# Enhancing Clinical Sentiment Analysis with a Novel Stochastic Model Based on Brownian Motion

**Maria El-Badaoui**
Lasti Laboratory, National School of Applied Sciences, Sultan Moulay Slimane University, Khouribga, Morocco
maria.el-badaoui@usms.ac.ma (corresponding author)

**Noreddine Gherabi**
Lasti Laboratory, National School of Applied Sciences, Sultan Moulay Slimane University, Khouribga, Morocco
n.gherabi@usms.ma

**Fatima Qanouni**
Lasti Laboratory, National School of Applied Sciences, Sultan Moulay Slimane University, Khouribga, Morocco
qanouni.fatima@gmail.com

*Received: 23 July 2025 | Revised: 6 September 2025 and 15 September 2025 | Accepted: 18 September 2025*

 *DOI: https://doi.org/10.48084/etasr.13609*

## ABSTRACT

**Depression profoundly affects emotional states, behaviors, and overall quality of life. In the digital era, patients increasingly share their experiences through online comments, offering valuable yet complex emotional narratives. However, traditional sentiment analysis methods, which often assign a single emotional label to an entire text, fail to capture the nuanced intra-comment emotional fluctuations, overlooking the evolving and sometimes contradictory nature of emotional expression. This study introduces a novel sentiment analysis framework that models intra-comment emotional flow as a Brownian-like trajectory. Inspired by the random motion of particles in physics, the proposed method treats each sentence as a time step and assigns it a sentiment polarity score. The cumulative sequence of these scores constructs an emotional curve that reflects the dynamic affective progression throughout the comment, effectively simulating a Brownian path. From these trajectories, quantitative indicators—including drift, variance, $Z$-signal changes, trajectory length ($L$), maximum emotional level ($M$), minimum emotional level ($m$), and amplitude — are extracted, providing a detailed and structured understanding of emotional instability. To uncover patterns across comments, the K-Means clustering algorithm is applied, enabling the automatic grouping of comments with similar emotional dynamics. In addition, Principal Component Analysis (PCA) is used to reduce dimensionality and facilitate the visualization and interpretation of emotional profiles. The proposed approach is applied to a corpus of comments from depressed patients to identify characteristic emotional patterns and examine differences according to the type of treatment. This work highlights the relevance of stochastic modeling for advancing sentiment analysis in complex psychological contexts and offers a tool to support remote monitoring, personalized interventions, and early detection of emotional relapse.**

***Keywords-sentiment analysis; Brownian motion model; emotional trajectories; K-means***

## I. INTRODUCTION

Depression is one of the most widespread mental health disorders in the world, profoundly affecting people's emotional state, behavior, and overall quality of life. In today's digital landscape, patients increasingly express themselves through online comments, health forums, and treatment reviews. The automatic analysis of such discourse provides a valuable opportunity to understand their emotional experience. Two main categories of approaches are generally used: supervised learning and unsupervised learning. Supervised learning is based on the exploitation of annotated data and the use of Machine Learning (ML) algorithms (such as random forests [1] or SVM [2]) and Deep Learning (DL) models such as

Convolutional Neural Networks (CNN) or Recurrent Neural Networks (LSTM [3], GRU). These techniques have been widely used in the classification of feelings in social networks or online evaluations [4-6]. At the same time, unsupervised learning allows exploration of hidden structures in unlabeled data. Methods such as Principal Component Analysis (PCA) and clustering algorithms such as K-means or DBSCAN are widely used to discover emotional or behavioral profiles, particularly when the data is complex, subjective, or partially annotated. This approach is particularly useful in the analysis of patient stories, where evolving emotional dynamics do not always lend themselves to rigid classification.

In [7], a hybrid multilingual approach was proposed to classify tweets into sentiments (positive, negative, neutral). This method was based on the calculation of two measures, positivity and negativity, based on semantic similarity using WordNet, combined with a fuzzy logic system (fuzzification, inference rules, and defuzzification). To manage massive volumes of data, this approach was parallelized with Hadoop and MapReduce. Experiments showed that this approach significantly improved accuracy (95%) compared to traditional sentiment analysis methods. The integration of fuzzy logic means that the ambiguity of sentiments expressed in tweets can be dealt with more effectively.

In [8], a method was proposed to classify feelings related to the COVID-19 pandemic in Indonesia using a combination of fuzzy logic and the Naïve Bayes classifier. Data were collected from Twitter in the period between January and April 2020 and classified into three classes: positive, negative, and neutral. After preprocessing the tweets and applying the Fuzzy Naïve Bayes model, an accuracy of 83.1% was obtained at a training/test distribution of 70:30. The analysis shows that positive tweets were 36.7% of the total, followed by negative tweets at 35.0%, while 28.3% showed a neutral sentiment. The use of fuzzy logic improves the management of textual multiplicity, usually found in opinions posted through social networks.

The method in [9] was based on a binary analysis of feelings, using classic models such as Random Forest, Logistic Regression, and Decision Tree, the first achieving an accuracy of 97.78%. In addition, a BERT (Bidirectional Encoder Representations from Transformers) model was used for the fine classification of emotions (joy, sadness, neutrality, fear, and anger), achieving an average accuracy of 94%. In addition, Latent Dirichlet Allocation (LDA) thematic modeling was used to identify topics correlated with the emotions detected. In addition, a web platform integrated these models to predict users' emotional state in real time.

