OPEN ACCESS
Indo-JC Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

IND. JOURNAL ON COMPUTING
VOL. 4, ISSUE. 2, SEPT 2019. PP. 119-126
DOI:10.21108/INDOJC.2019.4.2.330

Named Entity Recognition for an Indonesian Based Language Tweet using Multinomial Naive Bayes Classifier

Ramadhyni Rifani [#1], Moch Arif Bijaksana [*2], Ibnu Asror [#3]

[#] *School of Computing, Telkom University*
jalan Telekomunikasi No. 1, Terusan Buahbatu, Bandung, Indonesia

[1] dhynirifani@gmail.com

[2] arifbijaksana@telkomuniversity.ac.id

[3] iasror@telkomuniversity.ac.id

Abstract

In Natural Language Processing (NLP), Named Recognition Entity (NER) is a sub-discussion that is quite widely used for research. The main task of the Named Entity Recognition (NER) is to help identify and detect the entity name of a word contained in the sentence. The data sources used are Indonesian language tweets that are real time, often occur, and the number of words each tweet is limited to 280 characters. The required words on Indonesian language tweets can refer to the name of the entity of a person, location, organization. Therefore, the entity name is determined by considering the word pattern around it. In Indonesia, the average an account posts a tweet at least 1-3 tweets every day containing formal and informal sentences. This is a difficult challenge to provide the right entity naming. In this study, we conducted the naming of entity for Indonesian tweet using the Multinomial Naive Bayes Classifier algorithm. The system uses precision, recall, and f-measure as evaluation metrics. The naming of this entity is able to classify with the highest f-1 value of 80%.

Keywords: Natural Language Processing, NLP, Named Entity Recognition, NER, Tweets, Multinomial Naive Bayes Classifier,

Abstrak

Dalam Natural Languange Processing (NLP), Named Entity Recognition (NER) merupakan sub-bahasan yang cukup banyak digunakan untuk penelitian. Tugas utama dari Named Entity Recognition (NER) yaitu membantu mengidentifikasi dan mendeteksi nama entitas dari suatu kata yang terdapat dalam kalimat. Sumber data yang kami gunakan yaitu tweet bahasa Indonesia yang bersifat real time, sering terjadi, dan jumlah kata setiap tweet dibatasi yaitu 280 karakter. Kata yang terdapat pada tweet bahasa Indonesia dapat merujuk nama entitas orang atau lokasi atau organisasi, sehingga untuk menentukan nama entitas tersebut harus mempertimbangkan terlebih dahulu dengan melihat pola kata disekitarnya. Di Indonesia, rata-rata suatu akun memposting tweet paling tidak 1-3 tweet setiap hari yang berisikan kalimat formal dan informal. Ini merupakan tantangan yang cukup sulit untuk memberikan penamaan entitas yang tepat. Pada penelitian ini kami melakukan penamaan entitas tweet bahasa Indonesia dengan menggunakan algoritma Multinomial Naive Bayes Classifier. Sistem menggunakan presicion, recall, dan f-measure sebagai metrik evaluasi. Penamaan entitas ini mampu mengklasifikasi dengan nilai f-1 tertinggi yaitu 80%.

Kata Kunci: Natural Language Processing, NLP, Named Entity Recognition, NER, Tweet, Multinomial Naive Bayes Classifier,

RECEIVED ON AUGUST 2019. ACCEPTED ON SEPT 2019

I. INTRODUCTION

The definition of social media according to the online dictionary Merriam-webster is a form of electronic communication where users build online communities for various information, personal messages, and other contents such as video [10]. With the existence of social media, we can interact and exchange information anywhere and anytime. One example of social media that is quite widely used in Indonesia, namely Twitter. Twitter is a text microblogging network created by Jack Dorsey [9]. Twitter not only can post tweets along 280 characters but also can post pictures and share video links. Twitter users also vary from organizations, governments, and individuals whose purpose is microblogging, media promotion, news media, and exchanging knowledge. On Twitter, there are tweets with a variety of texts, ranging from formal texts to informal texts. Formal text is text that conforms to the rules of grammar, using standard words, and usually used in formal or official situations. While the informal text is a text that deviates from the rules of grammar by using non-standard words and used in informal or informal situations.

