International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 12 Issue 02 February 2024, Page no. – 3999-4007

Index Copernicus ICV: 57.55, Impact Factor: 8.316

DOI: 10.47191/ijmcr/v12i2.01

Clustering Fake News with K-Means and Agglomerative Clustering Based on Word2Vec

Izhar Muhammad Tianda[1], Mohammad Noufal Ubadah[2], M. Fariz Fadillah Mardianto[3], Said Agil Al Munawwarah[4], Nurhalisa Ishak[5], Dita Amelia[6], Elly Ana[7]

[1,2,3,6,7]Statistics Program, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

[4]Statistics program, Faculty of Mathematics and Natural Science, University of Hasanuddin, Makassar, Indonesia

[5]Mathematics program, Faculty of Mathematics and Natural Science, State University of Makassar, Makassar, Indonesia

ARTICLE INFO

Published on:
03 February 2024

Corresponding Author:
M. Fariz F Mardianto.

ABSTRACT

Fake News on digital platforms is a major problem in this digital age. Many people want to find methods to detect Fake News. This research looks at a way to group Fake News articles using K-Means and Agglomerative Clustering techniques, using the semantic representations from Word2Vec embeddings. The researchers use natural language translation methods and advanced machine learning to improve the accuracy and efficiency of Fake News detection. The study involves getting meaningful features from textual data, turning them into vector representations using Word2Vec, and then applying clustering algorithms to sort similar articles. The methodology aims to improve the most recent state of the art in Fake News detection, helping to create more reliable and robust tools to fight misinformation in the digital age, In the comparative analysis of clustering metrics, K-Means clustering exhibits a Purity Score of 88.09% and an Adjusted Rand Score of 58.03%. Conversely, Agglomerative Clustering with the Ward method yields a Purity Score of 85.13% and an Adjusted Rand Score of 49.36%.The Purity Score of 88.09% for K-Means suggests a strong ability to form clusters where the majority of data points share the same true class. Agglomerative Clustering with Ward, though slightly lower at 85.13%, also demonstrates effective class separation within clusters. When considering the Adjusted Rand Score, which accounts for chance and measures the agreement between true and predicted labels, K-Means significantly outperforms Agglomerative Clustering with Ward. The scores are 58.03% and 49.36%, respectively.

KEYWORDS: Agglomerative, Clustering, Fake News, K-Means, Word2Vec.

I. INTRODUCTION

In recent years, the proliferation of Fake News has become a serious challenge in the era of information overload. The dissemination of falsehoods through various online platforms poses threats to public discourse, decision-making processes, and the overall integrity of information ecosystems. As a response to this pressing issue, the application of advanced machine learning techniques has gained prominence in identifying and clustering Fake News. This study focuses on employing two popular Clustering methods, K-Means, and Agglomerative clustering , to categorize and group Fake News articles based on their semantic content represented by Word2Vec embeddings. Word2Vec, a widely used word embedding technique, captures the contextual relationships between words, allowing for a nuanced understanding of the semantic structure within textual data. The purpose of this research is to look at the efficacy of clustering algorithms in discerning patterns and similarities among fake news articles. Recently, Fake News has drawn a lot of focus on a national and international level It comes up in political contexts a lot, but it's also talked about in relation to business and health organizations and their decisions about stakeholders[1]. By leveraging the semantic information embedded in Word2Vec representations, The goal of the suggested method is to improve clustering's accuracy and effectiveness, contributing to the development of robust tools for detecting and organizing Fake News content. The major reason why distributed word representations improve performance in numerous NLP applications is their ability to

capture semantic regularities[2]. The outcomes of this research are anticipated to offer insights into the potential application of clustering algorithms for effective Fake News identification, thereby contributing to the ongoing efforts to mitigate the adverse effects of misinformation in the digital age. Prior research has underscored the escalating challenges posed by Fake News propagation and emphasized the need for sophisticated techniques to address this issue.These days online Fake News tend to be meddling and assorted in terms of themes, styles and stages[3]. Traditional methods, such as rule-based approaches, have demonstrated limitations in adapting to the dynamic and evolving nature of misinformation. Consequently, the integration a new class of machine learning algorithms has evolved as a promising avenue in order to identify patterns and distinguishing Fake News from genuine information. Recognize and detail related essential hypotheses over different disciplines to energize intrigue inquire about on Fake News[4]. K-Means clustering, a widely employed algorithm for unsupervised learning, has shown success in various text clustering applications. The K-Means computation is dependent on the value of k, which should always be specified in order to undertake any clustering research. Clustering using diverse k values will inevitably create distinctive comes about[5]. By partitioning data into distinct groups based on similarity, K-Means can reveal underlying structures within datasets. Within the context of Fake News detection, leveraging K-Means to cluster articles may offer a systematic approach to discerning shared linguistic features indicative of misinformation. Agglomerative clustering, another prevalent method, constructs hierarchical clusters by iteratively merging the most similar data points. In this research, we look at the effectiveness of hierarchical clustering, which is one of the most common clustering approaches because it is adaptable to most types of data. When compared to partition clustering methods such as K-Means[6]. This approach aligns with the hierarchical structure inherent in language and may capture nuanced relationships between news articles. The hierarchical nature of agglomerative clustering provides a potential advantage in revealing finer-grained distinctions among various types of fake news content.

