VIVA-Tech International Journal for Research and Innovation
ISSN(Online): 2581-7280

Volume 1, Issue 1 (2018)
Article No. 4
PP 1-6

VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

Tweet Summarization and Segmentation: A Survey

Siddhi Naik[1], Swati Mamidipelli[1], Shruti Lade[1], Ashwini Save[2]
[1](B.E. Computer Engineering, VIVA Institute of Technology/ Mumbai University, India)
[2](Assistant Prof. Computer Engineering, VIVA Institute of Technology/ Mumbai University, India)

Abstract : *The use of social media is increasing day by day. It has become an important medium for getting information about current happenings around the world. Among various social media platforms, with millions of users, twitter is one of the most prominent social networking site. Over the years sentiment analysis is being performed on twitter to understand what tweets that are posted mean. The purpose of this paper is to survey various tweet segmentation and summarization techniques and the importance of Particle Swarm Optimization (PSO) algorithm for tweet summarization [1][2].*

Keywords – *Tweet-Summarization, segmentation, Particle swarm optimization.*

1. INTRODUCTION

Social media is a platform/technology that can be used for creating and sharing various information, which can be accessed from any corner of the world. It is one of the best feasible way through which marketing can be done, current affairs can be known also it can be used to know the perspective of different people about an ongoing issue around the globe. Over the years social networking sites have evolved, one of which is Twitter. Use of twitter has grown enormously over the decade. It has been efficiently serving the users for interaction and information sharing.

Numeral tweets are posted on twitter on daily basis, which are analysed by sentiment analysis to draw summary of opinions expressed by the user and categorize them to know the views of the user on an issue. The problem with sentiment analysis is that there are millions of users with alterations in opinion to test [4]. There are also practical tests to sentiment analysis. It may happen that someone tweets something that may not be relevant to others, in this case summarization comes into picture and plays a significant role. For this various technique have been developed by researchers over the years [3], some of which are discussed in module II.

For providing user with better results tweets are not only summarized but they are first segmented. Segmentation helps in conserving semantic meaning of tweet. Tweets are segmented (i.e. divided into parts) using various techniques are also discussed in module II.

There are various clustering algorithms available. One such optimal clustering algorithm is the Particle Swarm Optimization (PSO) Algorithm. The paper further surveys how this paper can also be used for clustering and how it is better than other clustering algorithms in module III. Module IV is the Analysis table for the Literature survey done. The final module V concludes the survey with conclusion followed by the references.

2. SUMMARIZATION AND SEGMENTATION

Quite a lot of Tweets are posted on twitter on daily basis. Many of the times it happens that a person may tweet something irrelevant, and because of this understanding what the user intends to say becomes a very difficult. For this purpose, summarization and segmentation are used. The purpose of these are to draw a summary from the tweet posted by user and mine his/her opinion. Segmentation is dividing of tweets into meaningful sentences and gaining their meanings [5] whereas, summarization clusters a group of similar tweets and draws a summary from this groups and provides it to the user.

Various summarization and segmentation techniques have been developed by researchers over the years few of which perform both segmentation and summarization together on a tweet[7]. Some either summarization or segmentation. The various techniques for summarization include various clustering algorithms like K-means, ACO, etc. some methods are graph formation for clustering of similar tweets [3]. Different

Volume 1, Issue 1 (2018)
Article No. 4
PP 1-6

segmentation methodologies use frameworks like the Hybrid-seg frame work for segmentation. Tweets are segmented taking several factors into consideration like the parts of speech, linguistics, etc. these segments are then searched locally and/or globally, for which they make use of dictionaries like the word-net. Following are some papers which deliberate some of these techniques [4][5][6].

2.1 A Graph Based Clustering Technique for Tweet Summarization [1].

Twitter is a prominent person to person communication site utilized by millions to share data. A user can discover tweets identified with any occasion however it ends up plainly troublesome for the user to peruse every one of the tweets. This paper consequently proposes a tweet summarization system which will decrease users efforts to a degree. In this system, comparative tweets identified with an occasion are considered, developed on which is a graph for that occasion. This diagram comprises of a bunch of comparable tweets. Numerous such groups are formed lastly a tweet from each group is incorporated into the summary. It utilizes WordNet and community identification in graphs. In WordNet things, verbs, descriptors, qualifiers and equivalent words are all around distinguished and assembled together as synsets. Eg. Automobile, car frame a synset. A graph of comparative tweets is framed in which, every node is a tweet and edges represent similarity among the tweets

.

2.2 A Tweet Summarization Method Based on a Keyword Graph [3].

Tremendous number of posts are posted on microblogging destinations, for example, twitter and consequently it is a critical wellspring of data. The paper suggests a keyword based tweet summarization framework. It utilizes a graph in which every node is a keyword and co-occurances are the edges.

