Pure and Applied Mathematics Journal
2025, Vol. 14, No. 5, pp. 114-119
https://doi.org/10.11648/j.pamj.20251405.12

Research Article

Machine Learning (ML) and Artificial Intelligence (AI) Approaches to Unstructured Data

Farha Khan[1, *], Pratima Ojha[1], Ghizal Firdous Ansari[2]

[1]Department of Mathematics, Madhyanchal Professional University, Bhopal, India
[2]Department of Physics, Madhyanchal Professional University, Bhopal, India

Abstract

This study explores the application of machine learning (ML) and artificial intelligence (AI) techniques to analyze unstructured textual data, focusing on topic modeling, sentiment detection, and behavioral prediction. We employ multinomial document models and unsupervised learning strategies to extract latent topics and evaluate the emotional and conversational drivers social media posts. A major contribution is the implementation of Behavior Dirichlet Probability Model (BDPM) which user moods and behaviors through unstructured textual data. The results validate the hypothesis of the model's ability to and guess behavior patterns with high accuracy, providing actionable insights for digital marketing strategies, techniques to enhance user interaction and mental wellness evaluation.

Keywords

Machine Learning, Artificial Intelligence, Unstructured Data, Multinomial Document Model, Predicting Users'

1. Introduction

Nowadays, unstructured data makes up the great bulk of the digital world's data. The term "unstructured data" may describe any kind of information that is not neatly arranged or has not been subject to a previous data model. Data that does not adhere to the conventional structure of relational databases (RDBMS) is known as unstructured data. You may hear it called qualitative data or approximated data from time to time. It can be highly obscure and difficult to understand, despite this, valuable conclusions can be drawn. Owing to the lack of a coherent data framework and simple, identifiable organization, the unstructured nature of the data complicates automated processing and manipulation by computers. It is incompatible with relational databases due to the absence of a predefined structure. While the file type of unstructured data can be identified, its actual content may still be unknown. Everyone can understand what's in these files. Their incompatibility with relational databases makes them challenging to examine. International Data Corporation predicts that by 2025, 80 percent of all data will lack an organized format. Companies are starting to pay more attention to unstructured data, and they want to use it to their advantage. This is especially true given the amount of information that is contained in and hidden beneath unstructured data. This analysis is necessary for firms to extract useful information from unstructured data and make relevant business choices. Based on these findings, choices will be made that impact consumers' perceptions of future planning, provide light on customer requirements, and lead to contributions that better address those needs.

[*]Corresponding author: farhakhan50705@gmail.com (Farha Khan)

Received: 20 July 2025; **Accepted:** 12 August 2025; **Published:** 25 September 2025

Analyzing unstructured data using machine learning algorithms and natural language processing methods allows a model to be trained to work for a particular goal. Suicide claims the lives of one person every forty seconds on a global scale, amounting to 800,000 avoidable fatalities annually. Improving clinical psychometrics, reducing stigma associated with mental health, conducting focused research, raising awareness of suicide as a preventable cause of death, and providing responsive professional training have all contributed to a lack of consistency and accuracy in predicting suicidal behavior.

2. Literature Review

Pasrija [5] investigates AI-driven personalization in social media, illustrating the behavioral prediction potential of AI systems, which is consistent with our BDPM approach.

Similarly, Sarker [4] outlines real-world applications of machine learning algorithms, aligning with our focus on practical deployment of AI models.

Najafabadi et al. [3] discuss deep learning applications and challenges in big data analytics, further justifying our emphasis on the complexities of unstructured data.

Rahmani et al. [2] provide a systematic study of AI approaches for big data analytics, offering methodological foundations relevant to our work on unstructured textual data.

Ravi et al. [1] highlight how AI can extract essential insights from unstructured data, which supports the analytical framework developed in this study.

Aldoseri et al. [10] emphasize the capabilities of AI systems, particularly in handling unstructured data—relies significantly on the presence and quality of core data resources. One of the most significant findings of this study is the identification of critical challenges that AI encounters when processing unstructured and heterogeneous data formats. Key issues include inconsistent data structuring and obstacles in preserving the reliability and accuracy of data, and the ethical concerns involved in analyzing sensitive and unstructured information. According to the authors, these issues pose substantial limitations on the effectiveness of AI in fields like healthcare and finance, where unstructured data (e.g., clinical notes or transaction logs) exists in large quantities but is inherently complex. They advocate for adaptive data strategies -encompassing improved pre processing pipelines and the integration of ethical AI frameworks as a means to address these challenges. This aligns closely with the present research's objective—to develop machine learning models capable of analyzing and extracting meaningful patterns from unstructured text.

