Social Networking Text Classification in Big Data Environment

Amit Mittal
Department of Computer Engineering
Institute Of Engineering & Technology, Devi Ahilya University, Indore, India.

Digendra singh Rathore
Department of Computer Engineering
Institute Of Engineering & Technology, Devi Ahilya University, Indore, India.

Abstract- The data mining is a technique of which offers the computer algorithms to compute patterns and find the category of data. Data mining is performed in both the techniques supervised and unsupervised learning. The selection of algorithm is depends upon the type and behavior of data. The data in nature found in two popular categories structured and unstructured data. The structured data is available in form of relational manner and the unstructured data is not organized in a given relativity. But if the data is well defined in their previous patterns then that can be utilized with the supervised learning algorithms. In this presented work the identified data twitter data set is used to perform analysis. For the classification of twitter data two class classification problems is considered. Therefore the entire input data samples are required to classify in two classes namely positive orientation and negative orientation. Therefore a binary classifier namely decision tree is utilized for making analysis and performing the classification task.

Keywords- Text mining, analysis technique, semantics, Bayesian classifier, dataset.

I. INTRODUCTION

The use of digital text is increases as the social media increases their effect in daily life. A number of research groups and individual researchers are working to finding the patterns on these data. In this presented study the social media text analysis and sentiment analysis techniques are investigated and a new classification technique is proposed for enhancing the performance of text classification. The given chapter provides an overview of the proposed work and involved investigation.

Data mining is a technique of mining information from the raw data. Here the information is a term that is relevant to the data which is required by a data miner or application. In this presented work the text data is mined for extracting the semantic based similar text from a raw set of text data. Basically text data is found in an unstructured manner and labeling of data is complicated task therefore most of the application are utilizing the cluster analysis techniques for categorizing data. But if the data is well labeled then that can be used with the classification algorithms also. Therefore the microblog data can be used with the classification.

In this presented work the text data for microblog analysis is used for preparing the classification algorithm. Basically the microblogs are frequently used now in these days with small amount of communication data. But frequent use of this communication channel increases the amount of data for manual analysis. The presented system is an automated text analysis technique that works on the labeled data and provides the outcomes in two step process. In first the data is processed in order to obtain the text features and then the learning on evaluated features are performed.

II. OBJECTIVES

The presented study is focused on finding the most optimum techniques for text analysis and classification according to their semantic patterns. Therefore some essential key works are involved with the current objective.

2.1. Investigation of text analysis technique: in this phase a strong literature is collected for finding the most appropriate classification technique which promises to provide the text pattern analysis for finding the similar semantic attributes from both the class distributions.

2.2. Design and implementation of semantics based classifier: after concluding the different articles and research on the semantics based text classification technique a new approach based on improved decision tree is prepared

2.3. Performance analysis of proposed classification technique: in this phase the developed classifier is tested on a sequence of real world data which is taken from the twitter. Additionally the performance of the system is reported on accuracy, error rate, time, and space complexity.

This section draws the key objectives involved in the current domain of study and in next section the motivation of the proposed study is provided...

III. REVIEW OF RELATED WORK

This section provides the detailed study about the text processing and their utilization on the different domains of study. Additionally the different research work is also reported by which the new generation text analysis work is efficiently performed.

Text Mining History

Manual text mining approaches, required much effort in order to find specific kind of data first introduced in mid-1980s, but technological progress have allow the domain to improve from the previous issues. Text mining is a divers domain that having a wide range of applications on information retrieval, machine learning, data mining, statistics, and computational semantics. Most of the information is recently stored as text data. Now in these days most of the research is progressing in the direction of multiple language support: this kind of system able to gain information across languages and also capable to group similar data from different kind of language sources according to their original semantics [1].

