IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Analysis Of Trending Twitter Topics Using a Multiview Clustering Approach

[1]Rutuja Joshi, [2]Neha Kothawade, [3]Prem Pagare, [4]Bhushan Rajput, [5]Jyoti R. Mankar

[1,2,3,4] Students, Department of Computer Engineering

[5]Assistant Prof, Department of Computer Engineering

K. K. Wagh Institute of Engineering Education and Research, Nashik, India

Abstract: Social Media has emerged as a popular medium for individuals to express themselves and share their ideas with others. The number of users across various social networks is growing at an exponential rate. Twitter has become a vital source of real-time information and communication, with a plethora of trending topics being discussed by users at any given moment. These topics cover a wide range of subjects, reflecting people's daily lives and interests. Therefore, it is essential to implement an effective methodology to identify short-term, high-intensity discussion topics. The proposed work utilizes a Multiview clustering approach, which has been shown to be highly effective in unfolding underlying patterns. The Multiview clustering approach used in the analysis of trending Twitter topics considers different views of the data, which include the text content of tweets, like count, retweet count, quote count, and reply count. This analysis can offer valuable insights into the latest trends and have practical applications, such as monitoring public opinion, recommending hot products, and detecting incidents. Additionally, artificially intelligent services like web search systems or recognition systems can also benefit from this analysis. The Twitter data is retrieved and a dataset is created for further analysis using Python's snscrape library. The aim of this work is to cluster topics based on multi-view data to identify the most discussed topics during a specific time period. Experiments on real datasets indicate that the Multiview clustering approach is more effective than single-view clustering in detecting trending topics with a reasonably close approximation to the expected outcome. Nonetheless, the proposed approach has some limitations and challenges, such as the requirement for sufficient and diverse data to ensure the accuracy of the clustering results.

Index Terms - Multi-View Clustering, Snscrape library, Feature Extraction, Dimensionality reduction, Keyword Extraction

I. INTRODUCTION

Clustering is one of the most critical unsupervised learning techniques, which has been widely applied for data analysis, such as social network analysis, gene expression analysis, heterogeneous data analysis, and market analysis. The goal of clustering is to partition a dataset into several groups such that data samples in the same group are more similar than those in different groups. Clustering plays an important role in mining the hidden patterns. The real-world data are always captured from multiple sources or represented by several distinct feature views. Different views of data describe different features of data. Clustering with complementary information provided by multiple views would obtain better results than clustering only on a single view.

Traditionally it was understood that one particular subset would be sufficient for data mining, and multiple views were often regarded as redundant. However, research has now illustrated that these multiple views are often complementary and help gain a better understanding of the data structure. Multiple view learning has two advantages: a better performance can be obtained by integrating the multiple views rather than a single view, and the accuracy of the knowledge produced can be cross-verified from multiple views. Hence there is a need for multi-view clustering to handle multi-view data.

The field of multi-view clustering, like many other multi-view, or multi-modal, machine learning fields rests on the idea that not only more data, but incorporation of different types of data, can lead to better outcomes. In the case of multi-view clustering, many techniques have been developed around Multiview data of images or genetic interactions. To be more specific, each view corresponds to one source of information. For example, web pages can be described by both the page-contents (one view) and the hyperlink information (another view). Besides, different facets of a datum can also be treated as different views. For instance, an image can be characterized by its shape, colour, and location.

A good example to understand the importance of MVC, or Multiview learning is “the blind men and the elephant” story where each blind man (a single view of the subject) may not acquire the true picture of the subject, thus only collecting Multiview data can recover the whole picture of the subject. Multi-view algorithms deal with each view of the data independently and then merge the solutions to obtain a complete, robust pattern which is superior compared to its single-view representation.

