Data homogeneity dependent topic modeling for information retrieval

Keerthana Sureshbabu Kashi[1], Abigail A. Antenor[2], Gabriel Isaac L. Ramolete[3], and Adrienne Heinrich[4]

Aboitiz Data Innovation, Goldbell Towers, 47 Scotts Road, Singapore
keerthana.sureshbabu@aboitiz.com[1], abigail.antenor@aboitiz.com[2]
gabriel.ramolete@aboitiz.com[3], adrienne.heinrich@aboitiz.com[4]

Abstract. Different topic modeling techniques have been applied over the years to categorize and make sense of large volumes of unstructured textual data. Our observation shows that there is not one single technique that works well for all domains or for a general use case. We hypothesize that the performance of these algorithms depends on the variation and heterogeneity of topics mentioned in free text and aim to investigate this effect in our study. Our proposed methodology comprises of i) the calculation of a homogeneity score to measure the variation in the data, ii) selection of the algorithm with the best performance for the calculated homogeneity score. For each homogeneity score, the performances of popular topic modeling algorithms, namely NMF, LDA, LSA, and BERTopic, were compared using an accuracy and Cohen's kappa score. Our results indicate that for highly homogeneous data, BERTopic outperformed the other algorithms (Cohen's kappa of 0.42 vs. 0.06 for LSA). For medium and low homogeneous data, NMF was superior to the other algorithms (medium homogeneity returns a Cohen's kappa of 0.3 for NMF vs. 0.15 for LDA, 0.1 for BERTopic, 0.04 for LSA).

Keywords: Topic modeling · Topic Discovery · Technique selection · Information retrieval · NMF · LDA · LSA · BERTopic · Homogeneity · Heterogeneity

1 Introduction

Topic modeling methods are a set of text-mining and information retrieval approaches that have been vastly utilized in natural language processing (NLP) for segmented text analysis. It serves best for several use cases: organizing vast volumes of text, the retrieval of information from unstructured or semi-structured documents, feature extraction through creating representations of latent classes, and clustering of documents. Topic model algorithms scan a set of documents called a corpus, examine how words and phrases co-occur, and group the words that best describe the document. These words often represent a coherent theme, categorized in a topic. A variety of topic models, from Bayesian probabilistic topic models (BPTMs) such as Latent Dirichlet Allocation (LDA) [1] to neural topic models (NTMs) such as variational autoencoders [2,3] and generative adversarial nets (GANs) [4,5,6], have been utilized in language models, text summarization, and text generation, identifying concealed semantics in heaps of semantic data [7,8].

Numerous researchers leverage on topic modeling techniques to easily derive topics from corpora. Often, the corpora used for topic modeling contain enormous amounts of text data resulting from open-ended survey questions and manual logging of complaints. For instance, Nguyen and Ho aim to evaluate how Latent Dirichlet Allocation (LDA) analyzes experience in customer service [9]. Various algorithms have attempted to demystify social media posts on Facebook and Twitter according to human interpretation [10,11,12].

While topic modeling can help automate the conception of ideas and abstractions, it is inevitable for these algorithms to pick up on noise or linguistically meaningless correlations of words found in documents. Recent papers which review topic modeling methods on different text lengths and domains suggest that not all topic modeling algorithms perform consistently. Vayansky and Kumar suggest that LDA, while versatile and often gravitated upon, does not befit complex data relationships [13]. Sbalchiero and Eder experiment the capability of different algorithms on different text

lengths [14]. Hu et al. even propose an interactive topic modeling framework which inputs user feedback on initial model outputs to further optimize the topic modeling solutions [15]. Validation metrics such as accuracy, Cohen's kappa coefficient, and coherence scores can be used to optimize these models and may return numerically sufficient performance, but vague semantic outputs may hinder or misguide human interpretation.

To illustrate this further, consider the topic keywords generated with Non-negative Matrix Factorization (NMF) [10], a commonly known topic model, and its subjective observation as shown in Table 1. This is applied to a set of complaints sourced from a dataset further discussed in Section 3.1. It can be observed that when data consisting of only two complaint categories is considered ('Late payment' and 'Fraud'), a user can easily assign a description and action point to the topic inferred from the topic key words.

Topic ID	Topic Key words	Topic Description	Impact
0	payment, interest, fee, late	Late payment	Complaints are sent to late payments dept.
1	report, theft, identity theft	Fraud	Complaints are sent to fraud dept.
2	card, charge, credit, fraudulent	Fraud	Complaints are sent to fraud dept.

Table 1. Keywords generated by NMF for two complaint categories

However, in Table 2, using the same NMF topic modeling technique now on three complaint categories ('Late payment', 'Fraud' and 'Closing account'), some topic keyword groups can result in mixed context words; in Topic ID 0, the mixed content seems to address both 'fraud' and 'late payment'. It may be seen that as the variation in the data increases, the value of the interpretations arising from the topic keywords can alter.

Topic ID	Topic Key words	Topic Description	Impact
0	charge, late, payment, merchant, card, fraud	Mixed context words.	Some of the complaints in this cluster are sent to late payments instead of fraud dept.
1	statement, payment, past due, found late	Late payment	Complaints are sent to late payments dept.
2	account, credit, close, credit card, close account	Closing account	Complaints are sent to closing accounts dept.
3	account, information, card, contact, fraudulent	Fraud	Complaints are sent to fraud dept.
4	payment, fee, late, late fee, pay, paid, interest, due	Late payment	Complaints are sent to fraud dept.

Table 2. Keywords generated by NMF for three complaint categories

Similar examples of the issue illustrated above could be seen in the contrasting best performance models for large text [16], sentence classification [17], and short text topic modeling [18]. Although previous papers can suggest that techniques such as LDA, NMF, and even Latent Semantic Analysis (LSA) perform best for certain scenarios [17,18], no technique appears to have the best average performance on a variety of data assortments and purposes.

