Eduvest - Journal of Universal Studies
Volume 4 Number 10, October, 2024
p- ISSN 2775-3735- e-ISSN 2775-3727

CLASSIFICATION OF BANJARESE HULU AND KUALA DIALECTS IN BANJARESE PROSE TEXTS

Muslihul Aqqad[1], Nova Rijati[2]
[1,2] Dian Nuswantoro University, Indonesia
Email: muslih.aqqad@gmail.com, nova.rijati@dsn.dinus.ac.id

ABSTRACT

This research focuses on classifying the Hulu and Kuala Banjarese dialects in the prose text "Datu Kandangan and Datu Kartamina". These dialects represent linguistic variations resulting from geographical, social, and cultural differences among language communities, particularly in South Kalimantan, Indonesia. Language analysis methods such as Python Natural Language Toolkit (NLTK), NumPy, and Latent Dirichlet Allocation (LDA) Visualization (LyLDAvis) were employed to classify the dialects, involving data preprocessing steps like tokenization, punctuation removal, stop word normalization, and stemming. The research findings reveal the superiority of the "Naive Bayes" method over the "Boolean Query," achieving high accuracy in identifying positive examples and classifying texts into Upper and Lower Banjar dialects. The "Naive Bayes" method outperforms the "Boolean Query" with precision and recall values of 0.955563 and 0.956098, while the "Boolean Query" only reaches 0.021416 and 0.146341. This study makes a significant scholarly contribution to understanding language and cultural diversity in South Kalimantan, opening opportunities for further exploration in developing Natural Language Processing (NLP) technology for Indonesian regional languages.

KEYWORDS *Banjarese Dialect, Boolean Query, Classification, Naïve Bayes, Dialect Classification*

INTRODUCTION

Natural Language Processing (NLP) has become one of the most exciting research areas in recent years. NLP aims to understand human language and enable computers to interact with human language more effectively. In this study, we will discuss several important tools used in language analysis, namely Python Natural Language Toolkit (NLTK), NumPy, and LDA Visualisation (LyLDAvis), to

How to cite:	Muslihul Aqqad, Nova Rijati. (2024). Classification of Banjarese Hulu and Kuala Dialects in Banjarese Prose Texts. Journal Eduvest. *4*(10): 9676-9687
E-ISSN:	2775-3727
Published by:	https://greenpublisher.id/

support the process of classifying Hulu Banjar and Kuala Banjar dialects in the prose text “Datu Kandangan and Datu Kartamina”.

Dialects are language variations that emerge as a result of geographical, social, and cultural differences between language communities. In the context of Indonesian, Banjar dialect is one of the dialect variations used by people in South Kalimantan. In Banjar dialect itself, there are two significant main variants, namely Hulu Banjar and Kuala Banjar dialects. Studies of language dialects have important value in understanding the linguistic diversity and sociolinguistic aspects of a region. Through dialect classification, researchers can better identify the differences and characteristics of each dialect variant. With the development of technology, the use of methods and techniques in the field of Natural Language Processing (NLP) has opened up new opportunities in language analysis, including in dialect classification.

The purpose of this study is to classify Hulu Banjar and Kuala Banjar dialects in prose texts, especially in the work "Datu Kandangan and Datu Kartamina". This work was chosen as the object of research because it contains narratives and dialogues that reflect the use of Banjar dialect in a distinctive context. Through dialect classification in this text, it is hoped that a deeper insight into the use and proportion of each dialect in the work can be obtained. In addition, morphology is one of the branches of science in the field of language that is important to study. Morphology studies the structure of word formation, including how words are formed from basic words, as well as semantic shifts in language. In Banjarese, there are three ways to form words from basic words, namely through affixation or the addition of affixes, reduplication or repetition, and through composition or compounding (Hapip et al., 1981).