To outline two exemplary domains, observe that healthcare systems often capture the same disease condition using different medical sensors (e.g., EEG, fMRI, PET are different ways of capturing neurological information), and criminal records often represent the same crime using modalities such as textual narratives, CCTV footages, audio tapes and photographs has been widely used in image categorization and motion segmentation tasks. Several methods are used in MVC like Multiview K-means, Multiview spherical k-means, co-regularized multi view spherical clustering, Multiview spectral clustering. This paper focuses on Multi-view K-Means, Multi-view Spectral Clustering and Multi-view Hierarchical Clustering.

II. LITERATURE SURVEY

Following studies were conducted previously on various Multiview Clustering techniques by various researchers, as described below in detail.

In a study by Zeel Doshi, Subhash Nadkarni, Kushal Ajmeraand Neepa Shah, TweerAnalyzer is implemented [7], it is capable of finding out the top ten trending hashtags and users at any given point in time and plotting them against their frequency using a bar graph. The "TweerAnalyzer " tool for analysing Twitter trends is described. It was developed using basic Python and JavaScript programming technique. The top ten trending hashtags and users at any given time can be found using TweerAnalyzer, and their frequency can be plotted using a bar graph. Additionally, a map visualisation is shown, which extracts the user's location coordinates and uses popup markers to display the tweets on a global map. This tool's simplicity makes it ideal for use even by those without technical backgrounds. The model presented in this work can be enhanced to enhance the user experience, offer new features, and maximise computing power.

In a more recent study by the author Iain J. Cruickshank[6] proposed a new hybrid paradigm to clustering multi-view data. Tested the new hybrid paradigm techniques along with several state-of-the-art intermediate integration and late integration techniques using social-based, multi-view data. Focused multi-view clustering technique to group hashtags from the COVID-19 pandemic twitter data that was gathered. A social media innovation called hashtags enables users to find and take part in discussions about topics of interest thus, groups of Insight into social media users' current hot topics can frequently be gained through hashtags. The first-ever multi-view clustering of hashtags from twitter data gathered during and about the COVID-19 outbreak in this chapter. The outcomes of the multi-view clustering show that users' use of hashtags evolves over time and exhibits temporal trends. Additionally, The outcomes of the multi-view clustering show that users' use of hashtags over the length of a pandemic exhibit temporal patterns and changes. Additionally, the multi-view clusters provide clear insight into the topical focus areas for twitter users during the pandemic and how different themes have varied user bases.

Edi Irawan; Teddy Mantoro; Media Anugerah Ayu; M. Agni Catur Bhakti; I Komang Yogi Trisna Permana [3] proposed a study where Twitter users are grouped according to the terms, they use the most frequently to respond to a hot political topic. Here, the word trend or primary theme of the problem is provided and addressed, as well as a comparison between the hierarchical clustering and k-means clustering approaches is done. Lokesh Mandloi; Ruchi Patel [5], studied and compared various machine learning methods like, Support Vector Machine Classification, Naïve Bayes Classification, and Maximum Entropy Classification method. The accuracy and precision of this classification algorithm's analysis were studied by the author.

A single view of the sample is typically insufficiently thorough, but multiple views can more effectively represent the features of the samples by complementing one another. As a result, the topic of multi-view data clustering is receiving increased attention from researchers. For multi-view clustering, certain researchers investigate non-negative matrix factorization (NMF) based methods using subspace learning concepts. These techniques, however, frequently struggle to deal with data noise, and the majority of them are insufficiently reliable without taking the distribution of the original data into account. To address these issues, Liang Zhao, Xiao Wang, along with other authors [1] suggested a novel deep probability multi-view feature learning (DPMFL) approach in this study.

