Determination of quality television programmes based on sentiment analysis on Twitter

IOP Conf. Series: Journal of Physics: Conf. Series **978** (2018) 012117 doi:10.1088/1742-6596/978/1/012117

Determination of quality television programmes based on sentiment analysis on Twitter

A Amalia[1], W Oktinas[2], I Aulia[2] and R F Rahmat[2]

[1]Departement of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia.
[2]Departement of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

*Email: amalia@usu.ac.id

Abstract. Public sentiment from social media like Twitter can be used as one of the indicators to determine the quality of TV Programmes. In this study, we implemented information extraction on *Twitter* by using sentiment analysis method to assess the quality of TV Programmes. The first stage of this study is *pre-processing* which consists of *cleansing, case folding, tokenizing, stop-word removal, stemming*, and *redundancy filtering.* The next stage is weighting process for every single word by using TF-IDF method. The last step of this study is the sentiment classification process which is divided into three sentiment category which is positive, negative and neutral. We classify the TV programmes into several categories such as news, children, or films/soap operas. We implemented *an improved k-nearest neighbor* method in classification 4000 twitter status, for four biggest TV stations in Indonesia, with ratio 70% data for training and 30% of data for the testing process. The result obtained from this research generated the highest accuracy with k=10 as big as 90%.

1. Introduction

Television Industries in Indonesia is growing so fast. It can be seen by many television stations established in Indonesia recently. The quantities of TV stations also affected with increasingly of TV programmes that can be classified into some categories like news, movie/soap opera, kids and else. The variety and quality of TV programmes often will be discussed by many people in social media like *Twitter*. *Twitter* as one of microblogging with more than 500 million account users and can generates 400 million tweets a day [1]. On *Twitter*, people can reveal his/her opinion mining including to assess the quality of TV programmes on TV stations. These set of opinions can be a consideration to TV stations to improve the quality of their TV programmes. Opinion mining or sentiment analysis is a computational study from people's opinion, sentiment and emotion that are expressed in the text [2] [3]. Sentiment analysis will classify polarity of text to mine whether the opinion is positive, negative or neutral opinion about something [4]. Recently, many researchers utilize *Twitter* to reveal society opinion about something because *Twitter* provides *Twitter* API to facilitate crawling data in large quantities [5]. Sentiment analysis is a classification process, but classification process on twitter data is more challenging than document classification because Twitter data contains unstructured text. The first stage of twitter classification is to classify whether the status is an opinion or not [2]. The next step is to organize the view into positives or negative sentiment. This study aims to determine the quality of TV programmes based on public opinion on Twitter. We implemented an Improved K-

Nearest Neighbor method for classification public opinion based on Twitter status in Bahasa Indonesia. *Improved k-Nearest Neighbor* is a modification of k-Nearest neighbor method. Improved k-Nearest Neighbor modify the k-values determination, for each category has various values. The diversity of k-values for each category is adjusted with training document in the category. The purpose of this modification is a to improve the k-Nearest Neighbor method that has issued in accuracy because all class has same k-values without considering the number of training document for each category. Meanwhile, distribution of training documents in training stage has not had same quantity.

2. Related Works

Many previous studies related to sentiment analysis and classification methods. Studies that utilize classification methods had been done by [6]. Earlier studies like marketing and social study utilize Twitter because Twitter can be used efficiently [7]. Study of public sentiment analysis about government policy had been done by [8], the study use comparison of some classification methods like k-Nearest Neighbor, Naïve Bayes and Support Vector Machine. The result of this study shows that Support Vector Machine raised a better accuracy which is 86% [9]. Another sentiment analysis study had been done by [10], this study implemented Naïve Bayes and unigram feature to observe sentiment analysis on President election in the United States. This study gets 59% accuracy [9]. Another sentiment analysis study had been done by [5], this study observed customer satisfaction based on emoticon on Twitter. This study compares three learning methods which are naïve Bayes, support vector machine and maximum entropy. This study use unigram feature and bigram, the research conducted the accuracy increase for Bigram feature for Naïve Bayes and maximum entropy. Sentiment analysis for public opinion in President election in India had been done by [11], this study use combination of *Naïve Bayes, Hidden Markov Model, and SentiWordNet*. This study gets 71.48% of accuracy.