Danielle Hopkins et al. [9] Government health programs have helped reduce stigma associated with mental health issues, yet suicide is still a major cause of avoidable death globally [9]. The use of algorithms to mimic and reproduce human intellect is known as machine learning (ML), a subfield of AI. New methods of predicting suicidal thoughts and behaviors have emerged thanks to technological developments, as clinician-based suicide prediction has not improved over the years. Incorporating and comparing results between structured data (only machine interpretable, like electronic health records) and unstructured data (human interpretable, like psychometric instruments), this systematic review and meta-analysis sought to synthesize current research regarding data sources in ML prediction of suicide risk. We looked for research on ML and suicide risk prediction in academic journals, online databases, and grey literature. Thirteen teen studies met the criteria. With structured data, the AUC was 0.873 and with unstructured data, it was 0.866; the overall result was an AUC of 0.860. The research' sources were unable to be characterized, leading to high variation. Predictions of suicide risk behavior were often accurate, according to the research. While the input data for structured and unstructured data sets were varied in terms of amount and kind, meta-analysis revealed that the results were similarly accurate.

Perifanis et al. [8] The use of artificial intelligence (AI) into business and IT strategy offers great potential for enterprises to create new business models and gain competitive advantages. While some trailblazers are making good use of AI, most companies are unable to capitalize on the prospects for value generation. A total of 139 peer-reviewed papers were examined using the study technique outlined by Webster and Watson (1). The literature indicates that previous studies have highlighted the performance benefits, success criteria, and challenges of using AI. This review's findings highlight the gaps in our current understanding of AI and the areas that need greater investigation before we can build AI capabilities, incorporate them into business and IT plans, and ultimately increase value across all areas of a company. Only by meticulously embracing and executing these state-of-the-art technologies can organizations hope to thrive in the modern age of digital transformation alignment. This review seeks to address the complexity of resource orchestration and governance in this dynamic environment, which is a pressing issue in the early stages of research on the strategic implementation of AI in organizations. By doing so, it hopes to assist present and future organizations in effectively enhancing various business value outcomes, despite the revolutionary advantages that AI capabilities may promote.

Sou âd Demigha et al. [7] The term "Big Data" describes information that is so massive that it defies conventional application processing capabilities. This is because of the difficulties inherent in collecting, storing, transporting, accessing, quickly processing, and updating this data. Analytics using AI, ML, and DL are commonplace in the Big Data paradigm. This article delves into how Big Data has altered the use of AI strategies and tools.

Mohsen Soori et al. [6] The area of high-tech robots has been utterly transformed by developments in AI, ML, and DL in the last few years. By enhancing their intelligence, efficiency, and adaptability to complicated tasks and settings, AI, ML, and DL are revolutionizing the area of advanced robotics.

Autonomous navigation, object detection and manipulation, predictive maintenance, and natural language processing are just a few examples of how AI, ML, and DL are being used in advanced robotics. Collaborative robots (cobots) that can learn from people and adapt to new activities and surroundings are another product of these technological advancements. Modern transportation systems may benefit from the use of AI, ML, and DL to make travel safer, more efficient, and more convenient for everyone involved. Manufacturing assembly robots are also benefiting greatly from AI, ML, and DL, which are enhancing their intelligence, safety, and efficiency on the job. In addition, they may be used in many different ways in aviation management, which helps airlines save money, work more efficiently, and provide better service to their customers. In addition, taxi firms may benefit from AI, ML, and DL to enhance client service by making it safer, more efficient, and more reliable. The study delves into the latest advancements in AI, ML, and DL as they pertain to sophisticated robotics systems. It also explores different ways these systems might be used to modify robots. To bridge the gaps between current studies and published articles, more study on the use of AI, ML, and DL in advanced robotics systems is also recommended. It is feasible to study and improve the performance of advanced robots in different applications by examining the uses of AI, ML, and DL in advanced robotics systems. This would increase production in advanced robotic industries.

Chen et al. [11] examine integrated AI models that leverage both rule based natural language processing and deep learning methods to enhance the semantic interpretation of social media content, with a focus on mood recognition and topic identification.

Liu and Zhang [12] examine user behavior on digital platforms through the analysis of multimodal unstructured data. Their findings supported the efficacy of probabilistic models in capturing user engagement patterns, reinforcing our approach using BDPM.

3. Material

Given a training dataset *(x(i), y(i))* where every *x(i)* is a vector and *i* ranges from 1 to *n* and *y(i)* the number of categories or themes that are linked with it, which may be an integer between 1 and *k*. At its core, the problem at hand is a multi-category classification job, the goal of which is to assign each input realization *x* to one of the *k* potential categories that y might take. The linked job becomes a binary classification problem when *k = 2*.