The goal of this paper is to investigate and propose a topic modeling selection method that is dependent on the data homogeneity. With this in place, the best topic modeling techniques can be identified for particular scenarios given a broad set of data variations in topic-laden corpora. The text content variety is measured based on a homogeneity score. Suggesting the best performing algorithm for certain scenarios will help researchers and organizations mitigate the risk of misidentifying topics on various corpora, which consequently lessens wasted resources and time.

In this paper, Section 2 describes the state-of-the-art approaches regarding preprocessing steps, the intuition behind different topic modeling algorithms, and the formulation of a homogeneity score used in the experimentation. Section 3 details the experimental setup while Section 4 discusses the results of the experiment. Section 5 concludes on the observed performance of the algorithms and how our proposed method can help users in selecting an appropriate algorithm based on data homogeneity.

2 Review of Related Literature

2.1 Prevalence of topic modeling in scientific and corporate settings

Topic modeling has applications in many fields, such as literature review, software engineering, and linguistic science. For example, organizing qualitative works, including opinion and sentiment analysis for marketing, discourse analysis in media studies, and blog usage for sociological studies, have shown to improve through the usage of an Interval Semi-supervised Latent Dirichlet Allocation (ISLDA) approach, with the help of defining a term frequency-inverse document frequency (TF-IDF) coherence score [19]. A variety of authors, such as DiMaggio et al. [20], Grimmer [21], Quinn et al. [22], Jockers and Mimno [23], Baum [24], Elgesem et al. [25], have shown the usage of topic modeling approaches for identifying concepts, subjects of discussion, or sentiments through data such as newspapers, press releases, speeches, books, blogs, and tweets. Other researchers have also shown topic modeling expertise with respect to software traceability [26], coupling of classes in object-oriented software systems [27], and mining source code histories [28,29].

Outside of natural and social sciences, topic modeling use cases are also prevalent in corporate settings. Voice-of-customer processing with topic modeling for extracting actual customer needs has been remarked by Özdağoğlu et al. [30]. Barravecchia et al. [31]. At the same time, topic modeling has also been observed to define service quality attributes, content marketing topics, sentiment analysis towards products, and issues in customer reviews [32,33,34,35]. Other applications of topic modeling were studied by Asmussen & Møller, focusing on the Latent Dirichlet allocation model in particular . A bibliometric analysis performed over topic modeling-focused researched papers in 2000-2017 observed that while computer science and engineering comprised of the majority of publications, topic modeling has emerged in other subjects such as medical informatics, telecommunications, business economics, operations research, biochemistry and molecular biology, remote sensing, and photographic technology [36].

2.2 Homogeneity and heterogeneity

The heterogeneity of texts in a corpus can be perceived using the concept of entropy, a measure of the level of complexity and randomness of any system with various interdependent components [37]. The formula for Shannon's Entropy is shown in Eq. (1).

$$S = -\sum_{i=1}^{T} p_i ln(p_i), \tag{1}$$

where S indicates entropy, p denotes the probabilities of occurence, T corresponds to types of different components, and i conform to the number of tokens of a text [37,38].

In this system, the texts are observed as communicative functions that drive the evolution of the corpus' entropic process. The complex nature of the corpus is shown in various aspects such as semantic, syntactic, discourse, chaotic interactions, and possible self-organizational mechanisms. In essence, homogeneity produces criteria for classifying texts with common linguistic characteristics. They may share the same features, such as lexical distance, word choice, or genre [39].

Using Shannon's Entropy, a homogeneity score denoted as H is calculated. A lower homogeneity score is computed when the topic variation increases with the number of categories. The formula is shown in Eq. (2).

$$H(X) = \sum_{i=1}^{N} P(x_i) \cdot \log \frac{1}{P(x_i)} \tag{2}$$

Considering a dataset X with elements x_i, fitting the formula to this use case, let N_C be the number of categories (labels) and C_i be a set of data points in a category (label). Defining probability for each element $P(x_i) = \frac{|C_i|}{N}$, the entropy equation can be transformed to Eq. (3).

$$H(X) = \sum_{i=1}^{N_C} \frac{|C_i|}{N} \cdot \log \frac{N}{|C_i|} \tag{3}$$

With this, the homogeneity score of the data set could be derived as the inverse of its Shannon's entropy shown in Eq. (4).

$$h(x) = \begin{cases} \infty & \text{if } H(X) = 0 \\ \frac{1}{H(X)} & \text{otherwise} \end{cases} \tag{4}$$

If $h(X) = \infty$, then X is fully homogeneous. So, the higher the value of $h(X)$, the data set X is said to be more homogenous. A lower homogeneity score is computed when the topic variation increases with the number of categories.

2.3 Preprocessing Techniques

Before modeling, preprocessing techniques are used to transform text inputs into machine-readable elements. Such steps are described below:

Tokenization Tokenization is a preprocessing technique that cuts text into sentences and breaks them down into words. Uppercase letters are converted to lowercase letters, then punctuation marks are removed [40].

Part-of-Speech (POS) Tagging This preprocessing technique tags words by their context in a sentence, whether they are nouns, verbs, adjectives, or other forms [41]. In this paper, POS taggers are used to study large tagged text corpora and make an abstract level of analysis from reliable low-level information.

Word Lemmatization Lemmatization is a linguistic term defined as grouping words with the same root or lemma. While there are concerns that the real meaning of words in the sentence could be out of context, lemmatization preempts this complication. Therefore, it is a preferred option for topic modeling compared to stemming. Stemming only cuts off the derived affixes of the word, thus losing the true intention of the word in the sentence [40].

Term Frequency-Inverse Document Frequency (TF-IDF) Words with a high relative term frequency value have high importance in documents. TF-IDF illustrates the correlation and relationship of words on the corpus by calculating similar degrees among documents and deciding on search rank. With this, getting the inverse of document frequency means we are identifying the terms with high importance [42].

