Science World Journal Vol. 19(No 4) 2024 https://dx.doi.org/10.4314/swj.v19i4.8
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

RELIABILITY OF SUPERVISED TOPIC MODELS OVER UNSUPERVISED TOPIC MODELS FOR THE PREDICTION TASK

*[1]Maku T.O., [2]Adenomon M.O., [2]Adehi M.U. & [2]Chaku S.E.

[1]Department of Statistics, Federal University, Otuoke, Nigeria
[2]Department of Statistics, Nasarawa State University, Keffi, Nigeria

*Corresponding Author Email Address: tolutayomaku@gmail.com

ABSTRACT
The study investigated the depth of machine learning's capacity to perform prediction tasks. The study used textual data, specifically the daily actions of cryptocurrency (Bitcoin) dealers, which were found in news articles. The data was employed merely because it produced crowd knowledge of trade from News articles that affected the market price trend. For the goal of making predictions, 4073 pre-processed, scraped news articles from CNBC's market section website were analysed using the Latent Dirichlet Allocation (LDA) model and its variation, the Supervised Latent Dirichlet Allocation Model (sLDA). The document-term matrix and "k" with different values ranging from 3 to 200 were used to train and test the models. The study used four metrics for evaluation because of our multinomial classification method: mean absolute percentage error (MAPE), mean absolute error (MAE), root mean square error (RMSE), and R^2. The outcome demonstrated that for label prediction for unlabeled new documents, the sLDA model performed better than the LDA model plus (classification or regression model). The response variable, which was tagged "users' or traders' interest," was the daily closing price of each corresponding document.

Keywords: Topic, Supervised Topic models, Unsupervised Topic models, Consumer News and Business Channel 's market section website

INTRODUCTION
A prediction task for any purpose aims to set a guild or warning against future occurrences, especially in financial areas. For the sake of this research, the study will use new novel of Topic models which is a machine-learning process to predict the trend of cryptocurrency price trend.
The two primary methods that stand out in artificial intelligence and machine learning are supervised learning and unsupervised learning. Researchers need to be able to differentiate between the two and choose the optimal approach when dealing with a certain scenario because each strategy has distinct characteristics and applications. While unsupervised learning searches for patterns and structures in data without prior knowledge of the intended output, supervised learning employs labelled data to train models for classification or prediction (Tishan *et al.,* 2023).
In this study, the distinction between supervised and unsupervised topic models will be thoroughly revealed, allowing aspiring machine learning researchers to take advantage of their benefits, and overcome the challenges posed by various real-world situations. Topic models are probabilistic generative models used in machine learning and natural language processing (Liu *et al.,* 2016).
"Topics" refers to the vague, unclear relationships that exist between vocabulary words and their usage in writing. A document is thought of as a collection of topics. Topic models identify the collection's latent themes and annotate the articles by them. Every word is thought to originate from one of those subjects. Finally, a distribution of document coverage of topics is produced, offering a fresh approach to data analysis of the subjects' points of view. Unsupervised topic models are used to identify hidden topics in textual data and to illustrate the connections between various topics and the papers or articles that revealed them. (Blei and Jordan. 2003).
Latent Dirichlet allocation (LDA), one of the unsupervised topic models, was primarily used for finding hidden topics in arrays of unlabeled documents. Its variant, the correlated topic model (CTM), was also used for finding hidden topics and topic correlations by utilizing the posterior covariance matrix of the topic. In the past, unsupervised LDA produced a tool for creating classification features. Insofar as they reduce the data dimension, LDA was supposed to be useful for classification (Blei). Fitting an unsupervised topic model may not be the best option when considering a prediction job. This led to our investigation into which model between the LDA and sLDA performs better on the prediction task.
Supervised topic models employ their built-in regression and classification tools to identify hidden topics in labeled documents and then predict labels for newly unlabeled documents. Additionally, compared to the LDA, the sLDA learns more cohesive subjects. On the other hand, new unlabeled documents can be labeled using the unsupervised topic model. Regression, classification, and support vector machine models can be used in conjunction with the unsupervised topic models to accomplish this. The study compared and chose the best model based on predictability strength between the supervised topic model and the unsupervised topic model that was jointly trained using a multinomial logistic regression. Support vector machine (SVM) models and other classification models can also be used to jointly train the unsupervised topic model.
The study "A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science" was conducted by Alloghani *et al.* (2020). They looked at scholarly publications published between 2015 and 2018 that discussed or applied supervised and unsupervised machine-learning techniques in several problem-solving paradigms. Using the PRISMA components, the review process identified 84 scholarly articles that had been published in different journals. Despite their meta data indicating that they were published in 2015, six of the 84 articles were published before that year. It was found that the six articles were included in the final papers due to indexing errors. However, it appeared from the reviewed papers that the decision tree, support vector machine, and Naïve Bayes algorithms were the most often used, discussed, and quoted by supervised learners.

However, unsupervised learning methods like k-means, principal component analysis, and hierarchical clustering also gained popularity. With the current developments in data science and machine learning, the investigation also found other popular algorithms, including ensembles and reinforcers, which may be the focus of more thorough research in the future.

"Supervised and Unsupervised Machine Learning Approaches; A Survey," Varma and Parasad (2023) concentrated on the two primary types of machine learning tasks: supervised and unsupervised approaches. In supervised learning, a lot of data (labeled datasets) was used to train the model, and the outcome was predefined. Their main goal was to predict the outcome. There were issues with categorization and regression. Additionally, they employed autonomous, unsupervised learning, which does not correlate input to output. The main objective of their study was to give readers a comprehensive understanding of pseudocodes for both supervised and unsupervised machine learning methods.

