Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

TOPIC MODELING USING DOCUMENT PIVOT APPROACH WITH FIXED WINDOW

MUSHTAQ AHMED
Department of Computer Science, the Islamia University of Bahawalpur Pakistan.
Email: malikmushee7@gmail.com

RAFAQAT KAZMI*
Department of Software Engineering, the Islamia University of Bahawalpur, Pakistan.
Corresponding Author Email: rafaqat.kazmi@iub.edu.pk

NADEEM SARWAR
Department of Computer Science, Bahira University Lahore Campus Pakistan.
Email: nsarwar.bulc@bahira.edu.pk

MUHAMMAD MURAD KHAN
Government College University Faisalabad. Email: muhammadmurad@gcuf.edu.pk

ALI SAMAD TAUNI
Department of Data Science. The Islamia University of Bahawalpur. Email: ali.samad@iub.edu.pk.

SUNNIA IKRAM
Department of Software Engineering, the Islamia University of Bahawalpur.
Email: sunnia.ikram@iub.edu.pk

AMNA IKRAM
Department of Computer Science & IT, Government Sadiq College Women University, Bahawalpur, Pakistan. Email: amnaikram@gscwu.edu.pk

Abstract

Twitter is very popular social media and micro blogging platform. The tweet of twitter are limited to 140 character and people are using this platform for getting information and keep themselves up to date in term of latest information and events. Topic modeling means to find the Topic Headings from the textual dataset. This research performed topic modeling using the document pivot approach on Twitter data, as this method was not suitable for generating the topic from Twitter stream data. This research implemented the document pivot using a Fixed Window approach by using Term Frequency-Inverse Document Frequency (TF-IDF) method in such a way on the twitter stream data that it will generate results in terms of topics high accuracy volume and solve the issue of sort of noise, rapidly changing contents and a very short size of the document. This research also conducted a cross-verification using the probabilistic topic modeling method's algorithm that is known as Latent Dirichlet Allocation (LDA). The result of both the Latent Dirichlet Allocation approach and the Term Frequency-Inverse Document Frequency algorithm compared. This comparative analysis shows that Topic Modeling using Document Pivot Approach with Fixed Window Operation shows more accuracy results and perform well processing on short text micro blogging data, which was limitation and gaps in previous researches. This accuracy results of this study are higher than currently using algorithm of Latent Dirichlet Allocation.

Index Terms: Topic Modelling, Document Pivot Approach, Fixed Window, LDA (Latent Dirichlet Allocation), TF-IDF (Term Frequency-Inverse Document Frequency).

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

1) INTRODUCTION

To handle emergency situation, event reporting is essential to respond quickly and minimize damage. Terrorist attack, public marches, and bushfire are some of the examples of emergency situation which need police, ambulances, and firefighters to be ready and cope with the situation as soon as possible to save the live sported on social media such as “twitter” by several individuals. In order to confidently decide that the event took place, the process of event detection requires addressing the keywords linked with each event along with the minimum count of each word. In this study we purpose a new spike matching approach to classify the keywords and to determine the possibility of each event, probability classification method is used by counting the volume of each keyword expression.

In this modern era, events are identified and investigated thoroughly by using a platform known as “social media”. The most predictive system in social media is word-based content such as likes, shares, comments, and retweet about that event [1]. The text is used either by monitoring keyword temporal trends, sorting words into items, or analyzing sentiment scores and polarity characteristics. Social media specially in tweeter, the main challenge in keyword-based models [2] is to decide if words, particularly when people use terms in a non-standard way should be used in first place or not.

The document pivot approach is a traditional topic modeling method that is very less applicable for a twitter which is known as a micro blogging platform [3]. The document pivot approach cannot be simply implemented on the twitter data because of short of noise, rapidly changing contents and a very short size of the document. As topic modeling uses a data cluster document based on the similarity approach but the tweet data is not enough to build a complete cluster based on the similarity of words. Topic modeling [4]is suitable for large blogging system and comments data Like Facebook and Reddit.

For Topic modeling on twitter data [5], the other approaches are suitable like the feature pivot method and probabilistic topic modeling method [6]. In this research, we will use the document pivot approach on Twitter to generate topics from the data of Twitter tweets. The main gap of existing research is that topic modeling using a document don't gives any suitable result in term of finding models from the twitter data.

The main objective of this research is to perform the topic modeling on Twitter data by using document pivot approach with Term Frequency-Inverse Document Frequency method. The research objective that is to be fulfilled during this study is as follows.

To implement the Document Pivot Approach on Tweets of Twitter data and reduce the noise in rapidly changing contents with very short size of the document. 2. To perform topic modeling on Twitter data by using Term Frequency-Inverse Document Frequency method and generate high accuracy results. 3. To implement Fixed widow constraint on the data set and perform cross-verification using the probabilistic topic modeling method's algorithm that is known as LDA (Latent Dirichlet Allocation) and compare the result of both LDA approach and the TF-IDF algorithm.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

The implementation of document pivot approach on Twitter data is very challenging because the document pivot approach is a traditional topic modeling method and is very incompatible on the Twitter data stream because of noise in rapidly changing contents and a very short size of the document. As topic modeling uses a data cluster document based on the similarity approach but the tweet data is not enough to build a complete cluster based on the similarity of words. During this research, the following research questions will arrive and this study will try to answer this question.

