Comparison of Feature Extraction for Sarcasm on Twitter in Bahasa

Nurul Afifah Arifuddin
Postgraduate Student of Electrical Engineering Department
Hasanuddin University
Makassar, Indonesia
afifah.nurul97@gmail.com

Indrabayu
Informatics Department
Hasanuddin University
Makassar, Indonesia
indrabayu@unhas.ac.id

Intan Sari Areni
Electrical Engineering Department
Hasanuddin University
Makassar, Indonesia
intan@unhas.ac.id

***Abstract*—This study aims to detect the text of sarcasm in Bahasa. Sarcasm detection is very important in the field of affective computing and sentiment analysis because expressions of sarcasm can reverse the polarity of sentences. Sarcasm is difficult to detect in text because there is no intonation of sounds and facial expressions. Therefore, in this study, a system is created to recognize the sentence of sarcasm in text. The data consist of 480 train data and 120 test data collected by crawling on Twitter. Then, the data passed through the preprocessing and feature extraction stages. Classification of sarcasm and non-sarcasm sentences uses the Support Vector Machine (SVM) algorithm. Experiments are done by comparing the accuracy of N-gram, POS Tag, Punctuation, Pragmatic and combining all features. Our proposed approach reaches the highest accuracy of 91.6% with a precision of 92% when all features are combined.**

Keywords—sarcasm, crawling, support vector machine, Ngram, POS Tag, Punctuation, Pragmatic.

I. INTRODUCTION

The survey conducted by the Indonesian Internet Network Providers Association (APJII) throughout 2017 revealed that 87.13% of Indonesians use the internet to access social media [1]. One of the most popular social media is Twitter. In 2016 Indonesia was the country with the third largest Twitter user in the world ("Twitter," n.d). A survey conducted by APJII showed that the most interesting topic was social/environmental news, with a percentage of 50.25%, followed by political news with 36.94% [1]. Twitter is a valuable resource for knowing public opinion because many users use it to post their reviews, and the data can be used to see public opinion on products, services, or express religious and political opinion. However, Twitter is not capable of giving a conclusion automatically about people's sentiments. Therefore, an analysis is needed to find out whether public responses are positive, negative, or neutral. The results of this analysis can taken into consideration, both for the community, product owners, services, or politicians. The Obama administration uses sentiment analysis to measure public opinion before the 2012 presidential election [2]. Since then, many politicians and brand owners have used sentiment analysis to get clearer insights to understand public attitudes and opinions.

Sentiment analysis is used to analyse expressions and opinions in text, and this can help us to understand the emotions and opinions conveyed by the writer [3]. Sentiment analysis has many challenges, and the detection of sarcasm is one of the main problems. Sarcasm detection is very important in many fields, such as affective computing and sentiment analysis because expressions of sarcasm can reverse the polarity of sentences. Sarcasm can be considered a bitter expression or ridiculous. The example of sarcasm statements is "I work 40 hours a week to be a poor person !!!".

Sentiment analysis system that is unable to detect sarcasm sentences will classify tweets inaccurately, which means sarcasm tweet that should recognise as a negative sentence will recognise as a positive sentence [4]. Sarcasm is a type of sentiment where the public expresses their negative emotions using positive words in the text [5].

Analysing public opinion that is on social media is not easy, because opinions on social media are mostly written in non-standard words [6].Also, sometimes, people use sarcasm in their opinions. Sarcasm is a word that has the opposite meaning of what said that is used to mock or show resentment [7]. Various impacts can arise as a result of the use of sarcasm on social media. Among other things, (1) The emergence of the analogy perception that in general, the Indonesian people like to use sarcasm expression. (2) Indirectly social media has educated the public to use sarcastic language. (3) The propaganda that using sarcastic language on social media is common, and is no longer seen as a violation of social ethics. (4) The Indonesian people accept (permissive) this problem [8].