Sun *et al.* (2022), studied A comprehensive comparison of supervised and unsupervised methods for cell type identification in single-cell RNA-seq. Eight supervised and ten unstructured cell-type identification methods were evaluated in this study using 14 publicly accessible scRNA-seq datasets from different tissues, sequencing techniques, and species. Numerous factors, such as the total number of cells, the number of distinct cell types, batch effects, reference bias, imbalance in the cell population, unknown/novel cell type, sequencing depth, and computer efficiency and scalability, were analysed. Instead of only comparing techniques, they focused on how variables affected the wide category of supervised and unsupervised procedures. They found that the supervised approaches outperformed the uncontrolled ones in most circumstances, except for the identification of unknown cell types.

A study on "Supervised topic models for multi-label classification" was conducted by Li et al. (2015). Numerous recent studies have demonstrated that generative modeling techniques, or topic models, performed admirably on multi-label classification, especially when applied to skewed data sets. This work built two supervised topic models for multi-label classification tasks. Two models, Frequency-LDA (FLDA) and Dependency-Frequency-LDA (DFLDA), expand Latent Dirichlet Allocation (LDA) based on two observations: label frequencies and label dependencies. They trained the models with the Gibbs sampler technique. Their two models outperformed the most sophisticated techniques, according to the results of the trials conducted on well-known collections.

Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification was conducted by Sathya *et al.* (2013). In relation to higher education, they conducted a comparative study of supervised and unsupervised learning models along with evaluation of how effectively they categorized patterns. They found that the unsupervised learning model' Korhonen Self Organizing Map, offers an efficient solution and classification, whereas the supervised learning model's error back-propagation learning method is very effective for many non-linear real-time applications.

"Supervised Machine-Learning Techniques: A Comparison" was conducted by Mohamed *et al.* (2022). They compared a few of the tools available for each supervised machine-learning approach and listed some of them in this paper. They outlined the possible application domains and gave a general overview of machine learning for comparative purposes.

Lehr *et al.* (2021) conducted research on "A comparison of supervised and unsupervised learning for optical inspection applications in quality control." For instance, they believed that quality monitoring of newly made products or the return of old and used components is a crucial component of a successful quality management system in enterprises. Their study assessed the effort required to get training data and compared it with the detection accuracy of the different approaches to ascertain the relative benefits of using unsupervised learning techniques. Printer cartridges, both new and old, were used for this. The image data came from 18 different models of printer cartridges. After that, they were fully labeled (annotated). A clever separation of training, validation, and test data allowed for the training of supervised and unsupervised methods and a comprehensive evaluation of the effort for data collection, annotation, and accuracy of fault detection.

Working on "A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised," Yang *et al.* (2019) contrasted various supervised and unsupervised methods for detecting credit card fraud. Their study looked at six supervised classification models: Logistic Regression (LR), K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGB). They also studied four unsupervised anomaly detection models: One-Class SVM (OCSVM), Auto-Encoder (AE), Restricted Boltzmann Machine (RBM), and Generative Adversarial Networks (GAN). using a dataset of public credit card transactions from the Kaggle website, which comprised 284,807 total transactions, 492 of which were fraudulent. Each of these models was trained by them. The transaction labels were only utilized by supervised learning models. To evaluate each model's performance in terms of the Area under the Receiver Operating Curves (AUROC), five-fold cross-validation was employed. Among supervised approaches, XGB and RF produced the best results, with corresponding AUROC values of 0.989 and 0.988. However, with an AUROC of 0.961, RBM fared better than unsupervised methods, and GAN came in second with an AUROC of 0.954. The experimental results showed that supervised models in their study performed marginally better than unsupervised ones. However, unsupervised algorithms continue to be effective for identifying credit card fraud transactions because of the lack of proper annotation and the issue of data imbalance in real-world applications.

Maetschke *et al.* (2013) examined "Unsupervised, semi-supervised, and supervised inference of gene regulatory networks." presentations on bioinformatics. Although many methods have been developed to achieve this goal. They acknowledged that identifying the gene regulatory network from expression data was a challenging task. However, there was no comprehensive evaluation that covers supervised, semi-supervised, and unsupervised methods and provides suggestions for their practical application. They reviewed inference methods in detail and used both simulated and real expression data. The results demonstrated that the Z-SCORE method on knockout data demonstrated significantly higher prediction accuracy than unsupervised alternatives, which had poor prediction accuracy. Even in a semi-supervised setting with small amounts of only positive data, the supervised approach achieved the highest accuracy and outperformed the unsupervised strategies in every other situation.

The study train the LDA model + multinomial regression model of different "k" values and the sLDA model of different "k" values using the same labelled pre-processed textual training data set. It also

https://dx.doi.org/10.4314/swj.v19i4.8
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

test the models using the same new test data, which is unlabeled, to perform our predictability strength comparison between the supervised topic model and the unsupervised topic models. After evaluating the metrics of the tested and trained models, we would select the best prediction model based on its greatest performance values.

MATERIALS AND METHODS

News articles about cryptocurrency-related activities that appeared in foreign media between 2016 and 2022 made up most of the research population. The Consumer News and Business Channel (CNBC) carries these articles. This study only included a sample of Bitcoin because it is so well-known compared to other cryptocurrencies.

The study's secondary data source is the corpus of news articles. All 6,000+ news pieces and articles that were pulled from the internet between 2016 and 2022 were authored in English. Using relevant meta-data from the previously described media source, it was feasible to rapidly scrape the text data using a custom Python script called "beautiful Soap" that was visualized using Jupyter Notebook. The following meta-data, which was saved in comma-separated (CSV) format, was present in the pages:

(i). Article Summary (ii). Article section (iii). Article link (iv). Article date (v). Article summary (vi). Article Body (vii)Opening Price(viii)Closing Price.

