Microblogging Sentiment Analysis with Lexical Based and Machine Learning Approaches

Warih Maharani
Faculty of Informatics
Telkom Institute of Technology
Bandung Indonesia
wrh@ittelkom.ac.id

Abstract— **The Digital World encounters rapid development nowadays, especially through the proliferation of social media in Indonesia. Twitter has become one of social media with expanded users within every sectors of society. There are so many part both individual as well as organization/enterprise which utilize twitter as tool for communication, business, customer relation, and other activities. Through the twitter's ever-expanding users with those particular purposes, the precise method to effectively and efficiently analyzing opinion-contained sentences become crucially needed. Therefore this research made for method analyzing through lexical based and model based approaches by machine learning to classify opinion-contained tweets using those 2 methods. The tested machine learning method are Support Vector Machine (SVM), Maximum Entropy (ME), Multinomial Naive Bayes (MNB), and k-Nearest Neighbor (k-NN). Based on the test outcome, lexical based approach highly depended on lexical database which became opinion classification matrix. Whilst machine learning approach can produce better accuracy due to its capability in new training data modeling based on outcome model. However, machine learning model based approach depends on various factors in analyzing sentiment.**

Keywords— Twitter, tweet, lexical based, machine learning, Support Vector Machine, Maximum Entropy, Multinomial Naive Bayes, k-Nearest Neighbor

I. INTRODUCTION

Opinion mining has been vastly developed along with increasing popularity of social media. Before the web proliferation, sometimes we need to gather opinion from relatives, friends, and other people around us. When a company needs public opinion for market research, 79% of public surveys are conducted to obtain outsiders opinion [1]. Within the rapid development of web that has reached 630,795,511 sites according to Netcraft surveys, people's method in expressing opinion and feeling is surely changed. They can contribute their feedback, opinion, or else in internet, group, blog, and other currently trending social media.

The research in opinion mining has been conducted before. It starts by indentifying tendentious words of certain subjectivity which representing opinion and semantics oriented identification (both positive and negative).

II. PREVIOUS WORKS

There are two possible approaches for opinion mining, which are lexical based and model based through machine learning. Dictionary based approach or what is called lexical based is language model-based approach. In Ding, X., Liu, B. and Yu [2] and Kanayama [3] research holistic lexicon benchmark and fully automatic lexicon expansion is used for opinion mining. Where in other hand, lexical benchmark which is utilized for English language generally uses Sentiwordnet [4][5][6][7]. Sentiwordnet established based on Wordnet lexical database as result of Princeton University and based on gloss quantitative analysis that connected to synset. This particular research has also being developed within 3.0 versions for optimizing current lexical database.

Lexical Sentiwordnet database that being produced has been used in a research Yan Dang et al [8] for sentiment classification in online product review case, using sentiwordnet and determines review subjectivity [9]. In previous research, researcher has developed lexical database for Bahasa Indonesia as an enhancement of Sentiwordnet [10].

The downside of lexical based method is that it relatively depends on established lexical database and language model architecture. However, there is an upside where it takes no training process. Therefore, machine learning is used as possible alternative approach. The most often utilized method in machine learning approach is SVM, Naive Bayes, Maximum Entropy and k-Nearest Neighbor (k-NN) [11]. Therefore this research is conducted to compare opinion mining performance by lexical based and model based approach within Indonesian language microblogging twitter data cases.

III. OPINION MINING

Within the process of opinion mining and sentiment analysis, there are several approaches that can be done such as lexical based opinion mining and model based or machine learning utilization.

978-1-4673-4992-5/13/$31.00 ©2013 IEEE

A. *Lexical based opinion mining*

This method of approach is one of possible alternative way used in opinion mining by utilizing lexical resource as benchmark source in determining opinion sentence subjectivity. One of the benefits from this method is that data training is no longer needed as model for predicting opinion classification. However, the determination of opinion subjectivity will not only been done based on available lexical, but also matching language model.

Lexical database provides opinion polarity information for every Indonesian word. That information will be processed based on English Sentiwordnet, Thesaurus and Indonesian dictionary.

