|| Volume 3 || Issue 6 || June 2018 || ISSN (Online) 2456-0774

INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

Survey Paper on Event Evaluation In Social Media

Ms. Kavita K. Cholke [1],Ms.Archana Panhalkar[2], Mr. R. S. Bhosale [3]

Student ,IT Department, Amrutvahini College of Sangamner, Ahmednagar India[1]

Assistant Professor,IT Department, Amrutvahini College of Sangamner, Ahmednagar India[2,3]

kavitacholke123@gmail.com[1], archana10bhosale@gmail.com[2], bhos_raj@rediffmail.com[3]

Abstract— **Twitter is a microblog benefit that creates an immense measure of literary substance day by day. This substance should be investigated by methods for content mining, characteristic dialect handling, data recovery, and other techniques. The large number of tweets submitted every day overpower clients who think that its hard to recognize substance of enthusiasm uncovering the requirement for occasion recognition calculations in Twitter. Such calculations we are proposed in this system covering both short (recognizing what is as of now happening) and long haul periods (inspecting the most striking as of late submitted occasions). For the two situations, we propose fluffy spoke to and auspicious developed tweet-based theoretic data measurements to display Twitter flow. The Riemannian separation is likewise abused as for words' marks to limit fleeting impacts because of accommodation delays. Occasions are identified through a multi-assignment diagram parceling calculation that: 1) ideally holds the greatest rationality inside a bunch and 2) while enabling a word to have a place to a few groups (occasions). Test comes about on genuine living information show that our approach beats different strategies.**

I INTRODUCTION

Microblogging is a broadcast medium that allows users to exchange small digital content such as short texts, links, images, or videos, Although it is a relatively new communication medium compared with traditional media, microblogging has gained increased Attention among users, organizations, and research scholars in different disciplines. The popularity of microblogging stems from its distinctive communication services such as portability, immediacy, and ease of use, which allow users to instantly respond and spread Information with limited or no restrictions on content. Virtually any person witnessing or involved in any event is nowadays able to disseminate real-time information, which can Reach the other side of the world as the event unfolds. For instance, during recent social Up heavals and crises. Social media refers to a set of different web sites that allow users to create, share and exchange content, such as social network sites, blogs, micro blogs, video shares, bookmarks, among others. The content generated is important in many research areas because there is information about various subjects in different contexts created from the users' point of view. Examples of applications of social media data mining include helping individuals and organizations to discover the acceptance level of products, the detection of disasters and anomalies, the forecast of the performance of politicians in election campaigns, the monitoring of diseases, to name but a few potential applications. When the database is composed of written documents, methods based on text mining, natural language processing, and information retrieval are usually applied. In the specific case of text mining approaches, tweets present different word distributions from one time period to another, as new trends appear for the discussed topics. This implies time varying document frequency metrics. Additionally, tweets are generated from different authors having different target audiences and/or writing styles, and they contain a number of extra symbols, misspelled or abbreviated words, resulting in a noisy estimation of the term frequency metric. Finally, Twitter's following and retweeting are important and they should be taken into account in the analysis. After text characterization, the second step is to extract events (equivalently sets of keywords) from the tweet posts by detecting common temporal similarities in their words' time series signals; the words of an event present a synchronized behavior in their appearance count. The research challenges at this stage are the following.

1) Tweet messages are often of unstructured meanings and the words of an event do not appear under a synchronized manner. This requires new forms of representation to compensate for the vagueness introduced by these temporal variations.

2) In contrast to conventional one-class assignment clustering methods, multi assignment clustering approaches are needed, since one word may belong to several events.

3) As words may belong to several events, clustering is not well-separable requiring the use of advanced methods, like graph partitioning.

Objective:-

The main objective of the system is

1) Optimally retains maximum coherence within a cluster.
2) While allowing a word to belong to several clusters (events). Experimental results on real-life data

demonstrate that our approach outperforms other methods.