The query "Bitcoin daily reports" was used to highlight the articles that were needed and helpful. According to Blei et al. (2017), top topic models assume a bag-of-words document representation. Each text is displayed here as a bag of its terms, with no consideration for word order or grammar. Numerous methods in the domains of natural language processing and information retrieval make use of this reduced paradigm. The NLP step was often broken down into four stages: (1) loading the news article as input data; (2) pre-processing the data; (3) turning texts into bag-of-words vectors; and (4) training the sLDA and LDA models. The news items (now called documents) were transformed into a format suitable for the modeling framework rather than being fed into the model as free text or raw data. Normalization, tokenization, stemming/lemmatization, and stop-word removal are common pre-processing techniques for text data. Following the collection and compilation of the articles, the text was pre-processed in Python using the SpaCy, Gensim, and Pandas modules. Before using NLP on the text, it must be preprocessed. The texts of the articles were then standardized by switching to lowercase. Then, punctuation and other non-ASCII characters were eliminated, along with foreign characters and word elongations. Then, frequently used non-informative stop-words like "the," "is," "I," and "did" were eliminated using stop-words from the Python genism module. Token words were then lemmatized using Python's genism module. Lemmatization is a type of text normalization that involves classifying inflected words into their base or dictionary root terms, or lemma. The terms "trouble," "troubling," and "troubled," for example, can be lemmatized to produce the lemma "trouble." The traditional stemming of tokens was used in response to Schofield et al. (2017)'s assertion that topic coherence is rarely enhanced between the pre-stemming and post-steaming Topic models. In the end, whitespace was removed to decrease the content's overall size. Each document has fewer than fifty (50) words deleted. Furthermore, words that appeared in less than 70% of the corpus were removed.

The LDA Model

The Latent Dirichlet Allocation (LDA), a comprehensive generative probabilistic model of a corpus, is based on the notion that a document comprises several themes. Conversely, a topic is just the arrangement of ideas within a specific vocabulary (words). The LDA states that 'K' themes are associated with a collection of documents and that these topics are displayed in varying proportions within each document (Blei 2003). Furthermore, the LDA assumes the term exchangeability, or a "bag of words," implying that the order of a term is unimportant (Aldous, 2009). Furthermore, because LDA is predicated on the idea that documents are interchangeable, it ignores the order in which they appear within a corpus.

The LDA generative process is as below:

1. for each topic k ε (1,..,k), draw a multinomial distribution over words $\beta_k \sim Dir(\eta)$
2. for each document, d ε (1,…, D), draw a multinomial distribution vector of topic proportions $(\theta_d) \sim Dir(\vec{\alpha})$
3. For each word position;
   i. draw a topic assignment $Z_{d,} \sim (\theta_d)$,
   ii. Draw a word $W_{d,} \sim (\beta z_{d,n})$,

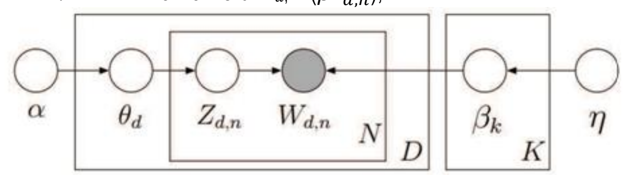


Figure 1 Graphical model representation of LDA

the observed words for document *d are* w_d, where $w_{d,n}$ is the n_{th} word in document *d*, which is an element from the fixed vocabulary.

$\beta_{1:K}$	:	Topics distribution over words
$\theta_{1:D}$	:	Per-document topic proportions
$z_{1:D,1:N}$	:	Per-word topic assignments

The Dirichlet formula for the hidden and observable variables is a key component of the LDA:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right) \quad (1)$$

The Dirichlet distribution has density;

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \quad (2)$$

Where the parameter α is a *k*-vector with components α_i>0 and θ_i> 0; $\sum_{i=1}^{k} \theta_i = 1$

The expected value of θ is given as; $E(\theta_i) = \frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i}$

The following is the posterior distribution of the hidden variable and the observed word:

$$p(\theta_{1:D}, z_{1:D}, \beta_{1:K} | w_{1:D}, \alpha, \eta) = \frac{p(\theta_{1:D}, z_{1:D}, \beta_{1:K} | w_{1:D}, \alpha, \eta)}{\int_{\beta_{1:K}} \int_{\theta_{1:D}} p(\theta_{1:D}, z_{1:D}, \beta_{1:K} | w_{1:D}, \alpha, \eta)} d\theta d\beta$$

$$= \frac{p(\theta_{1:D}) \prod_n^N p(Z_n|\theta) p(W_n|Z_n, \beta_{1:K})}{\int_\beta \int_\theta p(\theta_{1:D}) \sum_Z \prod_n^N p(Z_n|\theta) p(W_n|Z_n, \beta_{1:K}) d\theta d\beta} \quad (3)$$

Due to the intractability of the denominator in equation 2.3, a metropolis-hasting process which uses the variational process was used to control the model.

The sLDA Model
The supervised latent Dirichlet allocation model (sLDA) performs better when implementing such a plan for response-document pairs.
In topic models, which are distributions over collections of documents, each document is represented by a set of discrete random variables, $W_{1:n}$, which are its words. In topic models, which are a collection of unknown distributions over the vocabulary, the words in a document are viewed as emerging from a set of latent themes. All documents in a corpus share the same ‘K’ topics, but each document employs a different mix of subjects with topic proportions that are unique to it. In contrast to standard document mixing models that associate each document with a single, undefined subject. They are referred to as mixed-membership models by Erosheva et al. (2014). When deciding on labels for newly unlabeled documents, each document has a matching response as a covariate that is jointly modelled for prediction. Supervised Latent Dirichlet Allocation (sLDA) extends the LDA model to a supervised learning environment by allowing a response to be linked with each document and simultaneously modelling the response variable and the corpus of documents. Blei et al. (2017) claim that this allows it to predict future unlabeled articles and even determine which latent topics are most predictive of the response variables in the training set. let y represent a response variable from a generalized linear model with parameters η and δ. Should we take into account the subsequent fixed;$\beta_{1:K}$: the k topics with each β_k a vector of term probability, η and δ and the Dirichlet hyperparameter for the per-document topic proportion θ.
For every document and response variable, the generative process assumed by the sLDA is as follows:

1. Draw topic proportion, $\theta| \alpha \sim Di(\alpha\ )$
2. for each word,
   (a) Draw a topic assignment Z_n | θ ~*Mult*(θ)
   (b) Draw word w_n | Z_n ,$\beta_{1:K}$ ~*Mult*(β_{z_n})

3. Draw a response variable y| $z_{1:N}, \eta, \delta \sim$GLM ($\overline{z}, \eta, \delta$), where $\overline{z} = \frac{1}{N}\sum_{n=1}^{N} z_n$.