The challenge is to take advantage of the large amount of enterprise information that available in "unstructured" manner. In Oct 1958 an article is given by H.P. Luhn for A Business Intelligence System where text mining for unstructured data is addresses as the major issue, which describes a system that will:

"Utilize data-processing machines for auto-abstracting and auto-encoding of documents and for creating interest profiles for each of the 'action points' in an organization. Both incoming and internally generated documents are automatically abstracted, characterized by a word pattern, and sent automatically to appropriate action points". [2]

In the 1960s initially management information systems developed, and as Business Intelligence appeared in the '80s and '90s as a software category and arena of preparation. The prominence was on numerical data stored in relational databases. Additionally, text in "unstructured" documents is complex to practice. The arrival of text analytics in its existing form stops from a redeploying of research in the 1990s from algorithm improvement to application, according to Prof. Marti A. Hearst in an article of Untangling Text Data Mining:

The computational linguistics community has found that large text information storages as a resource for producing better text analysis techniques. A new prominence is the use of large online text storages to find new facts and developments about the world itself. For making development do not need completely automated text analysis; rather, a hybridization of computational and user-interactive analysis may open the door to get innovative results.

Text Mining

Text mining deals with the computational analysis of text for knowledge discovery and data pattern analysis. These techniques provide ease in information extraction, natural language processing, and information retrieval. Additionally, include these domains with algorithms and KDD methodologies. [3] Therefore, nota similar procedure can be followed with this domain of KDD process, where not data in general format and similar to each other. Therefore, text documents are required to seriously analyze. From this a new question for data mining techniques is that the data modeling perception for unstructured data sets [4].

If making effort for finding the definition of text mining, then that can be nearer to related research domain's or application's specific. At this point each of them can provide a different meaning of text mining, which is inspired by the specific viewpoint of the application area [4]:

- Text Mining = Information Extraction.

In this context the text mining is essentially parallels to information extraction, mean to say the information extraction from texts.

- Text Mining = Text Data Mining.

Text mining can be also defined by approximating to data mining as the application of algorithms and technique from the domain of machine learning and statistics analysis over text documents with the aim of discovering fruitful patterns. For that motive that is required to pre-process the text documents more appropriately. Different researchers use natural language processing, information extraction techniques or other simple pre-processing algorithms to find meaningful information from texts.

- Text Mining = KDD Process.

In the knowledge discovery process model, that is commonly found that the text mining is a process with a series of fractional steps, among the use of data mining or statistical analysis. In general way, the extraction of not yet discovered information in collections of texts documents also text mining as procedure orientated methodology on texts.

IV. APPLICATION OF TEXT MINING

There are a large area of applications for text mining implementation, not all the applications are possible to list their some of them are listed in this section.

4.1 Enhancing Web Search

One way to improve the users' performance and knowledge is Web search on other words meta-search engines. Usually, meta-search engines were comprehended to discourse different problems related to general-purpose search engines. Including Web data analysis, user query specific search results, and their appearance a common technique for alternate presentation of results is by sorting them into (a hierarchy of) domain clusters. The clustered data may be represented to the user in different kinds of visualization techniques, for example as a separate expandable tree or arcs that connect Web pages in graphically concentrated "maps". On the other hand, topics created by clustering may not verify acceptable for every query [5].

4.2 Mining Bibliographic Data

Notable properties of the text data collection are its incompleteness and diversity of techniques to provide references. Information being provided in free text format, The work included in this context:

(1) Extraction of basic information related to researchers, organizations, and complete lists of bibliographic references from the overall collected documents,

(2) Finding the duplicate references in the given forms equivalent to different researchers, thus creating co-authorship relationships, and

(3) Extract information from IST World Web portal and consume it with visualization techniques for analysis of the extracted information [5].

4.3Sentiment Classification

The domain of sentiment analysis handles the analysis of sentiments found in text documents. A basic task is sentiment classification, where a definite amount of text is sorted into classes which relate to, e.g. positivity or negativity of expressed ideas. This is an application of dimensionality reduction techniques for two sentiment classification issues:

(1) document polarity classification, where documents representing complete reviews are classified into positive or negative, and

(2) Sentence polarity classification, which deals with polarity classification of individual sentences [6].

Techniques of Text Mining

There are different kinds of techniques available by which the text pattern analysis and mining is performed. Some of the essential techniques are discussed in this section.

Information Extraction

A initial point for computers to analyze unstructured text is to use information extraction. Information extraction software recognizes key phrases and associations within text. This is performed for finding the predefined sequences in text data, that process is also called pattern matching. The software concludes the relationships among all the identified objects to provide the user relevance information from text data sources. This methodology can be very useful when a large volume of text data is required to analyze. Classical data mining accepts that the information to be "mined" is already in the form of a relational database [6].

