Comparison Study Of Term Weighting Optimally With SVM In Sentiment Analysis

Amril Mutoi Siregar[1], Sutan Faisal[1], Tukino[1], Adam Puspabhuana[2], Manase Sahat H Simarangkir[3]
{amrilmutoi@ubpkarawang.ac.id[1],sutan.faisal@ubpkarawang.ac.id,tukino@ubpkarawang.ac.id[3], adambhuana@stmik-kharisma.ac.id[4], manasemalo@politeknikmeta.ac.id[5]}

[1]Deparment of Technology and Computer Science, Buana PerjuanganUniversity,Jalan HS. Ronggo WaluyoTelukjambe Timur, Karawang, Indonesia
[2]Department of Information System Diploma, STMIK of Kharisma, JalanPangkalPerjuangan KM 1, Karawang, Indonesia
[3]Department of ComputerEngineering, Meta Industry Polytechnic, JalanInti I Blok C1 No. 7LippoCikarangCibatuCikarang, Bekasi, Indonesia

Abstract: The rapid of internet and social media users have changed the way people interact in their daily activities. For example, banking and retail began to use various social media, especially online media such as tweeter. The problem that arises is how to get information from thousands and even million data generated through social media, to be a decision as in predicting consumer satisfaction of the service or product.Another problem is the social media users in communicating using slang or local language. In sentiment analysis to predict the sentiment is not easy because it must be able to identify the words. In sentiment analysis, to overcome these problems the method used is text mining so as to process opinions from social media. The proposed approach is to analyze optimal term weighting between TF-IDF, frequency term (TF) and Binary Term Occurrence (BTO), using SVM algorithm. Target feature extraction for selection of datasets by predicting positive and negative sentiments. The result of weighting of terms approaching sentiment is using TF-IDFwith SVM.

Keywords: TF – IDF, TF, Binary Term of Occurrence, SentimentAnalysis, SVM, Twitter

1. Introduction

One of the social networking microblog that is currently in demand by many people is Twitter which has a feature to send and read a short message called tweets. Tweets can be read freely but can also be set only visible to other users who follow his Twitter or so-called follower.

Twitter acts as a media spread of information very quickly as Twitter users increase. The information generated from twitter is so free and varied as news, opinions, comments, criticism both positive and negative. Everyone can see the opinions of others on a problem through sentiment analysis, so it can help us in making a decision more quickly and accurately.

Sentiment analysis is a part of Text Mining or extracting text data, among which there is a process of processing and extracting textual data automatically to obtain information(BPL 8, Feldman, Ronen and Sanger, 2007).

The concentration of this research is to analyze the optimal term weighting between several methods such as TF-IDF, TF and BTO. These methods are classified using SVM algorithm.The result of this researchwill become an assessment to predict a text or tweets towards positive or negative.

ICASI 2019, July 18-19, Banda Aceh, Indonesia

DOI 10.4108/eai.18-7-2019.2288508

In this research,it will be generatedthe comparison of several term weighting methods. The best result of term weighting methods which is closer to either positive or negative, it would be used for the next research related to text mining.

2. Theories

2.1 Sentiment Analysis

Sentiment analysis is a method for analyzing data to know the sentiment of people. Sentiment analysis can be categorized into three tasks, namely informative text detection, information extraction and sentiment interestingness classification (emotional, polarity identification).

Sentiment classification (negative or positive) is used to predict polarity sentiment based on user sentiment data(Pan, S., Ni, X., Sun, J., Yang, Q., & Chen, 2010).Textual sentiment analysis is widely used not only in the area of scientific research but also for the needs of business marketing and technology(Chintala, 2012).

Sentiment analysis is a prominent and growing active research area that is influenced by the rapid growth of social media technologies. Through social media there are opportunities to access opinions from a number of people ondifferent types of businesses, world issues and social issues(Go, A., Huang, L,Bhayani, 2009).