Lexical based method relatively easy to be implemented due to its ability to replace manual architecture of lexical sentiment and doesn't require data training for opinion sentence model. Classification within lexical Based Approach is done based on previous research [9] :

$$\text{if } \sum_k \text{score(k)} \geq 0 \text{ then positive,}$$

$$\text{if } \sum_k \text{score(k)} < 0 \textit{ then negative} \quad (1)$$

B. *Machine learning based opinion mining*

Model based approach is an oncoming of machine learning method, which can be seamlessly done, both by supervised or unsupervised learning [11].

Supervised learning approach is done after studying the characteristics that possibly possessed by documents at particular class. This approach generally named after classification technique. This classification will divide the existing documents into training and testing document. By using the training document, this approach builds analysis model to be able to determine in which class a certain document belongs to. This particular model architecture done after performing feature selection, in which diction or part of document will be utilized as flag for determining class of a certain document. This class classification analysis, using machine learning, can be performed through several existing method, such as Naïve Bayes, Support Vector Machine, k-Nearest Neighbor and Maximum Entropy. Therefore this research implements those methods due to its usage frequency for opinion classification [11].

IV. SYSTEM DESIGN

A. *Lexical based Opinion Mining Design*

One of problems in microblogging data is abundances of information that is not included in the focus of opinion classification.

Example:

"Someone @Telkomsel flash pascabayar kartuHalo sjk hr minggu knp jd cepet bgt"

Therefore cleaning data process is required by erasing characteristics of tweet and existing symbols. It transforms the tweet above into:

"flash pascabayar kartuHalo sjk hr minggu knp jd cepet bgt"

Moreover, due to character limitation for tweet, there are many grammatically incorrect words resulted from abbreviation/simplification. Therefore, a special treatment should be applied to overcoming that particular problem.

After going through data cleaning process, the Part of Speech (POS) for each word within the sentence will be used to figures sentiment score of each particular word. Each sentiment score is acquired from lexical database for Indonesian language which referred to English SentiWordNet [10].

Acquired value will be used for calculating average sentiment score. It should be done since every word in SentWordNet have multiple definition or sense [8]. That average sentiment score will be calculated by the formula of:

$$Score(word = pos)_i = \frac{\sum_{k \in SentWordNet(word=pos\ \&\ polarity=i)} SentWordNet_{Score(k)_i}}{|synsets(word=pos)|} \quad (2)$$

Whereas pos ϵ POS for each word (as an adjective, adverb, verb), i ϵ {positive, negative, objective}, and k representing synsets of targeted word for particular definition or sense.

Following are examples of scoring weight from several opinion words in Indonesian language which stored in lexical database:

id	PScore	NScore	synset	gloss
00001740	0	0	bernapas	mengisap dan mengeluarkan napas#
00001740	0	0	hidup	mengisap dan mengeluarkan napas#
00001740	0	0	berasimilasi	mengisap dan mengeluarkan napas#
00001740	0	0	bernyawa	mengisap dan mengeluarkan napas#
00002724	0.375	0.375	tersedak	tersesat atau salah jalan (tentang air dsb yang di...
00002724	0.375	0.375	terteguk	tersesat atau salah jalan (tentang air dsb yang di...
00002724	0.375	0.375	sedak	tersesat atau salah jalan (tentang air dsb yang di...
00002724	0.375	0.375	kesedakan	tersesat atau salah jalan (tentang air dsb yang di...
00002724	0.375	0.375	terselak	tersesat atau salah jalan (tentang air dsb yang di...
00002724	0.375	0.375	keselak	tersesat atau salah jalan (tentang air dsb yang di...
00003380	0	0.25	mengganggu	menggoda; mengusik# merintangi; menyebabkan tidak ...
00003380	0	0.25	menggoda	menggoda; mengusik# merintangi; menyebabkan tidak ...
00003380	0	0.25	mengusik	menggoda; mengusik# merintangi; menyebabkan tidak ...
00003380	0	0.25	merintangi	menggoda; mengusik# merintangi; menyebabkan tidak ...
00003380	0	0.25	merundung	menggoda; mengusik# merintangi; menyebabkan tidak ...
00003380	0	0.25	menggelisahi	menggoda; mengusik# merintangi; menyebabkan tidak ...
00003380	0	0.25	memerusa	menggoda; mengusik# merintangi; menyebabkan tidak ...
00003380	0	0.25	memecah	menggoda; mengusik# merintangi; menyebabkan tidak ...
00003380	0	0.25	merembet	menggoda; mengusik# merintangi; menyebabkan tidak ...
00003380	0	0.25	grecok	menggoda; mengusik# merintangi; menyebabkan tidak ...
00003380	0	0.25	menegur	menggoda; mengusik# merintangi; menyebabkan tidak ...

