Sentiment Analysis on the Impact of MBKM on Student Organizations Using Supervised Learning with Smote to Handle Data Imbalance

Lailatul Cahyaningrum[1], Ardytha Luthfiarta[2*], Mufida Rahayu[3]

[1,2,3]*Informatics Engineering Department, Universitas Dian Nuswantoro, Semarang, Indonesia*

[1]lailacahya22@gmail.com

[2]ardytha.luthfiarta@dsn.dinus.ac.id(*)

[3]mufidarahayu2002@gmail.com

Received: 2023-12-07; Accepted: 2024-01-05; Published: 2024-01-15

Abstract— Recently, there has been a decline in student interest in joining organizations. One of the causes is the *MBKM program "Merdeka Belajar Kampus Merdeka*". With this program from the government, more and more students are interested in entering because it is considered more profitable. Responses regarding this were conveyed by students through questionnaires, Twitter crawling, and YouTube comments. The data obtained was 1,770 (negative, positive, and neutral labeling) using Sastrawi, Nazief & Adriani, and Arifin Setiono stemming. There is an imbalance of data in labeling, so it is necessary to do SMOTE to balance the data. The algorithms used in the research focus on modeling the Naïve Bayes Classifier, Support Vector Machine, and Decision Tree with the split random method, with the best results using Support Vector Machine. Of the three algorithms, the highest results were obtained from the results of Arifin Setiono's data setmming, using a Support Vector Machine with 91% accuracy, obtained from 90% training data and 10% testing.

Keywords— *MBKM*; Sentiment Analysis; SMOTE; Stemming; Classification; SVM.

I. Introduction

Students become agents of change in the world in the future, especially during the golden age in 2045, when Indonesia reaches a century and experiences a demographic bonus [1]. In addition to being agents of change, students are also iron stock, the next generation of the nation who will be able to replace government leaders for the better [2]. On the way to becoming agents of change and iron stock, students can gather in student organizations in addition to lectures because organizations are also a means of learning to develop their intellectual, social, and religious abilities [3]. Organizations are also useful for increasing student awareness and involvement in society, increasing professionalism, instilling creativity, and increasing critical thinking, which will be useful for post-campus life [4]. However, today the presence of *Organisasi Kemahaisswaan (ormawa) and Unit Kegiatan Mahasiswa (UKM)* on Indonesian campuses has decreased in existence due to the *Merdeka Belajar Kampus Merdeka (MBKM)* program, one of which is a program of the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia *(Kemendikbud Ristek)*, the programs in *MBKM* are now increasingly numerous branches [5].

In September 2022, a survey at Jenderal Soedirman University revealed that 20.5% of active organization enthusiasts participated. Additionally, the percentage of those enthusiasts who preferred the *MBKM* program was four times higher than the overall organizational participation rate [6]. The survey results from several campuses in Semarang and Yogyakarta realized that with the program, the interest in Ormawa and UKM decreased, and the decline was an obstacle to the management's regeneration and performance. In the Tourism Education Student Association organization for the 2022/2023 period, out of 20 respondents, 64.7% were not interested in joining the organization's management [5]. They prefer to join the *MBKM* program because the benefits offered are considered more beneficial for them in the academic realm and future career preparation.

Students who participate in the organization convey opinions about the influence of *MBKM* on the organization, some of which they convey through Twitter, YouTube, and distributing questionnaires. The opinion data they convey is processed into sentiment analysis by understanding, managing data, and extracting computations using text mining techniques to produce positive, neutral, or negative categories [7]. Various algorithms are applicable for handling sentiment analysis, encompassing the Naïve Bayes Classifier (NBC), K-Means clustering, Decision Tree, and Support Vector Machine (SVM). The Decision Tree algorithm is easy to implement using a recursive algorithm. However, the Decision Tree also has the disadvantage of being applied to a large amount of data because it has the potential to experience overfitting. Overfiting will make the algorithm's performance decrease[8].

In several studies related to sentiment analysis on the implementation of the MBKM, as much as 475 data were obtained: 99.22% for Naïve Bayes, 96.90% for K-Nearest Neighbors, and 37.21% for Decision Tree[9]. A total of 849 data regarding sentiment analysis related to the *Merdeka Belajar Kampus Merdeka (MBKM)* Program produced a Support Vector Machine of 84.76% higher than the Decision Tree, which only produced an accuracy of 72.86% [10]. The comments in the telegram group of the supervisors obtained 591 comments, which showed an accuracy of 99.30% for Naive Bayes and 97.20% for K-Nearest Neighbors [11]. From November 20, 2021 to December 19, 2021, sentiment analysis was conducted by crawling data on Twitter. A total of 5980 data

DOI : https://doi.org/10.25139/inform.v9i1.7484

points were collected, giving the SVM algorithm a success rate of 73.12% and the Naïve Bayes classifier method a success rate of 67.92% [12]. Twitter sentiment analysis of the *MBKM* program with 1,212 data resulted in an accuracy value of 74.25 using the Naïve Bayes Classifier Algorithm [13].

It is necessary to conduct a sentiment analysis of the impact of *MBKM* on the sustainability of organizational life for students. Several studies related to sentiment analysis on the MBKM program have been carried out. However, those that discuss the impact on student organizations are still very minimal or can be said to have never been done by previous researchers. For this reason, researchers plan to take opinion data from questionnaires, crawling Twitter and YouTube, and then model it with positive, negative, and neutral classes using a lexicon with Sastrawi, Nazief & Adriani, and Arifin Setiono stemming. Due to the potential data imbalance, this research will apply the oversampling method with the SMOTE technique.

II. Research Methodology

This research uses a comparison of the Naïve Bayes Classifier, Support Vector Machine, and Decision Tree algorithms to see the best accuracy of reviews about the influence of *MBKM* on the organization. Implementing the comparison model requires several stages to get the best accuracy results. The following in Figure 1 describes the research stages.