1) How the document pivot approach will be implemented on Twitter data and reduce the noise in rapidly changing short contents of tweets?
2) How the complete cluster based on the similarity of words will be built and how the Fixed Window constraint be used with TF-IDF (Term Frequency-Inverse Document Frequency)?
3) How the higher accuracy of Topic modelling results will be achieved and how cross-validation performed among probabilistic topic modelling method LDA and TF-IDF method?

The main contribution of this research is that, this research will perform topic modeling using the document pivot approach on Twitter data, as this method currently is not suitable for generating the topic from Twitter stream data. This research will implement the document pivot using a Fixed Window approach by using TF-IDF (Term Frequency-Inverse Document Frequency) method in such a way on the twitter stream data that it will generate results in terms of topics high accuracy volume and solve the issue of sort of noise, rapidly changing contents and a very short size of the document.

This research will also conduct a cross-verification using the probabilistic topic modeling method's algorithm that is known as LDA (Latent Dirichlet Allocation). The result of both the LDA approach and the TF-IDF algorithm will be compared and finally, that result will show that our research will have a greater number of accuracies in terms of finding the Topics data from the twitter stream.

2) RELATED WORK

The definition of an event varies from one discipline to another. A complete and organized concept of an event is unsatisfactory. Event is defined by Stanford Philosophy Encyclopedia as something that happens in real, such as childbirth, deaths, thunderclaps, rituals, and storms. "The desire to plan and implement actions and to bring in changes in the environment" is identified as an event in the real- world from a social prospect.

In the late 90s research has been done and an event is described as "something that occurs with effects at a certain time and location." That result plus effects boost users to conduct such acts on social media while circulating information related to activities in online social networks. A parallel concept is delivered that an event is "something vital that occurs at a particular time and location." Users on social media share material about real happenings online. Further, Lee and Weng explain an event sense while "a series of

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

posts swapping within a short time about the same subject and language." Another study generalized the notion as "an event that is a real-world happening with a time and a stream of Twitter messages addressing the case within the time." Moreover, another researcher Panagiotou et al. explains this "in the setting of online social networks, event e is something that generates a massive number of online social network acts" [7].

2.1 Real-Life Event

As we have studied from literature, that "an event is a method of responding to a measurable incident that consists of or put some effect on a group of people in a social network at a particular time and place." The event existence and time interval explained that how much this event being influenced globally by social media. Global events remain on the social media for a longer time as compared to the local event. The global event basically consists of many local events. For instance, a natural disaster that may affect worldwide includes global events along with many local events. Whenever an event occurs, it takes time to get a “hype”, maintain its importance for a specific period of time, and then decline from social media.

2.2 Virtual Event

Sometime there are some virtual events, which are not like real life events. These events can be described as topic of discussion between people who communicate and cooperate online without being physically present. Such events draw the attention of other users and not reflect the real life event such as # friendship, # meals, and # joyful. The other main topics of virtual events are sports, natural disasters, politics, showbiz and many more [7].

2.3 Event Detection

Internet networking provides an opportunity for the individuals to post the stories about different events that they witness in their real life and spread it worldwide. These real-life events have been detected continuously known as event detection by examining the event content that is reported publicly. Events have been detected by using different event detection techniques according to their nature.

2.3.1 Specified Event Detection Technique:

The processing of data covering associated information (such as position, keywords, interval, and user) to achieve event information is termed as Specified Event Detection (SED) while a social event is then known or scheduled. It processes already explained information and features which are purposed to be appeared in the data to reflect an event. This already explained data represents as a basic building block to the actual event [7].

2.3.2 Unspecified Event Detection Technique:

In the absence of prior information, Unspecified Event Detection (UED) methods was used to detect a new or novol event. This method uses the fixed time period in which the

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

tweeter stream has been raised to identify the keywords and impressions that can highlight the event [7].

2.3.3 New Event detection technique (NED)

New event detection technique is used for continuous monitoring the live data [8]. The live data streaming on social media detect the hot trending story [9]. This method is mainly about explaining the data in short interval that can identify a notable shift in knowledge utilizing a semi supervised or unsupervised learning method [7].

2.3.4 Retrospective event detection (RED)

This method can detect the historically major events [10]. The historical data is arranged in groups and classified to explain the major event that happened in the past [7].

2.3.5 Network event detection technique (NED)

Network event is described here as a surge of new material linked to a specific subject or incidence. These events can be viewed as a popular trend in a related manner. Network Event Detection (NED) believes that the presence of a trending topic or media story is defined by changes in the value of three types of named objects: persons, places and organizations. In a stream of data such as articles or tweets, NED defines those entities and develop a series of networks having a specific entites of nodes and edges that represent the co-occurrence of it within a text. This method has two stages: 1. recognition of events 2. characterization / summarization of events [11].

2.4 Entity Detection

Provided a streamed document (articles, tweets), NED aims to notice significant events. To recognize three types of entities, we use a NER technique for each document: Persons, Organizations and Places. We decided to search papers with the “Stanford NER classifier” because it is equipped with CoNLL2 dataset of news wire papers from Reuters, which is suitable in our situation. We used a standard classifier for detecting designated entities in tweets. Tweets vary qualitatively to news article, they are a extreme of 140 characters long, they contain several confusing word (typs, phrases, lower-case word), and frequently absence of adequate meaning to evaluate the type of an individual. We used the Ritter et al. classifier for object identification in tweets, which are according to the researchers, outperforms the Stanford NER framework for this role.

