate tools in detecting and understanding the behaviour of network communities is crucial as this can be used to model the dynamism of the domain they belong [4]. Different authors have applied diverse clustering techniques to detect communities on social network sites [26] ;[27] ; [10] with Hierarchical clustering being mostly used [31]. This technique is a combination of many techniques used to group nodes in the network to reveal strength of individual groups which is then used to distribute the network into communities. Vertex clustering belongs to hierarchical clustering methods, graph vertices can be resolved by adding it in a vector space so that pairwise length between vertices can be measured. Structural equivalence measures of hierarchical clustering concentrate on number of common network connections shared by two nodes. Two People on social network sites with several mutual friends are more likely to be closer than two people with fewer mutual friends on the network. Users in the same social network community often recommend items and services to one another based on the experience on the items or services involved.

* 1. **SEMANTIC WEB OF SOCIAL NETWORK SITES**

The Semantic Web platform makes knowledge sharing and the re-use possible over different applications and community edges. Discovering the evolvement of ***Semantic Web (SW)*** enhances the knowledge of the prominence of ***semantic web community*** and envisages the synthesis of the semantic web.

The work in [29] employed ***Friend of a Friend (FOAF)*** to explore how local and global community level groups develop and evolve in large-scale social network sites on the semantic web. The study revealed the evolution outlines of social structures and forecasts future drift. Likewise [30] application model of ***Semantic Web-based Social Network Analysis Model*** creates the ontological field library of social network analysis combined with the conventional outline of the semantic web to attain intelligent retrieval of the Web services. Furthermore, ***Voyeur Server*** [28] improved on the ***open-source Web-Harvest framework*** for the collection of online social network sites data in order to study structures of trust enhancement and of online scientific association***. Semantic Web*** is a relatively new area in social network analysis and research in the field is still ***evolving.***

* 1. ***OPINION ANALYSIS ON SOCIAL NETWORK SITES***

***According to Technorati, about 75,000 new blogs and 1.2 Million new posts*** giving opinion on products and services are generated every day [25]. Massive raw data generated every minute on common social network sites are laden with opinion of users as regards diverse subject ranging from personal to global issues [24]. Users’ opinions on social network sites can be referred to as discovery and recognition of positive or negative expression on diverse subject matters of interest. These opinion are often convincing and their indicators can be used as motivation services or even endorsement of political candidate during election [21], [23]. Even though online opinion can be discovered using traditional methods, this form is conversely inadequate considering the large volume of information generated on social network sites. This Fact underscores the relevance of data mining techniques in mining opinion expressed on social network sites.

Various methods have been developed to analyze the opinion arising from product, services, events or personality review on social network sites [36]. Data mining tools already used for opinion and sentiment analysis include collections of simple counting methods to machine learning. Categorizing opinion-based text using binary distinction of positive against negative [16], [21], [20], [17], is found to be insufficient when ranking items in terms of recommendation or comparison of several reviewers opinion [11] ( e.g , using actors starred in two different films to decide which of them to see at the cinema). Determining players from documents on social network has also become valuable as influential actors are considered as variables in the documents [12] when applying mining techniques on social network sites.

* 1. ***ASPECT-BASED/ FEATURE-BASED OPINION MINING***

**Aspect-based** also known as Feature-based analysis is the process of mining the area of entity customers has reviewed [8]. This is because not all aspects/features of an entity are often reviewed by customers. It is then necessary to summarise the aspects reviewed to determine the polarity of the overall review whether they are positive or negative. Sentiments expressed on some entities are easier to analyse than others, one of the reason being that some reviews are ambiguous.

**According to [7] aspect-based opinion** problem lies more in blogs and forum discussions than in product or service reviews. The aspect/entity(which may be a computer device) reviewed is either “thumb up” or “thumb down”; thumb up being positive review while thumb down means negative review. Conversely, in blogs and forum discussions both aspects and entity are not recognized and there are high levels of insignificant data which constitute noise. It is therefore necessary to identify opinion sentence is positive or negative [4]. Opinion sentences can be used to summarize aspect-based opinion which enhances the overall mining of product or service review.

An opinion holder expresses either positive or negative opinion [13], [9], [8] on an entity or a portion of it when giving a regular opinion nothing else [8]. However, [3] put necessity on differentiating the two assignments of finding out neutral from non-neutral sentiment and also positive and negative. This is believed to greatly increase the correctness of computerised structures.