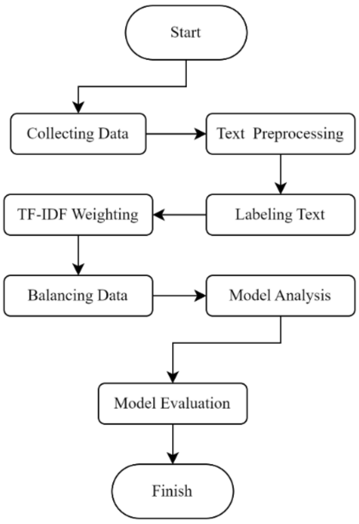

Figure 1. Workflow of Research Stages

A. Collecting Data

Three different sources were utilized to collect the data: crawling Twitter and YouTube, as well as questionnaires. Twitter was selected because the most recent information is always readily available on Twitter, particularly considering that approximately 69% of journalists engage with Twitter [14]. If we choose videos that match the problem, comments from YouTube will get precise results, and YouTube users have increased quite dramatically, around 73% [15], after this pandemic. The questionnaires were distributed to several public and private universities, which in their questionnaire answers are more detailed in describing how the impact is felt.

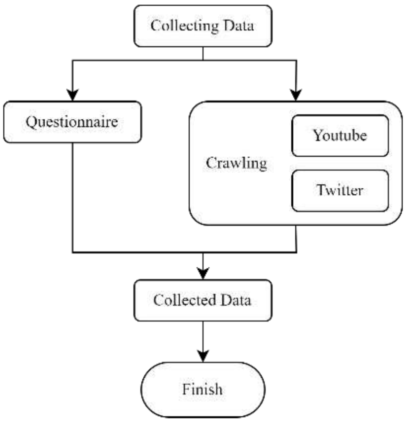

Figure 2. Workflow of Collecting Data

Figure 2 shows that this research uses data from crawling Twitter to get 45 data, questionnaires as much as 215 data, and YouTube comments as much as 1,510, so a total of 1,770 data is obtained. Crawling Twitter with keywords "*MBKM vs. Ormawa*" and "*MBKM atau Organisasi*" and crawling YouTube with API using Python.

B. Text Preprocessing

Text preprocessing is used to obtain more quality information, which is usually done at an early stage. The description is in Figure 3 and is done at the beginning to analyze sentiment with data from Twitter, YouTube, and questionnaires that can affect classification performance [16].

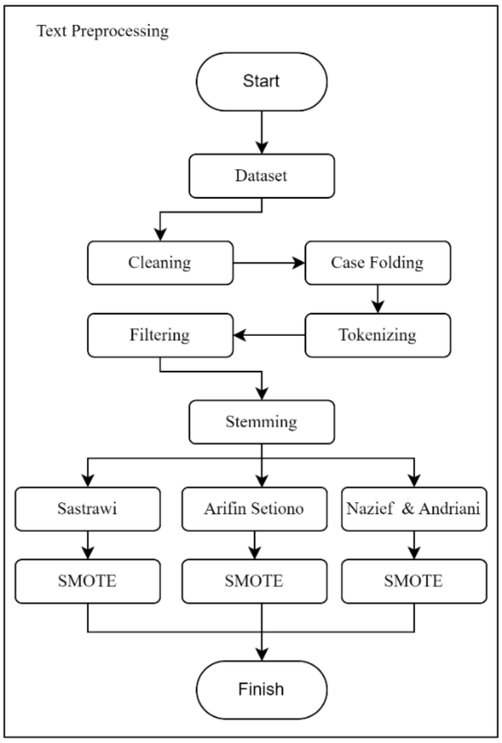

Figure 3. Workflow of Text Preprocessing

1) Cleaning: Cleaning is the process of removing usernames (@), hashtags #, numbers, URL (http://), delimiters such as commas (,), periods (.), and also other punctuation marks filtered from useless data in the data or commonly known as stopwords [17].

2) Case Folding: Data processing functions to convert all characters, such as capital letters, into lowercase letters [18].

3) Tokenizing: Processing sentences in the data into several words separated into important words only [19].

4) Filtering: Filtering is the process after tokenization to retrieve words that do not represent the content of a text document [20].

5)Stemming: Removal of affixes in the form of prefixes, suffixes, and confixes in each word so that it becomes a base word to homogenize the word form [20]. This research uses Sastrawi, Nazief, and Adriani, and Arifin Setiono. As a stemmer package, Sastrawi has not been designed to normalize abbreviations; for example, the word "*yg*" will not be known to mean "*yang*" by Sastrawi [21]. Bobby Nazief and Mirna Adriani developed the Nazief and Adriani algorithm in the stemming process. Additional rules, such as reduplication, prefixes, and suffixes, increase each word's accuracy [22]. The stemming of Arifin Setiono has a similar process to the stemming of Nazief & Adriani [23].

C. Labelling Data

Preprocessing makes the data only contain opinions that have been cleaned, and then positive, negative, and neutral labeling is done by determining the score of each sentence first [18].

D. TF-IDF weighting

The TF-IDF weighting method is common for generating vector sentences based on word vectors [24]. TF-IDF is a group of words or phrases to be numerically reduced by identifying the most important words in a document. Once converted into a numerical value, the numerical value is seen as the frequency of occurrence of the word. [25].

The result of the TF value does not provide important information in the converted word. This is because sometimes less useful conjugated or common words are counted to obtain the highest TF result of these words. Therefore, the Inferse Document Frequency (IDF) technique is required after using TF [26]. The Equation (1) is for the weighting of the TF-IDF.

$$Tf.IDF = TF_{ij}\,x\,IDF_{ij} = TF_{ij}\,x\log\frac{N}{DFj} \qquad (1)$$

In Equation (1) about TF IDF, N is the number of documents in the collection, TF term frequency, and IDF inverse document frequency.

E. Balancing Data

Data imbalance is still challenging because sometimes classes with minority numbers are more important [27]. This condition can later affect the prediction accuracy, decreasing with less data because it will tend towards one class and disadvantage the other [28]. Data balancing using the Synthetic Minority Oversampling Technique (SMOTE) operator can handle the problem of unbalanced data where the negative, positive, and neutral sentiment labels in the data set are not in proportional numbers [29]. Using the SMOTE operator, it was found that SMOTE can perform better than ADASYN [30].

F. Model Analysis

The analysis of the algorithm model used compares Naïve Bayes, SVM, and Decision Tree according to Figure 4, and then the three models are sought which produce the best accuracy value.