We increment entity recognition with a variant spellings process of news articles. That substitute single terms such as first or last names or with the complete name of the individual for private entities. Each record is searched for individual entities. It is swapped by the latest corresponding multiple word Individual entity phrase which was noticed when a particular word Person entity is originate. The single-word name is maintained where no match is made. We disambiguate for Position and Entity organizations by applying abbreviations and references to individually collated exceptions dictionaries [11].

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

2.5 Graph Time Series Analysis

In the news event identification handling the large weighted degree, the total of the weights of entity domain nearby to the domain edges, as a predictor. The dataset with All the News is divided from 14/6/2016 to 26/6/2016 into 13 one-day blocks. The final data collection of the FA Cup is discretized into one-minute sections covering the game time. From all blocks in each, a series of information graphs are generated [12]. Dataset, dataset. For all entity nodes from the graph list, a weighted degree time series is then generated. We aim to recognize events based on differences in network configuration, exposed in entity time-series, by tracking the development of weighted degrees for all nodes across time. We measure the first variations in the time series to detect major shifts. Delete averages, then measure the average and standard deviation of X blocks from a moving block. An entity node whose weighted level crosses a threshold of standard deviations beyond the rolling mean is then identified as a 'peaking entity' [11].

2.6 Summarizing the Detected Events

Event classification starts by gathering the collection of information for every section that addresses the top entities after recognizing the time frame. Because several events can co-occur, in order to collect reliable data, we need to distinguish related events. This extract all information which is not important to topic we also noticed that are trending. Doing so offers a significant quantitative and qualitative increase in the efficiency of our strategy.

We, therefore, select noun phrases from the study, collection, or construct the next generation of graphs acknowledged as Key-Graphs to help differentiate specific events in a similar time block.

Placed, the same graphs continued with nouns and noun-phrases are Key-Graphs [13], designed only of records concerning a topmost person. We use the Top-Mine algorithm for extracting individual noun sentences.

Next, to classify candidate events as groups in the Key-Graphs, we used the "Louvain group identification “algorithm, which can be applied on non-directed weighted diagrams. In each group, the individuals and noun phrases make a bag of terms summarizing the identified case. We get a detailed description of the case by ordering this bag of terms using the weighted degree of all node.

3) MATERIALS AND METHODS

A Twitter data source is used to identify trending topics, where implemented a nonparametric Bayesian model, called "Hierarchical Dirichlet Processes (HDP)"[14]. For this approach, by applying the Hierarchical Dirichlet Processes model on tweets, a subject vector is originally identified such that a distribution of subjects is determined for each tweet by exploiting the vector of subjects. In a distribution of subjects, the subject with the greatest likelihood is called a leading subject. Clusters are clustered into tweets with common trending issues. The Gibbs sampling algorithm is used by the authors in their concept. A dictionary of words produced using "Yet another Better Ontology (YAGO)" is

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

used by the framework to integrate semantic knowledge that can help cluster trending topics.

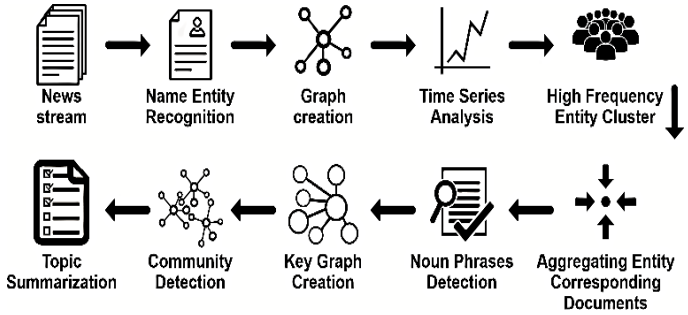

Fig. 1. Diagram of the Main steps of the method

3.1 Tweet Preprocessing

The workflow begins by gathering published tweets that can be set as an input parameter (for example, 1 hour) within a specified time frame. Then, to clean the input tweets, we add typical text preprocessing routines. We use a unique twitter tokenizer that handles user comments, hash tags, and emotions as single tokens.

Next, we remove retweets, IP addresses, characters, and emoticons that are not ASCII. It is worth noting that we do not conduct stop word elimination at this point because stop words (e.g. United States of America) may be part of NEs. As for hash tags, it is possible to describe a series of hand-crafted rules to break them into words.

Using SymSpell1, which matches incorrect tokens with WordNet sync sets, we also attempt to fix mispelled words [15].

3.2 Named Entity Recognition and Linking

To retrieve NE mentions in the tweets, we use "NERD-ML" where a Twitter-specific "Named Entity Recognizer (NER)" tool is a proved that NER is the best working tools on Twitter data [16]. In addition, “Named Entity Recognition and Linking (NERD-ML)”not just identifies the general categories of the article (people, association, and place), but also aims to connect any word mentioned during outer information bases. Then are connected in the ontology of the NERD with semantic classes [17].

3.3 Graph Generation

Previous research to model relationships between words in text using graph-based approaches treated all nodes in the text document as words and locating their place in text to establish [18]. These techniques may generate a complicated diagram, which is requiring high computing expenses to be prepared.

