The Scientific Temper (2024) Vol. 15 (3): 2692-2699
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.37
https://scientifictemper.com/

RESEARCH ARTICLE

A bigdata analytics method for social media behavioral analysis

Muhammed Jouhar K. K.[*], **K. Aravinthan**

Abstract

Twitter on web-based entertainment has become an important part of everyday life. This medium provides a list of current events in real-time, most of which is difficult to understand, so it must be sorted to find useful information. Human biology, pharmacology, and experimental factors influence their behavior. Twitter tweets are a text store that can reflect human emotions and sentiments. Behavior analytics (BA) is analyzing the behavior of individuals. BA can be used to filter useful information from tweets in healthcare and business applications. This paper presents the analysis of human behavior using Twitter data and a proposed social media behavior analysis big data analytics (BASMBA) algorithm. The proposed algorithm uses several techniques in its pre-processing, feature selection, and classification of tweets using BIGDATA. Additionally, the accuracy of the algorithm is verified using the precision factor and recovery time.

Keywords: Social media, Big data, Twitter, Machine learning, Behavior analysis.

Introduction

Behavior analysis uses learning principles to identify behavioral changes. While some branches strive to understand the cognitions underlying human behavior, the Behavior analytics (BA) focuses on human behavior. The BA has practical applications in healthcare, tissue monitoring and trade promotion. User-generated information is a good source of opinion and can be important for different applications that require understanding public opinion on a concept. A classic example illustrating the importance of public opinion refers to companies that are able to capture customer opinions about their products or competitors (Dubey *et al.,* 2020). BA is a science based on the fundamentals and principles of behaviorism. Behavior analysis can be significant in various different ways. Analysis can be based on the experimental study of behavior, which involves studying behavior and applying it to an individual, social, or cultural context. BA studies also address philosophical, historical, hypothetical and systemic issues in direct assessment. The experimental BA involves fundamental research aimed at adding to the body of knowledge about behavior, while the applied BA focuses on applying these behavioral principles to the real world. Those working in the field of applied conduct examination are keen on behavior and its relationship to the environment. BA is the computational study of people's opinions, attitudes, and emotions toward entities. Entities can represent persons, events, or subjects. Online entertainment is a great place to study human behavior from your tweet data. Social platforms have become important venues for global dialogue. Individuals everywhere are using the Internet to communicate and express themselves freely due to the ease and availability of gadgets. Social networks show how people interact with events and express opinions about everything that happens in the world. Twitter is quite possibly one of the most visited and used social networks. It's a really great resource for data about people's interests. To get useful information on a particular topic, tweets related to that topic can be analyzed. There are many techniques in BA and before applying any of them, the data is transformed into a structured form, which increases the performance level of the analysis process and consumes less time and less storage space. BA can identify the behavior of a particular text and then denote it as negative and positive. Applied conduct experts utilize applied exploration to make and carry out programs in schools, families, and communities that have been shown to be powerful in tending to ways of behaving related with mental imbalance and other formative disabilities.

Department of Computer Science, Adaikalamatha College, affiliated to Bharathidasan University, Vallam, Thanjavur, Tamil Nadu, India.

[*]**Corresponding Author:** Muhammed Jouhar K. K., Department of Computer Science, Adaikalamatha College, affiliated to Bharathidasan University, Vallam, Thanjavur, Tamil Nadu, India., E-Mail: muhammedjouhar87@gmail.com

How to cite this article: Jouhar, M. K. K., Aravinthan, K. (2024). A bigdata analytics method for social media behavioral analysis. The Scientific Temper, **15**(3):2692-2699.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.37

Source of support: Nil

Conflict of interest: None.

Received: 10/07/2024 **Accepted:** 10/08/2024 **Published** : 10/09/2024

The rationale for this approach to therapy suffering from mental imbalances and other formative problems is data-driven decision-making to inform behavior modification treatment choices in diverse populations and settings. This paper proposes a BA technique called BASMBA that uses rule mining, decision trees, and clustering.

Related Work

In general, behavioral analysis is applied to Twitter data that can be processed through natural language processing. The analysis of Twitter data is based on a taxonomic level of word and phrase learning. Classification of Twitter messages is similar to sentence-level sentiment analysis (Sharma *et.al.,* 2020). However, Twitter sentiment analysis is a unique task in the microblogging domain due to the casual and informal language used in tweets. A problem in the microblogging domain is how to use sentiment analysis techniques on well-formed data (Anspach *et. al.,* 2020). Many researchers include part-of-speech features, but results are still mixed. They conduct research in various ways, such as automatically collecting tweepy API training data.

