Published *online* on the journal's web page : http://innovatics.unsil.ac.id

Innovation in Research of Informatics (INNOVATICS)

| ISSN (Online) 2656-8993 |

Sentiment Analysis of Societal Attitudes Toward the Childfree Lifestyle Using Latent Dirichlet Allocation (LDA) and Support Vector Machines (SVM)

Ratna Andini Husen[1], Agustin[2], Susi Erlinda[3], Junadhi[4], Thinagaran Perumal[5]

[1,2,3,4]*Department of Informatic Engineering, STMIK AMIK Riau, 28294, Pekanbaru, Indonesia*
[5]*Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia*

[1]*husendini151100@gmail.com,* [2]*agustin @sar.ac.id,* [3]*susierlinda@stmik-amik-riau.ac.id,* [4]*ejunadhi@sar.ac.id,* [5]*thinagaran@upm.edu.my*

ARTICLE INFORMATION

Article History:
Received: September 27, 2024
Last Revision: January 01, 2025
Published *Online*: March 30, 2025

KEYWORDS

Analysis Sentiment,
Childfree,
Intent Sentiment,
Latent Dirichlet Allocation,
Support Vector Machine

CORRESPONDENCE

Phone: +6287898574895
E-mail: 2010031802011@sar.ac.id

ABSTRACT

This research investigates societal perspectives on the childfree lifestyle through Intent Sentiment Analysis, combining Latent Dirichlet Allocation (LDA) and Support Vector Machine (SVM) techniques. The childfree lifestyle, a deliberate decision by individuals or couples to remain childless, has spurred extensive public discourse, particularly on platforms like Twitter. This research aims to analyze sentiments and intentions within these discussions to uncover their implications for social dynamics and familial relationships. Using LDA, dominant topics were identified from a dataset of Twitter comments on the childfree topic. LDA uncovered hidden themes by modeling topics as mixtures of words, which were subsequently classified into positive, negative, and neutral sentiments using SVM. Data preprocessing included cleaning, tokenization, and stop word removal, while oversampling with SMOTE addressed class imbalances. The optimal number of topics was determined using coherence scores, with the highest coherence value of 0.400 achieved at one topic. The findings revealed that positive sentiments were classified more effectively than negative and neutral sentiments when using LDA and SVM with SMOTE. The top 10 topics primarily reflected societal commentary on the childfree lifestyle. Challenges included incomplete preprocessing, suboptimal clustering of similar themes, and imbalanced data, which limited the effectiveness of topic modeling and classification. Addressing these issues through improved feature selection, parameter optimization, and data augmentation could enhance performance for underrepresented categories. This research provides valuable insights into public attitudes toward the childfree lifestyle, offering implications for social research and policy development in the context of evolving societal norms.

1. INTRODUCTION

Lifestyle changes are increasingly developing in all aspects. One of them is the lifestyle of not having children or being childfree. Childfree is the decision of a person or couple not to have children [1]. The term childfree is of particular concern, people talk about this topic through uploading comments and opinions on social media including twitter. Twitter is a popular social media platform today. Twitter *is* used to express opinions because Twitter users are known to be critical in responding to a topic of conversation through tweets [2]. This childfree topic is one of the things that greatly influences the way people think so that it becomes one of the supporting factors that influence the high divorce rate, birth rate and marriage. There were 516,344 divorce cases recorded until 2022, which increased by 15.3% from the previous year [3]. The prevalence of this topic in social creates opportunities to analyze individual perspectives and sentiments related to this choice.

Intent Sentiment Analysis in this research is used to find out which these messages are in social network. Conversations related to the topic of childfree can be

classified using topic modeling methods to generate focused data on topics that are frequently discussed by Twitter users. The topic modeling used in this case is Latent Dirichlet Allocation (LDA). LDA uses the Bag of Words method to identify hidden topics in large document collections. Topic modeling with Latent Dirichlet Allocation allows accurate determination of keywords and topics [4]. Applying LDA to comments on childfree-related social media platforms, we can reveal the most dominant and deep topics in the conversation. In addition, by integrating machine learning approaches, so can analyze the sentiment contained in the text and identify the purpose or *intent* behind each statement.

Support Vector Machine (SVM) is a supervised learning algorithm useful for decision analysis. SVMs are designed to consider a variety of common factors and minimize structural risk when determining the optimal hyperplane (decision boundary) to separate data from predefined classes [5]. Research conducted under the title Topic Modeling Analysis of the Use of Twitter social media by State Officials using the Latent Dirichlet Allocation (LDA) method. The results of the analysis model are evaluated using perplexity and coherence score calculations. The model evaluation resulted in a perplexity value of -8.069 and a coherence score of 0.375 for a total of 7 topics. This shows that the model used is good for analyzing and finding topics in tweets [6].