2.4 Modeling Techniques

With the advent of data-driven approaches toward understanding customer feedback efficiently, various topic modeling techniques have been developed to optimize the topic identification problem in information retrieval. State-of-the-art and commonly used topic algorithms include Non-negative Matrix Factorization, Latent Dirichlet Allocation, Latent Semantic Analysis, and BERTopic [10].

NMF Non-negative Matrix Factorization (NMF) is a linear algebraic method on decompositional non-probabilistic algorithms [10]. The goal of NMF is to decrease the data dimension and to identify its principal components to represent the chaotic and complex data as an interpretable information cluster [43].

NMF breaks down the input term-document matrix V to term-topics matrix W and topics-document matrix M. This factorization was initiated by Paatero and Tapper from 1994 to 1997 [44], commenced with Lee and Seung in 1999 to 2001 [45]. It incorporates non-negative constraints that enhance semantic explainability with the parts-based representation [43]. See Eq. (5).

$$V_{i\mu} \approx (WM)_{i\mu} = \sum_{a=1}^{r} W_{ia} M_{a\mu} \tag{5}$$

Compared to LDA, this method utilizes Term Frequency-Inverse Document Frequency (TF-IDF) to represent the importance of each word in a corpus and not just solely rely on the raw word frequency. According to Egger and Yu, NMF outperforms LDA by producing clearly defined topics but is relatively standard compared to algorithms working on word embeddings [10].

LDA Latent Dirichlet Allocation (LDA) is an unsupervised generative probabilistic model of a corpus, representing topics by word probabilities [1]. LDA assumes that each document is a probabilistic distribution over latent subjects and that topic distribution in all documents share a common Dirichlet prior. Each latent topic in the LDA model is also a probabilistic distribution over words. Thus, the word distributions of issues share a common Dirichlet prior.

LSA Latent Semantic Analysis (LSA) is another older method for extracting the contextual meaning of words and phrases in sizeable text compilations. Known for reducing the dimensionalities of matrices during information retrieval procedures [46], LSA aims to take advantage of an implicit latent semantic form found in sentences and documents, estimated through singular-value decomposition to remove random "noise". When larger portions of data are in play, terms that may not appear in an individual document may still be linked to the document semantically. [47].

BERT and BERTopic The Bidirectional Encoder Representations from Transformers (BERT), classified as a transformer-based NLP technique, aims to improve on other fine-tuning approaches in applying pre-trained language representation [48]. Variations of BERT went with both clustering procedures and a class-based variation of the TF-IDF algorithm to create logical topic representations, coined BERTopic by Grootendorst in 2022 [49]. BERTopic utilizes pre-training or creating general-purpose language representation models fine-tuned on smaller-data NLP activities. These activities include next sentence prediction, sentiment analysis, question-answer responses, and other tasks found in the General Language Understanding Evaluation (GLUE) benchmark [50].

The architecture of BERT follows a self-attention mechanism [51] and is pre-trained using a Masked Language Model (MLM) and Next Sentence Prediction (NSP), two unsupervised tasks. Each document is converted to its embedding representation by clustering a version of the embeddings with reduced dimensionality and extracting the topics with a class-based TF-IDF iteration. This accounts for the clusters of documents as shown in Eq. (6).

$$W_{t,c} = tf_{t,c} \cdot log(1 + \frac{A}{tf_t}), \tag{6}$$

wherein the frequency of a term t inside a class c, or a collection of documents transformed into a single document cluster, is modeled. The logarithm of the mean words of each class A divided by t across all classes made, adding one to positive output values, is used to measure the amount of information the term t provides to the class c [49].

2.5 Evaluation Metrics

To determine the most suitable topic modeling algorithm for various levels of topic variation, accuracy and Cohen's kappa coefficient were used as evaluation metrics and to find the optimal number of topics for various variations in the data set coherence score was applied.

Coherence Score A coherence score is applied to find the degree of semantic similarity between high-scoring words in the topic. To calculate, the formula used is shown in Eq. (7).

$$C(t : V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{7}$$

where $D(v)$ denotes the document frequency of word type v, and $D(v, v')$ corresponds to the co-document frequency of the word types v', and v, and $V^{(t)}$ depicts the list of M most probable words in topic t. When the most significant and comprehensive words found for a certain topic have a high rate of co-occurrence, this results in high coherence score [52].

Accuracy A common evaluation metric for text classification is accuracy. It measures the correct classifications over the total number of classifications. The formula for accuracy is shown in Eq. (8).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

where TP stands for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative [53].

Cohen's Kappa Coefficient This formula calculates the kappa coefficient, denoted by κ shown in Eq. (9).

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{9}$$

where Pr(a) is the proportion of units in which the raters agreed, and Pr(e) is the proportion for which the agreement is expected by likelihood, as found in Eq. (9). For topic modeling purposes,

the agreements could be interpreted as how close the predicted value and the true value concur with one another. As κ positively increases, there is a larger chance of the predicted value and the ground truth agreeing. If there is no agreement, then κ is negative [54]. Along with accuracy, Kappa's coefficient is an additional evaluation metric for classifiers as it considers the presence of imbalanced data [55]. Therefore in this experiment, the better-performing classifiers should have higher values of κ [56].

3 Experimental Setup

3.1 Data source

The consumer complaint data used in the experiments was sourced from the Consumer Financial Protection Bureau of 2012-2017, particularly the "Credit Card complaints.csv" dataset [57]. It contains thirty (30) unique categories such as Billing disputes, Balance transfer, Delinquent accounts, Identity theft / Fraud / Embezzlement, and Late fee in the 'Issue' column. These were considered annotated compliant categories, while the topic modeling algorithms used the 'Consumer complaint narrative' column to generate topic keywords.

