TWEET SUMMARIZATION: A NEW APPROACH

[1] Sunil Katkar, [2]Maitreyee Phadke, [3]Anagha Pasalkar, [4]Vrishali Patil
[1] Assistant Professor, [2] Student, [3] Student, [4] Student
[1] BE Computer Engineering,
[1] Vidyavardhini's College of Engineering and Technology, Vasai West, India

Abstract: The utilization of online networking is expanding step by step. It has turned into an imperative mechanism for getting data about current happenings around the globe. Among different online networking stages, with millions of clients, twitter is a standout amongst the most prominent social networking site. Throughout the years sentiment analysis is being performed on twitter to comprehend what tweets that are posted mean. The motivation behind this paper is to overview different tweet division and synopsis strategies and the significance of Particle Swarm Optimization (PSO) algorithm for tweet summarization [1][2].

Keywords: Tweet Summary, Segmentation, Particle Swarm Optimization

1. INTRODUCTION

Web based life is a stage/innovation that can be utilized for making and sharing different data, which can be gotten to from any side of the world. It is a standout amongst the best doable route through which advertising should be possible, current issues can be known additionally it tends to be utilized to know the point of view of various individuals about a progressing issue far and wide. Throughout the years social networking sites have advanced, one of which is twitter. Utilization of twitter has become immense through the decade. It has been proficiently serving the clients for collaboration what's more, data sharing.

Numeral tweets are posted on twitter on everyday basis, which are analyzed by sentiment analysis to draw outline of feelings communicated by 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 additionally reasonable tests to sentiment analysis. It might happen that somebody tweets something that may not be applicable to other people, for this situation summary comes into picture and assumes a huge job. For this different methods have been created by specialists throughout the years [3], some of which are talked about in later module 2.

For providing users with better outcomes tweets are outlined as well as they are first segmented. Segmentation helps in monitoring semantic significance of tweet. Tweets are divided (i.e isolated into parts) utilizing different strategies are likewise talked about in module 2.

There are various clustering algorithms available. One such ideal clustering algorithm is the Particle Swarm Optimization (PSO) Algorithm. This paper further reviews how PSO can be utilized for clustering. It also depicts how it is superior to other clustering algorithms in module 3. Module 4 is the Examination table for the literature overview done. The last module V finishes up the review with end pursued by the references.

2. RESEARCH METHODOLOGY

A considerable amount of Tweets are posted on twitter on everyday schedule. On a large number of the occasions it happens that an individual may tweet something unessential, and due to this, understanding what the user intends to state becomes very difficult. For this reason, summarization and segmentation are utilized. The reasons for these are to draw an outline from the tweets posted by the user and mime his/her feelings. Segmentation is division of tweets into significant sentences and picking up their implications [5] while, summarization clusters a group of comparable tweets and draws a rundown from these groups and gives it to the user.

Various summarization and segmentation methods have been developed by scientists over the years few of which perform both segmentation and summarization together on a tweet [7] or some either of the one. The different methods for summarization incorporate various clustering algorithms like K-means, ACO, and so forth. A few of these strategies are also used for graph formations of a cluster for similar tweets [3]. Different segmentation methodologies use frameworks like the Hybrid-seg framework for segmentation. Tweets are fragmented thinking about a few components like the grammatical features, etymology, and so forth. These portions are then looked for locally as well as globally, for which they make utilization of lexicons like the word-net. Following are a few papers which deliberate a portion of these systems [4][5][6].

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

Twitter is a prominent person to person communication site used by millions to share information. A user can find tweets related to any event; however it winds up evidently troublesome for the user to examine all of the tweets. This paper

subsequently proposes a tweet synopsis framework which will diminish user endeavours to some extent. In this framework, similar tweets related to an event are considered, created on which is a chart for that event. This outline involves a bundle of similar tweets. Various such gatherings are framed in conclusion and a tweet from each gathering is consolidated into the summary. It uses WordNet and network ID in charts. In WordNet things, action words, descriptors, qualifiers and equal words are all around recognized and gathered together as synsets. For example: Automobile, car outline a synset. A diagram of near tweets is encircled in which, each hub is a tweet and edges speak to likeness among the tweets.

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

Huge amount of posts are posted on micro blogging sites, for instance, twitter and subsequently it is a critical wellspring of information. The paper recommends a keyword based tweet summarization framework. It uses a graph in which each hub is a keyword and co-occurrences are the edges.
A lot of catchphrases are achieved which happen intermittently in tweets. Number of co-events of any two words in the keyword set is determined. 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 Hash tag Recommendation [2].

The paper proposes a graph based methodology for finding comparative hash tags. A heterogeneous graph is utilized which contains users, tweets and hash tags as its nodes. This heterogeneous graph is then changed over into a homogeneous diagram comprising of hash tags as it were. The hash tags are positioned utilizing arbitrary walk and setting closeness measure. For each hash tag, the hash tag content likeness links all tweets and gauges the words. The homogeneous diagram model's hash tag co-events in tweets and edges speak to recurrence of co occurrences.