Topic Tracking

A topic tracking works by keeping user interest as profiles and, according to their interest the documents the user views, additionally other documents is also predicted for user. Yahoo allows users to select keywords when news relating to those topics becomes available. Topic tracking technology does have limitations, however. For example, if a user creates an alert for "text mining" user will catch news stories, and topics that are actually required. Some of the better text mining tools let user choose particular domain of interest or the software automatically can the user's interests based on user's browsing history and click information [7].

Categorization

Categorization techniques involve the identification of the main subjects of a document by providing the document into a pre-defined set of subjects. When, categorization take place to a document, than it will treat the document as a "bag of words". It does not try to process the original information as information extraction does. On the other hand, categorization only calculates words that are found and, using this calculations, the main topics identifies. Categorization usually depends on a dictionary where the topics are defined, and associations are recognized by observing for frequent terms, synonyms, and interrelated terms. Categorization techniques commonly have a method for ranking the documents by which documents have the most content on a particular subject [5].

V. RELATED STUDY

This section provides the recent efforts and algorithms that are contributing in the small text data processing and accurate sentiment analysis.

Micro blogging, like Twitter1, has become a popular platform of human expressions, through which users can easily produce content on breaking news, public events, or products. The massive amount of Micro blogging data is a useful and timely source that carries mass sentiment and opinions on various topics. Existing sentiment analysis approaches often assume that texts are independent and identically distributed(i.i.d.), usually focusing on building a sophisticated feature space to handle noisy and short messages, without taking advantage of the fact that the microblogs are networked data. Inspired by the social sciences findings that sentiment consistency and emotional contagion are observed in social networks, Xia Hu et al [9] investigate whether social relations can help sentiment analysis by proposing a Sociological Approach to handling Noisy and short Texts (SANT) for sentiment classification. In particular, they present a mathematical optimization formulation that incorporates the sentiment consistency and emotional contagion theories into the supervised learning process; and utilize sparse learning to tackle noisy texts in Microblogging. An empirical study on two real-world Twitter dataset show the superior performance of given framework in handling noisy and short tweets.

Feather selection is a process that extracts a number of feature subsets which are the most representative of the original meaning from original feature set. It greatly reduces the text processing time and increases the accuracy because of removing some data outliers. With the rapid development of Web 2.0 and the further evolution of the Internet, short text like micro-blog plays an important role in people's daily life. However, existing feature selection methods cannot effectively extract these short text features, and greatly reduce the classification and clustering performance of short text. In this regard, Zitao Liu et al [10] propose a novel feature selection method based on part-of-speech and How Net. According to the composition of the text property, we choose the words with larger amount of information by different part-of-speech, and then expand the semantic features of these words based on How Net, in this way the short text has more useful features. They use test data set collected from sina micro-blog and adopt the micro average and macro average of F1-Measure to evaluate the effects of short text classification. The results show that the short text feature selected by method has a good amount of information, as well as good classification results.

ApoorvAgarwal et al [11] examine sentiment analysis on Twitter data. The contributions of this paper are: (1) first introduce POS-specific prior polarity features. (2) and explore the use of a tree kernel to obviate the need for tedious feature engineering. The new features (in conjunction with previously proposed features) and the tree kernel perform approximately at the same level, both outperforming the state-of-the-art baseline.

Microblogs are a tremendous repository of user-generated content about world events. However, for people trying to understand events by querying services like Twitter, achronological log of posts makes it very difficult to get a detailed understanding of an event. In this paper, Adam Marcus et al [12] present TwitInfo, a system for visualizing and summarizing events on Twitter. TwitInfo allows users to browse a large collection of tweets using a timeline-based display that highlight speaks of high tweet activity. A novel streaming algorithm automatically discovers these peaks and labels them meaningfully using text from the tweets. Users can drill downto subevents, and explore further via geolocation, sentiment, and popular URLs. Author contributes a recall-normalized aggregate sentiment visualization to produce more honest sentiment overviews. An evaluation of the system revealed that users were able to reconstruct meaningful summaries of events in a small amount of time. An interview with a Pulitzer Prize-winning journalist suggested that the system would be especially useful for understanding a long-running event and for identifying eyewitnesses. Quantitatively, our system can identify 80-100% of manually labeled peaks, facilitating a relatively complete view of each event studied. Social media streams, such as Twitter, Facebook, and SinaWeibo, have become essential real-time information resources with a wide range of users and applications. The rapidly increasing amount of live information in social media streams has important societal and marketing values for large corporations and government organizations. There is a strong need for effective techniques for data gathering and content analysis. This problem is particularly challenging due to the short and conversational nature of posts, the huge data volume, and the increasing heterogeneous multimedia content in social media streams.

