Available Online: 2 April 2025 DOI: 10.54254/2977-3903/2025.21918

A comprehensive review of probabilistic and statistical methods in social network sentiment analysis

Chenyiqiu Zhen

University College London, London, United Kingdom

chenyiqiu.zheng.23@ucl.ac.uk

Abstract. In the era of rapid digital transformation, social networks generate huge amounts of textual data every day, making sentiment analysis an essential tool for understanding public opinion. This study focuses on the application of probabilistic and statistical methods to sentiment analysis in social networks, highlighting their effectiveness in dealing with uncertainty and modeling the distribution of emotions. The main objective is to evaluate the role of Naïve Bayesian (NB), Hidden Markov Models (HMMs), and Bayesian networks in emotion classification, emotion propagation, and dynamic emotion tracking. Through literature review and comparative analysis, this study examines the existing research, computational efficiency, and real-world applications of probabilistic classification models. The results show that Naive Bayes is computationally efficient and effective for large-scale emotion classification, while HMM and Bayesian networks excel in sequential emotion prediction and user behavior modeling. The study highlights the advantages of probabilistic methods in sentiment analysis, while acknowledging their limitations, such as their reliance on probabilistic assumptions and the challenges of capturing deep contextual semantics. Future research should explore hybrid approaches that combine probabilistic models with deep learning techniques to improve the predictive performance and scalability of real-time sentiment analysis.

Keywords: sentiment analysis, Naïve Bayes, Hidden Markov Models, Latent Dirichlet Allocation

1. Introduction

Sentiment analysis is a subfield in Natural Language Processing (NLP), studies on identifying and categorizing emotions, attitudes, and opinions expressed in textual data [1]. It is rapidly evolving and presently finds applications in several areas: from academic research to government policy analysis, business intelligence, and organizational decision-making [2]. Meanwhile, with the exponential proliferation of social network sites like Twitter, Facebook, and Reddit, great amounts of information and data are generated in a single day, which give great insight into public opinion and sentiment on an issue [3]. Ensuring its analysts and interpretations can be as effective as possible has important ramifications for trend predictions, market analytics, political discussions, and even crisis management. It achieves these goals by embedding a wide range of mathematical methods into its development, hence providing a theoretical and computational basis for many techniques. The methods discussed will not only increase the efficiency and accuracy of the sentiment analysis but also widen the scope of varied and complex data sets such as those created from social media platforms.

A range of mathematical fields has contributed to sentiment analysis. Linear algebra plays a foundational role through methods like word embeddings (e.g., Word2Vec, GloVe), which transform text into vector representation, and matrix factorization techniques (e.g., PCA, SVD), which reduce data dimensionality while preserving key features [4, 5]. Probability theory introduces approaches like Naïve Bayes and Bayesian inference for text classification and sentiment intensity modeling, as well as probabilistic graphical models and Hidden Markov Models (HMM) for capturing sequential patterns and sentiment propagation [6]. Statistics brings feature selection techniques like Chi-Square tests and Information Gain, regression analysis for continuous sentiment scoring, and robust model evaluation metrics such as cross-validation and hypothesis testing [7]. Beyond these, graph theory has been employed for social network sentiment analysis through graph embeddings and Graph Convolutional Networks (GCNs), enabling the modeling of user interactions and sentiment propagation [8]. Optimization methods ensure effective parameter tuning in machine learning models, with approaches like Genetic Algorithms (GAs) and convex programming optimizing hyperparameters and feature selection [9]. Tensor analysis supports multimodal sentiment analysis by integrating text, image, and audio data [10]. In addition, differential equations for temporal modeling and topological data analysis for mining data emerged, further broadening the mathematical toolkit [11, 12].