The results of this research are providing information about the most dominant topics in the childfree dataset through the Support Vector Machine (SVM) algorithm and Latent Dirichlet Allocation (LDA) topic modeling as well as confusion matrix and coherence score evaluation models. In addition, this research can provide information about the focus of Twitter users' perspectives on the topic of childfree using the LDA method with public sentiment towards the idea of not having children. The results of this research can facilitate the community and can be a means for readers to find out the public's views on childfree.

2. RELATED WORK

The concept of a childfree lifestyle, defined as the decision by individuals or couples not to have children, has sparked significant discourse on social media platforms, particularly Twitter. This lifestyle choice has been associated with various social dynamics such as divorce and birth rates, with highlighting perceptions of Generation Z in East Java towards this trend [1]. Twitter, a platform known for critical public discourse, serves as a rich data source for sentiment analysis and topic modeling [2]. Sentiment analysis, combined with topic modeling techniques like Latent Dirichlet Allocation (LDA), is effective in uncovering hidden topics within large text corpora, as shown in educational video comments analysis [4]. The Support Vector Machine (SVM), a supervised learning algorithm, has proven effective for text classification task highlighting its utility in research categorization. Addressing the common challenge of data imbalance in sentiment analysis, the Synthetic Minority Over-sampling Technique (SMOTE) has been employed [7], to enhance the performance of machine learning models by balancing data distribution. Evaluation metrics like coherence scores and confusion matrices are critical for assessing the effectiveness of topic models and classifiers in their topic analysis on Twitter data [6]. Previous studies integrating LDA and SVM, such as those in EdLink application reviews in analyzing COVID-19 vaccination refusal, underscore the versatility and effectiveness of these methods across different domains [8]. This study aims to leverage these established techniques to provide valuable insights into societal attitudes towards the chidfree lifestyle, contributing to the broader field of social sentiment analysis.

3. METHODOLOGY

The methodology provides a detailed outline of the technical steps and procedures that will be implemented throughout the research process.

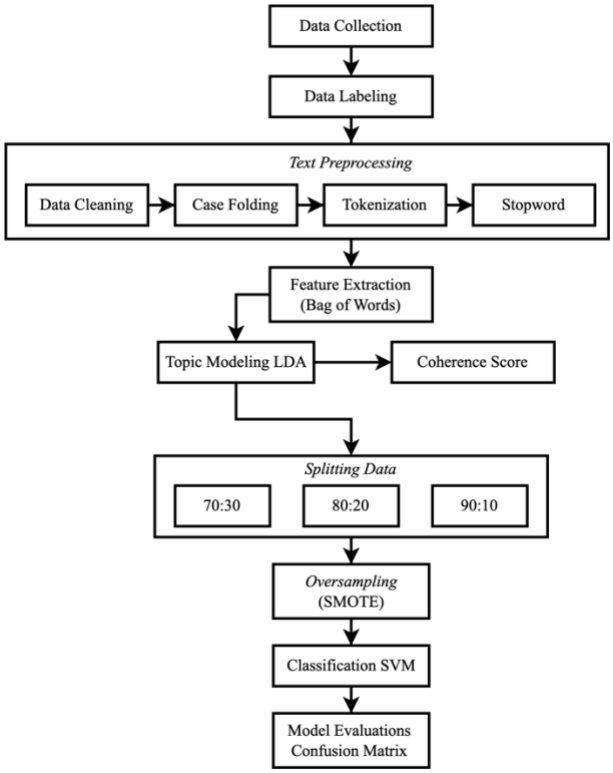

FIGURE 1. RESEARCH METHODOLOGY

3.1 Data Collection

Figure 1 illustrates the stages undertaken in this research, with the first stage being the collection of the dataset sourced from Twitter via the Drone Emprit Academy platform. The dataset is stored in a .csv file format, and this raw dataset will later be processed in this research.

3.2 Data Labeling

The data obtained from Twitter is then labeled. In the data labeling process, the polarity of the data is calculated first, and then the data is divided into three sentiment classes: positive, negative, and neutral. The positive class is assigned a value of 1, the negative class is assigned a value of -1, and the neutral class is assigned a value of 0.

3.3 Text Preprocessing

This stage helps improve data quality and ensure accuracy in the analysis process. As shown in Figure 1,

data is first cleaned through a data cleaning process, then all text is converted to lowercase in the case folding step. Essentially, the tokenization process aims to separate each word in the document. During this process, characters such as emojis, punctuation marks, links, hashtags, and URLs are removed. After tokenization, the next step is the removal of stopwords, where words that do not have significant meaning or influence are eliminated. The purpose of removing stopwords is to increase the signal-to-noise ratio in unstructured text, thereby enhancing the statistical significance of potentially important terms [7]. The stemming process concludes with the goal of removing prefix and suffix affixes. This technique uses grammar by extracting the root word or base word that cannot be further broken down [8].