Sarcasm is difficult to analyze automatically even by humans. Sarcasm can be identified from the intonation and facial expressions when the person is talking, but tweets have no tone, facial expressions, and background information [9]. Therefore, detection of sarcasm is still considered a difficult problem in sentiment analysis. Detection of sarcasm sentences is expected to improve the performance of sentiment analysis and affective computing.

Therefore, this research created a system that can be used to detect sarcasm sentences for any topic on twitter. Twitter Indonesia data source was chosen because research on the detection of sentence sarcasm in Indonesian has not been done much. Data from Twitter contains many slang words. Therefore, in this study, a corpus of slang words is made so that tweets can be classified using the SVM method. In this study, the authors used Ngram, POS-Tag, Pragmatic, Punctuation, and a combination of all features. The proposed feature is expected to improve the accuracy of the sarcasm detection system on Twitter. The method used to classify sarcasm sentences is Support Vector Machine (SVM).

The remaining of this paper is structured as follows. Section II describes some state of the art work related to our proposed approach. Section III presents detail of our proposed method. The conclusion and our future work is presented in section IV and our research conclusions are presented in section V.

II. Related Work

Research on the detection of sarcasm sentences has been carried out by many previous researchers, especially in English. Carvalho and Sarmento, in their study, revealed that the detection of sarcasm requires specific oral or gestural instructions. In the text, these instructions will be replaced by emoticons, punctuation marks, quotes, and interjection. All of these features are a sign that text comments or posts on social media try to mimic real conversations [10]. Prawira et al. obtained that the results of sarcasm detection with 1500 training data can increase the accuracy of sentiment analysis by 1.20%. The detection of sarcasm can also increase precision by 2.43%. However, the increase in accuracy and precision of sentiment analysis has resulted in a decline of the recall by 0.27% [11].

Bouazizi and Ohtsuki proposed a pattern-based approach using Part-of-Speech tags to detect sarcasm on Twitter and used four sets of feature extraction: sentiment-relate, punctuation-relate, lexical and syntactic, and pattern-related. The proposed approach has good results, although according to the authors, the results would be much better if the system is built using a larger training tool because patterns taken from the current pattern might not cover all sarcastic patterns [12].

Afiyati et al. also used the same features as Bouazizi, the difference is, Afiyati et al used Indonesian language data from WhatsApp group conversation [13]. The same system was also created by Lunando and Purwarianti, using Unigram, Negativity, Number of Interjection words and question words as features to detect sarcasm using data from the Indonesian twitter feed. However, the system produced low accuracy because there are many sarcasm texts that do not have a global topic so it is difficult to detect whether the sentence is sarcasm or not [7].

Meanwhile, in 2017, Prasad et al. compared various classification algorithms such as Random Forest, Gradient Boosting, Decision Tree, Adaptive Boost, Logistic Regression and Gaussian Naïve Bayes to detect sarcasm in tweets from Twitter Streaming API. Mapping of emoji and slang dictionaries was a new idea introduced in this study and used in the preprocessing stage. The features used in this study were: blob polarity, subjectivity, capitalisation, positive sentiment, and negative sentiment. The results showed that the Gradient Boosting algorithm had the best performance among other classification algorithms [14]. Ibanez et al. detected sarcasm using lexical and pragmatic features, the authors found that three pragmatic features: To User, smiley, and frown were among the ten most discriminatory features in the task of classifying sarcasm sentences. The results of this study indicated that using lexical features without other features was not sufficient to identify sarcasm and that pragmatic and contextual features were worthy of further study [15].

In 2018, Rahayu et al. made a study on the detection of sarcasm in Indonesian by combining interjection, punctuation, Bag of Words, and Naïve Bayes performance. The results obtained with F-Measure are 82%. To improve the performance of the sarcasm detection system, Rahayu et al. suggested to try other classification methods and add more sophisticated features [16].

III. Proposed Method

Sarcasm detection in this system consists of several stages. First, data is collected by crawling on Twitter and then stored in CSV format, then preprocessing is carried out to obtain structured data and feature extraction using N-gram, Post Tag, Pragmatic, and Punctuation. Furthermore, the classification of sarcasm and non-sarcasm sentences uses the support vector machine (SVM) algorithm. The proposed method can be seen in Figure 1.