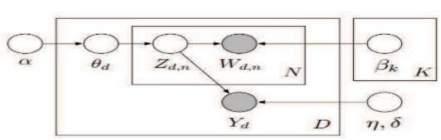


Figure 2: Graphical Model representation of Supervised Latent Dirichlet Allocation (sLDA)

The response variable distribution is a generalized linear model (McCallum *et al.,* 2005),

$$p(y|z_{1:N}, \eta, \delta) = h(y, \delta)\exp\{\frac{(\eta^T\overline{z})y - A(\eta^T\overline{z})}{\delta}\} \quad (4)$$

Given a natural parameter response variable, the random component of equation (2.4) assumes an exponential dispersion family distribution $\eta^T\overline{z}$ and a dispersion parameter δ. Canonical link functions are the sole ones used in the sLDA paradigm. The function known as the canonical link is the one that changes the mean μ= E (y_i) concerning the exponential family of distribution's natural exponential (location) parameter, such as normal, binomial, negative binomial, Multinomial, Poisson, and gamma distribution which are mostly used in GLM. Notably, $h(y, \delta)$ is the base measure, y is the sufficient statistics, and $A((\eta^T\overline{z})$ is the normalization log. Due to the GLM framework's versatility, sLDA may be used to represent a variety of response variable types, whose distribution can be expressed in equation (2.4)'s exponential dispersion form. For example, for a normally distributed random variable y, $h(y, \delta) = \exp\frac{1}{\sqrt{2\pi\delta}}\{-\frac{y^2}{2}\}$ and $A(\eta^T\overline{z})$=$\frac{(\eta^T\overline{z})^2}{2}$ here, mean μ is $\eta^T\overline{z}$ and variance is δ.
Blei & McAuliffe (2017) serve as the foundation for our calculations. We use variation inference to approximate the posterior density by computing the posterior distribution of the document-level latent variable θ, the topic proportions, and the topic assignment $Z_{1:N}$ given the words $W_{1:N}$ and the corpus-wide model parameters.

$$p(\theta, z_{1:N}|\ w_{1:N}, y, \alpha, \beta_{1:K}, \eta, \delta) = \frac{p(\theta|\alpha)(\prod_{n=1}^{N} p(z_n|\ \theta)p(w_n|z_n, \beta_{1:K}))p(y|\ z_{1:N}, \eta, \delta)}{\int p(\theta|\alpha)\sum_{z_{1:N}}(\prod_{n=1}^{N} p(z_n|\ \theta)p(w_n|z_n, \beta_{1:K}))p(y|\ z_{1:N}, \eta, \delta)\, d\theta} \quad (5)$$

For a Gaussian random variable with an identity link function, the expected mapping from the natural parameter to the mean parameter is;

$$\mathbb{E}[Y|w_{1:N}, \alpha, \beta_{1:K}, \eta, \delta) \approx \eta^T\mathbb{E}[\bar{Z}] \quad (6)$$

Fitting the LDA and sLDA models
The models were fitted using the variational expectation-maximization (VEM) technique. The "Document-Term matrix, "K" (the number of topics), and "control" (the Latent Dirichlet Allocation (LDA)-VEM) were among the parameters used to fit the LDA. The parameters were used to determine the maximum number of iterations for the conjugate gradient method, which alternates between the E-step and M-step to maximize the corpus's probability, as well as the convergence tolerance for the variance and E-M algorithms, respectively. In the M-step, the procedure establishes the upper limit concerning the model parameters (the topics and the multivariate normal parameters), and in the E-step, it establishes the upper limit concerning the latent variables (the topic proportions and the topic assignments Z). The multinomial logistic regression model for label or response prediction was trained using the LDA posterior covariance matrix. One instance of this procedure was performed for every value of "k." 70% of training data and 30% of test data were used in a data split on the 4,073 pre-processed publications and articles. The training data set was used to train LDA and sLDA models, and the test data set was used to evaluate the prediction of our response variable. The response variable, which is essentially of relevance to traders or consumers, is the time-series closing price of Bitcoin that corresponds to the documents. Additionally, numerically categorize response variable as "low = 1," "fairly-low = 2," “high = 3” and “fairly-high = 4” with thresholds of "less than or equal to 10000," "less than or equal to 20000," "less than or equal to 40000," and "less than or equal to 60000," respectively. Variational expectation maximization (VEM) is the technique used to fit the sLDA model. Several parameters were used to adjust the algorithm's rate of convergence. These parameters included the "Document-Term matrix," "K" (the number of topics), "vocab" (vocabulary words associated with word indices used in documents), "e.iteration," "m.iteration," "alpha," "eta," "var" (variance of the response

 https://dx.doi.org/10.4314/swj.v19i4.8

variable), "annotations" (response variable), and others. Regression and a classifier are included in the sLDA model to predict labels or responses.

RESULTS AND DISCUSSION

Using 70% and 30% thresholds, respectively, a corpus (data) of 4073 documents was divided into training and testing data. As previously stated in Section (2.0), the split corpus was pre-processed before the split. Because the words cannot be utilized directly as input for the models, they were then transformed into a machine-learning language and then transformed into a document-term matrix (DTM). The prediction job was then completed by training and testing the sLDA and the LDA + regression model.

LDA and sLDA Model Metric Evaluation.

Using model assessment measures, the best model for our prediction task out of the LDA and sLDA models was found. The classification method we employed multinomial classification techniques which led to the selection of our model evaluation. Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RSME), and R^2 are the evaluation measures.