The process of affixation in Banjarese consists of several forms, such as prefixes (ba-, di-, maN-, sa-, ta-, and paN-), infixes (-al-, -ar-, -ul-, and -ur-), suffixes (-akan, -an, -i, -nya), and also confixes (ba-an, ka-an, and paN-an). Each form of affixation gives additional meaning to the basic word and reflects the richness and complexity of Banjarese. By understanding the process of morphology in Banjarese, it is hoped that this research can provide a more comprehensive contribution to understanding the Hulu Banjar and Kuala Banjar dialects in the prose text "Datu Kandangan and Datu Kartamina". The results of the dialect classification analysis and morphological understanding are expected to provide a significant scientific contribution to the development of linguistics, especially in understanding the linguistic diversity and culture in South Kalimantan.

In the research process, we will use several important tools that have been mentioned earlier, namely Python NLTK, NumPy, and LyLDAvis. Python NLTK will be used to perform text processing and language analysis in the early stages. NumPy will help in processing text data into vector or matrix form that is suitable for use in classification models. LyLDAvis will be used to visualize the results of the topic classification model (LDA) and help in understanding the representation of Hulu Banjar and Kuala Banjar dialect topics in the prose text "Datu Kandangan and Datu Kartamina". In dialect classification, we will apply two different approaches, namely boolean query and naive Bayes. The boolean query approach will be used as a simple classification method to identify text that contains keywords or phrases that are specifically related to Hulu Banjar or Kuala Banjar dialects. Meanwhile, the naive Bayes method will be used as a probabilistic classification method that utilizes Bayes' theorem with the assumption of independence of features in text.

By combining powerful language analysis tools such as Python NLTK, NumPy, and LyLDAvis, and applying boolean query and naive Bayes classification approaches, this research is expected to provide a deeper understanding of the use and proportion of Hulu Banjar and Kuala Banjar dialects in the literary work "Datu Kandangan and Datu Kartamina". The results of the research are expected to provide a significant scientific contribution to the understanding of linguistic diversity and sociolinguistic aspects in South Kalimantan, and potentially become the basis for further research in the field of NLP and language analysis.

RESEARCH METHODS

This study used a dataset of 428 reviews taken from the prose book “Datu Kandangan and Datu Kartamina”. The dataset was then divided into three classes: Hulu reviews, Kuala reviews, and neutral reviews. A total of 1,151 features were used in this study. The following are the stages of the research flow as described in the figure :

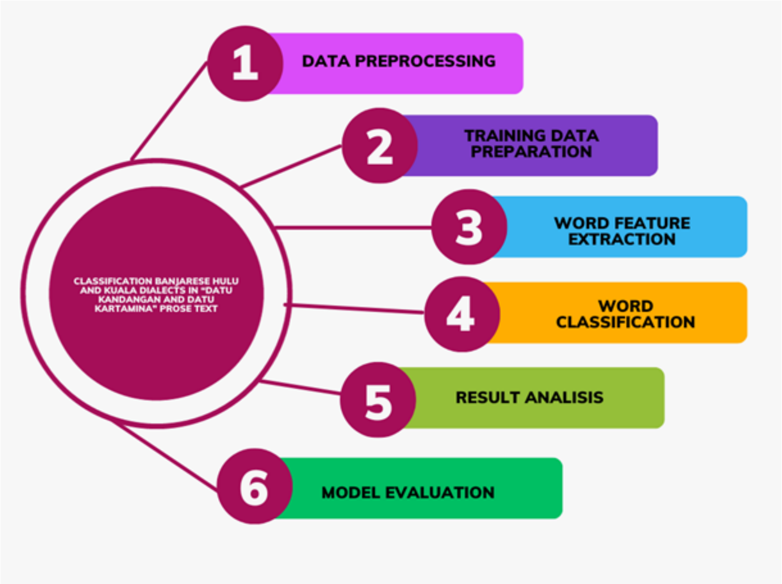

CLASSIFICATION BANJARESE HULU AND KUALA DIALECTS IN "DATU KANDANGAN AND DATU KARTAMINA" PROSE TEXT
1 DATA PREPROCESSING
2 TRAINING DATA PREPARATION
3 WORD FEATURE EXTRACTION
4 WORD CLASSIFICATION
5 RESULT ANALISIS
6 MODEL EVALUATION

Figure 1. Research Flow

Data Preprocessing

This stage aims to prepare the prose text "Datu Kandangan and Datu Kartamina" before the classification of Hulu Banjar and Kuala Banjar dialects is carried out. The pre-processing stage includes tokenization, punctuation removal, stop word normalization, and stemming. In the tokenization stage, the text is divided into token units such as words. Then, punctuation is removed to focus on the relevant words. After that, stop word normalization is carried out to remove common words that do not provide significant contributions in classification. The last stage is stemming, where words are extracted to their base forms.