Moreover, as the focus of “conversation” often shifts quickly in social media space, the traditional keywords based approach to gather data with respect to a target brand is grossly inadequate. To address these problems, YueGao et al [13] propose a multi-faceted brand tracking method that gathers relevant data based on not just evolving keywords, but also social factors (users, relations and locations)as well as visual contents as increasing number of socialmedia posts are in multimedia form. For evaluation, they set up a large scale microblog dataset (Brand-Social-Net) on brand/product information, containing 3 million Microblogs with over 1.2 million images for 100 famous brands. Experiments on this dataset have demonstrated that the proposed framework is able to gather a more complete set of relevant brand-related data from live social media streams. Authors have released this dataset to promote social media research..

VI. PROPOSED WORK

The proposed work is intended to find the classification algorithm for text classification that provide the semantically analysis of the input text data. Thus the presented chapter provides the understanding of the entire methodology to process the text data for semantically text analysis.

6.1 Domain overview

In this age of technology most of the computational algorithms and applications are hosted on remote servers. Users consume the remote data using a global information network known as internet. This network provides service and information 24X7 therefore that becomes a part of new generation life. Use of internet also connects us with the imaginary social world such as twitter, facebook and others. These internet based applications are sometimes called the social networks. In these social networking web applications a huge amount of text, image and video data is generated. The manual analysis of such huge amount of data is a challenging task, therefore computational or statistical techniques are applied on these data to find targeted patterns for this data.

In order to identify the patterns of data from the social networking web application according to their hidden pattern more specifically semantically patterns analysis technique is required. On the other hand for accurate classification of these data the traditional algorithm is not able to provide effective solution. Therefore a plug-in for traditional data mining technique is developed for providing ease in classifying text data.

Therefore the proposed work involves the classification of text data patterns according to their semantics in two different classes’ namely negative orientation and positive orientation. This classification helps to find the mood of a user during communication over the social networking web applications.

6.2 Algorithm study

This section provides the detailed study about algorithm which are used to develop the proposed sentiment text classification technique.

6.2.1 Bayesian classifier

The Naive Bayes classification algorithmic rule is a probabilistic classifier. It is based on probability models that incorporate robust independence assumptions. The independence assumptions usually don't have an effect on reality. So they're thought of as naive. You can derive probability models by using Bayes' theorem (proposed by Thomas Bayes). Based on the nature of the probability model, you'll train the Naive Bayes algorithm program in a very supervised learning setting. In straightforward terms, a naive Bayes classifier assumes that the value of a specific feature is unrelated to the presence or absence of the other feature, given the category variable. There are two types of probability as follows [20]:

- Posterior Probability [P (H/X)]
- Prior Probability [P (H)]

Where, X is data tuple and H is some hypothesis. According to Baye's Theorem

$$P\left(\frac{H}{X}\right)=\frac{P\left(\frac{X}{H}\right)P(H)}{P(X)}$$

6.3 Methodology

The proposed system for the sentiment text analysis and their accurate evaluation a new system is prepared using the traditionally available techniques. The organization of methodologies for obtaining sentiment based text analysis is given using figure 3.1.

Twitter dataset: The machine learning techniques required to phase of processes involved in the classification techniques first the learning through the previous examples and then utilizing the previous examples for classifying the data of the similar pattern. The given twitter data is used for the learning in the proposed technique of classification

