Proceedings of the 6th International Conference on Communication and Electronics Systems (ICCES-2021)
IEEE Xplore Part Number: CFP21AWO-ART; ISBN: 978-0-7381-1405-7

Comprehensive Survey on Techniques of Topic Evolution Mining

Pramod Bide,Varun Magotra, Jainam Jain,Pritam Rao,Kunal Jain
Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, India
{pramod_bide,varum.magotra,jainam.jain,pritam.rao, kunal.jain}@spit.ac.in

2021 6th International Conference on Communication and Electronics Systems (ICCES) | 978-1-6654-3587-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCES51350.2021.9488965

***Abstract*—With the tremendous increase in use of social media, information is constantly propagating online. Certain topics or events are popular at a given time, and are said to be trending on social networks such as twitter. Given the ease and pace at which information is generated on twitter, the trending topics are constantly evolving. When a twitter user starts tweeting with respect to the root event, they give way to multiple sub-events and subtopics, which themselves continue to evolve. Evolution of topics in such highly co related topics is quite interesting. Mining the evolution of such topics is an intriguing task which can yield interesting facts about the socio-political, economical or cultural climate of a place or community. It has applications in marketing, political campaigns, trend analysis and so on. Such research can also enable the tracking and prediction of twitter activity in the future. Due to the massive data, it has become increasingly difficult to cluster the interesting events from vast social media data. This paper comprehensively describes and summarizes the previous research in topic evolution of twitter data and provides a comparative analysis of the works in this domain, highlighting areas of future improvement.**

***Index Terms*—Social media analysis,Event clustering, Twitter activity, Data mining, Topic evolution, Natural language data, Topic detection**

I. Introduction

In recent years, many social medias sites have emerged like Flickr, YouTube, Facebook, and Google News. People can easily generate and share social content online in the form of various media. Therefore, these social media platforms serve as a huge repository of textual information generated by users, often about the events happening at a given time. A popular event that is taking place can spread very fast, and there are substantial amounts of social events with multi-modality (e.g., images, videos, and text) in Internet.

In real-world scenarios, most of the content and media posted on these networking sites is associated with social events happening and is related to specific topics. It is cumbersome for an analyst to manually identify topics and cluster the events. For example, a user wishes to know the entire theme evolutionary process of the event "Occupy Wall Street" for beginning to finish. When they search for the particular event on Google News or Flickr using the search engine given well-defined queries, they could be provided much more information of related events and topics. It would be highly beneficial if we could identify evolutionary trends of social events and visualize the progression over time of various themes, which is the goal of event tracking and event evolution.

Natural language data, particularly that found on social networks, is mostly unstructured. [1] This makes it tougher for machine learning models extract information from such data. Moreover, the incredible rate at which tweets are generated means that topics are constantly evolving. For example, the event of the Pulwama Attack of 2019 generated floods of tweets online, and also had many related events which were trending. Multiple topics and sub-topics such as Indo-Pak war, war crimes, the capture of soldiers, calls for blood donation for Indian soldiers trended over a period of time after the attack.

Topic evolution is usually a two step processing that first consists of identifying what the topics are and then keeping track of the changes in topics over time. There are various existing methodologies for identifying topics from text data. These include keyword-based approaches that find frequent n-grams. [2] [3] Also exist clustering based techniques as cited in [4]. Machine learning methods have become increasing popular, such as Latent Dirichlet Allocation for topic modelling. [5] In recent years, neural nets have been increasingly used to learn vectorized representations of words and sentences. [6] This makes semantic clustering of words easier, unlike frequency based methods where two words spelt different but of the same meaning will be treated differently.

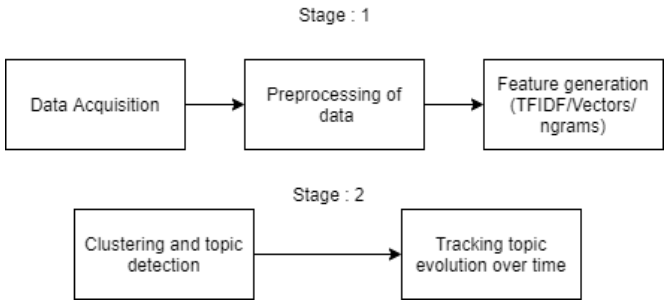

Fig. 1: General pipeline for tracking topic evolution

Topic evolution explores the change with time of topics identified with the methods described above. Previous work on topic evolution deals with detecting evolving topics for individual users as well as entire twitter communities, for unrelated as well as correlated events. This paper particularly

focus of evolution of topics within an existing event as opposed to unrelated events. The creation and evolution of subtopics is discussed. A total of 15 prominent research works pertaining to topic evolution have been compared and described in detail, from data acquisition, preprocessing, event detection and topic identification, to tracking of evolving topics in temporal data, as well as predicting possible future topics. The advantages as well limitations of these papers have been described in a further section.