3.2 Methodology

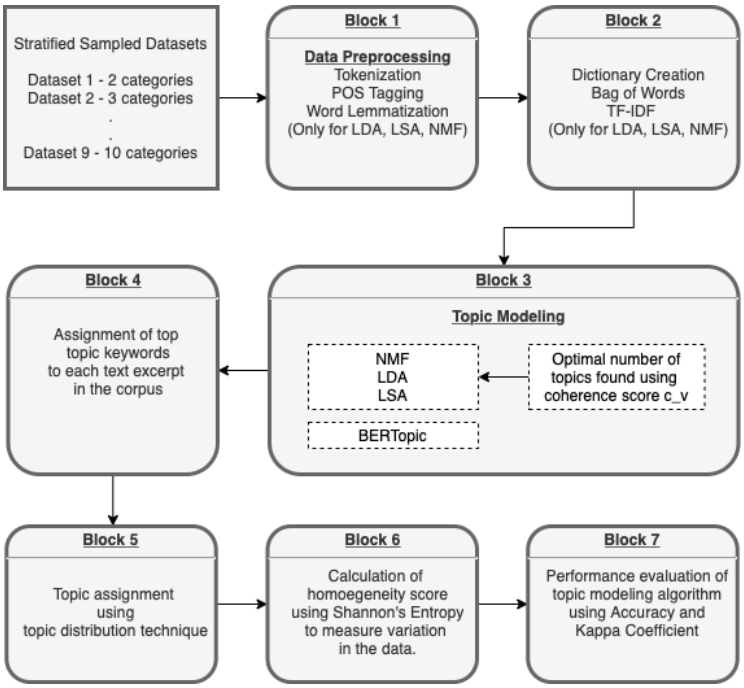

Fig. 1. Experimental Setup

Figure 1 explains the setup used in this experiment. Nine (9) stratified sampled data sets, containing complaint categories between two and ten, are obtained from the complaints dataset discussed in Section 3.1. After the dataset is created, a set of preprocessing steps, shown in Block 1, is applied to each stratified dataset sample. These include Tokenization, Parts-of-Speech Tagging, and Word Lemmatization. The preprocessing steps mentioned are only required for NMF, LDA, and LSA, not for BERTopic, as the latter uses pre-trained language and representation models.

After data is preprocessed, a dictionary and Bag of Words (BOW) are created, containing all words in the corpus represented as keys and the frequency of the occurrence of the words identified as values. Following the BOW is the TF-IDF identification, which carries information on which words are more critical. After documents are vectorized, the topic modeling method is implemented. In our case, the topic modeling techniques NMF, LDA and LSA utilize the coherence score calculation to find the optimal number of topics, while BERTopic internally calculates the optimal number of topics to be used.

The topic keywords, generated by executing the topic modeling algorithms, are then mapped against the annotated categories as shown in Blocks 4 and 5. As an example from our dataset, topics such as Advertising and marketing, Closing/Cancelling Accounts, Identity Theft, Rewards/Memberships, will be assigned based on the maximum value of the distribution among all annotated topics within each cluster formed by the topic modeling algorithms.

Advertising and Marketing	Closing /Cancelling account	Identity theft / Fraud	Rewards /Memberships	Categories with highest frequency of text excerpts	Topic ID
7%	8%	3%	**10%**	Rewards /Memberships	0
2%	8%	**11%**	0%	Identity theft /Fraud	1
0%	**36%**	3%	2%	Closing /Cancelling account	2

Table 3. Sample distribution of annotated categories for each generated topic by algorithm NMF. The first four columns denote the text excerpt frequency.

Topic ID	Topic Keywords	Subjective Observation
0	express, american express, american, point, membership reward, spend, card, offer, express credit, receive	Talks about Rewards and membership
1	capital, fraud, report, charge, bureau, line	Talks about Fraudulent activities
2	account, close, balance, account, open, without, credit, citibank	Talks about account closure

Table 4. Top keywords generated by NMF for each topic

An example of keywords generated by NMF is shown in Table 3. The first four columns show the distribution of existing annotated categories with each cluster of documents. The cluster of documents formed by the generated Topic ID 0 has 10% of total Reward/Memberships complaints/inquiries, 8% of inquiries on closing accounts, 7% on Advertising and marketing, and 3% of Identity theft/Fraud complaints. Since the majority of complaints within this cluster of Topic ID 0 are inclined towards inquiries on Reward/Memberships, the generated topic is assigned to Rewards/Memberships.

To verify the topic assigned to each topic ID, the model-generated top topic keywords, as shown in Table 4, are also manually interpreted using eyeballing techniques. Table 3 shows that the manual inference is matched with the aforementioned topic allocation using the distribution method.

Steps from Block 1 to Block 6 were ran nine (9) times for all homogeneity scores, generated by the equations in Section 2.2. This set of iterations was performed for each topic modeling algorithm for further comparisons. Accuracy and Cohen's kappa coefficient, found in Eq. (8) and (9), were used to evaluate the performance of each topic modeling algorithm, as shown in Block 7.

4 Results and Discussion

4.1 Outcome of experiments

Table 5 shows how the performance of NMF, LDA, LSA, and BERTopic vary depending on heterogeneity of the data. Classification accuracy and Kappa's coefficient as calculate. Eq. (8) and (9) are used as the metrics to evaluate the performance of these algorithms. The accuracy and homogeneity scores are normalized to set a common scale for a new diversified set of corpora. It can be seen that, when the homogeneity score changed from 1 to 0.47, the accuracy for LDA dropped by 55%, BERTopic by 30%, LSA by 14%, and NMF by 17% . Even though LSA has the most negligible change, it is relatively less performing than NMF, to begin with. Furthermore, as the homogeneity scores changed from 0.47 to 0.32, the normalized accuracy for LSA and BERTopic continued to plummet. For LSA, it dropped by 28%. For BERTopic it decreased by 25%. These drops in normalized accuracy are highlighted in red text in Table 5.