3. Methodology

The Methodology in determining TV Programmes' quality based on opinion analysis can be described in Figure 1. This process contains four big stages. The stages are Crawling data, Pre-processing Stage, word weighting and classification stages.

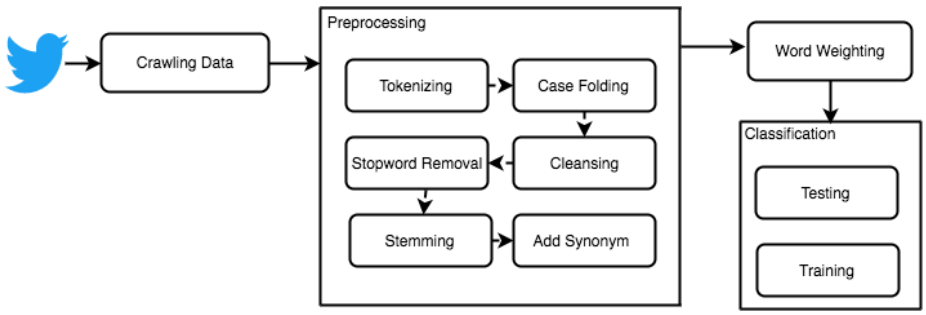

Figure 1. Methodology.

3.1. Crawling Data

In this study, we observe public opinion in Bahasa Indonesia on twitter to four biggest TV station in Indonesia. We crawl all tweet-mention to the official account twitter of these five TV stations through search API. These accounts are @officialRCTI, @whatsonANTV, @GlobalTVseru and @official_MNCTV. We got 4000 data or 1000 twitter status for each account mention.

3.2. Preprocessing

Preprocessing in this study is a process to prepare the data or tweets into 'a-bag-of-words' form. The purpose of this conversion is to reduce data dimension but still preserve the meaning of the tweets. We also eliminate a redundant data and fix inconsistent data from the dataset. This process contains tokenizing, case folding, cleansing, stopword removal, stemming and filter redundant tweets.

Tokenizing is a process to break the tweets into word by word. The next step is case folding process. This process is a process to convert these set of words into lower case. The aim of this action to convert all characters in the same way. The next process is Cleansing, which is a process to remove noise from tweets like hashtag(#), username(@username, URL address, emoticon and email address. The next stage is stopword removal. This stage is a process to eliminate un-important words like conjunctions, prepositions, and pronouns. We implemented Tala dictionary as a reference to un-important words. Tala dictionary is a list of words in Bahasa Indonesia that are not important (Tala 2013). However, we have not eliminated all words as Tala reference, because some words have essential meaning in opinion analysis, particularly for negative meaning. The next step is stemming, stemming is a process to convert a word into its basic words or remove the affix of the words. We implemented Nazief and adriani stemming algorithm. The last stage of preprocessing is adding synonym functions. This process is an action to check synonym of the words in 'a-bag-of-words' for each tweet. Some of these synonym words are slang or abbreviation words. Twitter is social media that only has 140 characters limitation for each tweet. This limitation makes some people used abbreviation words, for example, people use "ga" instead of "enggak" to represent "not". The additional of this synonym words will optimize frequency of occurrences calculation for words which have the same meaning.

3.3. Word Weighting

Words weighting is a process to measure or evaluate how important a word is to a document. We implemented TF-IDF measurement. TF-IDF or Term Frequency-Inverse Document Frequency is TF value multiply to IDF value. TF is a measurement of word quantity occurrences in a document. Meanwhile, IDF is a process to calculate the word weighting based on documents quantities which contain a certain word. The greater TF-IDF value of a word weighting calculation the more important that word is. The equation of TF-IDF can be seen in equation 1 and 2.

$$IDF\ (w) = \log\left(\frac{N}{DF(w)}\right) \quad (1)$$

$$\text{TF -IDF (w,d)} = \text{TF (w,d)} \times \text{IDF(W)} \quad (2)$$

Where :

TF-IDF (w,d)	: weighting of a word in a whole document
Tf(w,d)	: Frequency occurrences of a word w in a document
IDF(w)	: inverse DF from a word
N	: Numbers of documents
DF (w)	: Numbers of documents containing word w

3.4. Classification

The next step is classification process. We implanted Improved K-Nearest Neighbor algorithm for classification. These steps contains training stage and testing stage. We divided dataset to training and testing process with ratio 7 : 3. The quantities of the divided dataset for each TV station can be seen in Table 1.