The integration of Word2Vec embeddings into clustering frameworks holds promise for enhancing the semantic understanding of textual data. Word2Vec representations encapsulate semantic relationships between words, enabling algorithms to discern contextually similar terms. The proposed method is adaptable to many expert systems connected to text mining and can be a competitive option for improved topic modelling to provide direction for future research on technological trend analysis. [7]. By incorporating Word2Vec embeddings into the clustering process, the proposed methodology aims to augment the discriminatory power of clustering algorithms, facilitating more accurate categorization of fake news articles

II. LITERATURE REVIEW

In the realm of Fake News detection, various approaches have been explored, each with its unique set of methods and accuracies Notably, introduced models based on speech characteristics and predictive models, deviating from conventional methods. utilized a Naive Bayes classifier, achieving a 74% accuracy in detecting Fake News from diverse sources employed a combination of machine learning algorithms, but their reliance on an unreliable probability threshold yielded accuracies ranging from 85% to 91% used Naive Bayes for fake news detection on social media, but accuracy suffered for untruthful sources. obtained data from Kaggle, achieving an average accuracy of 74.5%. employed Naive Bayes for detecting Twitter spam senders, attaining accuracies between 70% and 71.2%. experimented with different approaches, achieving an accuracy of 76% explored Naïve Bayes, Neural Network, and Support Vector Machine (SVM), with Naïve Bayes reaching 96.08% accuracy in detecting fake messages, while Neural Network and SVM reached an impressive 99.90% combined KNN and random forests, improving results by up to 8% in a mixed false message detection model. focused on the 2012 Dutch elections' fake news on Twitter, finding that the decision tree algorithm performed best with an F score of 88% presented a counterfeit detection model using N-gram analysis, achieving the highest accuracy at 92%[25]. Inlight of this diverse landscape, our research builds upon these methodologies, employing K-Means and Agglomerative Clustering, as introduced by Z. Khanam. These clustering techniques offer a unique perspective on Fake News detection, leveraging the semantic representations from Word2Vec embeddings. The research aims to contribute to the evolving field of Fake News detection by providing insights into the effectiveness of clustering algorithms in uncovering patterns and similarities among news articles. The advantage of using Word2Vec for clustering lies in its ability to capture semantic relationships between words, allowing for a more nuanced representation of text data. This, in turn, can lead to more meaningful clusters that reflect the underlying semantic structure of the documents. The clustering technique K-Means is classified as a part of partitional clustering method. The segmentation of provided datasets into discrete clusters is accomplished by reducing the squared error between individual data points and the mean (centroid) of the related cluster. [17]. The algorithm accomplishes this by assigning each data point repeatedly to the center of cluster that is closest to it in terms of the Euclidean distance, effectively optimizing the partitioning of the dataset into coherent and internally consistent clusters.

$X = \{xi\}$ where $I = 1,2,\ldots,n$ of d-dimensions data points of size n, X partitioned into 'k' cluster $C = \{cj\}$ where $j = 1,2,\ldots,k$

$$J(Ck) = \sum_{Xi \in Ck} \|xi - \mu k\|^2 \quad (1)$$

$$J(C) = \sum_{K=1}^{K} \sum_{xi \in ck} \|xi - \mu k\|^2 \quad (2)$$

The K-Means algorithm begins by randomly selecting a specified number, k, of centroids from the dataset. Following that, the method computes the distance between each data point and each chosen centroid and assigns each data point to the cluster associated with the closest centroid. When a new member is added to a cluster, the centroid of that cluster is recalculated. The K-Means method iterative procedure continues until the cluster memberships achieve a stable state with minimal changes in following iterations. [17].

The basic steps in k-Means clustering are as follows:

1. Selecting the first division with "k" clusters.
2. Create a new division by attaching each of the patterns to the nearby cluster center.
3. Continue with steps 2 and 3 until the cluster membership is steady.

Agglomerative Clustering algorithm with Ward's method for Fake News prediction. This is a type of Hierarchical clustering algorithm that operates in a bottom-up manner[20].

Initially, every information point considered as a distinct group Then, the algorithm iteratively merges the two clusters that are most similar based on a specific similarity metric, until the appropriate number of clusters is attained [21].

The Ward method is a specific instance of agglomerative clustering that aims to minimize the total within-cluster variance at each step. This means that at each step of the clustering process, the two clusters that are merged are the ones that result in the slightest increment in the total within-cluster variant. This method is particularly effective when the data is globular[21]. which, at each phase, attempts to reduce the overall within-cluster variance. Nonetheless, this requirement can be expressed as

$$dist(Cp, Cq) = \frac{|Cp||Cq|}{|Cp| + |Cq|} \|MCp - MCq\|^2 \quad (3)$$

Where *MCp* stands for the cluster *Cp* centroid vector As a result, the Ward technique additionally merges the two clusters at each stage at the closest possible distance, where the inter-cluster The distances across clustered means, adjusted using an equation of the cluster sizes, are the definition of distances[22].