Table 1. Metric evaluation of the sLDA model with varying values **of "k"**

K_VALUES_slda	MAPE_slda	MAE_slda	RMSE_slda	R_SQRED_slda	ADJ_R_SQRED_slda
3	66.99009	1.061556	1.996749	0.009554	0.009311
10	49.8341	0.802849	0.98224	0.336044	0.335881
20	46.44281	0.761462	0.940588	0.391162	0.391012
30	49.07126	0.789546	0.977847	0.34197	0.341809
50	42.03424	0.710612	0.905843	0.435311	0.435172
100	42.60092	0.738343	0.942975	0.388067	0.387917
200	42.67802	0.737328	0.933316	0.400539	0.400392

The numerical values of the sLDA assessment across the metric lines are shown in Table 3.1 above. A close examination of the table reveals that the sLDA with k = 50 performs better than other "k" values. Its MAPE, MAE, RSME, and R^2 values are the lowest, lowest, and highest, respectively. Among other models, this makes the sLDA with k = 50 the best one.

Table 2. Metric evaluation of the LDA model with varying values of "k"

K_VALUES_lda	MAPE_lda	MAE_lda	RMSE_lda	R_SQRED_lda	ADJ_R_SQRED_lda
3	84.34563	1.889674	2.289739	-0.21992	-0.22025
10	88.30725	1.902674	2.7906	-0.22096	-0.22128
20	85.34525	1.899967	2.31898	-0.22031	-0.22032
30	82.30725	1.676737	2.189678	-0.21985	-0.21322
50	87.45665	1.822674	2.418988	-0.20097	-0.20042
100	85.34525	1.899967	2.31898	-0.22031	-0.22032
200	86.67848	1.856774	2.191617	-0.32187	-0.33298

Table 3: Evaluation of metrics to select the best model between the sLDA and LDA models

Metric	lda30	slda50
MAPE	82.30725	42.03424
MAE	1.676737	0.710612
RMSE	2.189678	0.905843
RSQRD	-0.21985	0.435311
ADJRSQRED	-0.21322	0.435172

The numerical values of the LDA assessment across the metric lines are shown in Table 3.2 above. A close examination of the table reveals that the LDA with k = 30 performs better than other 'k' values. Its MAPE, MAE, RSME, and R^2 values are the lowest, lowest, and highest, respectively. As a result, among various models, the LDA with k = 30 is adjudged the best model.

The numerical values of the optimal models from Tables 3.1 and 3.2 are shown across the metric lines in Table 3.3 above. A further examination of the table reveals that the sLDA with k = 50 performs better than the LDA with k = 30. Its MAPE, MAE, RSME, and R^2 values are the lowest, lowest, and highest, respectively. As a result, for fresh unlabelled texts, the sLDA with k = 50 is the best model fit for predicting the response variable (user interest). The implication of the results appears below in section 2 showing better prediction pattern of the sLDA than the LDA.

Science World Journal Vol. 19(No 4) 2024 https://dx.doi.org/10.4314/swj.v19i4.8
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Outcomes of Ideal Models (LDA30 & sLDA50) Applications

Table 4: Table showing the first Seven (7) Topics with Seven (7) words/terms from the LDA30 model output

Topic 1	marking	fcau	signal	poshmark	prosecute	facilitating
Topic 2	glut	drone	mirror	macron	privatelyheld	backer
Topic 3	separate	dominance	hsiao	moonpay	tame	amassed
Topic 4	worm	pbocs	crunching	diatribe	proactively	saudi
Topic 5	inherit	tumblr	riskreward	softening	swanky	chipmakers
Topic 6	dismantle	conagra	unbelievably	spanning	shot	introduces
Topic 7	peer	impersonator	cattle	hayek	dislike	brody

Tale 4 lists the likely 30 Topics from LDA30 model along with their corresponding terminology. The appendix lists the additional 27 themes that the LDA30 model has found.

Table 5: Table showing the first Seven (7) Topics with Seven (7) words/terms from the sLDA50 model output

Topic 1	around	managed	goldbacked	collapse	magic	harvey
Topic 2	china	chinese	yuan	beijing	country	pboc
Topic 3	sure	entering	mission	anticipated	raise	boost
Topic 4	goldman	street	wall	bank	morgan	sachs
Topic 5	facebook	libra	facebooks	social	project	association
Topic 6	cramer	stock	host	money	twitter	question
Topic 7	trump	president	bill	house	senate	congress

Table 5 shows the probable 50 topics from the sLDA50 model with their terms. The appendix shows 47 other topics uncovered by the sLDA50 model.

Table 6 displays the predicted labels for the first 50 documents in the testing corpus out of the first 200 documents by both LDA30 and sLDA50.

1	label/Price	documents	predictionlda30	predictionslda50
2	2	2	1.719972874	1.616772874
3	1	4	1.109713345	1.156771335
4	1	13	1.198891815	0.839795532
5	2	14	1.54963007	1.51910483
6	2	20	2.381261952	1.917880422
7	2	24	1.940833783	1.582023391
8	1	25	2.586547898	1.06982221
9	2	28	2.871271564	1.681407118
10	3	30	3.31314923	0.899112571
11	3	33	1.159389263	1.736931326
12	1	46	0.822885111	1.656455462
13	2	48	1.917880422	1.176013205
14	2	50	1.582023391	1.424496127
15	2	51	1.06982221	2.031319639
16	2	60	1.51910483	1.258942692
17	3	64	1.881108469	2.398612232
18	2	69	1.767672375	1.514849818
19	2	71	2.856183344	2.042250995
20	1	76	1.964051884	0.341652222
21	4	78	2.051931643	3.256558619
22	2	79	1.747156775	1.442965254
23	2	81	1.434950029	1.347948236
24	2	82	1.244809315	1.93211296
25	1	83	0.78992493	0.78992493
26	1	85	1.037789395	1.037789395
27	3	87	1.253041388	2.732591753
28	3	89	1.399839	2.977118025
29	4	97	2.71578609	3.31314923
30	1	104	2.954822559	1.159389263
31	1	112	2.326610371	0.822885111

