ISSN: 2321-7782 (Online)

Volume 4, Issue 3, March 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

A Survey on Tweet Analysis and Real Time Detection for H1N1 Virus Reporting System

Chandan Chaurasia[1]
UG Student
Department of Computer Science & Engineering
BMS college of Engineering
Bangalore, India

BaninSensha Shreshta[2]
UG Student
Department of Computer Science & Engineering
BMS college of Engineering
Bangalore, India

Himanchal Patel[3]
UG Student
Department of Computer Science & Engineering
BMS college of Engineering
Bangalore, India

Vishal Kesharwal[4]
UG Student
Department of Computer Science & Engineering
BMS college of Engineering
Bangalore, India

Vikrant B.M.[5]
Assistant Professor
Department of Computer Science & Engineering
BMS college of Engineering
Bangalore, India

Abstract: Twitter has got so much consideration these days. A significance of Twitter is its continuous nature. We research the constant connection of occasions, for example, swine influenza or H1N1 infection assault in Twitter and propose a calculation to screen tweets and to identify an objective occasion. To recognize an objective occasion, we add to a classifier of tweets taking into account elements, for example, the watchwords in a tweet, the quantity of words, and their setting. We view every Twitter client as a sensor and apply molecule sifting, which are generally utilized for area estimation. The molecule channel works superior to anything other than equivalent techniques for assessing the areas of target occasions. As an application, we build up a H1N1 infection reporting framework. On account of the across the board of influenza and the vast number of Twitter clients all through the nation, we can recognize an influenza with high likelihood just by observing tweet.

I. INTRODUCTION

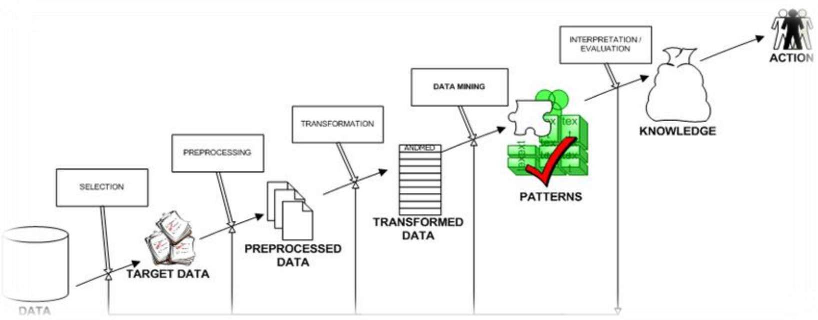

Fig 1: Basic steps involved in process.

For the most part, information mining (at times called information or learning disclosure) is the procedure of investigating information from alternate points of view and shortened it into valuable data - data that can be utilized to expand income, cuts costs, or both. Information mining programming is one of various investigative instruments for examining information. It permits clients to break down information from various measurements or edges, classify it, and condense the connections distinguished. Actually, information mining is the procedure of discovering connections or examples among many fields in expansive social databases.

While substantial scale data innovation has been developing separate exchange and systematic frameworks, information mining gives the connection between the two. A few sorts of investigative programming are accessible: factual, machine learning, and neural systems. For the most part, any of four sorts of connections are looked for:

Classes: For instance, an eatery network could mine client buy information to decide when clients visit and what they commonly arrange. This data could be utilized to expand movement by having every day specials.

- **Clusters**: Data things are gathered by connections or customer inclinations. For instance, information can be mined to recognize market sections or shopper affinities.
- **Associations**: Information can be mined to recognize affiliations. The lager diaper case is an illustration of familiar mining.

Sequential patterns: Information is mined to foresee conduct examples and patterns. For instance, an outside hardware retailer could anticipate the probability of a being bought in light of a purchaser's buy of resting packs and trekking shoes.

Data mining consists of five major elements:

a) Extract, change, and load exchange information onto the information stockroom framework.

b) Store and deal with the information in a multidimensional database framework.

c) Provide information access to business experts and data innovation experts.

d) Analyze the information by application programming.

e) Present the information in a helpful organization, for example, a chart or table.