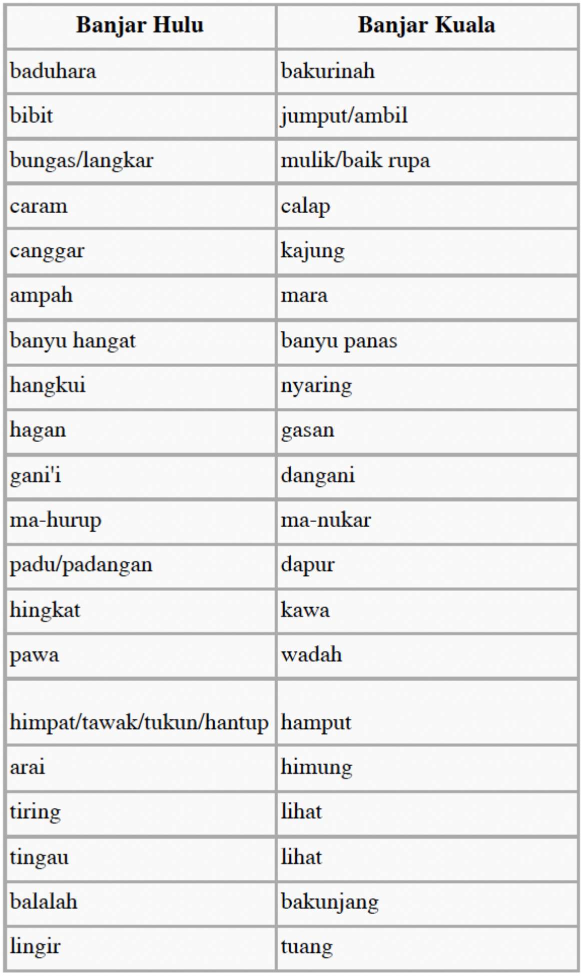
TF-IDF(t, d, D) = TF(t, d) * IDF(t, D)

a. TF(t, d) : The term frequency (TF) value of the word "t" in document "d". TF measures how often the word "t" appears in document "d".
b. IDF(t, D) : The inverse document frequency (IDF) value of the word "t" in the document collection "D". IDF measures how common or rare the word "t" appears in the entire document collection "D".
c. t : The word or term whose TF-IDF weight is to be calculated.
d. d : The document being analyzed.
e. D : The collection of all documents in the corpus or dataset.

Training Data Preparation

After the prose text is pre-processed, the training data is formed by selecting words that are contained in the Hulu Banjar and Kuala Banjar dialects. These words are obtained from an external file that contains a list of words that represent each dialect. Next, the training data is formed by associating these words with the appropriate dialect label based on the Hulu Banjar and Kuala Banjar dictionaries.

Table 1. Table of Banjar Hulu and Kuala Dialects



Banjar Hulu	Banjar Kuala
baduhara	bakurinah
bibit	jumput/ambil
bungas/langkar	mulik/baik rupa
caram	calap
canggar	kajung
ampah	mara
banyu hangat	banyu panas
hangkui	nyaring
hagan	gasan
gani'i	dangani
ma-hurup	ma-nukar
padu/padangan	dapur
hingkat	kawa
pawa	wadah
himpat/tawak/tukun/hantup	hamput
arai	himung
tiring	lihat
tingau	lihat
balalah	bakunjang
lingir	tuang

Word Feature Extraction

After the training data is formed, word feature extraction is performed. The frequency of word occurrence in the entire training data is calculated using the FreqDist function from NLTK. Words with high frequencies are considered to be relevant features. These features are then used in the formation of the training data in a format that is accepted by NLTK.