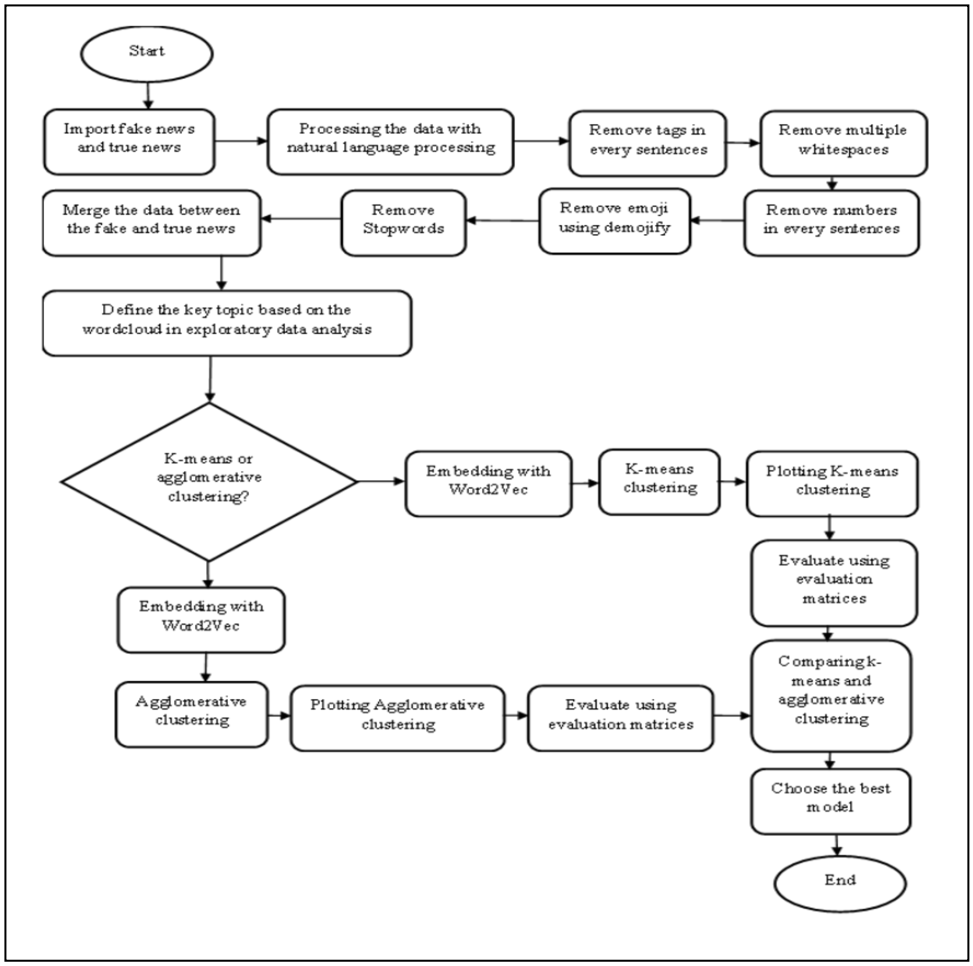

Fig. 1. Research procedure flowchart illustrates the systematic progression of activities

III. RESEARCH PROCEDURE

Adjusted Rand Score or adjusted Rand index was the preferred method for comparing two partitions of a set of observations (e.g., clustering)[12]**.** and purity Score for: Clustering with low quality has a purity value that is nearly 0, whereas clustering with high quality has a purity value that is nearly 1 [13].This offer valuable insights into the effectiveness of clustering algorithms. While the Adjusted Rand Score focuses on the overall agreement between true and predicted labels, the Purity Score emphasizes the homogeneity of clusters in terms of class membership. Researchers and practitioners often use these metrics in tandem to obtain a thorough knowledge of clustering quality and to guide the selection of the most suitable algorithm for a given dataset refer table 1.

Table 1: Evaluation metrics for model goodness

Metrics evaluation	Description
Purity Score	Purity score is a metric used in cluster analysis to assess the caliber of clusters produced by a clustering algorithm The purity cluster quality.
Adjusted Rand Score	The Adjusted Rand score is a measure. in cluster analysis to evaluate the quality of a clustering solution. It assesses how well-defined and homogeneous the clusters are by measuring the degree to which the majority of data points within each cluster belong to a single class or category.