	Kappa Coefficient				Normalized Accuracy			
Normalized Homogeneity Score	**LSA**	**BERTopic**	**LDA**	**NMF**	**LSA**	**BERTopic**	**LDA**	**NMF**
1.00	0.06	0.42	0.25	0.35	**0.77**	**1.00**	**0.82**	**0.91**
0.47	0.14	0.24	0.04	0.35	**0.63**	**0.70**	**0.27**	**0.74**
0.32	0.06	0.24	0.07	0.35	**0.35**	**0.45**	0.23	0.68
0.15	0.03	0.16	0.07	0.39	0.16	0.35	0.26	0.65
0.11	0.04	0.10	0.15	0.30	0.12	0.26	0.32	0.50
0.09	0.09	0.12	0.13	0.35	0.20	0.11	0.15	0.55
0.05	0.03	0.15	0.16	0.28	0.11	0.19	0.26	0.41
0.02	0.02	0.13	0.07	0.26	0.00	0.19	0.04	0.37
0.00	0.04	0.13	0.05	0.23	0.03	0.13	0.01	0.33
Average Score	**0.06**	**0.19**	**0.11**	**0.32**	**0.26**	**0.38**	**0.26**	**0.57**

Table 5. Homogeneity and accuracy values for NMF, LDA, LSA and BertTopic

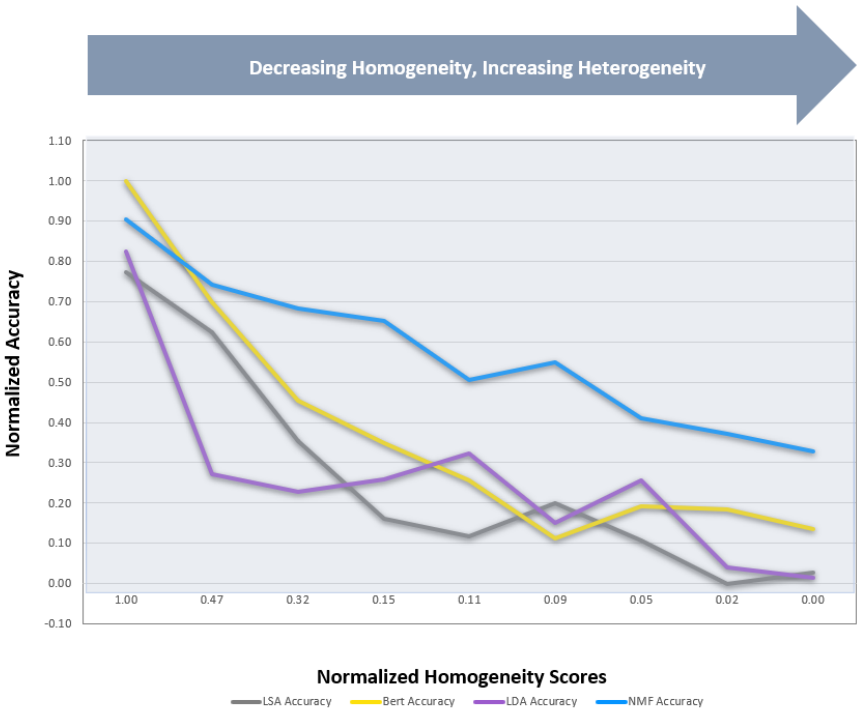

Fig. 2. Normalized homogeneity vs Normalized accuracy for the investigated topic modeling algorithms. BERTopic outperforms other algorithms for highly homogeneous data and NMF is superior to other algorithms for low-medium homogeneity scores

The results shown in Figure 2 and Figure 3 add to this claim, illustrating that as the homogeneity score decreases, the normalized accuracy and Kappa's coefficient also decrease. However, there are exceptions. For instance, when the homogeneity scores range between 0.15 and 0.05, the normalized accuracy bounces back in Figure 2 for some algorithms.

As seen in Figure 2, NMF seems to be the most accurate and suitable algorithm as topic variation increases. While its normalized accuracy decreases, it stays above the other algorithms tested somewhere after the 0.47 mark of the normalized homogeneity. Moreover, the gap for the normalized accuracy between NMF and the others after this point ranges from 15% to 47%. This significant difference further solidifies the claim that NMF is better in increasingly varying topics. However, when the homogeneity score is at its highest value, or when the variability in the topic is least, BERTopic outperforms NMF by 9%. Thus for topics with more similarities, it can be claimed that BERTopic is the best algorithm to use, followed by NMF, LDA, then LSA.

The performance of the algorithms between homogeneity scores of 0.05 and 0.15 shows some insight. While NMF still surpasses the others, LDA appears to be the second best option, beating BERTopic and LSA. It can therefore be concluded that for the corpus with moderate topic variations, NMF and LDA are the desirable algorithms to use.

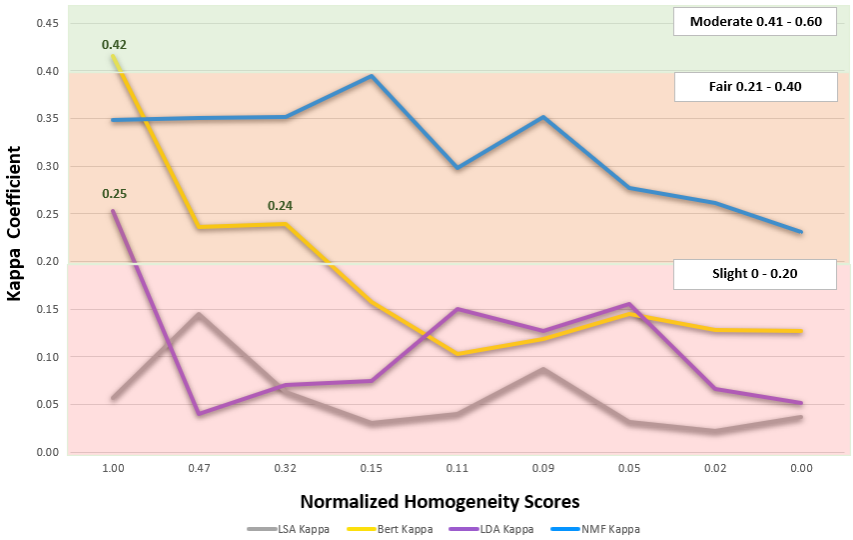

Fig. 3. Normalized homogeneity vs Kappa coefficient for investigated topic Modeling algorithms.