Although each of these mathematical methods has helped develop sentiment analysis, this paper will be restricted to probabilistic and statistical methods. These have gained widespread usage because of their robustness in handling uncertainty and modeling sentiment distribution effectively. By providing a comprehensive overview of probabilistic and statistical methods in sentiment analysis, this paper aims to summarize their current applications, challenges, and potential future directions. This study provides valuable insights into the efficiency and applicability of probabilistic models in sentiment analysis. Its findings provide valuable insights for business, policymaking, and AI development, while also serving as a foundation for future research integrating probabilistic and deep learning approaches.

2. Probabilistic methods in sentiment analysis

As a crucial field and branch of mathematics, probability theory has been providing a strong foundation for sentiment analysis, providing robust and reliable models for text uncertainty and helping make predictions. Probability theory offers solutions to text classification, temporal emotion modeling and user behavior analysis. These probabilistic techniques are widely used for handling noisy, large-scale sentiment data, particularly in domains such as social media monitoring, customer feedback analysis, and opinion mining. The process of sentiment analysis is generally divided into aspect term extraction, aspect classification detection, opinion term extraction and aspect emotion classification. Probabilistic methods are particularly good at capturing the dependencies and variations of entities, which effectively improves the accuracy of aspect sentiment classification. This section explores key probabilistic methods in sentiment analysis, including classification models, topic modeling, sentiment propagation, sentiment intensity modeling, and sequence prediction techniques.

2.1. Probabilistic models for sentiment classification

Probabilistic classification is an important component of sentiment analysis, providing a structured approach to classify the text based on the likelihood of belongings to given sentiment classes. It predicts a probability distribution over a set of classes based on observed language. It has the benefits of low cost of implement and fast in computation [13]. Among these models, Naïve Bayes (NB) and Hidden Markov Models (HMMs) are two of the most commonly used approaches.

2.1.1. Naïve Bayes (NB)

Naïve Bayes is a simple yet effective probabilistic classifier. It is a supervised machine learning approach, which operates under the assumption of conditional independence between features given the sentiment label in the same class [13]. Based on Bayes Theorem, it determines the probability of a specific event to have occurred providing other particular events happened [14]. Despite this assumption being a strong simplification, Naïve Bayes remains widely used in sentiment classification due to its computational efficiency scalability. It could be applied to both single and multiclass tasks, also small datasets and massive data records [15].

According to the Naïve Bayes theorem:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{1}$$

where $P(H|X)$ is the probability of hypothesis H being true given a specific event X happened (also known as posterior probability), $P(X|H)$ is the probability of event X given hypothesis H is true (or the likelihood probability), $P(H)$ is the probability of hypothesis H is true (or the prior probability), $P(X)$ is the probability of event X (or marginal probability).

In the situation of sentiment classifier, hypothesis H is sentiments, event E is aspects or features of text. As the features of a category could be many, say $B_1, B_2, \dots, B_n$, the formula turns into:

$$P(A|B_1, B_2, \dots, B_n) = \frac{P(B_1|A)P(B_2|A)\dots P(B_n|A)P(A)}{P(B_1)P(B_2)\dots P(B_n)} = \frac{P(A)\prod_{i=1}^{n} P(B_i|A)}{\prod_{i=1}^{n} P(B_i)} \tag{2}$$

As $\prod_{i=1}^{n} P(B_i)$ is constant, the formula could also be written as:

$$P(A|B_1, B_2, \dots, B_n) \propto P(A)\prod_{i=1}^{n} P(B_i|A) \tag{3}$$

As a sentiment classification model, the purpose is to find all possible probabilities for A and make the maximum value as the output:

$$A = argmax_A P(A)\prod_{i=1}^{n} P(B_i|A) \tag{4}$$

There are several variations for Naïve Bayes, including Multinomial Naive Bayes, Bernoulli Naive Bayes, Gaussian Naive Bayes, Complement Naive Bayes, and Categorical Naive Bayes.