Embarking on the journey of data processing through Natural Language Processing (NLP) with Gensim is akin to navigating a realm where text possesses a complexity that requires purification for clear and meaningful interpretation.The story unfolds in the midst of a forest of words and digital symbols, where our text corpus emerges as a map still dusted with traces from the digital civilization. As we scrutinize more closely, we discern tags, spatial intricacies, the inevitable presence of numbers, enlivening emojis, and lurking stopwords hiding among the lines of words.Gensim, acting as a wise guide, steers us through this cleansing process. First, we traverse the alleys of text to liberate sentences from the shackles of attached digital tags**.** Removing tags in The text data set may contain unnecessary the Gensim Library[14]. Like plucking unwanted wildflowers, Gensim delicately cleanses the text of the long-lingering digital imprints.After freeing itself from tags, the journey moves into more refined terrain. Gensim transforms into a layout expert, tidying up irregular spaces and replacing them with an orderly layout. Words, once scattered, now assemble in formation, ready to be arranged into coherent sentences. However, the true challenge is not yet over. Numbers, often unwelcome guests in the narrative, now become the focus of Gensim's attention. With mathematical precision, Gensim sifts through and discards the numbers that previously crowded every sentence. Now, the text shines unburdened by unnecessary numerals, Emojis as symbols of modern expression, become the next focal point in the cleansing process. Gensim utilizes the magic of demojify to transform emojis into their textual forms or gracefully removes them altogether. With this enchanting touch, the text is liberated from visual embellishments that might disturb its intrinsic meaning.Finally, we confront stopwords Hence, the approach of eliminating stop words is viable, as we aim to exclude unnecessary words. In the preprocessing stage, Gensim offers methods for removing stop words[15]. those often steadfast companions that can disrupt clarity of meaning. With the wisdom of Gensim, we witness the removal of these stopwords, providing an opportunity for the remaining words to carry meaning with sincerity.As the sun sets on the horizon of the narrative, we witness text that has undergone metamorphosis. From a corpus once bustling with digital traces, the text has transitioned into a clean and meaningful narrative. Gensim, as the magical tool in hand, successfully guides us through the forest of words and unveils the hidden stories within the purified lines of text.In the context of clustering, a scientific process is employed to assign numerical labels to news articles, differentiating between fake news and true news. This process involves representing the data, typically news articles, into a format suitable for clustering algorithms. Various Algorithms like K-Means and Hierarchical Clustering, are then applied to discern patterns and structures within the dataset.The representation of data often includes the extraction of relevant features from the text, such as the creation of word vector representations using methodologies like Word2Vec. In this model, we utilized the Word2Vec model to represent the text, effectively capturing the semantic context of the text. [16]. The chosen clustering algorithm, along with the specified number of clusters, endeavors to group the news articles into distinct categories. The labeling aspect involves assigning the numerical label 0 to articles categorized as fake news and the numerical label 1 to those identified as true news. This labeling process is a crucial step in the clustering methodology, allowing for the subsequent analysis and interpretation of the clustered groups. Overall, this scientific process of assigning labels to differentiate between fake and true news within a clustering framework contributes to a nuanced understanding of the inherent structures and patterns present in the diverse landscape of news articles.The process involves training a Word2Vec model

using a substantial collection of textual information. During training, the model learns to Encode words as vectors in a continuous vector space, where the positioning of vectors reflects the semantic relationships between words. Once the model is trained, each word in the vocabulary is associated with a dense vector.In the context of clustering, the Word2Vec embeddings can be used as feature vectors to represent documents or sentences. For example, consider a collection of news articles. Each article can be represented as the average or sum of the Word2Vec vectors of its constituent words. This results in a vector representation that captures the semantic content of the article.The vector representations obtained from Word2Vec can then be fed into a clustering algorithm, For example, algorithms like K-Means or Hierarchical clustering. These algorithms will group together documents or sentences that have similar vector representations, indicating semantic similarity in the underlying text.

IV. EXPLORATORY ABOUT THE DATA

In examining the dataset, The data obtained in this study were taken from the website twitter.com[8]. it becomes evident that it comprises two distinct categories of news articles: "politicsNews," with a count of 11,272, and "worldnews," totaling 10,145 articles. These numerical representations offer a quantitative lens through which to understand the distribution of news topics within the dataset.The higher count in the "politicsNews" category, standing at 11,272, suggests a pronounced emphasis on political matters within the dataset. This numerical prevalence implies a media landscape that prioritizes coverage of political events, likely influenced by their direct impact on societal structures and governance. Conversely, the "worldnews" category, with 10,145 articles, portrays a significant presence of global events in the dataset. Although slightly fewer in number compared to "politicsNews," this count indicates substantial attention given to diverse international issues.The numerical difference between the two categories, with 1,127 more articles in "politicsNews," beckons further investigation. This variance prompts curiosity regarding potential factors influencing the distribution, such as editorial decisions, regional relevance, or the perceived significance of political events within the media landscape. Based on Figure 2, this quantitative breakdown sets the stage for a more nuanced exploration of the dataset. The dominance of political news and the substantial representation of global events hint at the media's thematic priorities, inviting a deeper analysis of the content and potential underlying influences shaping the portrayal of political and global events in this dataset.

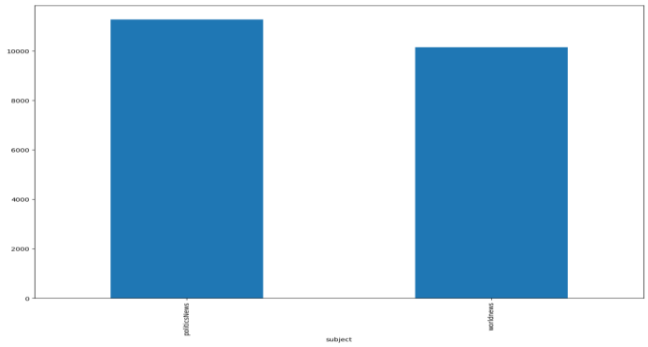

Fig. 2. Plot of the distribution on politics News and World News Based on True News

