Polarity of Indonesian Regional Election (Pilkada) 2020 Related Tweets

R Latifah*, W Rahmawati, N Rosanti, Y Adharani, P Meilina, A Mutholib, N Amri

Informatics Engineering Study Program, Universitas Muhammadiyah Jakarta, Jakarta, Indonesia

*retnani.latifah@umj.ac.id

Abstract. Simultaneous regional election is a hot topic in 2020 in which a lot of people voiced their opinion through social media, including Twitter. Not all regional election related tweets that were posted has positive sentiment, some were negative, and some were neutral. Sentiment analysis could be used to learn about the polarity of those tweets. This paper did a sentiment analysis of regional election related tweets from six cities by using several classification methods SVM, kNN, LR and MNB. The methods were used in labelled dataset which automatically labelled using Indonesian Sentiment Lexicon (InSet), resulting in 49% of tweets are negative, 42% are positive and 9% are neutral. The best classifier is SVM with 78% of F1 score and 79% of accuracy. However, this classifier along with LR suffered an overfitting because there are still unnecessary features that are used in building the model. Reducing the dimension manage to resolve the overfitting problem for SVM and LR to some extent although the performance is down to around 50%.

1 Introduction

These days, it is common to see a discussion about certain issues in social media such as Facebook, Twitter, Instagram, Youtube etc. Whenever there is a happening, either in domestic or international, people often voiced and discussed their thought through those sites. Furthermore, most of news provider have turn to social media to increase their site's engagement, resulted in easier access for people to learn about what happen in the world [1]. However, this discussion sometime went in a bad direction.

In Islam, we are taught to speak nicely otherwise we better keep silent (HR. Bukhari). This should be implemented not only in real life interaction, but also in social media. However, because social media is a free platform and borderless, sometime people could not restrain themselves and tweet in negative words, either directly or indirectly. It could lead in a debated which sometime could hurt each other feelings and even affecting mental health [2].

Sentiment analysis is a computational research about classifying textual contents according to their tendency, whether it is positive, negative or neutral [3][4][5]. There are many researches about sentiment analysis and in recent days most of its data source come from social media. This is due to the huge number of users using social media to speak about what is in their mind daily. As of January 2021, there are 4.2 billion active users in social media including Twitter [6].

Twitter is a popular microblogging where people could post their thought in less than 280 characters. As of Januari 2021, Indonesia rank 6 in the country with the most Twitter users with around 14 million users [7]. With this huge number of users, there are many tweets coming from Indonesian and certain topics often get trended worldwide. They often speak about political affairs, entertainment, global issues and other things.

Simultaneous regional election (Pilkada 2020) is a topic that often being discussed and debated in Twitter in 2020. It was held on 9 December 2020, yet months before the election people already talked about it in Twitter. Even after the election day, it was still a hot issue and there are various hashtags accompanying the topic each day. Like other topics, people also talk about this election in both positive and negative ways.

In order to learn about the sentiment of regional election related tweets during and after election day, this study has collected several tweets as initial corpus dataset. This paper presents the process of dataset collection and its automatic labelling using Indonesian Sentiment Lexicon (inSet). It is also addressed the polarity of the dataset using several classification methods and how reducing the dimension could managed the overfitting to certain extent with some limitation.

2 Related work

Sentiment analysis is a topic that has been widely research because the number of data and subjects to be analyzed keep increasing each day. One topic that often being research in sentiment analysis is about election. A research about Jakarta Election in 2017 was conducted using the combination of textual tweets and emoji detection using Multinomial Naive Bayes, which resulted in the improvement of the accuracy [1]. The dataset used in this study focus on each candidates of the election and only consist of opinionated tweet about the candidates. The dataset was manually labelled.

Another research about electoral topic has also been done in 2018 about East Java Governor Electoral. This research uses Naive Bayes classifier, while there is no notable new technique in this paper, the performance of the classifier is good, more than 90% of F1 score, and could concluded that the first candidate got more attention than the second one in Twitter which in line with the actual result [8]. This paper did not provide the labelling process of the dataset; thus, we could not determine whether it is manually or automatically labelled dataset.

In 2019, Indonesia held its presidential election. A sentiment analysis about the two president candidates was conducted using Facebook fanpage as its data source [5]. This research used automatic labelling by calculating the sentiment score of each post. If there is a positive word then add one, if negative, then subtract with one. The Naive bayes classifier has 74.6% F1 score. This is might due to the data correction they did before pre-processing the data.

