SociNewsRank: Recognizing and Ranking Prevalent News Topics using Media Factors

M. Sravanthi[1] and Anita Kumari Singh[2]

[1]*M.Tech, Department of Computer Science and System Engineering, Andhra University College of Engineering, Visakhapatnam, India*

[2]*Research Scholar, Department of Computer Science and System engineering, Andhra University College of Engineering, Visakhapatnam, India*

[1]*(E-mail:sravanthimarpu562@gmail.com)*

[2]*(E-mail:anitasinghani@gmail.com)*

Abstract: *In early days, mass media sources such as news media provide us about daily events. In the recent period, unlike news media, social media services like Twitter generates a huge amount of user-generated data, which contains valuable information and become a famous research area in the latest technology. For these resources to be useful, we must predict the interactions between social media and traditional news streams by filtering out the unwanted noise and finding the prevalent news based on its similarity to news media. Even after noise is removed information overload may still exist in remaining data. To achieve prioritization, the information must be ranked in order of estimated importance considering three factors. First, the Media Focus (MF) of a topic the temporal prevalence of topic in news media. Second, User Attention (UA), the temporal prevalence of a topic in social media. Last, User Interaction (UI), the temporal prevalence of a topic in a social-user relationship. We proposed an unsupervised framework-SociNewsRank, which captures both news and social media data to recognize the news topics prevailing in both social media and the news media and then ranks them by popularity using their degrees.*

Index-Terms: *Information Filtering, Social computing, Keyword extraction, Topic Identification, Topic Ranking.*

I. Introduction

The extracting of valuable information from online sources has become a prominent research area in information technology in recent years. In the past, a knowledge that intimates the general public of daily events has been provided by mass media sources, specifically the news media which produce both hard-copy and Internet versions simultaneously. On the other hand, in social media, the Internet is being a free and open forum for exchanging information. Microblogs have become one of the most popular social

media outlets. One microblogging service like Twitter is used by millions of people around the world, providing vast amounts of user-generated data.

News media sources are considered as genuine because they are published by professional reporters, who are held liable for their content. In social media, one may assume that this source potentially contains information with equal or greater value than the news media, but one must also assume that because of regular, non-journalist users are able to publish unverified content and express their interest in certain events; much of its contents are useless. For social media data to be of any use for topic identification, we must find a way to filter uninformative information and capture only information which, based on its content similarity to the news media, may be considered useful or valuable.

Social media services like Twitter can also provide additional or supporting information to a particular news media topic. In summary, truly valuable information may be thought of as the area in which these two media sources topically intersect. Unfortunately, even after the removal of unimportant content, there is still information overload in the remaining news-related data, which must be given priority.

To aid in the prioritization of news information, news must be ranked in order of estimated importance considering three factors: First, Media Focus (MF) the temporal prevalence of specific topic in news media indicates that it is widely covered by news media sources, Second, User Attention (UA), the temporal prevalence of a topic in social media, specifically in Twitter, indicates that users are interested in the topic and can provide a basis for the estimation of its popularity. Last, User Interaction (UI), the temporal prevalence of a topic in a social-user relationship, the number of users discussing a topic and the interaction between them also gives insight into topical importance. By combining these three factors, we gain insight into the topical importance and are then able to rank the news topics accordingly.

Integrated, filtered, and ranked news topics from both proficient news providers and individuals have several benefits. The most evident use is possible to improve the quality and coverage of news recommender systems, adding user popularity feedback.

To achieve its goal, SociNewsRank uses two datasets to extract keywords from news media sources for a specified period of time to identify the overlap with social media from that same period. We then find a top topic and show how the prevalence of topics varies across news articles and tweets. The correlation between topics are identified, positive correlation shows that both topics are likely to be discussed. A graphical network shows how closely related topics are to one another. Finally, the topics are ranked and show how news influences social media.

II. Related Works

Related works are the basic step in preparing the new methodology for the particular area of the subject. Many researchers have been made their work on the latest advancements in the technology for the improvements in finding and ranking prevalent news topics using Social media. The main research areas applied in this paper include topic identification, topic ranking, keyword extraction.

