Sundanese Language Level Detection Using Rule-Based Classification: Case Studies On Twitter

Ade Sutedi, Dede Kurniadi, Wiyoga Baswardono

Abstract : Along with the history of the Sundanese tradition, language has an important role to show the existence of Sundanese culture, especially in Banten and West Java. Today, the use of Sundanese language are decreased due to a competition of regional languages with national languages even with foreign languages. In addition, the divergence of language in the society cause disparities between young people and older people. The native speaker are reduced due to social developments in society that are increasingly wide open. This issue becomes popular in the last decade due to the death of language especially for regional language. To discover the existence of Sundanese language in social media today, Twitter was used to analyze as a parameter that indicate the existence of a Sundanese language used by the people. The objectives of this research are: (1) to identified the existence of Sundanese language in social media; (2) to classify the word levels of Sundanese language used and comparing their levels to get the summary of characteristic of Sundanese language for every region. In this research, classification process taken from Sundanese vocabulary which divided into three levels: Ribaldry level (Loma), Standard level (Hormat ka sorangan), and Polite level (Hormat ka batur). Classification involves the word n-grams (unigram, bigram, and trigram) features with rule-based classification to determine Sundanese and non Sundanese language with their levels. In this research, the data was retrieved from Twitter user based on their region especially in Banten and West Java provinces. The result shows that the use of Sundanese language among the people still exists and also used and in social media with ribaldry level dominated. Prediction score for several feature is smaller than previous research. But, we consider the precision value of the experimental results obtained score 0.841 which can be used to determine the predictive value close to the actual positive value.

Index Terms : Classification, Language levels, Sundanese, Twitter.

1. INTRODUCTION

Language is one of the communication media used to be able to connect each other. Indonesia is one of the country with variety of languages based on Summer Institute of Linguistics (SIL) [1] which scattered in many regions in Indonesia. One of the most widely used is Sundanese language. This secondary language used in several region of Banten and West Java Provinces and becomes the language that popular and widely used in large population [2]. The existence of Sundanese language cannot be separated from the culture that developed in their community. The different of location causes the diversity and levels in terms of dialect for each region such as Banten, Cirebon, and “Priangan” [3].

Fig. 1. *The region of West Java and Banten province [19]*

Today, the use of Sundanese language in the community has decreased due to competitions [1] of Sundanese Language usage between young people. So it cause “The death of a language” issue over the community. This evidenced by 40% of the population of West Java who can understand the Sundanese language [2, 21]. in order to maintain the use of Sundanese language among the community, the local government has taken steps that can strengthen the culture and use of mother tongue.Today, Sundanese Langugage was widely introduced in both of media and services to preserve the use of language in the community. It is very important to use the Sundanese in daily living now. Concrete efforts can be made by using Sundanese as a medium of communication in the family environment [1]. In regional health centers [4] between staff and patients in order to create effective and efficient communication. Make regulations on the use of Sundanese language every Wednesday in certain areas in West Java [5]. Therefore, research in Sundanese language was conducted several contribution, i.e. WordNet [2], Sundanese Unicode [6], and Stemmer Algorithm [7], etc.

- *Ade Sutedi is currently lecturer at Department of Informatics, Sekolah Tinggi Teknologi Garut, Jalan Mayor Syamsu 1, Garut 44151, Indonesia. E-mail: adesutedi@sttgarut.ac.id*
- *Dede Kurniadi is currently lecturer at Department of Informatics, Sekolah Tinggi Teknologi Garut, Jalan Mayor Syamsu 1, Garut 44151, Indonesia. E-mail: dedekurniadi@sttgarut.ac.id*
- *Wiyoga Baswardono is currently lecturer at Department of Informatics, Sekolah Tinggi Teknologi Garut, Jalan Mayor Syamsu 1, Garut 44151, Indonesia. E-mail: wiyoga_ baswardono @sttgarut.ac.id*

However, the response of the community in using local languages, especially Sundanese, is still very low. As stated in [2], the process of determining the semantic relations in WordNet by the participant was always tricky. It means that not all participants [2] understand the Sundanese language. In addition, the use of Sundanese language is more popular among communities like Fiksimini Basa Sunda who have a commitment in maintaining Sundanese culture by writing mini fiction in accordance with Sundanese grammar correctly [8]. In order to discover the existence of Sundanese language, this research was used a Twitter social media as source for data based on their location (Latitude and Longitude) associated with the region in West Java and Banten which depicted in table 1.

TABLE 1.