DB technology uses a predefined dictionary for sentiment classification, but it has limitations in classification. In database-based systems, there is a lack of area-based semantics because of the utilization of the bag-of-words concept. Conversely, MLB frameworks have domain-specific sentiment rating training data. Linguistic differences and class imbalances in text can be addressed using bootstrap techniques (Ozbay *et al.,* 2020) to explain the combination of SVM and classification of the Twitter dataset. The attitude and sentiment of the company are shown in the tweets based on the social and business news on the Twitter account. Jena, Rabindra (Jena and Rabindra, 2020) proposed a structure Twitter sentiment analysis by manipulating irregularities in casual performance. So, use hashtags to categorize tweets into likes and dislikes. Roland T. Rust *et al.*, 2021 proposed a neural network technique for artificial intelligence based on content classification. For example, a user's Facebook page has content of different polarities. Meier, Adrian, and Leonard Reinecke 2021 proposed a data collection strategy for opinion examination. Online comment and tweet analysis with KNIME. Rathje, Steve, Jay J. Van Bavel, and Sander Van Der Linden. 2021 Analyzing stock data tweets using a combination of neural networks and support vector machines. Saura, Jose Ramon, Daniel Palacios-Marqués, and Domingo Ribeiro-Soriano 2021 use the Naive Bayes algorithm to identify state-specific sentiment on Facebook. Therefore, such analysis is content-based

Methodology

Twitter is a popular social network with many types of users from different countries. Many organizations and news providers use Twitter to share information. Twitter generates a large database containing real-time information. Hidden information found in this data can be used for a variety of purposes. However, the results depend on the choice of a suitable set of functions. This document uses three distinct steps in its analysis of tweet behavior. Pre-process the dataset first for best results. The pre-processed data goes through a feature selection process before being classified by BASMBA. The architecture is shown in Figure 1.

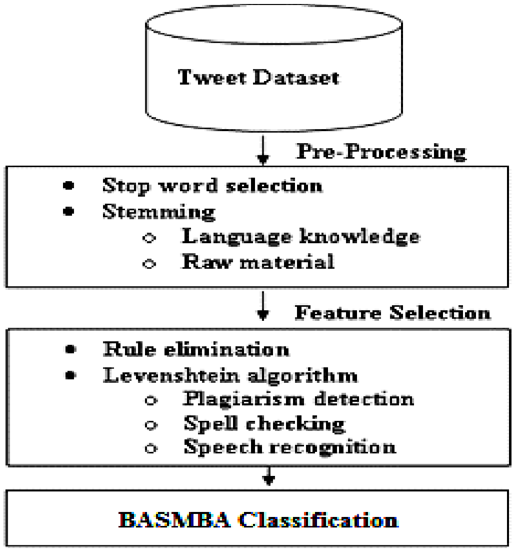


Figure 1: BASMBA architecture

Stop word selection is the first process in tweet data pre-processing. Stop words are less discriminative common words used in English communication, such as "the," "is," "at," "which," and "on". Tweet data contains countless stop words, which makes analysis difficult, and is usually filtered before processing BA text because it affects output performance. The study in (Onan, Aytug and Mansur Alp Tocoglu, 2021) examined spent tweets from six different datasets to find out the effect of removing stop words on effective BA results. Because, BASMBA identifies the stop words to be removed and only further processes the remaining information.

Tweets contain words with corresponding meanings. Stemming is the process of finding the root of a word. Derivation can be seen as a crude heuristic process that simply cuts off the end of words. Unlike lemmatization, lemmatization does not involve lexical analysis. Stemming is a pre-processing task that is performed before calling other application-specific tasks. One of the earliest applications of stemming was information retrieval (IR). Searching for the keyword "explosion" failed to retrieves documents indexed by the term "explosives". Stemming solves this problem because indexing will be done using the root word Online web search tools, for example, Google Search, use a bypass. A 2009 analysis of Google searches showed that a few postfixes, for example, "-s," "-ed," and "-ing," were thought to be closely related to word stems. The suffixes "-able", "-tive", "-ly", and "-ness" are considered less relevant. BASMBA uses dictionaries as raw material to look up language meanings.

Table 1 outlines the diverse methodologies utilized in the BASMBA algorithm for analyzing social media behavior, specifically Twitter data. It succinctly encapsulates the multifaceted approach taken in the research, starting with the initial pre-processing steps of stop word removal and stemming, which aim to enhance data quality and normalize text for analysis. Feature selection and rule elimination follow, ensuring that only relevant attributes are considered for classification. Levenshtein's algorithm is then applied for string matching and spell checking, contributing to the accuracy of the analysis. Decision trees facilitate feature selection and classification, while association rule mining uncovers meaningful correlations within the data. Finally, document clustering aids in grouping similar tweets together, fostering better organization and analysis. Collectively, these functions and techniques enable the BASMBA algorithm to effectively uncover behavioral patterns and trends within Twitter data, offering valuable insights for social media analysis and research.