A set of keywords are attained which occur recurrently in tweets. Number of co-occurrences of any two words in the keyword set is calculated. A graph is created such that

Nodes → Keywords

Co-occurrences → Edges

For the graph clustering process, the proposed system makes use of K-Clique algorithm. A K-Clique is a sub-graph which has K-nodes connected to others.

Thus, the tweets which contain all the keywords of a clique will have a higher probability to be like each other.

.2.3 Graph Summarization for Hashtag Recommendation [2].

The paper proposes a graph based approach for finding similar hashtags. A heterogeneous graph is used which contains users, tweets and hashtags as its nodes. This heterogeneous graph is then converted into a homogeneous graph consisting of hashtags only. The hashtags are ranked using random walk and context similarity measure. For each hashtag, the hashtag content similarity concatenates all tweets and weighs the words. The homogeneous graph model's hashtag co-occurrences in tweets and edges represent frequency of co-occurrences.

Thus, the graph is summarized to a homogeneous that only includes relations between hashtag. To rank the vertices in the hashtag graph Random Walk with Restart is been used.

2.4 Abstractive tweet stream summarization using Natural language processing [11].

The paper discusses generation of an abstract summery of live tweets stream using Natural Language Processing. The live tweets are pre-prepared and subject based tweet cluster vectors are framed by incremental clustering. A transformation matrix is developed utilizing the tcv's centroids and the live tweets, at that point an abstractive synopsis is produced utilizing the proposed summarization strategy. After collecting the tweets, they are pre-processed and keyword occurrences are calculated in tweets. The tweets are then clustered using incremental tweet clustering algorithm followed by construction of transition matrix which records the co-occurrences of keywords.

2.5 A Survey on Online Tweet segmentation for Linguistic Features [6].

Tweet summarization is analysed to provide information to the users who are unaware of the topic which is currently been posted by other users, so that they can also continue to comment on that topic. Once the tweet data is taken from the twitter data source, the words in the tweets are analysed by pre-processing the data. Stop Word Removal method is used to discard the words which does not provide precise meaning to the

sentence. The tweets are partitioned into different segments. Clustering of tweets is done by using Parts of Speech (POS) taggers and then the tweets are put under specific class labels. Joint based Named Entity Recognition is used which is more accurate than work based Named Entity Recognition. Natural Language Processing (NLP) is a tool which detects the linguistic features such as complexity, expressivity, etc by mapping the input text.

Joint Named Entity Recognition is computational in linking the recognized entities.

Drawback in this system is that each opinion cannot provide complete information about the topic tweeted. Also, execution time required for analysing each opinion will be more.

2.6 Tweet Segmentation and its application to Named Entity Recognition. [5].

In the proposed system, Tweets are separated into germane segments by conserving the information and are easily mined. To obtain high quality tweet segmentation HybridSeg framework is used. HybridSeg framework divides tweets in batch mode. Tweets from a battered twitter stream are grouped into batches using a fixed interval of time. Each batch of tweets are then segmented by HybridSeg. To demonstrate tweet segmentation benefits Named Entity Recognition (NER) is used. There are two methods of NER, namely:

1. Random-Walk (RW) based NER.
2. Part of Speech (POS) based NER.

A segment graph is built based on random walk method. The node in this graph is a identified segment. An edge exists between two nodes if they co-occur in some tweets.

In part of speech based NER method noun phrases are considered as named entity using segment rather than word as a unit.

2.7 Multi-Criterion real time summarization based upon adaptive threshold [4].

Monitoring the tweets that describes the event properly or referring to an entity is time consuming and may provide irrelevant information. For this, summarization of tweets is done which highlights relevant and redundant information related to event as soon as it occurs. The decision of selecting and rejecting the tweets is done when tweets are made available. Instead of using predefined threshold for decision making, threshold is estimated as soon as new tweet arrives in real time. Novelty detection and in formativeness are the methods used for decision making. The main purpose of increasing the amount of information with respect to what the user has already known is achieved.

A real-time summarization is provided instead of categorizing sub-events.

Drawback in this system is that, if the novelty scores vary while new tweets arrives than the predefined threshold could be inappropriate.

2.8 Summarization of tweets and named entity recognition from tweet segmentation [7].

The system proposed in this paper divides tweets into segments using the Hybrid-Seg framework. The framework divides the tweets taking numerous factors into consideration like the parts of speech. The tweets that are segmented are then checked locally as well as globally. For searching globally various dictionaries like word-net. This increases the value of segmented tweet and helps find proper meaning of the tweets. The segmented tweets are then summarized using various clustering algorithms. As the tweets get segmented and then summarized this improves the quality of summary provided to the user. The problem with this methodology is that it uses various clustering algorithms like K-means which is fast and easy to implement but is not accurate enough.

