SENTIMENT ANALYSIS OF HOAX NEWS JOB VACANCY WITH NAIVE BAYES METHOD

[1st] Dwi Sartika Simatupang
Informatics Engineering
Nusa Putra University
Sukabumi, Indonesia
dwi.simatupang@nusaputra.ac.id

[2nd] Siti Nursinta
Informatics Engineering
Nusa Putra University
Sukabumi, Indonesia
siti.nursinta_ti20@nusaputra.ac.id

[3rd] Andres Gunawan
Informatics Engineering
Nusa Putra University
Sukabumi, Indonesia
andres.gunawan_ti20@nusaputra.ac.id

[4th] Yance Wainauw
Informatics Engineering
Nusa Putra University
Sukabumi, Indonesia
yance.wainauw_ti20@nusaputra.ac.id

Abstract— Hoax news or fake news has become a significant problem in today's digital era, the news in question is information on job vacancies on social media, one of which is Twitter, which is widely discussed by job seekers. The spread of hoax news related to job vacancies can cause harm to job seekers who rely on this information to find work. Therefore, sentiment analysis of hoax job vacancies is important to help users distinguish between valid and hoax information. This study aims to analyze sentiment towards hoax news related to job vacancies on Twitter. The method used is the Naive Bayes method, which is a classification method commonly used in sentiment analysis. Hoax news dataset collected from job vacancy information will be processed and features that are relevant to job vacancies will be identified. Furthermore, the Naive Bayes sentiment analysis model will be developed to classify hoax news sentiment as positive or negative. The results of sentiment analysis will help users identify and avoid hoax news related to job vacancies. In addition, the results of sentiment analysis also improve the quality of the content displayed and strengthen user confidence in the information provided. It is hoped that this research will contribute to overcoming the problem of hoax news on Twitter, improve the quality of available information, and assist job seekers in making better decisions in finding work.

Keywords: sentiment analysis, hoax news, twitter, job vacancies, naive bayes.

I. INTRODUCTION

The use of social media has become a personal need for the community, one of which is Twitter which is a source of information for job seekers. The amount of information obtained quickly becomes the basis for spreading fake news or hoax carried out by irresponsible persons [1]. One of them is the spread of hoax news related to job vacancies. The impact of hoax news itself can change people's perceptions to become misguided [2]. The spread of hoax news is very detrimental to society because many parties feel disadvantaged by the spread of hoax news [3].

This can be seen from the large amount of inappropriate job vacancy information spread on Twitter social media.

The jobs currently provided are informed through information technology [4]. Twitter is one of the information media used by the public in finding work. However, some people do not know that much of this information is inconsistent with its authenticity.

Sentiment analysis or sentiment analysis is the process of processing text data automatically, to understand information about opinions or responses [5]. Sentiment analysis allows users to evaluate and understand the sentiments contained in hoax news. With this understanding, users can distinguish between fake news and valid information, so that they can make more informed and accurate decisions when looking for work. To assist in sentiment analysis, it is necessary to have a method for determining valid data or information based on the presentation of negative values and positive values.

Sentiment analysis is used to find the information needed from unstructured data, so it is hoped that in this study public sentiment can be identified regarding hoax news on Twitter [6]. In addition, sentiment analysis can help improve the quality of the content displayed. By understanding the sentiments that arise related to hoax news, managers can identify and remove content that is invalid or detrimental to applicants. This will increase user confidence in job vacancy information on Twitter and increase credibility as a reliable source of job information.

However, sentiment analysis of hoax news on job vacancies has its own challenges because irresponsible individuals often use manipulative and emotional strategies. Therefore, sentiment analysis is needed to produce more valid and accurate information.

Based on the above problems, the authors try to analyze hoax news about job vacancies on Twitter using the Naive Bayes method. The Naive Bayes algorithm is a classification technique that utilizes probability and statistical methods. This algorithm is one of the popular classification methods and is included in the top ten algorithms in data mining according to the IEEE International Conference on Data Mining (ICDM'06) in Hong Kong [7]. Naive Bayes is a branch of mathematics known as probability theory to find the greatest chance of a possible classification by looking at the frequency of each classification in the training data [8].

II. METHODOLOGY

The method used in this study is the naive Bayes method. The Naïve Bayes method is a classification method based on the Bayes theorem [9] with the assumption that all the features used are independent of one another [10].

