

Impact of smoothing techniques for text classification: implementation in hidden Markov model

Norsyela Muhammad Noor Mathivanan^{1,2,3}, Roziah Mohd Janor¹, Shukor Abd Razak⁴,
Nor Azura Md. Ghani¹

¹School of Mathematical Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia

²School of Computing and Creative Media, University of Wollongong Malaysia, Shah Alam, Malaysia

³UOW Malaysia KDU Penang University College, Georgetown, Malaysia

⁴Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Kuala Terengganu, Malaysia

Article Info

Article history:

Received Sep 25, 2024

Revised Aug 18, 2025

Accepted Sep 7, 2025

Keywords:

E-commerce products
Job title classification
Occupational data mining
Product classification
Sequential data
Spam filtering
Supervised learning model

ABSTRACT

A hidden Markov model (HMM) is widely used for sequence modeling in various text classification tasks. This study investigates the impact of different smoothing techniques, such as Laplace, absolute discounting, and Gibbs sampling on HMM performance across three distinct domains: e-commerce products, spam filtering, and occupational data mining. Through the comparative analysis, Laplace smoothing consistently outperforms other techniques in handling zero-probability issues, demonstrating superior performance in the e-commerce and SMS spam datasets. The HMM without any smoothing technique achieved the best results for job title classification. This divergence underscores the dataset-specific nature of smoothing requirements, where the simplicity of parameter estimation proves effective in contexts characterized by a limited and repetitive vocabulary. Hence, the findings suggest that tailored smoothing strategies are crucial for optimizing HMM performance in different textual analysis applications.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Nor Azura Md. Ghani
School of Mathematical Sciences, Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
40450 Shah Alam, Selangor Malaysia
Email: azura@tmsk.uitm.edu.my

1. INTRODUCTION

Text classification is a core task in natural language processing (NLP) that involves assigning predefined labels to text documents. It is widely used in applications such as information retrieval [1], [2], sentiment analysis [3], [4], product classification [5], [6], spam detection [7], [8], and document categorization [9]. The performance of text classification methods greatly affects the efficiency and accuracy of many automated systems, making it important to improve and evaluate different techniques. One well-known method in text classification is the hidden Markov model (HMM) [10], which is effective in modeling sequences due to its probabilistic structure. However, HMMs often struggle with sparse data and unseen events [11], common issues in large text datasets. Smoothing techniques help address these problems by adjusting probability estimates for rare or missing data [12], improving generalization and model reliability.

Several smoothing methods, such as Laplace smoothing, Good-Turing discounting, and backoff models have been extensively studied in NLP. These methods reduce the risk of assigning zero probabilities to unseen events, which could otherwise cause errors during classification. Recently, more advanced techniques have been developed to improve this process. For example, Ren *et al.* [13] introduced

discrimination-aware label smoothing, which dynamically learns label distributions to handle class imbalance and noisy data. Wu *et al.* [14] proposed text smoothing, which uses pre-trained masked language models to convert one-hot vectors into more informative representations, improving performance in low-resource settings. Fetta *et al.* [15] explored semantic graph smoothing, using semantic relationships to enhance sentence embeddings for better classification and clustering. In HMMs, smoothing enhances both transition and emission probability estimates, improving classification accuracy. Wu *et al.* [14] showed that smoothed representations can outperform one-hot encodings in data augmentation, suggesting similar benefits for HMMs. Smoothing emission probabilities is also essential to avoid zero values, as shown in metadata extraction from bibliographic references [16]. Using the entire vocabulary with appropriate smoothing has outperformed conventional feature selection in ensuring reliable parameter estimation [17]. Furthermore, fuzzy smoothing of state transitions has improved classification rates in uncertain environments, such as speech recognition, with applications in text classification [18]. HMMs have also shown success in domains like biomedical text and document classification, particularly when enhanced with smoothing techniques [19].

This research aims to analyze the impact of different smoothing techniques on the performance of HMMs in text classification tasks. By systematically implementing and comparing these methods, we seek to identify the most effective strategies for enhancing the accuracy and robustness of HMM-based classifiers. The study also explores the trade-offs associated with each technique, providing insights into their practical applications and potential areas for further improvement. In the following sections, we will review the theoretical foundations of HMMs and smoothing techniques, describe our experimental setup, present the results of our comparative analysis, and discuss the implications of our findings. Through this comprehensive evaluation, we hope to contribute to the ongoing efforts in optimizing text classification methodologies and advancing the field of NLP.

2. METHOD

2.1. Data description

Department of statistics Malaysia (DOSM) has collected product information from one of the major online store websites through the STATSBD project known as price intelligence (PI) using its prototype web scraper. A few leaf nodes were used to represent the chosen categories from the browse tree of the website. Table 1 presents the description of the four corpora selected for this study which incorporated datasets from three different domains. The first domain is e-commerce products and there are two datasets used i.e. non-food and household products under this domain. The two categories under the non-food dataset are cooking & dining (407 instances) and party accessories (80 instances). On the other hand, the five categories under the Frozen dataset are frozen food (291 instances), yogurt (162 instances), ice cream (147 instances), cheese (85 instances), and juices (87 instances).