1. ***CONCLUSION AND FUTURE DIRECTION***

The technical details of the system is highlighted including spam features, algorithm and efficient implementation. The hypothesis behind the design will be analysed, especially in accuracy issues. In order to justify and show how this system can handle a huge number of posts and monitor real-time social activities in social network sites to identify spams. This method currently detects spam only I textual features but can further accommodate other features as well. The complexity of this approach is low and it can be used in reality easily,

Different data mining techniques have been used in social networks sites and its analysis, The techniques ranges from unsupervised to semi-supervised and supervised learning methods, previous research work on data mining and algorithms have been identify by our survey, The diverse experimental results have also confirmed the relevance of data mining techniques in retrieving valuable information and contents from huge data generated on social network sites. Future survey will tend to investigate novel state-of-the-art data mining techniques for social network sites. The survey will compare similar data mining tools and explore other enhanced algorithms with variable large datasets.

1. ***REFERENCES***
2. Balogun Abiodun Kamoru., Azmi Bin Jaafar Omar., Marzanah A Jabar., Masrah Azrifah Azmi Murad., AbdulMajd.B.Umar; A Spam Detection Issues and Spam Identification of Fake profiles on Social Networks. Journal of Theoretical and Applied Information Technology. 15th November 2017. Vol.95.No.21. ISSN 1992-8645.
3. Kaplan, A.M and Haelein,M:. Users of the world unite! The challenges and opportunities of social media. Science direct, 59-68,2010.
4. Xin Jin.,Cindy Xide Lin., Jiawei Han.,Jiebo Luo.,; A Data Mining-based Spam Detection System for Social Media Networks.JCIS, 2017.
5. Shekar .C., Wakade. S., Liszka.K. J., and Chan. C.C.,: Mining Pharmaceutical spam from twitter. In ISDA, pages 813-817,2010.
6. Asur, S., and Huberman, B.:” Predicting the future with social network”. Web Intelligence and Intelligent Agent Technology(WIIAT),2010 IEEE/WIC/ACM International Conference on Vol.1.IEEE,2010.
7. GAD: General Activity Detection for Fast Clustering on Large Data; Xin Jin, Sangkyum Kim, Jiawei Han., LiangLiang Cao., Zhijun Yin., University of Ilinois at Urbana-Champaign.
8. Liu, B: Sentiment Analysis and Opinion Mining; AAAI-2011,San franciso, USA,2011,
9. Hu, M., Liu, B.,: Mining and Summarizing Customer Reviews.KDD ’04. In: Proceedings of the tenth ACM SIGKDD International Conference, 2004.
10. Kim, S., Hovy, E.: Determining the sentiment of Opinions. In: Proceedings of International Conference on Computational Linguistics (COLINGS\_2004), 2004.
11. Papadopoulous, S., Kompatsiaris, Y., Vakali, A., Spyridonos, P.: Community Detection in Social Media. Data Mining and Knowledge Discovery, 24(3), 515-554, 2012.
12. Pang, B., and Lee, L.,: Seeing stars: Exploiting class relationship for sentiment categorization with respect to rating scales,. In: Proceeding of the Association for Computational Linguistics (ACL), pp.115-124, 2005.
13. Zhou, D., Councill, I., ZHa, H.., Giles, C.: Discovering temporal communities from social network documents. Data Mining, 2007. ICDM 2007. In: Proceeding of the 4th ACM International Conference on Web search and data mining (pp.347-354). ACM,2011.
14. Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., Jurafsky, D.,: Automatic of Opinion Propositions and their Holders. In : Proceedings of the AAAI Spring Symposium on Exploring Altitude and Affect in text, 2004.
15. Paul Heymann., Georgia Koutrika., Hector Garcia-Molina,.: Fighting Spam on Social Web sites: A Survey Approaches and Future Challenges.IEEE’07 Internet Computing. 2007.
16. Kaur, G., : Social Network Evaluation Criteria and Influence on Consumption behaviour of the youth segment.2013.
17. Hatzivassiloglou, V., Mckeown,K.: Predicting the semantic Orientation of Adjectives. In: Proc.8th Conference on European Chapter of the Association for Computational Linguistics 174-181,1997