The rest of the paper is organized as follows. In Section II, the related literature is reviewed. The surveyed literature is compared and contrasted in section III, the Discussion. The conclusions derived from the discussion area explained in section IV, summarizing which approaches work best, and possible areas of future work.

II. Literature Survey

Claudia Lauschke et. al. propose a method for analyzing Twitter users and creating a profile of their activity based on long term interests and short term trends(everyday events or reactions). The aim of this study is to monitor evolution of user profiles over time on blogs and microblogs. [7] For user modeling, a collection of documents D is monitored at discrete time periods. A summary of the content published is generated for each time period. These summaries are used to track changes over time. Using only keywords for user modeling would result in a coarse summar y, so instead, a more elaborate model of the profile is created by using the topics that exist in the collection. The change in two topics is measured using KL-divergence of their keyword distributions. KL scores range from 0 to infinity, where 0 denotes identical topics. Evolution between two topics exists if the distance lies below a particular threshold.Thus, an evolutionary transition graph is created, which suggests how stable or volatile the profile is. The observation time periods for each user are determined dynamically based on the number of tweets over time, since every user posts at different rates. For topic extraction from tweets, the bisecting k-Means clustering technique is used. After summarizing the topics and monitoring changes between consecutive time periods, changes are reported such as number of survivals, number of appearances and number of disappearances. This methodology was applied on a real self-crawled Twitter dataset of a predefined list of users which included famous people in various fields. The list consisted of a politician, a journalist, a pop star and two scientists/professors. The results show how topic survival rates over time periods vary widely for different types of users. Some topics appear in almost every time period, whereas some emerge and disappear again based on current events. A major challenge posed was the usage of Internet slang, which makes preprocessing of tweets and topic detection difficult. Thus future work will have to focus on improved preprocessing techniques for the dataset and better methods for detection of topic and topic continuation discovery. Also, the size of the dataset can be increased for better results [7].

Topic modelling has been used to identify growing trends by analysing social media platform. Conventional methods such as LDA can be used to identify topics from a standard corpus of text. Methodologies based of LDA such DTM and Online LDA are based off LDA to identify trends over texts that are generated sequentially over time. However they suffer from the drawback that topic number or the total number of topics must be predetermined as an external parameter. On a more complicated note, it is not necessary for the topic number to remain constant as time passes. The number of topics can increase or decrease over many epochs. Another drawback that is observed is that these models are designed for long texts, whereas posts on social media are usually short texts. A non-parametric dynamic topic modelling approach suggests the use of the Chinese restaurant process as a basis [8]. The co-occurrence model is used to tackle the data sparsity problem. The short text is broken down into bi-terms, an unordered pair of two words which exist in the text. The assumption made here is that as the two words exist in the same text, they relate to the same context, and hence the topic. The bi-terms belong to different time epochs. To model topic distribution across time epochs, the recurrent Chinese restaurant process is used as a non parametric technique to evaluate the prior distribution of topics based on the bi-terms. This distribution is then used to generate bi-terms for the corresponding topic. The Chinese restaurant process or CRP for short, works on the myth that a Chinese restaurant always has an empty table. So when a new customer enters, the customer can sit at one of the occupied existing table, probability of it happening being directly proportional to the number of customers already sitting there, or the he can choose to sit at a new table with some other probability. By using temporal information as well, the method is modified to create the Recurrent Chinese Restaurant Process. As a distribution gets older, it is given less importance as a factor in the current topic distribution, which is demonstrated using exponential decay. In this methodology, there are two main factors on which topic distribution in the current epoch depends on. First the past distributions for all topics, and secondly, the topic distributions of all other topics except the one being considered. Gibbs Sampling was done to infer the parameters, and the methodology showed improved results over the baseline DTM and Online LDA.