Word Classification

After word feature extraction is performed and the frequency of word occurrence is found in the entire training data, the Naive Bayes model is then used to classify the text and determine whether the text belongs to the Banjar Hulu dialect, the Banjar Kuala dialect, or neutral.

Query with Boolean Method

In the boolean query method, a function called boolean_query() is created that takes text and keywords as input. The text is checked to see if it contains all the keywords (keywords). If so, the text is classified as a Banjar Hulu or Kuala dialect based on the dictionary, otherwise, it is classified as a neutral dialect. The formula for boolean query is as follows:

P = First condition (term 1)
Q = Second condition (term 2)
AND = Logical operator AND (conjunction)
OR = Logical operator OR (disjunction)
NOT = Logical operator NOT (negation)

Using boolean logical operators, the formula for boolean query can be written as :

AND Query: P AND Q

Meaning, search for documents that contain both conditions (term 1 and term 2).

OR Query: P OR Q

Meaning, search for documents that contain one or both conditions (term 1 or term 2).

NOT Query: NOT P

Meaning, search for documents that do not contain a specific condition (term 1).

Combined Query: (P AND Q) OR (R AND NOT S)

Meaning, perform a combination of several conditions using the logical operators AND, OR, and NOT.

Query with Naive Bayes Method

In addition to the boolean query method, a query is also performed using the Naive Bayes method. A function called process_text_naive_bayes() is created that takes text as input. The text is processed by tokenizing, extracting word features, and using a Naive Bayes model to predict the dialect label. In the case of text classification, the Naive Bayes Classification formula can be written as follows :

$$P(C|F1, F2, ..., Fn) = (P(F1|C) * P(F2|C) * ... * P(Fn|C) * P(C)) / P(F1, F2, ..., Fn)$$

Explanation:

1) P(C|F1, F2, ..., Fn) is the posterior probability that text data X belongs to category C, given features F1, F2, ..., Fn in the text data X.

2) P(Fi|C) is the probability that feature i appears in category C. This probability is calculated from the training data by counting the frequency of occurrence of feature i in text data that falls into category C divided by the total text data that falls into category C.
3) P(C) is the prior probability of category C, which is the probability that text data randomly belongs to category C. This probability is also calculated from the training data by counting the number of text data that fall into category C divided by the total text data.
4) P(F1, F2, ..., Fn) is the prior probability of features F1, F2, ..., Fn in the text data X. In Naive Bayes, features are assumed to be independent, so this probability can be calculated as the product of the probabilities of each feature. In the case of text, this probability can be calculated as the product of the probabilities of the occurrence of each word in the text data X.

Result Analisis

To analyze the classification results, the prose text “Datu Kandangan wan Datu Kartamina” was used as a corpus. The code with python and the required libraries will be processed to classify each text in the corpus and store the classification results in a list. The classification results are then stored in a CSV file that contains the columns 'origin' (original text), 'normalized' (processed text), and 'class' (determined dialect class).

Model Evaluation

The predictions from both methods are evaluated by comparing them with the actual labels. The evaluation is carried out using metrics such as accuracy, precision, and recall. The evaluation results provide an overview of the performance of both methods in classifying Banjar Hulu and Banjar Kuala dialects in the prose text “Datu Kandangan wan Datu Kartamina”.

Precision = True Positives / (True Positives + False Positives)

1) True Positives (TP) is the number of positive data that are correctly predicted by the model.
2) False Positives (FP) is the number of negative data that are incorrectly predicted as positive by the model.
3) Accuracy = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)
4) True Positives (TP) is the number of positive data that are correctly predicted by the model.
5) True Negatives (TN) is the number of negative data that are correctly predicted by the model.
6) False Positives (FP) is the number of negative data that are incorrectly predicted as positive by the model.
7) False Negatives (FN) is the number of positive data that are incorrectly predicted as negative by the model.

In both of the above formulas, True Positives, True Negatives, False Positives, and False Negatives are values that are obtained from the results of evaluating a classification model by comparing the model's prediction results with the actual labels of the data. Precision measures how many of the data that are predicted positive by the model are actually positive, while accuracy measures how many of the entire data are predicted correctly by the model. The higher the values of precision and accuracy, the better the performance of the classification model in predicting data.