18. Turney, P.: "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," In: Proceeding of the Association for Computational Linguistics (ACL),pp. 417-424,2002.
19. Yoshida. K., Adachi. F., Washio, H., Homma.T., Nakashima. A., Fujikawa. H., and Yamazaki.K., : Density-based spam detector. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ,Pages 486-493. 2004.
20. Byun .B., Lee. H., Webb. S., and Pu.C.,: A discriminative classifier learning approach to image modelling and spam image identification. In CEAS,2007.
21. Pang, B., Lee. L., and Vaithyanathan. S.,: Thumb Up? Sentiment Classification Using Machine Learning Techniques. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadephia, July 2002, 79-86. Association for Computational Linguistics, 2002.
22. Kaschesky, M., Sobkowicz, P., Bouchard, G., .: Opinion Mining in Social Network: Modelling, Simulating, and Visualizing Political Opinion Formation in the Web. In: The Proceedings of the 12th Annual International Conference on Digital Government Research,2011.
23. Godbole, N., Srinivasaiah, M., Steven, S.: Large Scale Sentiment Analysis for News and Blogs. In: Proceeding of the Conference on WebLogs and SM (ICWSM),2007.
24. Pang, B. and Lee, L.,: Using very Simple Statistics for Review Search: An Exploration, .In: Proceedings of the International Conference on Computational Linguistics (COLING), 2008.
25. Tepper, A. : How Much Data is Created Every Minute? (INFOGRAPHIC). 2012. [http://mashable.com/2012/06/22/data-created -every-minute/.Retrieved on 16/10/2013 at 19.00](http://mashable.com/2012/06/22/data-created%20-every-minute/.Retrieved%20on%2016/10/2013%20at%2019.00).
26. Kim, P.: The Forrester Wave: Brand Monitoring, Q3.2006. Forrester Wave (White paper),2006.
27. Girvan, M., Newman, M.E., : Community Structure in Social and Biological Networks. Proceedings of the National Academy of Sciences. 99(12),7821-7826,2002.
28. Fortunato, S.: Community Detection in Graphs. Physics Reports, 486(3), 75-174,2010.
29. Murty, D., Gross, A., Takata, A., Bond, S.,: Evaluation and Development of Data Mining Tools for Social Network Analysis In Mining Social Networks and Security Informatics (pp.183-202), Springer , Netherlands,2013.
30. Zhai, Z., Liu, B., Xu, H., & Jia, P.: Clustering Product Features for Opinion Mining . In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (pp.347-354). ACM,2011.
31. Ruan, X.H., Hu, X., Zhang, X.,: Research on Application Models of Semantic Web-Based Social Network Analysis. In Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Volume 2 (pp.455-460) Springer Berlin Heidelberg, 2014.
32. Newman, M.,:Networks: An Introduction. Oxford University Press, 2010
33. Scott, J.: Social Network Analysis: Developments, Advances and Prospects. Social Network Analysis and Mining, 1(1),21-26,2011.
34. Santorini, B. : Part-of-Speecch Tagging Guidelines for The Penn Tree Bank Project(3rd Revision, 2nd Printing) Technical Report, Department of Computer and Information Science, University of Pennyslvania,1995.
35. Zhou, L., Ding, L., and Finnin, T. : How is the Semantic Web Evolving? A Dynamic Social Network Perspective. Computers in Human Behaviours, 27(4) 1294-1302, 2011.
36. Sindhawani, V and Melville, P.: Document-word co-regularization for semi-supervised sentiment analysis. 8th IEEE International Conference on Data Mining, 2008.
37. Rao, D and Ravichandran, D.: Semi-Supervised Polarity Lexicon Induction. In: Proceeding of the European Chapter of the Association for Computational Linguistics (EACL),2009.
38. Goldberg, A., and Zhu, X., : Seeing Stars when there aren't many stars: Grap-based semi-supervised learning for sentiment categorization. In HLT-NAACL.2006 Workshop on Text graphs: Graph-based Algorithm for Natural Language Processing, 2004.
39. Pang, B and Lee, L.: Seeing stars: Exploiting Class Relationships for sentiment categorization with respect to rating scales. In: Proceedings of the Association for Computational Linguistics (ACL) ,pp 115-124, 2005.