Tabel I
THE DIFFERENCE BETWEEN FORMAL AND INFORMAL TWEETS IN THE DATASET

Tweet	Type
Apakah kita tetap akan disini ?	Formal
Dunia ini sangat berwarna	Formal
kalo gini kan enak, tiap hari gretong	Informal
besok mager ngapa-ngapain help @amanda	informal

Named Entity Recognition or NER is part of the Text Mining and Natural Language Processing process which is very useful in the process of extracting information [1]. Information extraction is a process of finding information from a documents or natural language as its input from the results of useful information in the form of structured information with a certain format. The main task of the NER is to identify and classify names in text into predetermined classes. Table II is an example of an entity detected in a tweet.

Tabel II
NAMED ENTITY RECOGNITION

#	Entity	Information	Example
1	PER	Name Person	Nani Sumarni, Anhar Mahmud
2	ORG	Name Organization	BMKG, BTN
3	LOC	Name Location	Indonesia, rumah
4	O	Other	jalan, di

For the example in Indonesian language tweets the system can recognize 'Nani Sumarni' and 'Anhar Mahmud' as person's names and 'rumah' as the location in the sentence 'Nani Sumarni dan Anhar Mahmud sedang di rumah'. Multinomial Naive Bayes is used to solve the problem of entity identification on tweets.

Multinomial Naive Bayes Classifier is an algorithm developed from the Classifier Naive Bayes theorem [17]. Naive Bayes Classifier is a type of statistical classification [7]. The main theory of Naive Bayes is predicting the probability of class membership. The more training, the better the level of accuracy that will be produced.

In this study, we want to apply entity naming to Indonesian tweets using the Multinomial Naive Bayes Classifier method, starting with collecting data manually from Indonesian tweets, preprocessing, feature extraction, and NER using Multinomial Naive Bayes Classifier.

II. RELATED WORK

Previous research related to Recognition of Named Entities for the Indonesian Language by using 700,000 news articles and the Indonesian Wikipedia to practice the word embeddings. This model is created using Word Vector Representation, Coonvolutional Neural Networks, and Long Short Term

Memory which consists of Bidirectional-LSTM and Layers Connected Layers. The names of the entities used in this study are the names of people, organizations, locations, events, and others. Based on the results of this study got an average f-1 value is 77% [8].

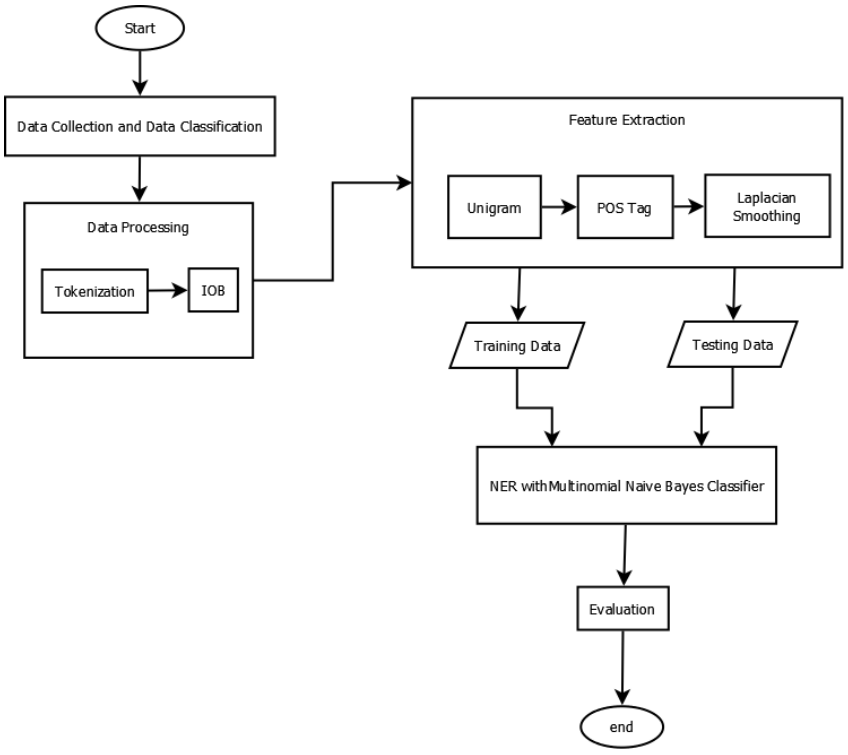
The other studies related to Named Entity Recognition for Indonesian language tweets using 8000 formal and non-formal tweets for data train and 2000 formal and non-formal tweets for data testing. The data preprocessing stage uses a lower case and tokenization. This model is created using the NER standard that implements CRF Classifier. The name entities used in this study are people's names, location names, organization names, and others. Based on the results of the study by combining formal and non-formal sentence models, it produces 62% recall value and 87% precision value for formal tweets, while 36% recall value and 90% precision value for non-formal [14].