RESULT AND DISCUSSION

Data Pre-processing

After pre-processing the prose text "Datu Kandangan wan Datu Kartamina", feature extraction was performed from 428 corpus to perform classification of Banjar Hulu and Banjar Kuala dialects. The results of feature extraction revealed 1151 unique features consisting of various words or terms in the corpus. Next, an analysis was conducted on the words that appear most frequently in the corpus, known as most frequent words. The most frequent words identified are 'tuti', 'datu', 'urang', 'kampung', 'inya', 'imah', 'kumandan', 'imbah', 'kandangan', 'amun', 'banar', 'walanda', 'kartamina', 'hidin', and 'napang'. These words have a high frequency in the corpus and can provide an overview of the characteristics of the language and dialect that are often used in the prose text.

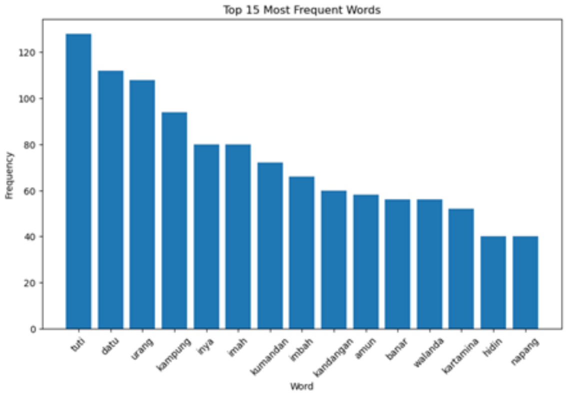


Figure 2. Frequent Words

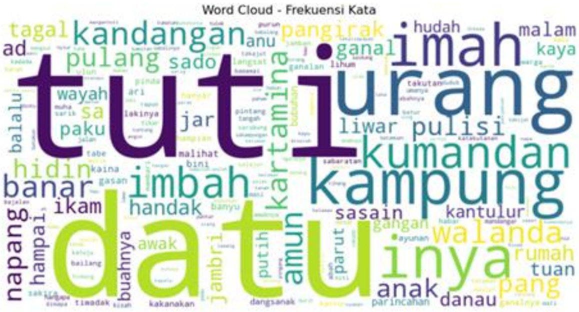
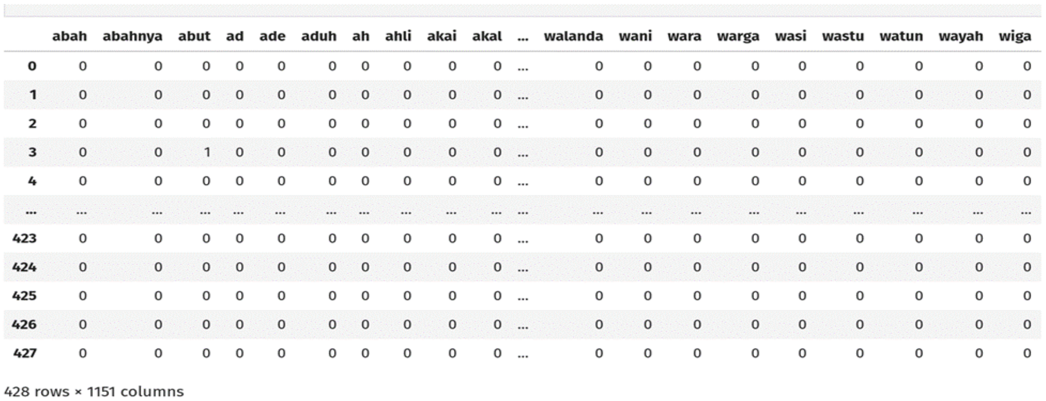


Figure 3. Word Frequency in wordcloud

Feature Extraction

A total of 1151 features were obtained from the results of feature extraction on 428 corpus. These features are a collection of relevant and diverse words that can be used in further analysis to classify Banjar Hulu and Banjar Kuala dialects.

