Comparison of Naive Bayes Smoothing Methods for Twitter Sentiment Analysis

Rif'at Ahdi Ramadhani, Fatma Indriani, Dodon T. Nugrahadi
Faculty of Mathematics and Natural Science
Universitas Lambung Mangkurat
Banjarbaru, Indonesia
rifat.ramadhani@live.co.uk, f.indriani@unlam.ac.id, dodonturianto@unlam.ac.id

***Abstract*— In sentiment analysis, the absence of sample features in the training data will lead to misclassification. Smoothing is used to overcome this problem. Previous studies show that there are differences in performance obtained by the various smoothing techniques against various types of data. In this paper, we compare the performance of Naive Bayes smoothing methods in improving the performance of sentiment analysis of tweets. The results indicated that Laplace smoothing is superior to Dirichlet smoothing and Absolute Discounting with the micro-average value of F1-Score 0.7234 and macro-average F1-Score 0.7182.**

Keywords—Sentiment Analysis, Data Mining, Naive Bayes, Smoothing, Laplace, Dirichlet, Absolute Discounting

I. INTRODUCTION

Twitter, a text-based social media, has a significant amount of user-generated content, with 200 million tweets created daily, covering a variety of topics [1]. With this significant amount of data, it has the potential of research related to text mining. One that can be explored on social media data is opinion and sentiment analysis. Naive Bayes is a method that can be used for classification. In the process of text classification, the lack of features sample in training data will provide an absolute zero probability value, causing errors in the classification process. Smoothing methods are used to minimize the possibility of calculation errors on Naive Bayes if the features of test data did not occur on the training data.

Previous research [2] [3] indicated that smoothing methods for Naive Bayes give the different performance for text classification. Our contribution in this paper is the study of smoothing methods for sentiment analysis of Twitter which has not been done before. The organization of this paper is as follows. In section II, we survey related work in sentiment analysis and smoothing methods. In section III, we explain the methodology of the research. Section IV presents and discusses experimental results. Section V concludes the paper.

II. RELATED WORK

The previous study of sentiment analysis has been conducted by Pang et al. [4] using machine learning algorithms to classify a movie review. This research deals with the classification of related sentiment film review whether a film is rated positively or negatively. In the pre-processing stage, neither were performed stemming and elimination of stop words. The learning algorithms used were Naive Bayes, Maximum Entropy (MAXENT), and Support Vector Machine (SVM). The results indicated that Naive Bayes algorithm with unigram features obtained 78.7% accuracy and SVM with unigram features obtained 72.8% accuracy.

Later, Franky and Manurung [5] replicated the study with the data from a movie review on Pang et al. [4] by doing translating film reviews from English into Indonesian. Feature selection is made with some features include 2000 features unigram with the highest frequency of occurrence, 5000 unigram features with the highest frequency of occurrence, all unigram features, and 25 features unigram at the end of a movie review. The results were quite interesting demonstrated with the use of 25 feature unigram at the end of a movie review that gave an accuracy advantage compared to other feature selections. The values of accuracy obtained were 78.38% (Naive Bayes), 79.16% (Maximum Entropy), and 79.54% (SVM).

Recent research on Indonesian tweet sentiment analysis has been done in [6] with data collected from emoticons and national media accounts data. Other research on sentiment analysis is tweets about television shows [7] [8], "Kurikulum 2013" (*2013 Curriculum)* [9], the presidential candidate [10] [11], and public policy [12]. These research employ various machine learning algorithms including SVM, Naive Bayes, and C4.5 with various features.

Vohra and Teraiya [13] explained that in sentiment analysis, also, to apply machine learning algorithms such as Naive Bayes, Support Vector Machine, and C4.5 there were known other methods such as lexicon method for identifying the polarity of sentiment. Taboada et al. [14] applied the lexicon method by providing polarity value of the features of a particular word based on the type of position words (Part of Speech) in a sentence. The development of research related to sentiment analysis, a hybrid method or the incorporation of machine learning methods and lexicon method had also been developed as the study of Zhang et al. [15].