This study relates to the performance of the Naive Bayes Multinomial Classifier by using news data monitored by the Processing Institute and Information Providers (PPI), Director General of Information and Public Communication (IKP), and the Ministry of Communication and Information. The text preprocessing stage uses case holding, tokenization, stemming, and filtering used for classification. Before entering the classification stage, a feature selection process is carried out using DFThresholding to reduce the data dimensions and calculate the number of documents. The next step is determine the threshold, if the data is less than the threshold, it cannot be used in the classification process. In the classification process using the Multinomial Naive Bayes Classifier method with an accuracy of 92% [15].

III. METHODOLOGY

A. System Overview

There are three main process in the system, that is preprocessing, feature extraction, and Named Entity Recognition using Multinomial Naive Bayes Classifier. Fig 1 is an overview of the Named Entity Recognition general process.



Gambar 1. System Design of Named Entity

B. Data Collection and Data Classification

This study uses data from Indonesian Language tweets taken manually from Twitter. That has been labeled with 1000 tweets for training data and 300 for testing data. There are three types of entities that can be recognized, that are the name of Person (PER), Organization (ORG), and Location (LOC). But the system will not identify a username (such as @ramadhyni) as a person name entity and hashtag (such as #BuahBatu) as a location entity.

C. Data Processing

Data processing is a process to generate words that can improve the system's performance. There are three processes carried out in data processing, which are Tokenization, giving IOB notation, and POS Tag. Tokenization is the process of splitting a sentence or text into words. IOB notation is marking each token with one of the other three tags. The three tags are I (inside), O (other), and B (begin), whereas POS Tag shows the class of a word in a sentence. After tokenization process and marking IOB notation are done, it proceed to the POS Tag process. These tokenization process and POS Tag are combined into one. Tabel III is an example of Labeling.

Tabel III
Example of Labeling

Tweet	Tag	Entity
Gedung	NOUN	B-LOC
Sate	NOUN	I-LOC
berada	VERB	O
di	CONJ	O
Bandung	PROPN	B-LOC
apakah	WP	O
kalian	PRON	O
tertarik	VERB	O
untuk	CONJ	O
datang	VERB	O
kesini	VERB	O
bersama	ADV	O
Danial	NOUN	B-PER
?	PUNCT	O
#BandungJuara	PUNCT	O
#Bandung	PUNCT	O
@amanda	PUNCT	O
@annisa	PUNCT	O

For the example, the tweet contains an location entity (Gedung Sate is the historic building in Bandung city), (Bandung) an location entity, (Danial) is person entity, and besides (Gedung Sate, Bandung, and Danial) is other. In this study the process of neutralizing formal text to informal text is not used, because the Multinomial Naive Bayes is Supervised Learning. Supervised Learning is an approach where there are already data which has been trained.