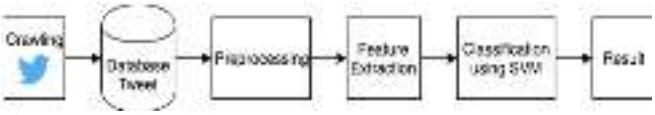

Fig. 1. Proposed Method

A. Input Data

The input data used in this study are tweets in Bahasa obtained by crawling. The data consist of 480 tweets for train data and 120 tweets for test data.

B. A Scenario for Data retrieval

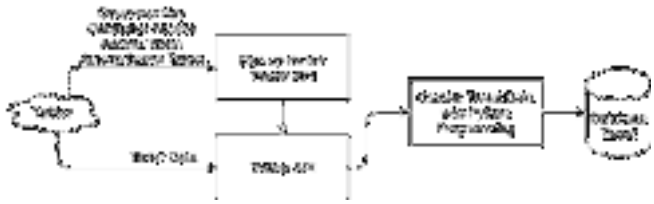
Fig. 2. Crawling Process on Twitter

Twitter provides an API that can be used to search for specific words. This study uses REST Application Programming Integration (API) to collect tweets using several keywords:

- #sarcasm: Tweets with hashtags #sarcasm are assumed to be sarcasm sentences.
- #fakta: Tweets with hashtag #fakta are supposed to be facts and therefore, can be consider as non-sarcasm data.

Based on Figure 2, to crawl data on Twitter, first enter consumer_key, consumer_secret, Access_token, and access_token_secret obtained when registering the Twitter API. Second, run a crawling program that has been made with the Python programming language, then enters keywords as mentioned above. Crawling data stored in CSV format and then the preprocessing stage is carried out. At this stage, 300 sarcasm data and 300 non-sarcasm data were collected.

C. Data Preprocessing

Data crawled from Twitter sometimes contain usernames, hashtags, URLs, repetitive tweets and do not use standard words. Moreover, Twitter users often abbreviate a word because of the limited character in a tweet. Therefore, preprocessing plays an important role in this research. The Tweet will be normalised so the classifier can process it. There are several processes in this stage as shown in figure 3.

Fig. 3. Preprocessing Stage

- Remove username (@user): Twitter allows users to reply to other tweets and mention the account name in a tweet. Therefore, in the preprocessing stage, tweets containing username are deleted.
- Remove repetitive tweet: there are several repetitive tweets due to the "retweet" feature on Twitter. A retweet is useless because they have the same content as the original tweets. Tweets that are double or contain RT letters (Retweet) deleted at this stage.
- Remove #: Twitter hashtags (#) are used to show the context of tweets. This character needs to remove because it is not a feautere. Therefore, a word containing "#" needs to be deleted.
- Remove URL: A tweet sometimes contains URLs, especially tweets that contain images, where the image on the tweet after crawling will also change to the URL. URLs considered useless because they do not convey meaningful features unless the URL is opened and processes the information from the referenced URL. This study focuses on the detection of sarcasm sentence text, and then the URL is deleted by detecting tweets containing "www" and "https: //".
- Remove stop word: Stop Stop words are common words that usually appear in large numbers and are considered to have no meaning. Stop words generally used in task information. In this study, the stop words removal is done to reduce the number of words that must be processed.
- Tokenize: Tokenization is separating words, symbols, phrases, and other important entities into word parts called tokens. This study uses the NLTK library for tokenisation.
- Change slang word to standard words: Indonesian Indonesian Tweets crawled from Twitter tend to use slang words rather than standard words, such as replacing the alphabet with numeric and use general informal words to replace formal words. Therefore, the author built a dictionary and used it to translate slang words into formal word.