This study also utilized two additional datasets from different domains, namely spam filtering and occupational data mining. The dataset related to spam filtering was retrieved from the UCI repository, which provides a widely recognized collection of data for machine learning applications. This dataset comprises labeled instances of emails categorized as spam or non-spam, allowing for the evaluation of text classification models in distinguishing between unsolicited and legitimate messages. Meanwhile, the dataset from Github was used for classifying job titles according to their job categories. This dataset consists of various job titles with corresponding categories, offering valuable insights for machine learning models aimed at automating job classification. The inclusion of these datasets ensures the robustness of the study by covering diverse domains and real-world applications, enhancing the generalizability of the findings.

Table 1. Summary description of datasets

Dataset	Category	Instance	Number of features	
			TF	TF-IDF
Non-food	2	487	461	459
Frozen food	5	772	656	654
SMS spam	2	5567	5903	5637
Job title	4	8586	1925	1919

2.2. Data characteristics

The product title lengths across the datasets show a relatively short and consistent distribution, as seen in Figure 1. The non-food products subset has a mode of 8 characters, and the frozen food products subset has a mode of 6 characters. This indicates that the product titles in these datasets are typically shorter, which may influence the effectiveness of different smoothing techniques.

Figure 2 shows the distribution of text lengths for the SMS spam and job title datasets. The SMS spam text length distribution shows a mode of 7 characters, with the majority of messages being relatively short. This pattern is similar to the product title lengths seen in Figure 1. In contrast, the job title dataset exhibits a mode of 3 characters, with a significant portion of the data consisting of very short and repetitive titles.

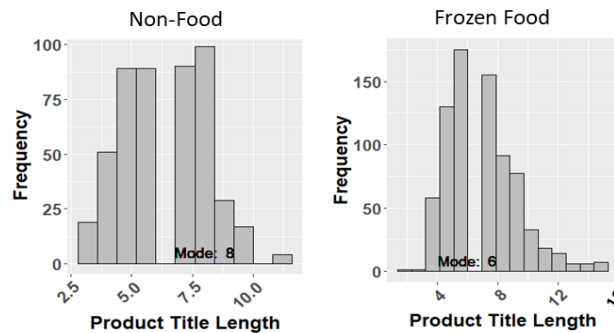


Figure 1. Non-food and frozen food products title lengths

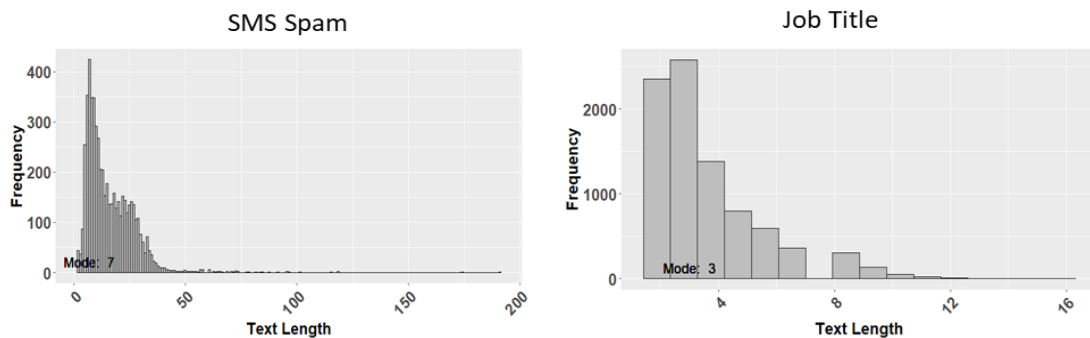


Figure 2. SMS spam and job title text lengths

2.3. Research design

There are several steps needed before classifying the data, as shown in Figure 3. These steps are essential in text classification research. After extracting the data, three preprocessing steps are applied i.e., tokenization, stop word removal, and stemming [20]. Tokenization splits product descriptions into words, stops word removal filters out common words, and stemming reduces words to their root forms, ensuring standardized data. The study used term-frequency (TF) and term-frequency inverse document-frequency (TF-IDF) for feature extraction and applied the correlation feature selection technique. These selected features are then used as inputs for HMMs with different smoothing techniques.

