STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LXVII**, Number 2, 2022
DOI: 10.24193/subbi.2022.2.05

COMPARISON OF DATA MODELS FOR UNSUPERVISED TWITTER SENTIMENT ANALYSIS

SERGIU LIMBOI

Abstract. Identifying the sentiment of collected tweets has become a challenging and interesting task. In addition, mining and defining relevant features that can improve the quality of a classification system is crucial. The data modeling phase is fundamental for the whole process since it can reveal hidden information from the textual inputs. Two models are defined in the presented paper, considering Twitter-specific concepts: a hashtag-based representation and a text-based one. These models will be compared and integrated into an unsupervised system that determines groups of tweets based on sentiment labels (positive and negative). Moreover, word-embedding techniques (TF-IDF and frequency vectors) are used to convert the representations into a numeric input needed for the clustering methods.

The experimental results show good values for Silhouette and Davies-Bouldin measures in the unsupervised environment. A detailed investigation is presented considering several items (dataset, clustering method, data representation, or word embeddings) for checking the best setup for increasing the quality of detecting the sentiment from Twitter's messages. The analysis and conclusions show that the first results can be considered for more complex experiments.

1. Introduction

In the last years, social media has gained ground based on the fact that people can express their feelings, ideas, and attitudes regarding almost everything. An interesting platform is Twitter, where users can write short messages (of maximum 280 characters), called tweets, and can follow other users and observe opinion trends or reviews about politics, social events, pop stars, etc. According to these new tendencies, the Sentiment Analysis domain becomes suitable for analyzing and detecting the hidden sentiment from messages of short lengths. Hence, the main goal of a lot of designed systems was

Received by the editors: 4 April 2023.

2010 *Mathematics Subject Classification.* 68T30, 68T50.

1998 *CR Categories and Descriptors.* I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Text analysis.*

Key words and phrases. Sentiment Analysis, Twitter, Data Representation, Hashtags, Clustering.

to identify if a given piece of information reflects a specific sentiment (e.g., positive, negative, neutral). Exploring and handling texts is a difficult task due to the free style of writing, colloquial language, and the use of abbreviations. Besides the preprocessing step, an essential phase is represented by the way researchers model the data based on the collected messages. So, it could be the case that unrevealed features (attributes that are not part of the actual text- e.g., metadata, emoticons, etc.) can have a high impact on the sentiment detection process and can be related to other built attributes. According to [3], several features are used in literature: lexicon-based (derived from the use of sentiment lexicons), linguistic attributes (number of nouns, adverbs, adjectives), part of speech tagging or others like number of curse words, greeting words, or question marks. All these new features can be combined and define more complex models (e.g. a model has both linguistic features and part of speech tagging) that are very important for the entire process that handles textual information.

Our designed system has the goal of determining if different models or data representations are suitable to identify the sentiment of tweets in the unsupervised context. It is well known that clustering techniques aim to determine groups of instances (in this case, messages) where objects from the same group are very similar and different from the objects of the other groups. Therefore, an analysis of the relevance of the two types of features (hashtag and tweet text) using unsupervised techniques is proposed. Applying different clustering algorithms, we want to determine two groups of messages (one positive and one negative) by using two new models. Bearing in mind that in Twitter's world, hashtags represent an important feature since they are indicators of the message, we define a hashtag representation that will use this concept determined from the tweet. On the other hand, a text-based representation is built based on the idea that maybe the text (without hashtags) composes a relevant input for the sentiment detection problem. Furthermore, in the numerical experiments, we will determine which model is better for sentiment classification in the unsupervised context.

Finally, the original contributions of this paper are the following. Two data representations are defined based on standalone features extracted from tweets: text-based and hashtag-based. The defined representations are applied in the unsupervised context for detecting two groups of messages: one with positive tweets and one with negative messages. The main contribution is represented by comparing the two representations to determine which fits best in the unsupervised scenario. According to our previous experiments for the supervised approach [6], the presented representations are not new but crucial for the starting point of defining more complex and interesting features based on tweets. Moreover, a detailed analysis is conducted to check if the clustering technique impacts the process in combination with two word-embedding methods (TF-IDF and frequency vector).