II. Literature Survey

Why Twitter: Understanding Micro blogging Usage and Communities

A.Java[13] suggested that micro blogging is another type of correspondence in which clients can portray their present status in short posts dispersed by texts, cellular telephones, email or the Web. Twitter, a well knownMicro blogging instrument has seen a great deal of development since it was dispatched in October, 2006. In this paper, we introduce our perceptions of the Studying so as to micro blogging wonders the topological and topographical properties of Twitter's informal organization. We find that individuals use Micro blogging to discuss their day by day exercises and to look for on offer data. At last, we examine the client goals related at a group level and show how clients with comparable aims join with one another.

2) Social Networks that Matter: Twitter under the Microscope

Huberman[14]proposed Researchers, promoters and political activists see large online interpersonal organizations as a representation of social connections that can be utilized to examine the spread of thoughts, social security flow and viral showcasing, among others. Be that as it may, the connected structures of informal communities don't uncover real associations among individuals. Lack of consideration and the day by day rhythms of life and work make individuals default to associating with those few that matter and that respond their consideration. An investigation of social cooperation’s inside of Twitter uncovers that the driver of use is a fresh and wearing system of association’s fundamental the "proclaimed" arrangement of companions and supporters.

3) Tweet, tweet, Retweet: Conversational Aspects of Retweeting on Twitter.

Danah-Golderexplained Twitter as a Micro blogging administration that empowers clients to post messages ("tweets") of up to 140 characters - pillow an assortment of open practices; members use Twitter to talk with people, bunches, and the general population everywhere, so when discussions develop, they are frequently experienced by more extensive groups of onlookers than simply the conversationalists. This paper [15]analyzes the act of retweeting as a route by which members can be "in a discussion." While retweeting has turned into a tradition inside Twitter, members retweet utilizing distinctive styles and for various reasons. We highlight how creation, attribution, and open constancy are arranged in different ways. Utilizing a progression of contextual investigations and experimental information, this paper maps out retweeting as a conversational practice.

4) Micro blogging for Language Learning: Using Twitter to Train Communicative and Cultural Competence.

Borau[16]proposed how our work breaks down the value of small scale blogging in second learning utilizing the case of the informal community Twitter. Most learners of English don't require significantly more uninvolved info in type of writings, addresses or recordings, and so forth. This info is promptly accessible in various structures on the Internet. What learners of English need is the opportunity to effectively create variety of language and the opportunity to utilize English as instrument of correspondence. We examine the understudies' messages and show how the use of Twitter prepared open and social fitness.

Glivia[1]surveys the convenience of twitter hashtags in conclusion investigation. They separated 10,173,382 tweets identified with the Brazilian Presidential races in 2010. They examined these tweets and watched that the positive conduct of the tweeters crosswise over time was in agreement to the theory that hashtags conclusions coordinate the general people opinion. They moreover confirmed that the data production in twitter takes after a course model where individuals settle on their choices deliberately or not, in view of another person's estimations and decisions.

Alec Go[2] presented a system for ordering twitter messages. First, the inquiry term is standardized so that the question term without anyone else's input is not one-sided. Positive and negative tweets are gathered by utilizing ":)/:-) "And ":(/:- (". 80,000 positive and 80,000 negative tweets were gathered as preparing set. The tweets are then preprocessed. The graphical texts are then stripped from the preparation information since texts negatively affect the correctnesses of SVM and Maximum Entropy classifiers yet little impact on Naive Bayes classifier. They investigated the use of unigrams, bigrams, unigrams and bigrams, and Parts of Speech components. The accompanying were the correctness watched utilizing distinctive classifiers.

Apoorv Agarwal[3]have manufactured models for two arrangement: a paired undertaking of grouping feeling into positive what's more, negative classes and a 3-path of characterizing feeling into positive, negative and nonpartisan classes. Experimentation is finished with unigram model, component based model and tree piece based model. For the tree part based model they outlined a new tree representation for tweets. They utilized the unigram model for estimation investigation for Twitter information, as their gauge. Their analyses demonstrate that a unigram model is without a doubt a hard gauge accomplishing more than 20% over the chance standard for both grouping undertakings. Their element based model that uses just 100 elements accomplishes comparable exactness as the unigram model that utilizes more than 10,000 elements. Their tree portion based model beats both these models by a noteworthy edge. They too explored different avenues regarding a mix of models: joining unigrams with components and consolidating highlights with the tree part. Both these blends were found to beat the unigram pattern by more than 4% for both grouping assignments.