In this work, we presume that its meaning is described by the words, NE is mentioning surrounding to a tweet [19]. To build event graphs, we rely on the NE sense, constructed as follows:

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

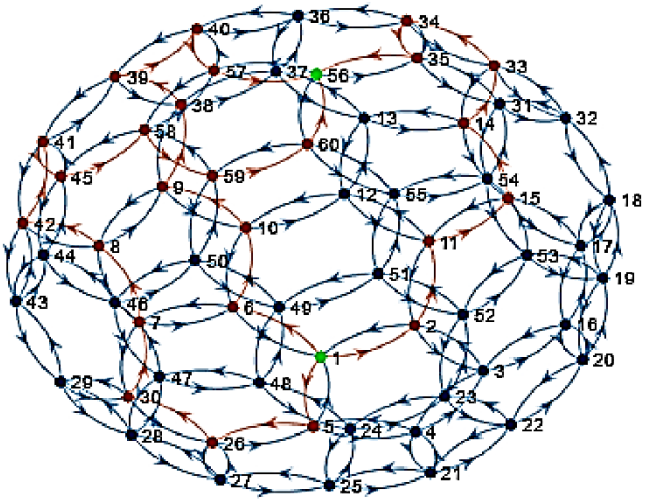

Fig. 2. Cluster Graph Nodes and Edges

Nodes: In a tweet, we analyze k and NE terms that lead and conduct their nodes as mention, where k is less than 1 is considered the number of words was forming the NE interest encompassing a NE.

Edges: Nodes are joined by an edge in which the graph occurs in the sense of a NE.

Weight: for understanding the NE, the edge weight is the sign of co-occurrences among words. Also, each edge maintains the tweet list from which the feature is recognized by cooperation.

Let (v, $\mathcal{E}$) is a graph with a vertices collection v and edges $\mathcal{E}$. Such as $\mathcal{E} \subset v \times v$. Let ($vi$) be the collection of vertices that pointing out to vi for every $vi \in$, and Out (vi) be the collection of vertices that pointing out to vi.

Let $\mathcal{E}i$= (vj,) be an edge connecting the vj to vk node, we'll define! In tweets published within a time window, as the weight of $\mathcal{E}i$, that is expressed relations among vj to vk are recognized.

3.4 Graph Partitioning

The relationships among terms in the NE and graph of the event are designed for the model. The partition of the graph into sub-graphs, which will be recognized as "event candidates".

A few common keywords are typically exchanged by tweets referring to the same events [20]. This concept is illustrated in case graphs by closer relations among nodes that are connected to the identical occurrence.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

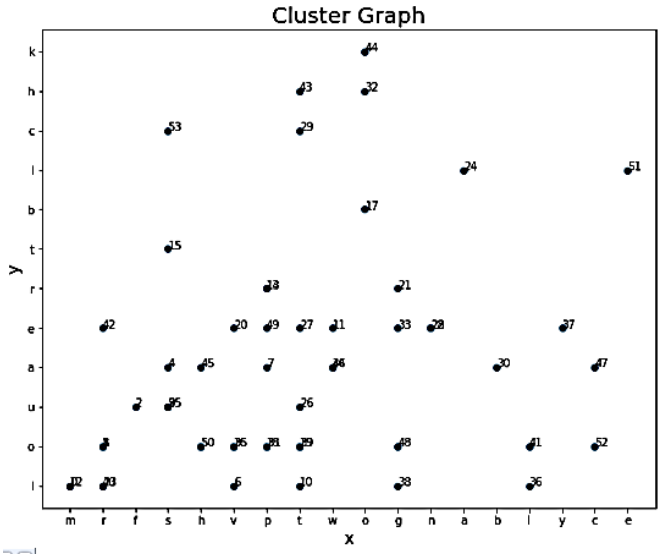

Fig. 3. Cluster Graph Each Cluster Graph Node Value shows number of Edges

The edge weight connecting the similar terms of the tweet to the same events which are higher, the edge weight among nodes connecting the similar terms of the tweet to distinct events. The partitioning of the graph is eliminated when huge graph G is divided into subgraphs.

In a closely linked graph $\mathcal{G}$, let $\mathcal{E}$ = (υ1, 1), (υ2, $\mathcal{W}$2)... (υn,) be a set of pairs of vertices. λ As the edge smallest number whose expulsion from $\mathcal{G}$ will distribute $\mathcal{G}$ into similar subgraphs. Likewise, the connectivity of edge λ ($\mathcal{G}$) of $\mathcal{G}$ of the set an edge $\mathcal{S}\subset\mathcal{E}$ is defined as the smallest cardinality in such a way that $\mathcal{G}$ − $\mathcal{S}$ − is no extended strongly related [17].

3.5 Event Detection

The events in which subgraphs are multiple are unbounded to each other. Each subgraph is stored independently in the event detection sub-module.