According to the fundamental problems related to HMM, only two steps should be taken to implement a supervised HMM, i.e., estimating the parameters by learning their initial probability matrices and decoding the sequence to find the best-hidden sequence [21]. Both steps are done to solve learning and decoding problems, respectively. The likelihood problem is typically ignored when using supervised HMMs because the algorithm used for the forward algorithm is employed for likelihood computation in semi-supervised or unsupervised learning. The Viterbi algorithm is similar to the forward algorithm, but instead of using the sum of previous path probabilities, it uses the maximum value. The forward algorithm lacks a back pointer component and computes the observation likelihood, while the Viterbi algorithm finds the most likely state sequence by tracking the path of hidden states leading to each state [21].

The learning problem in HMMs involves adjusting the model parameters. Using the training set of observations, the aim is to find the best way to predict the states. By comparing the predicted states to the known states, the prediction accuracy can be estimated based on the correctly decoded states in the test set. Since the states are known, maximum likelihood estimates (MLEs) are used to maximize the complete-data likelihood and obtain the HMM parameters. There are three HMM parameters which are A , B , and π that can be learned given an observation sequence o and the set of states in the HMM. The initial probabilities are

denoted by π_s are the number of times for state s labeled in the dataset. The initial parameter computation is presented by,

$$\text{Initial probabilities, } \pi_s = \frac{p_s}{\sum_{s'}(p_{s'})}$$

Next, the transition probabilities are denoted by $A_{s_i s_j}$ are the number of times transition s_i to s_j was taken among the sequences. The transition parameter computation is presented by,

$$\text{Transition probabilities, } A_{s_i s_j} = \frac{a_{s_i s_j}}{\sum_{s_j'}(a_{s_i s_j'})}$$

The emission probabilities are denoted by $B_s(k)$ are the number of times k was emitted while in state s . The emission parameter computation is presented by,

$$\text{Emission probabilities, } B_s(k) = \frac{b_s(k)}{\sum_{k'}(b_s(k'))}$$

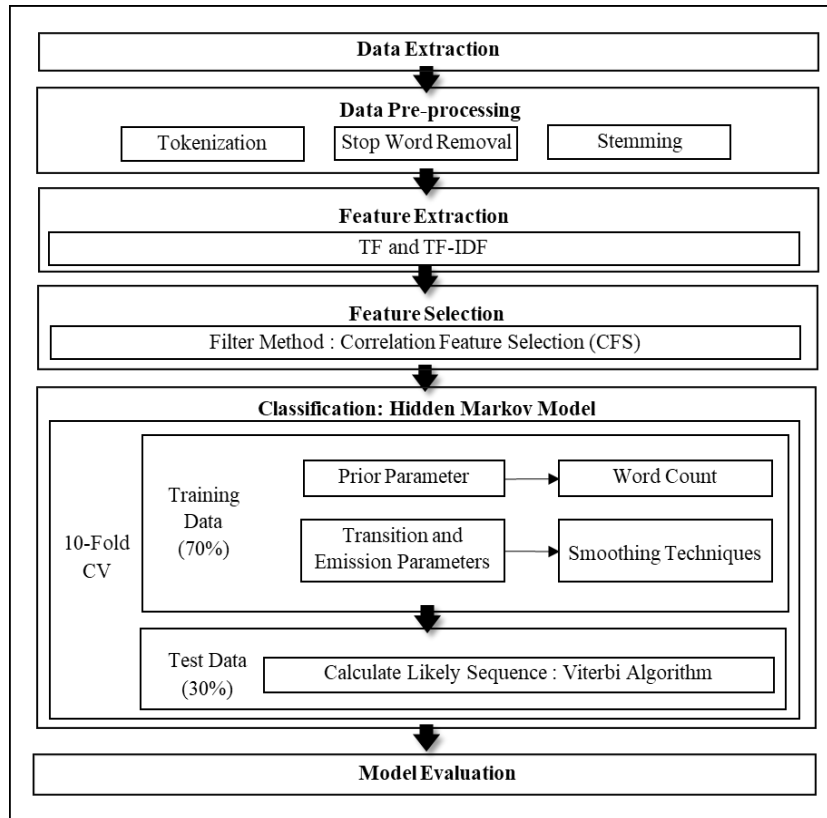


Figure 3. Research framework

In a supervised HMM, emission probabilities are computed to categorize all words in the dataset, and transitions between categories for a sequence of observations are counted. These two adjustable parameters are key for achieving a high-performance classification model and are often the focus of researchers. The pseudocode to compute all three parameters is shown in Figure 4.