3.4 Bag of Words

The technique you are referring to is the "Bag of Words" (BoW). BoW is an approach that represents a text object, such as a sentence or document, as a collection of words, disregarding grammar and word order. This approach focuses on the presence of words in the document without considering their structure or context. It is useful for text analysis where the information related to the presence of words is more relevant than the order or structure of the sentences.

TABLE 1. FORMATION BAG OF WORDS

Words	Frequency
Desire	1
Really	1
Cotton Candy	1
Compassion	1
Follower	1
Cult	1
Childfree	1

This process involves collecting unique words from the text, where each word is counted based on its frequency of occurrence in the document. These unique words are sorted once in different orders, and the frequency of each word is used to understand its importance in the context of the document. This is the initial step in understanding the distribution of words and their relevance in text analysis.

3.5 Latent Dirichlet Allocation (LDA)

LDA is a general probabilistic model that assumes each topic is a combination of a collection of potential words, also referred to as tokens, and each document (corpus) is a combination of a collection of probabilistic topics known as latent topics [7]. The advantage of LDA is its ability to automatically group keywords to identify several topics that emerge from various opinions in each class [9]. The LDA algorithm works by initializing the parameters as follows:

a) Total documents (M)
b) Total topics (K) to be displayed
c) Total iterations (i)
d) Total words in the documents (N)
e) Coefficient LDA (α, β)

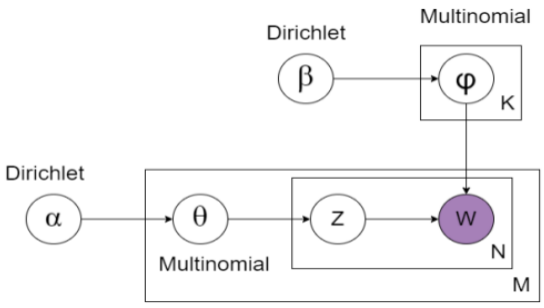

FIGURE 2. LDA WORKFLOW

LDA assigns labels to each word in the dataset according to the topic specified in the document (topic assignment) through word distribution. In this research, topic modeling aims to identify aspects within tweets. The number of topics becomes an important factor with the coherence score serving as the evaluation metric. The higher the coherence score, the better the resulting model will be. In this research, the modeling is conducted using the following steps:

1. Data that has been processed includes results from preprocessing that are labeled with negative, neutral, and positive sentiments.
2. The construction of lexicons is carried out to identify unique terms that are used within the corpus construction.
3. Calculation of coherence values between various topic distributions is performed using the LDA method. The process continues with the aggregation of data labeled with sentiments.
4. By calculating the topic presence contribution in each text. Finally, the resulting data set is interpreted through graphs, word clouds, and analysis.

3.6 Coherence Score

The coherence score is used to evaluate topic models. For a coherence value of 4 segments, it generally involves segmentation that groups data into clusters of words. Estimating the likelihood of analyzing the possibility of clustering those words together. To determine the overall coherence score, the measure of coherence indicates some quality of one group of words supporting another.

3.7 Splitting Data

Before constructing the Support Vector Machine (SVM) classification model, the data was divided into training and testing sets, with the testing data comprising 10-30% and validated three times. Subsequently, resampling was performed to optimize the data imbalance and achieve a more equitable data distribution [11].

3.8 Oversampling (SMOTE)

In this research, to address imbalance in the distribution of data, the SMOTE oversampling method was employed. Imbalance refers to some classes having relatively few data points, while others have significantly more. The objective of oversampling is to balance the number of samples between minority and majority classes by duplicating instances from the minority class. This approach also helps create a more balanced dataset and enhances the model's performance in identifying minority classes.

3.9 Support Vector Machine Classification Modeling

In this research, data were divided into 10-30% during the data splitting process, followed by resampling using SMOTE oversampling technique. The resulting dataset was then classified using SVM modeling to produce an optimal model. The method to analyze data and find patterns used for classification with Support Vector Machine involves training data and subsequently selecting the most accurate classification [12]. Support Vector Machine (SVM) is one of the classification methods aimed at finding the maximum margin hyperplane (MMH) [13].