In [10], a DL approach was employed to analyze the dynamics of sentiments expressed by breast cancer patients on Breastcancer.org. Combining manual annotation, semi-supervised co-training, and deep models (CNN, LSTM, BiLSTM), the BiLSTM with sentiment embeddings achieved an F1-score of 91.9%. The results show that rapid, positive responses in forums improve emotional well-being and that a specialized medical lexicon optimizes the detection of semantic polarities. In [11], DL was used to analyze emotions and classify cancer-related medical texts, in particular from messages posted on social networks. A distributed framework (Apache Spark) was used to process large volumes of data from three corpora: (1) cancer-related tweets, (2) a medical news dataset, and (3) scientific abstracts. The process included rigorous preprocessing (cleaning, lemmatization, polarity dictionaries), feature extraction using TF-IDF, Word2Vec, and Doc2Vec, dimension reduction (PCA, Chi2, SVD), and classification using various models. The LSTM model achieved an accuracy of up to 97.84% on the corpus of tweets.

In [12], a sentiment classification system was applied to user reviews of Indonesian health applications (Halodoc, Alodokter, Klikdokter) collected from the Google Play Store. The corpus consisted of 9,310 reviews (4,950 positive and 4,360 negative). This approach was based on the IndoBERT model, a pre-trained version of BERT optimized for Indonesian language processing. After rigorous preprocessing (cleaning, tokenization, normalization), the model was refined using supervised learning. The experimental results showed an accuracy of 96%, recall of 96%, precision of 95% and an F1-score of 95%, outperforming previous approaches such as Fast Large-Margin or BiLSTM.

In [13], a sentiment analysis model was proposed to predict the risk of suicide from microblog text, integrating not only textual content but also social contexts (relationships between users) and thematic contexts (hashtags). This approach introduced the notion of structure similarity to model indirect relationships between users, complemented by analysis of the thematic proximity of messages. This was formalized using Laplacian matrices to regularize the classification. Experiments carried out on two Twitter datasets showed that incorporating the social context significantly improves accuracy compared to methods based solely on text. The proposed model (SASS-T) achieved an accuracy of 82.1% and 83.4% on two datasets, outperforming traditional methods such as SVM, NB, LASSO, and previous models such as SANT and SMSC.

In [14], the unsupervised clustering algorithms K-Means and DBSCAN were used to analyze tweets about the COVID-19 pandemic. The K-Means algorithm groups tweets according to their semantic similarity, whereas DBSCAN works to find dense clusters and identify outliers. Performance evaluation involved standard metrics, such as precision, recall, F1 score, and accuracy. The results showed that DBSCAN performed a bit better than K-Means, with a 78% accuracy score compared to 74% in the two-cluster configuration, and an F1 score of 0.80 compared to 0.76 for K-Means. This study concluded that DBSCAN is more effective at detecting latent sentiment structures, particularly when the number of clusters is not known in advance. In [15], 18000 tweets relating to the #AFP hashtag were analyzed using unsupervised learning techniques, in particular the K-Means algorithm. After rigorous preprocessing (cleaning, tokenization, lemmatization) and a classification of sentiments using the NRC dictionary, the results reveal a predominance of negative sentiments (74%), followed by positive (22%) and neutral (4%). The use of the elbow method to determine the optimal number of clusters confirms the effectiveness of K-Means in structuring and analyzing unannotated text corpora.

In [16], multiclass analysis of sentiment in healthcare texts was performed using a hybrid framework that combined unsupervised and supervised learning techniques. As a first step, data from Twitter and pharmaceutical journals were cleaned by removing punctuation and empty words, and lower-casing. Then, K-Means was used to automatically label the datasets into five categories: treatment and drugs, prevention, symptoms and causes, news, and other. The text was converted into numerical representations by extracting features using TF-IDF. Next, supervised classifiers, such as Logistic Regression, SVM, and Random Forest, and DL models, such as CNN and a CNN-LSTM hybrid, were used to label them. The models were evaluated using accuracy, precision, recall, and F1-score to assess their ability to accurately classify health-related social media content. In [17], Natural Language Processing (NLP) and ML techniques were used to analyze Reddit posts related to depression and suicide. Data were extracted from the "SuicideWatch" subreddit, preprocessed, and used to train several classification models, including SVM, Logistic Regression, Naïve Bayes, and Random Forest. The performance of these models was 77.29% for Logistic Regression, 74.35% for Naïve Bayes, 77.12% for SVM, and 77.30% for Random Forest.