The choice of BASMBA function is followed by the use of rule elimination furthermore, the utilization of Levenshtein's algorithm. Grammatical rules affect polarity, which can significantly improve BA. The advanced system also considers POS tags, applies rejection rules, and detects irony. While polarity determination as one of BA tasks is an area that has been studied, some language-specific aspects (negation, irony, metaphor) still present challenges and areas for improvement. The study in identified negation and created rules to improve BA output. The method predicts behavior as an unsupervised method that computes the polarity and strength of words and phrases. The study showed a positive relationship between's the extremity determined by the system and the words assigned to the participants.

Levenshtein distance, It is named after Vladimir Levenshtein and is a string metric that measures the difference between two sequences. Informally, the alter distance between two words is the base number of single-character alters (i.e., insertions, deletions, or substitutions) expected to transform single word into another.It is also known as alter distance, in spite of the fact that it can also represent a bigger group of distance measures and is firmly connected with pairwise string alignment. Mathematically, the edit distance between two strings a, b (of length |a| and |b|) is given by cam,b(|a|,|b|) where:

$$lev_{a,b}\,(\mathrm{i,j}) = \begin{cases} \max\,(i,j) & if\min(i,j)=0 \\ min\begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(ai\neq bj)} \end{cases} & otherwise \end{cases}$$

Where 1(ai≠bi) is a marker capability, equivalent to 0 when ai≠bi,otherwise equal to 1, cam, , b(i,j) is the distance between the first i characters of a and the first j characters of b. The main component in the base relates to cancellation (from a – b), the second element element corresponds to inclusion, and the third element matches or does not match, depending on whether the respective symbols are the same. BASMBA dynamically uses this algorithm for string matching and spell checking. It is used because Levenshtein's algorithm calculates contingent upon whether the particular images are the equivalent expected to adjust string to get another string.

Decision Tree Algorithm

The decision tree algorithm in machine learning is a methodical approach to construct a predictive model from data. Mathematically, let *D* denote the training dataset comprising feature vectors **x***i* and corresponding target labels *yi*, where *i*=1,2,…,*N*. Each feature vector

Table 1: Functions and techniques employed in the BASMBA algorithm

Function/Technique	*Description*
Stop word removal	Identifying and removing common stop words (e.g., the, is, at) from tweet data to improve analysis accuracy and performance.
Stemming	Process of finding the root of words to reduce them to their base or root form (e.g., running becomes run) for normalization.
Feature selection	Selecting relevant features or attributes from pre-processed tweet data for further analysis and classification.
Rule elimination	Applying rules or heuristics to eliminate irrelevant or redundant features from the dataset, enhancing classification accuracy.
Levenshtein's algorithm	String metric for measuring the difference between two sequences, used for string matching and spell checking in tweet data processing.
Decision trees	Utilizing decision tree algorithms for feature selection and classification of tweet data based on attribute values and rules.
Association rule mining	Discovering interesting associations or correlations between large numbers of data items, aiding in decision-making and analysis.
Document clustering	Grouping similar documents or tweets together based on their characteristics or features, enhancing data organization and analysis.

$\mathbf{x}i$ is represented as $\mathbf{x}i=(xi1,xi2,\ldots,xid)$, where d is the dimensionality of the feature space.

The decision tree algorithm recursively partitions the feature space by selecting the best attribute A and threshold θ at each internal node to minimize impurity or maximize information gain. In a binary split, the dataset D is divided into left (Dleft) and right (Dright) subsets based on the condition $xj\leq\theta$, where j is the index of the selected attribute.

Formally, at each step of the recursion, the algorithm aims to find the attribute A and threshold θ that minimize a given impurity measure Impurity(D) or maximize information gain $IG(D,A)$:

$$A,\theta = \arg\min_{A,\theta} \text{Impurity}\left(D_{\text{left}}\right) + \text{Impurity}\left(D_{\text{right}}\right)$$

Or

$$A,\theta = \arg\max_{A,\theta} IG(D,A)$$

Where, impurity (D) represents the impurity of dataset D and $IG(D,A)$ is the information gain when splitting on attribute A.

Common impurity measures include Gini impurity:

$$\text{Impurity}_{\text{Gini}}\ (D) = 1 - \sum_{k=1}^{K} \left(p_k\right)^2$$

$$\text{Impurity}_{\text{Entropy}}\ (D) = -\sum_{k=1}^{K} p_k \log_2\left(p_k\right)$$

where K is the number of classes and pk is the proportion of samples in class k in dataset D. The decision tree algorithm continues recursively until a stopping criterion is met, such as reaching a maximum depth, achieving a minimum number of samples in each leaf node, or when further splits do not significantly improve the model's performance.