An example of a tweet before going through the preprocessing stage can be seen below:

RT@fnabila: Asiiiiiik bentar lagi sekolah!!!!!!!!! ga sabar !!!!!!!! #sarkasme

In the preprocessing stage, "RT" and "@" are omitted, the slang word "ga" changed to "tidak", and a post tag is given to each word and symbol. Result of the preprocessing stage can be seen below:

[('asiiiik', 'NN'), ('bentar', 'NN'), ('lagi', 'RB'), ('sekolah', 'NN'), ('!, 'Z'), ('!', 'Z'), ('!, 'Z'), ('!', 'Z'), ('!, 'Z'), ('!', 'Z'), ('!, 'Z'), ('!', 'Z'), ('tidak', 'NEG'), ('sabar', 'VB'), ('!', 'Z'), ('!', 'Z'), ('!', 'Z'), ('!', 'Z')]

D. Feature Extraction

Feature extraction is one of the most important parts of sarcasm detection. In this study, four features are used to detect sarcasm.

a) Ngram: The Ngram feature was chosen in this study because, in Indonesian, many phrases did not consist of only one word [17]. Also, the research conducted by Piyoros et al. produced high accuracy in detecting sarcasm sentences in English using the Ngram feature. N-gram refers to the sequence of tokens in sentences where N refers to the number of words in the sequence. In this study, the authors use unigrams (N = 1) and bigrams (N = 2). To use N-gram, the author uses the NLTK library, and in classifying, the tweet is tokenised as described in the preprocessing stage, lemmatised to normalise the word based on the basic form which is the shape of the lemma, and uncapitalised to change uppercase letters to lowercase.

b) Pos tag Features: POS-Tag is a common process carried out in NLP. The POS-Tag (Part of Speech) feature in our study inspired by the work of Bouazizi et al. who used the English POS-Tag to detect sentences of sarcasm. POS-tags is the categorisation of word classes, such as nouns, verbs, adjectives, etc. POS Tagger is a method that can automatically annotate part of speech tags for each word in the document. The Corpus POS Tag used in this study consists of 22 tags and a collection of Indonesian languages of more than 250,000 lexical tokens made by Dinakaramani et al., which can be seen in table 1[18].

c) Pragmatic features: Pragmatic is a branch of linguistics that examines expressions in depth and analyses implicit and literal meanings. In some previous studies, pragmatic is one of the features used by researchers to detect sarcasm in tweets. Pragmatics learn languages that are not used directly, in this case using instructions. The pragmatic features used are:

- The number of capital letters: Capitalization is used to give extra emphasis on the emotions conveyedin the text, so the number of capital letters in a tweet is calculated then used as a feature.
- Number of Emoticons: Emoticons are commonly used on all social media platforms to express the sentiment.
- Amount of Slang Expression: Because sarcasm is intended to have an element of humour, the appearance of a higher slang is considered to have the potential to indicate sarcasm. Common slang words like 'lol', 'rofl' and 'lmao' quite commonly used.

TABLE 1. INDONESIAN TAGSET [18]

Tag	Description	Example
CC	Coordinating conjucnction	tapi, namun, dan
CD	Cardinal number	empat, juta, delapan, 8497, seperdua, 1, 309, banyak, ratusan
OD	Ordinal number	pertama, kedua, ke-3, keempat
DT	Determiner/article	Sang, Para, Si
FW	Foreign Word	Meaning, terms & conditions
IN	Preposition	Oleh, sebelum, selama, di, dari, setelah