Traditional sentiment analysis methods often reduce a text to a single emotional label, overlooking the internal complexity and variability of the emotions expressed. This study presents an innovative approach to sentiment analysis based on the Brownian motion model [18], a concept rooted in stochastic process theory. In mathematics, Brownian motion is defined as a family of random variables $X(t)$ representing the evolution of a random phenomenon over time. Inspired by its physical interpretation, this model describes the erratic behavior of a particle suspended in a fluid, whose trajectory is shaped by countless random and invisible collisions. This concept is adapted to sentiment analysis by considering a textual comment as an unstable emotional trajectory. Each sentence is treated as a time step assigned an emotional polarity score. The progressive accumulation of these scores allows the construction of an emotional curve that simulates a Brownian motion, capturing the affective variations unfolding throughout the discourse. This model effectively reflects the internal, sometimes contradictory, emotional shifts—frequently observed in the narratives of people experiencing depression. A set of quantitative indicators is extracted from each trajectory, including drift (overall tendency), variance (emotional instability), $Z$-signal changes (changes in emotional direction), trajectory length ($L$), maximum emotional level ($M$), minimum emotional level ($m$), and amplitude (intensity of variation), which provide a fine-grained understanding of intra-comment emotional dynamics. The K-Means clustering algorithm [19] is applied to analyze and structure these emotional profiles, enabling the automatic grouping of comments based on the similarity of their emotional trajectories. Principal Component Analysis (PCA) is employed to reduce the dimensionality of the feature space and to facilitate the interpretation of the resulting emotional clusters. This approach was applied to a corpus of depressive patient comments to identify characteristic emotional profiles and compare affective expressions based on variables such as the type of treatment mentioned.

The objective of this study is twofold: first, to demonstrate the relevance of the Brownian model for enhancing sentiment analysis in complex psychological contexts and second, to propose a tool to support the interpretation of emotional states expressed by depressive patients, with potential applications in remote monitoring, personalized treatment, and early relapse detection.

## II. PROPOSED METHOD

### A. Data Preparation and Sentiment Scoring

The data used in this study were collected from a variety of sources in which patients freely express their experiences of depression and drug treatments, including: specialist health forums, medical discussion platforms (such as 'ask-a-patient', drugs.com, etc.), and posts on social networks, where users spontaneously share their feelings [20]. The initial dataset consisted of 9600 comments. After removing comments that contained fewer than two sentences, a dataset of approximately 8300 comments was obtained. These textual corpora reflect a diversity of points of view, levels of language, and discourse structures, making an in-depth preprocessing process essential.

#### 1) Filtering Out Insufficient Comments

To ensure the reliability of the statistical and emotional analyses, comments containing fewer than two sentences were eliminated, removing comments that are too brief or anecdotal and ensuring that the emotional trajectory contains enough points to reflect significant variation.

#### 2) Syntactic Segmentation

Each comment was segmented into sentences, which constitute the basic units of the emotional trajectory. This segmentation was carried out using automatic NLP tools, such as SpaCy or NLTK, guaranteeing consistent syntactic separation even in informal language.

#### 3) Lexical Clean-up and Standardization

The comments collected often contain non-standard content (abbreviations, special characters, hashtags, emojis). Therefore, a linguistic clean-up was performed to ensure textual uniformity and facilitate sentiment analysis (Figure 1):

- Remove HTML tags, links, special characters, or textual noise,
- Correct typographical anomalies (excessive punctuation, double spaces, etc.),
- Convert all text to lowercase to standardize processing.
- Depending on the needs of the analysis modules, certain stopwords were removed or retained to preserve emotional nuances.

#### 4) Assigning Polarity Scores Sentence by Sentence

The emotions expressed in each comment were then thoroughly examined. Unlike traditional methods that assign an overall score to the comment, this approach is based on a granular analysis, carried out sentence by sentence. For this purpose, TextBlob [18] was used, a Python NLP library that calculates polarity and subjectivity scores for each sentence.

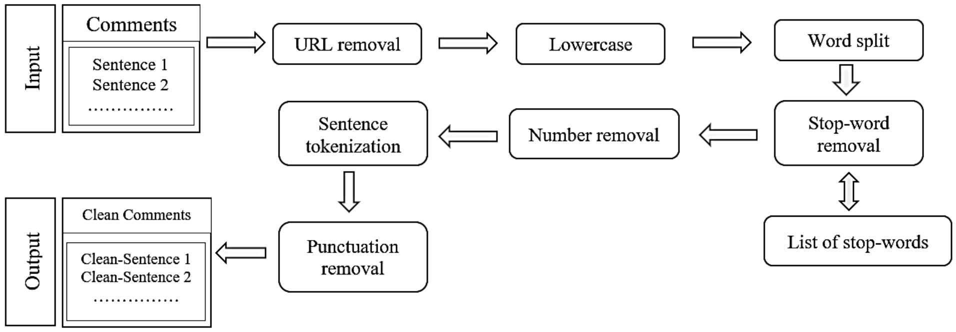


Fig. 1. The main steps in the data processing process.