Within the dataset under scrutiny, a diverse array of news articles has been categorized into six distinct groups: "News," "Politics," "Left News," "government news," "US News," and "Middle-east." These categories are represented by varying counts, each shedding light on the distribution and thematic emphasis within the dataset.The broad category of "News" encompasses the largest share, comprising 9050 articles. This general classification likely encapsulates a broad spectrum of news topics, reflecting the dataset's overall informational diversity. As we move into more specialized classifications, the "politics" category stands out with 6841 articles, suggesting a notable emphasis on political topics, a common focal point in news reporting. The category label "left-news" follows with 4459 articles, indicating a specific ideological orientation within the dataset. This particular classification may represent news sources or content with a left-leaning perspective, adding a nuanced dimension to the dataset's content.For news directly related to governmental activities, the "Government News" category is identified, comprising 1570 articles. This focused classification likely includes official statements, policy updates, or legislative developments. The "US_News" and "Middle-east" categories, with 783 and 778 articles respectively, signify a concentration on news related to the United States and the Middle East. These focused categorizations highlight specific geographical and geopolitical considerations within the dataset.In interpreting these findings, the dataset exhibits a noteworthy emphasis on political topics, as evidenced by the significant count in the "politics" category. Additionally, the presence of category-specific news classifications, such as "Government News," "US_News," and "Middle-east," underscores a focus on distinct aspects of governance and global regions. This categorical breakdown lays the groundwork for deeper exploration, particularly in the context of identifying the prevalence of false information within specific news classifications[9]. The distribution of articles across these categories provides valuable context for understanding the dataset's thematic priorities and sets the stage for more detailed analyses of content and potential misinformation.

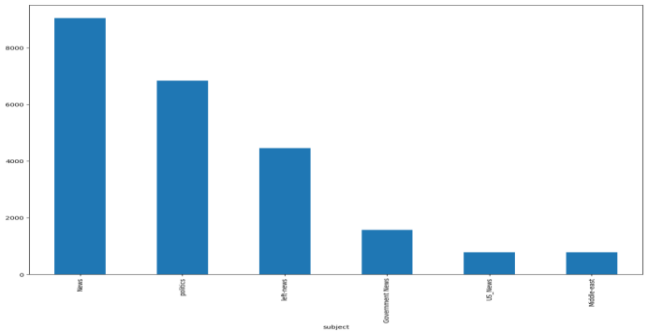

Fig. 3. Plot distribution several subject Based on Fake News

The word cloud analysis of true news articles within the political domain reveals a conspicuous prominence of certain terms, with "Trump" and "Republican" standing out prominently. Therefore, this research was conducted to examine the usefulness and problems when applying the big data word cloud method, which **is** one of the text data analysis methods**[**10**]**. The frequent appearance of these terms underscores their significance and prevalence in the political discourse captured by the dataset. In scholarly terms, the salient manifestation of the term "Trump" implies a substantial focus on the political figure associated with this name, likely referring to the former President of the United States, Donald Trump. This prominence suggests a noteworthy emphasis on his actions, policies, or statements within the analyzed true news articles, indicative of his influential role in contemporary politics. Similarly, the clear visibility of the term "Republican" in the word cloud signifies a recurrent association with the political party. This suggests a substantial coverage of Republican-related topics, encompassing party ideologies, activities, or key figures affiliated with the Republican Party. The recurrent presence of this term suggests an integral role in the true political news narrative encapsulated in the dataset. The cumulative effect of these observations paints a picture of a political landscape where news coverage is notably centered around Donald Trump and the Republican Party. This emphasis might be reflective of the period in which the data was collected, capturing the political dynamics and events involving Trump and the Republican Party during that timeframe. The word cloud, as an illustrative tool, effectively distills the essence of the political news corpus, highlighting the significant role played by these key terms in shaping the narrative of true political news articles.