Figure 2 illustrates the decision tree algorithm's process utilized within the research methodology. Initially, the training set is initialized for analysis. The algorithm checks if all tuples belong to the same class; if so, it returns a single-node tree root with the corresponding label. Alternatively, if the attributes are empty, the algorithm returns a single-node tree root with the most common class label. If neither condition is met, the algorithm proceeds to select the best attribute for data splitting and creates a decision tree root with the chosen attribute. This process continues recursively, with the algorithm splitting the data into subsets based on selected attribute values and calling itself for each subset until all branches are labeled. Finally, the algorithm concludes with a stop sign, indicating the end of the decision tree construction process. This flowchart provides a clear representation of the decision tree algorithm's iterative and recursive nature in analyzing and classifying data.

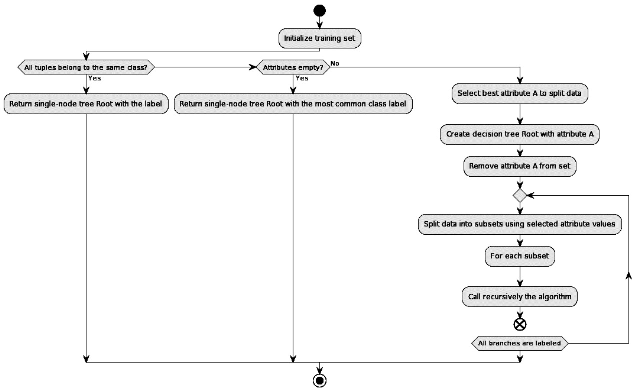

Initialize training set
All tuples belong to the same class?
Yes
Attributes empty?
No
Yes
Return single-node tree Root with the label
Return single-node tree Root with the most common class label
Select best attribute A to split data
Create decision tree Root with attribute A
Remove attribute A from set
Split data into subsets using selected attribute values
For each subset
Call recursively the algorithm
All branches are labeled

Figure 2: Decision tree construction algorithm flowchart

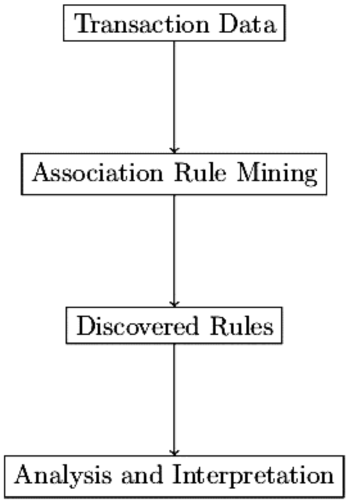


Figure 3: Flowchart of association rule mining process

Figure 3 illustrates the sequential steps involved in association rule mining. It begins with transaction data, which serves as the input for the association rule mining process. The data undergoes analysis using association rule mining techniques, resulting in the discovery of association rules. These rules represent patterns or relationships within the transaction data. Subsequently, the discovered rules are subjected to analysis and interpretation, where their significance and implications are evaluated. This process aids in extracting meaningful insights and actionable knowledge from the transaction data, facilitating informed decision-making and strategy formulation. Overall, the diagram encapsulates the iterative and analytical nature of association rule mining, highlighting its role in uncovering valuable patterns and associations in large datasets.

BASMBA then processes the selected features using decision trees, rule mining, and document clustering for classification. A decision tree (DT) is used to represent possible values specified at nodes. A decision tree is a greedy algorithm that builds a decision tree in a top-down recursive divide-and-conquer approach. The algorithm starts with tuples in the training set and selects the best attributes that yield the most classification information. It generates a test node for this, and then top-down decision tree induction splits the current set of tuples in view of the current test attribute value. Generation stops when all tuples in the subset have a place with a similar class, or on

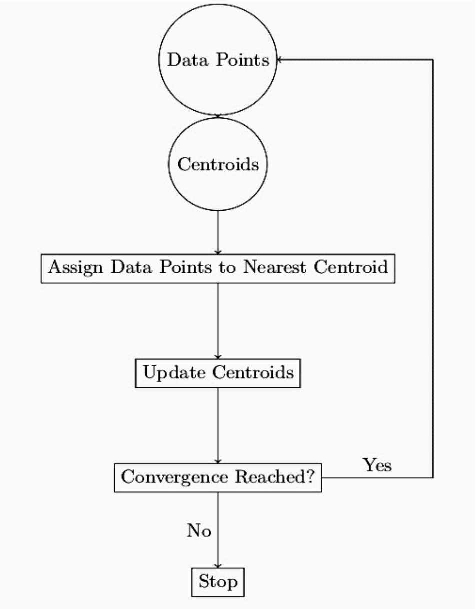


Figure 4: K-means clustering process

the other hand if further separation into more subsets is not worth continuing.