Figure 1. Sampe of contain within Indonesian language lexical database

978-1-4673-4992-5/13/$31.00 ©2013 IEEE

B. *Machine Learning based*

In general, this opinion classification is represented in 2 classes which are positive and negative. In principal, opinion classification becomes text classification problem, which can be solved using supervised and unsupervised learning methods. The SVM and Naive Bayes methods have proven to be eligible to classify movie review data into positive and negative opinion [12][13][14].

This research is using SVM, Naive Bayes, k-NN and Maximum Entropy methods of approach to classify opinion within twitter.

Process performed on machine learning approach consists of data cleaning, stemming, POS tagging, tokenization and term weighting using subjective weights on Indonesian lexical database [10]. Numerical data generated from the process will be classified using machine learning approach [13].

1) *SVM*

In the implementation of SVM method as classifier, Indonesian language lexical database is used for acquiring scoring weight from existing terms within opinion sentence.

In the simple way, SVM seek for the best hyperplane which will be functioned as classes separator in input space [15]. In the learning process, SVM seek for the best hyperplane among several existing alternative hyperplanes (discrimination boundaries). In this microblogging problematic, linear kernel SVM will be used [12]. In linear SVM, there are two classes which should be separated by hyperplane and the SVM should seek for the best hyperplane to separating two classes of positive and negative.

Preprocessing stage produces the numerical data that will be used as training data to generate a model of SVM. Based on this model, test data was tested by applying the selected parameters and the resulting SVM model [12].

2) *Multinomial Naive Bayes*

Within MNB, a random variable of Z representing words emersion in vocabulary V. Each document d is described as a multinomial word distribution, N_{it} calculated from the number of occurrences of the word w_t occurs in document d. So the possibility of a given document is a class. The probability of mentioned document in assumption that the probability of each word is independent and c_j class is known, will be defined as [14] :

$$P(d_i|c_j;\theta) = P(|d_i|)|d_i|!\prod_{t=1}^{|v|}\frac{P(w_t|c_j;\theta)^{N_{it}}}{N_{it}!} \quad (3)$$

Probability will be calculated based on :

$$\theta_{w_t|c_j=P(w_t|c_j;\theta)},\ 0 \le \theta_{w_t|c_j} \le 1,\ \sum_t \theta_{w_t|c_j} = 1 \quad (4)$$

3) *K-Nearest Neighbor*

KNN Method algoritm works in the nearest range from query instance to label sample in order to find the KNN. Label sample is projected to N-dimensional space, where each dimensions are representing data features. This space will be divided into several parts based on label sample classification. A dot in this space is marked by x class, if x class is the most common in the nearest k number of neighbor from that dot. The length of gap between neighbors generally calculated based on euclidean distance [16].

4) *Maximum Entropy*

The principal of Maximum Entropy classification is to maximize the probability of a document to be classified. Within the process of classification, conditional probability value p(a|b) will be searched from a class if document b is known, for a compilation of class A = $\{a_1, a_2, ..., a_p\}$ and B = $\{b_1, b_2, ..., b_q\}$.