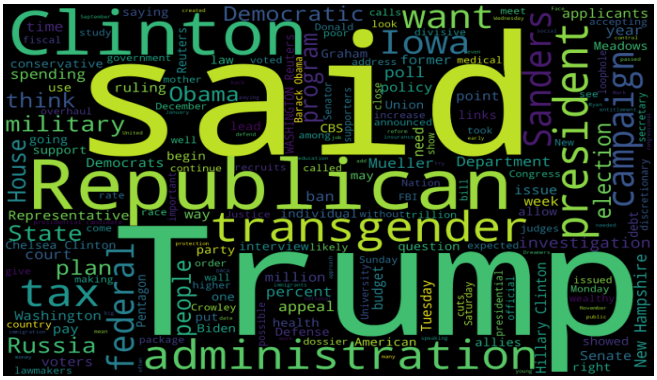

Fig. 4. Wordcloud of the Politics True news helping identify potential Word

The analysis of the word cloud within the realm of fake news in the political domain illuminates distinct patterns, with the prominent presence of certain terms, notably "Trump" and "Republican." The conspicuous recurrence of these terms underscores their significance and prevalence in the narrative of misinformation encapsulated by the dataset.In scholarly terms, the evident prominence of the term "Trump" suggests a substantial thematic concentration on the political figure associated with this name, presumably referring to the former President of the United States, Donald Trump[11]. This distinct visibility implies a noteworthy emphasis on his actions, policies, or statements within the fabric of fabricated political news articles, reflecting the manipulation of information related to his persona.Likewise, the discernible visibility of the term "Republican" in the word cloud signals a recurrent association with the political party. This implies a systematic inclusion of Republican-related topics, encompassing distorted portrayals of party ideologies, activities, or key figures affiliated with the Republican Party. The recurrent appearance of this term suggests a deliberate skewing of information to manipulate perceptions surrounding the political party. Collectively, these observations paint a picture of a fabricated political narrative wherein misinformation is strategically centered around Donald Trump and the Republican Party. This emphasis may be indicative of a deliberate effort to exploit these political figures and entities for the propagation of false information, possibly for the purpose of shaping public opinion or advancing specific agendas. The word cloud, as an analytical tool, effectively captures the essence of the misinformation landscape, highlighting the deliberate focus on Trump and the Republican Party within the corpus of fake political news articles.

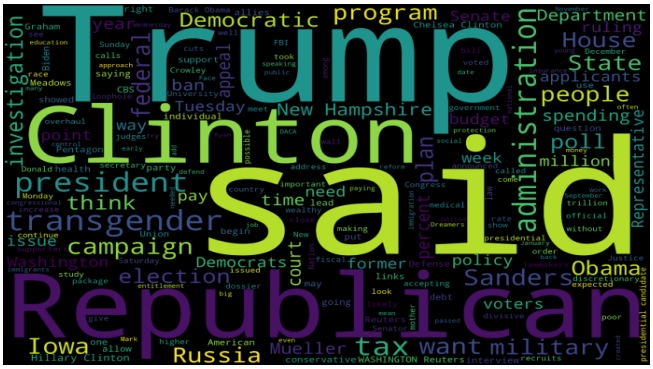

Fig. 5. Wordcloud of the Politics Fake news helping identify potential Word

V. RESULT AND DISCUSSION

In the context of utilizing the K-Means method on a set of data, particularly when referring to Table 2 displaying prediction results and true labels, there are often some inaccuracies in predicting whether an observation belongs to the cluster labeled as fake or true.Scientifically, it can be explained that K-Means method is a partitional clustering method that aims to group data into a specified number, k, of clusters based on measured characteristics[18]. The optimization objective is to keep the overall squared error between every single point of data as low as possible and its respective cluster centroid. However, the success of K-Means greatly relies on data distribution and In addition to the option to choose of the parameter k. When examining Table 2, which presents prediction outcomes and true labels, some mismatches may arise. These discrepancies could be attributed to natural variations in the data or incongruities between the actual distribution of data clusters and the assumptions made by the K-Means algorithm.

Table 2: Dataframe head of prediction and labels K-Means clustering

Sentences	Labels	Prediction
hamilton, star, makes, jokes, twitter, "black...	0	0
obama, races, set, gitmo, terrorists, free…le...	0	1
msnbc, anchor, makes, sexist, comment, men, c...	0	0
china, says, resurfacing, tensions, korean, p...	1	1
new, york, states, trump, energy, efficiency,..	1	1
obamacare, taking, rural, hospitals, let, fac...	0	1
hillary, clinton, 'most, corrupt, militaristi...	0	0

In relation to Figure 6 depicting K-Means clustering, the visual representation suggests a notable distinction in the distribution of data points between true and fake labels. The clusters formed for true labels appear to outnumber those for fake labels, indicating a relatively higher prevalence or concentration of true observations within the dataset.The spatial arrangement of clusters on the graph reflects the grouping patterns identified by the K-means algorithm. True labels, likely characterized by similar features, exhibit a more extensive and cohesive presence in comparison to the clusters associated with fake labels.This clustering outcome holds significance in understanding the inherent structure of the dataset. The prevalence of true labels in more numerous and well-defined clusters suggests a higher degree of similarity or coherence among these data points. On the contrary, the fewer and possibly more dispersed clusters associated with fake labels may indicate a greater variability or heterogeneity in the features of these observations. Analyzing the distribution and relative sizes of clusters provide useful insights regarding the nature of the information and aids in discerning patterns or trends that may be present[19]. It is essential to interpret these findings in the broader context of the dataset's characteristics and the specific features driving the clustering outcomes.

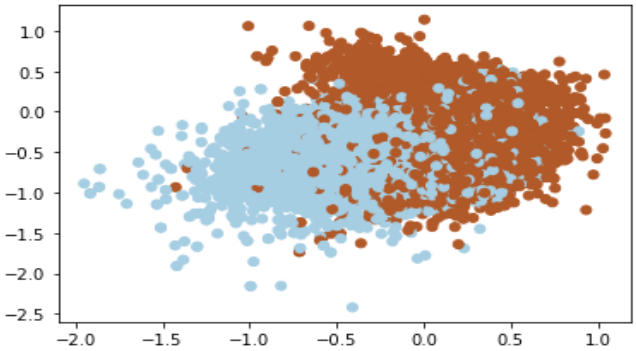

Fig. 6. Plot K-Means clustering to visually represent how data points are assigned to different clusters