3 Method

3.1 Dataset Collection and Preparation

In Twitter sentiment analysis, we need to collect the data by crawling before doing pre-processing and cleaning [8]. The data in this research were collected by crawling Twitter's posts (tweets) using 'pilkada' (regional head election) as keyword. The data are collected from December 9th, 2020 until December 17th, 2020 and got around 25000 tweets. The data then being cleaned by deleting duplicate tweets and only using the tweets contained one of six areas which held the election, "Medan", "Manado", "Surakarta / Solo", "Semarang", and "Surabaya". These cities were chosen because the areas are considered as big cities and a lot of people took interest in the result. We used information retrieval with vector space model to retrieve the tweets from each area by using the name of the cities as query. The total tweets from each area and after deleting the duplicate is shown in table 1.

Table 1. Total tweets for dataset

Areas	Retrieved Tweets	Post duplicate deletion
Medan	5498	1860
Semarang	627	230
Surakarta/Solo	5591	2173
Surabaya	1500	864
Makasar	994	502
Manado	2845	1588
Total	17055	7217

In order to train the data with a classifier, it is required to label the tweets with sentiment's type. The labelling was done automatically using sentiment score [5]. Instead of adding or subtracting the score with one every time positive or negative words appear, we used the Indonesian sentiment lexicon by Koto and Gemala (2017) to calculate the score [9]. It has 3609 positive words and 6609 negative words with each words have score between -5 and +5. The sentiment score will be calculated using formula (1). If the score is more than one then it will be assigned as positive, if less than one is negative and if zero, it is neutral [5]. Example of how it works can be seen in table 2.

$$score = \sum positiveword - \sum negativeword \tag{1}$$

Table 2. Example of labelling process

Tweet	**Positive words**	**Negative words**	**Total score**

Sekedar referensi pilkada besok. Untuk Tangsel dan Surabaya coblos paslon no 1. Tetap jaga jarak dan pakai masker ya!	tetap = 3
	jaga = 2
	jarak = 3
		Jaga = -3
		jarak = -3
			8 – 6 = 2 , positive

Solo uda ga usah pakek pilkada udah tau siapa yang menang. Agak kesel jg sih....	solo = 2
	siapa = 1
	menang = 4
		tau = -4
		siapa = -2
		kesel = -5
			7 – 11 = -4, negative

3.2 Pre-processing Data

The labelled dataset underwent a pre-processing before it can be classified using classification models. Depend on the text pre-processing techniques that are used, the performance of the models could differ. Some research gest a better result when some pre-processing techniques are used, while sometime complex techniques also resulted in bad evaluation [10].

In text mining, some methods that could be used as in pre-processing are cleansing, case folding, filtering, tokenizing, stemming and extracting features [5][11][2][12]. Cleansing is a process to remove punctuations and unwanted characters. In this step, we also removed digit, http links, mentions (@username) and retweet indication (RT). Case folding is changing the capital letters into lowercase and filtering is removing common words called stopwords. The list of stopwords used in this research is from Jelita Asian (2007) [13] taken from Adikara (2012) website [14].

Stemming is a method to change the word into its root. This research used sastrawi python library to do the stemming process which has the rules from Nazief and Adriani stemmer, Confix Stripping Stemmer, Enhance Confix Stripping Stepper and its modification [15]. The stemmed tweets then breaks into tokens. If the token only has less than 3 characters, it will be removed. Table 3 showed the result from cleansing to tokenizing.

Table 3. Example of Each Pre-processing Step

Pre-processing	Tweet	Result
Cleansing and case folding	pilkada di Solo dan Medan lah yang menjadi perhatian publik saat ini. \n\nMeskipun orang sudah tahu hasil nya terutama di Solo, Medan mungkin masih ada harapan. \n\n#JangantakutkeTPS katanya	pilkada di solo dan medan lah yang menjadi perhatian publik saat ini n nmeskipun orang sudah tahu hasil nya terutama di solo medan mungkin masih ada harapan n n jangantakutketps katanya
Stopword removal	pilkada di solo dan medan lah yang menjadi perhatian publik saat ini n nmeskipun orang sudah tahu hasil nya terutama di solo medan mungkin masih ada harapan n n jangantakutketps katanya	pilkada solo medan perhatian publik n nmeskipun orang tahu hasil nya solo medan harapan n n jangantakutketps
Stemming	pilkada solo medan perhatian publik n nmeskipun orang tahu hasil nya solo medan harapan n n jangantakutketps	pilkada solo medan perhati publik nmeskipun orang hasil nya solo medan harap n n jangantakutketps
Tokenization	pilkada solo medan perhati publik nmeskipun orang hasil nya solo medan harap n n jangantakutketps	“pilkada” “solo” “medan” “perhati” “publik” “nmeskipun” “orang” “hasil” “nya” “solo” “medan” “harap” “jangantakutketps”