Andersen et al. discovered that a powerful distribution of a graph can be accomplished by dividing higher vertices from low rated ones if the nodes in the graph have dissimilar values,

In the event graph the rate vertices use a PageRank-like algorithm as follows:

$$S(V_i) = ((1-d) + d\sum_{v_j} \frac{w_{ij}}{\sum_{v_k \in Out(v_k)} w_{jk}} S(V_j))\epsilon_i \quad (I)$$

Where! $wij$is the edge weight that connects Vi to Vj, d is a factor that is dumping normally, and a node i is a parameter of penalization. T The penalization parameter is assumed to be a uniform distribution in previous approaches, instead of specifying its TF-IDF performance. In various time windows, the rate of the nodes can be trending words by skewed, due to unnecessary features in tweets. Thus, the TF-IDF rate to reduce the influence of trending words in a selection of tweet.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

We allocate an initial valueτ= 1 /n to every vertex in the graph before measuring the score with equation 1, where all number of nodes in the graph is n. When the concentration is accomplished by each estimate emphasizes by each node to the wanted degree. By evaluating the separation between the modern and earlier emphasis, that the concentration of degree can be node reached. It based on a calculated parameter, we begin by breaking the vertex set into high ranked and low ranked vertices.

Next, we prepare the high-ranked vertices subset that determines the highest weighted antecedents and replacements for every applicant as keywords for event applicants. After excluding the keywords from the list of the edges, we additionally examine the keywords of separated nodes for the event applicant if it converts detached.

We split the keywords relevant to an occurrence into the following subsets depending on the semantic class given by the NER method, what (the form about occurrence), where (the place of event where they occur), who (the entity or individual associated). As for the year, the oldest tweets describing the occurrence are chosen. We process the event candidates. Next, duplicate case candidates are combined that they share well-known words and the place where the candidates in the time window are supposed. Thus, a novel case is generated by the fusion of the two case candidates' words and individuals.

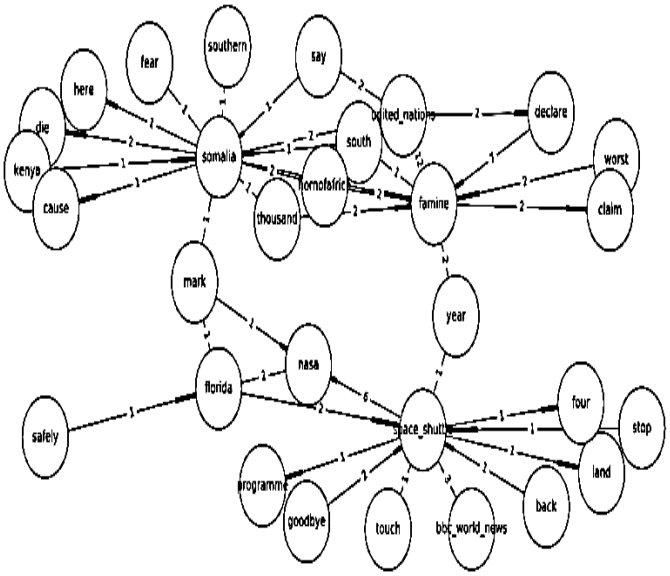

Fig. 4. Graph from a sample of famine events in Somalia tweets

We split the keywords relevant to an occurrence into the following subsets depending on the semantic class given by the NER method, what (the form about occurrence), where (the place of event where they occur), who (the entity or individual associated). As for the year, the oldest tweets describing the occurrence are chosen. We process the event candidates. Next, duplicate case candidates are combined that they share well-known words and the place where the candidates in the time window are supposed. Thus, a novel case is generated by the fusion of the two case candidates' words and individuals.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

An occurrence is assumed to be true if there is at least one NE involved and if the specified number of tweets given as an input parameter exists [9].

Fig 3.1: Graph from a sample of famine events in Somalia tweets and the space shuttle to Mars produced on day '2011-07-07' [14].

3.6 Event Merging

They include the related keywords, entities (e.g. individual, entity, and place) while an interval of k days, where the parameter of input is k, and we consider events in separate time-windows as duplicates. When a new occurrence is observed as a repeat, it is combined with the prior one that has been identified [15].

4) EXPERIMENTAL SETUP

Data Set Used

In this research, the twitter dataset that is named as Super Tuesday 2012 is used. This data set contains microblogging data with a short context value. By performing operations on short context values this research will evaluate our proposed model that either the previous graph of research is fulfilled or not like short of noise, rapidly changing contents and a very short size of the document.

4.1 Data Set Info

The data set named as Super Tuesday 2012 contains microblogging data with the short code text value. This data set contains the following columns.

1. Published date
2. ID
3. Tweet

4.2 Preprocessing on Dataset

Before implement, any research work and Research model, the data set is a preprocessed in which data is normalized and any independent values are identified. The attributes on which our research work is not dependent is removed. Only the dependent attributes keep remains in the data set so that data set values should be in a clean format and the size of the dataset is reduced by eliminating the non-dependent attributes from the dataset.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