2.2 Twitter

Twitter is a microblogging service that was officially released on July 13, 2006(Mostafa, 2013). Twitter's main activity is to post something short (tweet) via the web or mobile. The maximum length of a tweet is 140 characters. Twitter has become an almost limitless source used in text classification.

There are many characteristics on twitter tweets(Go, A., Huang, L,Bhayani, 2009). The messages on twitter have many unique attributes, which distinguish them from other social media:

- Twitter has a maximum character length of 140 characters.
- Twitter provides data that can be accessed freely using the Twitter API, simplify the process of collecting tweets in large numbers.
- Language model - twitter users post messages through many different media. Frequency of misspellings, slang language and abbreviations higher than other social media.
- Twitter users send short messages on various topics that are tailored to a particular topic and it applies globally. Over the past few years, twitter has become very popular. The number of twitter users has risen to 190 million and the number of tweets published on Twitter every day is over 65 million (Ravichandran, M., & Kulanthaivel, 2014).

2.3 Term Weighting

In text mining, word weighting is crucial to get sentiment in the form of news and opinions. In weighting the word will determine whether opinions include positive or negative sentiments. In this research some weighting of words used.

2.3.1 TF-IDF

TF-IDF weights are the weight of every word in each document. To obtain TF-IDF value using Equation 2. To get the value of TF-IDF required IDF value. Because the term "the" is so common, term frequency will tend to incorrectly emphasize documents which happen to use the word "the" more frequently, without giving enough weight to the more meaningful terms "brown" and "cow".

The term "the" is not a good keyword to distinguish relevant and non-relevant documents and terms, unlike the less-common words "brown" and "cow". Hence an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

Karen Spärck Jones conceived a statistical interpretation of term specificity called Inverse Document Frequency (IDF)(Spärck Jones, 1972), which became a cornerstone of term weighting. The specificity of a term can be quantified as an inverse function of the number of documents in which it occurs.

The IDFT value can be searched with Equation 1.

$$idf_t = log_{10}\left(\frac{N}{df_t}\right) \qquad (1)$$

Where IDFT is the number of documents containing a term and N is the total document tested.

$$Tf.Idf = tf\ x\ idf_t \qquad (2)$$

IDF is the number of documents containing the term. The log function of IDF to provide some smoothing. In this case, each document is regarded as a vector with 1 componentcorresponding to each term present in the dictionary along with the weights of each component. For terms that do not appear in the document, then the weight is 0.

2.3.2 Term Frequency

Suppose we have a set of English text documents and wish to rank which document is most relevant to the query, "the brown cow". A simple way to start out is by eliminating documents that do not contain all three words "the", "brown", and "cow", but this still leaves many documents.

To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its term frequency. However, in the case where the length of documents varies greatly, adjustments are often made (see definition below).

The first form of term weighting is due to Hans Peter Luhn which may be summarized as:The weight of a term that occurs in a document is simply proportional to the term frequency(Luhn, 1957).

$$Tf_{i,j}\frac{Tf_{i,j}}{\max(Tf_{i,j})} \qquad (3)$$

2.3.3 Binary Term Occurrences

Binary term occurrence is the term weighting that use the occurrence of term. if a word appears, one or more will be assigned a value of 1 and if none exists it will be assigned a value of 0.

$$\begin{matrix} 1,\ if\ Tf_{i,j} > 0 \\ 0,\ if\ Tf_{i,j} = 0 \end{matrix} \qquad (4)$$

2.4 Support vector machines (SVM)

Support Vector Machines (SVM) is a learning method tool that analyzes data and recognizes patterns, used for classification and regression analysis. The original SVM algorithm was created by Vladimir Vapnik and the current standard derivative of Soft Margin (Cortes, C. Vapnik, 1995).

The standard SVM takes the set of input data, and predicts, for each given input,possible input is a member of one of the classes of the two existing classes, which creates an SVM as a binary linear nonprobability classifier. Because SVM is a classifier, then assigned a training set, each marked as belonging to one of two categories, an SVM training algorithm constructs a model that predicts whether new data falls into a category or another.