Wei Liang et al. use a combination of topic modelling and mining of social network for research topic evolution. A collection of textual documents, the profile of the researcher and the researcher's information gathered from social media are initially pre-processed to eliminate the irrelevant, incomplete and noisy data and then fed as an input to the proposed topic evolution model consisting of 3 layers. The fluctuation of research topics with time depend on the scientific activity of the researcher. Thus, the input data is used to extract scientific activity with is defined by characteristics such as time, location, field, category, impact factor and researcher. Further, clusters of similar research topics are formed and graphical structures are used to represent relationships between researchers in a specific time slot. Topics and authors of 500

research papers from 2001 to 2010 are used as the data set and the time unit is considered to be 1 year. Thus, the proposed topic evolution model is able to analyze how the research topics vary over a time period of ten years [9].

The authors of this paper [10] have created a state of the art framework, Sumblr. It is dynamic and works on large datasets, this gives it an edge over the traditional frameworks used. IT supports continuous stream summarization. It has three major parts. The first one, Tweet Cluster Vector deals with clustering of tweets and keeping the statistics. The second one is a Tweet Cluster Vector-Rank summarization to generate online and historical summaries of time spans. The last part deals with topic evolution detection. The promising results boast about its efficacy over the traditional methods.

Ruoran Liu et. al. integrated temporal pattern learning along with topic analysis for mining phase evolution of real time events. The number of posts of a particular topic are identified along with the occurrence, development, climax, decline and ending phases of the topic evolution and a quantitative model is developed. Real world datasets are obtained from social media platforms such as Weibo and News. Temporal sequence of the number of post's local maximal and local minimal points are clustered using k-means algorithm. The clustering returns two new subsequences of maximal and minimal clusters which are then used to identify development, decline and ending patters. Peak interval detection using burst detection technique and dynamic programming is carried out to identify fine grained attention. Topic analysis is used to extract keywords from different posts of each phase to describe the topics using Summarization technique such as TextRank, and it comprises of generation of word filter, construction of word association network and word importance ranking. Uniform partition of 5 phases and Extended burst detection is utilized for comparison. Evaluation metrics such as Overlapping ratio and Inverse Phase Frequency are used and Weibo gives the value 90.3% and 3.91 whereas News has a value of 67.7% and 3.22 respectively [11].

Organizations and retailers are strongly looking into microblogging platforms like twitter to know and understand their consumers and thereby predict buying patterns. Although, sometimes a surface level analysis does not provide truly valuable insight and must contain more detail to it to produce actionable insight. An LDA based approach to identify topics evolving out of a tweet combined with a regional analysis of the information retrieved can provide a much-detailed insight into what a consumer expects or how he might behave. Also, the regional distribution of topics can be further analyzed to identify clusters of regions and understand how certain geographical regions behave similarly. A region-topic matrix is constructed and then is subjected to singular value decomposition and a meta-clustering approach for a region-topic affinity and clustering. There is a strong tendency of culturally similar or similar speaking language countries or regions to talk about the same topics, which businesses can leverage proactively [12].

In [13], the temporal aspects of a user's behaviour are explored based on the frequency of activity of the user. Usually the maximum number of tweets come from a small portion of the users and it is also found that their tweeting time changes if an important event has occurred. They introduced the concept of "tweet strength which gives more weightage to users who do not tweet frequently. The active users who tweet frequently are given less weightage. The topic evolution mechanism is studied through this and the users are identified depending on the evolution pea and topic popularity. This can be helpful to gain insights about an anomaly event.The replicated topics of tweets are studied which indicate it has been copied.

The empirical analysis suggests the relationship between topic, infrequent and frequent users. They conclude that particular topics should disperse into different communities instead of staying localized. Also the number of people one follows should correspond to the tweeting activity. A very detailed analysis has been done on the temporal nature of users.