In the context of data mining using a DataFrame depending on the outcomes of Agglomerative Clustering with the Ward method, attention is drawn to Table 3, which presents columns labeled Agglomerative Clustering Ward prediction and Labels Notably, the observed data reveals several inaccuracies in predicting whether the labels are true or false. These discrepancies indicate mismatches between the predictions generated by the Agglomerative Clustering algorithm and the actual labels in the dataset. The algorithm, employing the Ward method, endeavors to group data into clusters based on proximity and similar characteristics.However, the success of the algorithm relies on the intrinsic structure of the data and the selection of specific parameters. If there is a misalignment between the actual distribution of data clusters and the assumptions made by the algorithm, it can lead to errors in label predictions[23].In this analysis, it becomes imperative to assess the underlying properties contributing to these prediction errors. Natural variations in the data or complexities in patterns that may not be fully

accommodated by a specific algorithm could be contributing factors.

Table 3: Dataframe head of prediction and labels Agglomerative clustering with ward's

Sentences	Labels	Prediction
hamilton, star, makes, jokes, twitter, "black...	0	1
obama, races, set, gitmo, terrorists, free…le...	0	1
msnbc, anchor, makes, sexist, comment, men, c...	0	1
china, says, resurfacing, tensions, korean, p...	1	0
new, york, states, trump, energy, efficiency,...	1	0
obamacare, taking, rural, hospitals, let, fac...	0	0
hillary, clinton, 'most, corrupt, militaristi...	0	1

In the context of Figure 7, illustrating Agglomerative Clustering with the Ward method, the visual representation suggests a notable contrast in the distribution of data points between fake and true labels. The clusters formed for fake labels appear to outnumber those for true labels, indicating a relatively higher prevalence or concentration of fake observations within the dataset.The spatial arrangement of clusters on the plot reflects the grouping patterns identified by the Agglomerative Clustering algorithm. Fake labels, likely characterized by similar features, exhibit a more extensive and cohesive presence in comparison to the clusters associated with true labels. This clustering outcome holds significance in understanding the inherent structure of the dataset. The prevalence of fake labels in more numerous and well-defined clusters suggests a higher degree of similarity or coherence among these data points. On the contrary, the fewer and possibly more dispersed clusters associated with true labels. May indicate a greater variability or mismatch vector representation withWord2vec in the features of these observations[24].

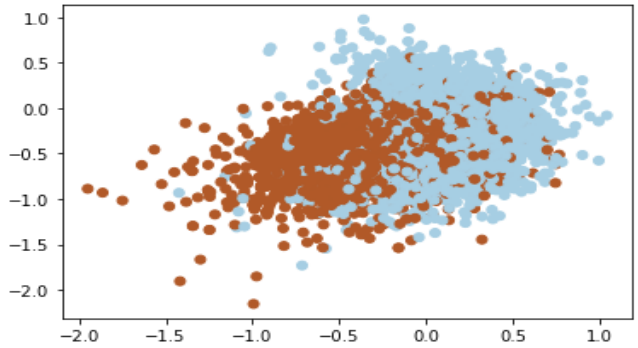

Fig. 7. Plot Agglomerative clustering to visually represent how data points are assigned to different clusters

In the comparative analysis of clustering metrics, K-Means clustering exhibits a Purity Score of 88.09% and an Adjusted Rand Score of 58.03%. Conversely, Agglomerative Clustering with the Ward method yields a Purity Score of 85.13% and an Adjusted Rand Score of 49.36%.The Purity Score of 88.09% for K-Means suggests a strong ability to form clusters where the majority of data points share the same true class. Agglomerative Clustering with Ward, though slightly lower at 85.13%, also demonstrates effective class separation within clusters.When considering the Adjusted Rand Score, which accounts for chance and measures the agreement between true and predicted labels, K-Means significantly outperforms Agglomerative Clustering with Ward. The scores are 58.03% and 49.36%, respectively. In the context of clustering Fake News articles, the emphasis is often on achieving high Purity to ensure that most articles within a cluster are indeed fake. K-Means, with its higher Purity Score and superior Adjusted Rand Score, appears to be the more effective model for this specific task. The argument for the superiority of K-Means is reinforced by its ability to form clusters with a high degree of homogeneity in terms of the true class, specifically focused on identifying clusters of fake news. The higher Adjusted Rand Score further supports K-Means' ability to accurately predict labels while considering chance. Therefore, based on the evaluation metrics and the specific focus on clustering fake news, K-Means clustering emerges as the preferred model for this analysis. Its higher Purity and Adjusted Rand Scores indicate a more robust performance in grouping similar fake news articles compared to Agglomerative Clustering with the Ward method as reflected in Table 4.

Table 4: Evaluation metrics of K-Means Clustering and Agglomerative Clustering with Ward's classes

Clustering class	Adjusted rand score	Purity score
K-means clustering	58.03%	88.09%
Agglomerative clustering with ward's	49.36%	85.13%

VI. CONCLUSION

In conclusion, the K-Means clustering model emerges as the superior algorithm, showcasing better performance according to the evaluation metrics employed in this analysis. However, acknowledging the limitations associated with word2vec embeddings, our future research endeavors will center on the integration of transformer embeddings. This strategic move aims to elevate both the Adjusted Rand Score and Purity Score, leveraging the advanced capabilities of transformers to enhance the standard of vector representations. In summary, our current analysis has illuminated the challenges linked to word2vec embeddings within the clustering context. The proactive adoption of transformer embeddings represents a crucial step in refining

the vector representation process, with the ultimate goal of improving clustering outcomes. Our forthcoming efforts will be dedicated to implementing these enhancements, striving for more accurate and contextually relevant clustering results.