3. PARTICLE SWARM OPTIMIZATION (PSO) ALGORITHM

Particle Swarm Optimization Algorithm or popularly known as the PSO algorithm is an optimization algorithm used to give optimal results [8]. This algorithm is inspired from the behaviour of bird flocking or fish schooling. This algorithm mainly consists of two entities swarm and particle, particle is an individual entity knowing its best position/value and swarm is group of particles knowing position/value of best particle. As this is an optimal algorithm, it can be used for clustering tweets in the process of tweet summarization [10]. Following are few papers that compare PSO against other clustering algorithms, proving PSO as a better algorithm for clustering over others. Also, how some algorithms can be improved by combining with PSO

3.1 Real time clustering of tweets using adaptive PSO and map reduce [10]

A large amount of data is generated by social media/ social- networking sites such as Twitter, Facebook, etc. these types of data have complex structure which causes difficulty in capturing, storing, analysing, clustering visualization of data. For clustering, this type of data different clustering algorithms is used. Distinct algorithms like k-means cluster data. But there is a need to use algorithm that is able to cluster data in less amount of time. For this purpose PSO is preferable. The paper implements PSO for clustering data in twitter using Hadoop Map reduce framework. The outcome illustrates that PSO performs better than K-means. The results show that the accuracy of K-means is 62%, whereas that of PSO comes out to be 90.6%.

3.2 Comparative Analysis of clustering by using Optimization Algorithm [8]

Data mining has become very prominent for the extraction and manipulation several data and establishing patterns in gigantic and chaotic datasets. Clustering of data is an important technique in organizing data. In this paper various optimal clustering algorithms like the Genetic Algorithm, Ant Colony Optimization and Particle Swarm Optimization are compared to find the better of out of these algorithms. For comparing these algorithms different metrics like weighted arithmetic mean, standard deviation is used. The results prove the accuracy of PSO over GA, ACO. The parameters are implemented in MATLAB 7.11.0.

3.3 A clustering algorithm based on integration of k-means and PSO [9]

Clustering data remains one of the major problems faced in data mining that has attracted a lot of attention. One of the eminent algorithm in this field is K-Means clustering that has been effectually applied to numerous problems. But this method has its own downsides, such as the necessity of the competence of this method to initialization of cluster centres. To improve the quality of K-Means, hybridization of this algorithm with other methods is suggested. Particle Swarm Optimization (PSO) is one of the optimization algorithms that has been united with K-Means. Both algorithms the K-means and PSO are combined using their métiers. Most of the methods introduced in the context of clustering, that hybridized K-Means and PSO, used them sequentially, but this paper, applies them entwined. The results show the ability of this approach in clustering analysis. From the results, amended k-means using PSO is better than ordinary k-means algorithm, the results also prove that the algorithm turned out to be better as both k-means and PSO were united using their strengths.

3.4 The improved k-means clustering using proposed extended PSO [12]

Clustering plays a vital role in various fields of research and science. A popular algorithm for this is K-means. The algorithm has the strengths like high speed and an easy implementation but, this algorithm lacks local optimality and for this purpose, an extended version of PSO is suggested in this paper. The production of initial population is based on chaos trail whereas it is randomized trial. This suggested algorithm is further evaluated against genetic algorithms, hybrid k-means using UCI datasets and result show that the proposed algorithm is better than others. The results after comparison prove suggested algorithm is better than others. The experimental results are performed using 3benchmark datasets.

4. ANALYSIS TABLE

The following table gives the analysis of techniques and methods discussed in this paper on tweet segmentation and summarization and the particle swarm optimization (PSO) algorithm.

Sr. No.	Title Of Paper	Technique	Database Used	Accuracy/Efficiency
1.	Graph based clustering technique for Tweet Summarization [1].	WordNet, Clustering, Info-Map, Sum-basic Algorithm.	Twitter API	Proposed system is better than Sum-basic algorithm.
2.	Tweet Segmentation and its Application to Named Entity Recognition. [5].	Random Walk POS, Tweet Segmentation: Hybrid-Seg.	Twitter API	Local linguistic features are more reliable than term dependency in guiding segmentation process.
3.	Real time clustering of tweets using	K-Means, PSO, Hadoop and Map	Twitter API	Accuracy of K-Means is 62% and PSO is 90.6%