Table 1. Training Data Proportion

TV Station	Quantity	
	Training Data	Testing Data
ANTV	700	300
Global TV	700	300
RCTI	700	300
MNCTV	700	300

On training stage, 70% of the dataset is classified or labelled manually as positive, negative or neutral opinion. These classifications that will be used to create a model of sentiment analysis. 30% of the rest of data will be classified based on the model that got from training step. We implement an *improved k-nearest neighbor algorithm.* The first step, we calculate the similarity of two documents with cosine similarity method. Cosine similarity is a document vector (D) measurement with query vector (Q). The more similar D to Q, the more appropriate the document with the query. Cosine similarity equation can be seen in equation 3.

$$cosSim(X, dj) = \frac{\sum_i^m = 1\ xi.dji}{\sqrt{(\sum_i^m = 1\ xi)^2}.\sqrt{(\sum_i^m = 1\ dji)^2}} \quad (3)$$

Where X is testing document, dj is training document, xi and dji is word weighting value.
We classify the testing document with *Improved k-Nearest Neighbors algorithm*. We calculate similarity comparison of training data and testing data for each category and determine the category based on the best probability. Equation 4 shows the formula of this method.

$$p(x, c_m) = argmax_m \frac{\sum_{dj \in top\ n\ knn(cm)} sim(x,d_j) y(dj,c_m)}{\sum_{dj \in\ top\ n\ k\ NN(cm)} sim(x,d_j)} \quad (4)$$

where:
p(x,cm) : probability document X in category of cm
sim(x,dj) : similarity of document X with training document dj
top n kNN : top n neighbour
y(dj.cm) : attribute function from suitable category

we tried ten times trials, to find out the influence of training set proportion for each category and k value to classification effectiveness.

Table 2. Training Data Proportion

Training Data Proportion			
Positive	Negative	Neutral	Quantity
1385	700	724	2800

Table 2. Show differences proportion of training data quantity that is used in the classification process. We tried various of k values for 1200 test data. Average accuracy value for 10 times trial can be seen in Table 3. From 2800 of training data with proportion quantity 1385 positives, 700 negatives and 724 neutral. From table 3 show that the highest accuracy is 90% for k = 10 and the lowest accuracy is 85% for k = 60.

Table 3. Testing System Based On K Values

k	n (new *k-values*)			Accuracy
	Positive	Negative	Neutral	
5	2	1	1	89%
10	5	2	2	90%
15	7	4	4	88%
20	10	5	5	88%
25	12	6	6	88%
30	15	8	8	87%
35	17	9	9	87%
40	20	10	10	87%
50	25	13	13	86%
60	30	15	15	85%

4. Result and Discussion

Based on the previous stage, we conclude that the best accuracy obtained for k =10. The excerpt of the result of the experiments for 1200 data testing with k = 10 can be seen in Table 4.

Table. 4 System Test Result

No	Training Data	Classification	
		System	Manual
1.	@whatsonANTV wajib di tonton Karena Alur ceritanya pasti bikin Baper dan selalu bikin penasaran, Jadi selalu tak in🔲 https://t.co/Fql8XWcyYv	Positif	Positif
2.	Arigatou buat yang udah nonton @merlynasun di @LensorANTV edisi pagi ini, wassalamu'alaikum dan BRAVO OLAHRAGA!!! https://t.co/wiQv6VHCSP	Positif	Positif
…	…	..	..
1200	Mantap jiwa dubbingnya keren. Walau tayangnya malem kalo nonton Jodha Akbar mata tetep melek @Official_MNCTV #JodhaAkbarMNCTV	Positif	Positif

Based on this study, from 1200 data testing, the system can classify 1080 opinion mining categorization correctly. Accuracy classification of this study can be calculated :

$$Accuracy = \frac{1080}{1200} \text{ x } 100\% = 90\%$$

Based on this study we can conclude accuracy of opinion classification using *Improved K-Nearest Neighbor* is 90%.