REGION AND THE LATITUDE AND LONGITUDE [20]

Region	Latitude	Longitude
Bandung	-6.914744	107.609810
Bekasi	-6.230833	107.013611
Bogor	-6.594444	106.789167
Ciamis	-7.3257	108.3534
Cianjur	-6.822222	107.139444
Cirebon	-6.7252	108.5678
Garut	-7.2024	107.8878
Indramayu	-6.326389	108.32
Karawang	-6.303333	107.305556
Kuningan	-6.975833	108.483056
Lebak	-6.65	106.21667
Majalengka	-6.836111	108.227778
Pandeglang	-6.3084	106.1067
Purwakarta	-6.556944	107.443333
Serang	-8.5432	115.1708
Subang	-6.57	107.756667
Sukabumi	-6.918056	106.926667
Tangerang	-6.178306	106.631889
Tasikmalaya	-7.3274	108.2207

Based on its geographical location, Sundanese language has a diversity of dialects in terms of pronunciation and vocabulary [3]. Language level also appears (Undak-usuk) in some ethnic of Sundanese society stages in language [3] which depicted in the table 2 for several word example.

TABLE 2.

EXAMPLE OF SUNDANESE VOCABULARY [9] BASED ON THEIR LEVELS [3].

Ribaldry (commonly)	Standard (to our self)	Polite (to others)
Abus (get in)	Lebet (get in)	Lebet Lebet (get in)
Acan (not yet)	Teu Acan (not yet)	Teu Acan (not yet)
Bapa (father)	Pun Bapa (father)	Tuang Rama (father)
Cicing (stay)	Matuh (stay)	Linggih (stay)
Ngomong (talk to, speak)	Nyanggem (talk to, speak)	Nyarios (talk to, speak)
Embung (no way, don't want)	Alim (no way, don't want)	Teu Kersa (no way, don't want)
Imah (house)	Rorompok (house)	Bumi (house)

Therefore, by looking at the vocabulary in table 2 that is associated with the location in table 1 it can be used as a reference to look for possible language levels from Twitter data sources. The data separate to segment of words to produce a Sundanese language vocabularies then classified it by comparing vocabularies that already exist in the database. This result becomes a reference of Sundanese language level from users on social media, especially Twitter.

2 METHODOLOGY

In this section, we explained several methods that have been developed in previous research as well as the methods in this research to determine the classification of tweets that containing Sundanese language. In previous studies, text classification in Twitter has been developed for various purposes. Including the Arabic text classification was successfully implemented word N-grams feature [10] with best result using unigrams. Rule-based and statistical method was implemented to conduct an emotion classification [11] in Indonesian language to regional languages such as Sundanese and Javanese [12]. Besides that, twitter also modeled by Naive Bayes and Support Vector Machine (SVM) to classified the informative and uninformative tweets used as a medium for disaster information [13]. Named-Entity Recognition, Event, Sentiment, and Emoticon with Unigram feature and Part of Speech to determine misinformation and detect rumor in Twitter data [14]. Although in [11] has achieved a good accuracy, it still faces difficulties in processing words in tweet that are not complete. Another disadvantage on Twitter classification is uninformative tweets considered informative [13]. Then, the difficulties of classification can happen because the tools are still weak in the preprocessing [14]. Therefore, the classification process using Twitter data still needs to be studied and developed by finding other methods to find more information from raw data to solve existing problems and utilized that for further processing. In order to solve the existing problems, this research applies several methods which have previously been discussed with a simple approach that includes several stages. Including data collection, data pre-processing, and implementation of rule-based classification to determine the class of Sundanese and non Sundanese language and its level. All of these stages are illustrated in figure 2. The database of Sundanese vocabulary [3] was built by manually input process as a comparison data to test the presence of Sundanese words in the tweets. This data, divide into three categories i.e. vocabulary of word, n-grams type (unigrams, bigrams, trirams), and level categories. These categories will be used to determine whether the tweet tested is indeed Sundanese language or not. if the tweet is Sundanese, the level of language used will be determined. The results obtained are adjusted to the user's location feature on twitter.