Multinomial Naïve Bayes is a frequency-depended model focused on multinomially partitioned data. As Naïve Bayes is based on conditional probabilities while multinomial Naïve Bayes based on multinomial distribution:

$$f(x) = \frac{n!}{n_1!n_2!n_3!\ldots n_m!} {p_1}^{n_1} {p_2}^{n_2} {p_3}^{n_3} \ldots {p_m}^{n_m} \tag{5}$$

where n is the total number of occurrences of all events, n_i represents the number of occurrences of outcome i, p_i represents the probability that the outcome i occurs.

It is very effective when dealing discrete and discontinuous features. It has calculation as:

$$P(x_i|c_k) = \frac{count(x_i|c_k)+\alpha}{(\sum_{x\in V} count(x|c_k))+\alpha|V|} \tag{6}$$

where $P(x_i|c_k)$ refers to the probability of the word x_i in the text is in class k, $count(x_i|c_k)$ refers to the number of times that the word x_i appears in the text, α is the smoothing parameter, $\sum_{x\in V} count(x|c_k)$ is the counts of all words in class k, $|V|$ is the number of unique words in the vocabulary.

As Multinomial Naïve Bayes is related to frequency, Bernoulli Naïve Bayes instead efficiently deal with problems of binary concept. It assumes that the data is discrete and distributes as Bernoulli mode, focusing on the appearance or absence of the term in the file under specified consideration. Bernoulli distribution is:

$$f(x) = \begin{cases} p^x(1-p)^{1-x} & if\ x = 0{,}1 \\ 0 & otherwise \end{cases} \tag{7}$$

where p and $(1-p)$ are the probability of $x = 0$ and $x = 1$, respectively.

$$P(x_i|c) = P(x_i = 1|c)x_i + (1 - P(x_i = 1|c))(1 - x_i) \tag{8}$$

Gaussian Naive Bayes assumes that the data are all normally distributed. Then the Gaussian NB has formula:

$$P(x_i|c) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{9}$$

with mean μ and standard deviation σ.

Complement Naive Bayes is an adaptation of Multinomial Naïve Bayes. Complement Naïve Bayes focuses on computing the probability of a word not belonging to each class and adjusting accordingly to handle imbalanced data. First, calculate the probabilities if the word not belong to each class; Second, choose the minimum value along all classes. Therefore, the classification formula turns to be:

$$A = argmin_A P(A) \prod_{i=1}^{n} P\left(B_i\middle|\hat{A}\right) \tag{10}$$

where $\hat{A}$ refers to the situation when not belonging to class A.

Categorical Naïve Bayes assumes that each feature (word) has its own categorical distribution.

$$P(x_i = x|c_k) = \frac{count(x_i = x|c_k)+\alpha}{(\sum_{x'\in V} count\left(x_i = x'\middle|c_k\right))+\alpha|V|} \tag{11}$$

Naïve Bayes models have been extensively used in sentiment analysis due to their efficiency in handling text-based data and probabilistic nature. Naïve Bayes classifiers help in extracting subjective opinions from text, such as identifying whether the sentiment polarity of a text is positive, negative, or neutral. NB models could be used to predict user preferences in sentiment-based recommendation systems; determine sentiment in aspect-based sentiment analysis; track public sentiment in social media to help make business or organization decisions; support development of chatbot so that it could interpret and respond to customer queries.

2.1.2. Hidden Markov Models (HMM)

Hidden Markov Models (HMM), which is also known as a variation of dynamic Bayesian Network, is a modeling technique connecting with statistics. It allows for describing dependencies between different variables using an undirected graph structure that encodes conditional probability distributions [16]. It is easy to be implemented, good at handling sequential data and variable-length inputs. Thus, HMM is frequently applicated in domains such as speech recognition, facial expression recognition, gene prediction, gesture recognition, musical composition and Bio-informatics [6].