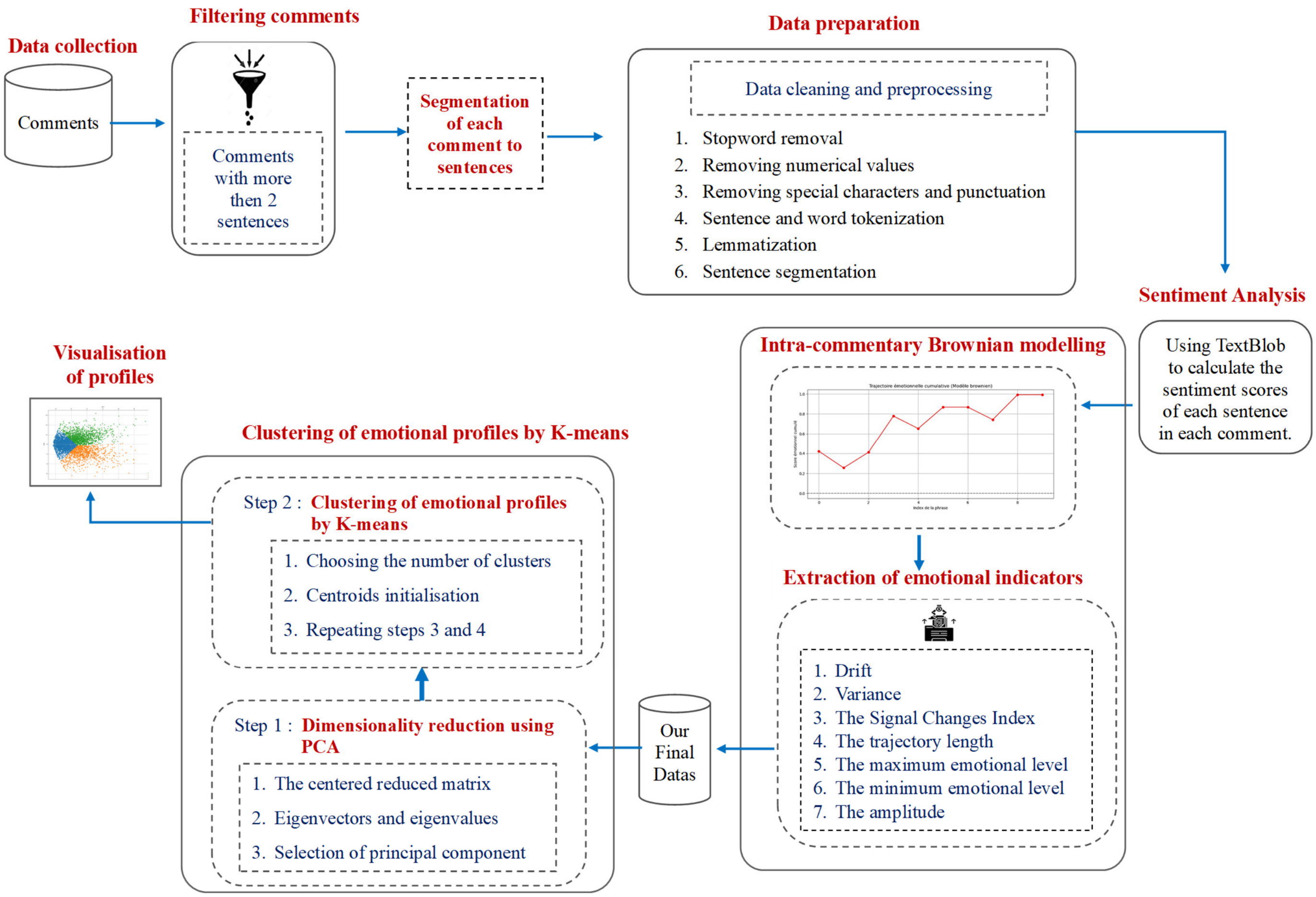


Fig. 2. Flowchart of the proposed method for sentiment analysis based on the Brownian motion principle.

### *B. Brownian Modeling of Emotional Trajectories*

#### *1) Application of the Brownian Motion Principle*

This study presents a novel approach to sentiment analysis by modeling emotional progression in user-generated texts, such as patient comments on their health, using the mathematical framework of Brownian motion. This method interprets the evolution of emotions as a discrete Brownian trajectory, capturing stochastic variations in sentiment across the text. This approach allows capturing the dynamic affective fluctuations expressed in patient comments, beyond a simple static classification (e.g., positive/neutral/negative), as illustrated in Figure 2. Brownian motion, also known as the Wiener process, is a continuous-time stochastic process $B(t)$ with the following characteristics:

- $B(0) = 0$
- $B(t)$ has independent increments.
- $B(t) - B(s) \sim \mathcal{N}\ (0, \sigma^2(t-s))$ for $0 \le s < t$ (1)
- $B(t)$ has continuous but nowhere differentiable paths.

A discrete approximation is used in practical applications.

$$X_k = \sum_{i=1}^{k} \epsilon_i \quad (2)$$

where $\epsilon_i \sim \mathcal{N}(\mu, \sigma^2)$ and $X_k$ represents the cumulative state at time step $k$.

In this adaptation, a textual comment is assimilated to a stochastic trajectory, where each sentence corresponds to a discrete time step and brings an emotional "shock" to the system. In other words, each comment is interpreted as a temporal sequence of emotional states, evolving over the sentences.