Sana Malik et. al. presented an use of binned topic models and statistical topic modelling to assemble tweets about relatively similar topics under automatically generated topics. Their application aimed to ensure that complex trends were not ignored as the size of Twitter data increased. Visualization of the development of the topics was performed by TopicFlow which is highly interactive tool. TopicFlow itself covers two sectors - automatic topic generation fro tweets and visualizing topic trends over a period of time. Latent Dirichlet Allocation (LDA) algorithm is used for carrying out the statistical topic modelling. The topics are then aligned and visualized using TopicFlow. To confirm the accuracy and the extent of usability of TopicFlow to explore Twitter data-sets, a preliminary usability test with 18 volunteers having ranging ages and including both the genders was performed. The volunteers were given a short description to the tool and a total of 7 separate tasks and once they had completed the tasks using the tools they were asked to rate the tool on a 20-point Likert scale where a higher rating would indicate better results. The NASA Task Load Index was used to choose the rating metrics : performance, effort, frustration, and effectiveness of the tool. The means and standard deviations of the results were found to be varied widely depending on the task. The results show that the interface provided by TopicFlow enabled the users to smoothly navigate through the tool and easily perform the given tasks which strengthens the initial hypothesis. It was found that for tasks which involved identifying details regarding the topics, time taken to perform the tasks was the least and was about 10 to 20 seconds on average. For tasks involving number of tweets in a specific topic, it took about 30-50 seconds on average which was greater than the previous scenario. The task which took the volunteers the longest was scanning through the tweets for analyzing trends, it took 81.2 seconds on average to perform the task. The test showed that the responsiveness of the main visualization tool to interactions was one of the favourite part of the participants and tweet list pane was their least favourite [14].

Sheikh Motahar Naim et. al. presented a unique approach which involves combination of various types of information

together to find hidden themes in text based documents. As there is a large amount of text data available for analysis, this approach may be really helpful for effective analysis. There are many algorithms that make use of labels and other annotations to find the hidden themes. The approach presented in this paper involves use of two different dimensions of information-timestamps and labels to reach the goal of finding the themes effectively. It is observed that this approach is faster than the traditional approaches. Only document-level annotations are considered in this paper and content-level might be considered in the future [15].

Shengsheng Qian et. al. [16] identified the difficulty in finding and organizing interesting events because of the growth of social media and such events being held via the Internet. They presented a novel approach in solving this problem. They designed a multi-modal event topic model, which captured the various different topics pertaining to social events and identified evolutionary trends of these events. This was achieved by modelling social media documents to understand the relation between textual data and related image data, which helped to separate visual and non visual representative topics. To apply it for social event tracking, they made use of incremental updation. The experiments on real world dataset demonstrates its efficiency as it performs better than all existing models. The use of domains like Youtube and Google News can further improve the performance of the model. Addition of tasks like summarization of events and identification of important event attributes through data mining can be used to further improve the utility of the framework.

The work by Mansoureh Takaffoli et al. [17] identifies the topic evolution problem with the help of communities, which are clusters of graphs of individuals densely connected at a particular time instance, and meta communities which are similar communities at different time instances. The evolution analysis in the communities is done dynamically by operations such as Form, Dissolve, Survive, Split and Merge, at each snapshot. They also use a community matching algorithm with the help of weighted bipartite graphs to consider preference for similar communities in closer temporal proximity.

Also covered in this survey was the work of Muhammad Abulaish et al [18], who proposed a word embedding based approach. Evolution events of five categories are identified for the topics. These include emergence, persistence, convergence, divergence, and extinction. For topic detection, LDA is used on the tweets. Therefore LDA modelling is used to find the topics at a given instance of time. The time period for which the evolution is being noted is divided into m sections. LDA is used to extract topics for each of these sections and the topics are represented as a m-partite graph. Then, the closeness between multiple topics is calculated using both explicit and implicit proximity. For explicit proximity, the co-ocurring words between two topics are considered to indicate that they are closely related. For topics which mean the same but are represented by different synonymous words, the methodology uses word embeddings from Glove. Glove enables us to map textual data into an n-dimensional vector space. Topics that are close are connected using edges showing transition of topics from one time interval to another Different thresholds are set to determine the various evolution events. The overall methodology seemed promising and detailed, and worked well on certain topics of social media data.

The paper by D.Lipika et al [19] applies this domain specifically to regional analysis of tweets which has applications in business analytics. Twitter4J API is used for collecting a large volume of tweets. Tweets are geo-tagged using time-zones, and geo-location meta data. Additionally, duplicates and near duplicates are grouped together into buckets to reduce the overall volume of distinct tweets to be processed. LDA is used to generate topics for a particular day, modeled as a distribution of words. Topic evolution is detected by finding correlation between topics a particular day and the days preceding it. In addition, the geo-tagging data extracted earlier is used to find regional affinities in the topics and find trends across regions globally.