Chen and Goodman [16] previously studied smoothing methods for language modeling, including additive smoothing (Laplace/Lidstone), Good-Turing, Jelinek-Mercer, Katz smoothing, Witten-Bell, Absolute Discounting, and Kneser-Ney. For comparison evaluation, they use the measure cross-entropy which is a common metric for evaluating language

models. The best smoothing method according to this study is their modification of Kneser-Ney smoothing.

Smoothing methods have also been studied for language models applied to information retrieval [17], which investigated Jelinek-Mercer, Dirichlet, and Absolute Discounting smoothing methods. The JM method gave the best average performance in query retrieval. This paper also proposes a two-stage smoothing method which gives better performance than single smoothing.

Research on smoothing methods has also been done for text classification with Naive Bayes. For spam email classification, Hafilizara [2] compared Laplace, Jelinek-Mercer, Dirichlet, Absolute Discounting, and Two-Stage smoothing. The results show that Dirichlet smoothing method provided the best performance. For question topic classification, Yuan et al. [3] considered four smoothing methods for Naive Bayes: Jelinek-Mercer, Dirichlet, Absolute Discounting, and Two-Stage. Their result showed that Absolute Discounting and TS performed the best. However, smoothing methods have not been specifically studied in sentiment analysis, especially short and informal text such as Twitter. This paper focuses on comparing various smoothing methods for Twitter sentiment analysis, namely Laplace, Dirichlet, and Absolute Discounting methods.

III. METHODOLOGY

A. Data

The data for this study were obtained by using the Twitter API based on keyword searches in the form of words related to public facilities in Indonesia. Sample data used include the data with keywords such as "Bandara" (*Airport*), "Jalan" (*Road*), "Puskesmas" (*Health Center*), and "Rumah Sakit" (*Hospital*). 2000 data were collected for each used keyword. The obtained data are stored in CSV (Comma Separated Value) as contained in TABLE 1.

Next, each tweet is labeled manually according to the type of sentiment present (neutral, positive, negative, unknown/unclear). The unknown class is removed, and the rest underwent preprocessing steps: cleansing, handling of emoticon and emoji, conversion of slang terms, stemming, stop words deletion, handling of the negative word, case-folding, tokenization, and elimination of duplicate data (i.e. retweet). Changes in the data after the labeling and preprocessing stage are presented in TABLE 2.

TABLE 1. CSV STRUCTURE OF DATASET

No	Attribute	Description
1	Tweet ID	Identification Index of Tweet text (Primary ID)
2	User ID	Identification Index of Twitter user (Primary ID)
3	Screen Name	Username of Twitter user started by "@"
4	Profile Image URL	Profile picture of Twitter User
5	Geolocation (Latitude)	Latitude location
6	Geolocation (Longitude)	Longitude Location
7	Time	Timestamp of Tweet
8	Text*	Tweet text
9	Source	Application used by Twitter user

***the primary attribute used in the study**

TABLE 2. CHANGES IN DATASET (AFTER PREPROCESSING AND CLASS LABELING)

Keyword	Class (sum of data)				Total[a]	Total[b]
	Neutral	*Positive*	*Negative*	*Unknown*		
Bandara	523	385	188	127	1223	1096
Jalan	197	382	193	358	1130	772
Puskesmas	520	689	437	130	1776	1646
Rumah Sakit	395	469	467	248	1579	1331
Total	1635	1925	1285	863	5708	4845

[a] **Total of all sentiment class**

[b] **Total of all sentiment class (without unknown class)**

The final dataset consisted of 4845 tweets and divided into training data (80%), and test data (20%). The division conducted equally (stratified-sampling) for each class of sentiment data. Division of training data and test data conducted are presented in TABLE 3. 10-Fold Cross-Validation is used in hyper-parameter optimization of Dirichlet smoothing and Absolute Discounting smoothing by using training data.

TABLE 3. DISTRIBUTION OF TRAINING DATA AND TESTING DATA

Keyword	Class Distribution							
	Training				Testing			
	Neut	*Pos*	*Neg*	*Tot*	*Neut*	*Pos*	*Neg*	*Tot*
Bandara	418	308	38	876	105	77	150	220
Jalan	158	306	39	618	39	76	154	154
Puskesmas	416	551	87	1317	104	138	350	329
Rumah sakit	316	375	93	1065	79	94	374	266
Total	1308	1540	257	3876	327	385	1028	969

B. Experiment

The process of data mining in this study involved a comparison of the three smoothing methods: Laplace, Dirichlet (Dir), and Absolute Discounting (AD). Dirichlet and Absolute Discounting methods have a parameter each which need tuning to get the best classification performance (sentiment analysis).