Alexander Pak[4]utilized Twitter API to gather twitter datasets in three classifications i.e positive, negative and unbiased. An investigation is performed by checking the appropriation of word frequencies in the body. The plot was found to comply with Zipf's law. The investigation is done in one more strategy. The parts of discourse of every word are labeled utilizing Tree Tagger and translated the distinction of label appropriations between sets of content (positive, negative, impartial or subjective, objective). The assumptions are characterized utilizing both Naive Bayes and SVM classifiers. Innocent Bayes was found to perform better. Two Bayes classifiers were prepared, one in view of n-gram and another in view of parts of discourse spreading. To expand the precision some regular n-grams are disposed of that don't emphatically demonstrate any supposition. The best execution was accomplished when a bigram was utilized.

Oshini Goonetilleke[5]addresses the issues around big information nature of twitter and the requirement for new information administration which confines the utilization of existing frameworks. For this reason, the framework essentially includes parts like Focused Crawling which is utilized as a part of powerful recovery and better scope, pre-handling of tweets should be possible utilizing Tokenization. Twitter zombie is a stage utilized for social occasion the information and taking apart. To give adaptability and productivity of handling huge measures of information, Tred Miner can be utilized for constant examination of tweets. Twitter permits the subsequent tweets to be gathered in clusters and after that stores them in social information base for further distinguishing proof of estimations. Essentially, one general stage is introduced for twitter utilizing comparable frameworks. Martha Arias et. al. [23] have concentrated on investigation of hashtag level of opinion grouping. The primary point of this is to consequently create the general estimation for a given hash tag in twitter. It is found that out of 0.6 million haphazardly chosen tweets 14.6% contained no less than one hashtag. A two stage SVM classifier is utilized to decide the extremity of assessment lastly, it is found that execution can be constantly expanded by Boosting loopy conviction spread with exactness expectation up to 77.72%. A basic standard methodology is created on the consequences of tweets utilizing basic voting technique where a paired quality was given to relating positive, negative and impartial tweets.

Po-Wei Liang[9]use Twitter API to gather twitter information. The preparation information falls in three distinct classifications (camera, cell telephone and film). The information is named as positive, negative and non-conclusions as opposed to using information that contains graphical texts to recognize the assessment. This is on account of textsthat are not generally reliable with the assumption. Pre-handling is done on the considered tweets. In the following step, the tweets containing suppositions are separated. This is finished by utilizing Naive Bayes classifier on the preparation set. Unigram Naive Bayes model is executed and the Naive Bayes rearranging autonomy presumption is utilized. In the following step, elements are dispensed with by utilizing the Mutual Information and Chi square element extraction strategy. Next, the introduction of a sentiment sentence is anticipated. i.e. positive or negative. It is found that tweets are delegated, obstinate and non-stubborn with a 76.8 %. Utilizing highlight determination, a precision of 96.6% is gotten. The test information set is characterized with a precision of 90.17% with the emphasis on just positive and negative information. Another preparing set is produced utilizing graphical data to signify positive or negative tweets. Utilizing the text prepared information set, the tweets are characterized with an exactness of 58.65%.

Spencer[11]has displayed an online apparatus Sentimentor to order live Twitter information into positive negative and target tweets. It has an interface that permits clients to effectively break down the word appropriations and pictorial representation of the assumptions in tweets. Twitter API is utilized for information extraction process. The gathered tweets were pre-prepared. The POS labeling of every word was done and unigrams and bigrams were removed. The Naive Bayes calculation was utilized for grouping with POS labels, unigrams and bigrams as components. It is found that the best precision of 52.31 % was acquired by utilizing bigram without POS labeling. The utilization of POS labels has negatively affected the exactness in view of the utilization of summation of POS labels over an expression as opposed to considering parallel events.