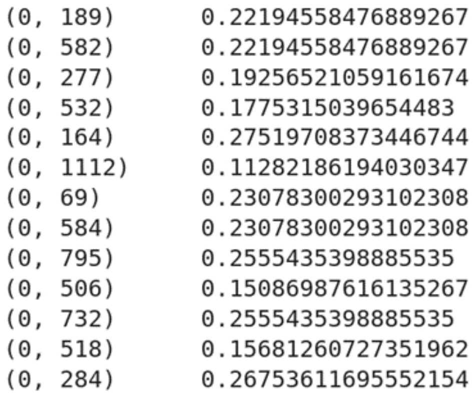


	abah	abahnya	abut	ad	ade	aduh	ah	ahli	akai	akal	...	walanda	wani	wara	warga	wasi	wastu	watun	wayah	wiga
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
423	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
424	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
425	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
426	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
427	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0

428 rows × 1151 columns

Figure 4. Document Feature Extraction Table

The feature extraction values show the occurrence of each feature in the documents of the corpus. Features with high extraction values indicate that the word appears more often in the prose text and can have an important contribution to the dialect classification process. This feature extraction information will be used to form the training data, select important features, and classify the dialects using the Naive Bayes method or the boolean query method. By utilizing the results of this feature extraction, researchers can identify the words that are most relevant in distinguishing between Banjar Hulu and Banjar Kuala dialects in the prose text "Datu Kandangan wan Datu Kartamina". The results of this analysis and classification are expected to provide deeper insights into the differences and characteristics of each dialect in the work.



```
(0, 189)      0.22194558476889267
(0, 582)      0.22194558476889267
(0, 277)      0.19256521059161674
(0, 532)      0.1775315039654483
(0, 164)      0.27519708373446744
(0, 1112)     0.11282186194030347
(0, 69)       0.23078300293102308
(0, 584)      0.23078300293102308
(0, 795)      0.2555435398885535
(0, 506)      0.15086987616135267
(0, 732)      0.2555435398885535
(0, 518)      0.15681260727351962
(0, 284)      0.26753611695552154
```

Figure 5. TF-IDF Results

The output results shown are a representation of the TF-IDF matrix. TF-IDF (Term Frequency-Inverse Document Frequency) is a common method used in text analysis to evaluate the importance of a word in a document or text corpus. The TF-IDF matrix represents each word in a document as a vector with a dimension that corresponds to the number of words in the corpus. Each entry in the matrix shows the TF-IDF weight of the word in the associated document.

In the given output example, the pair (i, j) shows the document index (i) and feature index (word) (j) in the TF-IDF matrix. The entry (i, j) shows the TF-IDF weight of the word indexed by j in the document indexed by i. For example, in the pair (0, 89), the value 0.22194558476889267 shows the TF-IDF weight of the word indexed by 89 in the document indexed by 0. This value reflects how important the word is in the document based on its frequency of occurrence in the document and the frequency of occurrence of the word in the entire text corpus.

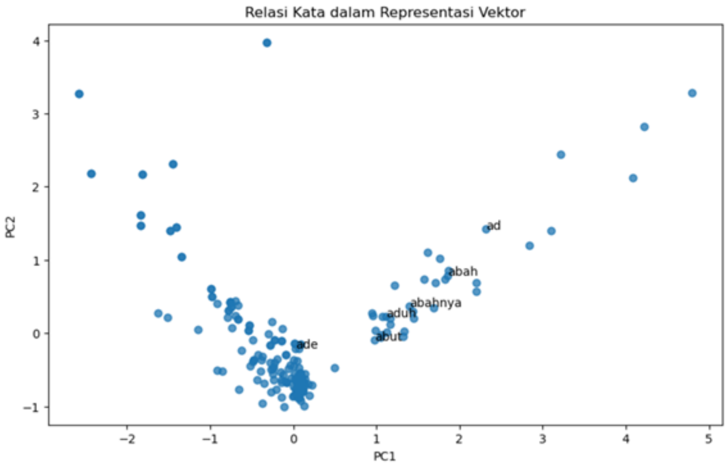


Figure 6. Word Relations in Vector Representation

The results shown are a visual representation of the relationships between words in vector space. Each row shows a specific word along with the vector coordinates that describe its position in the two-dimensional space, represented by the PC1 and PC2 axes.