D. Feature Extraction

Feature extraction is the process of extracting attributes from a text or training data that has been processed [6]. Feature extraction is also referred to a process of converting data from text into a numerical feature for machine learning processes. If this process is done, it will proceed to the calculation process with naive bayes [7]. In this study, the features to be used are as follows:

- POS Tag
  POS (Part of Speech) Tag is giving a tagset to a word [5]. POS Tagging is an activity which annotating every word or token such as nouns, verbs, adjectives, and others [13]. In this study, the purpose of using the POS Tag is to improve the value accuracy, and also to look for words that are most likely to have Named Entity [2]. POS Tag labeling in this study uses polyglot. There are several word classes that are used which are explained in Table IV. There are different type to use tags in the use of this research [14], as follows:

Tabel IV
PART OF SPEECH

POS	POS Name	Example
PROPN	Specific Noun	Indonesia, Bandung
VERB	Verb	Pergi, Jalan
FW	Foreign Word	Car, #hidupsehat
NOUN	noun	Mobil, Handphone
PUNCT	Symbols/Punctuation	(, :
NEG	Negation	Tidak, Gak
ADJ	Adjective	Cantik, Kuat
ADV	adverb	Sementara, Nanti
CONJ	Conjunction	Di, Ke, dari
CD	Number	satu, dua
PRON	Pronouns	Saya, Kamu
WP	Question	apa, bagaimana

- Unigram
  In this study, Unigram is used by comparing one word before and after involving two steps, which is dividing the string into overlapping n-gram and doing a check to get a substring that has the same structure [12]. Furthermore, the use of Capital letters in a text is important to determine the Named Entity class [13].
- Laplacian Smoothing
  Laplacian Smoothing is one of a smoothing method that can be used in the Naive Bayes algorithm [7]. The purpose of using this method is a solution to avoid to avoid zero values in probability calculations. One of the problems in the data is the limited data and causes a zero value to appear in the probability calculation which results in not being able to classify the data.

$$P(wi|Cj) = \frac{count(wi, Cj) + i}{\sum wEVcount(w, Cj) + |C|} \tag{1}$$

  Formula 1 is a formula to seek Laplacian Smoothing. Formula 1 is used to avoid zero values in Naive Bayes probabilities. *P(wi|Cj)* is the number of wi words occurred in Cj class and $\sum$*wEVcount (w, Cj)* is the number of all w words in Cj class.

E. Named Entity Recognition using Multinomial Naive Bayes Classifier

Naive Bayes Classifier or NBC is a classification method using simple probabilities that calculate a set of probabilities by summing frequencies and value combinations from the given data set [16]. The core of Naive Bayes is to find the highest probability value of a data [3].

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

Formula 2 have four function, first *Pcd* is a class c probability after d is entered to class c, *P(c)* is the probability of class c before, *Pdc* is a probability of d in class c, and the last is *Pd* is a probability of d.

In this study, the data train and the data testing are processed using the Naive Bayes Multinomial Classifier. Naive Bayes Multinomial is a supervised learning classification method that uses probabilistic models by implementing algorithms for multinomially distributed data used in the classification of text where the data is represented as the number of vectors [18]. Distributions are determined by vectors for each class. Multinomial Naive Bayes is applied regardless of the order of words or documents in general cite sustainable2017 analysis. According to Naive Bayes Multinomials, in general the probability of a document d, as part of class members c. Probability from document d against class c can be calculated by Formula 3.

$$Cmap = argmax_{c\epsilon c} P(c) \prod_{1}^{nk} P(w|c) \tag{3}$$

Formula 3 have *P(w|c)* is the probability of the occurrence of the word w in the class in all training documents and *P(c)* is prior to the probability of documents in class c.

$$P(c) = \frac{Nc}{N} \tag{4}$$

Formula 4 have *P(c)* is prior class probability c, *Nc* is the number of words contained in that class, and *N* is the total word.