21	4	78	2.051931643	3.256558619
22	2	79	1.747156775	1.442965254
23	2	81	1.434950029	1.347948236
24	2	82	1.244809315	1.93211296
25	1	83	0.78992493	0.78992493
26	1	85	1.037789395	1.037789395
27	3	87	1.253041388	2.732591753
28	3	89	1.399839	2.977118025
29	4	97	2.71578609	3.31314923
30	1	104	2.954822559	1.159389263
31	1	112	2.326610371	0.822885111
32	2	119	1.837586296	1.837586296
33	2	123	1.332439291	1.332439291
34	1	124	1.198891815	1.198891815
35	2	127	1.54963007	1.54963007
36	3	128	2.381261952	2.906004136
37	2	133	2.636498481	1.792093301
38	3	134	1.245308277	0.91699397
39	3	137	1.82994596	3.258988933
40	1	138	1.766091925	2.573662765
41	2	147	2.968571613	1.255542481
42	2	151	1.115207773	1.615650124
43	2	153	3.474868786	1.253041388
44	2	166	0.891766724	1.399839
45	3	167	1.624408573	2.71578609
46	3	168	1.283915216	2.954822559
47	3	170	1.558374697	2.326610371
48	2	172	2.873138033	1.384074022
49	2	177	1.478222113	2.042967445
50	2	178	1.920570633	2.513551254
51	3	184	2.723744794	1.538763414

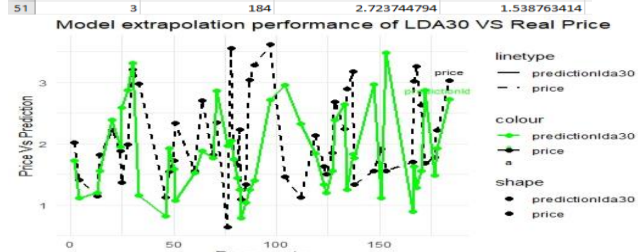


Figure 3: graphical representation of LDA30 label prediction in Table 6

Science World Journal Vol. 19(No 4) 2024 *https://dx.doi.org/10.4314/swj.v19i4.8*
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

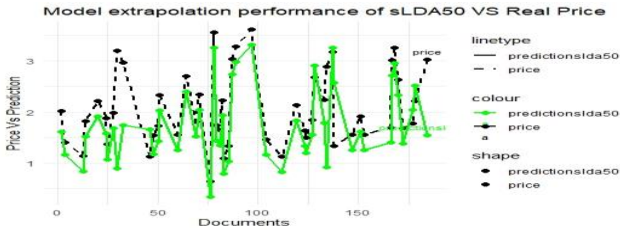


Figure 4: graphical representation of sLDA50 label prediction in Table 6

Figure 3 and 4 clearly show that the trend pattern looks more similar for sLDA50 than LDA30. This further strengthens the fact that sLDA models is reliable and do better in prediction tasks.

Conclusion

To compare and identify the best model between the LDA and the sLDA for predicting the response variable for unlabelled new documents, the study has used textual data (labelled documents) on Bitcoin cryptocurrency operations. The correlated topic model is also unsupervised, but it goes one step further than the LDA to demonstrate correlations between the topics, words, and documents. It is thought that the LDA model performs better in identifying Latent Topics from the documents. Compared to the unsupervised version, the sLDA is more effective at predicting response variables (labels) for unlabelled documents and reveals more cohesive hidden subjects.

Conversely, in this paper, the LDA can be concurrently trained with other regression or classification models for prediction purposes. The findings demonstrated that the sLDA performs better than the LDA model + (classification or regression model) when used for prediction. "K" number was chosen at random from 3 to 200 to compare which model performs better in prediction between the LDA and the sLDA. The models were trained using this technique for every value. When only applying a topic model to a given corpus, the coherence graph, perplexity graph, or both used to determine the optimal value of "k."

The best results obtained with a coherence graph of "k" with the largest coherence value; however, care must be taken if the value selected under or over-fits the model. It takes less time to converge when obtaining the correct value of "k" from a corpus through coherence, perplexity, or both when comparing models, although the two approaches have different goals and tasks. Not selecting the appropriate "k" value implies that the subjects will be less coherent for a high-quality result. To improve the production of cohesive topics, there was extreme caution when handling the hyperparameter tuning during the research.

Until the relative change in the probability was less than 10-6, variational inference was used, and until the relative change in the likelihood bound was less than 10-4, variational Expectation-Maximization was used. Similarly, the F-score and accuracy have always been employed by various literatures to assess how well their topic model's function. It was found that when utilizing the sLDA model for prediction, this metric evaluation performs better when employing a bi-classification approach rather than a multi-classification method. As a result, this led to the decision to use R-squared, RMSE, and MAE measures.

REFERENCES

Aldous, David. (2009). More Uses of Exchangeability: Representations of Complex Random Structures.

Alloghani, Mohamed & Al-JumeilyObe, Dhiya& Mustafina, Jamila & Hussain, Abir & Aljaaf, Ahmed. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. 10.1007/978-3-030-22475-2_1.

Blei D. and Jordan M. Modeling annotated data. In Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 127–134. ACM Press, 2003.

Blei D., Kucukelbir A. and McAuliffe J.D. (2017) Variational Inference: A Review for Statisticians. Journal of the American Statistical Association, 112:518, 859-877.

Erosheva E., Fienberg S., and Lafferty J. (2014), Mixed-membership models of scientific publications. Proceedings of the National Academy of Science, 97(22):11885–11892

Lehr, Jan & Philipps, J & Hoang, V & Wrangel, D &Krüger, Jörg. (2021). Supervised learning vs. unsupervised learning: A comparison for optical inspection applications in quality control. IOP Conference Series: Materials Science and Engineering. 1140. 012049. 10.1088/1757-899X/1140/1/012049.

Li, Ximing& Ouyang, Jihong& Zhou, Xiaotang. (2015). Supervised topic models for multi-label classification. Neurocomputing. 149. 811-819. 10.1016/j.neucom.2014.07.053.