Motivation:-

Large number of short text producing every day from various resources such as what Sapp or internet blogging. The short text may be leads to the misunderstanding of meaning of text. So there is need to get actual information and understand short text.

II LITERATURE SURVEY

1) "Shubhada Shimpi and Sambhaji Sarode", Survey on Tweet Timeline Generation and Summarization Methods, January 2017

Tweet are developed for the use of short text message and it is useful for both users and data analyst. Twitter which gets over 400 million tweets in line per day has emerged as a useful source of news, blogs, evaluations and extra. Our proposed work consists three components, first tweet stream clustering for clustering tweets using Bisect k -means cluster algorithm then second component tweet summarization cluster vector technique for generating rank summarization using greedy algorithm, therefore requires functionality which significantly different from traditional summarization and the third component is to detect and monitors the summary-based and volume-based variation to produce timeline automatically from tweet stream. Implementing continuous tweet circulation decreasing a text file is however not a simple assignment, for the reason that a huge range of tweets is worthless, unrelated and raucous in nature, due to the social nature of tweeting. Further, tweets are strongly correlated with their posted instance and up-to-the-minute tweets have a tendency to arrive at a very fast charge. Efficiency tweet streams are always very large in the stage, hence the summarization algorithm should be substantially successful. The flexibility it needs to offer tweet summaries of random moment periods. Subject matter evolution it must robotically locate sub- topic modifications and the moments that they appear.

2)"farzindar atefeh and wael khreich", a survey of techniques for event detection in twitter, number 1, 2015

Twitter is among the fastest-growing microblogging and online social networking services. Messages posted on Twitter (tweets) have been reporting everything from daily life stories to the latest local and global news and events. Monitoring and analyzing this rich and continuous user-generated content can yield unprecedentedly valuable information, enabling users and organizations to acquire actionable knowledge. This article provides a survey of techniques for event detection from Twitter streams. These techniques aim at finding real-world occurrences that unfold over space and time. In contrast to conventional media, event detection from Twitter streams poses new challenges. Twitter streams contain large amounts of meaningless messages and polluted content, which negatively affect the detection performance. In addition, traditional text mining techniques are not suitable, because of the short length of tweets, the large number of spelling and grammatical errors, and the frequent use of informal and mixed language. Event detection techniques presented in literature address these issues by adapting techniques from various fields to the uniqueness of Twitter. This article classifies these techniques according to the event type, detection task, and detection method and discusses commonly used features. Finally, it highlights the need for public benchmarks to evaluate the performance of different detection approaches and various features.

3)"Zhenhua Wang, Lidan Shou and Ke Chen", On Summarization and Timeline Generation for Evolutionary Tweet Streams, MAY 2015

Short-text messages such as tweets are being created and shared at an unprecedented rate. Tweets, in their raw form, while being informative, can also be overwhelming. For both end-users and data analysts, it is a nightmare to plow through millions of tweets which contain enormous amount of noise and redundancy. In this paper, we propose a novel continuous summarization framework called Sumblr to alleviate the problem. In contrast to the traditional document summarization methods which focus on static and small-scale data set, Sumblr is designed to deal with dynamic, fast arriving, and large-scale tweet streams. Our proposed framework consists of three major components. First, we propose an online tweet stream clustering algorithm to cluster tweets and maintain distilled statistics in a data structure called tweet cluster vector (TCV). Second, we develop a TCV-Rank summarization technique for generating online summaries and historical summaries of arbitrary time durations. Third, we design an effective topic evolution detection method, which monitors summary-based/volume-based variations to produce timelines automatically from tweet streams. Our experiments on large-scale real tweets demonstrate the efficiency and effectiveness of our framework.

4)"Willyan D. Abilhoa and Leandro N. de Castro", A keyword extraction method from twitter messages represented as graphs, 2014 Elsevier Inc.