Pablo[10]displayed a group of Naive Bayes classifiers for identifying the extremity of English tweets. Two diverse Naïve Bayes classifiers have been manufactured in particular Baseline (prepared to group the tweets as positive, negative and nonpartisan), and Binary. The components considered by the classifiers are Lemmas (things, verbs, modifiers and intensifiers), Multiword, and Polarity Lexicons from various sources and Valence Shifters. The preparation information set of tweets is gotten from SemEval Organization-2014 and extra commented on tweets from outside sources. Numerous mixes of the systems and elements are actualized. It is additionally reasoned that execution is best when twofold procedure is utilized with multiword and valence shifters highlights.

Marc Egger[6]have explored about content based User-Generated-Content which is helpful for corporate organizations. They have considered the procedures along three stages gathering, investigation (positive, negative or nonpartisan) and representation, alongside top down methodology is utilized starting with systems working on archive and boring down to lower levels of data extraction. Here theme displaying systems, for example, Probabilistic Latent Semantic Analysis or Latent designation are considered, which can be utilized to reveal unique subjects inside of archive. Development of Decision trees is observed to be more Complex for truncation for this Naïve Bayes that can be utilized to recognize the end of sentence. To guide every expression of content onto parts of discourse Distance based methodology is utilized.

Kumar[8]recover twitter information utilizing Twitter API. They preprocess the tweets and add weightage as per the quantity of shout imprints and the modifiers, verbs and qualifiers are labeled in every tweet. Descriptors and negative words are considered to ascertain the extremity of the entire expression. Extremity of the tweet is ascertained taking into account an equation. The framework proposed had attributes of seeing the notions in tweets. **Kouloumpis**performed assessment investigation utilizing Twitter hashtags (e.g., #bestfeeling, #newphone, #androidwhat) to recognize positive, negative, and nonpartisan tweets. They utilized three distinctive of Twitter messages in their trials a hash labeled information set, an information set and a physically explained information set ISIEVE. Their objective for these examinations was two-fold. In the first place, they needed to assess whether the preparation information with marks got from hashtags is helpful for preparing supposition classifiers for Twitter. Second, they need to assess the adequacy of the tweets after pre-preparing for slant examination in Twitter information. It was inferred that the investigations on twitter feeling examination demonstrate that grammatical form elements may not be helpful for conclusion investigation in the microblogging space.

Davidov[7]proposed a regulated supposition arrangement system which utilized information from Twitter. By using 50 Twitter labels and 15 smileys (:- o) as feeling names, this structure stays away from the requirement for work manual explanation, permitting recognizable proof and arrangement of differing notion sorts of short messages. The paper additionally assessed already created systems for estimation characterization and demonstrated that their structure effectively distinguished feeling sorts of sentences that were untagged. The nature of the estimation recognizable proof was additionally affirmed by genuine manual arrangement and cross approval. Conditions and cover between various estimation sorts spoke to by smileys and Twitter hashtags was likewise investigated. In this study, four distinctive component sorts (accentuation, words, n-grams and examples) for notion order were utilized and the commitment of every element sort for this undertaking was asses.

Narr[12]inspect a free supposition classifier. The tweets English, French, German and Portuguese are considered. Graphical texts as marks are utilized to gather preparing information since they trusted that in short messages such as tweets, the texts are frequently steady with the general assumption of the tweet. Gullible Bayes classifier is utilized from NTLK Natural Language Processing Toolkit. 10000 tweets are utilized for testing which was physically explained utilizing the help Mechanical Turk laborers into positive, negative, unbiased or immaterial. They utilized different blends of the classifier. With Unigram classifier a precision of 81.3% was accomplished in English tweets. Best exactnesses watched for the four language are: English - 81.3%, French - 74.9%, German - 79.8% and Portuguese - 64.9%.