Each token from each tweet is represented as TF-IDF weight, which is the most common weighting scheme in text mining and information retrieval [16]. This weight calculated the product of tokens’ frequency in each tweet (TF) and the invers of the number of tweets which has certain token (IDF) and stored it in a term document matrix. Each tweet represented by a vector of its token’s TF-IDF.

3.3 Classification

This study used TfidfVectorizer in NLTK modul [17] to generate the vectors from cleaned datasets. The data were randomly split with 80% training set and 20% testing, then modelled using several classification methods, kNN, Multinomial Naive Bayes (MNB), Support Vector Machine (SVM) and Logistic Regression (LR). We used RBF kernel for SVM because it has better performance in preliminary testing across the dataset. The performance of each methods was evaluated using precision, recall, f1 measure and accuracy.

The goal of this research is to learn about the polarity of tweets related to regional election in six areas. Thus, we separated the data according to its area and classifying them with several classifiers. We also classify the whole dataset and compared the result.

4 Results and discussion

4.1. Automatic Labelling

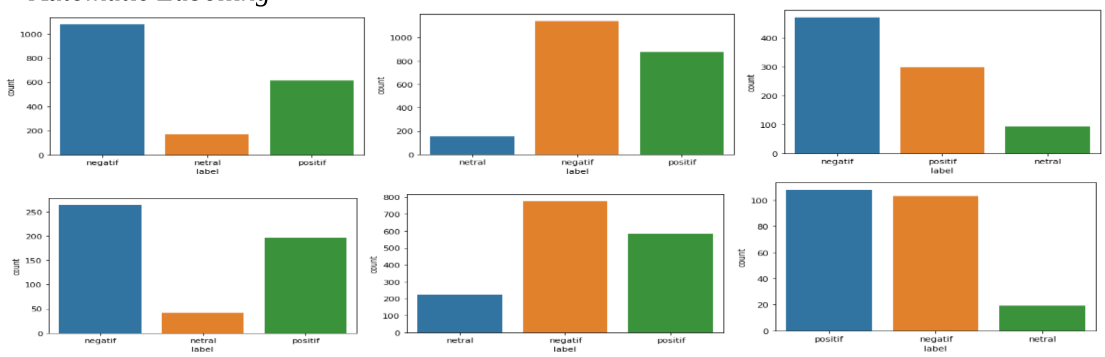

Figure 1. Polarity of regional election related tweets. From top left to bottom right, Surabaya (negative, positive, neutral), Solo (neutral, negative, positive), Medan (negative, neutral, positive), Manado (neutral, negative, positive), Makasar (negative, neutral, positive), Semarang (positive, negative, neutral).

The labelled dataset has 704 neutral tweets, 2677 positive tweets and 3836 negative tweets with the details for each area are in figure 1. Tweet with negative words take up about 49% and almost all areas have more negatives tweets compared to positive and neutral. There are only 9% of neutral tweets, this is due to the use of sentiment strength from Indonesian sentiment lexicon. If we used equal strength (adding and subtracting with one) for each positive and negative words, there might be more neutral tweets because most of the tweets rarely use the sentiment words explicitly. It is more about the nuance of the tweets and sometime sarcasm, which is difficult to capture.

The dataset with Medan, Solo, Surabaya, Makasar and Semarang has relevant words about Pilkada in its wordcloud including the head candidates. In Medan and Solo in particular, the name of Indonesian president is often included in the tweets due to his familial relationship with the two candidacies from the two cities. However, in Manado dataset, there are many unrelated words especially in negative tweets. It seems like some people used the election hype to include the word 'Pilkada' in their post, although it is completely unrelated.