The process in this method is as follows.

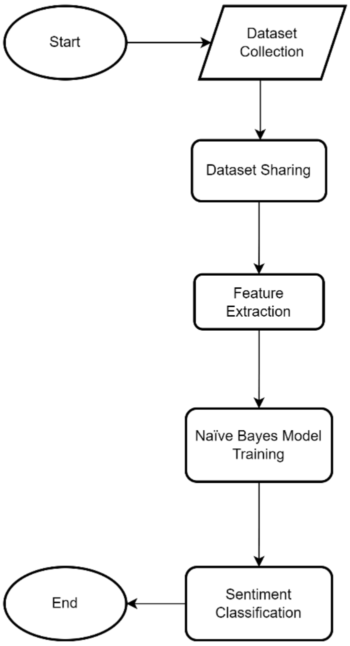

Figure 1. Methodology Flowchart

A. Dataset Collection

Collection of hoax news datasets related to job vacancies from social media or other relevant sources. This dataset includes various examples of hoax news with different sentiments, such as positive or negative.

B. Distribution of Datasets

Dividing the dataset into two subsets: the training set used to train the Naïve Bayes model and the testing set to test the performance of the trained model.

C. Feature Extraction

Feature extraction is performed to convert text into a numerical representation that can be used by the model.

D. Naïve Bayes Model Training

Naive Bayes model training using training subsets. The Naive Bayes model learns sentiment patterns from features and sentiment labels in the training data. The model will calculate the probability of sentiment (positive or negative) based on the existing features.

E. Sentiment Classification

Use a trained model to classify sentiment on a test subset. The model will calculate the sentiment probability for each hoax news text and assign a sentiment label based on the highest probability.

III. RESULTS AND DISCUSSION

A. Preprocessing Stage

1. Cleansing

Cleansing is the process of cleaning unnecessary words. Examples include username(@), email, hashtag(#), and URL.

Table 1. Cleansing Process

Training Data	
Input	**Output**
@sarifudhinz Hati-hati gaes.. ini nomernya barusan sy di wa +91 6291 369 236	Hati-hati gaes. Ini nomernya barusan sy di wa +91 6291 369 236

2. Case Folding

Case Folding is the process of changing words into the same form, either uppercase or lowercase.

3. Tokenizing

Tokenizing is the process of identifying words in a text that are cut off by spaces or special characters.

Table 2. Tokenizing Process

Training Data	
Input	**Output**
Harap berhati-hati terhadap pihak yang mengatasnamakan Pertamina. Tetap semangat dan sukses selalu ya sob	[Harap, berhati-hati, terhadap, pihak, yang,mengatasnamakan, Pertamina, tetap semangat, dan sukses, selalu, ya]

4. Stopword Removal

Stopword removal is the process of removing words that often appear and are general in nature. Such as the use of conjunctions: and, which, yes, and others.

5. Stemming

Stemming is the process of making affixed words (affixes me-, meng-, pe-, peng-, etc.) to become root words.

Table 3. Stemming Process

Training Data	
Input	**Output**
membuat	buat
mengatasnamakan	atas nama
pelatihan	latih
penglihatan	lihat

B. Naive Bayes Algorithm

1. Data Source

The data used in this study is in the form of comments taken directly from Twitter. Here's an example used

Table 4. Sample Data Used

Comment
@Linda66663301 Saya sudah kena tipu...
@r1en3goo Jangan sampai tertipu yaa
@pvrivasi Wihh mesti hati2 nih
@nindi_kirani Pintar[2] dlm memilih informasi ya guys

2. Preprocessing of Implementation

Before classifying naive Bayes, the data will be processed first at this stage so that at the classification stage the results will be more optimal.

Table 5. Preprocessing Result Data

Comment
[Saya, sudah, kena, tipu]
[Jangan, sampai, tertipu, yaa]
[Bulan_maret, mulai_beraksi, penipu_nya]
[Wihh, mesti, hati-hati]
[Pintar-pintar, dalam_memilih, informasi]

C. Application of Naive Bayes

This stage is a classification process based on existing sentiments. This stage includes two processes as follows.

1. Learning Naive Bayes Classifier (NBC) Process

There are three steps in the NBC learning process, namely:

- Shaping Features

The features in question are keywords that become the parameters of the training data unit, namely comments that will be classified into predetermined sentiment classes (positive and negative).