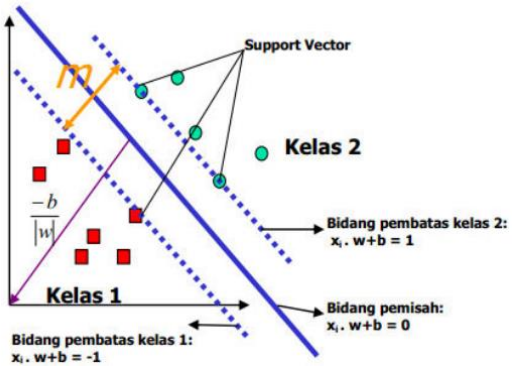

FIGURE 3. HYPERPLANE SVM

Both layers are separated by a parallel hyperplane. The first layer boundary is the first layer boundary, while the second layer boundary is the second layer boundary.

$$xi .w + b \geq +1 \text{ foryi} = +1$$
$$xi .w + b \leq -1 \text{ foryi} = -1 \quad (1)$$

Noted:

w : Normal vector

b : Relative position of the bias relative to the coordinate center

Generally, the working principle of SVM involves finding the optimal separation line (hyperplane) with two classes. The process of determining the optimal separation line continues until an optimal hyperplane is found. Therefore, SVM optimization is needed to find the maximum margin between hyperplanes with two classes. To enhance SVM, dual form optimization is used to find the hyperplane. The initial optimization form is the primitive form of SVM, and the second form is the SVM primal form [14]. The primitive form is not used in this study because it does not meet the required conditions.

$$C = S \times M \times P \times \Sigma \quad (2)$$

C = Coherence Value

S = Segmentaion

M = Confirmaion Meaure

P = Probaility Estimaion

Σ = Agregation

3.10 Confusion Matrix

Confusion Matrix is used to measure the performance of a classification system by comparing the classification results of the system with the actual classification. This table categorizes the number of test data that are correctly and incorrectly classified.

TABLE 2. CONFUSION MATRIX

Correct Classification	Predicted (+)	Predicted (-)
Actual "+"	True Positives	False Negatives
Actual "-"	False Negatives	True Positives

a. True Positives (TP) are the number of positive data records classified as positive.
b. False Positives (FP) are the number of negative data records classified as positive.
c. False Negatives (FN) are the number of positive data records classified as negative.
d. True Negatives (TN) are the number of negative data records classified as negative.

4. RESULT AND DISCUSSION

This research utilizes Python programming and tools available on Google Collab. The observational analysis focuses on a childfree dataset containing both pro and contra comments from Twitter users. Based on literature study, Support Vector Machine (SVM) is chosen as the classification model suitable for text data. Latent Dirichlet Allocation (LDA) topic modeling aids in identifying aspects within the modeling process. SMOTE oversampling assists in optimizing the dataset for balanced classification without losing non-representative data.

4.1 Data Collection

The dataset was collected in February-March 2023, obtaining 5000 tweets from Twitter. The data was saved in .csv format and analyzed using panda's library. The raw dataset in Table 3 consists of unprocessed tweets that have not undergone cleaning, which can influence the modeling process.

TABLE 3. EXAMPLE RAW DATASET

No.	Type	Mention	Date	Link
0	Mention	Pengen banget gw gulai rahim penganut sekte ch...	15/03/2023 03:00:46	https://twitter.com/web/statuses/163576...
1	RT	RT @t_gilik yak bagus, video ini semakin meneg...	15/03/2023 03:45:50	https://twitter.com/web/status...
2	RT	RT @Ienad_28 @tanyakank Nanti alasannya 'Rez...	15/03/2023 04:42:27	https://twitter.com/web/statuses/16357...
…	…	….	…..	…..

4.2 Data Labeling

The data was labeled with 3 categories of sentiment: negative, neutral, and positive. The results are shown in the following Table 4:

TABLE 4. PREPROCESSED DATA

No.	Date	Author	Mentions	Sentiment
0	2023-03-15 05:00	@ygman90s (yogaymjd)	Pengen banget gw gulai rahim penganut sekte...	Negative
1	2023-03-15 04:55	@TheLegitimateP1 (Courier 74)	RT @f_gilik yak bagus, video ini semakin meneg...	Positive
2	2023-03-15 04:42	@shtttshut (A¢AcA AgAsAgA*AVA'AAceAc)	RT @leinad_28 @tanyakanrl Nanti alesannya 'Rez...	Negative
3	2023-02-26 10:14	DewiFriciliat (si ayang)	RT@AREAJULID Orang yang notabennya ter...	Neutal
…	…	….	…..	…..

4.3 Text Preprocessing

The dataset used has not been directly processed by the system but has undergone several preprocessing steps to enhance the quality of the data.

4.4 Cleaning

The dataset has not undergone any preprocessing yet and remains in its original form. To achieve optimal results, the dataset needs to be cleaned first by removing unnecessary elements such as dropping columns that are not used. The cleaning results can be found in the following figure.