Twitter is a microblog service that generates a huge amount of textual content daily. All this content needs to be explored by means of text mining, natural language processing, information retrieval, and other techniques. In this context, automatic keyword extraction is a task of great usefulness. A fundamental step in text mining techniques consists of building a model for text representation. The model known as vector space model, VSM, is the most well-known and used among these techniques. However, some difficulties and limitations of VSM, such as scalability and sparsity, motivate the proposal of alternative approaches. This paper proposes a keyword extraction method for tweet collections that represents texts as graphs and applies centrality measures for finding the relevant vertices (keywords). To assess the performance of the proposed approach, three different sets of experiments are performed. The first experiment applies TKG to a text from the Time magazine and compares its performance with that of the literature. The second set of experiments takes tweets from three different TV shows, applies TKG and compares it with TFIDF and KEA, having human classifications as benchmarks. Finally, these three algorithms are applied to tweets sets of increasing size and their computational running time is measured and compared. Altogether, these experiments provide a general overview of how TKG can be used in practice, its performance when compared with other standard approaches, and how it scales to larger data instances. The results show that TKG is a novel and robust proposal to extract keywords from texts, particularly from short messages, such as tweets.

5)"Shaomei Wu", Who Says What to Whom on Twitter, April 1, 2011.

We study several longstanding questions in media communications research, in the context of the microblogging service Twitter, regarding the production, flow, and consumption of information. To do so, we exploit a recently introduced feature of Twitter known as "lists" to distinguish between elite users—by which we mean celebrities, bloggers, and representatives of media outlets and other formal organizations—and ordinary users. Based on this classification, we find a striking concentration of attention on Twitter, in that roughly 50% of URLs consumed are generated by just 20K elite users, where the media produces the most information, but celebrities are the most followed. We also find significant homophily within categories: celebrities listen to celebrities, while bloggers listen to bloggers etc; however, bloggers in general rebroadcast more information than the other categories. Next we re-examine the classical "two-step flow" theory of communications, finding considerable support for it on Twitter. Third, we find that URLs broadcast by different categories of users or containing different types of content exhibit systematically different lifespans. And finally, we examine the attention paid by the different user categories to different news topics.

6)"Hila Becker", Selecting Quality Twitter Content for Events, 2011

Social media sites such as Twitter contain large amounts of user contributed messages for a wide variety of real-world events. While some of these "event messages" might contain interesting and useful information (e.g., event time, location, participants, opinions), others might provide little value (e.g., using heavy slang, incomprehensible language) to people interested in learning about an event. Techniques for effective selection of quality event content may therefore help improve applications such as event browsing and search. In this paper, we explore approaches for finding representative messages among a set of Twitter messages that correspond to the same event, with the goal of identifying high quality, relevant messages that provide useful event information. We evaluate our approaches using a large-scale dataset of Twitter messages, and show that we can automatically select event messages that are both relevant and useful.

7)"Haewoon Kwak, Changhyun Lee", What is Twitter, a Social Network or a News Media?, April 2010

Twitter, a microblogging service less than three years old, commands more than 41 million users as of July 2009 and is growing fast. Twitter users tweet about any topic within the 140-character limit and follow others to receive their tweets. The goal of this paper is to study the topological characteristics of Twitter and its power as a new medium of information sharing. We have crawled the entire Twitter site and obtained 41.7 million user profiles, 1.47 billion social relations, 4, 262 trending topics, and 106 million tweets. In its follower-following topology analysis we have found a non-power-law follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from known characteristics of human social networks [28]. In order to identify influentials on Twitter, we have ranked users by the number of followers and by PageRank and found two rankings to be similar. Ranking by retweets differs from the previous two rankings, indicating a gap in influence inferred from the number of followers and that from the popularity of one's tweets. We have analyzed the tweets of top trending topics and reported on their temporal behavior and user participation. We have classified the trending topics based on the active period and the tweets and show that the majority (over 85%) of topics are headline news or persistent news in nature. A closer look at retweets reveals that any retweeted tweet is to reach an average of 1, 000 users no matter what the number of followers is of the original tweet. Once retweeted, a tweet gets retweeted almost instantly on next hops, signifying fast diffusion of information after the 1st retweet. To the best of our knowledge this work is the first quantitative study on the entire Twittersphere and information diffusion on it.