First, experiments using 10-fold cross-validation on training data are conducted to get the optimum parameter for Dir and AD. The optimum parameter values could be obtained by applying a grid search of the range of values. The range of parameters specified for optimum value search on Dirichlet and Absolute Discounting methods are presented in TABLE 4.

TABLE 4. SEARCHING RANGE OF SMOOTHING PARAMETER

Smoothing	Parameter	Minimum Value	Maximum Value	Interval
Dirichlet	μ	1000	3000	100
Absolute Discounting	δ	0	1	0,05

Next, experiments are conducted on various sizes of training data to determine the effect of training size on classification performance. The percentage variation of the training data is presented in TABLE 5.

TABLE 5. VARIATIONS OF TRAINING DATA USAGE

No	Percentage	Number of Training Data
1	1%	39
2	5%	194
3	10%	388

4	20%	775
5	30%	1163
6	40%	1550
7	50%	1938
8	60%	2326
9	70%	2713
10	80%	3101
11	90%	3488
12	100%	3876

Finally, evaluation comparison between Laplace, Dirichlet, and Absolute Discounting was made using the values of accuracy, precision, recall, and F1-Score on the test data. The final assessment performance smoothing technique was based on the micro-average F1-Score and macro-average F1-Score.

IV. RESULT

A. Hyper-Parameter Optimization

Search limit values set for Dirichlet parameter was from 1000 to 3000 and had the interval of 100. Meanwhile, the search limit values set for the parameters Absolute Discounting was from 0 to 1 and had the interval of 0.05. Optimum parameters obtained in Dirichlet was 2500 and the optimum parameters obtained in Absolute Discounting was 0.95. The obtained result was presented in Fig 1 and Fig 2.

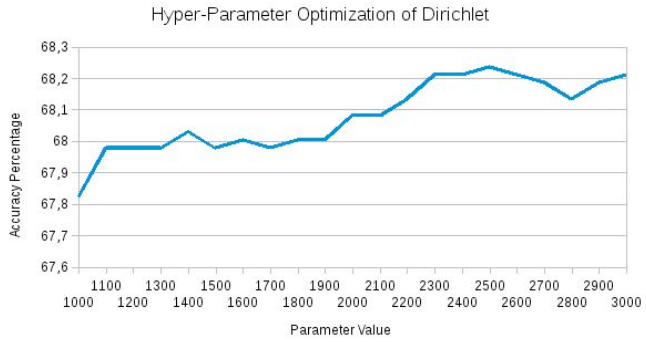

Fig 1. Hyper-Parameter Optimization of Dirichlet Smoothing

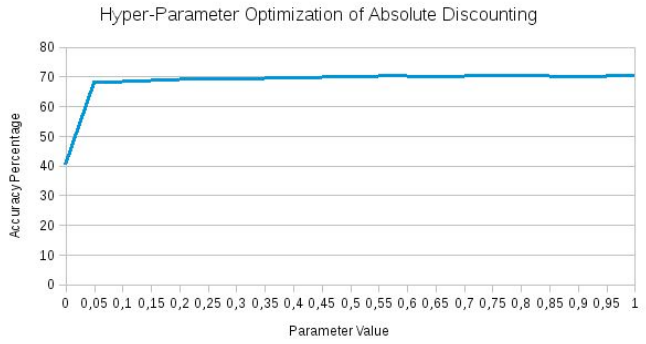

Fig 2. Hyper-Parameter Optimization of Absolute Discounting Smoothing

B. Effect of Varying Number of Dataset

Variations of the amount of the training data had been conducted to determine the effect of the amount of training data on the performance of the classification. Variations of the amount of the training data included training data as much as 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100%. The result of micro-averaged and macro-averaged F1-Score is presented in TABLE 6 and TABLE 7.