	adaptive PSO and Map Reduce [10].	Reduce Framework.		
4.	Multi-criterion real time Summarization based upon adaptive threshold [4].	Tweet Filtration, Novelty detection.	Twitter API	NA
5.	A Clustering Algorithm based on Integration of K-means and PSO [9].	K-Means, PSO.	UCI Machine Learning Repository (Benchmark).	Improves the speed of finding result :52.38% in normalized dataset and 52.30% in original dataset compared to K-Means.
6.	Tweet Summarization based on a Keyword graph [3].	K-Clique Algorithm for graph Clustering.	Twitter dataset from spinn3r	Less important words are removed and strong words are considered which makes the graph method more efficient.
7.	The Improved K-Means Algorithm Using Proposed Improved PSO [12].	K-Means PSO	3-Benchmark Datasets	proves that accuracy of K-means increases when fused with PSO.
8.	Summarization of tweet and Named Entity Recognition from Tweet Segmentation [7].	Hybrid-Seg Framework, Clustering for Summarization.	Twitter API	NA
9.	Comparative Analysis of clustering by using Optimization Algorithm [8].	Genetic Algorithm, PSO, Ant Colony.	UCI repository of Machine Learning Databases.	Accuracy of PSO is more than GA and ACO.
10.	A Survey on Online Tweet Segmentation for Linguistic Features [6].	Stop word removal method, Parts of Speech (POS) taggers.	Twitter API	POS makes the user understand the tweets more easily.
11.	Abstractive tweet stream summarization using Natural Language Processing [11].	Clustering of tweets, Transition matrix generation, calculating tf-idf matrix.	Twitter API	NA
12.	Graph Summarization for Hashtag Recommendation [2].	Construction of Heterogeneous graphs (Tweet User), Convert to Homogeneous graph (Nodes-Hashtag) .	Twitter API	NA

5. CONCLUSION

Sentiment analysis is a branch of data mining that deals with opinions, expressions and decision making. There are many systems that perform summarization on twitter to gain the different opinions expressed by user via tweets. The different summarization and segmentation technique, as discussed in this paper, make use of various clustering algorithms for summarization. Which are accurate, but lag some parameters in comparison to PSO. Hence, the user may not get a proper summary for a desired tweet. PSO is optimization algorithm which provides an optimal solution. If researchers try using PSO instead of other clustering algorithms, then it is expected to give more accurate results in comparison to other clustering algorithms.

Volume 1, Issue 1 (2018)
Article No. 4
PP 1-6

REFERENCES

[1] S. Dutta, "A graph based clustering technique for tweet summarization."Reliability, Infocom Technologies and Optimization *(ICRITO)(Trends and Future Directions), 2015 4th International Conference on. IEEE, 2015*.

[2] M. Al-Dhelaan and H. Alhawasi. "Graph Summarization for Hashtag Recoendation.", *Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on. IEEE, 2015*.

[3] Tae-Yeon Kim, "A tweet summarization method based on a keyword graph." *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication. ACM, 2014*.

[4] A. Chellal, B. Mohand and B. Dousset. "Multi-criterion real time tweet summarization based upon adaptive threshold." *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on. IEEE, 2016*.

[5] J. Weng, "Tweet Segmentation and its Application to Named Entity Recognition." (2015): 1-15.

[6] R. P. Narmadha and G. G. Sreeja. "A survey on online tweet segmentation for linguistic features." *Computer Communication and Informatics (ICCCI), 2016 International Conference on. IEEE, 2016.*

[7] C. Chavan and R. Suryawanshi. "Summarization of tweets and Named Entity Recognition from tweet segmentation." *Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on. IEEE, 2016*.

[8] P. Kataria, R. Navpreet and Rahul Sharma. "Comparative Analysis of Clustering by using Optimization Algorithms." *International Journal of Computer Science and Information Technologies 5.2 (2014): 1076-1081*.

[9] H. A. Atabay, M. J. Sheikhzadeh, and M. Torshizi. "A clustering algorithm based on integration of K-Means and PSO." *Swarm Intelligence and Evolutionary Computation (CSIEC), 2016 1st Conference on. IEEE, 2016*.

[10] A. P. Chunne, C. Uddagiri, and C. Malhotra. "Real time clustering of tweets using adaptive PSO technique and MapReduce." *Communication Technologies (GCCT), 2015 Global Conference on. IEEE, 2015, p. 26*.

[11] S. R. Annamalai and R. R. Thirumalai, "Abstractive tweet stream summarization using natural language processing." *International journal of advances in cloud computing and computer science 2 (2016)*.

[12] https://www.google.co.in/search?q=opinion+mining+and+data+mining&oq=opinion+mi ning+and+da&gs_l=psy, last accessed on 19/09/2017.

[13] https://en.wikipedia.org/wiki/Data_mining, last accessed on 19/09/2017.

[14] https://en.wikipedia.org/wiki/Sentiment_analysis, last accessed on 19/09/2017.

[15] M. Lashkari and M. H. Moattar. "The improved K-means clustering algorithm using the proposed extended PSO algorithm." *Technology, Communication and Knowledge (ICTCK), 2015 International Congress on. IEEE, 2015*.