In addition to accuracy, the Kappa coefficient is an added evaluation metric. As seen in Section 2.5, the Kappa coefficient κ helps users evaluate the performance of each topic modeling algorithm, becoming more reliable when there is an imbalance in the classes. As seen in Figure 3, the graph conveyed three ranges. According to Landis and Koch, [58], a kappa value less than 0 indicates no agreement, 0-0.20 draws as slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1 as the perfect agreement between the predicted values and the ground truth.

For datasets similar to the one at hand, NMF still is the desirable algorithm to use after the 0.47 mark in the homogeneity score. It consistently exhibits fair agreement, meaning that the predicted and actual values are concurring equitably. For the highest homogeneity score tested, BERTopic seems to provide a good Kappa score slightly above the fair range and over the minimum point in the moderate range, thus it can be used more than the NMF at this stage. BERTopic could be used until the 0.32 homogeneity score but not beyond. It already went under the slight range with LSA and LDA. LSA is almost always in the slight range, therefore this could be the last option to take across the whole range.

In summary, the accuracy of topic modeling algorithms must be viewed as dependent on the present homogeneity in the data. For this dataset, NMF is the best algorithm for low and medium homogeneity. But alternatively, LDA could be used for medium homogeneity and BERTopic for high homogeneity. BERTopic performs best for the most homogeneous data, but the performance deteriorates as the homogeneity in data decreases.

Alongside the comparison of evaluation metrics, a sample set of tables of an “Identity Theft / Fraud / Embezzlement” topic with its generated topic keywords are shown for a homogeneity score of 1.00 and a homogeneity score of 0.32. Aside from the accuracy and Kappa coefficient, a subjective observation will aid in validating the effectiveness of using the recommended algorithms. The subjective validation will also support the use of suitable performance measures for this study.

In Table 6, BERTopic and NMF are able to give 100% correct interpretation for the topic ’Identity theft/ Fraud/ Embezzlement.’ under a homogeneity score of 1.00. However, BERTopic performs

Homogeneity Score: 1.00			
Topic	**Topic Keywords**	**Subjective Observation**	**Remarks**
NMF			
Identity theft / Fraud /Embezzlement	account, report, credit, identity, debt, credit report, reporting, open, theft, identity theft	Correct Interpretation	100 % topics gives correct interpretation
Identity theft / Fraud / Embezzlement	card, charge, credit, credit card, company, call, fraudulent, receive, fraud, card company	Correct Interpretation	
Identity theft / Fraud / Embezzlement	capital one, capital, one, one credit, theft, reporting, merchandise, resolve, secure	Correct Interpretation	
LDA			
Identity theft / Fraud / Embezzlement	credit, account, card, charge, bank, call, get, information, would, number	Cannot Interpret	Only 25% of topics give correct interpretation
Identity theft / Fraud / Embezzlement	card, credit, one, payment, account, make, capital, month, sent, state	Cannot Interpret	
Identity theft / Fraud / Embezzlement	payment, balance, would, tv, delivery, purchase, make, best, pay, check	Cannot Interpret	
Identity theft / Fraud / Embezzlement	card, credit, account, charge, call, would, receive, time, fraud, told	Correct Interpretation	
LSA			
Identity theft / Fraud / Embezzlement	credit, card, account, charge, call, payment, would	Cannot Interpret	Zero topic with correct interpretation
BerTopic			
Identity theft / Fraud / Embezzlement	credit, card, account, not, fraud, charges, bank	Correct Interpretation	100% of topics give correct interpretation and clusters all the fraud into one definite cluster, instead of creating many clusters like NMF

Table 6. Topic Keywords for Identity theft / Fraud / Embezzlement with Homogeneity Score of 1.00

best as it could give precisely one definite cluster with correct interpretation from the keywords generated, while NMF gave three clusters with proper interpretation. On the other end, LDA showed 25% correct interpretations and LSA produced keywords that are not as interpretable. In Table 7, which showcases topic keywords for the same topic for the homogeneity score of 0.32, it is apparent that NMF performs the best among LDA, LSA, and BERTopic. NMF gave two-thirds topics worth of correct interpretations, while the three others had either mixed words or uninterpretable keywords.

Homogeneity Score: 0.32			
Topic	**Topic Keywords**	**Subjective Observation**	**Remarks**
NMF			
Identity theft / Fraud / Embezzlement	call, say, would, time, bill, money, get, phone, pay	Contains Mixed Words	66% topics give correct interpretation
Identity theft / Fraud / Embezzlement	capital, capital one, one, one bank, fraud, information, judgment, someone, wallet, charge back	Correct Interpretation	
Identity theft / Fraud / Embezzlement	identity, theft, identity theft, report, debt, victim, open, account, victim identity, information	Correct Interpretation	
LDA			
Identity theft / Fraud / Embezzlement	credit, account, card, report, charge, bank, call, close, fraudulent, receive	Contains Mixed Words	Zero topics with correct interpretation
LSA			
Identity theft / Fraud / Embezzlement	card, credit, bank, account, charge, call, one, capital, amex	Cannot Interpret	Zero topics with correct interpretation
Identity theft / Fraud / Embezzlement	one, charge, american, express, bank, call, paypal, time, make, capital	Cannot Interpret	
BERTopic			
Identity theft / Fraud / Embezzlement	credit, card, account, had, would, payment	Cannot Interpret	Zero topics with correct interpretation

Table 7. Topic Keywords for Identity theft / Fraud / Embezzlement with Homogeneity Score of 0.32

The boundaries and algorithm rankings may vary for other data sets unrelated to bank customer complaints or inquiries. A similar experiment should be conducted to understand the generalizability of this study's findings.