TABLE 6. VARIATION OF TRAINING SET AND MICRO-AVERAGE F1-SCORE

Training Set	**Micro-Average F1-Score**		
	Laplace	*Dirichlet*	*Absolute Discounting*
1%	0.5232	0.4582	0.5769
5%	0.6089	0.5604	0.5882
10%	0.646	0.5913	0.6233
20%	0.6347	0.5779	0.5986
30%	0.6481	0.5779	0.6089
40%	0.6161	0.5728	0.5666
50%	0.6677	0.6316	0.6594
60%	0.6873	0.6522	0.6605
70%	0.6708	0.6254	0.6316
80%	0.6914	0.6656	0.6729
90%	0.7224	0.6791	0.6966
100%	0.7234	0.6935	0.7007

TABLE 7. VARIATION OF TRAINING SET AND MACRO-AVERAGE F1-SCORE

Training Set	**Macro-Average F1-Score**		
	Laplace	*Dirichlet*	*Absolute Discounting*
1%	0.5535	0.4692	0.5955
5%	0.6197	0.5698	0.6043
10%	0.651	0.6055	0.6361
20%	0.6548	0.6137	0.6242
30%	0.6667	0.616	0.6365
40%	0.6434	0.6049	0.6048
50%	0.6705	0.6378	0.6622
60%	0.6876	0.657	0.6586
70%	0.6783	0.6424	0.6412
80%	0.6895	0.6699	0.6714
90%	0.7175	0.678	0.6926
100%	0.7184	0.6904	0.6963

Based on Fig 3 and Fig 4, it has been concluded that the amount of training data is influencing the performance of the classification. The more training data is used, the better the performance of the classification obtained. The amount of training data is one of the main factors affecting the performance of the classification.

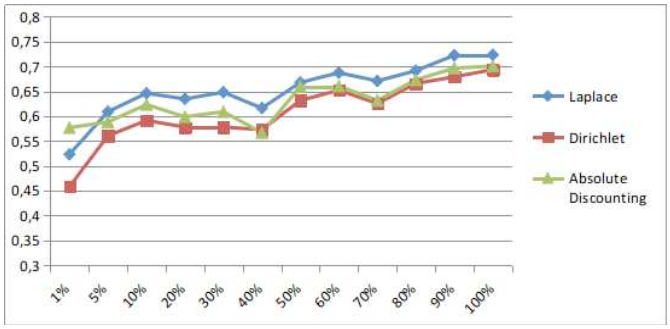

Fig 3. Variation of Training Set and Micro-Average F1-Score

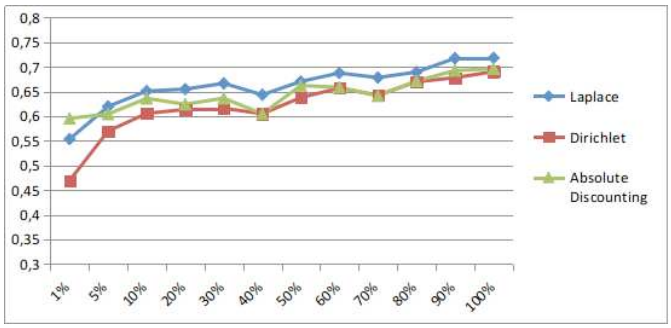

Fig 4. Variation of Training Set and Macro-Average F1-Score

C. *Evaluation and Comparison of Smoothing Performance*

Evaluation and comparison of smoothing performance were obtained from accuracy, precision, recall, and F1-Score metric. Various results were obtained for each sentiment class as presented in Fig 5 and Fig 6.

The best accuracy performance obtained in Laplace smoothing technique was 72.3426%. The best precision performance obtained in the neutral sentiment was Laplace smoothing with a precision value of 0.7218. The best precision performance obtained in the positive sentiment was Dirichlet smoothing technique with a precision value of 0.7666. The best precision performance obtained in the negative sentiment was Laplace smoothing with a precision value of 0.6996.