This model is used to represent the emotional variations in the text as a discrete Brownian trajectory, which reflects the emotional dynamics of each user. Specifically:

- Each $C$ comment comprises $n$ sentences: $C = \{S_1, \ldots, S_n\}$
- Each sentence $S_i$ is assigned a sentiment score $S_i \in [-1,1]$ using the TextBlob polarity scoring tool:

$$\begin{cases} s_i > 0, & \text{positive sentiment} \\ s_i < 0, & \text{negative sentiment} \\ s_i = 0, & \text{neutral content} \end{cases}$$

The definition of the cumulative emotional trajectory is based on:

$$E_k = \sum_{i=1}^{k} S_i, \text{ for } k = 1, \ldots, n \quad (3)$$

This trajectory is similar to a Brownian trajectory with a drift, where $\mu = \mathbb{E}[S_i]$ represents the average sentiment score and $\sigma^2 = \mathbb{v}ar(S_i)$ is the variance of the sentiment. Thus, $\mathbb{E}[E_k] = k\,\mu$, $\mathbb{v}ar(E_k) = k\sigma^2$.

*2) Extraction of Indicators*

From each emotional pathway, a characteristic vector of seven dimensions $\vec{X}$ is extracted, which captures both the global and local emotional behavior.

- Drift: The derivation in the proposed model consists of measuring the average sentiment score for all the sentences in a comment.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} S_i \quad (4)$$

where $s_i$ is the sentiment score of the $i^{th}$ sentence (from TextBlob, typically in $[-1, 1]$), and $n$ is the number of sentences in the comment.

- Variance ($\sigma^2$) measures the degree of emotional instability in a comment. Its calculation is based on the average of the sentiment scores at the sentence level compared with the overall emotional average (drift). High variance indicates fluctuating or contradictory emotional expression, while low variance indicates emotional stability and consistency throughout the text. It is defined as:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (S_i - \mu)^2 \quad (5)$$

where $S_i$ is the sentiment score of sentence $i$, $\mu$ is the mean sentiment (drift), and $n$ is the number of sentences.

- The $Z$ signal-changes index quantifies the frequency with which the sentiment polarity switches between sentences in a comment, whether it is from positive to negative or vice versa.

$$Z = \sum_{i=2}^{n} 1_{sign(S_i) \neq sign(S_i - 1)} \quad (6)$$

where $S_i$ is the sentiment score of the $i^{th}$ sentence, and is:

$$sign(s) \begin{cases} +1, & \text{if } s > 0 \\ -1, & \text{if } s < 0 \\ 0 & \text{if } s = 0 \end{cases}$$

- The trajectory length $L$ is used to measure the total amount of emotional movement in a comment, regardless of the direction of this movement (positive or negative). A low $L$ index indicates emotional stability with little variation, while a high $L$ index shows dynamic and expressive feelings with frequent emotional changes. It is defined as:

$$L = \sum_{i=2}^{n} |E_i - E_{i-1}| \quad (7)$$

where $E_i$ is the cumulative sentiment after sentence $i$, $|E_i - E_{i-1}|$ is the absolute change in emotion between two consecutive sentences, and $n$ is the number of sentences in the comment.

- The maximum emotional level ($M$) is the highest cumulative emotional point achieved during a comment. $E_k$ is the sentiment score obtained after the $i^{th}$ sentence. Using this measure, the moment when the user expresses the highest emotional positivity can be determined.

$$M = \max_{k} E_k \quad (8)$$

- The minimum emotional level ($m$) represents the lowest point in the emotional curve. It corresponds to the most negative emotional state experienced during the comment's progression, and is identified by:

$$m = \min_{k} E_k \quad (9)$$

- The amplitude expresses the range of the emotional fluctuation in a comment. A high amplitude indicates that the person who wrote the comment experienced or expressed strong emotional variations, for example, strong positivity at one moment and strong negativity at another. A low amplitude indicates that the emotional tone remained close to average and did not vary significantly. The amplitude is calculated as follows:

$$A = \max_{k} E_k \text{ - } \min_{k} E_k \quad (10)$$

where $E_k$ represents the emotional score obtained by combining the sentiment scores for the sentence in the commentary up to step $k$, $\min_{k} E_k$ is the highest emotional point reached, and $\max_{k} E_k$ is the lowest emotional point.

*3) Visualization of a Brownian Emotional Trajectory from a Real Commentary*

Figure 3 and Table I illustrate an example of emotional dynamics in patient comments using the analogy of Brownian motion. It represents a set of randomly scattered dots, symbolizing the emotional phrases expressed throughout the commentary. A red zig-zag path links these dots, representing the unpredictable emotional fluctuations of a depressed patient as he speaks. This emotional path, constructed from the polarity scores associated with each sentence, forms an emotional Brownian trajectory.

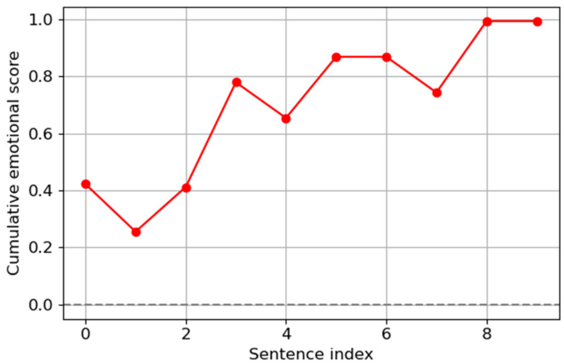


Fig. 3. Cumulative emotional pathway using the proposed Brownian motion model.