The remainder of the paper presents the related work in Chapter 2 and the whole methodology (architecture, steps, data models) in Chapter 3. The experimental setup is highlighted in Chapter 4, focusing on the analysis of results. In the end, conclusions and future work are specified in Chapter 5.

2. Related Work

In literature, various approaches define new features or models for detecting the sentiment of a collection of tweets. For example, one of the novel features is the one described in [7]. A flexible feature is built considering its proximity words extracted from a given tweet. An interesting survey is the one of Zhang [13] that presents different features for detecting mental illnesses from textual input, especially if people can present depression symptoms. Various attributes can be handled for this context, from depressive symptoms lexicons, emotion lexicons, and mood emoticons to emotion variability features. An attractive model is OL-DAWE presented in [10], where a tweet's sentiment is reversed if there are many negative words in the message. The system proposed by Chiong [2] uses three groups of features considering sentiment lexicons and platform-specific features for depression detection for tweets input. Therefore, several features are explored: the number of positive, negative, and neutral words, the number of links, negative terms or retweets, or linguistic attributes (e.g., the ratio of adverbs and adjectives). The system of [5] uses hashtags to detect different emotions (e.g., sadness, joy, etc.). A term frequency is computed for each hashtag, and four hashtag-based emotion lexicons are built and applied for the whole process.

The presented paper explores two basic features from collected texts and compares them to determine which fits best for the sentiment detection of tweets. The following section will describe the entire methodology of our approach.

3. Methodology

3.1. **Architecture.** The architecture of the entire system is illustrated in **Figure 1**. In the initial point, the relevant tweets are collected for the experiments. This step is enhanced with the sentiment label provided by the Vader lexicon (if the label is not already present in the dataset). Then, the data is pre-processed and modeled based on two representations: text-based and hashtag-based. These representations extract the relevant aspects from a tweet and pass the outcome to a word embedding step where the models are converted into numerical representations. These inputs will be handled by a clustering algorithm to determine groups of similar data. Next, the clusters can be evaluated and visualized. In addition, the initial data can be visualized for further comparisons. In the following subsections, every phase will be explained.

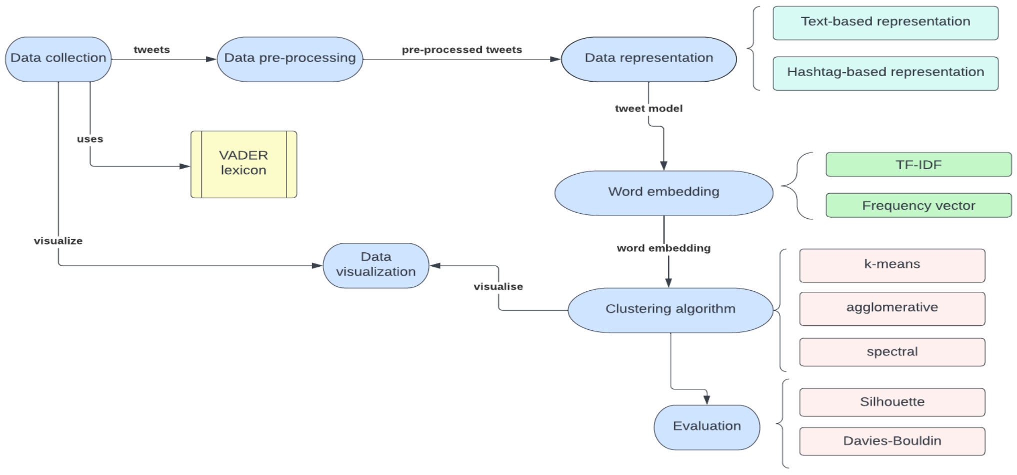


Figure 1. System Overview

3.2. **Pre-processing.** The data pre-processing phase is very important in the Sentiment Analysis process because the system handles textual information. Thus, some cleanup mechanisms are needed to provide the proper input for the data modeling phase. The following operations are used to pre-process the collected tweets: lowercasing, removal of punctuation, stop words, and special characters, and stemming by using the Porter Stemmer [1].

3.3. **Data Representation.** After the pre-processing of tweets, there is a need to extract valuable information from the data. So, various features can be built to handle tweets and to find a proper representation that can be converted into numerical input for the clustering algorithms.