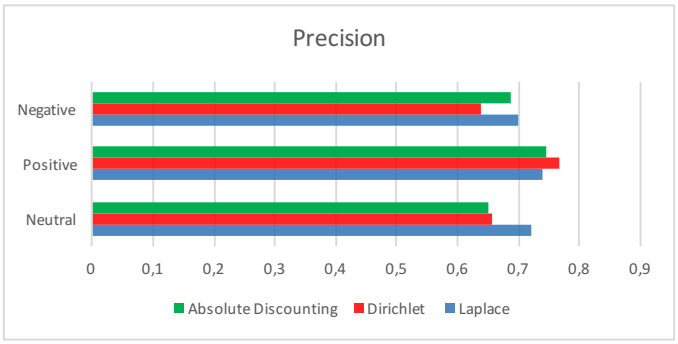

Fig 5. Precision Evaluation

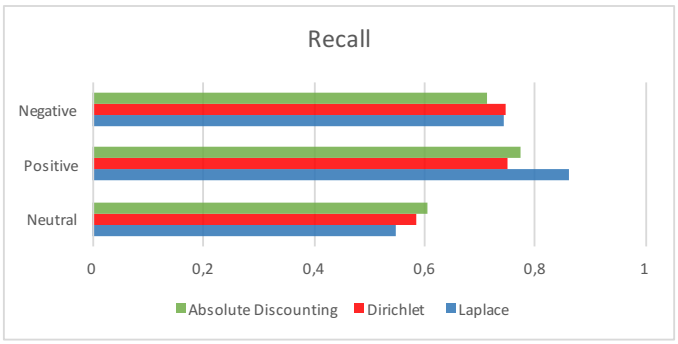

Fig 6. Recall Evaluation

The best recall performance obtained in the neutral sentiment was Absolute Discounting smoothing with a recall value of 0.6055. The best recall performance obtained in the positive sentiment was Laplace smoothing with a precision value of 0.8597. The best recall performance obtained in the negative sentiment was Dirichlet smoothing with a precision value of 0.7471.

The difference in performance was shown in the performance metric of precision and recall. The worst performance was shown in recall evaluation on neutral sentiment. To improve the accuracy, more feature to detect neutral text is required [18].

The final assessment of the performance of smoothing technique was performed by calculation and comparison of micro-averaged F1-Score and macro-averaged F1-Score. The best performance of micro-averaged F1-Score and macro-averaged F1-Score was obtained in Laplace smoothing with a value of 0.7234 and 0.7182 as presented in Fig 7.

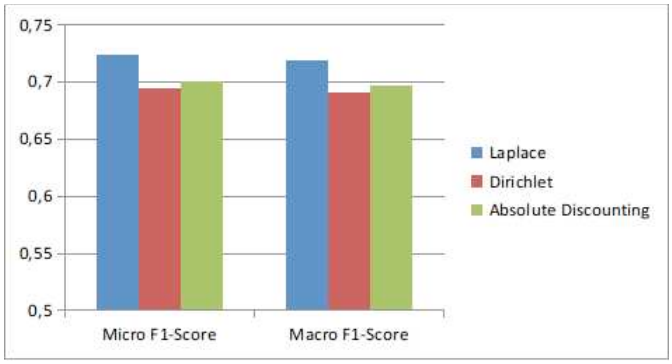

Fig 7. Comparison of Micro-Average and Macro-Average

Laplace performance was better than Dirichlet and Absolute Discounting. However, the difference of performance results can be seen in other researches conducted by Yuan et al. [3] and Hafilizara [2] where Laplace smoothing achieved lower performance when compared to other smoothing techniques. The difference in performance smoothing technique can occur due to factors of noise word features in the training data. Noise word features are words that do not strongly represent the characteristic of sentiment class resulting lower performance in classification.

V. CONCLUSION

The best performance of micro-average F1-Score and macro-average F1-Score were obtained in Laplace smoothing with a value of 0.7234 and 0.7182. The best accuracy performance obtained in Laplace smoothing technique was 72.3426%. There were some different results of precision and recall where the performance of precision and recall that have mixed results for each sentiment class.

The amount of training data could impact the performance classification. The more training data is used, the better the performance of the classification obtained. The amount of training data in this study was limited to 4845 from 8000 raw datasets after preprocessing had been done. The amount of data obtained in the data collection should be improved so that the amount of the final would be large enough that could improve the performance of classification.