HMM is an amplifying of Markov Chain which is a model offering probabilistic information about a sequence of states of a random variable, each of which can take a value of some sets. This is useful when events are observable, if the events are not, HMMs are helpful. HMM is a quintuple (S, O, A, B, π). It consists of N not directly observable hidden states S:

$$S = \{S_1, S_2, S_3, \ldots, S_N\} \tag{12}$$

Then there is M observable symbols per state:

$$O(t) = \{O_1, O_2, O_3, \ldots, O_M\} \tag{13}$$

A is the state transition probabilities matrix denoted by $A = \{a_{ij}\}$, where

$$a_{ij} = P\left(X_{t+1} = S_j \middle| X_t = S_i\right) \tag{14}$$

with $1 \leq i, j \leq N$, represents the transition probability of moving from state i at time t to state j at time $t+1$. It could also be written as $a\left[S_i, S_j\right]$.

B is the observation probabilities sequence denoted by $B = \{b_j(t)\}$, where

$$b_j(t) = P\left(O(t) \middle| X(t) = S_j\right) \tag{15}$$

with $1 \leq j \leq N$, represents the probability of emitting $O(t)$ from state S_j.

$$\pi = \{\pi_i = P(X_1 = S_i) | 1 \leq i \leq N\} \tag{16}$$

is a vector of the initial state probabilities referring to the initial state distribution, with $0 \leq \pi_i \leq 1$ and $\sum_i \pi_i = 1$.

HMM could also be written as $\theta = (\pi, A, B)$. It assumes that each observation or symbol is emitted by a hidden state, which produces the probability matrix.

While HMM has lots of advantages, there are also disadvantages. As artificial neutral network can retain a cumulative history of its transitions, HMM does not retain the label of the previous state over time, relying only on the value of the previous state. Also, if extending to more general scale, HMM is linear and shallow [17].

Thus, in the need of prediction based on several successive previous data, there is high-order HMM. If the distribution is extended to the previous n observations, the Markov Chain is an nth-order. The transition probability is now $P(x_k | x_{k-1}, x_{k-2}, \dots, x_{k-n+1}, x_{k-n})$.

2.2. Topic modeling with probabilistic methods

Probabilistic topic model is the most widely developed and applied model in various research fields, and has achieved good results in various application fields, especially in text classification and information retrieval. Latent Dirichlet Allocation (LDA) is one of the topic models which is designed to improve other topic models such as Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Indexing (PLSA). It is a generative probabilistic model of a corpus [18]. Thus, it has advantages such as stability and comprehension when interpreting latent semantic [19].

LDA models a collection of D documents using the following probabilistic generative process:

Each document d with N words is associated with a multinomial distribution over K topics, drawn from a Dirichlet prior:

$$\theta_d \sim Dirichlet(\alpha) \tag{17}$$

where θ_d represents the topic proportions for document d, α is a hyperparameter that controls the sparsity of topic.

Each topic is associated with a multinomial distribution over $|V|$ words in vocabulary, drawn from a Dirichlet prior:

$$\beta_k \sim Dirichlet(\eta) \tag{18}$$

where β_k represents the word probabilities for topic k, η is a hyperparameter that controls the smoothness of word distribution.

For each word w_n in document d, draw a topic assignment z_n from the document's topic distribution:

$$z_n \sim Multinomial(\theta_d) \tag{19}$$

with $z_n \in [1,2, \dots, K]$, and draw a word from the topic-specific word distribution:

$$w_n \sim Multinomial(\beta_k) \tag{20}$$

$\{\alpha, \beta\}$ are estimated by maximizing the marginal probability of document d over latent variables $\{\theta, z\}$:

$$p(d|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta)\right) d\theta \tag{21}$$

Since exact inference is difficult to be controlled, LDA employs Variance Inference (VB-EM) or Collapsed Gibbs Sampling to approximate the later distribution of latent variables [18]. LDA scales efficiently with Online LDA and Parallel Gibbs Sampling, making it well-suited for sentiment analysis applications such as opinion mining, aspect-based classification, and trend detection [20].