We define two data representations that will be analyzed during our experiments: **text-based** and **hashtag-based** representation. Since a hashtag is an important indicator for a tweet, highlighting the keywords of the short message, we consider that we can design a representation that will take into

[1] http://snowball.tartarus.org/algorithms/porter/stemmer.html

account this aspect. On the other hand, we want to observe the impact of removing these keywords from the tweet and keep only the text for further stages.

3.3.1. *Hashtag Representation.* Considering a collection of tweets $T = \{tweet_1, tweet_2, ..., tweet_t\}$ where t is the length of the tweet, and $tweet_i$ is a message that contains hashtags and text, the hashtag representation will be defined for a $tweet^i_{hash}$:

$$tweet^i_{hash} = \{hash_1, hash_2, ..., hash_h\} \tag{1}$$

where $hash_i$ is the *i-th* hashtag of the tweet from the collection T and h is the number of extracted hashtags from the tweet $tweet^i$,

For example, if there is the tweet "Donald Trump will be the next #president #Trump for 2016 #victory", the hashtag-based representation will contain the corresponding list of hashtags: {president, Trump, victory}.

3.3.2. *Text-based Representation.* The text-based representation will start from the collection of tweets T where every tweet will contain the textual information without the hashtags. So, a $tweet^i_{text}$ will be specified in the next way:

$$tweet^i_{text} = \{word_1, word_2, ..., word_w\} \tag{2}$$

where $word_i$ is the *i-th*word of the tweet from the collection T (it cannot be a hashtag) and w is the number of words that compose the tweet.

If we use the same example as in the hashtag-based representation, the corresponding list of words will be: {Donald, Trump, will, be, the, next, for, 2016}. Of course, if we apply the pre-processing rules, the list will be shortened.

3.4. **Word embedding Representation.** The next step is representing by the conversion of the previously defined models into a numerical representation, technique called **word embedding**. Two methods are used for the experiments: **TF-IDF** and **Count vectorizer**.

TF-IDF (term frequency-inverse document frequency) [1] is defined based on the next formulas

$$TF(term) = \frac{m}{M} \tag{3}$$

and

$$IDF(term) = log(\frac{N}{n}) \tag{4}$$

where m is the number of times the term (word/ hashtag, in our case) appears in the tweet, M is the number of terms in the tweet, N is the number of tweets and n is the number of tweets where the term appears in the collection T.

The **Count vectorizer** is the Python naming [2] for a frequency vector. Basically, it determines for every word/ hashtags the number of appearances. If we have the tweet "Donald Trump will be the next president. He is the best president", the words will have the frequencies: Donald-1, Trump-1, will-1, be-1, the-2, next-1, president-2, he-1, is-1, best-1.

3.5. **Clustering Algorithms and Evaluation Measures.** The next stage will be represented by the clustering algorithm that will use the numerical input modeled in the previous phase to determine relevant groups of data. The main idea is that information from the same group is very similar and different from data from other clusters. For our experiments, the goal is to determine two clusters: one with positive tweets and the other one with negative messages. As algorithms, three techniques are used: **k-means** [12], **agglomerative** [12] and **spectral clustering** [9].

The result of the clustering is evaluated via internal measures like **Silhouette** and **Davies-Bouldin** indices [11].

3.6. **Vader Lexicon.** We label the datasets, in case the sentiment is missing, to visualize the information and have a better view of tweets, with the corresponding polarity by using the Vader lexicon [4]. **Vader (Valence Aware Dictionary and Sentiment Reasoner)** lexicon determines a compound value for every word. Then, a so-called sentiment score of a message *tweet* will be the sum of the sentiment scores of the corresponding terms (word or hashtag):

$$score(tweet) = \sum_{i=1}^{q} score_{Vader}(term_i), \tag{5}$$

where q is the length of tweet *tweet* and $score_{Vader}(term_i)$ is the sentiment score of the i^{th} word.

All in all , the sentiment label of tweet *tweet* is determined as follows:

$$sentiment_{label}(tweet) = \begin{cases} positive, & \text{if } score(tweet) > 0.05 \\ negative, & \text{otherwise} \end{cases} \tag{6}$$

where 0.05 is a threshold computed taking into account different experiments from the literature. So, the dataset that does not contain the sentiment label will be enriched with this information via the Vader lexicon.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

3.7. **Data Visualization.** Data visualization is an important phase during the experiments. There is a need to visualize data before and after the clustering process in order to proceed with detailed analysis and comparison between representations (text and hashtag-based). So, the t-SNE technique is used for dimension reduction of the dataset [8]. In other words, each collection of tweets is reduced from the high-dimensional representations to two dimensions and visualized according to the two sentiments (positive and negative).