4.2. Classification Result

The performance of each classifiers shown in figure 2. The best classifier is SVM with 79% accuracy and recall when used whole dataset. Compare to other classifiers, logistic regression is the one with the closest result to SVM, with 75%. It followed by multinomial naive bayes with 68%. kNN has the worst result with only 33% of accuracy and recall although the precision is decent (73%). Overall, all classifiers have good precision which are around 73% to 78%. It means that when classifiers predict a tweet that has positive sentiment, it is correct 73% to 78% of the time. Meanwhile, when kNN only has 33% recall, it means that kNN could only correctly identify 33% out of all tweets with positive sentiments. It is different compared to SVM, which is able to identify 79% of tweets with positive sentiments.

F1-score and accuracy for SVM, LR, kNN and MNB, respectively are 78% and 79%, 73% and 75%, 39% and 33%, and 64% and 68%. F1-score is a common measure to learn about the accuracy of a classifier in high imbalanced dataset. However, from the result it can be seen that there is only a little difference between the two measurements. This is maybe because the dataset is not imbalance enough.

From the figure, it shows that the number of tweets affected the performance. The performance for Semarang, Makasar and Surabaya dataset are generally lower compared to Medan, Solo and Manado dataset. Medan dataset has the second-best evaluation after whole dataset. Overall, the difference between whole dataset and Medan dataset is only less than 3%. This means that the other dataset did not have significant contribution in improving the performance.

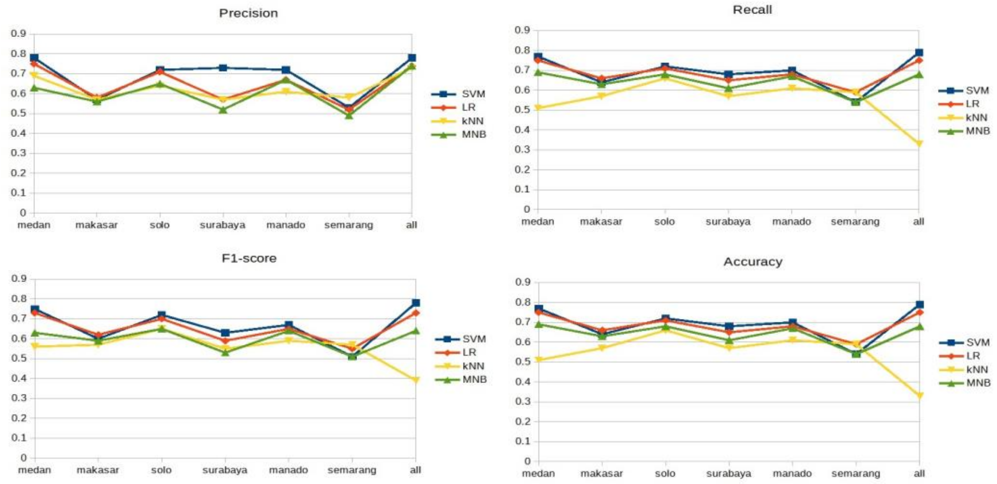

Figure 2. Performance of Classifiers

The performance of SVM and LR can be said as decent with more than 70% of F1-measure. However, we found that the model is overfitting as seen in figure 3. Training set has high-performance score compared to test set, which means the model was built too close to the training set and less generalized. There are unnecessary features that are still used which make the model become complex. This is maybe because the data itself came from social media where people tend to use abbreviate and misspelling the words which have not been addressed in this research.

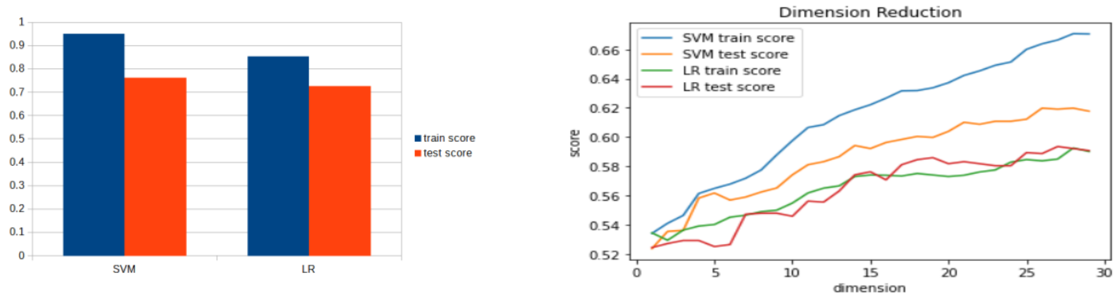

Figure 3. (left) Overfitting in SVM and LR classifier. (right) Dimensionality reduction to resolve overfitting.