TABLE 5. CLEANING RESULT

No.	Mentions_remove_user	Mentions_cleaning
0	Pengen banget gw gulai rahim penganut sekte ch...	Pengen banget gw gulai rahim penganut sekte ch...
1	RT yak bagus, video ini semakin menegaskan ba...	yak bagus video ini semakin menegaskan bahwa c...
2	RT Nanti alesannya 'Rezeki mah ada aja, seka...	Nanti alesannya Rezeki mah ada aja sekarang
3	RT : Video anti childfree tapi isinya contoh o...	Video anti childfree tapi isinya contoh yg ora...
…	…..	…..

4.5 Case Folding

Case folding converts all letters to lowercase and removes any characters other than alphabetic letters using the str.lower() function in Python programming. Table 6 shows the results of case folding.

TABLE 6. CASE FOLDING RESULTS

No.	Mentions_remove_user	Mentions_case_folding
0	Pengen banget gw gulai rahim penganut sekte ch...	Pengen banget gw gulai rahim penganut sekte ch...
1	RT yak bagus, video ini semakin menegaskan ba...	yak bagus video ini semakin menegaskan bahwa c...
2	RT Nanti alesannya 'Rezeki mah ada aja, seka...	Nanti alesannya Rezeki mah ada aja sekarang
3	RT : Video anti childfree tapi isinya contoh ora...	Video anti childfree tapi isinya contoh yg ora...
…	…..	…..

4.6 Tokenizing

The dataset that has undergone case folding subsequently was tokenized into words using the word_tokenize() function from the nltk.tokenize module. This result is shown in Figure 7 below.

TABLE 7. TOKENIZING

No.	Mentions_remove_user	Mentions_tokenization
0	Pengen banget gw gulai rahim penganut sekte ch...	[pengen, banget, gw, gulai, rahim, penganut, s…
1	RT yak bagus, video ini semakin menegaskan ba...	[yak, bagus, video, ini, semakin, menegaskan...
2	RT Nanti alesannya 'Rezeki mah ada aja, seka...	[Nanti, alesannya, rezeki, mah, ada, aja, seka…
3	RT: Video anti childfree tapi isinya contoh ora...	[Video, anti, childfree, tapi, isinya, contoh, ora...
…	…..	…..

4.7 *Stopwords*

Stopwords help maintain more accurate analysis by removing words that appear frequently but do not carry much meaning. Stopwords in Indonesian were handled using the Sastrawi module with nltk.corpus. The Stopword results are shown in Table 8.

TABLE 8. STOPWORDS RESULTS

No.	Mentions_remove_user	Mentions_case_folding
0	Pengen banget gw gulai rahim penganut sekte ch...	Pengen banget gw gulai rahim penganut sekte ch...
1	RT yak bagus, video ini semakin menegaskan ba...	yak bagus video ini semakin menegaskan bahwa c...
2	RT Nanti alesannya 'Rezeki mah ada aja, seka...	Nanti alesannya Rezeki mah ada aja sekarang
3	RT : Video anti childfree tapi isinya contoh ora...	Video anti childfree tapi isinya contoh yg ora...
…	…..	…..

4.8 Bag of Words

The Bag of Words model extracts feature from the text and classifies the text [15]. Bag of Words counts the frequency of words in each document. In this study, the library “CountVectorizer” was used to implement BOW.

TABLE 9. BAG OF WORDS RESULTS

No.	aa	aaa	yzvous	zalina	zaman	zeke
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
…	…..	…..	…..	…..	…..	…..
1838	0	0	0	0	0	0
1839	0	0	0	0	0	0
1840	0	0	0	0	0	0
1841	0	0	0	0	0	0
1842	0	0	0	0	0	0

4.9 Latent Dirichlet Allocation (LDA)

The LDA model can identify topics that are subjectively relevant; however, calculating coherence before starting topic modeling is necessary to ensure a more accurate interpretation of these topics. This process involves using a prepared corpus and a relevant lexicon.

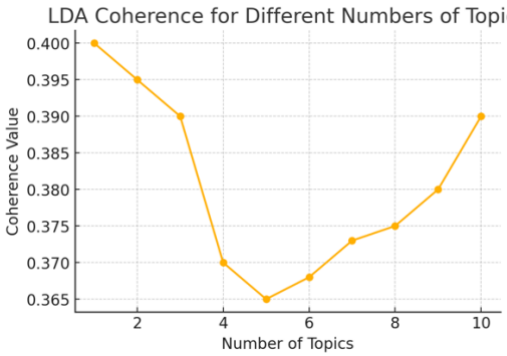

Figure 10. Coherence Score Graph

The coherence score decreases for a single topic, which is 0.400 when taking 10 topic categories. The most suitable number of topics is selected based on the highest coherence score [16]. Number of topics: 1 - Coherence score: 0.400. Number of topics: 2 - Coherence score: 0.386. Number of topics: 3 - Coherence score: 0.387. Number of topics: 4 - Coherence score: 0.365.