Assume, for the sake of argument simplicity, that all *x* are elements of the set $\{-1, +1\}^{d}$ for some number d that defines the number of "features" in the model. On the other hand, each element *xj*, where j is an integer between 1 and d, may take on one of the two possible values.

The following is the way the Naïve Bayes model may be expressed (Qin, Tang and Chen, 2012).

Assign the vector components *xi* and the category label *y* to a set of random variables *X1* through *Xd*. The goal then shifts to predicting the combined likelihood as

$$P(Y = y, X\_1 = x\_1, X\_2 = x\_2, \ldots, X\_d = x\_d) \quad (1)$$

as a pair with the characteristics of each label *y x1... xd*. Here is the main assumption made in the Naïve Bayes setup:

$$P(X\_1, X\_2, \ldots, X\_d \mid Y) = \prod\_{j=1}^{d} P(X\_j \mid Y) \quad (2)$$

The naiveté of equation (2) stems from the strong NB assumption. Even if the model in question still practically maintains pretty effective, this one is incredibly advantageous as it delivers a substantial reduction of the count of model parameters.

Based on equation (3), there are two types of model parameters.

$$q(y) = P(Y=y) \text{ for every } y \in \{1... k\}$$

$$qj\ (x|y) = P\ (Xj = x \mid Y = y)$$

Lastly, the likelihood of any *y, x1... xd* considered to be:

$$P\left(y \cap x_{\{1\}} \cap x_{\{2\}} \cap \ldots \cap x_{\{d\}}\right) = q(y) \times \prod_{j=1}^{d} q_j\left(x_j \middle| y\right) \quad (3)$$

The purpose of this test is to uncover a previously unknown aspect *x = {x1, x2,..., xd}*, In order to get the estimations, we have to maximize the.

One way to get the MLE of equation using the parameters

$$q(y) = \frac{\sum_{i=1}^{n}\left[\left[y^{i}=y\right]\right]}{n} = \frac{count(y)}{n} \quad (4)$$

y and 0 in all other cases.

It is also possible to obtain the MLE for equation as:

$$\hat{q}_j\ (x/y) = \frac{\sum_{i=1}^{n}[[y^{i}=y \cap x_j^{(i)}=x]]}{\sum_{i=1}^{n}[[y^{(i)}=y]]} = \frac{count_j(x|y)}{count(y)} \quad (5)$$

$$count_j\ (x\backslash y) = \sum_{i=1}^{n}[[y^{(i)} = y \cap x_j^{(i)} = x]]$$

4. Result

Imagine a set of papers that are all either related to sports *(S)* or informatics *(I)*, with each document falling into one of those categories. Finding an estimate for a Bernoulli document classifier to identify unknown documents relevant to Sports or Informatics is the goal here, given a training dataset of eleven documents.

Allow eight words to make up vocabulary *V* as

$$V=\begin{bmatrix} w_1 = goal \\ w_2 = tutor \\ w_3 = variance \\ w_4 = speed \\ w_5 = drink \\ w_6 = defence \\ w_7 = performance \\ w_8 = field \end{bmatrix}$$

Therefore, any document may be represented as an 8-dimensional vector. Row vectors mi, where m_{it} is the count of words w_t in D_i, may now be used to represent a document D_i.

Below you can see the training dataset shown as a matrix that corresponds to each subject or category. In this matrix, each row represents an 8-dimensional document vector.

$$B^{sport} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$B^{Inf} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Using a Naïve Bayes (NB) classifier, the goal now is to categorize the following vectors into one of the two subjects.

$$b_1 = (1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1)^T$$

$$b_2 = (0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0)^T$$

The training dataset is indexed as,

$N = 11$, $NS = 6$, $NI = 5$.

The training dataset may be used to estimate the prior probabilities, which are provided by

$$\hat{P}(S) = \frac{6}{11}; \quad \hat{P}(I) = \frac{5}{11}.$$

Table 1 displays the number of documents in the training dataset, denoted as *nk(w)*, along with estimates of the likelihoods of individual words.