III. Proposed Technique

Since the presentation of Web 2.0, the social Web is turning into a critical asset for substance extraction. There are a huge number of clients who are utilizing the social Web as a medium for distributed data, talking about political issues, sharing recordings, posting remarks, overseeing web journals, furthermore expressing their own supposition in continuous examinations. The advancement and wide appropriation of informal organizations among household clients has opened a totally better approach to examine subjects for humanist and every other analyst concentrating on social relations and/or client association saw from a totally new viewpoint and at an aggregate diverse scale. A few methodologies attempt to bunch gatherings of clients utilizing k-mean or hieratical grouping calculations utilizing as a criteria companionship relations. Different methodologies attempt to haphazardly separate client profiles to coordinate them in a worldwide social chart. Some take a more extensive methodology by looking at among themselves the most applicable informal communities these days.

To understand how current archiving tools are not fully up to the task of social Web archiving, consider the simplified architecture of a traditional Web crawler [1].

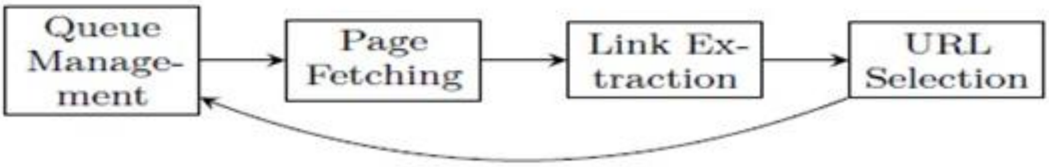

Figure 2: Traditional processing chain of a Web crawler

A **Web crawler** (also known as Web spider or robot) is a PC program that assesses the Web in a systematic way and recovers focused on archives. Customary Web crawlers creep the Web in a reasonably extremely straightforward way. They begin from a seed rundown of the URLs to be put away in a line (e.g., the beginning URL might be the landing page of a Web website). Site pages are then brought from this line in a steady progression and connections are removed from these Web pages in the event that they are in the degree.

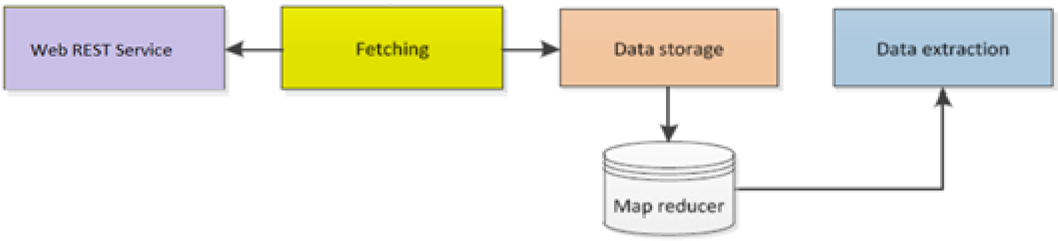

Figure 3: Crawler architecture

- **Crawling and Storing Techniques:** The crawler is composed of three main architectural blocks: a fetching module, a storing module and a data extraction module:

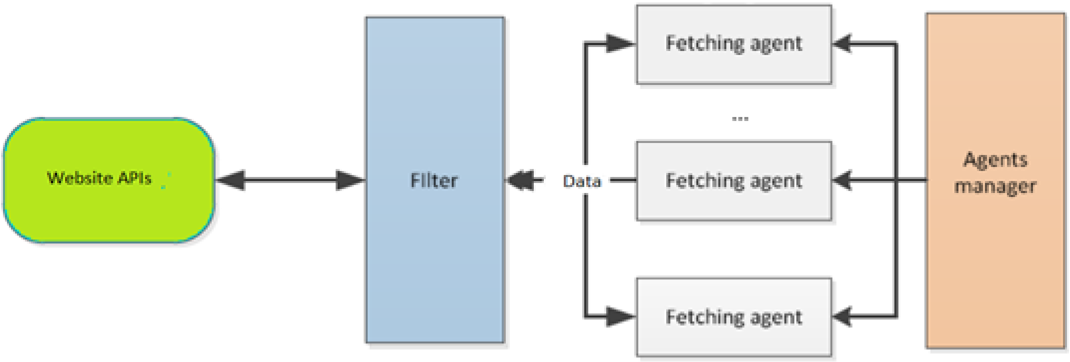

Figure 4: Multiple fetching agents

The first job has to be done is to fetch the data (in our case tweets). It has been found that fetching 100 tweets takes around 22-23 seconds, which is a huge amount of time. The fetching rate is improved by using multiple fetching agents which are handled by the agent manager. This improves the rate by 8 times. The fetching subcomponent allows us to filter the desired result by locations, locale, language or period of tweets.