In classification models like HMMs with hidden variables, the decoding task aims to find the optimal state sequence for a given observation sequence, revealing the hidden structure of the HMM. Typically, this involves running the forward algorithm to compute the likelihood of the observation sequence for each possible hidden state sequence. However, due to the exponential number of state sequences, directly using the forward algorithm becomes impractical. Instead, the Viterbi algorithm is commonly employed. It efficiently finds the optimal state sequence using dynamic programming and recursion.

```

Let
N = Number of observations in the dataset
M = Number of words in each observation
S = Number of states in the dataset
K = Number of words in the dataset
 $O_j^i$  = Observations in the dataset,  $i = 1,2,3,\dots,N$  and  $j = 1,2,3,\dots,M$ 
 $C^i$  = States in the dataset,  $i = 1,2,3,\dots,N$ 
 $C_s$  = Labelled category for each observation in dataset,  $s = 1,2,3,\dots,S$ 
 $W_k^s$  = Words in the dataset,  $k = 1,2,3,\dots,K$  and  $s = 1,2,3,\dots,S$ 

function INITIAL_HMM()

for each  $s$  in  $C$ 
  count  $O$ 
  denote number by  $p_s$ 
  calculate initial probabilities by using  $\pi_s = \frac{p_s}{\sum_{s'}(p_{s'})}$ 

for each  $s$  in  $W$ 
  for each  $k$  in  $W$ 
    count  $W_k^s$  in  $O$ 
    denote number by  $b_s(k)$ 
    calculate emission probabilities by using  $B_s(k) = \frac{b_s(k)}{\sum_{k'}(b_s(k'))}$ 

for each  $k$  in  $W$ 
  assign category by  $C(k) = \max(W_k^s)$ 

match  $C(k)$  with  $O_j^i$ 

for each  $s_i$  in  $W$ 
  for each  $s_j$  in  $W$ 
    count transition of  $C(k)$  from  $s_i \rightarrow s_j$  according to  $O_j^i \rightarrow O_{j+1}^i$ 
    denote number by  $a_{s_i s_j}$ 
    calculate transition probabilities by using  $A_{s_i s_j} = \frac{a_{s_i s_j}}{\sum_{s_j'}(a_{s_i s_j'})}$ 

```

Figure 4. A pseudocode for parameter estimations in HMM

2.4. Smoothing techniques

A smoothing technique is a flattering probability distribution process to ensure all word sequences can occur with some probabilities rather than having certain words with zero probabilities. The term smoothing refers technique for adjusting the maximum likelihood estimate of probabilities to provide more accurate probabilities. There are three smoothing techniques used in the study i.e. Laplace, absolute discounting, and Gibbs sampling smoothing techniques. Figure 5 shows the pseudocodes for parameter computations for each smoothing technique.

Specifically, Figure 5(a) illustrates the parameter computations for the Laplace smoothing technique, which is the simplest and most frequently used method for addressing data sparsity. The Laplace smoothing technique is most frequently used by previous researchers despite various smoothing techniques have been proposed [12]. It is the simplest and oldest technique to solve data sparseness problems. This technique also serves as a fundamental baseline concept for other smoothing techniques with the same parameters.

Figure 5(b) depicts the parameter computations for the absolute discounting smoothing technique, which adjusts transition and emission probabilities by discounting observed counts. It is a method commonly used in language modeling contexts to adjust probability estimates by discounting observed counts of events [22]. In the context of HMMs, absolute discounting adjusts both transition probabilities (the likelihood of moving from one hidden state to another) and emission probabilities (the likelihood of emitting observable symbols given a hidden state). The core idea behind absolute discounting is straightforward yet effective: it ensures that even if certain state transitions or emissions were not observed during training, they still retain a non-zero probability in the model. This is achieved by subtracting a fixed discount d from the observed counts of events, and redistributing this discount mass among all possible events for a given context.

Figure 5(c) presents the pseudocode for the Gibbs Sampling technique, a Markov chain monte carlo (MCMC) method used for estimating model parameters through iterative sampling. It is particularly well-suited for complex models and large datasets [23]. It is a Markov chain monte carlo (MCMC) method that generates samples from a joint probability distribution by iterative sampling from the conditional distributions of each variable. In the context of HMMs, Gibbs Sampling can be used to estimate the hidden states given the observed data and then to update the model parameters based on these sampled states. This iterative process allows for the exploration of the posterior distribution of the model parameters, providing a robust means of incorporating the variability in the data and avoiding overfitting to sparse observations.