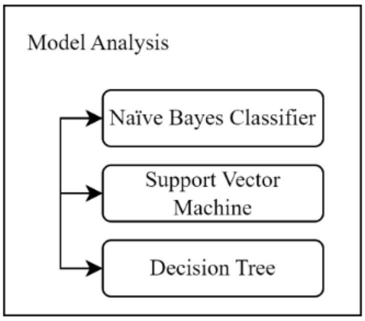

Figure 4. Workflow of Model Analysis

1) Naïve Bayes Classifier: Naive Bayes has a strong assumption of independence of each condition or event with a classification method based on Bayes' theorem [31].

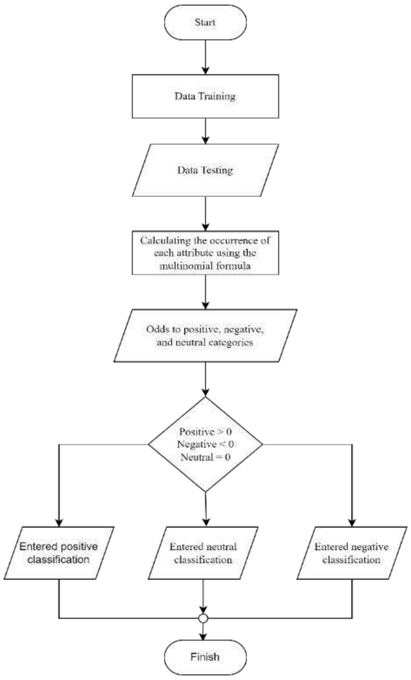
Start
Data Training
Data Testing
Calculating the occurrence of each attribute using the multinomial formula
Odds to positive, negative, and neutral categories
Positive > 0
Negative < 0
Neutral = 0
Entered positive classification
Entered neutral classification
Entered negative classification
Finish

Figure 5. Flowchart of Naïve Bayes Classifier

DOI : https://doi.org/10.25139/inform.v9i1.7484

Figure 5 is the flowchart of the Naïve Bayes Classifier and below is Bayes' theorem presented in Equation (2) *c*, *d* are events, *P(c|d)* = probability of *c* given *d* is true, *P(d|c)* probability of *d* if *c* is true, and *p(c), p(d)* independent probabilities of *A* and *B* respectively.

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \tag{2}$$

2) Support Vector Machine: SVM includes data classification using a hyperplane emphasizing risk minimization, which is the function estimation by minimizing the generalization error boundary [32]. SVM can overcome overfitting and produce a good classification model even though it is trained with relatively little data [17].

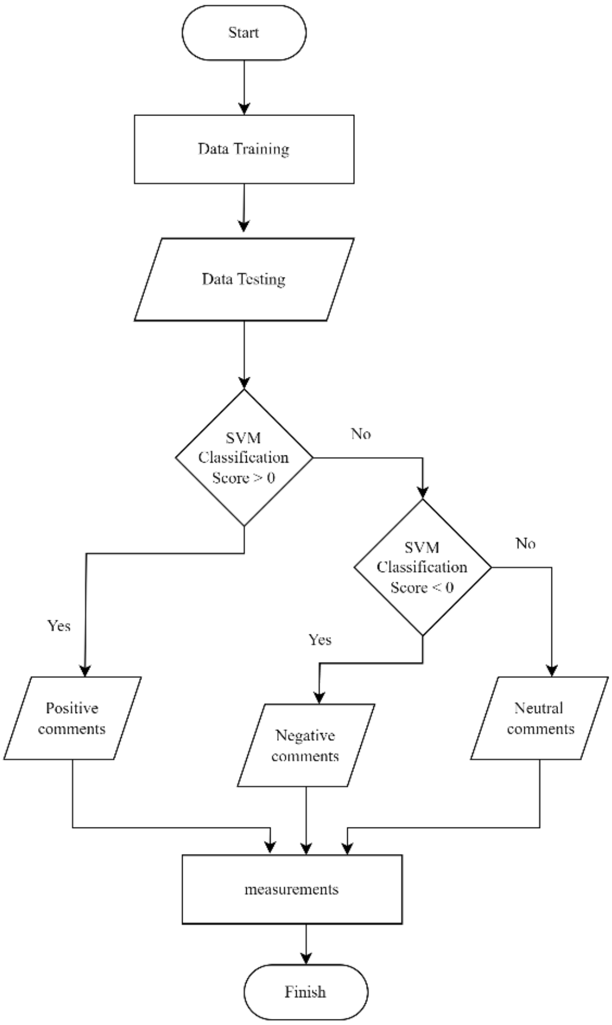

Figure 6. Flowchart of Support Vector Machine

Figure 6 is a flow chart of the Support Vector Machine, and below is the SVM Theorem presented in Equations (3) to (6) with the description Wi is vectors (W0, W1, W2, ... Wm), b biased term (W0), and X = variables.

$$y = W_0 + W_1W_1 + W_2W_2 \text{ ....} \tag{3}$$

$$= W_0 + \sum_{i=1}^{m} W_iX_i \tag{4}$$

$$= W_0 + W^TX \tag{5}$$

$$= b + W^TX \tag{6}$$

3) Decision Tree: The decision Tree classification algorithm uses a top-down decision tree structure to determine the research data class [33]. When attempting to measure the diversity or presence of a data set, it is necessary to have entropy and gain values. The decision tree obtains the entropy and gains values using herniations and equations.