Figure 4 visually depicts the sequential steps of the K-means clustering algorithm. It begins with the initialization of data points and centroids, where centroids represent the initial cluster centers. The algorithm then iteratively assigns each data point to its nearest centroid based on a distance metric, typically Euclidean distance. After the assignment step, the centroids are updated to the mean of the data points assigned to each cluster. This process continues iteratively until convergence is reached, which occurs when the centroids no longer change significantly between iterations or a predefined number of iterations is reached. The diagram illustrates the iterative nature of the K-means algorithm, emphasizing how data points are grouped into clusters based on proximity to centroids and how centroids are adjusted to optimize the clustering result. Ultimately, K-means clustering is a widely used method for partitioning data into coherent groups, making it valuable for various data analysis and machine learning tasks.

BASMBA then uses association rule mining (ARM) on the DT output, because ARM can find interesting associations or correlations between large numbers of data items. The discovery of interesting associations among large transaction log is used in the decision-making process. Let us consider the following assumption, representing association rules in mathematical representation: J = {i1, i2, …, im} is a bunch of elements. D = set of database transactions, where each transaction T is a bunch of elements, say T J. Every transaction is associated with an identifier named a TID. A, B = set of elements. A transaction T is said to contain A if and only if A T . An association rule is an entailment of the form A B where A, B and A B = the rule A B holds on the set D of transactions supporting S, where S is the percentage of transactions in D that include A B, namely

The standard A B has certainty C in the exchange set D in the event that C is the level of transaction in D containing A that likewise contains B, i.e.,

confidence (A B) = P (B A) = [support count(A B)/support count(A)].

A rule that Settle for Apoyo's Minimal Shade (min_sup), furthermore, a base certainty limit (min_conf) is called a strong rule. A group of elements is called an element set. A set of elements containing k elements is a set of k elements. The recurrence of an item set is the quantity of exchanges that contain the item set. This is moreover referred to simply as frequency, number of supports or set count. If the frequency of occurrence of the set of elements is greater than or equal to min_sup furthermore, the complete number of exchanges in D, the item set meets the minimum support. Therefore, the number of transactions required for an item set to satisfy the minimum support is called the minimum support. An itemset is a frequent itemset if it satisfies the minimum support.

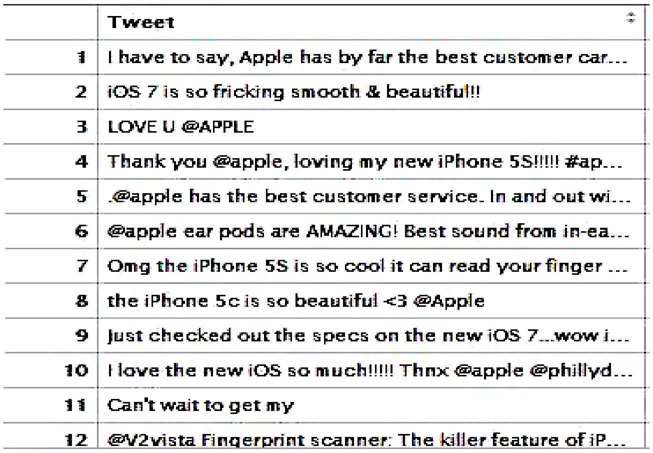
Document clustering has been studied in several different areas of text mining and information retrieval. Originally, document clustering was studied to improve the accuracy or recall of information retrieval systems (Sun *et al.,* 2021) what's more, as an effective way to find nearest neighbors of documents. Grouping has additionally been proposed for browsing document collections (Long *et al.,* 2021) or organizing results returned by search engines in response to user queries (Kaliyar *et al.,* 2021). Document grouping is also used to automatically generate groups of documents (Kaliyar *et al.,* 2021). Due to its efficiency, BASMBA uses K-means clustering.

Results

Experiments and results analysis were performed using an Intel i5-2410M CPU with a 2.30 GHz processor and 4 GB RAM running Windows operating system. For the resulting analysis, after loading the tweet dataset and performing BA on the collected tweets, use Python to process the data. Figure 2 shows a snapshot of the uploaded tweets.