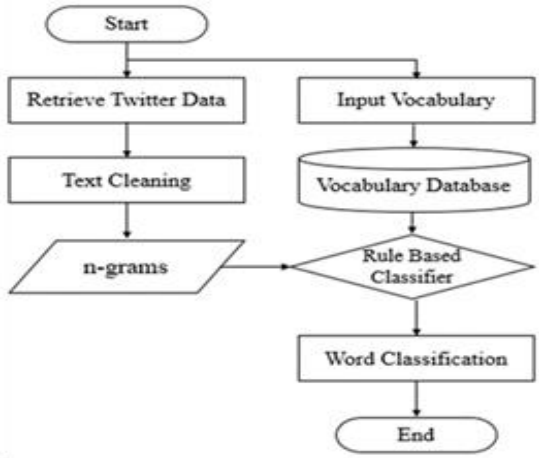

Fig. 2*. The stage of classification process*

2.1 Preprocessing

In this research, the data split into two data set. First, the data

manually input from Sundanese dictionary for database comparing. Second, the data was retrieve from Twitter users as a training data and testing data. Due to the tweets data from the user obtained does not have a suitable location (no latitude longitude position). So, we try to retrieve with the hashtag of region in West Java and Banten (i.e. #Bandung) based on the table 1. The retrieved for every region is 1000 tweets and the total should be 19000 tweets, but we obtain 8446 tweets which depicted in Fig 3. This data will be used as training data in the classification process.

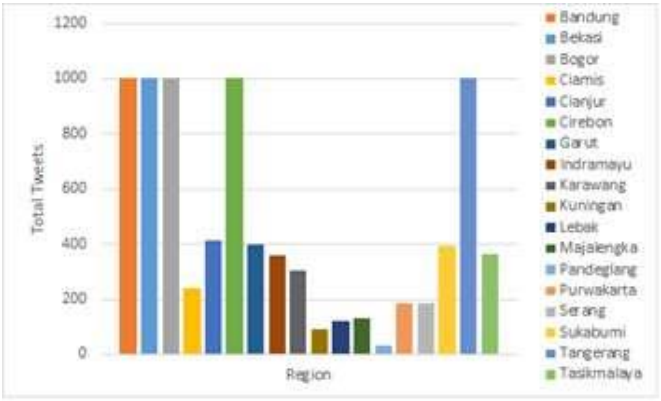

Fig. 3. *The tweets for all region in West Java and Banten*

2.2 N-grams

Classification involves the process of calculating words based on n-grams (Unigram, Bigrams, and Trigrams). Word level from n-grams features could be better result than single term [10]. This features becomes a supporter in the process of classifying texts that become the main constituents of words in Sundanese language. The same thing was done in previous research in the preprocessing process [11], [13], and [14]. N-grams is a feature that is widely used and useful in the classification process, especially in text processing. In this research, the database of vocabulary was built using n-grams features to difine Sundanese words into unigrams, bigrams, and trigrams associated with their levels which total 2387 of words based on [3] data. This database used as comparison data when performed the Twitter data retrieve to produce n-gram categories and Sundanese language levels used by the user.

2.3 Rule-based Classification

Rule-based classification is technique for classifying record using a collection of "if…then…else" rule [18] which represented normal for, R = (r_1 v r_2 v… r_k) where R is the rule set and r_i is the classification rule or conjunctions. Rule-based has been successful in many studies such as emotion classification [11], Sentence-Level Emotion Detection [15], Detecting Travel Modes [16], and Diagnosis of Heart Sounds [17]. In this research, the rule based applied to classify the vocabulary of Sundanese and non Sundanese language with their levels which meets the following rules.

R1 : IF ((Vocabulary = yes) ^ (Level{Ribaldry, Standard, Polite} = yes)) v (Region = yes) → Sundanese
R2 : Else IF ((Vocabulary = yes) ^ (Level{Ribaldry, Standard, Polite} = yes)) v (Region = no) → Sundanese
R3 : Else IF ((Vocabulary = yes) ^ (Level{Ribaldry, Standard, Polite} = no)) v (Region = yes) → Non Sundanese
R4 : Else IF ((Vocabulary = no) ^ (Level{Ribaldry, Standard, Polite} = no)) v (Region = yes) → Non Sundanese
R5 : Else IF ((Vocabulary = no) ^ (Level{Ribaldry, Standard, Polite} = no)) v (Region = no) → Non Sundanese