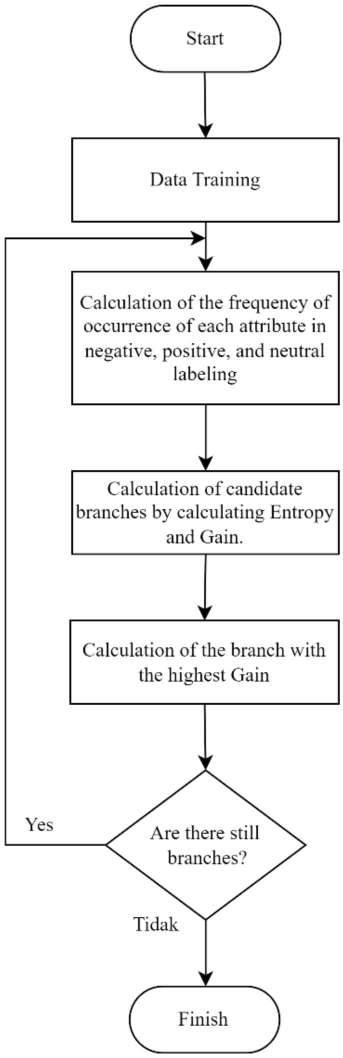
Start
Data Training
Calculation of the frequency of occurrence of each attribute in negative, positive, and neutral labeling
Calculation of candidate branches by calculating Entropy and Gain.
Calculation of the branch with the highest Gain
Are there still branches?
Yes
Tidak
Finish

Figure 7. Flowchart of Decision Tree

Figure 7 is the Decision Tree flowchart; below is the Decision Tree Theorem presented in Equations (7) and (8). Where the *S* variable is the Set of all possible outcomes, the *i* variable is individual outcomes, the *n* variable is probability of occurrence, the *pi* variable is probability, the *A* variable is the

DOI : https://doi.org/10.25139/inform.v9i1.7484

specific attribute, the $|Si|$ variable is the size of subset Si after splitting, and Si Entropy of each subset Si after splitting.

$$Entropy\ (S) = \sum_{i=1}^{n} - pi.\log_2 pi \tag{7}$$

$$Gain(S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si) \tag{8}$$

G. Evaluation

The evaluation process (Machine Learning) is carried out to test the classification results by measuring the truth value of the system. After obtaining the Naïve Bayes, SVM, and Decision Tree classification models, an evaluation process needs to be carried out before continuing to apply the model to the testing data by knowing the level of accuracy, precision, and recall [34].

$$precision = \frac{TP}{TP + FP} \tag{9}$$

$$recall = \frac{TP}{TP + FN} \tag{10}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

Equations (9), (10), and (11) explain that accuracy is the ratio of correct predictions (positive, neutral, and negative) to overall data, precision is the ratio of TP predictions to overall positive prediction results, and recall is the ratio of TP predictions to overall positive data, all of which can be obtained from the confusion matrix shown in the following formula [35]. TN is the opposite of TP, which means that both the system and the expert give a negative result. FP means that the system gives a negative result, but the expert gives a positive result; FN occurs when the system gives positive and negative results. [36].

III. Result and Discussion

In this study, the data on the influence of *MBKM* on organizations was processed using three types of stemming: Sastrawi, Nazief & Adriani, and Arifin Setiono. The algorithms that will be used are the Naïve Bayes Classifier, the Support Vector Machine, and the Decision Tree, which will be searched for the highest accuracy.

A. Collecting Data

The process of data collection is in 3 ways. The first was with a questionnaire through Google Forms, a total of 9 questions, but of the 9 questions are again grouped for those who follow *MBKM* and *Ormawa*, only *MBKM*, only *Ormawa*, and not both obtained 215 respondents' students across Indonesia. Twitter crawls with keywords "MBKM vs. Ormawa" and "MBKM atau Organisasi" yield 45 data. Then, crawling from YouTube comments using the programming language Python through Google Collaboratory gets 1,510 data, so the data obtained is 1,770.

B. Text Preprocessing

In this study, the data used from questionnaires, Twitter, and YouTube relates to the influence of *MBKM* on organizations, where this data has a variety of writing styles, so the text data obtained is unstructured data that is quite difficult to process. Before classification or tagging, the data needs to be transformed into more structured data, including negative, positive, or neutral. The stages of transforming unstructured data into structured information are called text preprocessing stages, and the steps are listed in Table I.

Table I
Text Preprocessing

Raw Data	
@UGM_FESS Lagian mbkm sebagian dikasih uang saku, beda sama beberapa organisasi yang setiap kali mereka bikin event lu sebagai panitia harus ngeluarin duit lumayan atau disuruh jualan sama upload story pp. Secara kebermanfaatan mbkm kadang lebih menarik dibanding organisasi	
Text Preprocessing	
Cleaning	*UGMFESS Lagian mbkm sebagian dikasih uang saku, beda sama beberapa organisasi yang setiap kali mereka bikin event lu sebagai panitia harus ngeluarin duit lumayan atau disuruh jualan sama upload story pp Secara kebermanfaatan mbkm kadang lebih menarik dibanding organisasi*
Case Folding	*ugmfess lagian mbkm sebagian dikasih uang saku beda sama beberapa organisasi yang setiap kali mereka bikin event lu sebagai panitia harus ngeluarin duit lumayan atau disuruh jualan sama upload story pp secara kebermanfaatan mbkm kadang lebih menarik dibanding organisasi*
Tokenizing	*ugmfess, lagian, mbkm, Sebagian, dikasih, uang, saku, beda, sama, beberapa, organisasi, yang, setiap, kali, mereka, bikin, event, lu, sebagai, panitia, harus, ngeluarin, duit, lumayan, atau, disuruh, jualan, sama, upload, story, pp, secara, kebermanfaatan, mbkm, kadang, lebih, menarik, dibanding, organisasi*
Filtering	*ugmfess lagian mbkm sebagian dikasih uang saku beda sama beberapa organisasi setiap kali mereka bikin event sebagai panitia ngeluarin duit lumayan disuruh jualan upload story pp secara kebermanfaatan mbkm kadang lebih menarik dibanding organisasi*
Stemming	*ugmfess lagi mbkm sebagian kasih uang saku beda sama berapa organisasi setiap kali mereka bikin event sebagai panitia keluar duit lumayan suruh jualan upload story pp secara manfaat mbkm kadang lebih tarik banding organisasi*

Table I shows the results of text preprocessing. The text preprocessing step uses cleaning, case folding, tokenizing, filtering, and stemming to get maximum text results before text labeling.

C. Labelling Text

Data that has gone through the text preprocessing stage is then labeled using a lexicon word dictionary. Labeling uses a lexicon with values less than 0 negative, more than 0 positive, and 0 neutral, as in Table II.