A study conducted by Hikmat Ullah Khan, Shumaila Nasir, Kishwar Nasim, Danial Shabbir and Ahsan Mahmood [4] suggested method identifies real-time Twitter trending topics and ranks the top phrases and hashtags. In addition to exploratory data analysis, the research paper investigates the Term Frequency-Inverse Document Frequency (TF-IDF), Combined Component Approach (CCA), and Biterm Topic Model (BTM) approaches for locating the topics and terms within given topics. The paper also discusses the motivation for trend prediction over social media

Yixiang Fang, Haijun Zhang, Yunming Ye and Xutao Li [9] have demonstrated MVTD and STVSM framework which is used for Multiview topic detection. It primarily focuses on three relations or views namely as semantic, social tags and temporal relations and generates k clusters. It performs (13.8%) percent better than single view clustering

Similarly, Cody Buntain and Jennifer [8] Golbeck proposed a method for automating fake news detection on Twitter by learning to predict accuracy assessments in two credibility-focused Twitter datasets: CREDBANK, a crowdsourced dataset of accuracy assessments for events in Twitter, and PHEME, a dataset of potential rumours in Twitter and journalistic assessments of their accuracies. The authors collected data from threads identified as containing fake news by human experts and extracted features to train a machine learning classifier. The results show that the classifier is able to accurately identify fake news threads with high precision and recall, and the approach generalizes well to new data. This work highlights the potential of using machine learning to combat the spread of misinformation on social media

III. RESEARCH METHODOLOGY

SYSTEM ARCHITECTURE:

The detailed system architecture is shown in figure below:

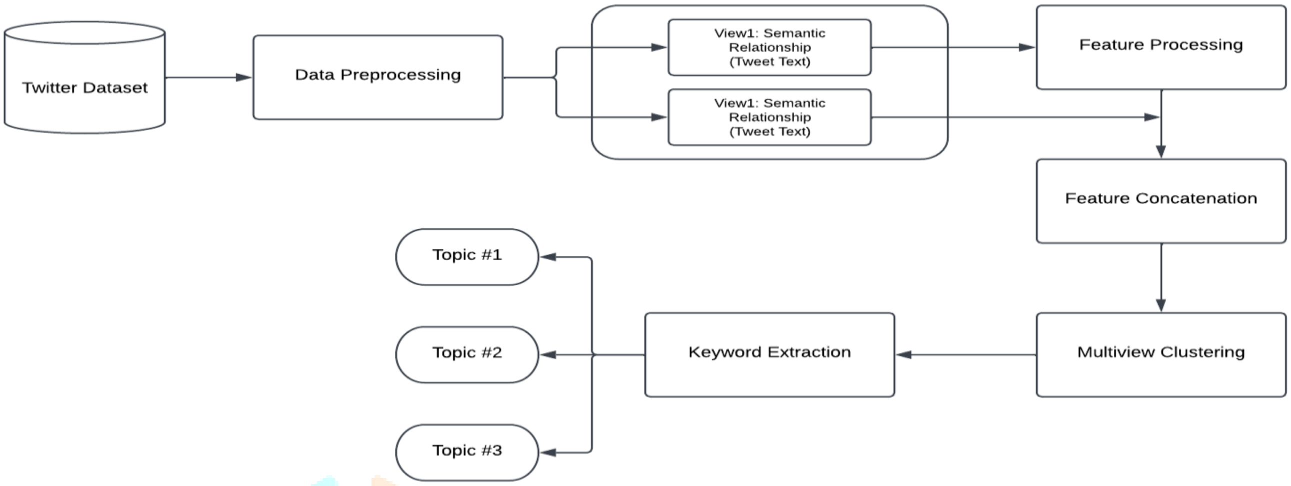

Fig. 1: System Architecture

The above architectural diagram represents the different steps involved in a proposed approach for Multiview clustering of Twitter data related to trending topics. Here is a brief explanation of each block in the diagram:

1. Data Collection: The first step involves collecting Twitter data related to trending topics using the web scraping technique with the help of a snscrape python library.
2. Data Pre-processing: The collected data is pre-processed by removing duplicates, handling missing values, removing the hashtag symbol, removing any non-ASCII character, and cleaning the text data by removing stop words, stemming, and performing other text normalization techniques.
3. Feature Processing: Feature processing involves a series of techniques used to prepare and manipulate data features for use in machine learning algorithms. Two important techniques in feature processing are feature extraction and dimensionality reduction.
    1) Feature Extraction: In this step, the tweet data is transformed into a numerical format by converting text data into a set of numeric features using Term frequency-inverse document frequency (TF-IDF) technique.
    2) Dimensionality reduction: To reduce the number of features in the data and make it more manageable for analysis, dimensionality reduction technique called Truncated SVD (Singular Value Decomposition) is used.
4. Feature Concatenation: This step involves integrating the extracted features from multiple data views to perform clustering.
5. Multiview Clustering: In this step, clustering algorithms such as k-means, hierarchical clustering are applied on the multiview data.
6. Keyword Extraction: The keyword extraction methods are applied to extract the representative keywords as topics from clusters.
7. Performance Evaluation: The performance of the proposed Multiview clustering approach is evaluated using metrics such as silhouette score, Calinski Harabasz index, Davies Bouldin index.
8. Visualization: In this step, the difference between single and Multiview clustering approaches is analyzed, and the results of the developed work are visualized using a scatter plot.

ALGORITHMS

1. Multiview K-means Clustering:

Multiview k-means clustering is a variant of the k-means clustering algorithm that is used to cluster data with multiple views or data sources. Each view represents a different set of features for the same set of data points. In the proposed work, the multiple views are combined and stored in a single variable, which is then used for clustering using the k-means algorithm.

The multiview k-means algorithm works by first determining the optimal number of clusters using some criteria, such as the elbow method or the silhouette score. In this work, the elbow method is used. Then, the k-means model is fitted to multiview data again with the optimal number of clusters. Finally, after predicting cluster labels for the multiview data using the fit_predict method, it displays the top clusters representing trending Twitter topics.

The advantage of multiview k-means clustering is that it can capture the complementary information contained in different views, which can lead to more accurate clustering results than using a single view. Additionally, it can be used in a wide range of applications, such as image classification, speech recognition, and bioinformatics.

2. Multiview Spectral Clustering:

Multiview spectral clustering is a clustering technique that can handle data with multiple views or modalities. In this approach, the data is represented as a similarity matrix, where each view corresponds to a different measure of similarity. The similarity matrices from each view are then combined into a single matrix, which is used as input to the spectral clustering algorithm. In the proposed work, the similarity matrix is computed using the Gaussian kernel for the combined data from all views.

It uses the eigenvectors of the similarity matrix to map the data points to a lower-dimensional space, where clustering is performed. The top clusters are then extracted, and keywords are extracted for each cluster to identify the most relevant tweets.

Spectral multiview clustering has been shown to be effective in cases where the different views are complementary and provide a more complete representation of the data. It has several advantages over single-view clustering. First, it can handle data with complex structures that may not be captured by a single view. Second, it can combine different types of information to improve clustering performance. Third, it can reduce the effects of noise and outliers by leveraging information from multiple perspectives.

3. Multiview Hierarchical Clustering

Multiview hierarchical clustering is a clustering technique that group`s data objects based on multiple views or perspectives of the same dataset. In other words, it takes into account the multiple representations of the same data to provide a more comprehensive and accurate clustering result.

The multiview hierarchical clustering algorithm builds a dendrogram by iteratively merging clusters based on the similarity between their views. It begins by treating each view as a separate dataset and performing hierarchical clustering on each view independently. Then, it combines the resulting clusters from different views based on a specified criterion, such as the average linkage or the complete linkage. In the proposed work, Ward linkage criteria are used to merge clusters.

The algorithm continues to merge clusters at different levels of the dendrogram until the desired number of clusters is reached. The resulting clusters are based on the collective information from all views, which can provide a more robust clustering result than using a single view.

IV. EXPERIMENTAL SETUP

DATASET:

The data in Twitter dataset which is retrieved using Python's snscrape library spans from January 1st to January 31st 2023. Data collection process involved gathering information from five distinct news channels, namely CNN News 18, NDTV, Times Now, ZoomTV, and Republic TV.

The dataset consists of 5000 data points stored as rows with 6 features in columns.