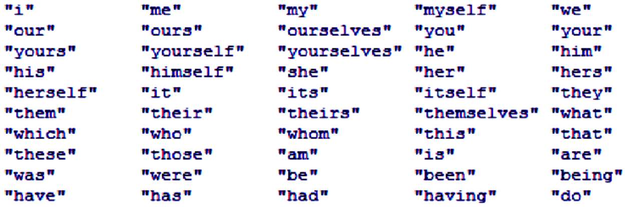
BASMBA Pre-processing

Since a single sentence can generate many attributes (for example, my sister and I swim together generate 15 attributes), and it is more difficult for the algorithm to deal with large data sets with high dimensions. Often, the dimensionality of the dataset needs to be reduced by diminishing the quantity of traits. The authors apply two methods to reduce dimensionality: Stopword removal and lemmatization. Single stop words (e.g., from, to, an, to) were removed from the dictionary to reduce dimensionality, resulting in a 15 to 20% reduction in data for various sports



	Tweet
1	I have to say, Apple has by far the best customer car...
2	iOS 7 is so fricking smooth & beautiful!!
3	LOVE U @APPLE
4	Thank you @apple, loving my new iPhone 5S!!!!! #ap...
5	.@apple has the best customer service. In and out wi...
6	@apple ear pods are AMAZING! Best sound from in-ea...
7	Omg the iPhone 5S is so cool it can read your finger ...
8	the iPhone 5c is so beautiful <3 @Apple
9	Just checked out the specs on the new iOS 7...wow i...
10	I love the new iOS so much!!!!! Thnx @apple @phillyd...
11	Can't wait to get my
12	@V2vista Fingerprint scanner: The killer feature of iP...

Figure 5: Tweets snapshot



```
"i"        "me"         "my"           "myself"      "we"
"our"      "ours"       "ourselves"    "you"         "your"
"yours"    "yourself"   "yourselves"   "he"          "him"
"his"      "himself"    "she"          "her"         "hers"
"herself"  "it"         "its"          "itself"      "they"
"them"     "their"      "theirs"       "themselves"  "what"
"which"    "who"        "whom"         "this"        "that"
"these"    "those"      "am"           "is"          "are"
"was"      "were"       "be"           "been"        "being"
"have"     "has"        "had"          "having"      "do"
```

Figure 6: BASMBA stop words list

activities, but stop words in phrases (e.g., "for" "Physical activity" in "Physical activity is" or "of") is retained. Stop words were removed using the stop word removal function. Figure 3 shows the BASMBA stop word list.

To accommodate readers with differences in color vision, these figures must still be usable when printed in grayscale. Use non-color terms to refer to graphic elements, such as "represented by a square" rather than "represented by blue". Use different patterns in bar charts, different line patterns in graphs, and different shapes in plots to distinguish groups of elements and reinforce color differences.

Stemming reduces dimensionality by identifying the root of a word and removing enough prefixes and prefixes from different word forms. One bypass application is to count sentiment word usage and perform basic BA. When used with a dictionary or spell checker, the stemmer can be used to suggest corrections when spelling mistakes are found. Stemming is based on the assumption that words have structure based on stemming and stemming. The study of words and their parts is called morphology. In an IR system, given a word, lemmatization is really about finding lexical variants. The term fusion denotes a combination of variants in a common root. Words can contain prefixes and suffixes, commonly called affixes. Stemming usually deals with suffixes. Figure 6 shows the output derived from BASMBA.

BASMBA also focuses on the handling of negative grammar rules, as it can help improve the accuracy of classifying feelings or behaviors. The processed rules represent a subset of the grammatical rules used for negation processing in the language, i.e. the most common rules in tweets. The proposed method uses Levenshtein in tweets to remove plagiarism. A step in the process is to verify that the copied tweet is actually the original.

BASMBA Classification

Decision trees are a popular method of inductive reasoning. They are used because they are resistant to noisy data and learn disjunctive expressions. A decision tree is a k-matrix tree in which each internal node specifies a test for some feature relevant to the tree's inductive selection. Decision tree induction is a decision tree classifier that learns to build a tree structure where each internal node (no leaf nodes) represents an attribute test. Each branch represents a test result, and each external node (leaf node) represents a class prediction. At each node, the algorithm selects the best partition data attributes for each class. The best attributes for partitions are selected by attribute selection with information gain. The attribute with the highest information gain splits that attribute. The information gain of the attribute is given by

$$\inf o(D) = -\sum_{i=1}^{m} p_i \log_2 (p)$$

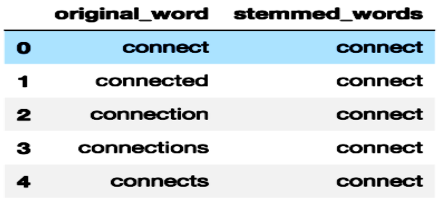
BASMBA rule mining then identifies frequently occurring elements (words) before using the document groupings listed in Figure 7 as its final output.