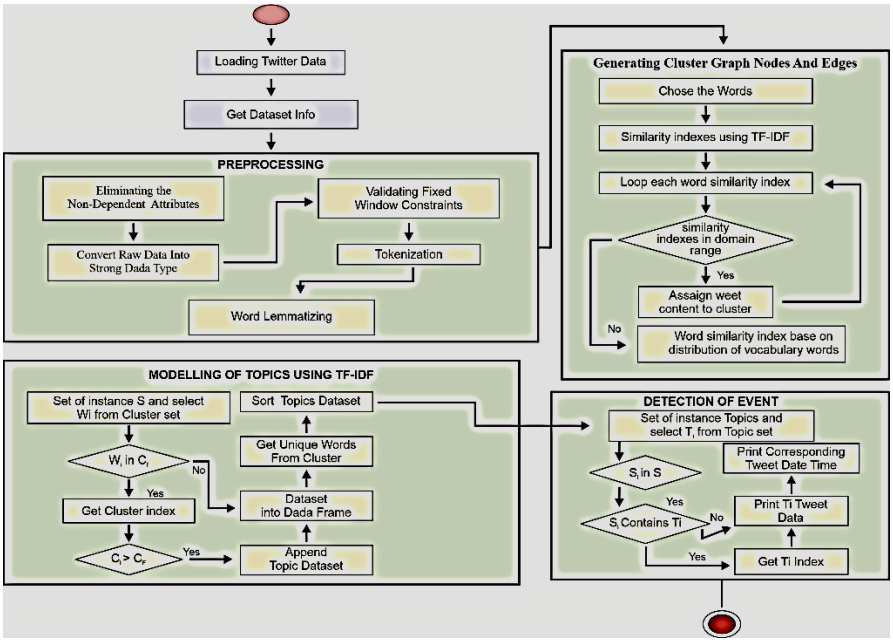

Fig. 5. The overall process of exploring topic modeling using document pivot approach with fixed window

4.2.1 Eliminating the Non-Dependent Attributes

This research is based on a fixed window operation on data set by using the document pivot approach so that for fix window the time attribute is the dependent attributes for defining the fixed window. The tweet attribute is used for performing the cluster range and other research work. So these above-listed attributes are dependent attributes for our research work. The tweet ID is not used in our search model so this attribute is non-dependent attributes. Therefore this attribute will be eliminated from the dataset.

4.2.2 Convert String Date Value to Date Time

For performing the operation by using a fixed window approach the time attribute is very important for this research. The data set contains the time attribute in a string format for performing any operation on this data set. The string value of the date is converted into the data type of date time.

4.2.3 Validating Fixed window Constraints

This research is based on fixed window operation using document prevent approach so this study will validate the fixed window constraints on the data set. In the current study, the fixed Windows depends on the date.

The limitation of the fixed window is to process the tweets of single date data set. So, this study will first convert the Date Time into the only date and find the unique date values if a unique date value is greater than 1 then it will exit and stop the further process.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

4.2.4 Tokenization

The first step is to create tokenization in the data set. For tokenization, the text is populated into the sentences and then these sentences will be converted or split into the words.

For implementing the similarity system in the text. The text is converted into the lower case so that there will not be an issue of matching the words. The punctuation words are removed from the sentences so that the words will only be used for finding the similarity. All words that are less than the three characters are removed from the tokens.

4.2.5 Word Lemmatizing

To find the similarity index from the sentences the words are converted from the third form of verbs to in the first form of the verb. The third person words are converted into first persons. So the benefit of this lemmatizing the words is that any sentence use in future or past tenses are converted into the present tense.

4.3 Generating Cluster Graphs Nodes and Edges

In the document pivot approach using fixed window operation, the clustering is done on the tweet stream data that divides data set into groups depending on the similarity of the words within a group. A large amount of data is divided into a smaller number of groups. Each data cluster has similar patterns of the words in each tweet. The remaining tweets without similarities will be grouped in another cluster.

The clustering technique is an unsupervised machine learning technique in which this method does not require any training for the learning phase.

Procedure.

Input: Number of Tweets M

1. Number of Words t
2. B vocabulary matrix
3. Cluster distribution based on similarity of words

Steps,

1. Choose the words
2. find similarity indexes using term frequency–inverse document frequency TF-IDF
3. loop each word similarity index
    a. condition if similarity indexes in domain range
        i. assign tweet content to cluster
4. the selection of word similarity index depends on distribution of vocabulary words

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

Table 1: The Algorithm 1

Algorithm 1
Input: S (set of instanced), W0First word, C(Cluster Count), Cf Final Cluster, WIWord Similarity index using term frequency–inverse document frequency **Output:** Similarity Index based Clustering. 1. Loop on set of instance S and select Wi from vocabulary dictionary. Computer Term frequency Similarity Index a. Loop Wi> Cf i. Select neighbor word last Inverse Frequency ii. Condition if Wi> Cf 1. Assign to Cluster 2. Break b. End loop 2. End loop

4.4 Modelling of Topics using TF-IDF

After generating the clustering graph, the next step is to generate the topic dataset using term frequency and Inverse document frequency method for this purpose the study used the corpora library of python with using gensim.

The number of topics that can be found through this fixed window size operation is equal to 15 in each iteration.

Table 2: The Algorithm 2

Algorithm 2
Input: S (set of Cluster), Topic (Set of topic),C0First word, C(Cluster Count), Cf Final Cluster, WIWord Similarity index using term frequency–inverse document frequency **Output:** Topic Dataset. 1. Loop on set of instance S and select Wi from Cluster set. a. Loop Wiin Cf i. Get Cluster index (Convert index to normalize number by multiplying 100 and subtract from 100) ii. Condition if Ci> Cf 1. Append Topic Dataset b. End loop 2. Convert Dataset into data frame 3. Get only unique Words from Cluster 4. End loop 5. Sort Topic dataset according to normalized index value

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

5) RESULTS AND DISCUSSION

Implementing the TF-IDF approach, the output of the topic dataset which was the required output by using the document pivot approach with fixed window operation. The accuracy level of this data set is given in table 3.