The work by Arpaci et al. [20] analyses data from Twitter and performs Evoluionary Clustering algorithms regarding the COVID-19 pandemic. The work makes use of around 43 miillion tweets, containing thousands of unigrams, bigrams and trigrams. The evolutionary method makes use of K-means clustering and determines the number of clusters using the Elbow method. Thus, the clusters of the previous day acted as the initial state for the present day. These clusters signify the frequency level of the terms associated with the topic. It was determined that unigrams showed a larger trend that bigrams and trigrams, and accoringly, a time series graph was obtained for the three n-grams, distributed in 6 clusters, based on the descening order of frequencies.

Cai et al. [21] proposed an LDA-based technique that performs well with microblogging sites like Twitter. Their work also takes makes use of vital components like topic tags, authors and microblog document relations through special symbols used by the sites. The authors established two ways of temporal topic dynamism - content evolution and intensity evolution, the results for which are described as a progressiive graph with respect to time. Furthermore, this model tends to perform significantly better than the traditional LDA approach, as determined from the Perplexity measure of the models.

III. Comparative Table and Discussion

After surveying the papers, we selected the best ones which help expound the topic of this paper. A comparative analysis of these has been done. Table I gives a detailed comparative analysis of the different methodologies surveyed:

The comparison above tells us that the scope and applicability of topic evolution techniques is varied, and the success of the previous research works differs based on the domain, use case and dataset.

Firstly, scope is very important as we may be finding common topics for individual users, communities, organizations or the entire social media population. While some papers cover broader topics like news and sports, others are specific to communities, such as topic evolution of research articles. The

TABLE I: COMPARATIVE STUDY

METHODOLOGY	***DATASET***	***ACCURACY***	***RESEARCH RESULTS***	***LIMITATIONS AND GAPS***
The methodology consists of user modeling based on topic, change detection (based on KL-divergence of corresponding keyword distributions) and monitoring and reporting changes between consecutive time periods [7].	A real world self-crawled dataset from Twitter, containing a predefined list of well-known people in various fields.	Each type of user posted ondifferent topics, each havingvarying survival rates overtime periods.	Some topic chains are stable with time while others evolve over a period. Some subjects define a user and are present in most time periods, others appear and then disappear soon.	Incorporating Internet slang and using more elaborate methods for topic detection and topic continuation discovery. The size of the dataset can also be increased.
The Word Co-occurrence model is used for short-texts, by considering the bi terms in the text. Recurrent Chinese Restaurant Process is used to calculate the priors. Each preceding topic distribution plays a role in determining the current topic distribution, however the influence decays exponentially over time. Gibbs Sampling is done for parameter inference. [8]	English tweets from month of July from the twitter7 dataset were used. The vocabulary had the top 10,000 frequent words and 10,000 tweets sampled daily for three weeks.	Topic Coherence and topic distinctiveness are used as parameters for evaluation. The proposed methodology showed higher scores for both over the selceted baseline methods DTM and Online LDA	The model was designed specifically for short-texts as multiple models and tackled the issue of identifying the topic number and determining the topic number for various epochs. It scored higher than both the baseline models.	No reason was given to selecting the baseline models. Those models were not designed for short texts, hence were at a natural disadvantage. The model considers that time epochs are of the same length which may or may not be true.
The phase-evolution mining model consists of Temporal Pattern Mining. Temporal sequences clustered using k-means, then phase partition and detecting peak interval carried out. Topic analysis for Filter Generation, word association and importance. Main topics are summarized using TextRank. [11]	Data about Rio Olympics obtained from Weibo and News dataset from August 1,2016 to August 31, 2016, containing 10,166 and 319,785 pieces respectively	Overlapping Ratio for Weibo and News dataset is 90.3% and 67.7% and Inverse Phase Frequency for the same is 3.91 and 3.22 respectively	Phase partition results help understand event evolution. Topic analysis of each phase is done to respond to public opinion in a timely manner. on Weibo, the growth rate is more than 200%, but less than 100% on News media platform. For News, the ending of climax phase is earlier than Weibo.	Temporal patterns can be enhanced to improve performance and accuracy of News dataset.
An LDA based generative model is used for topic extraction followed by performing a regional analysis by constructing a region-topic matrix and then performing singular value decomposition and taking a meta clustering approach [12].	Twitter4J API, that gathers twitter streams on the basis of specific keywords was used.	The method ends up clustering similar speaking language countries like Portugal and Brazil together and the Latin American Countries together as they speak Spanish	There is a high tendency of countries and regions that are like each other be discussing the same topics and having the same topics trending	Regional clustering approach needs a user input on the number of days two regions have to be talking similarly to cluster them
Visualization of topics is proposed using statistical topic modelling along with binned topic models. This is done by grouping relatively similar tweets into the topics that automatically generated using the interactive visualization tool, TopicFLow [14].	Current events such as presidential debates and Hurricane Sandy, common interests, communities, and some historical data sets.	Binned topic models were most accurate for events that occurred in real-time and a short span. 18 volunteers participated in the study and each participant had different ratings for different tasks. Means and standard deviations of these are used to get the final rating.	TopicFLow's features were extensively tried out by the participants through given tasks. Main Visualization and dynamic interactions and tips were the most liked features from the tool. Visual aids such as bar charts were liked by the participants while Tweet List pane and Filter Pane were least liked.	Proposed methodology is limited to Twitter data. Also, there is a restriction brought on by the screen space on the number of topics and bins that can be effectively visualized using the tool.
A collection of textual documents, and the profile and information of the researcher are initially preprocessed to eliminate the noisy data and fed as input to the topic evolution model. The input data is used to extract scientific activity defined by time, location, field, category, impact factor and researcher. Clusters of similar research topics are formed. [9]	Topics and authors of 500 research papers from 2001 to 2010, time unit is considered to be 1 year.	Trends for 4 topics : image compression, wireless sensor, wearable computing, support vector, were obtained using a time series diagram. Researcher relationships were plotted and red nodes represent researchers who have written a large number of papers in their fields.	The time series diagram showed that 'image compression' continued to decrease in popularity until it completely vanished in 2009. Wearable computing emerged as a new topic in 2004 and then kept increasing in popularity. 'Wireless sensor' remained popular consistently from 2001 to 2010.	Similar topics were grouped under the same category and insignificant topics were ignored. Few existing topics were discarded. In future the method can be used for plotting the trends of Facebook and Twitter groups and topics.
The methodology consists of adopting a multi-modal event topic model, that analyses social events on the basis of textual and visual data related to it present on social media [16].	Manually collected from social media websites for 8 different social events over past few years.	Model outperforms existing models for the same task.	The model identifies the correlation between textual and visual representation of data regarding social events, and proposes the use of an incremental strategy to apply it for social event tracking.	The domains considered are limited. Efficiency needs to be tested on bigger domains like Google News.