TABLE I. EXTRACTED INDICATORS

| | |
|---|---|
| Drift | 0.09927579 |
| Variance | 0.04060194 |
| Amplitude | 0.73720238 |
| Trajectory length | 1.403869048 |
| Max | 0.99275794 |
| Min | 0.25555556 |
| Sign changes | 8 |

## III. RESULTS AND DISCUSSIONS

### *A. Exploring Interrelationships Between the Extracted Features*

This stage aimed to quantify the emotional dynamics of each comment based on its cumulative trajectory by calculating a set of numerical descriptors known as emotional indicators. These synthetic variables reflect different dimensions of the affective variation perceived throughout the text.

The following seven emotional indicators were extracted for each trajectory: drift, variance, amplitude, trajectory length, maximum value, minimum value, and the number of sign changes. The correlation matrix (Figure 4) highlights several key structural relationships between these indicators derived from Brownian-based modeling of sentiment dynamics. There was an almost perfect correlation between the amplitude and the trajectory length (0.90), showing that they both capture the overall intensity of emotional fluctuations. The correlation between variance and amplitude (0.44) and trajectory length (0.46) is only moderate, suggesting that it captures a distinct dimension of the variability that is related to the sentiment dispersion around the drift instead of the amplitude of extreme deviations. The drift is highly correlated with both the maximum (0.73) and the minimum (0.75) values, which validates that the dominant polarity of a comment indirectly influences the emotional peaks and troughs reached in a trajectory.

It is observed that emotionally unstable trajectories, characterized by numerous polarity reversals, tend to be longer and more variable, with a moderate correlation between the number of sign changes and trajectory length (0.70). These observations demonstrate the internal coherence of the extracted indicators and highlight the multidimensional complexity of emotional expression in patient narratives.

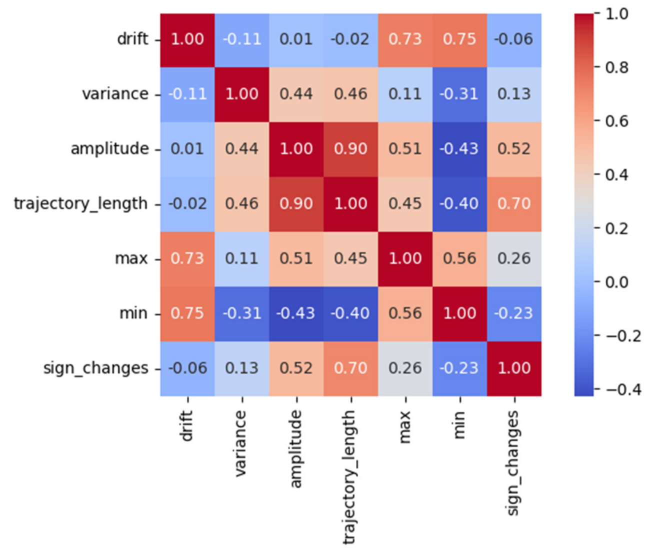


Fig. 4. Correlation matrix of initial variables.

### *B. Reducing Dimensionality and Emotional Profiling*

The Kaiser criterion was applied to determine the optimal number of principal components to retain in the analysis, retaining only those components with an eigenvalue strictly greater than 1, i.e., those that explain more variance than a standardized initial variable. Applied to the seven emotional indicators extracted from the Brownian trajectories, the spectral analysis revealed that only the first two principal components (PC1=3.04, PC2=2.38) satisfy this criterion, as shown in Table II. Their selection allows effective dimensional reduction while capturing the essence of the overall structure of the data.

TABLE II. EIGENVALUES ASSOCIATED WITH EACH PRINCIPAL COMPONENT

| Component | Eigenvalues |
|---|---|
| PC1 | 3.0416 |
| PC2 | 2.3811 |
| PC3 | 0.8715 |
| PC4 | 0.4453 |
| PC5 | 0.1907 |
| PC6 | 0.0710 |
| PC7 | 0.0000 |

However, despite this overall consistency, the correlation matrix shows that the variance is only weakly to moderately correlated with the other indicators (maximum correlation of 0.46), which means that it is not adequately represented by the first two principal components. Consequently, variance appears as a distinct dimension, reflecting local emotional stability, and independent of both overall intensity (amplitude, length) and dominant polarity (drift, extremes). This justifies the inclusion of a third principal component (PC3) to capture this residual variance and ensure more reliable and exhaustive modeling of intra-commentary emotional dynamics.

### *C. Visualization and Interpretation of Emotional Profiles*

The unsupervised learning process does not require labelled data and can be solved using clustering methods. Clustering is a typical illustration of unsupervised learning, where visual classifications can be found that match the hypothesis. Clustering aims to identify similarities, regardless of the type of data. The K-means algorithm was employed to identify similar emotional profiles, using three main steps iteratively:

- Random initialization of the centroids of the $k$ clusters.
- Assignment of each observation to the nearest centroid according to Euclidean distance.
- Updating the centroids by averaging the vectors of each cluster.