REFERENCES

1. Z. Zikrayanti, "PREVENTION OF ONLINE FAKE NEWS ON SOCIAL MEDIA DURING COVID-19 PANDEMIC: A LITERATURE REVIEW APPROACH," PROCEEDINGS 2022, 153–164.
2. H. Liu, L. Fang, J.-G. Lou, and Z. Li, "Leveraging Web Semantic Knowledge in Word Representation Learning," Proc. AAAI Conf. Artif. Intell. 33, 6746–6753 (2019).
3. X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," Inf. Process. & Manag. 57(2), 102025 (2020).
4. S. Tufchi, A. Yadav, and T. Ahmed, "A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities," Int. J. Multimedia Inf. Retr. 12(2) (2023).
5. M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," Electronics 9(8), 1295 (2020).
6. A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, and A. Song, "Efficient agglomerative hierarchical clustering," Expert Syst. With Appl. 42(5), 2785–2797 (2015).
7. S. Kim, H. Park, and J. Lee, "Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis," Expert Syst. With Appl. 152, 113401 (2020).
8. K. Chen, Z. Duan, and S. Yang, "Twitter as research data," Politics Life Sci. 2021, 1–17.
9. O. Abu Arqoub, A. A. Elega, B. Efe Özad, H. Dwikat, and F. A. Oloyede, "Mapping the Scholarship of Fake News Research: A Systematic Review," Journal. Pract. 2020, 1–31.
10. Y. Hwang, "Development of Big Data Teaching-learning Activities Using Wordcloud Based on Learners' Preferences," J. Converg. Sci., Technol., Soc. 1(1), 1–5 (2022).
11. R. R. Ferreira, "Benkler, Y., Faris, R., & Roberts, H. (2018). Network propaganda: Manipulation, disinformation, and radicalization in American politics. Oxford: Oxford University Press," Mediap. – Rev. Comun., J. Espac. Publico 2020(11), 103–105.
12. D. Steinley, M. J. Brusco, and L. Hubert, "The variance of the adjusted Rand index.," Psychol. Methods 21(2), 261–272 (2016).
13. "Application of Fuzzy and Possibilistic c-Means Clustering Models in Blind Speaker Clustering," Acta Polytech. Hung. 12(7) (2015).
14. A. Borg, M. Boldt, O. Rosander, and J. Ahlstrand, "E-mail classification with machine learning and word embeddings for improved customer support," Neural Comput. Appl. 2020.
15. V. A. Kozhevnikov and E. S. Pankratova, "RESEARCH OF TEXT PRE-PROCESSING METHODS FOR PREPARING DATA IN RUSSIAN FOR MACHINE LEARNING.," Theor. & Appl. Sci. 84(04), 313–320 (2020).
16. G. Mustafa, M. Usman, L. Yu, M. T. afzal, M. Sulaiman, and A. Shahid, "Multi-label classification of research articles using Word2Vec and identification of similarity threshold," Sci. Rep. 11(1) (2021).
17. A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data," Inf. Sci. 2022.
18. Hui Xiong, Junjie Wu, and Jian Chen, "K-Means Clustering Versus Validation Measures: A Data-Distribution Perspective," IEEE Trans. Syst., Man, Cybern.,B (Cybern.) 39(2), 318–331 (2009).
19. C.-P. Hu, J.-M. Hu, S.-L. Deng, and Y. Liu, "A co-word analysis of library and information science in China," Scientometrics 97(2), 369–382 (2013).
20. C. Gaiteri, M. Chen, B. Szymanski, K. Kuzmin, J. Xie, C. Lee, T. Blanche, E. Chaibub Neto, S.-C. Huang, T. Grabowski, T. Madhyastha, and V. Komashko, "Identifying robust communities and multi-community nodes by combining top-down and bottom-up approaches to clustering," Sci. Rep. **5**(1) (2015).
21. M. Haghir chehreghani, "Reliable Agglomerative Clustering," IEEE IJCNN; Int. Jt. Conf. Neural Netw. (IJCNN), Pp. 1-8, 2021; Doi 2018(eprint arXiv:1901.02063).
22. J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," J. Am. Stat. Assoc. 58(301), 236–244 (1963).
23. T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann, "Stability-Based Validation of Clustering Solutions," Neural Comput. 16(6), 1299–1323 (2004).
24. T. Liu, C. Wang, K. Huang, P. Liang, B. Zhang, M. Daneva, and M. van Sinderen, "RoseMatcher: Identifying the impact of user reviews on app updates," Inf. Softw. Technol. 2023, 107261.
25. Z. Khanam, B.N. Alwasel, H. Sirafi, and M. Rashid, IOP Conference Series: Materials Science and Engineering 1099, 012040 (2021).