Topic Identification:

Much research has been carried out in the field of topic identification—referred to more formally as topic modelling. Two traditional methods for detecting topics are LDA [1] and PLSA [2]. LDA is a generative probabilistic model that can be applied to different tasks, including topic identification. PLSA, similarly, is a statistical technique, which can also be employed to topic modelling. In these approaches, however, temporal information is lost, which is paramount in finding prevalent topics and is an important feature of social media data. Furthermore, LDA and PLSA only discover topics from text corpora; they do not rank based on popularity or prevalence.

Zhao et al. [11] carried out similar work by developing a Twitter-LDA model designed to identify topics in tweets. Their work, however, only considers the personal interests of users and not prevalent topics at a global scale.

Keyword Extraction:

Concerning the field of keyword or informative term extraction, many unsupervised and supervised methods have been proposed. Unsupervised methods for keyword extraction rely solely on implicit information found in individual texts or in a text corpus. Supervised methods, on the one hand, make use of training datasets that have already been classified. In this paper we used unsupervised methods; there are those that employ statistical measures of term informativeness or relevance, such as TFIDF [6], word frequency [7], n-grams [8], and word co-occurrence [9].

Topic Ranking:

Another major concept that is incorporated into this paper is topic ranking. There are several means by which this task can be accomplished, traditionally being done by estimating how frequently and recently a topic has been reported by mass media. Some works have made use of Twitter to discover news-related content that might be considered important. Sankaranarayanan et al. [10] developed a system called TwitterStand, which identifies tweets that correspond to breaking news. They accomplish this by utilizing a clustering approach for tweet mining

Wang et al. [12] proposed a method that takes into account the users' interest in a topic by estimating the amount of times they read stories related to that particular topic. They refer to this factor as the UA. They also used

an aging theory developed by Chen et al. [13] to create, grow, and destroy a topic. The life cycles of the topics are tracked by using an energy function. The energy of a topic increases when it becomes popular and it diminishes over time unless it remains popular. We employ variants of the concepts of MF and UA to meet our needs, as these concepts are both logical and effective.

III. Design and Implementation

The goal of our method—SociNewsRank—is to identify, consolidate and rank the most prevalent topics discussed in both news media and social media during a specific period of time. The system framework can be visualized in Fig. 1. To achieve its goal, the system must undergo three main stages.

1) **Pre-processing:** Key terms are extracted and filtered from news and social data corresponding to a particular period of time.
2) **Prevalent Topic Identification:** Top topic from news and social media are identified and show how the prevalence of topics varies across news articles and tweets. The correlation between topics are identified, positive correlation shows that both topics are likely to be discussed.
3) **Content Selection and Ranking**: Finally topics are ranked and show how tweets are related to news media.