Number of topics: 5 - Coherence score: 0.370
Number of topics: 6 - Coherence score: 0.372
Number of topics: 7 - Coherence score: 0.366
Number of topics: 8 - Coherence score: 0.368
Number of topics: 9 - Coherence score: 0.375
Number of topics: 10 - Coherence score: 0.390.

4.10 Oversampling SMOTE

DalTalSet was balanced using oversampling SMOTE from the imbalanced-learn module by importing SMOTE. This method significantly increases the number of samples representing the minority class [11]. The number of samples before oversampling was 1,843. The number of samples after oversampling was 3,222.

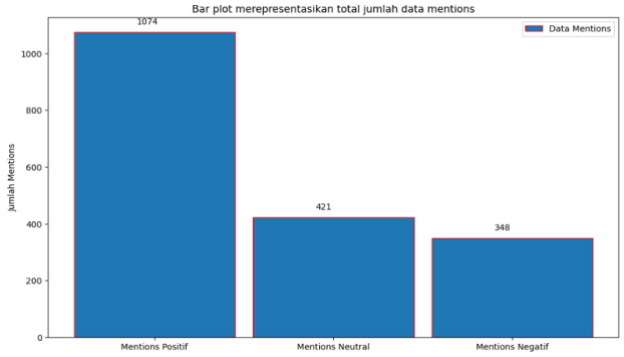

FIGURE 4. RESULT BEFORE SMOTE OVERSAMPLING

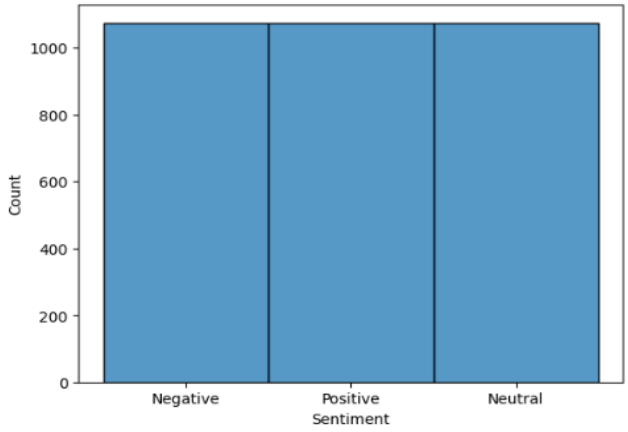

FIGURE 5. RESULT AFTER SMOTE OVERSAMPLING

Figure 5 shows the class distribution after oversampling. Balancing the distribution between minority and majority classes can improve the model's ability to predict minority classes more accurately. The oversampling method helps to correct class imbalance, as seen from the comparison of these two graphs, thereby enhancing the model's performance in the classification task for the childfree dataset.

4.11 Splitting Data

Data splitting is divided into 3 percentages after going through the Bag of Words and Latent Dirichlet Allocation processes. The highest result achieved was an accuracy of 67% with SMOTE oversampling after the LDA process, yielding the following results.

TABLE 10. SPLITTING DATA

Splitting Data	Without SMOTE	SMOTE Oversampling
90:10	57%	67%
80:20	54%	63%
70:30	54%	60%

4.12 Support Vector Machine (SVM)

In this research, text classification is performed using the Support Vector Machine (SVM) method with negative, neutral, and positive labels through the aid of Latent Dirichlet Allocation (LDA) and bag of words to achieve optimal results.

TABLE 11. SPLITTING DATA

Splitting Data	LDA+SVM	LDA+SVM SMOTE
90:10	64%	57%
80:20	61%	49%
70:30	59%	50%

The results indicate that SMOTE does not always improve performance when used in conjunction with LDA and SVM. After applying SMOTE, the accuracy decreased from 64% to 57% in the 90:10 data split. The same occurred in the 80:20 and 70:30 splits, where accuracy dropped significantly.

4.13 Model Evaluation with Confusion Matrix

The confusion matrix in Figure 6 shows that data not using SMOTE cannot be classified correctly because the model cannot learn the data well due to data imbalance. All data is included in the positive class, necessitating data balancing as shown in Figure 7.