1. User: This feature refers to the username of the user who posted the tweet. .
2. Tweet: This feature represents the actual text of the tweet, limited to 280 characters.
3. Likecount: This feature represents the number of times the tweet has been liked by other users.
4. Retweetcount: This feature indicates the number of times the tweet has been retweeted by other users.
5. Quotecount: This feature represents the number of times the tweet has been quoted by other users.
6. Replycount: This feature indicates the number of times the tweet has been replied to by other users.

PERFORMANCE PARAMETERS:

1. Silhouette score:

The silhouette score is a metric used to evaluate the quality of clustering results. It measures how well each data point in a given cluster is separated from other clusters in the data set. It ranges from -1 to 1, with higher values indicating better-defined clusters.

The silhouette score for a data point i in a cluster is calculated by comparing the average distance between i and all other data points in the same cluster (a_i) to the average distance between i and all data points in the nearest neighbouring cluster (b_i). The silhouette score for the entire cluster is then calculated as the mean silhouette score for all data points in the cluster.

A high silhouette score means that the data point is well-matched to its own cluster and poorly matched to neighbouring clusters, indicating that the clustering is appropriate. Conversely, a low silhouette score means that the data point is poorly matched to its own cluster and may be more suited to a different cluster.

It can be used to compare the quality of clustering results generated by different clustering algorithms or with different parameter settings, and can be used to optimize clustering results by selecting the best clustering algorithm and parameter settings based on the highest silhouette score. The silhouette score for a data point *i* in a cluster can be calculated using the following formula:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

where:

- a(i) is the average distance between point i and all other points in the same cluster
- b(i) is the average distance between point i and all points in the nearest neighbouring cluster (i.e., the cluster with the next highest membership for point i)

2. Calinski Harabasz index (also known as the "variance ratio criterion"):

The Calinski-Harabasz index is a metric used to evaluate the quality of clustering results. It compares the variance between clusters to the variance within clusters. A higher Calinski-Harabasz index indicates better clustering results. It quantifies the ratio of between-cluster variation to within-cluster variance.

A high Calinski-Harabasz index indicates that the clusters are well-separated and distinct since the between-cluster variation is greater than the within-cluster variance. A low Calinski-Harabasz index, on the other hand, signifies that the within-cluster variation is equal to or smaller than the between-cluster variance, suggesting that the clusters may be ill-defined or overlapping. The Calinski-Harabasz index is calculated as follows:

$$\mathrm{CH} = \frac{(\boldsymbol{B} \,/\, (\boldsymbol{k} - \mathbf{1}))}{(\boldsymbol{W} \,/\, (\boldsymbol{n} - \boldsymbol{k}))}$$

where:

- CH is the Calinski-Harabasz index
- B is the between-cluster sum of squares, which measures the variance between clusters
- W is the within-cluster sum of squares, which measures the variance within clusters
- k is the number of clusters
- n is the total number of data points

3. Davies Bouldin index:

The Davies-Bouldin index is a metric used to evaluate the quality of clustering results. It measures the average similarity between each cluster and its most similar cluster, with lower values indicating better-defined clusters.

A low Davies-Bouldin index means that the average similarity between each cluster and its most similar cluster is low, indicating that the clusters are well-separated and distinct. Conversely, a high Davies-Bouldin index means that the average similarity between each cluster and its most similar cluster is high, indicating that the clusters may be poorly defined or overlapping. The Davies-Bouldin index is calculated as follows:

$$\mathrm{DB} = \left(\frac{\mathbf{1}}{\boldsymbol{k}}\right) * \left[\Sigma\, i \left(\boldsymbol{max}_{\boldsymbol{j}} \frac{(\boldsymbol{R}_{\boldsymbol{i}} + \boldsymbol{R}_{\boldsymbol{j}})}{\boldsymbol{d}(\boldsymbol{c}_{\boldsymbol{i}}, \boldsymbol{c}_{\boldsymbol{j}})}\right)\right]$$

where:

- DB is the Davies-Bouldin index
- k is the number of clusters
- i and j are indices that range over the clusters
- R_i is the average distance between all data points in cluster i and the centroid of cluster i
- $d(c_i,\ c_j)$ is the distance between the centroids of clusters i and j

V. RESULTS AND DISCUSSION

Sr. No.	Algorithm	Views	Silhouette value	Calinski Harabasz Index	Davies-Bouldin Index
1.	K-Means Clustering Algorithm	Single view	0.157	113.077	2.700
		Multiview	0.976	4139.465	0.666
2.	Spectral Clustering Algorithm	Single view	0.177	143.309	2.665
		Multiview	0.984	1387.692	0.010
3.	Hierarchical Clustering Algorithm	Single view	0.133	139.013	2.130
		Multiview	0.958	3089.926	0.680

Table 1: Comparison of clustering algorithms using performance metrics

In this table, the performance metrics of three different multiview clustering algorithms: K-means, Spectral and Hierarchical multiview clustering. As shown in Table 1, the Spectral multiview clustering algorithm achieved the highest Silhouette score of 0.984 indicating that it performed the best among the three clustering algorithms. K-means multiview clustering also performed well with a Silhouette score of 0.976 and a Calinski-Harabasz index of 4139.465, while hierarchical had the lowest scores silhouette score compared to other two algorithms. Based on these results, we can conclude that the Spectral algorithm is the most effective clustering algorithm for the dataset used in the proposed work.

VI. Conclusion

The analysis of trending Twitter topics using a multiview clustering algorithm can provide valuable insights into the dynamics of social media conversations. By considering multiple perspectives, such as text content that is semantically related, like-count, retweet-count, reply count, and quote count, this approach can identify distinct clusters of related tweets that may represent different aspects of a larger conversation. However, there are two potential directions for extending the project. Firstly, additional types of relations, such as temporal, social, or geographical relations, could be taken into consideration. Secondly, the project could be expanded to include multiple social media platforms, such as Facebook, LinkedIn, and Instagram.

References

[1] Liang Zhao, Xiao Wang, Zhenjiao Liu, Hong Yuan b, Jingyuan Zhao, Shuang Zhou, "Deep probability multi-view feature learning for data clustering" Expert Systems with Applications 2023

[2] Ben Yang, Xuetao Zhang, Feiping Nie, and Fei Wang, "Fast Multiview Clustering with Spectral Embedding" IEEE, VOL. 31, 2022.

[3] Edi Irawan, Teddy Mantoro, Media Anugerah Ayu, M. Agni Catur Bhakti, I Komang Yogi Trisna Permana,"Analyzing Reactions on Political Issues in Social Media Using Hierarchical and K-Means Clustering Methods", IEEE May 2021.

[4] Hikmat Ullah Khan, Shumaila Nasir, Kishwar Nasim, Danial Shabbir, Ahsan Mahmood "Twitter trends: A ranking algorithm analysis on real time data" Expert Systems with Applications 2021.

[5] Lokesh Mandloi and Ruchi Patel "Twitter Sentiments Analysis Using Machine Learninig Methods" International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, 2020.

[6] Iain J. Cruickshank, “Multi-view Clustering of Social-based Data”, International Studies Review Publication, 2020.

[7] Zeel Doshi, Subhash Nadkarni, Kushal Ajmera and Neepa Shah, “TweerAnalyzer: TweeterTwitter trend detection and visualization”, IEEE Publication, 2017.

[8] Cody Buntain and Jennifer Golbeck, “Automatically Identifying Fake News in Popular Twitter Threads”, IEEE Publication, 2017.

[9] Yixiang Fang, Haijun Zhang, Yunming Ye and Xutao Li, “Detecting hot topics from Twitter: A multiview approach”, JIS Publication, 2014.

[10] Hamidreza Mirzaei, "A Novel Multi-View Agglomerative Clustering Algorithm Based on Ensemble of Partitions on Different Views" IEEE 2010