These steps are repeated until convergence is achieved. The objective of K-means is to reduce the sum of squared distances between intra-clusters, as measured by:

$$\min_{S} E(\mu_i) = \sum_{i=1}^{k} \sum_{x_j \in S_i} || x_j - \mu_i ||^2 \quad (11)$$

where $S = \{S_1, S_2, \ldots, S_k\}$ indicates the partition of the set of emotional vectors and $\mu_i$ is the centroid of the cluster $S_i$. Each centroid is updated by imposing the extreme condition on the quadratic function $E(\mu_i)$, which gives:

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \quad (12)$$

This method, which is both efficient and rapid, made it possible to group the 8,700 Brownian trajectories into $k = 3$ distinct emotional profiles (Figure 6). These clusters reflect recurring emotional dynamics among patients, such as profiles with stable drift, intense oscillations, or extreme variations. The choice of $k$ was guided by the elbow method, ensuring an optimal balance between intra-cluster cohesion and inter-cluster separation. The correlation circle (Figure 5) was used to better understand and interpret the positioning of comments in the reduced-dimensional space, highlighting the contribution of each initial emotional indicator to the principal components and facilitating the interpretation of comment distribution in relation to these variables.

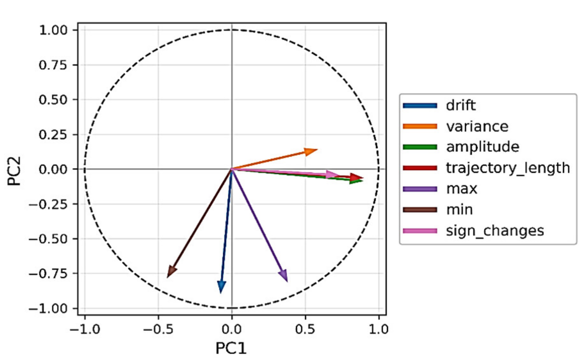


Fig. 5. 2D circle of correlations.

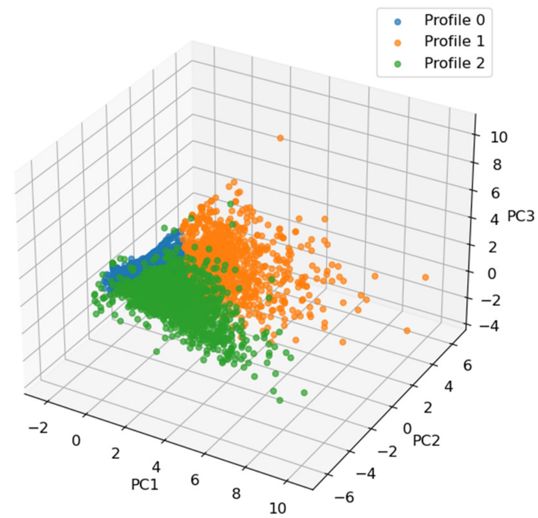


Fig. 6. 3D K-means clustering of sentiment profiles.

### *D. Word Cloud for Each Profile*

The first cluster is characterized by strongly positive words such as "good", "better", "great", "happy", "effective", "wonderful", "best", "glad", often combined with intensifiers ("really", "completely", "very"). The frequent use of "experience", "success", "positive", or "love" shows that patients in this group generally describe a significant improvement and emotional satisfaction with the treatment. The sentimental Brownian trajectory of these comments is ascending, with an optimistic dynamic dominated by a stable positive polarity.

The word cloud in the second cluster is heavily dominated by intensely negative words such as "horrible", "terrible", "worst", "awful", "sick", "bad", "tired", and "crying". There are also adjectives evoking psychological deterioration ("anxious", "angry", "crazy") and negative physical indicators ("cold", "ill", "painful"). This cluster reflects a general feeling of rejection of treatment, high dissatisfaction, or persistent suffering. It represents a profile of patients who express descending emotional trajectories, unstable, and critical of their therapeutic experience.

The word cloud in the third cluster shows a strong presence of neutral to slightly positive terms such as "better", "able", "normal", "fine", "effective", and "worth", often accompanied by modifiers such as "really", "slightly", and "pretty". The recurrence of the words "first", "due", "past", or "slowly" suggests nuanced and contextualized narratives, where the emotional evaluation evolves over the course of the commentary. This profile corresponds to a moderate emotional dynamic, marked by partial improvements or relative acceptance of the effects of treatment.

## IV. CONCLUSION

This study proposed a stochastic modeling framework inspired by Brownian motion, where each patient comment is represented as an emotional trajectory built from sentence-level polarity scores. Unlike traditional sentiment analysis approaches, which rely on global polarity labels and thus miss temporal variability, the proposed model captures the dynamics of emotional expression through seven quantitative indicators: drift, variance, amplitude, path length, maximum, minimum, and number of sign changes. Experiments on the Depression DataSet [20] demonstrated that this fine-grained modeling leads to clearer differentiation between varying levels of depressive narratives. Specifically, comments reflecting more severe depressive states were found to be characterized by higher emotional variance, more frequent changes in polarity signs, and longer path lengths—patterns that are not detected by classical sentiment classification. These findings provide empirical evidence that emotional instability and ambivalence, often described in clinical psychology, can be computationally observed in textual data, and that the degree of such fluctuations may serve as an indicator of depression severity.