F. Evaluation

This study uses F-measure to measure the performance of the process described above. F-measure is taken from recall and precision results. Evaluation of features in this study measures individual work features and combined features to determine the best performance. Before searching for recall and precision values, first create a Confusion Matrix table. Confusion matrix is a calculation method that consists of information about the results of classification carried out by the system, prediction lines are the results of classification classified by the system while the actual column is the result of actual prediction whisch are done manually [4]. The Matrix table is shown in the following table V:

Tabel V
Tabel Matrix

	Actual Positive	**Actual Negative**
Predected Postive	TP	FN
Predected Negative	FP	TN

$$p = \frac{TP}{TP + FP} \tag{5}$$

After create a Confusion Matrix and then calculate Precision using Formula 5 that value have ranges from 0 up to 1 [19].

$$r = \frac{TP}{TP + FN} \tag{6}$$

r is a recall value that ranges from 0 to 1. After the precision and recall values are known, then the F-Measure value can be found using Formula 7 [11].

$$f - 1 = \frac{2pr}{p + r} \tag{7}$$

The following is an explanation of each function in the formula above. *p* is precision, *r* is recall, *f-1* is F-measure, *TP* is the number of correct and positive predictions, *TN* is the number of correct and negative predictions, *FP* is the number of false and positive predictions, and *FN* is the number of false and negative predictions.

IV. Results and Discussion

This section will explain the steps and results of the study obtained. The evaluation metrics used to measure performance is Precision, Recall, and f-1.

Tabel VI
precission, Recall, and F-1 score of additional post tag features

Named Entity	**Precission**	**Recall**	**F-1**
PER	51%	84%	74%
ORG	61%	87%	71%
LOC	66%	50%	62%

Table VI shows the best results of precision, recall, dan f-measure from each Named Entity class in the study conducted by us. The highest f-1 score is generated by the PER tag 74% followed by ORG

71%, and LOC 62%. The LOC tag gets the lowest f-1 score compared to other entities. The following is an example of an error in the word marking as an LOC in the result of the study.

Tabel VII is the best result from the research conducted by us. The best performance of this research is the combination of many features.

Tabel VII
PRECISSION, RECALL, AND F-1 SCORE FOR ALL FEATURE COMBINATION

Features	Precission	Recall	F-1
Unigram + Smoothing	60%	67%	61%
Unigram + POSTag	70%	85%	75%
Smoothing + POSTag	60%	65%	63%
Unigram + Smoothing + POSTag	85%	80%	**80%**

On the Unigram and Laplacian Smoothing features, it produces poor performance, both of these features show that it doesn't work well as an individual feature. Because Laplacian Smoothing can only make the probability value not zero. From each calculation, data is added to one and will not make a significant difference in probability estimation.

In the POS Tag feature, it produces good performance, an individual feature and combined with other features can produce quite good performance. One way to increase performance value is by combining the use of Unigram, Laplacian Smoothing, and POS Tags by producing a performance value of 80%.

V. CONCLUSION

The problem found in this research is naming entities on twitter consisting of formal texts and informal texts. Based on these problems and other studies that have not used the Multinomial Naive Bayes Classifier algorithm, it becomes a difficult challenge to provide the right entity naming. In addition, informal texts are deviant from the rules of grammar.

Naming the entity by implementing the Naive Bayes Multinomial algorithm requires a lot of data train, because the more data train results the better the accuracy. This study uses data train as many as 1000 tweets, and 300 tweets for data test. Data train and data test tokenization and IOB notation prior to adding POS Tags, unigram, and Laplacian Smoothing. The Data Testing then calculates the probability, the highest probability result will be the output of Named Entity.

The experimental results in this study produced the best f-1 score of 80% using the compilation of all features. This shows that the addition of POS Tags is the best feature for Named Entity using Multinomial Naive Bayes with an increase in f-1 score of 19%. With this, it can be concluded that based on the above problems, the Multinomial Naive Bayes Classifier can solve the problem of naming entities using tweets in Indonesian.