4. Experimental setup

The experiments will cover all previously mentioned phases focusing on the datasets, results, and comparison between different experiments in order to highlight which model or representation fits properly for tweets.

4.1. **Data sets.** Three data sets are used for the numerical experiments. The first one contains tweets related to president Joe Biden [3]. This textual input will be referenced as **Joe Biden data set** in the experiment's details. The data collection consisted of 357 tweets that contain hashtags, divided into 298 positive messages and 59 negatives. As a remark, this is an unlabelled data set, so the Vader lexicon will be used to determine an initial sentiment for each message. The second data set is composed of 4.316 tweets with hashtags (3219 positive and 1097 negative) related to COVID-19, messages collected from the April-June period of 2020 [4]. In the experiments, this collection will be called **COVID-19 data set**.

The third data set has tweets from the 2016 USA presidential debate of the Republican Party [5]. It consists of 13.871 labeled messages that are positive, negative, or neutral. Since the focus is also on tweets that have hashtags, 10.323 are kept for the experiments, and only the positive and negative ones are used (the focus is on a binary sentiment). Hence, 2.180 messages are positive, and 8142 are negative. This fact leads to the idea that the data set is quite unbalanced. Therefore, we took 30% as the testing dataset (3097 tweets where 2180 are labeled as positive and 917 as negative messages). In addition, this collection will be called **Republican Presidential debate data set**.

4.2. **Experiments for Joe Biden Data Set.** Before applying the clustering techniques to the chosen dataset, a visualization step is used for checking how data is distributed according to the target classes (positive and negative tweets). Vader lexicon is used for determining the label since the collection does not have the sentiment. T-SNE is used for both previously defined data

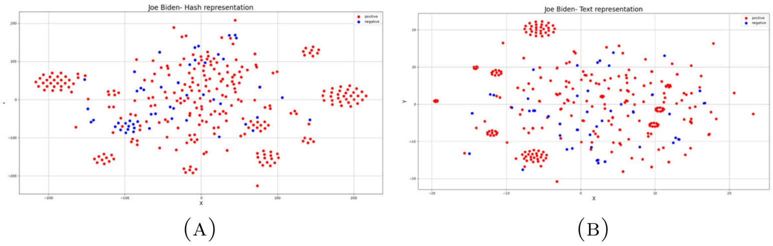
[3] https://www.kaggle.com/ibrahimrrz/tweeter-nlp

[4] https://www.kaggle.com/arunavakrchakraborty/covid19-twitter-dataset

[5] https://www.kaggle.com/datasets/crowdflower/first-gop-debate-twitter-sentiment

representations (hashtag-based and text-based) and word embeddings (TF-IDF and Count Vectorizer). This is a mandatory phase for proceeding with a detailed comparison and drawing relevant conclusions.

4.2.1. *Experiments using the TF-IDF embedding.* Figure **2** presents the dataset before the clustering process for both models. Then, the three algorithms (k-means, agglomerative and spectral) are used for determining two groups of tweets: one with positive messages and one with negative ones.



(A)

(B)

Figure 2. Initial Joe Biden dataset (A)Hashtag (B)Text

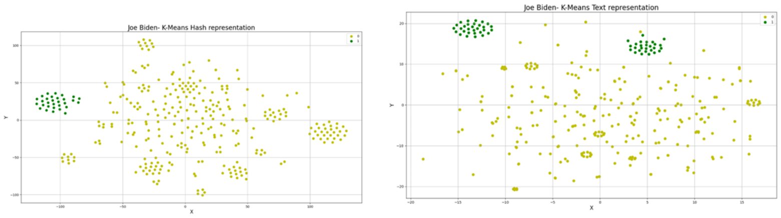
The results of the grouping are reflected in Table **1**, presenting the Silhouette and Davies-Bouldin values. A value closer to 1 is a better clustering for the Silhouette measure and a lower value is a good indicator for the Davies-Bouldin metric.