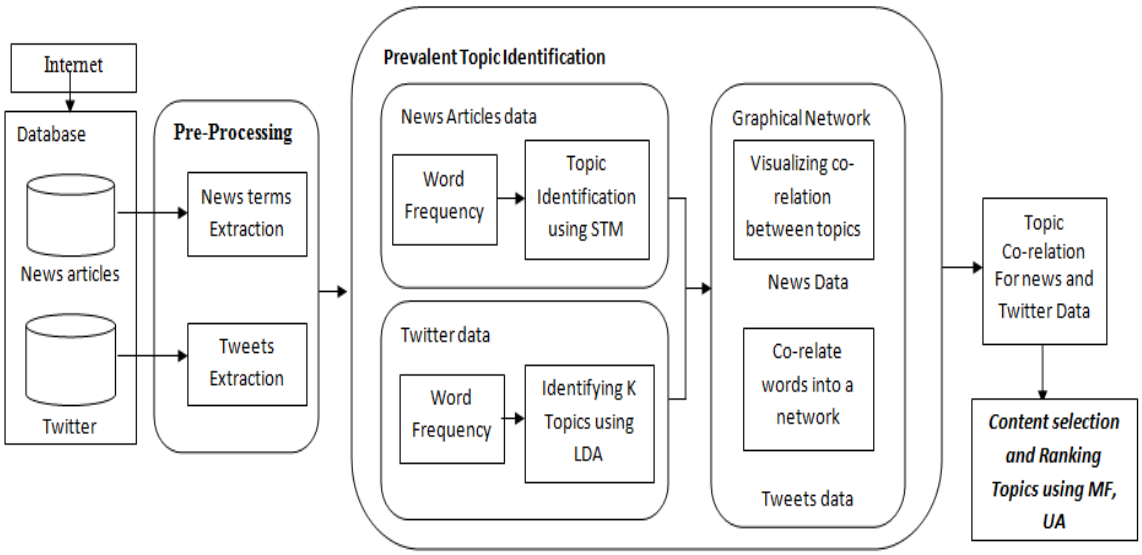

Fig.1 SociNewsRank Framework

Initially, news and tweets data are crawled from the Internet and stored in a database. News articles are obtained from specific news websites [5] and tweets are crawled from the Twitter public timeline [3], [4]. A user then requests an output of the top *k* ranked news topics for a specified period of time between start date d1 and end date d2.

A. Pre-processing

In the pre-processing stage, the system first queries all news articles and tweets from the database that fall within date *d*1 and date *d*2. Additionally, two sets of terms are created: one for the news articles and one for the tweets, as explained below.

i. **News Term Extraction**: The set of terms from the news data source consists of keywords extracted from all the queried articles. Due to its simple implementation and effectiveness, we implement a Structural Topical Model to extract the top 'k' keywords from news articles. The extracted keywords represents all unique terms that are added to set *N* which appear in news articles from date d1 to date d2. It is worth pointing out that, since *N* is a set, it does not contain duplicate terms.

ii. **Tweets Term Extraction*:*** For the tweets data source, the set of terms are not the tweets' keywords, but all unique and relevant terms. First, the language of each queried tweet is identified, disregarding any tweet that is not in English. From the remaining tweets, all terms that appear in a stop word list or that are less than three characters in length are eliminated. The part of speech (POS) of each term in the tweets is then identified using a POS tagger. This POS tagger is especially useful because it can identify Twitter-specific POSs, such as hash tags, mentions, and emoticon symbols. The extracted keywords added to set *T* represents all unique terms that appear in tweets from date's *d*1 to *d*2.

B. Prevalent Topic Identification:

In this component, top topics from news and social media are identified and a network graph is constructed.

i. **Term Frequency:** First, the frequency is calculated for the extracted keywords from both the news and social media. In the case of term set 'N' represents the document frequency of each term 'n' is equal to the number of news articles (from dates d1 to d2) in which n has been selected as a keyword; it is represented as df(n). The document frequency of each term't' in set 'T' represents the tweets data is calculated in a similar fashion. In this case, however, it is the number of tweets in which t appears; it is represented as df(t). For simplification purposes, we will henceforth refer to the document frequency as "occurrence." Thus, df(n) is the occurrence of term n and df(t) is the occurrence of term "t".

ii. **Topic Modelling:** Next, the topics from the both news and tweets are identified based on some models. These models identify the top topics that are prevalent in both news and social media. A documents length clearly affects the results of topic modelling. For extremely short texts (e.g. Twitter posts) or extremely long texts (e.g. books), it can build sense to concatenate/split single documents to receive longer/shorter textual units for modelling. As an unsupervised method, topic models are suitable for the exploration of data. The calculation of topic models aims to determine the proportionate composition of a fixed number of topics in the documents of a collection.

- In case of news articles we use Structural Topic Model (STM) which is a general framework for topic modelling with document-level covariate information, which can improve inference and qualitative interpretability by affecting topical prevalence, topic content, or both.
- In case of tweets, after pre-processing, a corpus is created for tweets data on which we calculate a Latent Dirichlet Allocation (LDA) topic model [1] to find the top topics from tweets which shows some of the word related to news.

iii. **Graphical Network:** Once Topic Modelling is done, we have to calculate pair wise correlations within topics. This function requires "igraph" R package.