DOI : https://doi.org/10.25139/inform.v9i1.7484

TABLE II
LABELLING TEXT

Clean_Text	Polarity _score	Polarity
karena lebih efisien dan fleksibel iya keuntungan nya dapat banyak teman relasi ilmu dan pengetahuan baru kerugian nya lebih cape saja karena menghabiskan banyak waktu	0	Neutral
pingin ikut mbkm tapi masih susah dapat informasi setiap ukm pasti memiliki keuntungan dan kerugiannya masing yang saya rasakan ikut ukm itu banyak memiliki keuntungan yaitu bisa menambah relasi menambah wawasan belajar tanggung jawab dsb	-6	Negative
mbkm vs ormawa halo kema unpad saat ini masih banyak perdebatan mengenai relevansi peran organisasi kemahasiswaan dalam pengembangan diri anggotanya belum lagi dengan adanya program merdeka belajar kampus merdeka yang dinilai lebih baik dalam mengembangkan soft skill	11	positive

The results of Table III are the visualization results of word clouds from texts on negative, positive, and neutral sentiments, as well as on different stem results, namely the results of Sastrawi, Nazief & Adriani, and Arifin Setiono.

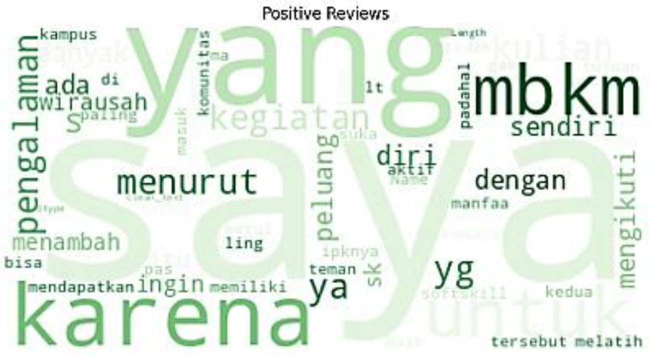

Figure 8. Word Cloud Positive

A positive word cloud in Figure 8 using the stemming Sastrawi yielded words that frequently occur: *'Saya', 'yang', 'karena', 'MBKM'*, and *'pengalaman'.*

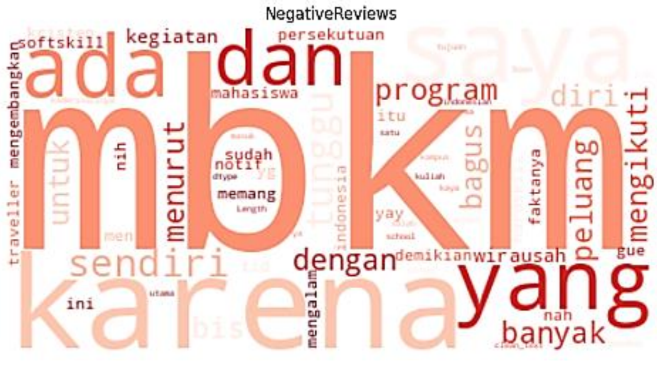

Figure 9. Word Cloud Negative

A negative word cloud in Figure 9 using Arifin Setiono stemming yielded frequently occurring words: *'MBKM', 'ada', 'karena', 'dan'*, and *'yang'.*

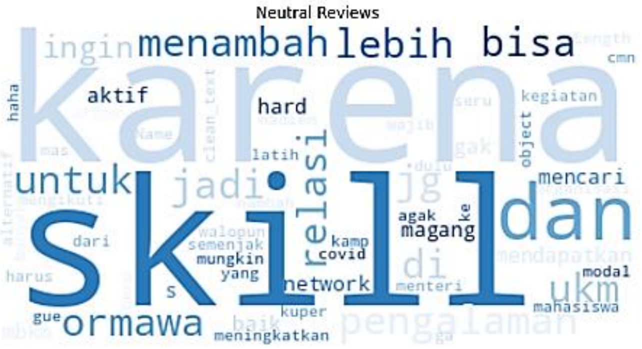

Figure 10. Word Cloud Neutral

A neutral word cloud in Figure 10 using the stemming Nazief and Andriani obtained words that often appear, namely "*karena*", "*skill*", "*dan*", "*ormawa*", and "*ukm*".

D. TF-IDF Weighting

TF-IDF is used to extract features using the Python Count Vectorizer library, namely the conversion of text features into a vector representation and the grinding of words using TF-IDF, after which the data can be tested by splitting it into test data and training data.

E. Data Balancing

Data balancing, using the Synthetic Minority Oversampling Technique (SMOTE) operator, addresses data imbalance issues where negative, positive, and neutral sentiment labels in datasets are not proportionate. In this research, the labeling process uses three types of stemming algorithms: Sastrawi, Nazief & Adriani, and Arifin Setiono. So that the results of the negative, neutral, and positive classifications of the three stemming algorithms are different. Here are some data results using the three stemming algorithms and before or after SMOTE.