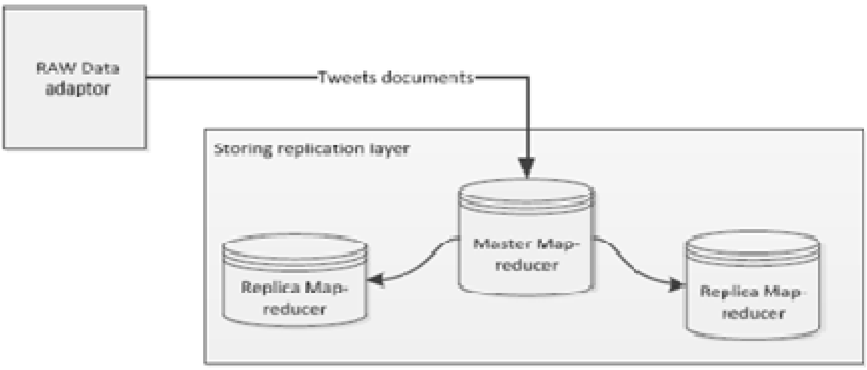

Figure 5.: Store module

This module interfaces with the bringing module and stores information in a guide administration. This module is made out of a RAW information connector, which changes information got from the operators to standard JSON objects that are later put away in the guide reducer group. We utilized the guide reducer rather than a customarily SQL database since: 1. map-reducer permits the putting away of non specific reports. 2. It has a replication worked in-instrument. 3. It permits keeping the full history of records.

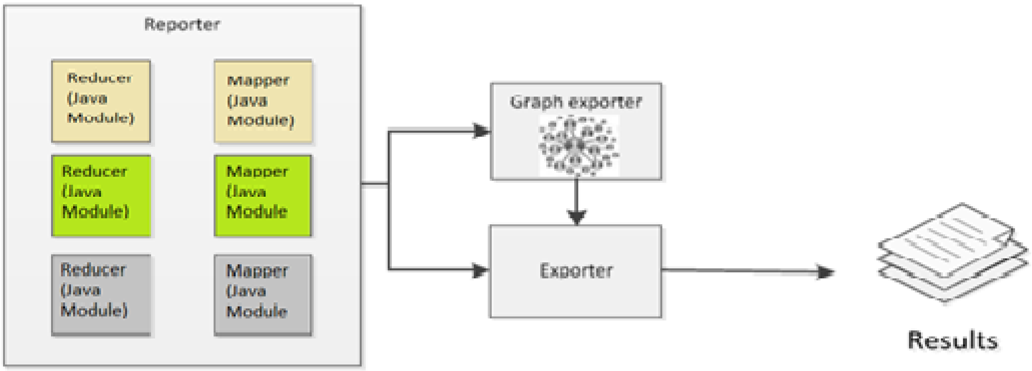

Figure 6.: Extraction module

This model is made out of three primary segments: chart manufacturer, content analyzer and center. The correspondent is the center (made out of java modules) which cooperates with the guide reducer to perform particular assignment. The mapper characterizes what data to be removed from guide reducer. The reducer characterizes the way information is totaled in the last result. The chart developer is the subcomponent that manufactures the tweets social diagram, we utilize this segment to perform investigation about the availability of the system, disengaged sub charts, topology of client gatherings and so forth. The exporter is a minor segment that is sending out information as reports.