We also classify the favourite TV stations and TV programmes Category based on positives opinion. The quality of TV station and TV programmes can be determined by comparing the quantities of positives sentiment analysis is higher than negatives and neutral view. The result can be described in Figure 2.

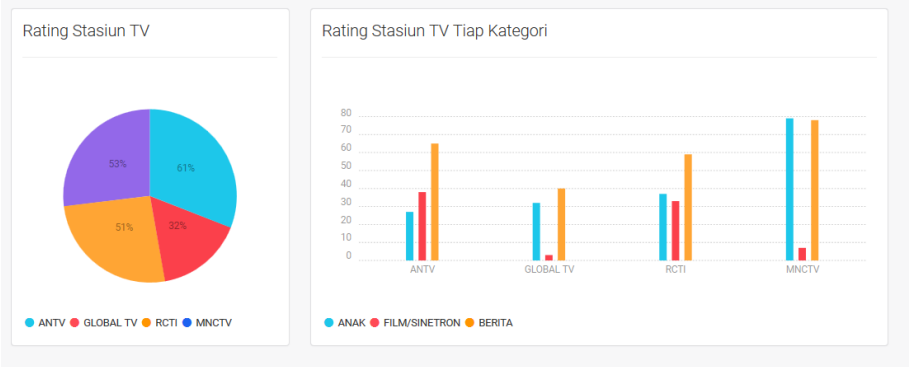

Figure 2. TV Stations Quality Based On Public Opinion

TV station that got the highest positive public opinion is ANTV with getting 61% of positives opinion. For TV programmes, MNCTV got two highest categories, for kid programmes category 80% and news category 79%.

5. Conclusion

Improved K-Nearest Neighbor (IKNN) method can be used to classify Indonesian sentiment into 3 categories: positive, negative and neutral sentiments. From the results of comparing positive opinion with negative and neural opinion, we can conclude the highest for news category is MNCTV, the highest result for the category of film / soap opera is ANTV, and the highest result for children category is MNCTV by 80%. The highest overall favourite for Station TV is ANTV. Sentiment analysis examination for Bahasa Indonesia with Improved K-Nearest Neighbor (IKNN) yield the highest accuracy for k =10 which is 90%.

6. Reference

[1] D. Farber, "Twitter Hits 400 Million Tweets per Day, Mostly Mobile. CNET," 2012.
[2] B. Liu, "Sentiment Analysis and Opinion Mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012.
[3] I. P. Cvijikj and F. Michahelles, "Understanding social media marketing," in *Proceedings of the 15th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '11*, 2011, p. 175.
[4] B. Pang, L. Lee, … S. V. the A.-02 conference on, and undefined 2002, "Thumbs up?: sentiment classification using machine learning techniques," *dl.acm.org*.
[5] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision."
[6] Amalia, D. Gunawan, M. S. Lydia, and Charlie, "The Usage Evaluation of Official Computer Terms in Bahasa Indonesia in Indonesian Government Official Websites," *J. Phys. Conf. Ser.*, vol. 755, p. 11001, 2016.
[7] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining."
[8] G. Stylios *et al.*, "Public Opinion Mining for Governmental Decisions," *Electron. J. e-Government*, vol. 8, no. 2, pp. 203–214, 2010.
[9] 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," *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, pp. 115–120, 2012.
[10] A. Wang, "Services-oriented URL filtering and verification Chao Wang *, Xi Li and Xuehai Zhou Nadia Nedjah," vol. 4, no. 4, 2013.
[11] R. Jose and V. S. Chooralil, "Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach," in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, 2016, pp. 64–67.