Table 1. Joe Biden dataset- Clustering results for TF-IDF

Clustering algorithm	Hashtag		Text	
	Silhouette	Davies-Bouldin	Silhouette	Davies-Bouldin
K-Means	**0.215**	**0.761**	**0.142**	**0.877**
Agglomerative	0.214	0.788	0.138	0.982
Spectral	0.168	0.788	0.138	0.982

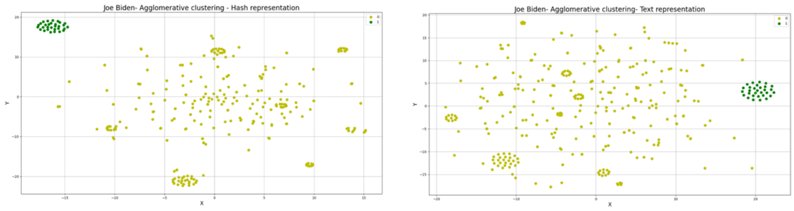
The figures 3 , 4 and 5 highlight the t-SNE representation of the clustering results for the defined models considering the three mentioned techniques.

4.2.2. *Experiments using the frequency vector/Count Vectorizer embedding.* The next word embedding used in the experiments is the frequency vector implemented via the Count Vectorizer library from Python. Figure **6** presents the initial dataset after the modeling with the hashtag and text representation and converting the representations into frequency vectors.



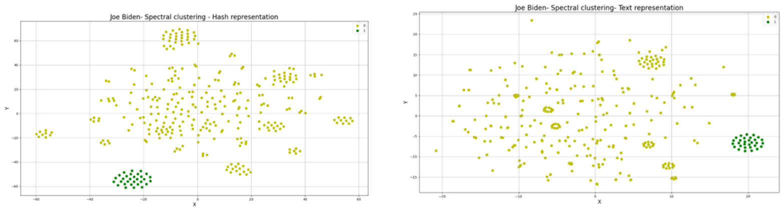
(A) (B)

Figure 3. K-Means TF-IDF (A)Hashtag (B)Text



(A) (B)

Figure 4. Agglomerative TF-IDF (A)Hashtag (B)Text



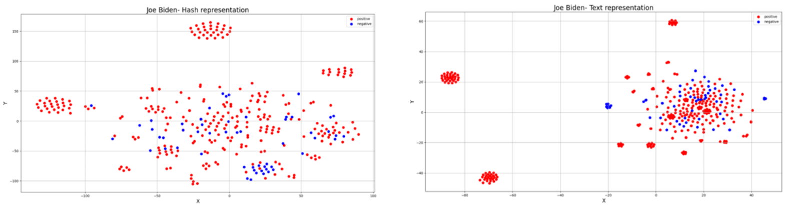
(A) (B)

Figure 5. Spectral TF-IDF (A)Hashtag (B)Text

Then, the clustering results for all the techniques are briefly described in Table **2**. It is noteworthy that there is no significant distinction between the used clustering algorithms.

In the end, the clustered data is illustrated in figures 7, 8, and 9.

4.3. **Analysis for Joe Biden dataset experiments.** From the t-SNE visualization of the Joe Biden dataset, we can notice that the collection is quite unbalanced. There are a lot of positive messages (marked with red color) and

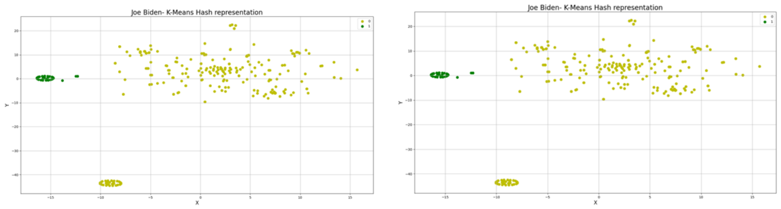


(A) (B)

Figure 6. Initial Joe Biden dataset with frequency vector: A)Hash B) Text

Table 2. Joe Biden dataset- Clustering results for Count vectorizer

Clustering algorithm	Hashtag		Text	
	Silhouette	Davies-Bouldin	Silhouette	Davies-Bouldin
K-Means	**0.102**	**0.971**	**0.095**	**0.998**
Agglomerative	0.101	0.982	0.079	1.123
Spectral	0.101	0.982	0.079	1.123