3 RESULT AND DISCUSSION

In this research, the data was taken from several account of Twitter user based on the hashtag of regions in table 1. Furthermore we proceed it by manually classifying and found that the Sundanese Twitter data only had in several regions with hashtag of Bogor, Cirebon, Garut, Majalengka, Purwakarta, and Tasikmalaya regions. Then, the regions that have a lot of tweets (such as Bandung, Bekasi, Cirebon), the Banten provincial area (Lebak, Pandeglang, Serang, Tangerang) it almost none of Sundanese words. Even though the data has provided information about locations with the hashtag feature, it just obtained around 0.34% from all of Sundanese tweets, so it is not possible to be used as a reference for training data. To overcome the lack of the training data, we added Sundanese tweets content by retrieved from the Twitter official account of the language community in social media namely @fikminsunda [8]. This obtained were 2333 data with containing of 0.81% Sundanese tweets and 0.19 is not Sundanese languages with annotating of yes an no data related to Sundanese language. Due to the lack of latitude and longitude feature, in this data training we did not find any location that were expected to indicate the location where the twit originated so that the process of determining the area was difficult. In order to determine Sundanese language tweets and their level, a rule-based algorithm was implemented for the data testing with the scenario in the previous section. The classification result depicted by confusion matrix in the tabel 3.

TABLE 3.
The Confusion Matrix Result

	Actual Positive	Actual Negative
Predicted Positive	907	171
Pedicted Negative	954	301

The table 3 shows the experiment result of prediction from total 2333 of tweets with 907 tweets is predict as it's actual positive, 171 predicted positive but actual negative, 954 predicted negative but actual positive, and 301 predicted negative as it's actual negative. In this process, the result of predicted negative with actual positive more higher than predicted positive with actual positive. We found that the problem caused of the Sundanese vocabulary in the tweets content is not exist in the vocabulary database that we've built. By calculating the prediction using rule-based classification, the results score obtained are presented in table 4 below.

TABLE 4.
The Result of prediction score

Measure	Score
Accuracy	0.517
Precision	0.841
Recall	0.487
F1 Score	0.617

In this research, prediction result for some feature has low score then previous research. However, we see the precision value in this research is 0.841 which can cover the positive

value for each predicted tweet. The n-grams feature especially unigram can predict language level in Sundanese language. In other case, bi-gram and tri-gram features in this study have not shown maximum results due to the limitations of existing data from the Twitter retrieve process. So, by applying the rule-based classification method for the experiments in this research was conducted predictions of 907 Sundanese tweets with their levels result (363 Ribaldry, 307 Standard, and 237 Polite) from a total of 2333 tweets that shown in Fig. 4.

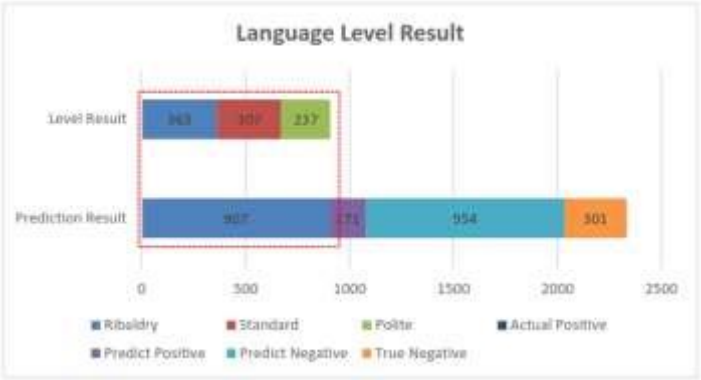

Fig. 4. *The Sundanese language level classification result*

Fig. 4 shows the results of the Sundanese language classification and its level which consists of two categories, namely prediction result, and level result. Prediction result states the Sundanese language and non-Sundanese language class. Level Result states the language level result for every tweets that predicted as Sundanese language. The dashed red line area states a total of Tweets that predicted as Sundanese language with a total of language level that taken from the prediction results of actual positive values.

4 CONCLUSION

The results of this research shows several concludes:

1. Sundanese language was capture as a second language which used by Twitter users who are located far away from the center of the capital city in the West Java and Banten provinces. It shows that the use of Sundanese language among the people still exists and also used and in social media with ribaldry level dominated. Due to the lack of the Latitude and Longitude feature, determining the classification of Sundanese language and their level for every region still faces obstacles and needs to be developed for further research.
2. Comparing the word level in Sundanese language was solved by a rule-based classification method with the result of Ribaldry level most widely used by the Twitter users then Standard level and Polite level.
3. Prediction score for several feature more smaller then previous research. But, we consider the precision value of the experimental results obtained score 0.841 which can be used to determine the predictive value close to the actual positive value.

5 RECOMMENDATION

Based on experimental results in this study, it highly recommends completing the Sundanese language vocabulary in further research. So, the process of comparing words in sentences more effective than the limit of vocabulary. For the future research, we will apply another method with focus in First Order Logic (FOL) to determine the relationship in the order of texts in Sundanese language form and combine with the statistical method such as Naive Bayes method to predict the independent relationship of vocabulary in a certain texts.