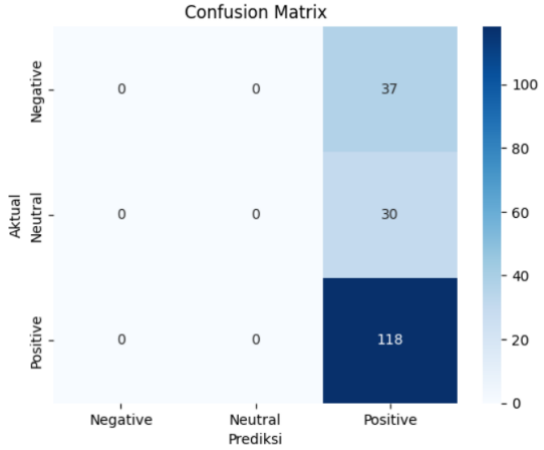

FIGURE 6. VISUALIZATION WITHOUT SMOTE

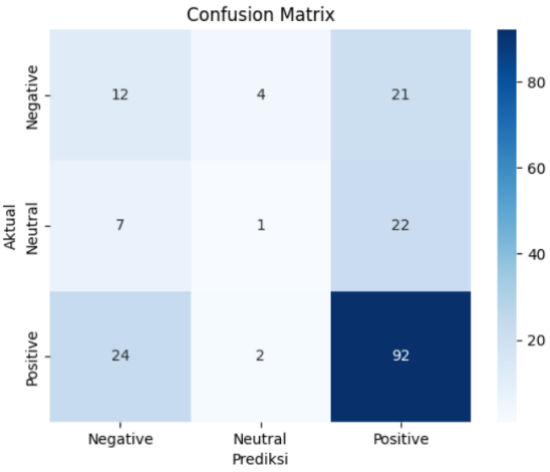

FIGURE 7. VISUALIZATION WITH SMOTE

The figure below shows that, in the negative class, out of 37 actual Negative samples, 21 were predicted as Positive. 12 samples were identified as Negative. In the neutral class, out of 30 actual Neutral samples, 22 were predicted as Positive. Only 1 sample was correctly identified as Neutral. In the positive class, out of 118 Positive samples, 92 were correctly predicted as Positive. 24 Positive samples were predicted as Negative. 2 Positive samples were predicted as Neutral.

FIGURE 8. WORD CLOUD VISUALIZATION

The word cloud displays dominant words in each user's comments on Twitter. The frequently appearing words include "single," "sleep," "sex," "LGBT," "hobby," "teacher," and "wife". Modeling topics with LDA through several text preprocessing steps resulted in differences in weight, the number of topics, and coherence scores. These differences occurred because LDA uses a probabilistic approach. SVM performs processing for classification with results that are obtained after LDA.

The modeling of LDA+SVM and LDA+SVM+oversampling SMOTE resulted in different accuracies and classification outcomes. The highest accuracy was achieved by LDA+SVM at 64%, while the accuracy for LDA+SVM+oversampling SMOTE was 57%. This occurred because the model could not optimally learn new data, especially since this research used the *Childfree* dataset. The research Using Latent Dirichlet Allocation (LDA) and Support Vector Machine (SVM) to Analyze Aspect-Based Sentiment in EdLink Application Reviews shows a coherence score of 0.487. The Lexicon-Based approach identified 1,223 reviews reflecting negative sentiment and 418 reviews reflecting positive sentiment. In testing using the Support Vector Machine (SVM) method with SMOTE, the accuracy reached 90.00%. The evaluation of the EdLink application indicates that improvements in reliability and performance, including feature updates and error corrections, are highly necessary.

5. CONCLUSIONS

In the confusion matrix, the performance of the Positive case is better than the Negative case and the Blind case with the help of LDAL+SVM+ in addition to SMOTE. The highest coherence value is generated at the number of topics 1 reaching 0.400. The best 10 topics generated by LDAL were related to comments on chidfree topics. Some of the challenges encountered include incomplete preprocessing, inaccurate customization, and LDAL topic coding that is less effective in clustering aspects that have similarities between topics. The handling of gas imbalance can be improved to get more optimal results. Improve feature selection and mode parameters. Using data augmentation techniques or other methods to improve performance on under-resourced keas.

REFERENCES

[1] V. & Audinovic and R. S. Nugroho, "Persepsi Childfree Di Kalangan Generasi Zilenial Jawa Timur," 2023.

[2] N. Trianasari and M. S. Ilmanizar, "Analisis Respon Pengguna Twitter Terhadap Tragedi Kanjuruhan Malang Menggunakan Setiment Analysis Dan Topic Modelling," vol. 11, no. 1, pp. 1–11, 2022.

[3] B. P. Statistik, "Jumlah Perceraian Menurut Provinsi dan Faktor," BADAN PUSAT STATISTIK. Accessed: Mar. 20, 2024. [Online]. Available: https://www.bps.go.id/id/statistics-table/3/YVdoU1IwVmlTM2h4YzFoV1psWkViRXhqTlZwRFVUMDkjMw==/jumlah-perceraian-menurut-provinsi-dan-faktor.html?year=2022

[4] Albert, "Analisis Topik dan Perbandingan Klasifikasi pada Kolom Komentar Video Youtube Edukasi Indonesia Menggunakan Pendekatan Latent Dirichlet Allocation," *Journal on Education*, vol. 05, no. 03, pp. 7418–7429, 2023.