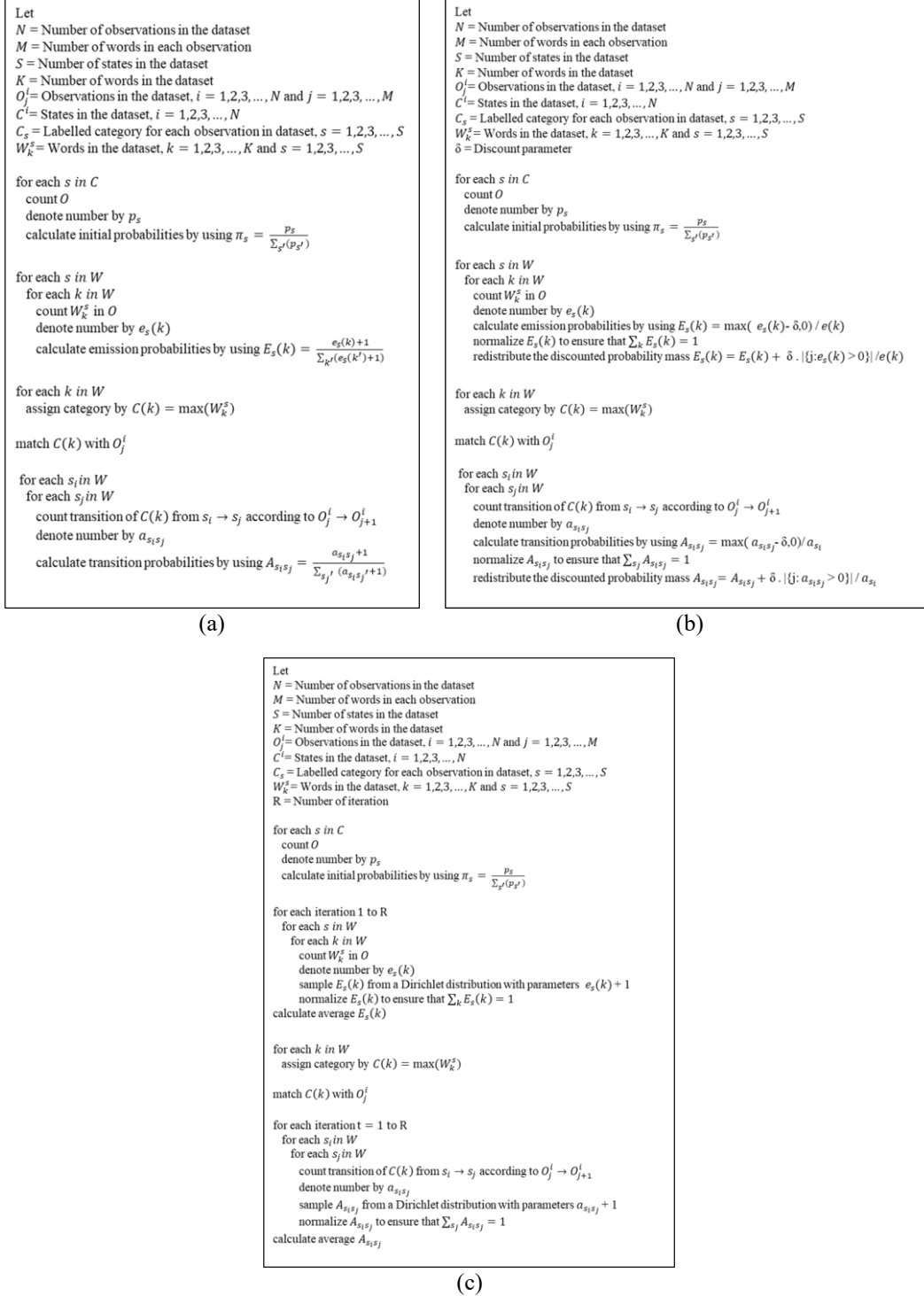


Figure 5. Pseudocodes for parameter estimations in HMM using: (a) Laplace smoothing technique, (b) absolute discounting smoothing technique, and (c) Gibbs sampling smoothing technique

3. RESULTS AND DISCUSSION

The comparative analysis was done by observing the HMM performances when applying different smoothing techniques. Table 2 shows the classification results for HMM variations applied to two e-commerce datasets: non-food products and frozen food products. The evaluation metric used is the F1-score, and two different embedding techniques, TF and TF-IDF are applied. For the non-food products dataset, the standard HMM without smoothing achieves an F1-score of 82.27% with TF embedding, which is slightly better compared to when using TF-IDF embedding. When absolute discounting is applied, the F1-scores increase

marginally to 82.97% for both TF and TF-IDF, indicating a slight benefit from smoothing. Gibbs sampling smoothing results in F1-scores of 85.47% with TF and 85.36% with TF-IDF, showing a moderate improvement over the standard HMM. Laplace smoothing shows a better improvement, achieving F1-scores of 87.01% with TF and 87.24% with TF-IDF, making it the best-performing technique for this dataset.

For the frozen food products dataset, the standard HMM achieves an F1-score of 66.85% with TF, improving to 66.96% with TF-IDF. Laplace smoothing again achieves the highest F1-score 67.50% with TF and 67.87% with TF-IDF, making it the most effective smoothing technique for this dataset. The performances of HMM with Absolute discount and Gibbs sampling smoothing techniques are lower compared to standard HMM. Absolute discount smoothing subtracts a fixed amount from the counts of observed events and redistributes it to unseen events. While this helps handle zero probabilities, it may over-smooth probabilities when observed events already reflect the underlying distribution well [22]. Gibbs sampling is a complex iterative technique, and it can introduce noise if not properly tuned or converged. These issues can result in a lower F1-score compared to using a standard HMM without smoothing [24].