Algorithm

Iterative algorithm

Step	*Description*
1	Set *k* to be the length of *L*. Set the length of *m* or *t*. If *k*=0, return my output. If *m*=0, return *k* and exit. Construct a matrix with 0..*m* rows and 0..*k* columns. Set *k* to be the length of *L*. Set the length of *m* or *t*. If *k*=0, return my output. If *m*=0, return *k* and exit. Construct a matrix with 0..*m* rows and 0..*k* columns.
2	Initialize the first row to 0..*k*. Initialize the first column to 0..*m*.
3	Check each character of *L* (*i* from 1 to *k*). Check each character of *t* (*i* from 1 to*m*).
4	If *L*[*i*] is equal to *t*[*j*], the cost is 0. If *L*[*i*] is not equal to *t*[*j*], the cost is 1.
5	Sets cell *d*[*i,j*] of the array equal to the minimum value: a. Add 1 to the next higher unit: *d*[*i*–1,*j*]+1. second. Add 1 to the left cell: *d*[*i,j*–1]+1. here. Add cost to cells diagonally up and to the left: *d*[*i*–1,*j*–1]+*cost*.
6	After completing iteration steps (3, 4, 5, 6), find distance *l* in cell *d*[*k,m*].

where k is the number of tweets, L is the length in characters, t is a single tweet, d is the distance, and m is the maximum value of tweets.



	original_word	stemmed_words
0	connect	connect
1	connected	connect
2	connection	connect
3	connections	connect
4	connects	connect

Figure 7: BASMBA stemming output

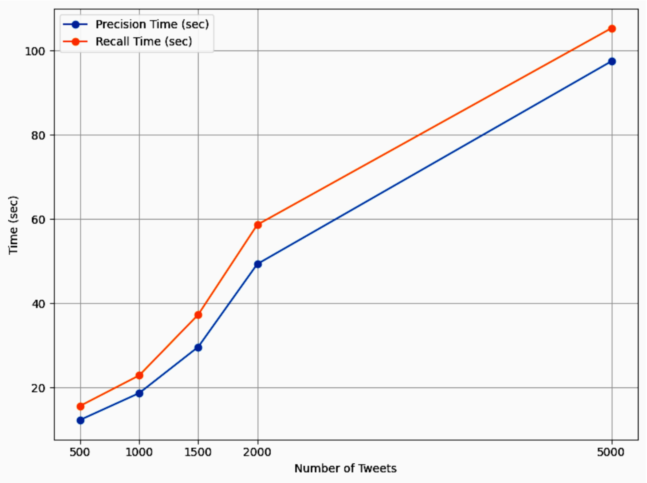


Figure 8: Final output performance graph

Table 2: Performance analysis of BASMBA algorithm: Precision and recall times

No. of tweets	*Precision time (in sec)*	*Recall time (in sec)*
500	12.33	15.66
1000	18.69	22.89
1500	29.66	37.33
2000	49.33	58.66
5000+	97.47	105.33

The effectiveness of BASMBA is measured using precision given by Precision = TP/ (TP+FP) and recall given by Recall = TP+FN, where TP-True Positives, FP-False Positives, FN-False Negatives. Table 1 lists the accuracy and recovery time values, while Figure 5 shows the performance in graph form.

Table 2 provides a comprehensive overview of the algorithm's performance across different dataset sizes. As the number of tweets increases, both precision and recall times also exhibit a noticeable increase, indicating the algorithm's ability to process larger datasets while maintaining relatively stable performance. For instance, with 500 tweets, the precision time stands at 12.33 seconds and the recall time at 15.66 seconds, whereas for datasets exceeding 5000 tweets, the precision time rises to 97.47 seconds and the recall time to 105.33 seconds. Despite the increase in processing time, the algorithm demonstrates consistent behavior, which suggests its suitability for handling datasets of various sizes with minimal degradation in performance. This analysis underscores the algorithm's reliability and efficiency in real-world applications, enabling users to make informed decisions regarding its implementation and scalability.

Figure 8 illustrates the performance of the BASMBA algorithm in terms of precision and recall times across varying numbers of tweets in the dataset. As the number of tweets increases, both precision and recall times exhibit a consistent upward trend, indicating that processing larger datasets requires more computational time. For instance, with 500 tweets, the precision time is approximately 12.33 seconds, and the recall time is about 15.66 seconds, while with 5000 tweets, the precision time rises to around 97.47 seconds, and the recall time increases to approximately 105.33 seconds. The graph underscores the algorithm's scalability, as it is capable of handling datasets of different sizes. However, it also highlights the trade-off between dataset size and processing time, providing valuable insights for optimizing algorithm performance and resource allocation in real-world applications.