For example, consider the first row:

The word “abah” has a PC1 vector coordinate of 1.8570833562196722 and a PC2 vector coordinate of 0.7987113961863019. This indicates that in the vector representation used, the word “abah” has a relative position with respect to the PC1 and PC2 axes of approximately (1.8570833562196722, 0.7987113961863019) in two-dimensional space. Similarly, the following rows show the vector coordinates of other words such as “abahnya”, “abut”, “ad”, and “ade”. This information provides a visual overview of how the relationships between these words are manifested in vector space. The distance and direction between the vector coordinates can be used to represent the degree of similarity or difference in meaning between these words in the representation used.

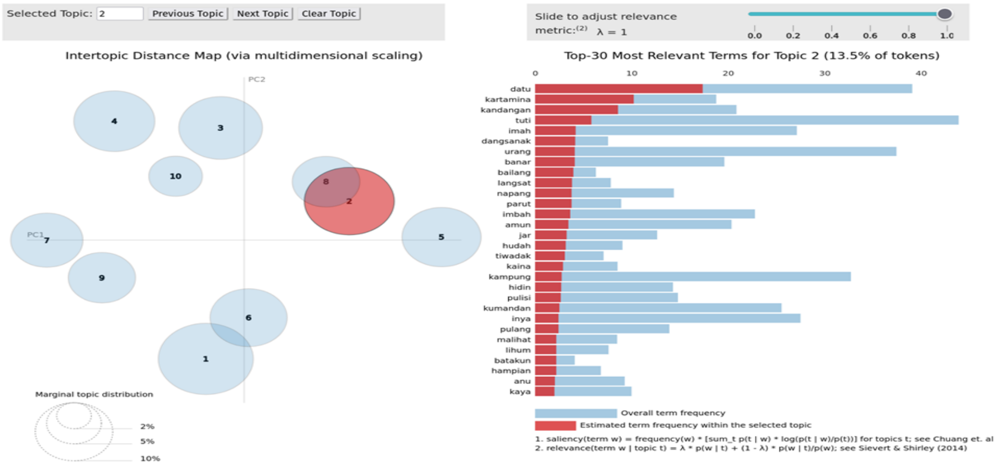
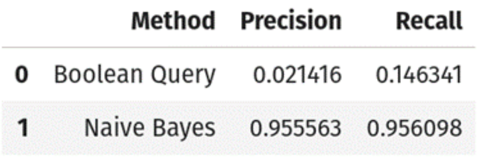

Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%
Slide to adjust relevance metric:(2) λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Top-30 Most Relevant Terms for Topic 2 (13.5% of tokens)
0
10
20
30
40
datu
kartamina
kandangan
tuti
imah
dangsanak
urang
banar
bailang
langsat
napang
parut
imbah
amun
jar
hudah
tiwadak
kaina
kampung
hidin
pulisi
kumandan
inya
pulang
malihat
lihum
batakun
hampian
anu
kaya
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 7. LDA Visualization

CONCLUSION

The results presented are a table showing the evaluation results for the two methods used, namely "Boolean Query" and "Naive Bayes". The table includes a column called "Method" that shows the name of the method, as well as columns called "Precision" and "Recall" that show the values of the associated evaluation metrics.



	Method	Precision	Recall
0	Boolean Query	0.021416	0.146341
1	Naive Bayes	0.955563	0.956098

Figure 8. Precision and Recall Table

Here is an explanation of each of the evaluation metrics listed in the table:

1. Precision : It is the ratio of the number of true positives (positives predicted correctly) divided by the total number of positive predictions (true positives + false positives). Precision measures how accurate the model is in identifying the positive class. The precision value ranges from 0 to 1, with a value of 1 indicating perfect precision. In this case, the "Boolean Query" method has a precision value of 0.021416, while the "Naive Bayes" method has a precision value of 0.955563. This shows that the "Naive Bayes" method has a much higher precision than the "Boolean Query" method.
2. Recall : Also known as sensitivity, recall measures how well the model does in identifying all true positive examples. Recall is calculated as the ratio of the number of true positives divided by the total number of positive examples (true positives + false negatives). The recall value also ranges from 0 to 1, with a value of 1 indicating perfect recall. In this case, the "Boolean Query" method has a recall value of 0.146341, while the "Naive Bayes" method has a recall value of 0.956098. This shows that the "Naive Bayes" method has a

much better ability to identify positive examples than the "Boolean Query" method.