methodology suggested in [7] is useful for cases when user-specific activity is to be noted, as in the definition of user profiles in [9]. Conversely, when the topic on social media as a whole is being studied, topic modelling techniques like LDA and word co-occurrence models, ad clustering techniques are relevant, as fine-grained insights on specific user activity are not needed.

As paper [11] demonstrates, the accuracy of results varies greatly based on the dataset, as noted by Weibo and News datasets in this example. Further, future approaches will also have to focus significantly on the pre-processing of data, as slang is very common on social media and confuses machine learning models. Therefore we can conclude that the choice and nature of dataset is very crucial.

The dynamic and evolving nature of topics means that an unsupervised approach is best suited, as the papers that didn't restrict the number of topics gave superior results against methods such as LDA which require the number topics to be predefined.

Considering that social media topics often need to be analysed for large audiences over massive data, the paper by Ruoran Liu et al [11] paper addresses the challenge of identifying temporal patterns also automatically mines topics of different phases. The use of K-means along with burst detection as well as Text Rank gave strong quantifiable results of 90% on the Weibo dataset, which was a large real-world dataset. This gave more tangible results as opposed to the other papers which demonstrated their results with examples but not quantifiable metrics.

IV. Conclusion

Thus this study covers a variety of methodologies for tracking evolution of topics on social media. Different methodologies perform well in varying scenarios. We are able to ascertain the best options for social media mining, and also analyse flaws in each research work to look for directions of future research. Refer Table I. Overall the techniques have good accuracy and are all-encompassing. They include the various social media trends, have performed topic classification, temporal analysis and user frequency estimation. These are the key factors that affect topic evolution mining. While these methods can be used for future experimentation, they need to be improved in certain areas. Insufficient "internet slang" data was the demerit we noticed across maximum research works. Today, majority of the users make use of the internet jargon, which is not included in majority of the datasets available. Another drawback noticed was that they restrict the data to twitter, other social media platforms like the upcoming one, Koo should be taken into consideration to widen the user demographic.

