

# Twitter Sentiment Analysis and Insight for Indonesian Mobile Operators

**Hansen Wijaya, Alva Erwin, Amin Soetomo, Maulahikmah Galinium**

Department of Information Technology, Faculty of Engineering and Information Technology, Swiss German University

---

**Keywords:**

Text Mining  
Sentiment Analysis  
Business Intelligence  
Business Analytic  
Knowledge Discovery

**ABSTRACT**

Twittering is one of the most common human activities around the world nowadays. Twitter is the name of the website that provides micro blogging services. Most people are tweeting whatever they feel. In this direction, the growth of mobile operators are also increasing, and competing to satisfy the customers, and the users of mobile operators are expressing their feelings about the mobile operators. The sentiment analysis then appears to analyze the mobile operator images from the customer side. This can be done by predicting the tweets from the mobile operator users by its sentiment. The training sets of data required to determine the sentiment of the testing data, and the training set is classified manually. This research is to propose an automatic sentiment for Indonesian mobile operators, which gathers insight from the customer side and the tweets are in Indonesian language and about several Indonesian mobile operators. The tweets for the training set are gathered in one month intervals. The accuracy of predicting the sentiment achieves 80%.

*Copyright © 2013 Information Systems International Conference.*

*All rights reserved.*

---

**Corresponding Author:**

Hansen Wijaya,  
Department of Information Technology, Faculty of Engineering and Information Technology,  
Swiss German University,  
EduTown, BSD City, Tangerang 15339, Indonesia.  
Email: hansen.wijaya@student.sgu.ac.id

---

## 1. INTRODUCTION

Since 2006 Twitter has become one of the highest growth of social media in the world. Twitter provides micro blogging services that are limited to 140 characters, at least 250 billion tweets are updated daily [1], with the growth of Twitter, it is becoming one of the greatest data sources that is available on the internet, and by digging into the Twitter data, it will be very useful for social marketers because the data can be extracted for mining opinions, views, moods and attitudes [2]. The growth of Twitter also has led many companies in the world to promote their brand, and keep the relationship with customers and there are plenty of companies that have created a Twitter account and tweet promotions of their brands, or some companies are creating a Twitter account for accepting complaints, comments, or suggestions from their users and reply to the users as part of their customer relation.

Mobile operators are one of the most important services in society and mobile operators should have a monitoring system with real time data, and one of the largest sources of real time data is Twitter because Twitter is the social media which is commonly used to express emotions, suggestions, or ideas by the users. Twitter can be used for many kinds of research, Twitter based research depends on the data captured from Twitter. One of the research that can be done with Twitter is sentiment analysis, which classifies tweets based on the sentiment that is expressed in a text [3]. The sentiment analysis will determine to what category the message is classified and it could be a positive sentiment, negative sentiment, or neutral sentiment.

In the research area, to classify the sentiment in Twitter, it needs unclassified data samples, which known as a training set. The training set is used to create a classification model to be compared with the unlabelled data. The training sets are the key to determine the accuracy of the unlabelled tweets and to create a good training set it needs to take lot of time because the good training sets should be made manually.

There is a method to reduce the time by creating manual training data for example by using the emoticon in the tweet, the emoticon can be a code whether the tweet is good, excellent, worse or terrible, but in this research, the emoticon can't be used because the tweet is based on the Indonesian language, and there are misuses of the emoticon by Twitter users. The research is scoped to analyze the sentiment and analyze the

hidden insight of the Indonesian top four mobile operators and the experiment will be considered in word-based approach.

The paper is organized into 4 different sections, the first section is the introduction, the second section is Research Method, the third section is Result and Discussion, and the last section is the conclusion of the paper.

## 2. LITERATURE REVIEW

### 2.1 Text Mining

Text mining is part of data mining, and it is a process to gather insight or knowledge of the user interaction and work with bunch of documents with the analysis tools, and it tries to gather information by extracting the pattern from the data sources [4].



Figure 1. Text Mining Process

The text collection steps consist of gathering the data from the original source, in this research the original source is Twitter. The preprocessing step is the step to make a structured model from Twitter data and transform the data gathered from Twitter. In the analysis step the data is analysed with several algorithms to get the relevant knowledge and pattern. In the validation step the results are presented and analysed for deeper insight.