For several training and feature data that used in the particular learning, we calculate conditional probability for a particular condition (a|b) as:

$$P(a,b) = \pi \exp\left(\sum_{i=1}^{n} \alpha_i^{f_i(a,b)}\right) \quad (5)$$

where π as normalization and α_i parameter is known using Generalized Iterative Scaling (GIS) algorithm [17]. Each parameter α_i, where $\alpha_i > 0$, corresponds to one feature f_i and can be interpreted as a weight for that feature. The probability p(a,b) is then a normalized product of those features that are active on the (a,b) pair. After that, class a as classification result of class b is expressed as:

$$a = \arg\max p(a,b) \quad (6)$$

V. EXAMINATION

To discover the performance result differences from 2 mentioned approached, tweet data from twitter is used as test data which is taken periodically within the range of October – November 2011. Data obtained using the engine crawler. There are 50.000 crawled Indonesian data from twitter used.

Following are several data samples regarding one of Telecommunication Provider Product in Indonesia.

comment	topik
@telkomsel utk aktivasi iphone flash unlimited *78...	telkomsel
@ipimaripi @wowogombel #kode coba bisa ke @telkoms...	telkomsel
@telkomsel mau tanya nih,tanda2 klo sim card kita ...	telkomsel
ngabuburit seru bareng opera mini & telkomsel....	telkomsel
@telkomsel udah dicoba tapi gak ada perubahan.	telkomsel
@telkomsel memang tdi malam ada telkoms poin 75 jt...	telkomsel
sudah, tapi semua indomaret yang saya datangi ngga...	telkomsel
@telkomsel boleh nanya gakk	telkomsel
koneksi internet @telkomsel flash pascabayar kartu...	telkomsel
sudah.. rt @telkomsel: @iphejhe jika paket telkoms...	telkomsel
promo diskon utk pembelian mjlh & buku di app ...	telkomsel
ikuti program gebyar bali festival, untuk kamu pen...	telkomsel

Figure 2. Tweet data sample

978-1-4673-4992-5/13/$31.00 ©2013 IEEE

Before continue to classification process, data will go through preprocessing stage by eliminating tweet and punctuation mark characteristics as well as elimination stop words. Tweet characteristics refer to @ symbol, hashtag that begins with #, and URL that point to particular page of a website.

A. Lexical based Opinion Mining Examination

Examination through lexical based approach is done by implementing with and without stemming process. Following are the average of classification accuracy for 1000 varied tweet data and the experiment being performed 5 times:

TABLE I. LEXICAL BASED OPINION MINING ACCURACY RESULT

Lexical based	Accuracy (%)					Average
Without stemming	76.7	72.1	76.7	74.9	72.56	74.59
With stemming	64.2	69.5	62.5	68.9	70.2	67.06

The effect of stemming process is resulting in the loss of information value of a certain word within a sentence. This condition will effecting on the determination of scoring weight that will be done to change text data into numerical data by using existing lexical data. The word which lost its information value in stemming process has no score or even wrongly valued. This will result to words' numerical data which will become input of classification in existing method.

B. Machine Learning based Opinion Mining Examination

Examination using SVM method, being done by experimenting several C score which valued 1 – 50, to aim optimum accuracy. The average results of classification accuracy using SVM, ME, MNB and k-NN methods which used in lexical based to the same 1000 tweet data are:

TABLE II. MACHINE LEARNING BASED OPINION MINING ACCURACY RESULT

Methods	Accuracy (%)					Average
SVM	84.0	78.45	79.60	82.6	82.50	81.43
ME	84.20	80.78	83.67	80.5	76.90	81.21
MNB	83.79	81.56	78.23	77.9	84.56	81.29
k-NN	72.66	75.20	72.92	70.0	70.10	72.1

Based on the experiment result using SVM method, it appears that the varied value of C parameter results in relatively stable accuracy value.

Moreover, there is no significant differences among experiment result using SVM, ME and MNB methods, showing that those methods is able to properly classifying opinion in Bahasa. Meanwhile for k-NN method, the system accuracy value is depending on how many K (amount of nearest neighbor) determined. After testing with various number of K, ranged from 1 -25, it obtains optimum K value of 19, as what is seen on the table below:

TABLE III. TESTING RESULT WITH VARIED K VALUE

K	1	3	5	7	9	11	13
	63.2	66.8	68.2	70	72.6	73.6	73.6

K	15	17	19	21	23	25
	73	72.2	75.2	73.2	73	70

The determination of K optimum number is depending on data set that being used, so different result can be obtained from varied 1000 tweet data by the average value of 72.1%.