The best F1-scores for both datasets are achieved with HMM using Laplace smoothing and TF-IDF embedding. In addition, Table 3 presents the classification results for HMMs with different smoothing techniques applied to two new text classification domains: spam filtering and job title classification. The best F1-score for classifying the SMS spam dataset is 69.17% and it is executed from HMM using Laplace smoothing and TF-IDF as the embedding technique. This result is in line with the best model obtained for classifying e-commerce product datasets used in the study.

Table 2. F1-scores of HMM models for e-commerce product datasets

Data	Classification model	Embedding technique	F1-score (%)	Best model
E-commerce	HMM	TF	82.27	HMM Laplace
Non-food products	HMM Laplace	TF-IDF	82.08	TF-IDF
		TF	87.01	
	HMM DISC	TF-IDF	87.24	
		TF	82.97	
	HMM GIBBS	TF-IDF	82.97	
		TF	85.47	
		TF-IDF	85.36	
		TF	66.85	HMM Laplace
	HMM Laplace	TF-IDF	66.96	TF-IDF
		TF	67.50	
Frozen food products	HMM DISC	TF-IDF	67.87	
		TF	65.00	
	HMM GIBBS	TF-IDF	65.06	
		TF	66.01	
		TF-IDF	66.31	

Table 3. F1-scores of HMM models for spam filtering and job title datasets

Data	Classification model	Embedding technique	F1-score (%)	Best model
SMS Spam Corpus	HMM	TF	68.80	HMM Laplace
		TF-IDF	67.14	TF-IDF
	HMM Laplace	TF	69.08	
		TF-IDF	69.17	
	HMM DISC	TF	68.30	
		TF-IDF	68.48	
	HMM GIBBS	TF	68.80	
		TF-IDF	69.07	
	HMM	TF	68.64	HMM
		TF-IDF	67.17	TF
Job title corpus	HMM Laplace	TF	66.33	
		TF-IDF	66.12	
	HMM DISC	TF	67.27	
		TF-IDF	67.15	
	HMM GIBBS	TF	66.91	
		TF-IDF	66.62	

However, the standard HMM without smoothing technique seemed to fit the best for classifying job titles. The highest F1-score is 68.64% using TF as the embedding technique. None of the smoothing

techniques helped in improving the HMM performances. Job titles often use a narrow vocabulary with repetitive terms, effectively captured by TF. In contrast, TF-IDF diminishes common terms, potentially reducing the importance of frequent job-related words. Avoiding smoothing allows the model to rely solely on observed data, beneficial for small and specific datasets like job title classification. Thus, the analysis reveals that smoothing techniques may not always improve HMM performances, as their effectiveness depends on the nature and complexity of the data. Nonetheless, the combination of Laplace smoothing and TF-IDF mostly provides the best results among the tested techniques.

The results showed that Laplace smoothing consistently outperformed other techniques in the e-commerce and SMS spam datasets. This can be attributed to the specific characteristics of these datasets, such as their relatively larger vocabulary size and higher degree of data sparsity. As shown in Figure 1 (product title lengths) and Figure 2 (SMS spam text lengths), the results align with these dataset features. Both product titles and SMS spam consist of relatively short text, making them suitable for Laplace smoothing. In these cases, Laplace smoothing, which adds a small constant to all observed counts, is beneficial in addressing zero-probability issues. It ensures that every possible event, even those not observed in the training data, has a non-zero probability [12]. This characteristic makes Laplace smoothing particularly effective for datasets with a more diverse vocabulary, as it prevents rare terms from being completely disregarded, thus improving the model's performance.

In contrast, job title classification did not benefit from Laplace smoothing, primarily due to the extremely short and repetitive nature of the titles, as shown in Figure 2 (job title text lengths). The job title dataset is characterized by a limited vocabulary, with a high frequency of term repetition. This type of dataset, combined with a smaller sample size, reduces the necessity for smoothing techniques. Laplace smoothing, in this instance, introduced unnecessary complexity without providing any clear advantage. The standard HMM model, relying on observed frequencies, was able to effectively capture the relevant patterns in the dataset. Furthermore, the TF embedding method, which retains the frequency of terms, proved more suitable for this dataset. Its ability to emphasize frequent terms contributed to better model performance in job title classification.

The varying performance of the smoothing techniques across datasets can also be attributed to the structural differences between them. In more variable datasets, such as those used for spam filtering, where there is a broader range of content and diversity in the vocabulary, smoothing methods like Laplace contribute to model generalization. By preventing overfitting to specific word patterns, Laplace smoothing enhances classification accuracy. However, in highly structured datasets like job titles, where the vocabulary is repetitive and narrowly defined, smoothing can reduce the impact of frequent, important terms [25]. In such cases, relying on the raw observed counts, as done in the standard HMM approach, proves more effective.