Further research is expected to be added with other features that can increase the value of accuracy as well as the dataset of tweets that are increasingly being multiplied in order to improve performance for NER tweets. And datasets can be grouped into two parts, formal and non-formal sentences. As well as adding other entity names, such as event names, time, and others.

PUSTAKA

[1] Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science and Business Media, 2012.

[2] Bayu Aryoyudanta. *Pendekatan Pembelajaran Semi-Supervised untuk Named Entity Recognition (NER) Bahasa Indonesia Menggunakan Algoritma Co-Training*. PhD thesis, Universitas Gadjah Mada, 2016.

[3] Garnis Berliana, Shaufiah Shaufiah, and Siti SaâĂŹadah. Klasifikasi posting tweet mengenai kebijakan pemerintah menggunakan naïve bayesian classification. *eProceedings of Engineering*, 5(1), 2018.

[4] Moch Arif Bijaksana, Siti Sa'adah, et al. Klasifikasi argumen semantik menggunakan kombinasi fitur named entities in constituent, head word pos, dan syntactic frame. *eProceedings of Engineering*, 2(2), 2015.

[5] David Christianto, Elisafina Siswanto, and Ria Chaniago. Penggunaan named entity recognition dan artificial intelligence markup language untuk penerapan chatbot berbasis teks. *Jurnal Telematika*, 10(2):8, 2016.

[6] Cahyo Darujati and Agustinus Bimo Gumelar. Pemanfaatan teknik supervised untuk klasifikasi teks bahasa indonesia. *Jurnal Bandung Text Mining*, 16(1):5–1, 2012.
[7] Sigit A Dayinta W W Putra P A. Named entity recognition (ner) pada dokumen biologi menggunakan rule based dan naÃŕve bayes classifier. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(11):4555–4563, 2018.
[8] Devin Hoesen and Ayu Purwarianti. Investigating bi-lstm and crf with pos tag embedding for indonesian named entity tagger. In *2018 International Conference on Asian Language Processing (IALP)*, pages 35–38. IEEE, 2018.
[9] Iwan Kosasih. Peran media sosial facebook dan twitter dalam membangun komunikasi. *Lembaran Masyarakat: Jurnal Pengembangan Masyarakat Islam*, 2(1):29–42, 2016.
[10] Nuning Kurniasih, S Sos, and M Hum. Penggunaan media sosial bagi humas di lembaga pemerintah. In *Forum Kehumasan Kota Tangerang*, 2013.
[11] Daniel T Larose and Chantal D Larose. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.
[12] Erick Alfons Lisangan. Implementasi n-gram technique dalam deteksi plagiarism pada tugas mahasiswa. *TEMATIKA, Journal of Informatics and Information Systems*, 1(2):24–30, 2013.
[13] Ony Naraulita Maringga. Pemeriksaan penggunaan huruf kapital pada teks bahasa indonesia menggunaan metode rule based. 2018.
[14] Y Munarko, MS Sutrisno, WAI Mahardika, I Nuryasin, and Y Azhar. Named entity recognition model for indonesian tweet using crf classifier. In *IOP Conference Series: Materials Science and Engineering*, volume 403, page 012067. IOP Publishing, 2018.
[15] Amelia Rahman, Wiranto Wiranto, and Afrizal Doewes. Online news classification using multinomial naive bayes. *ITSMART: Jurnal Teknologi dan Informasi*, 6(1):32–38, 2017.
[16] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York, 2001.
[17] Imanudin Shaufiah and Ibnu Asror. Android short messages filtering for bahasa using multinomial naive bayes. 2006.
[18] Muhammad Haerunnur Syahnur, Moch Arif Bijaksana, and Mohamad Syahrul Mubarok. Kategorisasi topik tweet di kota jakarta, bandung, dan makassar dengan metode multinomial naïve bayes classifier. *eProceedings of Engineering*, 3(2), 2016.
[19] Akhmad Zaini, M Aziz Muslim, and Wijono Wijono. Pengelompokan artikel berbahasa indonesia berdasarkan struktur laten menggunakan pendekatan self organizing map. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, 6(3):259–267, 2017.