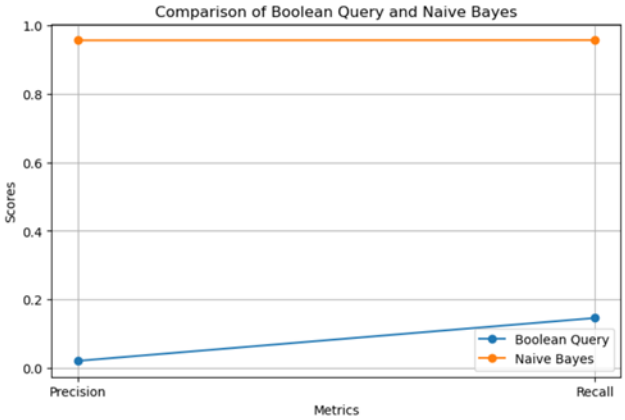


Figure 9. Comparison of Boolean Query and Naive Bayes

Therefore, based on the table, it can be concluded that the "Naive Bayes" method provides much better results in terms of precision and recall than the "Boolean Query" method.

REFERENCES

Tuhenay, D., & Mailoa, E. (2021). Perbandingan Klasifikasi Bahasa Menggunakan Metode Naïve Bayes Classifier (NBC) dan Support Vector Machine (SVM). Fakultas Teknologi Informasi, Program Studi Teknik Informatika, Universitas Kristen Satya Wacana.

Setiawan, A., Santoso, L. W., & Adipranata, R. (2020). Klasifikasi Artikel Berita Bahasa Indonesia Dengan Naive Bayes Classifier. Program Studi Informatika, Fakultas Teknologi Industri, Universitas Kristen Petra.

Hapip, A. D. (1977). Kamus Banjar-Indonesia. (Departemen Pendidikan dan Kebudayaan, Ed.) (1st ed.). Jakarta: Pusat Pembinaan dan Pengembangan Bahasa. Retrieved from http://repositori.kemdikbud.go.id

Hapip, A. D., Kawi, D., & Noor, B. (1981). Struktur Bahasa Banjar Kuala (Seri bb 71). Jakarta: Pusat Pembinaan dan Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan. Retrieved from http://repositori.kemdikbud.go.id/id/eprint/240

Hapsari, R. K., & Juli, Y. (2015). Stemming Artikel Berbahasa Indonesia Dengan Pendekatan Confix-Stripping. Prosiding Seminar Nasional Manajemen Teknologi XXII, 1–8.

Humaidi, A., Kamariah, & Harpriyanti, H. (2017). Infleksi dalam Bahasa Banjar. Jurnal Bahasa, Sastra, Dan Pengajarannya, 2(2), 262–272.

Kamus Bahasa Banjar dialek Hulu-Indonesia. Pusat Pembinaan dan Pengembangan Bahasa. Banjarmasin: Balai Bahasa Banjarmasin, 2008. Deskripsi Fisik: xv, 311 hlm.; 21 cm. ISBN: 978-979-685-776-0.

Sugiyono, & Maryani, Y. (2008). Kamus Bahasa Indonesia. Jakarta: Pusat Bahasa Departemen Pendidikan Nasional.

Sugono, D. (2017). Bahasa dan Peta Bahasa di Indonesia (1st ed.). Jakarta: Kementrian Pendidikan dan Kebudayaan. Retrieved from http://repositori.kemdikbud.go.id/7191/1/

Suseno, Y. S. A. (2012). Jendela Sastra Media Sastra Indonesia. Retrieved April 3, 2019, from http://www.jendelasastra.com/karya/prosa/datu-kandangan-wan-datu-kartamina