2.5 Research Methodology

In the research using the following methods:

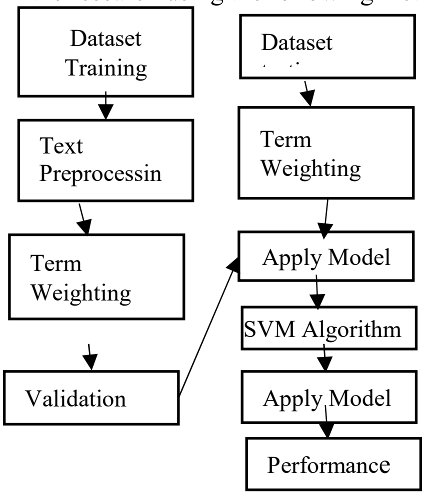

Fig.1. Research Methodology

2.5.1 Dataset Training

Dataset training is a collection of examples used for learning, that is in accordance with the parameters used. In this research using dataset training from tweeters as much as 200 tweets.

Table 1. Dataset Training I

Sentiment	Tweet
Negative	is so sad for my APL friend............
Negative	I missed the New Moon trailer...
Positive	omg its already 7:30 :O
Negative	Omgaga. ImsoooimgunnaCRy. I've been at this dentist since 11.. I was suposed 2 just get a crown put on (30mins)...
Negative	i think mi bf is cheating on me!!! T_T
Negative	or i just worry too much?
Positive	JuuuuuuuuuuuuuuuuussssstChillin!!
Negative	Sunny Again Work Tomorrow :-| TV Tonight

Positive	handed in my uniform today . i miss you already
Positive	hmmmm.... i wonder how she my number @-)
Negative	I must think about positive..

Table 2. Dataset Training II

Sentiment	Tweet
Positive	thanks to all the haters up in my face all day! 112-102
Negative	this weekend has sucked so far
Negative	jbisnt showing in australia anymore!
Negative	ok thats it you win.
Negative	&lt;-------- This is the way i feel right now...
Negative	awhhe man.... I'm completely useless rt now. Funny, all I can do is twitter. http://myloc.me/27HX
Positive	Feeling strangely fine. Now I'm gonna go listen to some Semisonic to celebrate
Negative	HUGE roll of thunder just now...SO scary!!!!

2.5.2 Dataset Testing

Dataset testing is a dataset independent of the training dataset but follows the same probability distribution as the training dataset. If the model corresponding to the training dataset also matches the test dataset well. In this study using dataset testing of tweeter as much as 10 tweets.

Table 3. Dataset Testing I

Doc	Tweet	RelevanJudgement
Doc 1	Magic ..thinking less than 50 % chance Hedo stays in Orlando. He's gonna go for the $$. They all do. Can't blame him though.	Negative
Doc 2	now who wud make this beautiful plc cry?	Negative
Doc 3	it is now the outer fat kid.	Negative
Doc 4	The servers are now backup, if you experience any more problems then please let me know Sorry about the delay...	Positive
Doc 5	&quot;yes, i am lol j/k luv this! @DJDolceVita: &quot;you are like a gentle breeze that has blown throug... â™« http://blip.fm/~7s7mn	Positive
Doc 6	no no se vale.	Negative

Table 4. Dataset Testing II

Doc	Tweet	Relevant Judgement
Doc 7	are my future rooooooooomies! NYC holler	Positive
Doc 8	I hope u have cheeSeburgerS 4 @Snubbmatic LMBO	Positive
Doc 9	Awesome	Positive
Doc 10	I am nothing of the kind and everything of the sort http://tr.im/mLsn	Positive

2.5.3 Text Preprocessing

Text preprocessing is a process of transforming unstructured data forms into structured data according to need, for further mining processes (sentiment analysis, summary, clustering of documents, etc.). or is the stage of the initial process of the text to prepare the text into data to be processed further.