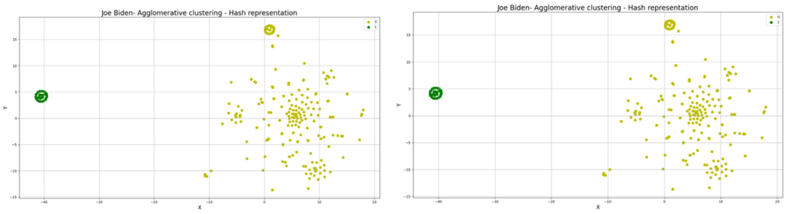


(A) (B)

Figure 7. K-Means A)Hashtag (B)Text

only a few negative tweets (the blue color indicates the negative sentiment). In addition, from the initial visualizations, we can notice that for the hashtag-based model, several sub-groups/ sub-clusters can indicate potential hashtag-based clusters. In other words, from the big group of positive messages, we can deduce small groups that have as highlights some relevant hashtags.

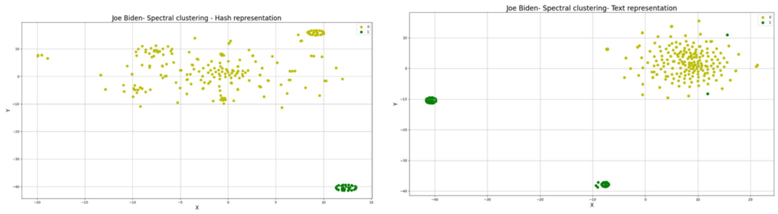
4.3.1. *Clustering algorithm analysis.* From the experiments, we can observe that the best values for the hashtag-based and text-based experiments, for both word embeddings, are the ones produced by the K-means algorithms.



(A)

(B)

Figure 8. Agglomerative clustering A)Hashtag (B)Text



(A)

(B)

Figure 9. Spectral clustering A)Hashtag (B)Text

Analyzing the values for the Silhouette and Davies-Bouldin indexes, there is no significant difference between the used clustering methods (the values are quite similar with only small variations). Therefore, the added value is not represented by the technique and the enhancement is reflected by how data is modeled (the defined representations). Also, regarding the t-SNE visualization of the clustering results, it can be highlighted the idea that there are a lot of positive messages and only a spot of negative tweets.

4.3.2. *Word embeddings analysis.* Comparing the results from TF-IDF and frequency vector embeddings, the best values, in terms of Silhouette and Davies-Bouldin, are the ones when the representations are converted into TF-IDF. This embedding brings relevance to our context since we work with texts of small lengths. When we convert textual input into frequency vectors we can face the issue that text is really small after the preprocessing phase. The evaluation in terms of embedding techniques is supported also by the t-SNE visualization. For the TF-IDF case, we distributed data in small sub-groups, but for the count vectorizer situation, after the clustering, there are a lot of dots (instances) distributed randomly and only two or three subgroups can be identified (so no semantic relevance can be deduced from the visualizations).

4.3.3. *Data representation analysis.* The best results are achieved by the hashtag-based representation in the unsupervised context (determining groups of similar tweets). This case is reflected also in the visualization part where we can identify subgroups from the dataset. All these things highlight the idea that hashtags are relevant concepts for the Twitter world and they can be drivers for defining the groups.

4.3.4. *Summary of the analysis.* **Table 3** sums up the conclusions of the analysis for the experiments conducted on the Joe Biden dataset.

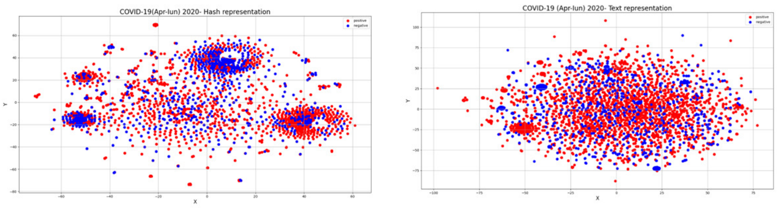
Table 3. Joe Biden dataset- Analysis summary

Concept	Conclusions
The clustering algorithm	K-Means produces the best values, but there is no significant difference between the used clustering algorithms
Word embedding	TF-IDF has better results than frequency vector
Data representation	For the unsupervised context, the hashtag-based is more relevant than the text-based model
Dataset	It is quite unbalanced: 298 positive tweets and 59 negatives.
Data visualization	Better visualization for the hashtag-based representation since we can identify relevant subgroups (conceptual/semantical clusters) in comparison with the text-based model.