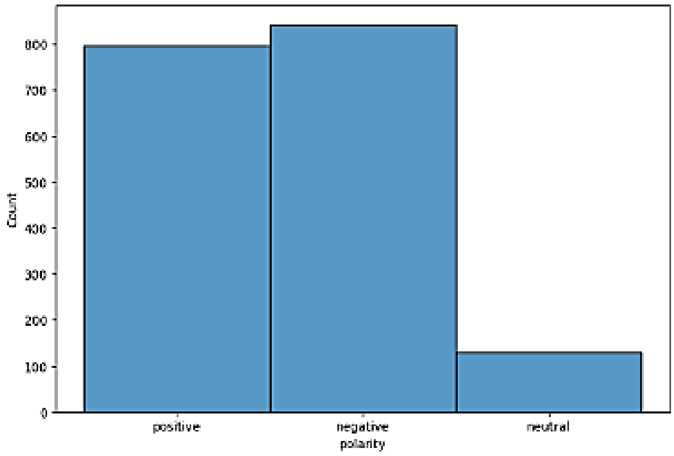

Figure 11. Before SMOTE Sastrawi

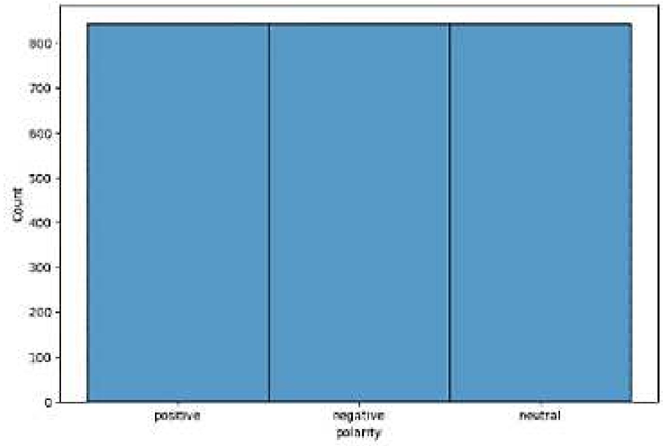

Figure 12. After SMOTE Sastrawi

By labeling the data with the steaming method, Sastrawi obtained 843 negative, 797 positive, and 130 neutral data in Figure 11. The SMOTE method was balanced to 843 positive, 843 positive, and 843 neutral data in Figure 12.

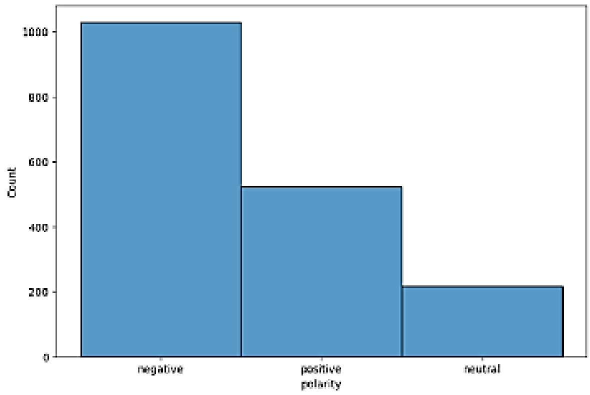

Figure 13. Before SMOTE Nazief & Adriani

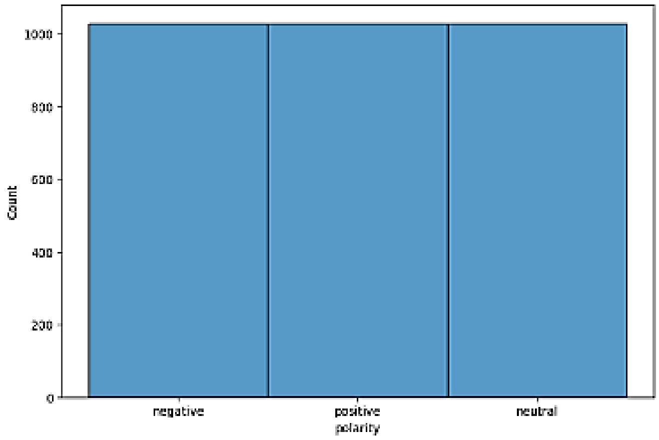

Figure 14. After SMOTE Nazief & Adriani

Labeling the data with the stemming Nazief and Andriani obtained 1,029 negative, 523 positive, and 218 neutral in Figure 13, after balancing with SMOTE to 1,029 positive, 1,029 positive, and 1,029 neutral in Figure 14.

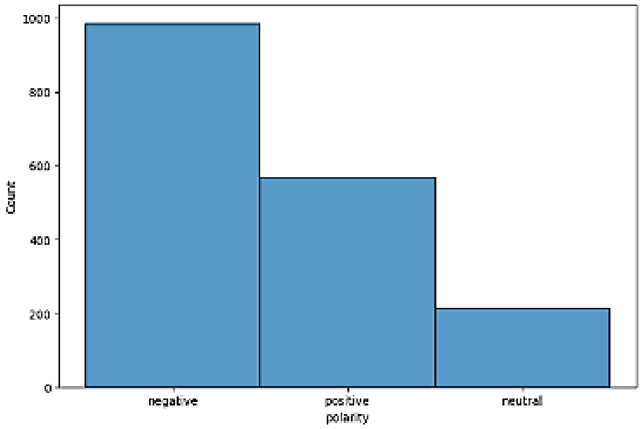

Figure 15. Before SMOTE, Arifin Setiono

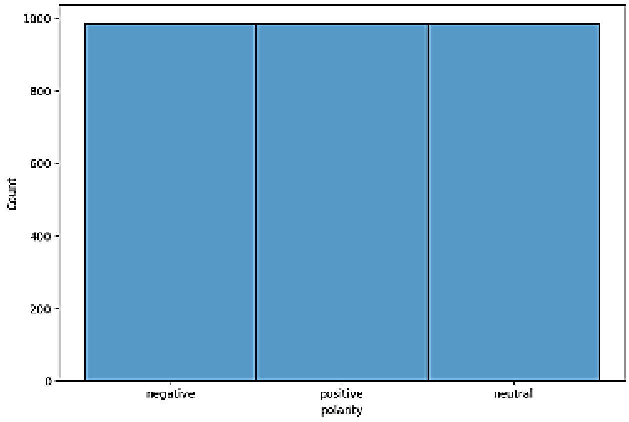

Figure 16. After SMOTE, Arifin Setiono

Labeling the data with Arifin Setiono obtained 987 negative, 569 positive, and 214 neutral in Figure 15, after SMOTE balancing to 987 positive, 987 positive, and 987 neutral in Figure 16.

F. Model Analysis

After the data are processed, labeled, and TF-IDF performed, the next step is to calculate the accuracy of the data using three algorithms, namely Naïve Bayes Classifier, Support Vector Machine, and Decision Tree, using data from the three types of stemming and random splitting.

TABLE III
ACCURACY SASTRAWI

Random Split		**Accuracy**		
Training	**Testing**	**NBC**	**SVM**	**DT**
90	10	77 %	83 %	61 %
80	20	76 %	82 %	63 %
70	30	75 %	82 %	62 %
60	40	75 %	80 %	68 %

DOI : https://doi.org/10.25139/inform.v9i1.7484

In Table III, the stemming result data using Sastrawi amount of 2,529 after the execution of SMOTE obtained the highest result on random split 90:10 or 90% data training and 10% data testing using support vector machine algorithm of 0.83%, until using random split 60:40 the maximum accuracy result is still using the support vector machine.