Tag	Description	Example
JJ	Adjective	Menarik, serasi, hidup, lucu, tepat, aneh, hitam, berani
MD	Modal and auxiliary verb	Akan, harus, boleh, sudah, perlu
NEG	Negation	Tak, belum, jangan, tidak
NN	Noun	Kelinci, mobil, pelatih, arah
NNP	Proper noun	Pramoedya, selat sunda, Indonesia, Bank BCA, Februari, Selasa, Idul Adha, Liga Indonesia
NND	Classifier, Partitive, and measurement noun	seorang, sebuah, ons, halaman, lembar, kilometer
PR	Demonstrative pronoun	Itu, situ, ini, sini
PRP	Personal pronoun	kalian, aku, saya, kami, kamu, dia, mereka
RB	Adverb	Saja, lebih, sering, bahkan, juga
RP	Particle	pun, -lah, -kah
SC	Subordinating conjunction	jika, supaya, begitu, agarm bahwa, kalau
SYM	Symbol	USD, IDR, +, -, !, #
UH	Interjection	mari, nih, ya, tuh, aduh, oh
VB	Verb	Mengusir, mengarahkan, menyebutkan, ditempatkan
WH	Question	Apa, dimana, berapa, siapa, mengapa, kenapa
X	Unknown	statemen

d) Punctuation features: Punctuation has many influences in text classification, especially in the field of sentiment analysis. The punctuation feature can be used to display aspects of behaviour such as low tones, facial movements, or exaggeration [19]. These aspects are translated into the use of certain punctuation or certain vocal repetitions when tweets were written. It has been observed that in sarcastic tweets, a large number of punctuation marks are used such as '!',? 'And' ... '.

IV. EXPERIMENTAL AND RESULT

Input data used in this study consisted of 600 data with a ratio of 80: 20 for training data and testing data. The text data of sarcasm and non-sarcasm sentences amounted to 300 data each with 240 training data and 60 testing data.

The system testing method used in this study is F-Measure, which is the value obtained from the measurement of precision and recall between the clustered class and the actual class contained in the input data. The following formula can achieve precision and recall:

$$Precision: \frac{TP}{TP+FP} (1)$$

$$Recall: \frac{TP}{TP+FN} (2)$$

After the values of precision and recall obtained, then F-Measure is calculated by the formula below:

$$F1\text{-}Score = 2.\frac{Precision .Recall}{Precision + Recall} (3)$$

The Accuracy can obtain by the following formula:

$$Accuracy : \frac{TP + TN}{TP + TN + FP + FN} (4)$$

In (1), (2), and (4), *TP* is true positive, *TN* is true negative, *FP* is false positive, and *FN* is false negative. The result of Sarcasm Text Detection on Twitter in Bahasa Using SVM by testing each feature was shown in Table 2.

TABLE 2. ACCURACY OF EACH FEATURE

Feature	Class	Precision	Recall	F1-Score	Accuracy
Ngram	Sarcasm	0.89	0.90	0.89	89.16
	Non-Sarcasm	0.90	0.88	0.89	
POS Tag	Sarcasm	0.53	0.40	0.46	52.5
	Non-Sarcasm	0.52	0.65	0.58	
Pragmatic	Sarcasm	1.00	0.07	0.12	53.33
	Non-Sarcasm	0.52	1.00	0.68	
Punctuation	Sarcasm	1.00	0.12	0.21	55.83
	Non-Sarcasm	0.53	1.00	0.69	

From the table 2, Ngram feature is the most important feature in the detection of sarcasm, where Ngram has the highest accuracy of 89.16%, followed by the Punctuation feature with an accuracy of 55.83%, Pragmatic 53.33%, and Post Tag 52.5%.

TABLE 3. THE RESULT of COMBINED FEATURES of NGRAM, POS-TAG, PRAGMATIC, & PUNCTUATION

Class	Precision	Recall	F1-Score	Accuracy
Sarcasm	0.92	0.92	0.92	91.66
Non-Sarcasm	0.92	0.92	0.92	

By combining the four features described on table 3, 91.66% accuracy are obtained.

V. CONCLUSION

The dataset used in this study is 480 tweets for training data and 120 tweets for test data obtained by crawling methods. The data must go through the preprocessing and feature extraction stages then the classification process is carried out by using the SVM method with a linear kernel.

The results of the sarcasm and non-sarcasm sentence classification process using Ngram feature yields 89.16% accuracy, the POS Tag feature produces an accuracy of 52.5%, the Pragmatic feature produces 53.33% accuracy, and the punctuation feature produces 55.83% accuracy.