Current work in deep learning and word-vectorization points towards vector based methods being used in the future. Word embedding techniques such as Word2Vec, Glove, Google's BERT capture semantic information much better than traditional approaches and can identify topics better. Hence future work in that direction can make use of such approaches.

References

[1] K. Ahmed, N. E. Tazi, and A. H. Hossny, "Sentiment analysis over social networks: An overview," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 2174–2179.

[2] K. Morabia, N. L. Bhanu Murthy, A. Malapati, and S. Samant, "SEDTWik: Segmentation-based event detection from tweets using Wikipedia," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 77–85. [Online]. Available: https://www.aclweb.org/anthology/N19-3011

[3] C. Li, A. Sun, and A. Datta, "Twevent: Segment-based event detection from tweets," ser. CIKM '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 155–164. [Online]. Available: https://doi.org/10.1145/2396761.2396785

[4] A. Sechelea, T. D. Huu, E. Zimos, and N. Deligiannis, "Twitter data clustering and visualization," in *2016 23rd International Conference on Telecommunications (ICT)*, 2016, pp. 1–5.

[5] E. S. Negara, D. Triadi, and R. Andryani, "Topic modelling twitter data with latent dirichlet allocation method," in *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 2019, pp. 386–390.

[6] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[7] C. Lauschke and E. Ntoutsi, "Monitoring user evolution in twitter," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012, pp. 972–977.

[8] Y. Zhang, W. Mao, and J. Lin, "Modeling topic evolution in social media short texts," in *2017 IEEE International Conference on Big Knowledge (ICBK)*, 2017, pp. 315–319.

[9] W. Liang, Z. Lu, Q. Jin, Y. Xiong, and M. Wu, "Modeling of research topic evolution associated with social networks of researchers," in *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, 2015, pp. 1169–1174.

[10] Z. Wang, L. Shou, K. Chen, G. Chen, and S. Mehrotra, "On summarization and timeline generation for evolutionary tweet streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1301–1315, 2015.

[11] R. Liu, Q. Li, C. Wang, L. Wang, D. D. Zeng, and H. Ma, "Mining phase evolution for hot topics: A case study from multiple social media platforms," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017, pp. 2814–2819.

[12] L. Dey, A. Khurdiya, and D. Mahajan, "Topical evolution and regional affinity of tweets," in *2013 International Symposium on Computational and Business Intelligence*, 2013, pp. 297–300.

[13] M. Jain, S. Rajyalakshmi, R. M. Tripathy, and A. Bagchi, "Temporal analysis of user behavior and topic evolution on twitter," in *Big Data Analytics*, V. Bhatnagar and S. Srinivasa, Eds. Cham: Springer International Publishing, 2013, pp. 22–36.

[14] S. Malik, A. Smith, T. Hawes, P. Papadatos, J. Li, C. Dunne, and B. Shneiderman, "Topicflow: Visualizing topic alignment of twitter data over time," in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, 2013, pp. 720–726.

[15] S. M. Naim, A. P. Boedihardjo, and M. S. Hossain, "A scalable model for tracking topical evolution in large document collections," in *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 726–735.

[16] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 233–246, 2016.

[17] M. Takaffoli, F. Sangi, J. Fagnan, and O. R. Zäıane, "Community evolution mining in dynamic social networks," *Procedia - Social and Behavioral Sciences*, vol. 22, pp. 49–58, 2011, dynamics of Social Networks. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877042811013784

[18] M. Abulaish and M. Fazil, "Modeling topic evolution in twitter: An embedding-based approach," *IEEE Access*, vol. 6, pp. 64 847–64 857, 2018.

[19] L. Dey, A. Khurdiya, and D. Mahajan, "Topical evolution and regional affinity of tweets," in *2013 International Symposium on Computational and Business Intelligence*, 2013, pp. 297–300.

[20] I. Arpaci, S. Alshehabi, M. , Alshehabi, M. Khasawneh, I. Mahariq, T. Abdeljawad, and A. E. Hassanien, "Analysis of twitter data using evolutionary clustering during the covid-19 pandemic," *Computers, Materials & Continua*, vol. 65, no. 1, pp. 193–204, 2020. [Online]. Available: http://www.techscience.com/cmc/v65n1/39561

[21] G. Cai, L. Peng, and Y. Wang, "Topic detection and evolution analysis on microblog," in *Intelligent Information Processing VII*, Z. Shi, Z. Wu, D. Leake, and U. Sattler, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 67–77.