TABLE IV
ACCURACY NAZIEF & ANDRIANI

Random Split		Accuracy		
Training	**Testing**	**NBC**	**SVM**	**DT**
90	10	81 %	88 %	72 %
80	20	79 %	86 %	70 %
70	30	79 %	85 %	69 %
60	40	77 %	85 %	70 %

On stemming data using Nazief and Andriani in Table IV, a total of 3,087 after SMOTE, the highest results were obtained on random split 90:10 or 90% data training and 10% data testing using Support Vector Machine algorithm of 0.88% until using random split 60:40 the maximum accuracy result was still using Support Vector Machine algorithm.

TABLE V
ACCURACY ARIFIN SETIONO

Random Split		Accuracy		
Training	**Training**	**NBC**	**SVM**	**DT**
90	10	82 %	91 %	77 %
80	20	81 %	87 %	73 %
70	30	80 %	85 %	67 %
60	40	79 %	83 %	63 %

In Table V, the data resulting from stemming using Arifin Setiono a total of 2961 after SMOTE, the highest results were obtained at a random split of 90: 10 or 90% of training data and 10% of testing data using the Support Vector Machine algorithm of 0.91%, until using a random split of 60: 40 the highest accuracy results are still using the Support Vector Machine algorithm. Compared to other stemming, stemming using Arifin Setiono, the difference between negative, positive, and neutral data is almost balanced, so that when oversampling is done using SMOTE, the results are still good, and when calculated using the SVM algorithm get maximum accuracy results.

G. Evaluation

After being tested, the models go through an evaluation process to obtain accuracy, precision, recall, and F1 score values. The results are in Table V.

TABLE V
MODEL EVALUATION

	Precision	Recall	F1_Score	Support
Negative	88 %	86 %	87 %	103
Neutral	94 %	97 %	96 %	99
Positive	90 %	89 %	90 %	95
Accuracy			91 %	297
Macro Avg	91 %	91 %	91 %	297
Weighted Avg	91 %	91 %	91 %	297

IV. CONCLUSION

In this study, the data shows many negative word classifications, and the word that appears most often is *MBKM* because, with this *MBKM,* students are reluctant to join organizations. Then the results of the comparison of the Naïve Bayes Classifier, Support Vector Machine, and Decision Tree algorithms with data that has been done by SOMTE using split random obtained the results of their respective accuracies for Naïve Bayes Classifier, namely 75% - 82% then decision tree with an accuracy of 61% - 73% and Support Vector Machine with an accuracy of 80% - 91% so that it can be concluded that it is true according to several researchers who have explained in the introduction, namely Support Vector Machine is the best algorithm between the two algorithms. The best Stemming used is Arifin Setiono, who managed to get 91% accuracy on SVM. However, it would be nice if the data is balanced. If it is not balanced, the accuracy of the results can go down. Future research can also use stratified sampling, k-fold, or hyperparameter tunning to get results with even better accuracy.

REFERENCES

[1] "Peran dan Inovasi Generasi Milenial dalam Mewujudkan Indonesia Emas 2045.pdf."

[2] A. A. Fauzi and T. Pahlevi, "Analisis Hubungan Keaktifan Berorganisasi Terhadap Hasil Prestasi Akademik Mahasiswa Fakultas Ekonomi Universitas Negeri Surabaya," *J. Pendidik. Adm. Perkantoran JPAP*, vol. 8, no. 3, pp. 449–457, Jul. 2020, doi: 10.26740/jpap.v8n3.p449-457.

[3] F. Fauziannor, "Faktor-faktor yang mempengaruhi minat mahasiswa dalam berorganisasi di kampus STIE Pancasetia," *Fair Value J. Ilm. Akunt. Dan Keuang.*, vol. 4, no. 8, pp. 3520–3533, Mar. 2022, doi: 10.32670/fairvalue.v4i8.1455.

[4] "Hidayah et al. - 2022 - Does reviving organizations serve an advantage for.pdf."

[5] H. Abdullah, F. Aziz, B. Firmansyah, K. Nabilah, and M. R. Adhani, "PENGARUH ORGANISASI MAHASISWA PENDIDIKAN PARIWISATA TERHADAP PRESTASI BELAJAR PADA ERA MERDEKA BELAJAR KAMPUS MERDEKA," vol. 6, no. 1, 2023.

[6] "Program MBKM Tinggi Peminat, Bagaimana Dampak Regenerasi Organisasi?" Accessed: November 27, 2023. [Online]. Available: https://www.cahunsoed.com/2022/09/program-mbkm-tinggi-peminat-bagaimana.html

[7] F. S. Mufidah, S. Winarno, F. Alzami, E. D. Udayanti, and R. R. Sani, "Analisis Sentimen Masyarakat Terhadap Layanan Shopeefood Melalui Media Sosial Twitter Dengan Algoritma Naïve Bayes Classifier," *JOINS J. Inf. Syst.*, vol. 7, no. 1, pp. 14–25, May 2022, doi: 10.33633/joins.v7i1.5883.

[8] Y. A. Singgalen, "Analisis Performa Algoritma NBC, DT, SVM dalam Klasifikasi Data Ulasan Pengunjung Candi Borobudur Berbasis CRISP-DM," *Build. Inform. Technol. Sci. BITS*, vol. 4, no. 3, Dec. 2022, doi: 10.47065/bits.v4i3.2766.

[9] A. Rozaq, Y. Yunitasari, K. Sussolaikah, E. R. N. Sari, and R. I. Syahputra, "Analisis Sentimen Terhadap Implementasi Program Merdeka Belajar Kampus Merdeka Menggunakan Naïve Bayes, K-Nearest Neighboars Dan

Decision Tree," *J. MEDIA Inform. BUDIDARMA*, vol. 6, no. 2, p. 746, Apr. 2022, doi: 10.30865/mib.v6i2.3554.

[10] L. A. Pramesti and N. Pratiwi, "Analisis Sentimen Twitter Terhadap Program MBKM Menggunakan Decision Tree dan Support Vector Machine," vol. 4, no. 4, 2023.