2.3. Sentiment propagation and user modeling

DBN uses a set of random variables to represent a hidden state, while HMM and LDA use a single random variable to do so. It also uses factorized or distributed methods to represent observations and uses graphical models to represent conditional independencies between these variables. When facing temporal data i.e. data generated sequentially by some causal process, directed graphical models are useful. Dynamic Bayesian Networks (NBW), which is an extension of Bayesian Networks and a generalization of HMM, is a very popular directed graphical model due to its intelligibility. DBN consists of a series of time intervals that represent the state of all variables at a given time, thus representing the evolution of a process over time [21]. In sentiment analysis, a DBN is used to model the temporal variation of sentiments and relationships between topics [22].

DBN models a topic i at time t associating with a distribution which is conditional on the values of a set of parents topic $Pa(i)$ at previous time-steps, for example time $t-1$. For topics $i = 1, 2, \dots, P$ at times $t = 1, 2, \dots, T$, set sentiment on topic i at time t as X_t^i.

$$p\left(X_t^i \middle| X_{t-1}, \theta\right) = f\left(X_{t-1}^{Pa(i)}, \theta\right) \tag{22}$$

where X_{t-1} is the sentiment of all topics at time $t-1$, f is a function of the parents of variable i at time $t-1$ with a set of parameters θ.

3. Applications and future development

3.1. Social media sentiment analysis

Sentiment analysis in social media is widely used in many fields, such as disaster relief, gathering and analyzing public opinion of international events, predicting disease outbreaks, and making strategic decisions in enterprises. Among them, the application of probability theory cannot be ignored. Hidden Markov Models (HMM) and Dynamic Bayesian Networks (DBN) model the temporal evolution of emotions, making it possible for applications such as detecting mood changes due to virus events or policy changes.

3.2. Multimodal sentiment analysis

As a more complex technology, multimodal sentiment analysis not only considers the data source of a single media, but also combines multi-modal text, audio and video to comprehensively analyze the emotion of objects. Bayesian Networks (BNs) are used to simulate the interactions between different modes to improve sentiment classification in noisy environments. Latent Dirichlet Allocation (LDA) is a probabilistic topic model that helps reveal emotion-related topics across modes, providing a deeper level of interpretability.

3.3. Dynamics sentiment analysis

Dynamic sentiment analysis is a technique that focuses on emotions that change over time. DBN and HMM are commonly used to track emotional trends, allowing probabilistic predictions of future emotional states. In addition, Kalman filters and particle filters, as Bayesian sequence estimation methods, can help improve emotion prediction in uncertain and changing environments. The Hierarchical Dirichlet Process (HDP) extends thematic modeling methods to capture changing emotional themes, making it useful for analyzing long-term emotional trends in public discourse and news media.

4. Conclusion

This paper primarily explores the application of probabilistic classification models, such as Naïve Bayes and Hidden Markov Models, in sentiment analysis, evaluating their effectiveness, computational efficiency, and real-world applicability in handling large-scale textual data. Probabilistic methods play a crucial role in sentiment analysis when proceeding with sentiment classification, temporal sentiment analysis, topic modeling, etc. From traditional Naïve Bayes classifiers to sophisticated Bayesian graphical models, these techniques continue to develop and improve along with evolving in deep learning and multimodal analysis. Although primarily including probabilistic methods, this paper does not extensively explore how these probabilistic models perform when integrated with deep learning approaches, such as combining Bayesian inference with transformer-based models like BERT or GPT because of the limitation of the length. This will help the uncertainty estimation in sentiment classification. More precise and flexible methods may be proposed in the future for adapting multimodal sentiment analysis and targeting more fine-grained emotions, among others.

References

[1] Ahongsangbam, D., & Haobam, M. D. (2025). A literature review on sentiment analysis. In M. Saraswat & R. Kumari (Eds.), *Applied Intelligence and Computing* (pp. 303–312). SCRS. https://doi.org/10.56155/978-81-955020-9-7-29

[2] Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions.* Cambridge University Press. https://doi.org/10.1017/CBO9781139084789

[3] Ruscica, G., Tucci, G., & Carneiro, B. (2025). TeleCatch: An open-access software for visualizing, filtering and extracting Telegram messages data. *Software Impacts, 23*, 100736.