4. CONCLUSION

The study compares the performance of HMMs using different smoothing techniques. The results consistently favored Laplace smoothing for the e-commerce and SMS spam datasets, demonstrating its efficacy in addressing zero-probability scenarios across diverse contexts. However, a notable departure emerged with the job title dataset, where the HMM without any smoothing technique yielded superior performance. This divergence can be attributed to the distinct nature of job titles, characterized by a restricted and repetitive vocabulary. The model benefited from exact match probabilities without the additional complexity introduced by smoothing methods. This finding underscores the effectiveness of simplicity in specific data settings, suggesting that simpler parameter estimation approaches may bolster an HMM performance in similar classification tasks. Future research could focus on combining different smoothing techniques to improve performance, such as integrating Laplace smoothing with other methods like absolute discounting. Exploring the impact of smoothing on advanced models, like deep learning classifiers, could also provide valuable insights. Additionally, automating the selection of the best smoothing technique based on dataset characteristics would be a useful direction. Finally, applying these techniques to specialized fields, like medical or legal text classification, could further enhance their practical use.

FUNDING INFORMATION

The research is funded by the University Teknologi MARA and Ministry of Education Malaysia under the Grant Scheme (FRGS/1/2018/STG06/UITM/01/1).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Norsyela Muhammad	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Noor Mathvianan														
RoZIAH Mohd Janor	✓	✓		✓				✓		✓	✓	✓		
Shukor Abd Razak			✓	✓		✓				✓			✓	✓
Nor Azura Md. Ghani	✓	✓				✓	✓		✓			✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The e-commerce product datasets are not publicly available as they are proprietary to the Department of Statistics Malaysia. The SMS Spam Collection is publicly accessible at: <https://archive.ics.uci.edu/dataset/228/sms+spam+collection>, and the Job Title Corpus can be accessed via: <https://www.kaggle.com/datasets/kshitizregmi/jobs-and-job-description>.





REFERENCES

- [1] T. Yang, L. Hu, C. Shi, H. Ji, X. Li, and L. Nie, "HGAT: Heterogeneous graph attention networks for semi-supervised short text classification," *ACM Transactions on Information Systems*, vol. 39, no. 3, 2021, doi: 10.1145/3450352.
- [2] S. E. V. S. Pillai and W. C. Hu, "Mobile text misinformation detection using effective information retrieval methods," in *Information Security and Privacy in Smart Devices: Tools, Methods, and Applications*, Palmdale, United States: IGI Global Scientific Publishing, 2023, doi: 10.4018/978-1-6684-5991-1.ch008.
- [3] A. Onan, "Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach," *Computer Applications in Engineering Education*, vol. 29, no. 3, 2021, doi: 10.1002/cae.22253.
- [4] L. Khan, A. Amjad, K. M. Afaq, and H. T. Chang, "Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media," *Applied Sciences*, vol. 12, no. 5, 2022, doi: 10.3390/app12052694.
- [5] D. Mensouri, A. Azmani, and M. Azmani, "Towards an e-commerce personalized recommendation system with KNN classification method," in *International Conference on Advanced Intelligent Systems for Sustainable Development*, 2023, pp. 364-382, doi: 10.1007/978-3-031-26384-2_32.
- [6] D. Pakpahan, V. Siallagan, and S. Siregar, "Classification of e-commerce product descriptions with the TF-IDF and SVM methods," *Sinkron*, vol. 7, no. 4, pp. 2130-2137, 2023, doi: 10.33395/sinkron.v8i4.12779.
- [7] S. Kaddoura, O. Alfandi, and N. Dahmani, "A spam email detection mechanism for English language text emails using deep learning approach," *IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pp. 193-198, 2020, doi: 10.1109/WETICE49692.2020.00045.
- [8] A. Sadia, F. Bashir, R. Q. Khan, A. Bashir, and A. Khalid, "Comparison of machine learning algorithms for spam detection," *Journal of Advances in Information Technology*, vol. 14, no. 2, pp. 178-184, 2023, doi: 10.12720/jait.14.2.178-184.
- [9] S. I. M. Ali, M. Nihad, H. M. Sharaf, and H. Farouk, "Machine learning for text document classification-efficient classification approach," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 1, pp. 703-710, 2024, doi: 10.11591/ijai.v13.i1.pp703-710.
- [10] P. Frasconi, G. Soda, and A. Vullo, "Hidden Markov models for text categorization in multi-page documents," *Journal of Intelligent Information Systems*, vol. 18, pp. 195-217, 2002, doi: 10.1023/A:1013681528748.
- [11] P. Hofmann and Z. Tashman, "Hidden Markov models and their application for predicting failure events," in *Computational Science*, 2020, pp. 464-477, doi: 10.1007/978-3-030-50420-5_35.
- [12] A. P. Noto and D. R. S. Saputro, "Classification data mining with Laplacian smoothing on naïve Bayes method," in *AIP Conference Proceedings*, 2022, doi: 10.1063/5.0116519.
- [13] H. Ren, Y. Zhao, Y. Zhang, and W. Sun, "Learning label smoothing for text classification," *PeerJ Computer Science*, vol. 10, 2024, doi: 10.7717/peerj-cs.2005.
- [14] X. Wu, C. Gao, M. Lin, L. Zang, and S. Hu, "Text smoothing: enhance various data augmentation methods on text classification tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, doi: 10.18653/v1/2022.acl-short.97.
- [15] C. Fattal, L. Labiod, and M. Nadif, "More discriminative sentence embeddings via semantic graph smoothing," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 8-13, 2024.
- [16] B. A. Ojokoh, O. S. Adewale, S. O. Falaki, "Improving on the smoothing technique for obtaining emission probabilities in hidden Markov models," *Oriental Journal of Computer Science and Technology*, vol. 1, no. 1, pp. 15-24, 2008.
- [17] D. Villar, H. Ney, A. Juan, and E. Vidal, "Effect of feature smoothing methods in text classification tasks," in *Proceedings of the 4th International Workshop on Pattern Recognition in Information Systems*, pp. 108-117, 2004.
- [18] F. Hosseynoost and M. Teshnehlal, "Improving hidden Markov model performance in phoneme classification by fuzzy smoothing," in *ESPOCO'05: Proceedings of the 4th WSEAS International Conference on Electronic, Signal Processing and Control*, 2007.
- [19] R. Fechner, J. Dörpinghaus, R. Rockenfeller, J. Faber, "A generic framework for hidden Markov models on biomedical data," *arXiv:2307.13288*, 2023.