Table 3: Accuracy of dataset

Topics	Accuracy	Topics	Accuracy
Newtgingrich	92.9	Vote	95.8
Romney	94.8	Barackobam	96.2
Santorum	94.8	Newt	79.8
Mittromney	95.0	Newtgingrich	87.2
Newt	95.6	Supertuesday	87.8
Today	94.3	Gingrich	91.8
Elect	94.8	Support	93.3
Vote	95.0	Ricksantorum	97.9
Ric	98.1		

The figure 6 is showing the Modelling Records with Accuracy using Term Frequency-Inverse Document Frequency approach in which x-axis is represent cluster of words and y-axis is represent the occurrence of words.

5.1 Cross Validation by Implementing LDA

This research is also conducting a cross-verification using the probabilistic topic modeling method's algorithm that is known as LDA (Latent Dirichlet Allocation). The result of both the LDA approach and the TF-IDF algorithm will be compared. After implementing the LDA algorithm the following topics are given as output.

Table 3: Accuracy of dataset with LDA

Topics	Accuracy	Topics	Accuracy
Santorum	97.0	Newt	79.8
Ronpau	97.7	newtgingrich	87.2
Today	84.8	supertuesday	87.8
Conserv	87.3	Gingrich	91.8
Live	90.4	Support	93.3
newtgingrich	92.9	Santorum	94.8
Romney	94.8	Vote barackobam	95.8 96.2
mittromney	95.0		

5.2 Comparison of Accuracy Level

This research conducted a cross-verification using the probabilistic topic modeling method's algorithm that is known as LDA (Latent Dirichlet Allocation). The result of both the LDA approach and the TF-IDF algorithm is finally compared.

This result shows that Topic Modelling using Document Pivot Approach with Fixed Window Operation shows more accuracy results and perform well processing on short text micro blogging data, which was limitation and gaps in previous researches as shown in figure 6.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

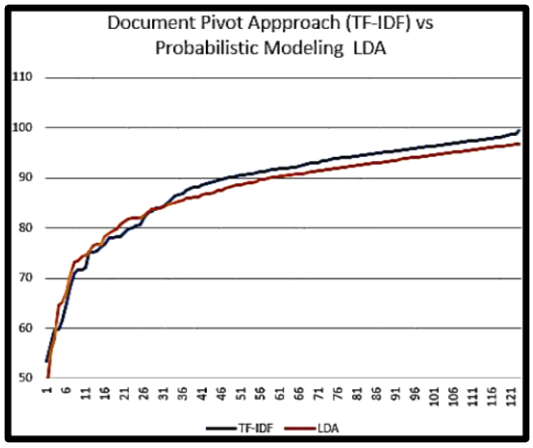

Fig. 6. Result Comparison for Topic Modelling with Document Pivot Approach using Fixed Window and (TF-IDF) with Probabilistic Modeling (LDA)

The result comparison of both algorithm the LDA algorithm and this study implemented Algorithm of TF-IDF.

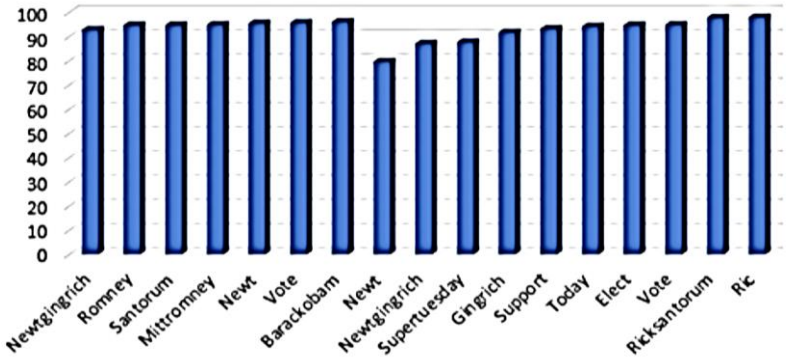

Fig. 7. Table and Graphs of Topic Modelling Records with Accuracy using TF-IDF

The graph represented in figure 7 is showing the Modelling Records with Accuracy using Term Frequency-Inverse Document Frequency approach in which x-axis is represent cluster of words and y-axis is represent the occurrence of words.

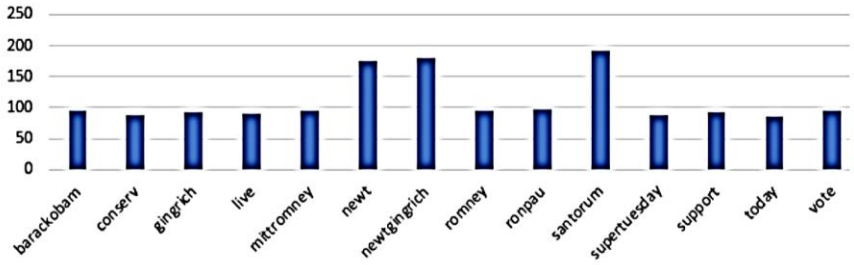

Fig. 8. Table and Graphs of Topic Modelling Records with Accuracy using LDA

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

The graph in figure 8 is showing the Modelling Records with Accuracy using Latent Dirichlet Allocation approach in which x-axis is represent cluster of words and y-axis is represent the occurrence of words.