Applying PCA and K-means clustering to vectorized emotional trajectories allowed us to uncover latent affective structures within the dataset. Specifically, clusters were identified, characterized either by oscillatory emotional patterns or by monotonic negative drift. These configurations correspond to established distinctions in psychopathology and point to potential associations between textual emotional dynamics and clinical states. In contrast to conventional sentiment analysis approaches, which rely primarily on static polarity scores, this method provides a multidimensional representation of affective expression, thereby enabling the identification of subtle emotional signals that are often overlooked by traditional techniques. In summary, the proposed Brownian-motion-based modeling not only enriches the computational analysis of depressive discourse but also contributes a novel methodological perspective to the intersection of NLP and medical sciences. It provides evidence that intra-comment emotional variability is a valuable marker for understanding patient experiences and may serve as a foundation for clinical decision-support tools aimed at monitoring treatment progress or detecting early warning signs of deterioration.

## REFERENCES


[1] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, https://doi.org/10.1023/A:1010933404324.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, https://doi.org/10.1007/BF00994018.

[3] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, https://doi.org/10.1162/neco.1997.9.8.1735.

[4] K. D. S. Devi, V. Sireesha, C. Sudha, M. Ravisankar, and P. D. K. Reddy, "A Novel Approach to Sentiment Analysis using GMM-Enhanced N-gram LSTM Networks," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23068–23073, Jun. 2025, https://doi.org/10.48084/etasr.10640.

[5] M. El-Badaoui, N. Gherabi, F. Quanouni, and M. Amnai, "A Novel Ontology-Based Approach for Analyzing Patient Sentiment Regarding Chronic Diseases," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 21, no. 08, pp. 155–170, Jun. 2025, https://doi.org/10.3991/ijoe.v21i08.53767.

[6] Z. Sabouri, N. Gherabi, M. Nasri, M. Amnai, H. El Massari, and I. Moustati, "Prediction of Depression via Supervised Learning Models: Performance Comparison and Analysis," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 09, pp. 93–107, Jul. 2023, https://doi.org/10.3991/ijoe.v19i09.39823.

[7] Y. Madani, M. Erritali, J. Bengourram, and F. Sailhan, "A multilingual fuzzy approach for classifying Twitter data using fuzzy logic and semantic similarity," *Neural Computing and Applications*, vol. 32, no. 12, pp. 8655–8673, Jun. 2020, https://doi.org/10.1007/s00521-019-04357-9.

[8] Z. Putri, S. Sugiyarto, and S. Salafudin, "Sentiment analysis using fuzzy naïve bayes classifier on covid-19," *Desimal: Jurnal Matematika*, vol. 4, no. 2, pp. 193–202, Jul. 2021, https://doi.org/10.24042/djm.v4i2.7390.

[9] T. Nijhawan, G. Attigeri, and T. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions," *Journal of Big Data*, vol. 9, no. 1, Dec. 2022, Art. no. 33, https://doi.org/10.1186/s40537-022-00575-6.

[10] A. Balakrishnan, S. M. Idicula, and J. Jones, "Deep learning based analysis of sentiment dynamics in online cancer community forums: An experience," *Health Informatics Journal*, vol. 27, no. 2, Apr. 2021, Art. no. 14604582211007537, https://doi.org/10.1177/14604582211007537.

[11] D. C. Edara, L. P. Vanukuri, V. Sistla, and V. K. K. Kolli, "Sentiment analysis and text categorization of cancer medical records with LSTM," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 5, pp. 5309–5325, May 2023, https://doi.org/10.1007/s12652-019-01399-8.

[12] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023, https://doi.org/10.14569/IJACSA.2023.0140813.

[13] E. R. Kumar and K. V. S. N. Rama, "Sentiment Analysis using Social and Topic Context for Suicide Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021, https://doi.org/10.14569/IJACSA.2021.0120249.

[14] S. D. Patravali and S. P. Algur, "COVID-19 Sentiment Analysis using K-Means and DBSCAN," *International Journal of Emerging Science and Engineering*, vol. 11, no. 12, pp. 12–17, Nov. 2023, https://doi.org/10.35940/ijese.L2558.11111223.

[15] O. Iparraguirre-Villanueva, V. Guevara-Ponce, F. Sierra-Linan, S. Beltozar-Clemente, and M. Cabanillas-Carbonell, "Sentiment Analysis of Tweets using Unsupervised Learning Techniques and the K-Means Algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, 2022, https://doi.org/10.14569/IJACSA.2022.0130669.

[16] A. Kumar and S. Shekhar, "Hybrid model of unsupervised and supervised learning for multiclass sentiment analysis based on users' reviews on healthcare web forums," *Journal of Autonomous Intelligence*, vol. 7, no. 4, Jan. 2024, https://doi.org/10.32629/jai.v7i4.971.

[17] P. Jain, K. Ram Srinivas, and A. Vichare, "Depression and Suicide Analysis Using Machine Learning and NLP," *Journal of Physics: Conference Series*, vol. 2161, no. 1, Jan. 2022, Art. no. 012034, https://doi.org/10.1088/1742-6596/2161/1/012034.

[18] N. Wiener, "Differential-space," *Journal of Mathematics and Physics*, vol. 2, no. 1–4, pp. 131–174, 1923.

[19] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, https://doi.org/10.1109/TIT.1982.1056489.

[20] "Mariabadaoui/Depression_DataSet." Sep. 05, 2025, [Online]. Available: https://github.com/Mariabadaoui/Depression_DataSet.