4.2 Recommended Data Homogeneity Dependent Topic Modeling Process

Figure 4 shows the recommended steps a user new to a dataset can apply to determine how homogeneous the data at hand is and which topic modeling algorithm should be considered depending on the homogeneity score. If the user's data is more homogeneous with a normalized homogeneity score greater than 0.5, BERTopic is recommended for similar datasets to the one discussed in this work. If the data moves to a homogeneity score of less than 0.5, NMF can be applied, followed by LDA. LSA is the last choice to take.

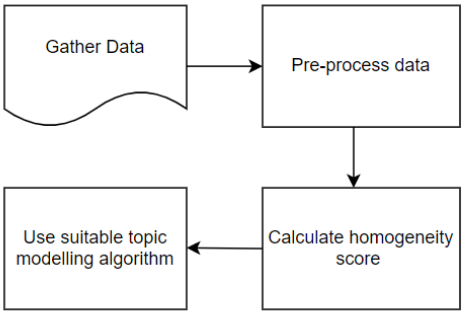

Fig. 4. Proposed methodology to select a topic modeling algorithm for a new data

5 Conclusion

Despite many years of advancements in topic modeling algorithms, there are still apparent drawbacks dealing with conceptually spurious or multi-context words generated by the topic modeling algorithms. These constraints escalate with increasing topic variation in given text corpora, which hinder coherence and usability especially in organizations with large free-text such as in banking and customer feedback. To address this limitation, a homogeneity score based on Shannon's entropy was formulated to capture the topic variation in a data set for each set of annotated categories. The performances of four commonly used state-of-the-art topic modeling algorithms, namely Non-negative Matrix Factorization (NMF), Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Bidirectional Encoder Representation from Transformers for Topic modeling (BERTopic), were evaluated using Accuracy and Cohen's kappa coefficient scores on different levels of data variation explained by the calculated homogeneity scores.

From the results above, it can be concluded that there is no single topic modeling algorithm among the four that perfectly works for increasing topic variation/heterogeneity. Comparatively, BERTopic outperforms other algorithms (Cohen's kappa of 0.42 vs. 0.06 for LSA) for high data homogeneity. For medium and low homogeneous data, NMF is superior to the other algorithms (medium homogeneity returns a Cohen's kappa of 0.3 for NMF vs. 0.15 for LDA, 0.1 for BERTopic, 0.04 for LSA). The methodology described in this paper aims to help users calculate the topic variation in their dataset which is derived from the proposed homogeneity score. The user can choose among most widely used algorithms, not limited to the four topic modeling techniques aforementioned, to get the best coherent interpretation from the topic keywords generated.

References

1. H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, 2019.
2. A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," *arXiv preprint arXiv:1703.01488*, 2017.
3. W. Joo, W. Lee, S. Park, and I.-C. Moon, "Dirichlet variational autoencoder," *Pattern Recognition*, vol. 107, p. 107514, 2020.
4. A. Jabbar, X. Li, and B. Omar, "A survey on generative adversarial networks: Variants, applications, and training," *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–49, 2021.
5. J. Glover, "Modeling documents with generative adversarial networks," *arXiv preprint arXiv:1612.09122*, 2016.
6. R. Wang, D. Zhou, and Y. He, "Atm: Adversarial-neural topic model," *Information Processing & Management*, vol. 56, no. 6, p. 102098, 2019.
7. H. Zhao, D. Phung, V. Huynh, Y. Jin, L. Du, and W. Buntine, "Topic modelling meets deep neural networks: A survey," *arXiv preprint arXiv:2103.00498*, 2021.
8. T.-N. Doan and T.-A. Hoang, "Benchmarking neural topic models: An empirical study," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4363–4368, 2021.
9. H.-H. Nguyen and Thanh, "Analyzing customer experience in hotel services using topic modeling," *Journal of Information Processing Systems*, vol. 17, pp. 586–598, 6 2021.