2.5.4 Term Weighting

Word weighting is done to get the value of the word / term successfully extracted. This research uses TF-IDF, TF, BTO method as a weighting process. At this stage, each document is manifested as a vector with as many elements of a word as successfully identified from the extraction step of the document above. The vector contains the weight of each word calculated based on the method used.

2.5.5 Apply Model

Apply The first model is trained on Example Set by another Operator, which is often a learning algorithm. After that, this model can be applied to another Example Set. Usually, the goal is to get predictions on invisible data or to transform data by applying a preprocessing model.

The Example Set in which the model is applied must be compatible with the Model attribute. This means that Example Set has the same number, order, type, and role Attribute as Example Set used to generate the model.

2.5.6 SVM Algorithm

SVM is SVM (Support Vector Machine) is a fast algorithm and good result for learning.These operators support different types of kernels including dots, radials, polynomials, neural, anova, epachnenikov, gaussian and multiquadric combinations. This kernel type explanation is given in the parameters section.

2.5.7 Validation

Validation is used to estimate how accurately a model will perform in practice.Validation Operator has two subprocesses: Training subprocesses and Testing subprocesses. The training subprocess is used to train the model. The trained model is then applied to the Test subprocess. Model performance is measured during the Testing stage.

The Example Set input is partitioned into a subset of k of the same size. From a subset k, a subset is retained as a test data set (ie an input of the Test subprocess). The remaining k - 1s are used as training data sets (ie subprocess input training). The cross validation process is then repeated at times, with each subset k used exactly once as the test data.

The result k of the iteration k is the average (or other combination) to produce a single estimate. The k value can be adjusted by using the number of fold parameters. Evaluation of model performance on independent test sets yields good performance estimates on invisible data sets.

2.5.8 Performance of Evaluation

Parameters used to evaluate the weighting performance of words in this study is the level of weighted accuracy to get the result of sentiment.

*Prediksi*Pos= Confidence pos> confidence Neg
*Prediksi*Neg = Confidence Neg> confidence Pos

Description:
*Prediksi*Pos = Positive Prediction
*Prediksi*Neg = Negative Prediction

Confidence pos = PositiveConfidence
Confidence Neg = Confidence Negative

Performance evaluation was performed onexperimental results of sentiment analysis system and on respondent sentiment analysis.

3. Results

The result of this research is the weighted performance of word to SVM algorithm following result:

Table 5. Term Weight Comparison Result

Name Document	TF-IDF		Term Frequency		Binary Term Occurrences	
	Confidence (Negative)	Confidence (Positive)	Confidence (Negative)	Confidence (Positive)	Confidence (Negative)	Confidence (Positive)
Doc 1	0.702	0.298	0.700	0.300	0.707	0.293
Doc 2	0.742	0.258	0.754	0.246	0.734	0.266
Doc 3	0.610	0.390	0.614	0.386	0.651	0.349
Doc 4	0.641	0.359	0.628	0.372	0.676	0.324
Doc 5	0.538	0.462	0.568	0.432	0.582	0.418
Doc 6	0.707	0.293	0.717	0.283	0.706	0.294
Doc 7	0.680	0.320	0.681	0.319	0.694	0.306
Doc 8	0.739	0.261	0.730	0.270	0.722	0.278
Doc 9	0.631	0.369	0.635	0.365	0.679	0.321
Doc 10	0.506	0.494	0.572	0.428	0.636	0.364

One example of the results of the above research is to show that doc 1 predicted negative with confidence negative = 0.707 and confidence positive = 0.293, so the conclusion that confidence negative>confidence positive, produce negative prediction.

The confidence value will change depending on the weighting of the word used, the weighting of words in sentiment analysis is very important to affect the weight of confidence.