### 2.2 Sentiment Analysis

Since 2003 sentiment analysis has been growing and it has become a part of text mining. Sentiment analysis is computational research based on the sentiment, emotion, opinion, comment and every expression that is expressed by text. The document is classified as D and the document is about an object, then the sentiment analysis has purpose to extract the attribute and component of the object that has been commented with a sentiment or expression and to specify the sentiment is either positive or negative [5]. Sentiment analysis is a part of a job that reviews everything related to the computational opinion, sentiment and subjectivity of a text [6]. Sentiment analysis is a tool to process a collection of search results that is objected to an attribute of a product (quality, feature, etc.) and process the aggregation of the opinions [7]. Based on the Oxford English Dictionary (2013) Sentiment is defined as an opinion or a view that is held or expressed. Sentiment is also defined as emotions or feelings that are stated with words. The sentiment analysis is focused on the reviews classification based on the polarity. Based on the classification, sentiment analysis is divided into two main groups: Document classification into opinion or facts class, or known as subjectivity classification. And Document classification to the positive or negative, or known as sentiment analysis. In this case there is an important process to define that the document has opinions and having sentiment also know the topic of the document to conclude the object and the sentiment positive or negative.

### 2.3 Twitter

Twitter is one of the biggest social media, the basic of Twitter is micro blogging services, Twitter is micro blogging because the user of Twitter can only sent and read a text based post with a maximum of 140 characters., and the 140 characters post is known as a tweet. Twitter has been online since March 21 2006 by Jack Dorsey, Noah Glass, Evan Williams and Biz Stone. Based on the Twitter official blog (2013) there are 200 million active users and create almost 400 million tweets each day. The Twitter is owned by Twitter Inc. based in San Francisco and the current chairman of the Twitter is Jack Dorsey.

## 3. RESEARCH METHODS

The tweets are collected using Google drive spreadsheet with the script written in Google Script language to contact the Twitter API and capture the data. The length of the query started from 1<sup>st</sup> of April to 5<sup>th</sup> of May 2013. The crawler only crawls tweets from Indonesian mobile operators such as Axis, Telkomsel, Indosat and XL. The result of crawl is 14,976 tweets were gathered.

For the prediction the research is done with the Rapidminer software, Rapidminer is a data mining open source software that is used in many researches. The model is started by creating the training set. The training set is created by joining two queries from the database to get balanced, negative, and positive tweets, each sentiment consists of 1200 tweets selected randomly for creating the training set. After joining the queries then the tweets are transformed as string attributes with Rapidminer operators, transforming the tweets as string attributes is needed because Rapidminer recognizes the data from the database as nominal, the next step is preprocessing the tweets with process document operators.

The pre-processed document is classified with several machine learning algorithms, in this research the algorithms are SVM (Linear), Naive Bayes and Decision Tree to decide which algorithm has the best accuracy and performance. Naive Bayes classifier uses a probability to define a document class, the classifier is using statistical approach, even though the classifier is not following the grammatical rules. In Naive Bayes Classifier the immersion of words does not affect other words immersion, and the absence of a word does not affect the absence of another word, and it does not decrease the accuracy of Naive Bayes Method [8]. While SVM is the classifier which divides the class with a linear classification and to solve the non-linear problem the kernel tricks is used. Decision tree is classify the data with a model and the model looks like a tree which has branches and leaves, the branch is stated for decision and the leaf as solution.

Naive Bayes has a method to define the best class in classification, which is called Maximum Posteriori (MAP) as stated in equation (1).

$$H_{\text{Bayes}}(d') = \underset{C_j \in C}{\operatorname{argmax}} \Pr(C_j | d') \quad (1)$$

Since naive Bayes uses probability to define a document class, the equation of counting probability is defined with equation (2):

$$\Pr(w_i | C_j) = \frac{1 + TF(w_i | C_j)}{|F| + \sum_{w' \in F} TF(w' | C_j)} \quad (2)$$

Where  $\Pr(w_i | C_j)$  is the probability of word in document in training set for j,  $TF(w_i | C_j)$  is the frequency of word in document in training set for j,  $|F|$  = unique frequency of every words in training set for j,  $\sum_{w' \in F} TF(w' | C_j)$  is unique frequency of the class. And based on the MAP, where  $H_{\text{Bayes}}$  is the maximum accuracy of the class, the final equation for Bayes is represented with equation (3):

$$H_{\text{Bayes}}(d') = \underset{C_j \in C}{\operatorname{argmax}} \frac{\Pr(C_j) \prod_{i=1}^{|d'|} \Pr(w_i | C_j)}{\sum_{C_j \in C} \Pr(C_j) \prod_{i=1}^{|d'|} \Pr(w_i | C_j)} \quad (3)$$

Where  $H_{\text{Bayes}}$  is Bayesian value for input  $d'$ ,  $\Pr(C_j)$  is  $\frac{|C_j|}{|D|}$ ,  $|C_j|$  is training sets amount for j category,  $|D|$  is the number of all documents in every category,  $\prod_{i=1}^{|d'|} \Pr(w_i | C_j)$  is the probability over the i-word frequency inside the document with j category and  $|d'|$  is the word frequency in a document.