- **News:** A graphical network display shows how closely related topics are to one another (i.e., how likely they are to appear in the same document).Positive correlation between topics indicate that both topics are likely to be discussed within a document.
- **Tweets:** We may be interested in visualizing all the relationships among words simultaneously, rather than just the top few at a time. As one common visualization, we can arrange the words into a network or graph. Here we will be referring to a "graph" not in the sense of visualization, but as a combination of connected nodes. Next, step is to identify the relevant topics from both social media and news media.

C. Content Selection and Ranking:

Now, the prevalent Topics from both the social and news media that fall within the date d1 and d2 have been identified, relevant content from the two media sources that is related to these topics must be selected and finally ranked .Related items from the news media will represent the Media Focus (MF) of the topic. Similarly, related items from the social media (Twitter) will represent the User Attention (UA)-more specifically, the number of unique Twitter users related to the selected tweets.

IV. Results

a) After extracting the news terms and tweet terms we have to find the most frequent words from both the Media sources and the following output screen shows the visualization of frequent words in the form of word cloud

Fig. 4.1 Word Cloud

b) Graphical Network for words and topics is shown in the following output screen:

The below screen results in finding the association between the word .

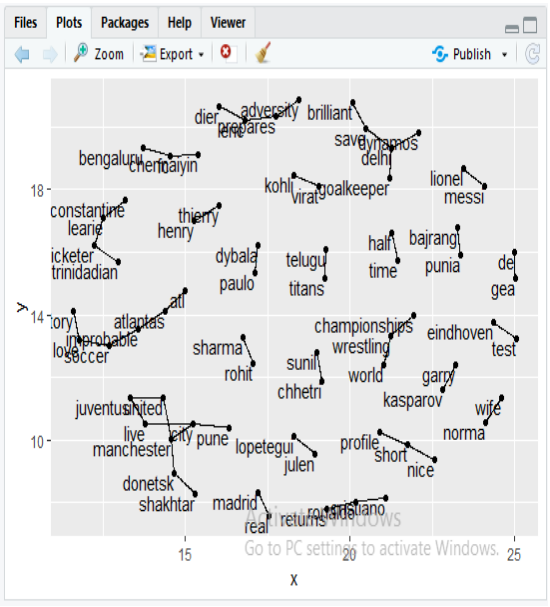

Fig. 4.2 Association between Words

The below output shows the correlation between the topics from different news articles and tweets.

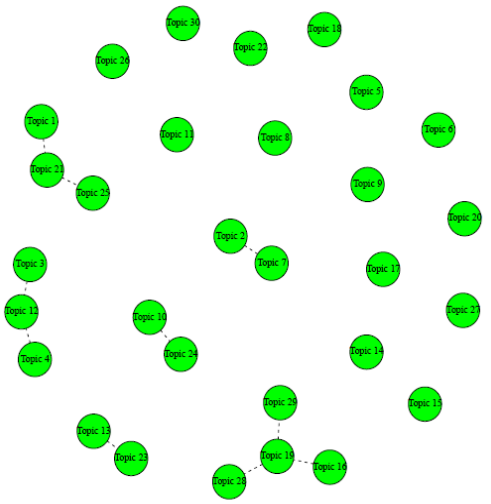

Fig. 4.3 Co-relation between Topics

c) Prevalent news that fall within the dates d1 and d2 have been identified, relevant content from the two media sources that is related to these topics must be selected and finally ranked. The ranked news topics are shown in the following output screen