Conclusion

Behavior analysis is an interesting field for applying natural language processing and automating text conclusions. It is used for social media trend analysis and sometimes for marketing purposes. Creating behavior analysis programs in Python is not a difficult task thanks to modern ready-to-use libraries. Text mining generally refers to the process of extracting valuable information from unstructured text. Hidden information on social networking sites, bioinformatics, and Internet security. Recognition through text mining is a major challenge in these fields. This paper proposes a novel BASMBA technique in which the DM technique is used in combination to identify the behavior of tweets. The results of an experimental study are also presented. K-means is used because of possible variations in agglomerative hierarchical clustering or possible mixed combinations with K-means. Furthermore, in the training dataset, although all summaries have almost the same length, they have slightly different numbers of frequent words after pre-processing. In order to avoid empty attribute values in any transaction in the transaction database collection, only the high-frequency words of each text are selected. The reason is that empty attribute values in the transaction set produce word sets that contain empty values. These phrases containing null values are not useful in classification. The discarded text is not considered as training data. This choice is made to increase the number of attributes for generating associated word sets, increasing the chances of generating more words in the word set, while also increasing the total number of word sets. Therefore, it can be concluded that BASMBA is a viable technique that can be used to automatically identify BAs from tweets.

References

Anshari, M., & Almunawar, M. N. (2021). Adopting open innovation for SMEs and industrial revolution 4.0. *Journal of Science and Technology Policy Management*, *13*(2), 405-427

Anspach, N. M., & Carlson, T. N. (2020). What to believe? Social media commentary and belief in misinformation. *Political Behavior*, *42*(3), 697-718.

Dubey, A. D., & Tripathi, S. (2020). Analysing the sentiments towards work-from-home experience during COVID-19 pandemic. *Journal of Innovation Management*, *8*(1), 13-19.

Ghermandi, A., Camacho-Valdez, V., & Trejo-Espinosa, H. (2020). Social media-based analysis of cultural ecosystem services and heritage tourism in a coastal region of Mexico. *Tourism Management*, *77*, 104002.

Grover, P., Kar, A. K., & Dwivedi, Y. K. (2022). Understanding artificial intelligence adoption in operations management: insights from the review of academic literature and social media discussions. *Annals of Operations Research*, *308*(1), 177-213.

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social science computer review*, *38*(1), 10-24.

Huang, M. H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science*, *49*, 30-50.

Jena, R. (2020). An empirical case study on Indian consumers' sentiment towards electric vehicles: A big data analytics approach. *Industrial Marketing Management*, *90*, 605-616.

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications*, *80*(8), 11765-11788.

Keles, B., McCrae, N., & Grealish, A. (2020). A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. *International journal of adolescence and youth*, *25*(1), 79-93.

Li, Y., & Xie, Y. (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of marketing research*, *57*(1), 1-19.

Long, C. (2021). I Just Like the Stock" versus "Fear and Loathing on Main Street": The Role of Reddit Sentiment in the GameStop Short Squeeze by Cheng Long, Brian M. Lucey, Larisa Yarovaya: SSRN. SSRN Scholarly Paper ID 3822315. *Social Science Research Network, Rochester, NY*.

Mariani, M., & Baggio, R. (2022). Big data and analytics in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management*, *34*(1), 231-278.

McClure, C., & Seock, Y. K. (2020). The role of involvement: Investigating the effect of brand's social media pages on consumer purchase intention. *Journal of retailing and consumer services*, *53*, 101975.

Meier, A., & Reinecke, L. (2021). Computer-mediated communication, social media, and mental health: A conceptual and empirical meta-review. *Communication Research*, *48*(8), 1182-1209.

Oh, S. H., Lee, S. Y., & Han, C. (2021). The effects of social media use on preventive behaviors during infectious disease outbreaks: The mediating role of self-relevant emotions and public risk perception. *Health communication*, *36*(8), 972-981.

Onan, A., & Toçoğlu, M. A. (2021). A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification. *Ieee Access*, *9*, 7701-7722.

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), e2024292118.

Ray, P., & Chakrabarti, A. (2022). A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. *Applied Computing and Informatics*, *18*(1/2), 163-178.

Saura, J. R., Palacios-Marqués, D., & Ribeiro-Soriano, D. (2021). Using data mining techniques to explore security issues in smart living environments in Twitter. *Computer Communications*, *179*, 285-295.

Sharma, P., Berwal, Y. P. S., & Ghai, W. (2020). Performance analysis of deep learning CNN models for disease detection in plants using image segmentation. *Information Processing in Agriculture*, *7*(4), 566-574.

Sun, H., Burton, H. V., & Huang, H. (2021). Machine learning applications for building structural design and performance assessment: State-of-the-art review. *Journal of Building Engineering*, *33*, 101816

Zhao, Y., Da, J., & Yan, J. (2021). Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *Information Processing & Management*, *58*(1), 102390.