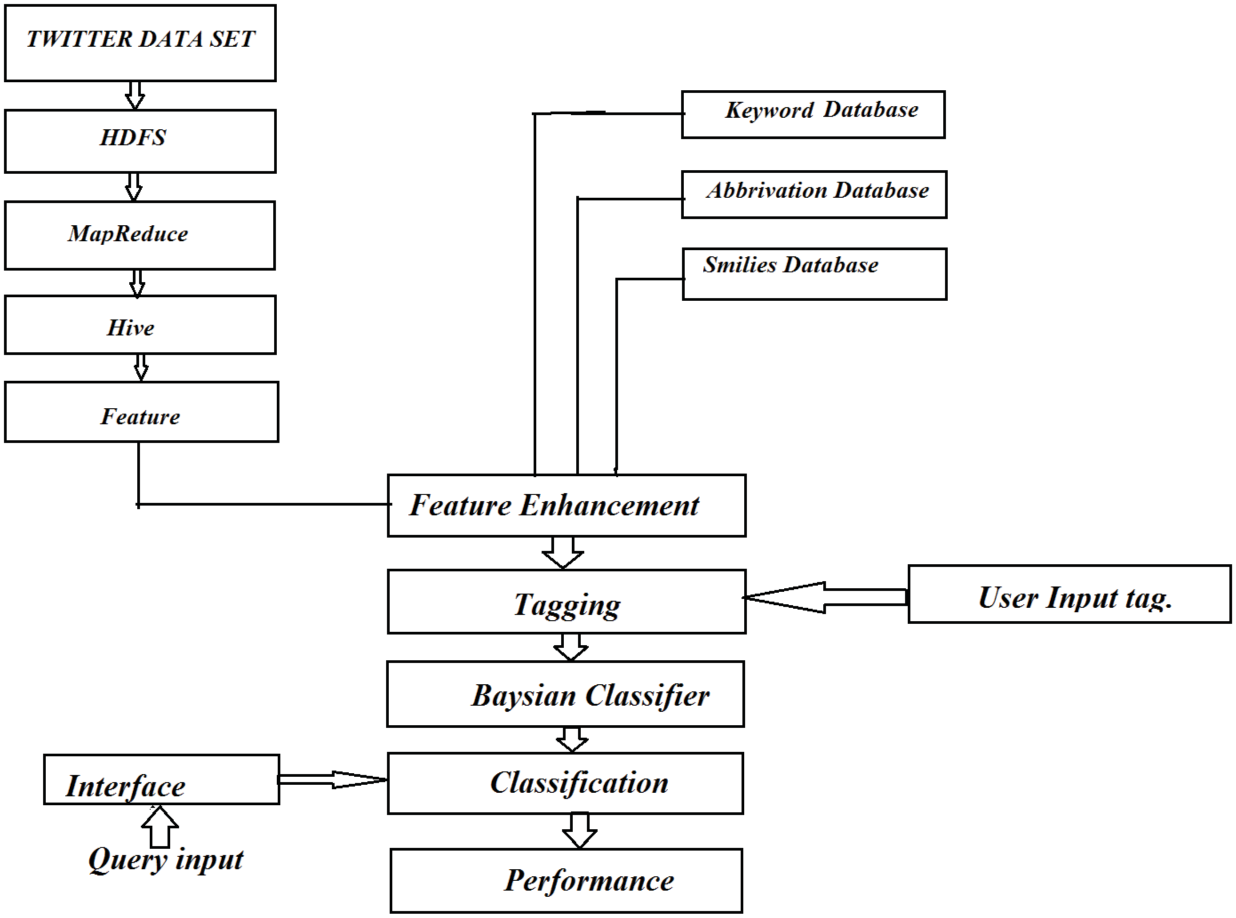

Figure : proposed methodology

HDFS: That is the big data repository which is used to store the data for learning and classification. Thus that is just storage architecture of data for processing in big data environment.

Mapreduce: That is a data processing tool used with the Hadoop infrastructure to process the data and reduce the amount of data from the actual amount of data using the mapping techniques. The training data placed in Hadoop directory is processed using the mapreduce and produced in next phase for storing them in to column arrangement.

HIVE: hive provides the data storage structure in the column manner thus the different number of attributes which are participating in classification is organized for the text data is used for further processing in the table manner.

Features: in this phase the column based data is processed to find the word occurrence frequency for the given or specified classes in the sentiments. Therefore the term frequency is provided by the following formula:

$$\text{word frequency} = \frac{\text{number of time accured a word}}{\text{total amount of word}}$$

*Feature enhancement***:** That is a feature enhancement technique or the data quality improvement technique by which the incomplete words and the different text symbols are recovered from three different data bases. These three different data are as follows:

- *Keyword database:* The keyword database contains the significant amount frequent words that are frequently occurred in any kind of text sentence organization such as, is, am, are, this, that, to and others. Using the feature enhancement technique the words are compared to the database and removed from the available features.
- *Abbreviation database:* In most of social network sites the number of abbreviations is used during the text communication such as for take care the people usage the terms TC. Thus a database is prepared with the Abbreviation and their full forms to reform again the entire aspect of the text data.
- *Similes database:* During text communication for expressing the moods and the emotions sometimes user are also usages the graphics or special characters. These special characters and graphics in the social network are known as the similes. Thus a database with the similes and their meaning is also prepared to enhance the features by completing the sentences.