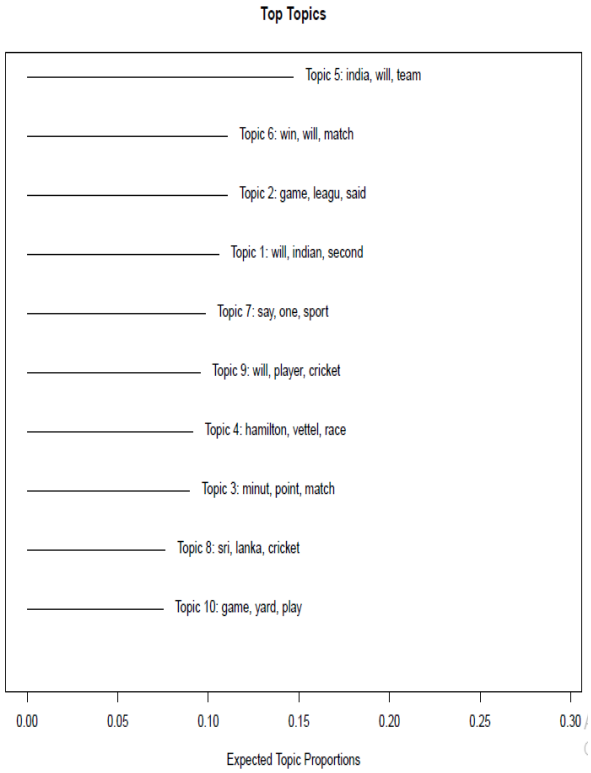

Fig 4.4 Topic Ranking

V. Conclusion

In this paper, we proposed an unsupervised method-SociNewsRank-which identifies news topics prevalent in both social media and news media, and then ranks them by using MF and UA as a relevance factors. This paper includes jointly topic modelling on two datasets that fall within the date d1 and date d2, which benefits the modelling performance for both long and short texts. We present the results of applying model to two datasets and show its effectiveness over non-trivial baseline. Based on the outputs of the model, further efforts are made to understand the complex interaction between news and social media data. Through extensive experiments, we find following factors: Even for the same events, focuses of new and Twitter topics could be greatly different. Topic usually occurs first in its dominant data source, but occasionally topic first appearing in one data source could be a dominant topic in another data set. Generally, news topics are much more influential than Twitter topics. Finally, the relevant topics from both the Medias are ranked using MF and UA .

References

[1] M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," J. Mach. Learn. Res., Vol.3,pp.993-1022,Jan.2003.

[2] T.Hofmann, "Probablistic Latent semantic Analysis," in Proc. 15th Conf. Uncertainity Artif. Intell., 1999,pp-289-296.

[3] httpps://medium.com/@rohitnair 94843/analysis-of twitter-data-using-r-part-1-twitter-authentication-d6cd67678ad7

[4] Twitter.[Online].Available:http://www.teitter.com,accessed Feb.2014.

[5] https://tm4ss.github.io/docs/Tutorial 1 Data acquistion.html

[6] G.salton,C.-S Yang, and C.T.Yu,"A theory of term importance in automatic text analysis," J.Amer.Soc.Inf.Sci,vol.26,no.1,pp.33-44,1975.

[7] H.P.Luhn,"Astatistical approach to machanised encioding and searching of literary information," IBMJ.Res. Develop.,vol,no.4, pp.v309-317,1957.

[8] J.D. Cohen, "Highlights:Language- and domain-independent automatic indexing terms for abstracting," J.Amer.Soc. Inf.Sci.,vol.46, no. 3, pp.162-1734,1995.

[9] Y. Matsuo and M.Ishizuka,"keywod extraction from a single document using word co-occurence statistical information," Int. J. Artif. Intell. Tools,vol.13,no. 1 . pp.157-169.

[10]J.Sankaranarayanan,H.Samet,B.e.teitl er,M.D.Liberman,andJ.Sperling,"TwitterStand:News in Twets," in Proc. 17th ACMSIGSPATLAL int. Conf. Adv. Geograph. Inf. Syst., Seattle,WA,USA,2009,pp. 42-51.

[11] W. X. Zhao et al., "Comparing Twitter and traditional media using topic models," in Advances in information retrieval. Heidelberg, Germany: Springer Berlin Heidelberg,2 011,pp.338-349.

[12] C. Wang, M. Zhang, L. Ru and S. Ma, "Automatic online news topic ranking using media focus and user attention based on aging theory," in Proc.17th conf. Inf. Knowl. Manag. Napa Country, CA, USA, 2008, pp.1033-1042.

[13] C.C. Chen. Y.-T. Chen, Y.sun, and M. C. Chen, "Life cycle modelling of news events using aging theory," in Machine Learning: ECML 2003. Heidelberg, Germany: Springer Berlin Heidelberg, 2003, pp. 47-59