[11] A. Rozaq, Y. Yunitasari, K. Sussolaikah, and E. R. N. Sari, "Sentiment Analysis of Kampus Mengajar 2 Toward the Implementation of Merdeka Belajar Kampus Merdeka Using Naïve Bayes and Euclidean Distence Methods," *Int. J. Adv. Data Inf. Syst.*, vol. 3, no. 1, Jun. 2022, doi: 10.25008/ijadis.v3i1.1233.

[12] M. Nashrullah and D. A. Efrilianda, "Sentiment Analysis of Kampus Merdeka Policy on Twitter Using Support Vector Machine and Naïve Bayes Classifier".

[13] M. Hermansyah, M. F. Firdausi, A. Wahid, and N. A. Prasetyo, "Twitter Sentiment Analysis for Exploring Public Opinion on the Merdeka Belajar-Kampus Merdeka (MBKM) 2023 with the Naïve Bayes Classifier Algorithm".

[14] "New Study Shows Twitter is the Most Used Social Media Platform Among Journalists," Social Media Today. Accessed: December 29, 2023. [Online]. Available: https://www.socialmediatoday.com/news/new-study-shows-twitter-is-the-most-used-social-media-platform-among-journa/626245/

[15] P. Suciu, "YouTube Remains The Most Dominant Social Media Platform," Forbes. Accessed: December 29, 2023. [Online]. Available: https://www.forbes.com/sites/petersuciu/2021/04/07/youtube-remains-the-most-dominant-social-media-platform/

[16] S. Riadi, E. Utami, and A. Yaqin, "Comparison of NB and SVM in Sentiment Analysis of Cyberbullying using Feature Selection," *sinkron*, vol. 8, no. 4, pp. 2414–2424, Oct. 2023, doi: 10.33395/sinkron.v8i4.12629.

[17] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," in *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, Moradabad, India: IEEE, Nov. 2019, pp. 266–270. doi: 10.1109/SMART46866.2019.9117512.

[18] R. Kusumawati, A. D'arofah, and P. A. Pramana, "Comparison Performance of Naive Bayes Classifier and Support Vector Machine Algorithm for Twitter's Classification of Tokopedia Services," *J. Phys. Conf. Ser.*, vol. 1320, no. 1, p. 012016, Oct. 2019, doi: 10.1088/1742-6596/1320/1/012016.

[19] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Semarang, Indonesia: IEEE, Oct. 2019, pp. 1–5. doi: 10.1109/ICIC47613.2019.8985884.

[20] L. Ardiani, H. Sujaini, and T. Tursina, "Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak," *J. Sist. Dan Teknol. Inf. Justin*, vol. 8, no. 2, p. 183, Apr. 2020, doi: 10.26418/justin.v8i2.36776.

[21] "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation.pdf."

[22] D. Soyusiawaty, A. H. S. Jones, and N. L. Lestariw, "The Stemming Application on Affixed Javanese Words by using Nazief and Adriani Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 771, no. 1, p. 012026, Mar. 2020, doi: 10.1088/1757-899X/771/1/012026.

[23] D. Mustikasari, I. Widaningrum, R. Arifin, and W. H. E. Putri, "Comparison of Effectiveness of Stemming Algorithms in Indonesian Documents:," presented at the 2nd Borobudur International Symposium on Science and Technology (BIS-STE 2020), Magelang, Indonesia, 2021. doi: 10.2991/aer.k.210810.025.

[24] H. Liu, X. Chen, and X. Liu, "A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis," *IEEE Access*, vol. 10, pp. 32280–32289, 2022, doi: 10.1109/ACCESS.2022.3160172.

[25] R. Wati, S. Ernawati, and H. Rachmi, "Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH," *J. Manaj. Inform. JAMIKA*, vol. 13, no. 1, pp. 84–93, Apr. 2023, doi: 10.34010/jamika.v13i1.9424.

[26] Imamah and F. H. Rachman, "Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regresion," in *2020 6th Information Technology International Seminar (ITIS)*, Surabaya, Indonesia: IEEE, Oct. 2020, pp. 238–242. doi: 10.1109/ITIS50118.2020.9320958.

[27] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6390–6404, Sep. 2023, doi: 10.1109/TNNLS.2021.3136503.

[28] "Raghuwanshi and Shukla - 2021 - Classifying imbalanced data using SMOTE based clas.pdf."

[29] Y. A. Singgalen, "Analisis Sentimen Top 10 Traveler Ranked Hotel di Kota Makassar Menggunakan Algoritma Decision Tree dan Support Vector Machine," Agustus 2023.

[30] "A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification.pdf."

[31] A. Deolika, K. Kusrini, and E. T. Luthfi, "ANALISIS PEMBOBOTAN KATA PADA KLASIFIKASI TEXT MINING," *J. Teknol. Inf.*, vol. 3, no. 2, p. 179, Dec. 2019, doi: 10.36294/jurti.v3i2.1077.

[32] S. Y. Pangestu, Y. Astuti, and L. D. Farida, "ALGORITMA SUPPORT VECTOR MACHINE UNTUK KLASIFIKASI SIKAP POLITIK TERHADAP PARTAI POLITIK INDONESIA," vol. 3, no. 1, 2019.

[33] C. A. Sari, A. Sukmawati, R. P. Aprilli, P. S. Kayaningtias, and N. Yudistira, "PERBANDINGAN METODE NAÏVE BAYES, SUPPORT VECTOR MACHINE DAN DECISION TREE DALAM KLASIFIKASI KONSUMSI OBAT," 2022.

[34] D. Nurmadewi, M. Amaliah, H. Hanifah, U. B. Purwanti, M. S. Arum, and N. W. Kusuma, "Sentiment Analysis of Jokowi's Candidate Discourse in Three Periods using the Naïve Bayes Method," *SISTEMASI*, vol. 12, no. 1, p. 166, Jan. 2023, doi: 10.32520/stmsi.v12i1.2413.

[35] "Hermansyah et al. - Twitter Sentiment Analysis for Exploring Public Op.pdf."

[36] N. Sevani, A. Setiawan, F. Saputra, R. K. Sali, and O. Sunardi, "Medical Diagnosis System in Healthcare Industry: A Fuzzy Approach," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 852, no. 1, p. 012149, Jul. 2020, doi: 10.1088/1757-899X/852/1/012149.