6) CONCLUSIONS

Topic modelling means to find the Topic Headings from the textual dataset. This research performed topic modeling using the document pivot approach on Twitter data, as this method was not suitable for generating the topic from Twitter stream data. This research implemented the document pivot using a Fixed Window approach by using TF-IDF (Term Frequency-Inverse Document Frequency) method in such a way on the twitter stream data that it will generate results in terms of topics high accuracy volume and solve the issue of sort of noise, rapidly changing contents and a very short size of the document. This research also conducted a cross-verification using the probabilistic topic modeling method's algorithm that is known as LDA (Latent Dirichlet Allocation). The result of both the LDA approach and the TF-IDF algorithm will compared. This result shows that Topic Modelling using Document Pivot Approach with Fixed Window Operation shows more accuracy results and perform well processing on short text micro blogging data, which was limitation and gaps in previous research. This accuracy results of this study are higher than currently using algorithm of LDA.

7) FUTURE WORK

By using the term frequency-inverse document frequency method for generating the topic modeling using the document pivot approach on Twitter data, this research generated a higher accuracy than the previously implemented LDA method. In the future, this search will continue to implement the topic modeling for generating the trends in social media and find the geographically based topics to find the impact of the economy on the basis of the occurrence of events. In most trading companies the trading values fluctuate with the events that occurred in that geographical area. So, by using our research on social media in the future we can predict the trading moves of the economy for that geographical area.

References

[1] E. Kross, P. Verduyn, G. Sheppes, C. K. Costello, J. Jonides, and O. Ybarra, "Social media and well-being: Pitfalls, progress, and next steps," Trends in Cognitive Sciences, vol. 25, pp. 55-66, 2021.

[2] A. Anwar, H. Ilyas, U. Yaqub, and S. Zaman, "Analyzing qanon on twitter in context of us elections 2020: Analysis of user messages and profiles using vader and bert topic modeling," in DG. O2021: The 22nd Annual International Conference on Digital Government Research, 2021, pp. 82-88.

[3] M. Kolmykova, N. Gavrilovskaya, M. Barsukova, and D. Kozlovskaya, "Use of Microblogging, Social Networking, and Short Messages in E-learning for Information Culture Building," International Journal of Emerging Technologies in Learning, vol. 16, 2021.

[4] M. Mujahid, E. Lee, F. Rustam, P. B. Washington, S. Ullah, A. A. Reshi, et al., "Sentiment analysis and topic modeling on tweets about online education during COVID-19," Applied Sciences, vol. 11, p. 8438, 2021.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 02: 2023
DOI10.17605/OSF.IO/S6N58

[5] A. P. Rodrigues and N. N. Chiplunkar, "A new big data approach for topic classification and sentiment analysis of Twitter data," Evolutionary Intelligence, vol. 15, pp. 877-887, 2022.

[6] M. R. Bhat, M. A. Kundroo, T. A. Tarray, and B. Agarwal, "Deep LDA: A new way to topic model," Journal of Information and Optimization Sciences, vol. 41, pp. 823-834, 2020.

[7] Z. Saeed, R. A. Abbasi, M. I. Razzak, and G. Xu, "Event detection in Twitter stream using weighted dynamic heartbeat graph approach [application notes]," IEEE Computational Intelligence Magazine, vol. 14, pp. 29-38, 2019.

[8] M. Hasan, M. A. Orgun, and R. Schwitter, "A survey on real-time event detection from the Twitter data stream," Journal of Information Science, vol. 44, pp. 443-463, 2018.

[9] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," in Proceedings of the 24th ACM international on conference on information and knowledge management, 2015, pp. 1867-1870.

[10] M. Fedoryszak, B. Frederick, V. Rajaram, and C. Zhong, "Real-time event detection on social data streams," in Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2774-2782.

[11] I. Moutidis and H. T. Williams, "Utilizing complex networks for event detection in heterogeneous high-volume news streams," in International Conference on Complex Networks and Their Applications, 2019, pp. 659-672.

[12] R. Koncel-Kedziorski, D. Bekal, Y. Luan, M. Lapata, and H. Hajishirzi, "Text generation from knowledge graphs with graph transformers," arXiv preprint arXiv:1904.02342, 2019.

[13] C. Rühlemann, "Visual Linguistics with R," Amsterdam: University of Freiburg, 2020.

[14] A. Madani, O. Boussaid, and D. E. Zegour, "Real-time trending topics detection and description from Twitter content," Social Network Analysis and Mining, vol. 5, pp. 1-13, 2015.

[15] C. Fellbaum, "A semantic network of english: the mother of all Word Nets," in Euro Word Net: A multilingual database with lexical semantic networks, ed: Springer, 1998, pp. 137-148.

[16] L. Derczynski, D. Maynard, G. Rizzo, M. Van Erp, G. Gorrell, R. Troncy, et al., "Analysis of named entity recognition and linking for tweets," Information Processing & Management, vol. 51, pp. 32-49, 2015.

[17] A. Edouard, E. Cabrio, S. Tonelli, and N. Le Thanh, "Graph-based event extraction from twitter," in RANLP17-Recent advances in natural language processing, 2017.

[18] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5410-5419.

[19] S. Nguyen, B. Ngo, C. Vo, and T. Cao, "Hot topic detection on twitter data streams with incremental clustering using named entities and central centroids," in 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF), 2019, pp. 1-6.

[20] A. J. McMinn and J. M. Jose, "Real-time entity-based event detection for twitter," in International conference of the cross-language evaluation forum for European languages, 2015, pp. 65-77.