Table 1. *Count of Documents and Estimates of Word Likelihood.*

Words	ns (w)	$\hat{P}$ (w | S)	n1(w)	$\hat{P}$ (w | I)
W1	3	$\frac{3}{6}$	1	$\frac{1}{5}$
W2	1	$\frac{1}{6}$	3	$\frac{4}{5}$
W3	2	$\frac{2}{6}$	3	$\frac{3}{5}$
W4	3	$\frac{3}{6}$	1	$\frac{1}{5}$
W5	3	$\frac{3}{6}$	1	$\frac{1}{5}$
W6	4	$\frac{3}{6}$	1	$\frac{2}{5}$
W7	4	$\frac{4}{6}$	3	$\frac{3}{5}$
W8	4	$\frac{4}{6}$	1	$\frac{1}{5}$

The next step in classifying the data is to calculate the posterior probability of the two test points.

$$\hat{P}(S|b_1) \alpha \ \hat{P}(S) \times \prod_{t=1}^{8}[b_{1t} \times \hat{P}(w_t|S) + (1 - b_{1t}) \times (1 - \hat{P}(w_t|S))] \propto \frac{6}{11}\left(\frac{1}{2} \times \frac{5}{6} \times \frac{2}{3} \times \frac{1}{2} \times \frac{1}{2} \times \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3}\right) = \frac{5}{891}$$

$$\approx 5.6 \times 10^{-3}$$

$$\hat{P}(I|b_1) \propto \hat{P}(I) \times \prod_{t=1}^{8}[b_{1t} \times \hat{P}(w_t|I) + (1 - b_{1t}) \times (1 - \hat{P}(w_t|I))]$$

$$\propto \frac{5}{11}\left(\frac{1}{5}.\frac{2}{5}.\frac{2}{5}.\frac{1}{5}.\frac{1}{5}.\frac{1}{5}.\frac{2}{5}.\frac{1}{5}\right) = \frac{8}{859375}$$

$$\approx 2.81 \times 10^{-4}$$

Therefore, b1 is classifiable as an element of S.

$$\hat{P}(S|b_2) \propto \hat{P}(S) \times \prod_{t=1}^{8}[b_{2t} \times \hat{P}(w_t|S) + (1 - b_{2t}) \times (1 - \hat{P}(w_t|S))] \propto \frac{6}{11}\left(\frac{1}{2}.\frac{1}{6}.\frac{1}{3}.\frac{1}{2}.\frac{1}{2}.\frac{1}{3}.\frac{2}{3}.\frac{1}{3}\right) = \frac{12}{42768}$$

$$\approx 2.81 \times 10^{-4}$$

$$\hat{P}(I|b_2) \propto \hat{P}(I) \times \prod_{t=1}^{8}[b_{2t} \times \hat{P}(w_t|I) + (1 - b_{2t}) \times (1 - \hat{P}(w_t|I))] \propto \frac{5}{11}\left(\frac{4}{5}.\frac{3}{5}.\frac{3}{5}.\frac{4}{5}.\frac{1}{5}.\frac{4}{5}.\frac{3}{5}.\frac{4}{5}\right) = \frac{34560}{4296875}$$

$$\approx 8.1 \times 10^{-3}$$

Thus, b_2 may be classified as belonging to,

Imagine a set of papers that are all either related to sports (S) or informatics (I), with each document falling into one of these categories. The goal now is to estimate a Naive Bayes classifier using the Multinomial model and to identify unknown documents as either Sports or Informatics, given a training dataset of eleven documents.

Assume the following eight-word vocabulary:

$$V=\begin{bmatrix} w_1 = goal \\ w_2 = tutor \\ w_3 = variance \\ w_4 = speed \\ w_5 = drink \\ w_6 = defence \\ w_7 = performance \\ w_8 = field \end{bmatrix}$$

This means that every document may be represented by an 8-dimensional vector. Row vectors mi, where m_{it} is the count of words *wt* in D_i, may now be used to represent a document D_i,

So, below you can see the training dataset organized into a matrix for each subject or category. Each row represents an 8-dimensional document vector.

$$M^{sport} = \begin{bmatrix} 2 & 0 & 0 & 0 & 1 & 2 & 3 & 1 \\ 0 & 0 & 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 2 & 1 & 0 \\ 1 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\ 2 & 0 & 0 & 0 & 1 & 0 & 1 & 3 \\ 0 & 0 & 1 & 2 & 0 & 0 & 2 & 1 \end{bmatrix}$$

$$M^{Inf} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Now, classify the following sample documents according to one of the two categories:

*D*1 = *w*5*w*1*w*6*w*8*w*1*w*2*w*6

*D*2 = *w*3*w*5*w*2*w*7

Representing the frequency of the term "w" in all documents related to the subject *k* using *nk(w)*, Here is how likelihood estimate is carried out:

$$\hat{P}(w|S) = \frac{n_s(w)}{\sum_{v \in V} n_s(v)},$$

$$\hat{P}(w|I) = \frac{n_i(w)}{\sum_{v \in V} n_i(v)}.$$

Table 2. *Parameter Estimates.*

	ns (w)	$\hat{P}$ (w | S)	n1(w)	$\hat{P}$ (w | I)
W1	5	$\frac{5}{36}$	1	$\frac{1}{16}$
W2	1	$\frac{1}{36}$	4	$\frac{4}{16}$
W3	2	$\frac{2}{36}$	3	$\frac{3}{16}$
W4	5	$\frac{5}{36}$	1	$\frac{1}{16}$
W5	4	$\frac{4}{36}$	1	$\frac{1}{16}$
W6	6	$\frac{6}{36}$	2	$\frac{2}{16}$
W7	7	$\frac{7}{36}$	3	$\frac{3}{16}$
W8	6	$\frac{6}{36}$	1	$\frac{1}{16}$

5. Conclusion

This study offers an analytical framework for gaining a competitive edge via the analysis of textual data. Steps in the suggested approach include identifying subnets, tokenizing N-grams, determining the document term matrix using inverse document frequency, topic modeling on each subnet to find the top N-grams, and lastly, determining the polarity and relevance of N-grams within the topics. By carrying out each step in turn, we are able to extract the positive and negative polarity of the main drivers from the textual data. In a broader sense, this method may be used to successfully achieve changing significance of hypernyms, which are drivers of textual data, in the event that we are studying textual data and want to understand its drivers.

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
BDPM	Behavior Dirichlet Probability Model
NB	Na ïve Bayes
AUC	Area Under Curve
RDBMS	Relational Database Management System

Conflicts of Interest

The authors declare no conflicts of interest.

Appendix

Critical Commentary and Link to Current Research

While the reviewed literature addresses the potential of AI and ML in handling large-scale unstructured data, few studies provide integrated frameworks that combine topic modeling, sentiment analysis, and behavior prediction. Our study builds upon these gaps by introducing BDPM—a unified probabilistic model that not only analyzes textual content but also infers user behavior trends. This contribution aims to operationalize theoretical advancements in practical, real-time systems for digital platforms.

References

[1] Ravi, Mr & Oza, Rajnikant & Shree, Bhavans & Doshi, H & Dipti, H & Domadiya, & Oza, Ravi & Domadiya, Dr. Dipti & Punjani, Dipti. (4). ROLE OF AI FOR GAINING ESSENTIAL INSIGHT FROM UNSTRUCTURED DATA.

[2] Rahmani AM, Azhir E, Ali S, Mohammadi M, Ahmed OH, Yassin Ghafour M, Hasan Ahmed S, Hosseinzadeh M. Artificial intelligence approaches and mechanisms for big data analytics: a systematic study. PeerJ Comput Sci. 2021 Apr 14; 7: e488. https://doi.org/10.7717/peerj-cs.488

[3] Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M. et al. Deep learning applications and challenges in big data analytics. Journal of Big Data 2, 1(5). https://doi.org/10.1186/s40537-014-0007-7

[4] Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160(6). https://doi.org/10.1007/s42979-021-00592-x

[5] Divij Pasrija, (7), "AI-DRIVEN AD PERSONALIZATION IN SOCIAL MEDIA PLATFORMS," © 2025 JETIR May 2025, Volume 12, Issue 5 www.jetir.org

[6] Mohsen Soori (8), "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," Cognitive Robotics Volume 3, 2023, Pages 54-70.

[7] Souâd Demigha (9), "The impact of Big Data on AI," 2020 International Conference on Computational Science and Computational Intelligence (CSCI).

[8] Perifanis, N.-A.; Kitsios, F. Investigating the Influence of Artificial Intelligence on Business Value in the Digital Era of Strategy: A Literature Review. Information 2023, 14, 85. https://doi.org/10.3390/info14020085

[9] Danielle Hopkins (10), "Structured data vs. unstructured data in machine learning prediction models for suicidal behaviors: A systematic review and meta-analysis," SYSTEMATIC REVIEW article Front. Digit. Health, 02 August 2022 Sec. Digital Mental Health Volume 4 - 2022 | https://doi.org/10.3389/fdgth.2022.945006

[10] Aldoseri, A.; Al-Khalifa, K. N.; Hamouda, A. M. Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. *Appl. Sci.* 2023, *13*, 7082. https://doi.org/10.3390/app13127082

[11] Chen, X., Wang, Y., & Li, Z. (11). Integrated AI models for semantic interpretation of social media content. Journal of Computational Social Science, 5(11), 211-230.

[12] Liu, H., & Zhang, K. (12). Multimodal unstructured data analysis for user behavior prediction on digital platforms. IEEE Transactions on Knowledge and Data Engineering, 33(12), 3456-3468.