There are 5 data sarcasm and non-sarcasm that are wrongly detected based on experiments by combining all features, this is because some words do not use standard words where the word is not in the corpus of slang sentences made by the author. Also, non-sarcasm data detected sarcasm because they contain features of Pragmatic and Punctuation. The resulted precision is 0.92, with an accuracy of 91.66%.

For further research, improving slang word dictionaries and using more training data will improve the accuracy of the sarcasm sentence detection system.

REFERENCES

[1] "Indonesian Internet Network Providers Association." [Indonesian] [Online]. Available: https://www.apjii.or.id/content/read/39/342/Hasil-Survei-Penetrasi-dan-Perilaku-Pengguna-Internet-Indonesia-2017. [Accessed: 30-Aug-2018].

[2] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle," in *Proceedings of the ACL 2012 System Demonstrations*, Jeju Island, Korea, 2012, pp. 115–120.

[3] Petrix Nomleni, "sentiment analysis using support vector machine," Sepuluh Nopember Institute of Technology, Surabaya, Indonesia, 2015.

[4] S. Albanie and Mm. Oxon, "Sarcasm Detection on Twitter: bolstering lexical features with contextual clues," University of Dublin, Trinity College, Dublin, Irlandia, 2013.

[5] P. Deshmukh and S. Solanke, "Sarcasm Detection and User Behaviour Analysis," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 6, no. 6, p. 6.

[6] G. A. Buntoro, T. B. Adji, and A. E. Purnamasari, "Twitter Sentiment Analysis with a Combination of Lexicon Based and Double Propagation," [Indonesian] p. 6, 2014.

[7] E. Lunando and A. Purwarianti, "Indonesian social media sentiment analysis with sarcasm detection," in *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Sanur Bali, Indonesia, 2013, pp. 195–198.

[8] F. Nugrahani, "Use Of Language In Social Media: Mirror Freedom Of Nation Characters," [Indonesian] *Stilistika*, vol. 3, No. 1, pp. 1–18, 2017.

[9] D. Maynard and M. A. Greenwood, "Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis," p. 6.

[10] P. Carvalho and L. Sarmento, "Clues for Detecting Irony in User-Generated Contents: Oh...!! It's "so easy" ;-)," p. 4.

[11] F. Prawira, "The effect of sarcasm detection on the quality of sentiment analysis on Twitter," Gadjah Mada University, 2017. [Indonesian]

[12] M. Bouazizi and T. Otsuki Ohtsuki, "A Pattern-Based Approach for Sarcasm Detection on Twitter," *IEEE Access*, vol. 4, pp. 5477–5488, 2016.

[13] Afiyati, E. Winarko, and A. Cherid, "Recognizing the sarcastic statement on WhatsApp Group with Indonesian language text," in *2017 International Conference on Broadband Communication, Wireless Sensors and Powering (BCWSP)*, Jakarta, 2017, pp. 1–6.

[14] A. G. Prasad, S. Sanjana, S. M. Bhat, and B. S. Harish, "Sentiment analysis for sarcasm detection on streaming short text data," [Indonesian] in *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)*, London, 2017, pp. 1–5.

[15] R. Gonzalez-Ibanez, S. Muresan, and N. Wacholder, "Identifying Sarcasm in Twitter: A Closer Look," p. 6.

[16] D. A. P. Rahayu, S. Kuntur, and N. Hayatin, "Sarcasm Detection on Indonesian Twitter Feeds," in *2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, [Indonesian], Malang, Indonesia, 2018, pp. 137–141.

[17] W. C. Indhiarta, "The Use of N-grams in Sentiment Analysis of Jakarta Regional Head Election Using the Naive Bayes Algorithm," [Indonesian] p. 18.

[18] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, "Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus," in *2014 International Conference on Asian Language Processing (IALP)*, Kuching, Malaysia, 2014, pp. 66–69.

[19] J. Subramanian, V. Sridharan, K. Shu, and H. Liu, *Exploiting Emojis for Sarcasm Detection*. 2019.