8) Short text conceptualization using a probabilistic knowledgebase

Most text mining tasks, including clustering and topic detection, are based on statistical methods that treat text as bags of words. Semantics in the text is largely ignored in the mining process, and mining results often have low interpretability. One particular challenge faced by such approaches lies in short text understanding, as short texts lack enough content from which statistical conclusions can be drawn easily. In this paper, we improve text understanding by using a probabilistic knowledgebase that is as rich as our mental world in terms of the concepts (of worldly facts) it contains. We then develop a Bayesian inference mechanism to conceptualize words and short text. We conducted comprehensive experiments on conceptualizing textual terms, and clustering short pieces of text such as Twitter messages. Compared to purely statistical methods such as latent semantic topic modeling or methods that use existing knowledge bases (e.g., WordNet, Freebase and Wikipedia), our approach brings significant improvements in short text understanding as reflected by the clustering accuracy.

III PROPOSED SYSTEM

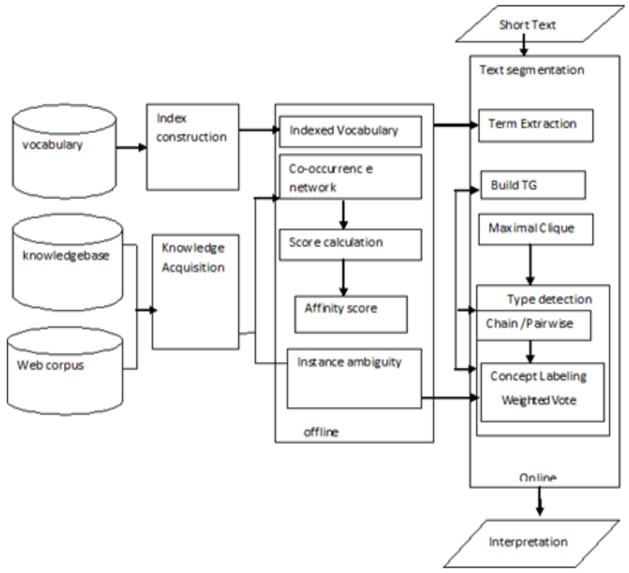

Figure 1 System architecture

Context-dependent conceptualization

Conceptualization seeks to map a short text (i.e., a word or a phrase) to a set of concepts as a mechanism of understanding text. Most of prior research in conceptualization uses human-crafted knowledge bases that map instances to concepts. Such approaches to conceptualization have the limitation that the mappings are not context sensitive. To overcome this limitation, we propose a framework in which we harness the power of a probabilistic topic model which inherently captures the semantic relations between words. By combining latent Dirichlet allocation, a widely used topic model with Probase, a large-scale probabilistic knowledge base, we develop a corpus-based framework for context-dependent conceptualization. Through this simple but powerful framework, we improve conceptualization and enable a wide range of applications that rely on semantic understanding of short texts, including frame element prediction, word similarity in context, ad-query similarity, and query similarity.Fig.1 illustrates our framework for short text understanding. In the offline part, we construct index on the entire vocabulary and acquire knowledge from web corpus and existing knowledge bases. Then, we pre-calculate semantic coherence between terms which will be used for online short text understanding. In the online part, we perform text segmentation, type detection, and concept labeling, and generate a semantically coherent interpretation for a given short text.