Test result shows that model based approach with machine learning produce better accuracy than lexical based approach. The lexical based approach is highly depended and influenced by referred lexical database in weight determination. Even though model based approach using scoring weight sheet from lexical data database but those scores will be referred as training data to obtain model base.

Moreover, lexical based approach is also depended to established language architecture. In this research, there is no treatment for classifying implicit opinion sentence, so the accuracy rate is still not optimum.

CONCLUSION

Based on the result of system tests and analysis it can be concluded that scoring result using lexical database for Indonesian language is able to classifying opinion into positive and negative

Model based approach with machine learning produce better accuracy rate than lexical based approach. This happens because of significant influence from lexical database which has been set as reference in determining positive and negative opinion. Meanwhile the accuracy rate using machine learning approach is depended on the dataset which become the training data and determine varied parameter for each method.

ACKNOWLEDGMENT

This research is one of research project result that being conducted in Telkom Institute of Technology Faculty of Informatics. Therefore we convey our highest gratitude to IT Telkom for all supporting facilities that has be provided.

REFERENCES

[1] Gandy, H.Oscar. Public Opinion Surveys and The Formation of Privacy Policy. Journal of Social Issues, Vol.59, No. 2, pp.283-299. 2003.

[2] Ding, X., Liu, B. and Yu, P. A Holistic Lexicon-Based Approach to Opinion Mining. *Proceedings* of *the first ACM International Conference on Web search and Data Mining* (*WSDM'08*), 2008.

[3] Kanayama, H. and Nasukawa, T. Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis. Proceedings of the 2006

Conference on Empirical Methods in Natural Language Processing (EMNLP'06), 2006.

[4] Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), pages 417–422, Genova, IT.

[5] Andrea Esuli and Fabrizio Sebastiani. 2007a. Randomwalk models of term semantics: An application to opinion-related properties. In Proceedings of the 3rd Language Technology Conference (LTC'07), pages 221–

225, Pozna´n, PL.

[6] Andrea Esuli and Fabrizio Sebastiani. 2007b. SENTIWORDNET : A high-coverage lexical resource for opinion mining. Technical Report 2007-TR-02, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT.

[7] Andrea Esuli. 2008. Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms, and Applications. Ph.D. thesis, Scuola di Dottorato in Ingegneria "Leonardo da Vinci", University of Pisa, Pisa, IT.

[8] Yan Dang, Yulei Zhang, Hsinchun Chen. A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews. IEEE Intelligent Systems 25(4): 46-53. 2010.

[9] Ohana, Bruno. *Opinion Minning with the SentWordNet Lexical Resource*. Dublin Institute of Technology. 2009.

[10] Maharani, W., Atastina I., Alfian. Indonesian Lexical Database for Opinion Mining. Jurtel Journal of IT Telkom Bandung. December 2012.

[11] Pang, Bo and Lee, Lillian. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval Vol. 2, No 1-2. 2008.

[12] Joachims, T. *Text Categorization with Support Vector Machines:Learning with Many Relevant Features*. Germany:University of Dortmund. 1998.

[13] Taylor and Cristianini. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press. 2000.

[14] Pang, Bo and Lee, Lillian. *Thumbs up? Sentiment Classification Using Machine Learnig Techniques*. USA:Cornell University. 2002.

[15] Gunn, S. (1998). *"Support Vector Machines for Classification and Regression"*. ISIS Technical Report, Image Speech & Intelligent Systems Group University of Southampton.

[16] Kozma, Laszlo. (2008). *K Nearest Neighbors Algorithm (kNN)*. Helsinki University of Technology, Special Course in Computer and Information Science.

[17] Mehra, Nipun & Khandelwal, Shashikant & Patel, Priyank. *Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews.*