IV. Conclusion

As portrayed in this paper, we examined the continuous way of Twitter, giving specific regard for occasion discovery. Semantic examinations were connected to tweets to characterize them into a positive and a negative class. We view every Twitter client as a sensor, and set the issue as location of an occasion taking into account tangible perceptions. Area estimation systems, for example, molecule sifting are utilized to appraise the areas of occasions. As an application, we added to a swine influenza reporting framework, which is a novel way to deal with tell individuals instantly of a H1N1 infection occasion. Microblogging has continuous qualities that recognize it from other online networking, for example, web journals and community oriented bookmarks. As portrayed in this paper, we introduced an illustration that influences the continuous way of

Twitter to make it valuable in taking care of a vital social issue: characteristic catastrophes. It is trusted this paper will give some understanding into the future reconciliation of semantic investigation with miniaturized scale blogging information.

ACKNOWLEDGEMENT

The work, reported in this paper, is supported by the college through the Technical Education Quality Improvement Programme [TEQIP-II] of the MHRD, Government of India.

References

1. Glivia A. R. Barbosa, Wagner MeiraJr, Ismael S. Silva, Raquel O. Prates, Mohammed J. Zaki, Adriano Veloso, "Characterizing the Effectiveness of Twitter Hashtags to Detect and Track Online Population Sentiment", ACM SIGCHI Conference on Human Factors in Computing Systems, Juried Worksin-Progress, May 2012.
2. Alec Go, RichaBhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision.
3. ApoorvAgarwal, BoyiXie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data", Workshop on Languages in Social Media, 2011, pp 30-38, ISBN: 978-1-932432-96-1.
4. Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", 7th conference on International Language Resources and Evaluation (LREC'10), Valletta, ISBN: 2-9517408-6-7, May 2010.
5. OshiniGoonetilleke, TimosSellis, Xiuzhen Zhang, SaketSathe, "Twitter analytics: A Big data management perspective", ACM SIGKDD Explorations Newsletter - Special issue on big data archive, Volume 16, Issue 1, June 2014, Pages 11-20, DOI: 10.1145/2674026. 2674029.
6. Marc Egger, Andre Lang, "Brief tutorial on how to generate user generated content", German Journal on Artificial Intelligence, pp 53-60, Feb 2013.
7. Davidov, Tsur, Rappoport, "Semi- Supervised Recognition of Sarcastic Sentences in Twitter and Amazon", 14th Conference on Computational Natural Language Learning, 2010, pp 107-116, ISBN: 978-1-932432-83-1.
8. Akshi Kumar, Teeja Mary Sebastian, "Sentiment Analysis on Twitter", International Journal of Computer Science Issues, Volume 9, Issue 4, No 3, July 2012, pp 372-378, ISSN (Online): 1694-0814.
9. Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social Media Data", IEEE 14th International Conference on Mobile Data Management, Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-4673-6068-5, http://doi.ieeecomputersociety.org/10.1109 /MDM. 2013.73
10. Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets", 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, Aug 23-24 2014, pp 171–175.
11. James Spencer, Gulden Uchyigit. Sentimentor: Sentiment Analysis of Twitter Data, The 1st International Workshop on Sentiment Discovery from Affective Data (SDAD 2012) pp 56-66, Bristol, UK.
12. SaschaNarr, Michael Hulfenhaus, SahinAlbayrak, "Language-Independent Twitter Sentiment Analysis", KDML workshop on knowledge discovery, data mining and machine learning, Dortmund, Germany, Sep 2012.
13. A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter: Understanding Microblogging Usage and Communities," Proc. Ninth WebKDD and First SNA-KDD Workshop Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07), pp. 56-65, 2007.
14. B. Huberman, D. Romero, and F. Wu, "Social Networks that Matter: Twitter Under the Microscope," ArXiv E-Prints, http://arxiv.org/abs/0812.1045, Dec. 2008.
15. G.L. Danah Boyd and S. Golder, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," Proc. 43rd Hawaii Int'l Conf. System Sciences (HICSS-43), 2010.
16. K. Borau, C. Ullrich, J. Feng, and R. Shen, "Microblogging for Language Learning: Using Twitter to Train Communicative and Cultural Competence," Proc. Eighth Int'l Conf. Advances in Web Based Learning (ICWL '09), pp. 78-87, 2009.