- [20] N. M. N. Mathivanan, R. M. Janor, S. A. Razak, and N. A. M. Ghani, "Feature substitution using latent Dirichlet allocation for text classification," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 1, pp. 1087–1098, 2025, doi: 10.14569/IJACSA.2025.01601105.
- [21] D. Jurafsky and J. H. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, Hoboken, United States: Prentice-Hall, Inc., 2000.
- [22] A. Svete, N. Borenstein, M. Zhou, I. Augenstein, and R. Cotterell, "Can transformers learn n-gram language models?," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9851–9867, doi: 10.18653/v1/2024.emnlp-main.550.
- [23] S. Qin, G. Zhang, Y. Wu, and Z. Zhu, "Bayesian grouping-Gibbs sampling estimation of high-dimensional linear model with non-sparsity," *Computational Statistics & Data Analysis*, vol. 203, 2025, doi: 10.1016/j.csda.2024.108072.
- [24] A. E. Gelfand, "Gibbs Sampling," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1300–1304, 2000, doi: 10.2307/2669775.
- [25] E. Senger, Y. Campbell, R. van der Goot, and B. Plank, "KARRIEREWEGE: A large scale career path prediction dataset," in *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 533–545, 2025.

BIOGRAPHIES OF AUTHORS







Norsyela Muhammad Noor Mathivanan     is a Ph.D. student at the School of Mathematical Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Malaysia, under the supervision of Nor Azura Md. Ghani and Roziah Mohd Janor. She is also a lecturer at the University of Wollongong Malaysia. Her research interests are related to natural language processing, text analytics, and machine learning. She can be contacted at email: norsyela.m@uow.edu.my.







Roziah Mohd Janor     is a retired Professor of Statistics who formerly served as the Vice-Chancellor of Universiti Teknologi MARA (UiTM) Malaysia. Her statistical advancements revolve around modeling involving multi-group latent variable models, structural equation modeling, and data envelopment analysis (DEA). She has taught advanced statistical modeling at the postgraduate level, exploring machine learning using R. Her current research interests focus on environmental, social, and governance (ESG) criteria and the sustainable development goals (SDG). She can be contacted at email: roziah.janor@gmail.com.



Shukor Abd Razak     is a Professor and Deputy Vice Chancellor (Research & Innovation) of Universiti Sultan Zainal Abidin (UNISZA), Terengganu, Malaysia. His research interests are on the security issues for Mobile Ad Hoc Networks, Mobile IPv6, Vehicular Ad Hoc Networks, and network security. He also actively conducts several research in digital forensic investigation, wireless sensor networks, and cloud computing. He is the author and co-author for many journals and conference proceedings at national and international levels. He can be contacted at email: shukorrazak@unisza.edu.my.



Nor Azura Md. Ghani     is a Professor at the School of Mathematical Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia, and Former Chair IEEE Computer Society Malaysia Chapter. Currently, she serves as Director at Research Management Center, Universiti Teknologi MARA, Malaysia. Her expertise is in big data, image processing, artificial neural networks, statistical pattern recognition, and forensic statistics. She is the author or co-author of many journals and conference proceedings at national and international levels. She can be contacted at email: azura@tmsk.uitm.edu.my.