Domain of sample data used in this study was limited to keywords related public facilities in Indonesia. Additional domains could be added by involving certain keywords so that

the characteristic pattern of sentiment analysis and the effect of smoothing on Twitter could be seen more widely.

At the stage of the preprocessing of data should also be done other phase such as feature selection (Chi-Square, Mutual Information, Information Gain, etc.) to reduce the dimensions of the resulting word on the dataset and reduce the noise so that the performance of classification could be improved. Hyper-parameter optimization should be done with alternative methods such as random search, Powell's, and Nelder-Mead that have a shorter computation time and efficient when compared with grid search.

ACKNOWLEDGMENT

The authors would like to thank Mr. Muliadi and Mr. Irwan Budiman for their valuable suggestions during this research. The research was contributed and supported by Universitas Lambung Mangkurat.

REFERENCES

[1] Twitter. 2011. 200 Million Tweets per Day https://blog.twitter.com/2011/200-million-tweets-per-day Accessed 2 November 2015

[2] Hafilizara, Mutia and J. Adisantoso. 2014. *Metode Smoothing dalam Naive Bayes untuk Klasifikasi Email Spam*. Bogor: Institut Pertanian Bogor. (in Bahasa Indonesia)

[3] Q. Yuan, G. Cong, and N. Thalmann, "Enhancing naive bayes with various smoothing methods for short text classification", *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, 2012.

[4] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?", *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, 2002.

[5] Franky and R. Manurung. Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews. In *Proceedings of the International Conference on Advanced Computational Intelligence and Its Application*, 2008.

[6] Aliandu, Paulina. Sentiment Analysis on Indonesian Tweet. *The Proceedings of The 7th ICTS*, Bali, May 15th-16th, 2013.

[7] Sentiaji, Aditia R and Adam M Bachtiar. Analisis Sentimen Terhadap Acara Televisi Berdasarkan Opini Publik. *Jurnal Ilmiah Komputer dan Informatika (KOMPUTA)*, ISSN : 2089-9033, 1-6, 2013.

[8] Tiara, M. Sabariah and V. Effendy, "Sentiment analysis on Twitter using the combination of lexicon-based and support vector machine for assessing the performance of a television program", *2015 3rd International Conference on Information and Communication Technology (ICoICT)*, 2015.

[9] Dyarsa, Pamungkas Singgih and Noor Ageng Setiyanto. 2015. "Analisis Sentiment Pada Sosial Media Twitter Menggunakan Naive Bayes Classifier Terhadap Kata Kunci "KURIKULUM 2013". Undergraduate Thesis. Faculty of Computer Science, Dian Nuswantoro University, Semarang. (in Bahasa Indonesia)

[10] H. Gemilang, A. Erwin and K. Eng, "Indonesian President candidates 2014 sentiment analysis by using Twitter data", *2014 International Conference on ICT For Smart Society (ICISS)*, 2014.

[11] M. Ibrahim, O. Abdillah, A. Wicaksono and M. Adriani, "Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation", *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015.

[12] D. Setyanugraha and A. Purwarianti, "Development of sentiment classification system for Indonesian public policy tweet", *2015 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2015.

[13] S. M. Vohra and J. B. Teraiya. A Comparative Study of Sentiment Analysis Techniques. *Journal of Information, Knowledge, And Research In Computer Engineering*, Volume – 02, Issue – 02, 313-317, 2013.

[14] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon-Based Methods for Sentiment Analysis", *Computational Linguistics*, vol. 37, no. 2, pp. 267-307, 2011.

[15] Khan, Aamera ZH, Mohammad Atique, and V. M. Thakare. "Combining lexicon-based and learning-based methods for Twitter sentiment analysis." *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)* (2015): 89.

[16] S. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling", *Computer Speech & Language*, vol. 13, no. 4, pp. 359-393, 1999.

[17] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval", ACM Transactions on Information Systems, vol. 22, no. 2, pp. 179-214, 2004.

[18] Lunando, Edwin, and Ayu Purwarianti. "Indonesian social media sentiment analysis with sarcasm detection." Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on IEEE, 2013.