10. R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in Sociology*, vol. 7, 2022.
11. S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, and X. He, "Interpreting the public sentiment variations on twitter," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 5, pp. 1158–1170, 2013.
12. Z. Xu, Y. Liu, J. Xuan, H. Chen, and L. Mei, "Crowdsourcing based social media data analysis of urban emergency events," *Multimedia Tools and Applications*, vol. 76, no. 9, pp. 11567–11584, 2017.
13. I. Vayansky and S. A. Kumar, "A review of topic modeling methods," *Information Systems*, vol. 94, p. 101582, 2020.
14. S. Sbalchiero and M. Eder, "Topic modeling, long texts and the best number of topics. some problems and solutions," *Quality & Quantity*, vol. 54, no. 4, pp. 1095–1108, 2020.
15. Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, "Interactive topic modeling," *Machine learning*, vol. 95, no. 3, pp. 423–469, 2014.
16. P. Suri and N. R. Roy, "Comparison between lda nmf for event-detection from large text stream data," in *2017 3rd International Conference on Computational Intelligence Communication Technology (CICT)*, pp. 1–5, 2017.
17. A. Anantharaman, A. Jadiya, C. T. S. Siri, B. N. Adikar, and B. Mohan, "Performance evaluation of topic modeling algorithms for text classification," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 704–708, 2019.
18. J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short text topic modeling techniques, applications, and performance: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 3, pp. 1427–1445, 2022.
19. S. I. Nikolenko, S. Koltcov, and O. Koltsova, "Topic modelling for qualitative studies," *Journal of Information Science*, vol. 43, no. 1, pp. 88–102, 2017.
20. P. DiMaggio, M. Nag, and D. Blei, "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding," *Poetics*, vol. 41, no. 6, pp. 570–606, 2013.
21. J. Grimmer, "A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases," *Political Analysis*, vol. 18, no. 1, pp. 1–35, 2010.
22. K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev, "How to analyze political attention with minimal assumptions and costs," *American Journal of Political Science*, vol. 54, no. 1, pp. 209–228, 2010.
23. M. L. Jockers and D. Mimno, "Significant themes in 19th-century literature," *Poetics*, vol. 41, no. 6, pp. 750–769, 2013.
24. D. Baum, "Recognising speakers from the topics they talk about," *Speech Communication*, vol. 54, no. 10, pp. 1132–1142, 2012.
25. D. Elgesem, I. Feinerer, and L. Steskal, "Bloggers' responses to the snowden affair: Combining automated and manual methods in the analysis of news blogging," *Computer Supported Cooperative Work (CSCW)*, vol. 25, no. 2, pp. 167–191, 2016.
26. H. U. Asuncion, A. U. Asuncion, and R. N. Taylor, "Software traceability with topic modeling," in *2010 ACM/IEEE 32nd International Conference on Software Engineering*, vol. 1, pp. 95–104, IEEE, 2010.
27. M. Gethers and D. Poshyvanyk, "Using relational topic models to capture coupling among classes in object-oriented software systems," in *2010 IEEE international conference on software maintenance*, pp. 1–10, IEEE, 2010.
28. S. W. Thomas, "Mining software repositories using topic models," in *Proceedings of the 33rd International Conference on Software Engineering*, pp. 1138–1139, 2011.
29. K. Tian, M. Revelle, and D. Poshyvanyk, "Using latent dirichlet allocation for automatic categorization of software," in *2009 6th IEEE International Working Conference on Mining Software Repositories*, pp. 163–166, IEEE, 2009.
30. G. Özdağoğlu, A. Kapucugil-Ikiz, and A. F. Celik, "Topic modelling-based decision framework for analysing digital voice of the customer," *Total Quality Management & Business Excellence*, vol. 29, no. 13-14, pp. 1545–1562, 2018.
31. F. Barravecchia, L. Mastrogiacomo, and F. Franceschini, "Digital voice-of-customer processing by topic modelling algorithms: insights to validate empirical results," *International Journal of Quality & Reliability Management*, 2021.
32. K. Ding, W. C. Choo, K. Y. Ng, and S. I. Ng, "Employing structural topic modelling to explore perceived service quality attributes in airbnb accommodation," *International Journal of Hospitality Management*, vol. 91, p. 102676, 2020.
33. Y. Putranto, B. Sartono, and A. Djuraidah, "Topic modelling and hotel rating prediction based on customer review in indonesia," *International Journal of Management and Decision Making*, vol. 20, no. 3, pp. 282–307, 2021.

34. A. Gregoriades, M. Pampaka, H. Herodotou, and E. Christodoulou, "Supporting digital content marketing and messaging through topic modelling and decision trees," *Expert systems with applications*, vol. 184, p. 115546, 2021.
35. M. J. Sánchez-Franco, F. J. Arenas-Márquez, and M. Alonso-Dos-Santos, "Using structural topic modelling to predict users' sentiment towards intelligent personal agents. an application for amazon's echo and google home," *Journal of Retailing and Consumer Services*, vol. 63, p. 102658, 2021.
36. X. Li and L. Lei, "A bibliometric analysis of topic modelling studies (2000–2017)," *Journal of Information Science*, vol. 47, no. 2, pp. 161–175, 2021.
37. M. M. Angel and J.-M. Rey, "On the role of shannon's entropy as a measure of heterogeneity," *Geoderma*, vol. 98, no. 1-2, p. 1–3, 2000.
38. A. A. Torres-García, O. Mendoza-Montoya, M. Molinas, J. M. Antelis, L. A. Moctezuma, and T. Hernández-Del-Toro, "Chapter 4 - pre-processing and feature extraction," in *Biosignal Processing and Classification Using Computational Learning and Intelligence* (A. A. Torres-García, C. A. Reyes-García, L. Villaseñor-Pineda, and O. Mendoza-Montoya, eds.), pp. 59–91, Academic Press, 2022.
39. Y. Zhang, "Modelling the lexical complexity of homogenous texts: A time series approach," *Quality amp; Quantity*, 2022.
40. C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2019.
41. R. Mitkov, *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2021.
42. S.-W. Kim and J.-M. Gil, "Research paper classification systems based on tf-idf and lda schemes," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, 2019.
43. Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336–1353, 2013.
44. P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, p. 111–126, 1994.
45. D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788–791, 1999.
46. S. T. Dumais *et al.*, "Latent semantic analysis," *Annu. Rev. Inf. Sci. Technol.*, vol. 38, no. 1, pp. 188–230, 2004.
47. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.
48. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
49. M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.
50. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.
51. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
52. D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 262–272, 2011.
53. J. Ge, S. Lin, and Y. Fang, "A text classification algorithm based on topic model and convolutional neural network," *Journal of Physics: Conference Series*, vol. 1748, no. 3, p. 032036, 2021.
54. J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, p. 37–46, 1960.
55. R. Adhitama, R. Kusumaningrum, and R. Gernowo, "Topic labeling towards news document collection based on latent dirichlet allocation and ontology," in *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, pp. 247–252, 2017.
56. S. M. Vieira, U. Kaymak, and J. M. Sousa, "Cohen's kappa coefficient as a performance measure for feature selection," *International Conference on Fuzzy Systems*, 2010.
57. C. F. P. Bureau, "Credit card complaints." `https://data.world/dataquest/bank-and-credit-card-complaints`, 2018.
58. M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, p. 276–282, 2012.
59. T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.
60. T. Hofmann, "Probabilistic latent semantic analysis," *arXiv preprint arXiv:1301.6705*, 2013.
61. C. B. Asmussen and C. Møller, "Smart literature review: a practical topic modelling approach to exploratory literature review," *Journal of Big Data*, vol. 6, no. 1, pp. 1–18, 2019.
62. S. Chen, C. Vidden, N. Nelson, and M. Vriens, "Topic modelling for open-ended survey responses," *Applied Marketing Analytics*, vol. 4, no. 1, pp. 53–62, 2018.