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

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

The paper examines generation of a unique summary of live tweet stream utilizing Natural Language processing. The live tweets are pre-prepared and subject based tweet group vectors are encircled by steady clustering. A transformation matrix is created using the tcv's ,centroids and the live tweets, by then an abstractive summation is delivered using the proposed synopsis methodology. Subsequent to gathering the tweets, they are pre-prepared and keyword events are determined in tweets. The tweets are then clustered using incremental tweet clustering algorithm pursued by construction of transition matrix which records the co occurrences of keywords etc.

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

Tweet summarization is examined to give data to the users who are unaware of the subject which is as of now been posted by different users, with the goal that they can likewise keep on remarking on that point. Once the tweet information is taken from the twitter information source, the words in the tweets are dissected by pre-processing the data. Stop Word Removal strategy is utilized to dispose of the words which does not give exact importance to the sentence. The tweets are divided into various portions. Grouping of tweets is finished by utilizing Parts of Speech (POS) taggers and after that the tweets are put under explicit class marks. 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.
Downside in this framework is that every assessment can't give total data about the point tweeted. Additionally, execution time required for analysing every feeling will be more.

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

In the proposed framework, Tweets are isolated into pertinent sections by rationing the data and are effectively mined. To get excellent tweet division HybridSeg system is utilized. HybridSeg system isolates tweets in cluster mode. Tweets from

a battered twitter stream are assembled into groups utilizing a fixed interval of time. Each batch of tweets is then divided by HybridSeg. To exhibit 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.

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 framework proposed in this paper separates tweets into fragments utilizing the Hybrid-Seg framework. The structure separates the tweets thinking about various variables like the grammatical forms. The tweets that are fragmented are then checked locally as well as globally. For seeking comprehensively different word references like word-net are used. This expands the estimation of sectioned tweet and helps find appropriate importance of the tweets. The sectioned tweets are then abridged utilizing different clustering algorithms. As the tweets get portioned and at that point abridged this improves the nature of summary given to the user. The issue with this approach is that it utilizes different clustering algorithms like K-means which is quick and simple to actualize yet isn't sufficiently accurate.

3. PARTICLE SWARM OPTIMIZATION (PSO) ALGORITHM

Particle Swarm Optimization Algorithm or famously known as the PSO algorithm is an optimization algorithm used to give ideal outcomes [8]. This algorithm is propelled from the behaviour of bird flocking or fish schooling. This algorithm basically comprises of two elements swarm and particle, particle is an individual element knowing its best position/esteem and swarm is gathering of particles knowing position/estimation of best particle. As this is an optimal algorithm, it tends to be utilized for clustering tweets during the process of tweet summarization [10]. Following are few papers that think about PSO against other clustering algorithms, demonstrating PSO as a superior algorithm for grouping over others. Also, how a few algorithms can be improved by combining with PSO.

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

A lot of information is produced by social media/ social- networking sites for example, Twitter, Facebook, and so on these kinds of information have complex structure which causes trouble in capturing, storing, analysing, clustering visualization of information. For clustering, these sorts of information distinctive clustering algorithms are utilized. 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 reason PSO is best.
The paper executes PSO for clustering information in twitter utilizing Hadoop Map Reduce framework. The result delineates that PSO performs superior to Kmeans. The outcomes demonstrate that the precision of K-means is 62%, while that of PSO turns out to be 90.6%.

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

Data mining has turned out to be noticeable for the extraction and manipulation of several data and building up examples in gigantic and chaotic datasets. Clustering of information is a vital strategy in arranging information. In this paper different optimal clustering algorithms like the Genetic Algorithm, Ant Colony Optimization and Particle Swarm Optimization are contrasted to locate the better out of these algorithms. For looking at these calculations distinctive measurements like weighted arithmetic mean, standard deviation is utilized. The outcomes demonstrate the exactness of PSO over GA, ACO.

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 algorithms 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 have been united with K-Means. Both algorithms the K-means and PSO are combined using their metiers. 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 assumes a fundamental job in different fields of research and science. A prominent calculation for this is Kmeans. The algorithm has the qualities like fast and a simple usage be that as it may, this algorithm needs local optimality and for this reason, an all-encompassing adaptation of PSO is proposed in this paper. The creation of initial populace depends on chaos trail while it is randomized preliminary. This proposed algorithm is further assessed against genetic algorithms; hybrid k-means using UCI datasets and result demonstrate that the proposed algorithm is superior to others. The exploratory outcomes are performed utilizing 3 benchmark 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 the 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 adaptive PSO and Map Reduce [10].	K-Means, PSO, Hadoop and Map Reduce Framework.	Twitter API	Accuracy of K-Means is 62% and PSO is 90.6%
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 Kmeans 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 KMeans Algorithm Using Proposed	K-Means PSO	3-Benchmark Datasets	proves that accuracy of Kmeans increases when fused with PSO.

	Improved PSO [12].			
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-Hash tag).	Twitter API	NA

5. CONCLUSION

Sentiment analysis is a part of data mining that manages sentiments, articulations and choice making. There are numerous frameworks that perform synopsis on twitter to pick up the diverse sentiments communicated by user through tweets. The diverse synopsis and division procedure, as talked about in this paper, make utilization of different clustering algorithms for an outline. Which are precise; however slack a few parameters in correlation with PSO. Henceforth, the user may not get an appropriate summary for a desired tweet. PSO is optimization algorithm which gives an ideal solution most of the times. 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.

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 Recommendation.", 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]. 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.