4.4. **Experiments for COVID-19 Data Set.** Considering the previous experiments and the ones conducted on the second dataset (COVID-19), we will present only the results for one clustering algorithm (K-Means) and the TF-IDF embedding. The initial visualization of the dataset for both representations (text-based and hashtag-based) is presented in **Figure 10**.

The clustering evaluation is given in the **Table 4** that illustrates the Silhouette and Davie-Bouldin values for the K-Means algorithm for both representations. The clustering visualization is presented in **Figure 11**.

4.4.1. *Analysis and conclusions.* The COVID-19 dataset is bigger than the previous one and data is more balanced than the Joe Biden set, but still quite unbalanced in terms of positive and negative sentiments. The clustering evaluation reflects the idea that the hashtag representation is better than the

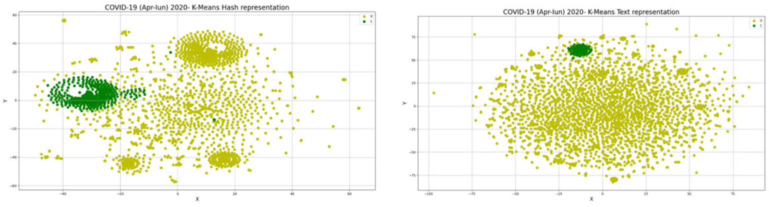


(A) (B)

Figure 10. Initial COVID-19 dataset A)Hashtag (B)Text

Table 4. COVID 19 dataset- Evaluation for TF-IDF embedding

Representation	Silhouette	Davies-Bouldin
Hashtag-based	0.166	1.030
Text-based	0.022	1.044

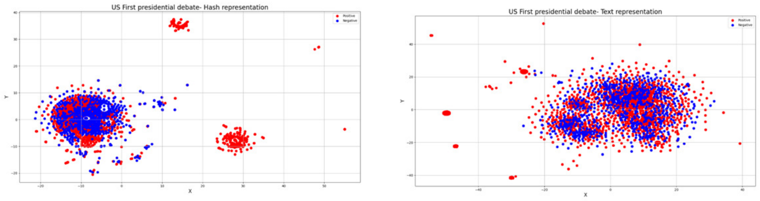


(A) (B)

Figure 11. K-Means A)Hashtag (B)Text

text-based one. This conclusion is strengthened by the data visualization since we have grouped data where we can identify more subgroups. The text-based visualization is a big cluster of positive tweets and a smaller group of negative instances.

4.5. **Experiments for 2016 Republican Presidential Debate Data Set.** The visualization of the initial dataset is presented in **Figure 12**. The test data is more balanced (2180 positive and almost 1000 negative) than the others, an idea reflected also in the t-SNE visualization. The experiments are conducted using the K-Means algorithm and TF-IDF embedding for text-based and hashtag-based representations.



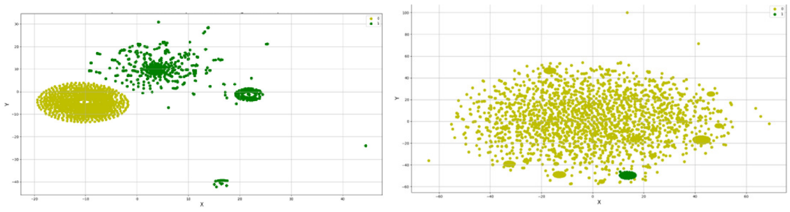
(A) (B)

FIGURE 12. Initial Republican Debate dataset A)Hashtag (B)Text

Table 5 defines the Silhouette and Davies-Bouldin values for the defined representations, using the K-Means technique. The clustering visualization is drawn in **Figure 13**. As evident from the evaluation measures' values and t-SNE visualization, the text representation gives pretty poor outcomes. Almost all instances are labeled as one class, and only a few are marked as the opposite one.