Tagging: In this place the user interaction with the feature list is required to supply the initial tags these tags are in terms of noun, pro-noun and other semantics of the text data.

Bays classifier: In this phase the tagged data from the previous phase is utilized to develop the statistical classifier. Which first prepare the training from the training data set and then the test set is used to provide the classification of the data according to the hidden sentiments.

Interface: The provision is made in order to simulate the training and test of the proposed sentiments based text classification model. That is a basic user interface design which is used to submit the training and testing dataset for classification and performance measurement.

Query input: That is a part of testing phase which is used to produce the unknown text for finding the orientation of the text communication without any class labels and that the responsibility of the data model to analyze the text data and provide the class labels for the unknown text according to the tagged data.

Performance: After the classification task the performance of the classification system is evaluated in terms of their accuracy, error rate and the other efficiency parameters.

VII. CONCLUSION

By the above study finally we can conclude over work in following point as:

- Improve the performance of unstructured text analysis.
- Improve the sentiment based text classification.
- Identifying the text patterns and user sentiments.

REFERENCES

[1] B. V. Rama Krishna, B. Sushma, "Novel Approach to Museums Development & Emergence of Text Mining", ISSN 2249-6343, International Journal of Computer Technology and Electronics Engineering (IJCTEE), Volume 2, Issue 2

[2] H. P. Luhn, "A Business Intelligence System", Volume 2, Number 4, Page 314 (1958), No topical Issue, IBM Research Journals

[3] Andreas Hotho, Andreas Nurnberger, Gerhard Paaß, FraunhoferAiS, "A Brief Survey of Text Mining", Knowledge Discovery Group Sankt Augustin, May 13, 2005s

[4] Hien Nguyen, Eugene Santos, and Jacob Russell, "Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization", IEEE Transactions on Systems, Man, and Cybernetics—PartA: Systems and Humans, Vol. 41, No. 6, November 2011

[5] Umajancy. S, Dr. Antony SelvadossThanamani, "An Analysis on Text Mining –Text Retrieval and Text Extraction", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013

[6] Miloš Radovanović, MirjanaIvanović, "Text Mining: Approaches And Applications", Abstract Methods and Applications in Computer Science (no. 144017A), Novi Sad, Serbia, Vol. 38, No. 3, 2008, 227-234

[7] Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009

[8] P. Bhargavi, B. Jyothi, S. Jyothi, K. Sekar, "Knowledge Extraction Using Rule Based Decision Tree Approach", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.7, July 2008

[9] Xia Hu, Lei Tang, Jiliang Tang, Huan Liu, "Exploiting Social Relations for Sentiment Analysisin Microblogging", WSDM '13, February 4–8, 2013, Rome, Italy, Copyright 2013 ACM 978-1-4503-1869-3/13/02

[10] Zitao Liu, Wenchao Yu, Wei Chen, Shuran Wang, Fengyi Wu, "Short Text Feature Selection for Micro-blog Mining", Conference Paper, January 2011, DOI: 10.1109/CISE.2010.5677015 · Source: IEEE Xplore

[11] Apoorv Agrawal, BoyiXie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data", Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 30–38,Portland, Oregon, 23 June 2011

[12] Adam Marcus, Michael S. Bernstein, Osama Badar,David R. Karger, Samuel Madden, Robert C. Miller, "TwitInfo: Aggregating and Visualizing Microblogsfor Event Exploration", CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.Copyright 2011 ACM 978-1-4503-0267-8/11/05

[13] YueGao, Fanglin Wang, Huanbo Luan, Tat-Seng Chua, "Brand Data Gathering From Live Social Media Streams", ICMR'14, April 01-04, 2014, Glasgow, United Kingdom.Copyright 2014 ACM 978-1-4503-2782-4/14/04