Modules

- **Indexing of vocabulary and knowledge acquisition.**
  Approximate term extraction aims to locate substrings in a text which are similar to terms contained in a predefined vocabulary. To quantify the similarity between two strings, many similarity functions have been proposed including token-based similarity functions (e.g., jaccard coefficient) and character-based similarity functions (e.g., edit distance). Due to the prevalence of misspellings in short texts, we use edit distance as our similarity function to facilitate approximate term extraction.
- **Text Segmentation.**
  We can recognize all possible terms from a short text using the tried-based framework described. But the real question is how to obtain a coherent segmentation from
  the set of terms. We use two examples to illustrate our approach of text segmentation. Obviously, fapril in paris lyricsg is a better segmentation of “april in paris lyrics” than fapril paris lyricsg, since “lyrics” is more semantically related to songs than two months or cities. Similarly, fvacation april parisg is a better segmentation of “vacation april in paris”, due to higher coherence among “vacation”, “april”, and “paris” than that between “vacation” and “april in paris”.
- **Type Detection.**
  Recall that we can obtain the collection of typed-terms for a term directly from the vocabulary. For example, term “watch” appears in instance-list, concept-list, as well as verb-list of our vocabulary, thus the possible typed-terms of “watch” are fwatch[c]; watch[e]; watch[v]g. Analogously, the collections of possible typed-terms for “free” and “movie” are f f ree[ad j]; f ree[v]g and fmovie[c]; movie[e]g respectively, as illustrated. For each term derived from a short text, type detection determines the best typed-term from the set of possible typed-terms. In the case of “watch free movie”, the best typed-terms for “watch”, “free”, and “movie” are watch[v], free[ad j], and movie[c] respectively.
- **Concept Labelling.**
  The most important task in concept labeling is instance disambiguation, which is the process of eliminating inappropriate semantics behind an ambiguous instance. We accomplish this task by re-ranking concept clusters of the target instance based on context information in a short text (i.e., remaining terms), so that the most appropriate concept clusters are ranked higher and the incorrect ones lower. Our

intuition is that a concept cluster is appropriate for an instance only if it is a common semantics of that instance and it achieves support from surrounding context at the same time. Take “hotel california eagles” as an example. Although both animal and music band are popular semantics of “eagles”, only music band is semantically coherent (i.e., frequently co-occurs) with the concept song and thus can be kept as the final semantics of “eagles”.

IV CONCLUSION

We propose a summed up structure to see short messages viably and proficiently. All the more particularly, we separate the undertaking of short content comprehension into three subtasks: content division, sort discovery, and idea marking. We detail content division as a weighted Maximal Clique issue, and propose a randomized estimation calculation to keep up exactness and enhance proficiency in the meantime. We present a Chain Model and a Pair astute Model which join lexical and semantic highlights to lead sort location. They accomplish preferable exactness over customary POS taggers on the named benchmark. We employ a Weighted Vote algorithm to determine the most appropriate semantics for an instance when ambiguity is detected. The experimental results demonstrate that our proposed Framework outperforms existing state-of-the-art approaches in the field of short text understanding.

REFERENCES

1) “Shubhada Shimpi and Sambhaji Sarode“, Survey on Tweet Timeline Generation and Summarization Methods, January 2017

2) “Farzindar Atefeh and wael khreich”, a survey of techniques for event detection in twitter, Number 1, 2015

3) “Zhenhua Wang, Lidan Shou and Ke Chen”, On Summarization and Timeline Generation for Evolutionary Tweet Streams, MAY 2015

4) “Willyan D. Abilhoa and Leandro N. de Castro”, A keyword extraction method from twitter messages represented as graphs, 2014 Elsevier Inc.

5) “Shaomei Wu”, Who Says What to Whom on Twitter, April 1, 2011.

6) “Hila Becker”, Selecting Quality Twitter Content for Events, 2011

7) “Haewoon Kwak, Changhyun Lee”, What is Twitter, a Social Network or a News Media?, April 2010

8)Short text conceptualization using a probabilistic knowledgebase Context-dependent conceptualization