[4] Rakshit, P., & Sarkar, A. (2025). A supervised deep learning-based sentiment analysis by the implementation of Word2Vec and GloVe embedding techniques. *Multimedia Tools and Applications, 84*, 979–1012. https://doi.org/10.1007/s11042-024-19045-7

[5] Zainuddin, N., Selamat, A., & Ibrahim, R. (2018). Hybrid sentiment classification on twitter aspect-based sentiment analysis. *Applied Intelligence, 48*, 1218–1232.

[6] Mor, B., Garhwal, S., & Kumar, A. (2021). A systematic review of hidden Markov models and their applications. *Archives of Computational Methods in Engineering, 28*, 1429–1448.

[7] Alshaer, H.N., Otair, M.A., Abualigah, L., Alshinwan, M., & Khasawneh, A.M. (2021). Feature selection method using improved CHI Square on Arabic text classifiers: analysis and application. *Multimedia Tools and Applications, 80*, 10373–10390.

[8] Phan, H. T., Nguyen, N. T., & Hwang, D. (2023). Aspect-level sentiment analysis: A survey of graph convolutional network methods. *Information Fusion, 91*, 149–172.

[9] Govindarajan, M. (2013). Sentiment analysis of movie reviews using hybrid method of naive bayes and genetic algorithm. *International Journal of Advanced Computer Research, 3*(4), 139.

[10] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1103–1114). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1115

[11] Tu, H. T., Phan, T. T., & Nguyen, K. P. (2022). Modeling information diffusion in social networks with ordinary linear differential equations. *Information Sciences, 593*, 614–636. https://doi.org/10.1016/j.ins.2022.01.063

[12] Almgren, K., Kim, M., & Lee, J. (2017). Mining social media data using topological data analysis. In *2017 IEEE International Conference on Information Reuse and Integration (IRI)* (pp. 144–153). IEEE. https://doi.org/10.1109/IRI.2017.41

[13] Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems, 226*, 107–134. https://doi.org/10.1016/j.knosys.2021.107134

[14] Ressan, M. B., & Hassan, R. F. (2022). Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets. *Indonesian Journal of Electrical Engineering and Computer Science, 28*(1), 375.

[15] Danyal, M. M., Khan, S. S., Khan, M., Ghaffar, M. B., Khan, B., & Arshad, M. (2023). Sentiment Analysis Based on Performance of Linear Support Vector Machine and Multinomial Naïve Bayes Using Movie Reviews with Baseline Techniques. *Journal on Big Data, 5*(1), 1–18. https://doi.org/10.32604/jbd.2023.041319

[16] Odumuyiwa, V., & Osisiogu, U. (2019). A systematic review on Hidden Markov Models for sentiment analysis. In *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICECCO48375.2019.9043297

[17] Perikos, I., Kardakis, S., & Hatzilygeroudis, I. (2021). Sentiment analysis using novel and interpretable architectures of Hidden Markov Models. *Knowledge-Based Systems, 229*, 107332. https://doi.org/10.1016/j.knosys.2021.107332

[18] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://doi.org/10.5555/944919.944937

[19] Putri, I. R., & Kusumaningrum, R. (2017). Latent Dirichlet allocation (LDA) for sentiment analysis toward tourism review in Indonesia. *Journal of Physics: Conference Series, 801*(1). IOP Publishing.

[20] Chien, J. T., Lee, C. H., & Tan, Z. H. (2018). Latent Dirichlet mixture model. *Neurocomputing, 278*, 12–22.

[21] Shiguihara, P., Lopes, A. D. A., & Mauricio, D. (2021). Dynamic Bayesian network modeling, learning, and inference: A survey. *IEEE Access, 9*, 117639–117648.

[22] Liang, H., Ganeshbabu, U., & Thorne, T. (2020). A dynamic Bayesian network approach for analysing topic-sentiment evolution. *IEEE Access, 8*, 54164–54174.