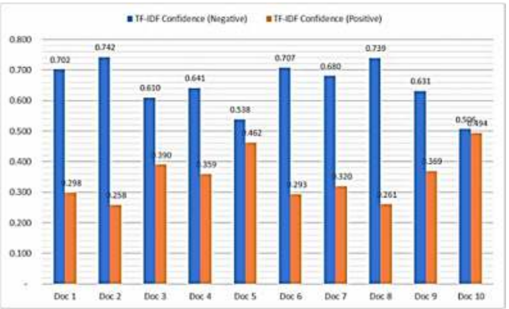

Fig.2. TF-IDF Positive and Negative Comparison Result Graphic

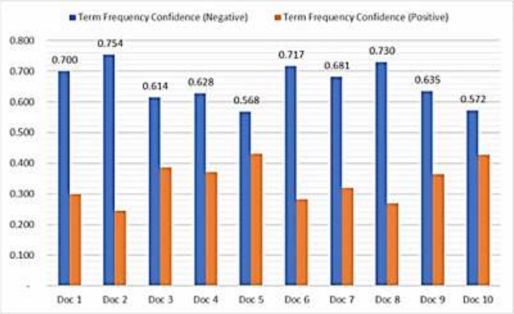

Fig.3. TF Positive and Negative Comparison Result Graphic

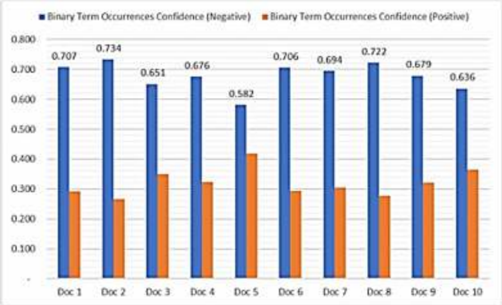

Fig.4. BTO Positive and Negative Comparison Result Graphic

4. Conclusions

The result of this research is the weightedThis studybuilds a model for doing Tweet Prediction based onword weighted combination sentence with SVM. The results obtained in this study are that TF-IDF performance is better. And TF and BTO have the same performance. The results of this experimental research using Rapid Miner Studio 7.6 show that accuracy with the term frequency feature provides better accuracy results than accuracy with the TF-IDF feature.

5. Future Plans

In this research it is necessary to use a larger Dataset Training to obtain optimal results in the use of word weighting in sentiment analysis.

Acknowledgements

We would like to express our deepest gratitude to Our God Allah SWT for making so many things possible and we believed that this research would never be done without Allah's blessing and Allah's guidance. Our deepest thanks to our families for their love and supports throughout our life. Thank you for giving us strengths, love and wishes to achieve our goals. Our great thanks to all our great lecturers that we could not mentioned all their names. Their supports and wishes become motivation in completing this research. Last but not least, deepest thanks go to all our friends.

References

[1] Chintala, S. (2012) *Sentiment Analysis using neural architectures*. New York University.

[2] Cortes, C. Vapnik, V. (1995) *Support-Vector Networks Machine Learning*.

[3] Go, A., Huang, L,Bhayani, R. (2009) *Twitter sentiment analysis*.

[4] Luhn, H. P. (1957) 'A Statistical Approach to Mechanized Encoding and Searching of Literary Information', *IBM Journal of research and development*, I(4), p. 315.

[5] Mostafa, M. (2013) 'More than words: Social networks' text mining for consumer brand sentiments'', *Expert Systems with Applications: An International Journal*, 40(10), pp. 4241–4251.

[6] Ravichandran, M., & Kulanthaivel, G. (2014) 'Twitter Sentiment Mining (TSM) Framework Based Learners Emotional State Classification and Visualization For E-Learning System', *Journal of Theoretical and Applied Information Technology*, 69(1), pp. 84–90.

[7] Spärck Jones, K. (1972) 'A Statistical Interpretation of Term Specificity and Its Application in Retrieval', *Journal of Documentation*, 28, pp. 11–21.