[5] R. Astuti, R. Andini Husen, A. Triono, M. Khairul Anam, P. Studi Teknik Informatika STMIK Amik Riau, and J. K. Purwodadi Indah, "Peningkatan Metode Support Vector Machines (SVM) pada Data Child-free Menggunakan Oversampling," vol. 2, no. 1, pp. 19–27, 2023.

[6] P. Patmawati and M. Yusuf, "Analisis Topik Modelling Terhadap Penggunaan Sosial Media Twitter oleh Pejabat Negara," *Building of Informatics, Technology and Science (BITS)*, vol. 3, no. 3, pp. 122–129, 2021, doi: 10.47065/bits.v3i3.1012.

[7] U. Malihatin S, Y. Findawati, and U. Indahyanti, "Topic Modeling in Covid-19 Vaccination Refusal Cases Using Latent Dirichlet Allocation and Latent Semantic Analysis," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 5, pp. 1063–1074, 2023, doi: 10.52436/1.jutif.2023.4.5.951.

[8] Anjali, G. Jivani, and M. Anjali, "A Comparative Study of Stemming Algorithm," *October*, vol. 2, no. 2004, pp. 1930–1938, 2007.

[9] N. L. P. M. Putu, Ahmad Zuli Amrullah, and Ismarmiaty, "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 123–131, 2021, doi: 10.29207/resti.v5i1.2587.

[10] R. P. F. Afidh and Syahrial, "Pemodelan Topik Menggunakan n-Gram dan Non-negative Matrix Factorization," *Jurnal Informasi dan Teknologi*, vol. 5, no. 1, pp. 265–275, 2023, doi: 10.60083/jidt.v5i1.385.

[11] Ridwan, H. E. Hermaliani, and M. Ernawati, "Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada," 2024. [Online]. Available: http://jurnal.bsi.ac.id/index.php/co-science

[12] Suhardjono, W. Ganda, and H. Abdul, "Prediksi Kellusan Menggunakan Svm Berbasis Pso," *Bianglala Informatika*, vol. 7, no. 2, pp. 97–101, 2019.

[13] Y. Kustiyahningsih and Y. Permana, "Penggunaan Latent Dirichlet Allocation (LDA) dan Support-Vector Machine (SVM) Untuk Menganalisis Sentimen Berdasarkan Aspek Dalam Ulasan Aplikasi EdLink," *Teknika*, vol. 13, no. 1, pp. 127–136, 2024, doi: 10.34148/teknika.v13i1.746.

[14] F. S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 1, no. 1, pp. 19–25, 2017, doi: 10.29207/resti.v1i1.11.

[15] T. Mardiana, "Bag of Words Clustering Using Weka," no. 2, pp. 0–5, 2016, doi: 10.13140/RG.2.1.4763.2807.

[16] Dinda Adimanggala, Fitra Abdurrachman Bachtiar, and Eko Setiawan, "Evaluasi Topik Tersembunyi Berdasarkan Aspect Extraction menggunakan Pengembangan Latent Dirichlet Allocation," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, pp. 511–519, 2021, doi: 10.29207/resti.v5i3.3075.

AUTHORS

Ratna Andini Husen

She is currently a researcher at STMIK AMIK Riau, Pekanbaru, Indonesia. Her research focuses on the areas of Informatics engineering, Machine learning and sentiment analysis.

Agustin

She is a dedicated Head of Study Program in Informatics Engineering. Study program in Informatics Engineering, STMIK AMIK Riau, Indonesia. With passion for the academic world, he is actively involved in teaching and mentoring students while contributing to the advancement.

Susi Erlinda

She is a dedicated lecturer in the Informatics Engineering study program in Informatics Engineering, STMIK AMIK Riau, Indonesia. With passion for the academic world, he is actively involved in teaching and mentoring students while contributing to the advancement.

Junadhi

He is a senior lecturer of vice chair III with a focus on fostering informatics engineering students in the fields of machine learning, mobile, website and data science. His field research focuses on Image Processing, NLP, and Information Retrieval. Information Retrieval.

Thinagaran Perumal

He received his B.Eng., M.Sc. and Ph.D. Smart Technologies and Robotics from Universiti Putra Malaysia in 2003, 2006, and 2011, respectively. Currently, he is an Associate Professor at Universiti Putra Malaysia. He is also Chairman of the TC16 IoT and Application WG National Standard Committee and Chair of IEEE Consumer Electronics Society Malaysia Chapter. Dr. Thinagaran Perumal is the Recipient of the 2014 IEEE Early Career Award from IEEE Consumer Electronics Society. His recent research activities include proactive architecture for IoT systems, development of the cognitive IoT frameworks for smart homes and wearable devices for rehabilitation purposes.