Liu, L., Tang, L., Dong, W. *et al.* An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* **5**, 1608 (2016). https://doi.org/10.1186/s40064-016-3252-8

Maetschke, Stefan &Madhamshettiwar, Piyush & Davis, Melissa & Ragan, Mark. (2013). Supervised, semi-supervised, and unsupervised inference of gene regulatory networks. Briefings in bioinformatics. 15. 10.1093/bib/bbt034.

Mohamed, Salem & Ashraf, Raghda& Ghanem, Amira &Sakr, Mohamed & Mohamed, Roaa. (2022). Supervised Machine Learning Techniques: A Comparison.

Schofield A., Magnusson M., Thompson L., Mimno D. (2017), Understanding Text
Pre-Processing for Latent Dirichlet Allocation. Retrieved from http://www.cs.cornell.edu/~xanda/winlp2017.pdf

Sun, X., Lin, X., Li, Z., & Wu, H. (2022). A comprehensive comparison of supervised and unsupervised methods for cell type identification in single-cell RNA-seq. *Briefings in Bioinformatics*, *23*(2), bbab567. https://doi.org/10.1093/bib/bbab567

Tishan, Yasiru. (2023). Understanding the Difference between Supervised and Unsupervised Learning Techniques. 10.13140/RG.2.2.36176.48641.

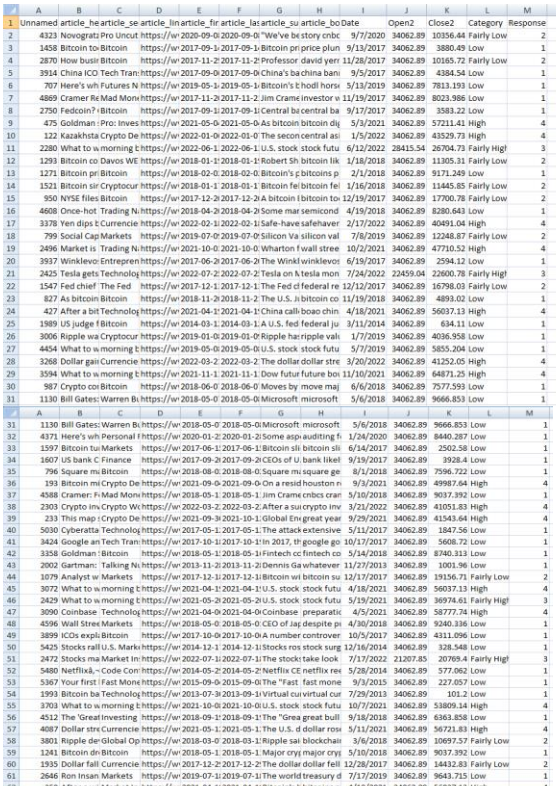
Varma, C. & Prasad, Puja. (2023). Supervised and Unsupervised Machine Learning Approaches—A Survey. 10.1007/978-981-19-5936-3_7.

Yang, Xulei& Wang, Li &Niu, Xuetong. (2019). A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised.

Science World Journal Vol. 19(No 4) 2024
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University
https://dx.doi.org/10.4314/swj.v19i4.8

APPENDICES

(A)



Tables showing the randomly attached data set from document 1 to document 4073

(B)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19
marking	glut	separate	worm	inherit	dismantle	peer	alibabas	vinci	ramada	netting	broadbase	rumor	insufficier	kovacevic	peer	offsetting	wellrecei	cunningh
fcau	drone	dominanc	pbocs	tumblr	conagra	imperson	pegged	schwabs	salvadora	emmer	williamso	shanghaib	thrive	stemmed	nayib	uncomfor	hiking	subpoena
signal	mirror	hsiao	crunching	riskrewar	unbelieva	cattle	bolton	fantasy	tatro	initially	serf	mpesa	philosoph	swathe	evolving	interpret	hummel	script
poshmark	macron	moonpay	diatribe	softening	spanning	hayek	symbolic	coincided	lithuania	pinto	tutorial	invesco	finger	governme	kickstarte	fanduel	lavish	teenage
prosecute	privatelyh	tame	proactivel	swanky	shot	dislike	pascal	underwer	cost	congestio	trevor	dirt	snoop	reconcilia	adami	intensity	nissan	crashing
facilitatin	backer	amassed	saudi	chipmake	introduce	brody	disdain	story	specialize	apple	nerd	crane	saved	comfortab	identify	minneapc	rhind	sneaker
prone	scoreboar	safeguard	stla	wallstreet	rebuke	migrate	matt	nerdwalle	horowitz	granitesha	lekander	dress	cirrus	unionize	adoption	behold	spying	davi
personalit	russo	accountin	brexit	stifling	rulesbase	peter	usbased	clarifying	mortar	copeland	mark	blockfi	bukowski	chuck	pulp	speak	pain	gupta
wakuda	westport	awarded	cheesecak	tmobile	chunk	headquar	lowerys	energyeff	gift	objective	pfau	fight	creep	zimbabwe	hassan	ethereum	resumptic	eagerly
sochi	carusocab	reclaimin	krug	winkler	difficult	mayor	relaunch	remembe	hydropow	weather	beacon	belfort	hammer	idle	barcelona	similar	piling	adopter

Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30
broadbase	rumor	insufficier	kovacevic	peer	offsetting	wellrecei	cunningha	davidson	brave	consortiu	dogefathe	evaluated	caused	deployme	recomme	pathway	playoff	simulated
williamso	shanghaib	thrive	stemmed	nayib	uncomfor	hiking	subpoena	defensive	bitdefend	awaitey	swapping	sextortior	roundthe	otis	launderin	uncharted	pleased	cayeron
serf	mpesa	philosoph	swathe	evolving	interpret	hummel	script	maduro	charting	spoofing	republica	cocreator	discovery	berkeley	endorsed	thicket	surveillan	improve
tutorial	invesco	finger	governme	kickstarte	fanduel	lavish	teenage	jpmorgan	fraudulen	santa	devaluing	theyll	higherenc	globe	tight	testimony	obsolete	patriarch
trevor	dirt	snoop	reconcilia	adami	intensity	nissan	crashing	posed	reinvent	pandora	congratul	purchase	beta	aficionad	specialist	wefs	granted	kimberley
nerd	crane	saved	comfortab	identify	minneapc	rhind	sneaker	polluting	slowness	divulge	cred	tapscott	electronic	clearly	singleserv	wellrecei	fica	pipeline
lekander	dress	cirrus	unionize	adoption	behold	spying	davi	smelting	sofis	nutshell	taxfree	jimmy	santa	plugin	joes	mess	combine	olive
mark	blockfi	bukowski	chuck	pulp	speak	pain	gupta	kick	pfizerbior	internetb	blowout	rizzutti	sandberg	clash	tamper	motivator	satisfies	schnorr
pfau	fight	creep	zimbabwe	hassan	ethereum	resumptic	eagerly	nakamoto	statemen	wismeijer	instagram	containin	illegally	quorum	pebble	noise	threemon	spoke
beacon	belfort	hammer	idle	barcelona	similar	piling	adopter	neil	inclusive	trendy	hows	impulse	arnott	florist	defies	genetic	showed	pearson

Table showing LDA30 model probable 30 topics

(C)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19
around	china	sure	goldman	facebook	cramer	trump	winklevos	dimon	share	civil	company	elliott	currency	loan	stablecoir	south	blockchai	country
managed	chinese	entering	street	libra	stock	president	tyler	jpmorgan	beijing	newsroon	beijing	prechter	bitcoin	debt	stablecoir	korea	technolog	russia
goldbacke	yuan	mission	wall	facebooks	host	bill	twin	jamie	cnbc	tyson	million	singer	bank	student	withdraw	north	network	russian
collapse	beijing	anticipate	bank	social	money	house	cameron	fraud	company	journalisn	coinbase	cause	beijing	home	celsius	korean	transactio	ukraine
magic	country	raise	morgan	project	twitter	senate	gemini	chase	stock	citizen	billion	half	digital	american	luna	greece	beijing	sanction
harvey	pboc	boost	sachs	associatio	question	congress	brother	davos	premarke	post	year	masucci	payment	according	voyager	greek	ripple	governme
vault	crackdowr	call	stanley	marcus	cramers	governme	space	conferenc	revenue	ryckman	firm	camel	bitcoins	mortgage	patent	nautilusco	ledger	state
larry	local	chairman	client	giant	facebook	republica	travel	gambling	earnings	banning	customer	theorist	central	university	billion	minister	project	salvador
deutsche	governme	jafari	jpmorgan	launch	dive	committe	virgin	casino	trading	system	venture	provided	transactio	school	reserve	ministry	applicatio	united
stake	hong	pull	chase	medium	deep	former	galactic	vega	quarter	reader	startup	zone	virtual	college	tether	missile	company	president

Science World Journal Vol. 19(No 4) 2024 https://dx.doi.org/10.4314/swj.v19i4.8
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Topic 20	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30	Topic 31	Topic 32	Topic 33	Topic 34	Topic 35	Topic 36	Topic 37	Topic 38
food	mining	beijing	case	nfts	riot	florist	long	life	heat	jonas	miller	cuban	square	crypto	buffett	beijing	future	stock
giza	miner	card	court	digital	filing	continue	trader	woman	gilbert	satellite	centrum	oleary	dorsey	cryptocurr	berkshire	people	bitcoin	market
fast	power	income	criminal	sold	orourke	kelley	kelly	people	monarch	origin	sharma	tank	mtgox	asset	warren	going	contract	investor
xbox	energy	account		million	release	finally	short	family	grow	matter	philosoph	shark	jack	ether	billionaire	money	product	year
ghana	bitcoin	credit	federal	sale	soon	message	brian	book	butterfly	china	mayweath	mark	bankruptc	token	hathaway	bitcoin	launch	trading
chicken	electricity	gain	attorney	token	sharehold	crazy	fast	business	guardian	bound	hypnosis	exclusive	karpeles	market	meeting	thats	cboe	growth
fike	mine	year	governme	nonfungib	john	gave	call	work	disaster	continued	trapani	miss	japanese	ethereum	chairman	thing	fund	price
neuner	state	saving	silk	collectible	longfin	keeping	buyer	medium	conclusion	alarm	structure	billionaire	dolev	cryptocurr	annual	year	commissi	equity
obesity	grid	plan	illegal	meme	comment	daneshgar	money	black	dangerous	sort	watch	sherman	tokyobase	wood	munger	really	exchanget	analyst
fried	industry	return	drug	auction	press	fewer	adami	million	recorded	tony	recovery	brady	tokyo	coin	gate	want	exchange	bubble

Topic 39	Topic 40	Topic 41	Topic 42	Topic 43	Topic 44	Topic 45	Topic 46	Topic 47	Topic 48	Topic 49	Topic 50
percent	tesla	overstock	apple	rating	beijing	beijing	bitcoin	beijing	thiel	investmer	chip
index	musk	byrne	google	tipranks	exchange	account	beijing	dollar	acciones	investor	pick
japan	dogecoin	overstock	amazon	small	security	attack	price	rate	para	fund	dalio
closed	twitter	saddingto	search	five	regulator	hacker	cryptocurr	inflation	peter	asset	week
lower	tweet	dealershi	food	operating	cryptocurr	security	digital	market	mercado	investing	squawk
session	elon	tzero	game	rated	coin	data	currency	bank	desde	financial	word
trade	vehicle	medici	whole	addition	investor	company	trading	week	hogan	invest	draft
gain	tweeted	lamborghi	iphone	gerdes	token	informatio	market	reserve	semana	portfolio	answer
close	electric	overstock	list	success	regulation	ransomwa	exchange	economy	bajista	advisor	show
reuters	medium	frigerio	microsoft	influx	regulatory	wallet	year	high	aumento	client	andrew

Table showing sLDA50 model probable 50 topics