Table 6. Formation of Training Data Features

Data	Feature	Sentiment
D1	kena(1), tipu(1)	negatif
D2	tertipu(1)	negatif
D3	beraksi(1), penipu(1)	negatif
D4	hati-hati(1)	positif
D5	pintar-pintar(1), memilih_ informasi(1)	positif

- Calculating Probability

After forming the features from the appearance of the training data, then calculate the probability of each class in the following way:

$$p(Ci) = \frac{fd(Ci)}{|D|}$$

Keterangan :

$fd(C_i)$ = Jumlah dokumen yang termasuk ci
$|D|$ = Jumlah data latih / jumlah komentar

Table 7. Probability Training Naive Bayes

Sentiment Class	Data					fd	p
	D1	D2	D3	D4	D5		
Negatif	1	1	1	0	0	3	3/5
Positif	0	0	0	1	1	2	2/5

- Determine Probabilities

After getting the probability, then calculate the features of each class in the following way :

$$p(Wk|Ci) = \frac{f(Wki,\ Ci) + 1}{f(Ci) + |W|}$$

Keterangan :

$f(w_{ki}, c_i)$ = Nilai kemunculan kata w_{ki} pada kelas c_i
$f(c_i)$ = Jumlah keseluruhan kemunculan kata pada kelas c_i

$|W|$ = *Jumlah keseluruhan n dari Wk*

Table 8. Trained Data Probability Models

Data f(Wki,Ci)	Sentiment Class	
	Negative	Positif
kena	$\frac{1+1}{3+8} = \frac{2}{11}$	$\frac{0+1}{2+7} = \frac{1}{9}$
penipu	$\frac{1+1}{3+8} = \frac{2}{11}$	$\frac{0+1}{2+7} = \frac{1}{9}$
beraksi	$\frac{1+1}{3+8} = \frac{2}{11}$	$\frac{0+1}{2+7} = \frac{1}{9}$
hati-hati	$\frac{0+1}{3+8} = \frac{1}{11}$	$\frac{1+1}{2+7} = \frac{2}{9}$
informasi	$\frac{0+1}{3+8} = \frac{1}{11}$	$\frac{1+1}{2+7} = \frac{2}{9}$

2. The Naïve Bayes Classifier (NBC) Classification Process

The flow of the naive Bayes classifier classification process used is as follows.

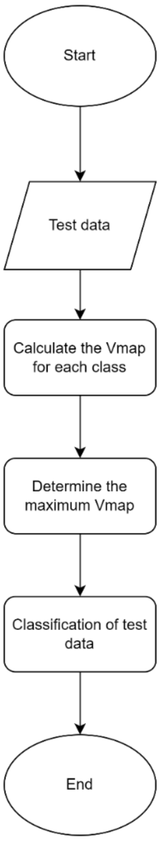
Start
Test data
Calculate the Vmap for each class
Determine the maximum Vmap
Classification of test data
End

Figure 2. Naive Bayes classification process flowchart

An example of one of the comments used as test data using a probability model.

Table 9. Classification Test Data

Comment	
Before Preprocessing	**After Preprocessing**
Harap berhati-hati terhadap pihak yang mengatasnamakan Pertamina. Tetap semangat dan sukses selalu ya	berhati-hati pihak mengatasnamakan semangat_sukses

After determining the test data, the next steps are :

- Calculate Vmaps

Vmap is a calculation used by the Naive Bayes classifier to determine the probability of each class based on the learning process. The highest value will be chosen as the probability value.

Based on the results of the training, the following is the calculation:

$$\text{Vmap} = \underset{\{negatif,positif\}}{argmax} \quad p(wk|c)x\,p(c)$$

$$\text{Vmap} = \underset{\{negatif,positif\}}{argmax} \quad p(wk|c)x\,p(c)$$

P(”berhati-hati”|ci) p(”pihak”|ci) p(”mengatasnamakan”|ci) p(”semangat”|ci) p(”sukses”|ci).