TABLE 5. COVID 19 dataset- Evaluation for TF-IDF embedding

Representation	Silhouette	Davies-Bouldin
Hashtag-based	0.445	0.566
Text-based	0.011	1.099



(A) (B)

FIGURE 13. K-Means A)Hashtag (B)Text

Based on the analysis, we can conclude that also for bigger datasets the hashtag-based representation is better than the text one, in the unsupervised context. Moreover, we notice that we do not have the small subgroups presented in the smaller datasets. Thus, the models tend to evolve into two main clusters with the defined labels: positive and negative.

4.6. **Comparisons and Summary.** **Table 6** presents the final conclusions of the conducted experiments on the three datasets: Joe Biden, COVID-19 and 2016 Republican Presidential Debate.

Table 6. Conclusions and summary

Concept	Conclusions
Clustering method	No significant difference between clustering technique
Word embedding	TF-IDF has better results than frequency vector
Data representation	For the unsupervised context, the hashtag-based is the relevant mode
Dataset	The Joe Biden dataset has 298 positive tweets and 59 negatives. The COVID-19 collection has 3219 positive and 1097 negative. The last one contains 2180 positive and 917 negative

5. Conclusions and Future Work

In the Sentiment Analysis area, there is a need to define new data representations and explore the valuable information the collected input offers. In this paper, we used two data representations for textual information of short lengths, in this case, tweets, that use the whole text or extract relevant platform-specific features: hashtag-based and text-based representations. Moreover, several clustering algorithms apply these two in the unsupervised learning context. The goal is to determine two main groups of tweets according to two sentiment labels: positive and negative. The experimental results reveal only a slight difference between the used clustering techniques. Therefore, the data representations bring the main enhancement. Regarding evaluation and data visualization, hashtag representations handle short messages better than text ones. Even though this is a simple methodology that uses two existing representations and analyzes which one fits better in the unsupervised context, our plan involves more interesting and complex work. The plan is to define topic-driven clusters based on the most popular and relevant hashtags collected from the data, exploring bigger datasets. Also, using the models defined in [6] and combined with more complex ones (e.g., BERT-based models) will be quite interesting. Overall, the first results are encouraging and design the steps for more exploratory and extensive experiments.

References

[1] Baeza-Yates, R., Ribeiro-Neto, B., et al. *Modern information retrieval*, vol. 463. ACM press New York, 1999.

[2] Chiong, R., Budhi, G. S., and Dhakal, S. Combining sentiment lexicons and content-based features for depression detection. *IEEE Intelligent Systems 36*, 6 (2021), 99–105.

[3] Hung, L. P., and Alias, S. Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection. *Journal of Advanced Computational Intelligence and Intelligent Informatics 27*, 1 (2023), 84–95.

[4] Hutto, C., and Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (2014), vol. 8, pp. 216–225.

[5] Koto, F., and Adriani, M. Hbe: Hashtag-based emotion lexicons for twitter sentiment analysis. In *Proceedings of the 7th Annual Meeting of the Forum for Information Retrieval Evaluation* (2015), pp. 31–34.

[6] Limboi, S., and Dioşan, L. Hybrid features for twitter sentiment analysis. In *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II 19* (2020), Springer, pp. 210–219.

[7] Pilar, G.-D., Isabel, S.-B., Diego, P.-M., and Luis, G.-Á. J. A novel flexible feature extraction algorithm for spanish tweet sentiment analysis based on the context of words. *Expert Systems with Applications 212* (2023), 118817.

[8] Van der Maaten, L., and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research 9*, 11 (2008).

[9] Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing 17* (2007), 395–416.

[10] Wang, W., Li, B., Feng, D., Zhang, A., and Wan, S. The ol-dawe model: tweet polarity sentiment analysis with data augmentation. *IEEE Access 8* (2020), 40118–40128.

[11] Xu, D., and Tian, Y. A comprehensive survey of clustering algorithms. *Annals of Data Science 2* (2015), 165–193.

[12] Xu, R., and Wunsch, D. Survey of clustering algorithms. *IEEE Transactions on neural networks 16*, 3 (2005), 645–678.

[13] Zhang, T., Yang, K., Ji, S., and Ananiadou, S. Emotion fusion for mental illness detection from social media: A survey. *Information Fusion 92* (2023), 231–246.

Babeş-Bolyai University, Faculty of Mathematics and Computer Science, 1 Mihail Kogălniceanu, Cluj-Napoca 400084, Romania

Email address: sergiu.limboi@ubbcluj.ro