ACKNOWLEDGMENT

The authors wish to thank to Sekolah Tinggi Teknologi Garut for any supported during this research in Departement of Infrormatics program.

REFERENCES

[1] C. Sobarna, “Bahasa Sunda Sudah Di Ambang Pintu Kematiankah?,” Makara Hum. Behav. Stud. Asia, vol. 11, no. 1, p. 13, 2007.
[2] S. D. Budiwati and N. N. Setiawan, “Experiment on building Sundanese lexical database based on WordNet,” J. Phys. Conf. Ser., vol. 971, no. 1, 2018.
[3] E. Z. Arifin, “Bahasa Sunda Dialek Priangan,” Pujangga, vol. 2, no. 1, pp. 1–44, 2016.
[4] E. Karlieni, A. Hamid, and T. Prabasmoro, “The Role of Sundanese Language in Therapeutic Communication The Oncology Clininc RSHS,” Int. Semin. Lang. Maint. Shift, vol. 549, pp. 542–549, 2017.
[5] A. C. Juwita, A. Kalimah, B. Sunda, and D. Twitter, “Agustina Chandra Juwita, 2014 Adegan Kalimah Basa Sunda Dina Twitter Universitas Pendidikan Indonesia | repository.upi.edu | perpustakaan.upi.edu,” pp. 56–58, 2014.
[6] I. Baidillah et al., “Direktori Aksara Sunda untuk Unicode,” p. 131, 2008.
[7] A. Purwoko, “Model stemming berbasis kamus untuk berbasis kamus untuk dokumen berbahasa sunda,” pp. 1–74, 2011.
[8] T. F. Djajasudarma, D. Indira, and T. Muhtadin, “Fiksimini Berbahasa Sunda dalam Media Sosial (Sundanese Minifiction in Social Media),” J. Komunikasi, Malaysian J. Commun., vol. 34, no. 2, pp. 293–308, 2018.
[9] Maman Sumantri, Atjep Djamaludin, Achmad Patoni, R. H. Moch. Koerdie, M. O. Koesman, and Epa Sjafei Adisastra, Kamus Sunda - Indonesia. 1985.
[10] A. Al-Thubaity, M. Alhoshan, and I. Hazzaa, “Using Word N-Grams as Features in Arabic Text Classification,” 2015, pp. 35–43.
[11] A. R. Atmadja and A. Purwarianti, “Comparison on the rule based method and statistical based method on emotion classification for Indonesian Twitter text,” 2015 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2015 - Proc., no. January 2018, 2016.
[12] A. A. Budiman, “PENDETEKSI BAHASA DAERAH PADA TWITTER DENGAN MACHINE LEARNING,” 2018.
[13] B. E. Parilla-ferrer, P. L. F. Jr, and J. T. B. Iv, “Automatic Classification of Disaster-Related Tweets,” 2015.
[14] S. Hamidian and M. T. Diab, “Rumor Detection and Classification for Twitter Data,” 2019.
[15] M. Z. Asghar, A. Khan, A. Bibi, F. M. Kundi, and H. Ahmad, “Sentence-Level Emotion Detection Framework Using Rule-Based Classification,” Cognit. Comput., vol. 9, no. 6, pp. 868–894, 2017.
[16] G. Xiao, Q. Cheng, and C. Zhang, “Detecting Travel Modes Using Rule-Based Classification System and Gaussian Process Classifier,” IEEE Access, vol. 7, pp. 116741–116752, 2019.

[17] M. E. Karar, S. H. El-Khafif, and M. A. El-Brawany, “Automated Diagnosis of Heart Sounds Using Rule-Based Classification Tree,” J. Med. Syst., vol. 41, no. 4, 2017.
[18] Tan P-N, Steinbach M and Kumar V 2006 An introduction to data mining: solution Manual
[19] https://tanahair.indonesia.go.id/portal-web/inageoportal/#/webmapid=215da7bb-b69b-448b-b018-0c1f17127f65 Acessed: 2 February 2020
[20] https://simplemaps.com/data/id-cities. Acessed: 27 January 2020
[21] Mulyanah A. Republika Online. [Online]. 2013 [cited 2020. Available from: http://nasional.republika.co.id/berita/nasional/jawa-barat-nasional/13/08/26/ms4nkw-bahasa-sunda-terancam-punah.