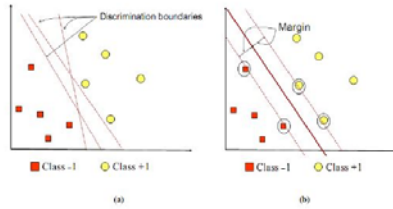


Figure 2. SVM classification with hyperlane [9]

Based on the Figure 2, SVM uses the linear classification especially when finding the line which separates two classes (hyperlane).

#### 4. RESULTS AND ANALYSIS

This section presents the results obtained by each algorithm used for classification in Rapidminer, with 14,974 tweets gathered.

Table 1. Algorithm Accuracy and Performance

Algorithm Name	Accuracy	Time Performed (s)
SVM	83.33%	26
Naive Bayes	72.22%	5
Decision Tree	66.27%	218

The table shows the algorithm used for creating the classification with the attribute of accuracy and time to process and classify the tweets. Based on the table, SVM holds the highest accuracy rate and the research concludes that SVM is the best algorithm, with industrial standard accuracy and without taking too much time to perform. The accuracy results are obtained from manual comparison through analysis of true positive/negative and false positive/negative, when the testing data set being applied to machine learning model. For instance, the testing set has a tweet “Telkomsel lemot” which has negative sentiment, however, machine learning model recognize as the opposite. This result of mistakenly identified true positive/negative and false positive/negative decreases the accuracy performance as whole.

The research also has a result for gathering statistics between the tweets and the words which represent with each mobile operators. Based on the tweets gathered, a balanced portion with randomly









## REFERENCES

- [1] Lima, Ana C. E. S. and de Castro, Leandro N. "Automatic sentiment analysis of Twitter messages". *IEEE* 2012; (): 52-57.
- [2] Stelzner, M.. "2012 Social Media Marketing Industry Report, How Marketers Are Using Social Media to Grow Their Businesses". Social Media Examiner. Report number: 1, 2012.
- [3] Pang, Bo, Lee, Lillian. "Opinion Mining and Sentiment Analysis". *Foundor Trends Information Retrieval* 2008; 2(1-2): 1-135.
- [4] Feldman R, Sanger J. "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data". Cambridge: Cambridge University Press; 2006.
- [5] Liu B. "Sentiment analysis and subjectivity". *Handbook of Natural Language Processing* 2010; (): 627-666.
- [6] Pang B, Lee L, Vaithyanathan S. "Thumbs up?: sentiment classification using machine learning techniques". *Proceedings of the ACL-02 conference on Empirical methods in natural language processing* 2002; 10(EMNLP '02): 79-86.
- [7] Dave K, Lawrence S, Pennock D. "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews". 2003; (): 519-528.
- [8] Asy'arie A, Pribadi A. "Automatic news articles classification in Indonesian language by using Naive Bayes Classifier method". *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services* 2009; iiWAS '09(): 658-662.
- [9] Nugroho A.S, Witarto A.B, Handoko D. "Support Vector Machine - Teori dan Aplikasinya dalam Bioinformatika". 2003.
- [10] Kwak H, Lee C, Park H, Moon S. "What is Twitter, a social network or a news media?". *Proceedings of the 19th international conference on World wide web* 2010; WWW '10(): 591-600.

## BIBLIOGRAPHY OF AUTHORS

	<p>Hansen Januar Fahrezasandy Wijaya  hansen.wijaya@student.sgu.ac.id  hansen.januar@gmail.com  8<sup>th</sup> Semester Student in Information System,  Faculty of Engineering and Information Technology,  Department of Information Technology,  Swiss German University  EduTown, BSD City, Tangerang, Indonesia</p>
	<p>Alva Erwin, M. Sc.  alva_erwin@yahoo.com  Education:  Bachelor Degree: Trisakti University, Jakarta, Indonesia  Master Degree: Pelita Harapan University, Tangerang, Indonesia  Doctor Candidate: Curtin University of Technology, Australia  Occupation: Lecturer of Information Technology Department at SGU</p>
	<p>M. A. Amin Soetomo, D. Sc.  mohammad.soetomo@sgu.ac.id  Education:  Bachelor Degree: University of Indonesia, Depok, Indonesia  Master Degree: George Washington University, Washington DC, USA  Doctoral Degree: George Washington University, Washington DC, USA  Occupation: Lecturer and Head of Master of Information Technology at SGU</p>
	<p>Dr. Maulahikmah Galinium, M. Sc.  maulahikmah.galinium@sgu.ac.id  Education:  Bachelor Degree: Swiss German University, Tangerang, Indonesia  Master Degree: Lund Universitet, Lund, Sweden  Doctoral Degree: University of Rome, Tor Vergata, Italy  Occupation: Lecturer and Deputy Head of Information Technology Department at SGU</p>