Vmap negative sentiment

Vmap (”negative”) =

P(”berhati-hati”|negatif) p(”pihak”|negatif) p(”mengatasnamakan”|negatif) p(”semangat”|negatif) p(”sukses”|negatif).

$$= \frac{2}{5} \times \frac{2}{11} \times \frac{2}{11} \times \frac{2}{11} \times \frac{1}{11} \times \frac{1}{11} = 0{,}018$$

Vmap positive sentiment

Vmap (”positive”) =

P(”berhati-hati”|positif) p(”pihak”|positif) p(”mengatasnamakan”|positif) p(”semangat”|positif) p(”sukses”|positif).

$$= \frac{2}{5} \times \frac{1}{9} \times \frac{1}{9} \times \frac{1}{9} \times \frac{2}{9} \times \frac{2}{9} = 0{,}012$$

- Determine Maximum Vmap

From the results of the Vmap calculation above, it can be seen that the negative Vmap value is greater than the positive Vmap value. So, it can be concluded that these comments are classified into the negative sentiment class.

IV. CLOSING

A. Conclusion

Sentiment analysis of fake job vacancies using the Naive Bayes method can make a significant contribution in identifying and distinguishing fake news from valid information. By understanding the sentiments contained in the hoax news, applicants can make more informed and accurate decisions in finding jobs. However, it is important to view the results of the analysis in the proper context and consider other factors that can affect the quality and accuracy of sentiment analysis.

B. Suggestion

There are several things that need to be added in the development of sentiment analysis on hoax job vacancies using the Naive Bayes method, namely the following.

1. Developing a larger and more representative dataset.

2. Using other methods to improve the accuracy and reliability of settlement analysis.

3. Pay attention to the context and novelty of information.

REFERENCE

[1] Sinta Peringkat, T., Dirjen Penguatan RisBang Kemenristekdikti, berdasarkan S., & Wati, R. (n.d.). *PENERAPAN ALGORITMA NAIVE BAYES DAN PARTICLE SWARM OPTIMIZATION UNTUK KLASIFIKASI BERITA HOAX PADA MEDIA SOSIAL*. www.bsi.ac.id

[2] Alpian, D., Krisna, N., & Salamah, U. (2022). PERBANDINGAN ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR UNTUK KLASIFIKASI BERITA HOAX KESEHATAN DI MEDIA SOSIAL TWITTER. *Jurnal Teknik Informatika Kaputama (JTIK)*, *6*(2).

[3] Sriyano, C. S., & Setiawan, E. B. (n.d.). *Pendeteksian Berita Hoax Menggunakan Naive Bayes Multinomial Pada Twitter dengan Fitur Pembobotan TF-IDF*.

[4] Taqwiym, A., Wijaya, N., Akuntansi, K., & Gi Mdp, S. (n.d.). *PERANCANGAN LOWONGAN KERJA ONLINE BERBASIS WEB PADA PT ANH*.

[5] Februariyanti, H., Firmansyah, M., Wibowo, J. S., & Utomo, M. S. (n.d.). *Terakreditasi "Peringkat 4 (Sinta 4)" oleh Kemenristekdikti*. *6*(2), 1–5. https://doi.org/10.5281/zenodo.4399381

[6] Giovani, A. P., Ardiansyah, A., Haryanti, T., Kurniawati, L., & Gata, W. (2020). ANALISIS SENTIMEN APLIKASI RUANG GURU DI TWITTER MENGGUNAKAN ALGORITMA KLASIFIKASI. *Jurnal Teknoinfo*, *14*(2), 115. https://doi.org/10.33365/jti.v14i2.679

[7] A. Oktian Permana, & Sudin Saepudin. (2023). Perbandingan algoritma k-nearst neighbor dan naïve bayes pada aplikasi shopee. *Jurnal CoSciTech (Computer Science and Information Technology)*, *4*(1), 25–32. https://doi.org/10.37859/coscitech.v4i1.4474

[8] Arifin, T. (2015). *METODE DATA MINING UNTUK KLASIFIKASI DATA SEL NUKLEUS DAN SEL RADANG BERDASARKAN ANALISA TEKSTUR: Vol. II* (Issue 2).

[9] Nugroho, G., Murdiansyah, D. T., & Lhaksmana, K. M. (n.d.). *Analisis Sentimen Pemilihan Presiden Amerika 2020 di Twitter Menggunakan Naïve Bayes dan Support Vector Machine*.

[10] *Prosiding Seminar Nasional Teknologi dan Informatika, 2017 : Kudus, 25 Juli 2017*. (n.d.).