Reducing the dimension of the vectors using SVD is able in overcoming overfitting despite the low performance, which is down to around 50%. From figure 3, for SVM reducing the dimension to below 5 has resolve the overfitting problem. Although as we raised the dimension, the sign of overfitting appears again. Meanwhile for LR, the lower dimension managed to resolve the overfitting issue up to lowering into 30 dimensions. This means that SVD can be used in addressing overfitting problem for this dataset.

5 Conclusion

This research collected regional election related tweets during and after the election day from six cities, Medan, Solo, Semarang, Makasar, Manado and Surabaya, with around 7217 tweets. From automatic labelling using InSet Lexicon, we found that 49% of the tweets are negative, 42% are positive and 9% are neutral. Except for Manado dataset, other areas dataset has more words that are related to regional election. This paper implemented four classifiers, SVM, LR, kNN and MNB which resulted in SVM has the best performance with 79% of accuracy followed by LR with 75%, kNN 33% and MNB 68%. However, SVM and LR suffered an overfitting with the difference between training and testing score is almost 20%. To overcome this problem, this research apply SVD as reduction dimension techniques and found that it is able to address the issue to certain extent although the performance has fall into around 50%.

Acknowledgments

We hereby thank you to Program Studi Teknik Informatika Universitas Muhammadiyah Jakarta for Supporting the publication of this research.

References

[1] Lestari A R T, Perdana R S and Fauzi M A 2017 Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji *J. Pengemb. Teknol. Inf. dan Ilmu Komput.* **1** 1718–24

[2] Razak Z I, Abdul-rahman S, Mutalib S and Abdul Hamid nurzeatul hamimah 2018 Web mining in classifying youth emotions *Malaysian J. Comput.* **3** 1–11

[3] Budiharto W and Meiliana M 2018 Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis *J. Big Data* **5** 1–10

[4] Fatyanosa T N and Bachtiar F A 2018 Classification method comparison on Indonesian social media sentiment analysis *Proc. - 2017 Int. Conf. Sustain. Inf. Eng. Technol. SIET 2017* **2018-January** 310–5

[5] Santoso E B and Nugroho A 2019 Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik Di Facebook *Eksplora Inform.* **9** 60–9

[6] Kemp S 2021 Digital 2021: Global Overview Report — DataReportal – Global Digital Insights

[7] Clement J 2021 • Twitter: most users by country | Statista *Statista*

[8] Hakimi F D D, Hamdi A Z, Ulinnuha N, Asyhar A H and Farida Y 2018 Analysis of Public Sentiment towards East Java Governor Election 2018 on Twitter using Text Mining *Proceeding Built Environ. Sci. Technol. Int. Conf. (BEST ICON 2018)* 262–7

[9] Koto F and Rahmaningtyas G Y 2018 Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs *Proc. 2017 Int. Conf. Asian Lang. Process. IALP 2017* **2018-January** 391–4

[10] Zin H M, Mustapha N, Murad M A A and Sharef N M 2017 The effects of pre-processing strategies in sentiment analysis of online movie reviews *AIP Conf. Proc.* **1891**

[11] Wicaksono A F, Vania C, Distiawan B T and Adriani M 2014 Automatically building a corpus for sentiment analysis on Indonesian tweets *Proc. 28th Pacific Asia Conf. Lang. Inf. Comput. PACLIC 2014* 185–94

[12] Camacho-Collados J and Pilehvar M T 2018 On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis *arXiv*

[13] Jelita A 2007 *Effective Techniques for Indonesian Text Retrieval*

[14] Adikara P P Kamus Kata Dasar dan Stopword List Bahasa Indonesia

[15] Rosid M A, Fitrani A S, Astutik I R I, Mulloh N I and Gozali H A 2020 Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi *IOP Conf. Ser. Mater. Sci. Eng.* **874**

[16] Markpoulos G, Mikros G, Iliadi A and Liontos M 2015 Sentiment Analysis of Hotel Reviews in Greek: A Comparison of Unigram Features *Cult. Tour. a Digit. Era, Springer Proc. Bus. Econ.* **9** 373–83

[17] Bird S, Klein E and Loper E 2009 *Natural Language Processing with Python* (O'Reilly Media Inc)