Text Mining of Twitter Data for Public Sentiment Analysis of Staple Foods Price Changes

Isti Surjandari[1], Muthia Szami Naffisah[1, 2], and M. Irfan Prawiradinata[1,2]
[1]Department of Industrial Engineering Faculty of Engineering, Universitas Indonesia
[2]Department of Economics, Universitas Indonesia
Email: isti@ie.ui.ac.id; {muthiaszami, irfan.prawiradinata}@yahoo.com

***Abstract*—Millions of users share their opinions on twitter, making it valuable platform for tracking and analyzing public opinion. Such analysis can provide critical information for decision maker in various domains. In this study, we examine public sentiment analysis of staple foods price changes in Indonesia based on twitter data. Text mining was used for classifying tweets into positive and negative sentiment. Then association between the type of staple foods and sentiment classes were analyzed using Chi Square test and Marascuillo procedure. Results show that Support Vector Machine (SVM) classifier produce higher accuracy than Na ïve Bayes and Decision Trees. Also, the price of milk, eggs and red onion had the most significant association to the negative sentiment compared to other commodities.**

***Index Terms*—text mining, sentiment analysis, marascuilo, food prices, twitter**

I. Introduction

Your The increasing amount of digital data is estimated to continuously grow by 40% annually. Unfortunately, there was only less than 3% of it being analyzed in 2012 [1]. Analysis of very large amount of data, or better known as *Big Data*, can be done using *Data mining*.

Text Mining is a variation of data mining that extract information from structured data. Text mining focuses on textual data, which is unstructured and difficult to deal with algorithmically, however, text is commonly used for information exchange by the today's society [2]. Large amount of textual data comes from social media and email. People post real time messages about their opinions on a variety of topics, making it valuable platform for tracking and analyzing public opinion on current situation, including their opinion on the food price changes. Despite greater interconnectivity, local impacts of shocks like food crises may not be immediately visible and trackable by traditional monitoring systems. So that it is often too late and more expensive to respond.

Opinion mining and sentiment analysis are related fields sharing common goals of developing and applying computational techniques to process collections of opinionated texts or reviews.

Timmer in 1996 has examined that there is no country that can sustain its economic growth without first solving its food problem [3]. The 2012 Global Food Security Index released by *Economic Intelligent Unit* has found that Indonesia's food security index is below 50 on a scale of 0-100. One of the causes of this condition is due the trend of increasing food commodities' prices. Considering the average percentage of expenditure on food consumption is relatively high in Indonesia (i.e., 47.71% of the total income per capita spent for food consumption [4]), therefore, a change in the price of staple foods is a problem that needs to be examined. This study attempt to examine public sentiment analysis of staple foods price changes on twitters data.

II. Methods

A. Text Mining

Text mining is a process of extracting unstructured information from a set of textual data [5]. In this study, public opinion about the changes of staple food prices in Indonesia was obtained from twitter. Twitter was selected because of its reputation as the world's and Indonesia's most popular microblog [6], [7].

TABLE I. Staple Foods

No	Staple Foods
1	Rice
2	Sugar
3	Cooking oil and butter
4	Chicken and beef
5	Egg
6	Milk
7	Corn
8	Kerosene and Liquefied Petroleum Gas
9	Salt
No	**Volatile Foods**
1	Red onion
2	Cayenne (red chili)
3	Rice
4	Orange
5	Chicken
6	Beef

Tweet is a message posted in *twitter website*. Tweets that are relevant to the staple foods subject are collected

Manuscript received July 1, 2012; revised November 5, 2014.

and processed using text mining, so that the textual data in the form of tweet can be analyzed numerically. Tweet collection is done by downloading messages from the twitter using an automated program that will scan twitter pages, and then create an index of data [8]. This study used *Scraperwiki*, which is an online tool to scrape data from multiple sources and copy it into a database [9]. Downloading data from tweeter has a time limit, i.e., data that can be downloaded only tweets within one week earlier [10]. For the purpose of this study, data is downloaded for seven weeks from April 14 until June 1, 2014.

Staple foods, as stated by the decree of Minister of Industry and Trade, Republic of Indonesia, are those commodities categorized as the basic needs of the society. Table I shows staple foods and other commodities that are classified as *volatile food* in 2013 [11].

Once tweets collected, they will go through a series of pre-processing steps to transform text data into numerical form of data, as shown in Fig. 1 [12].

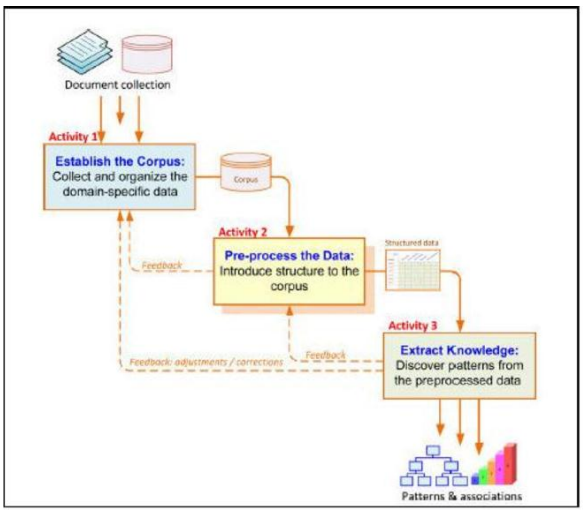

Figure 1. Text mining system diagram

Steps performed in the pre-processing stage along with the illustrations (Table II) are as follows:

- *Tokenization*, which separates the text in the *tweet* into pieces of word called *token*.
- *Filtering*, that is the elimination of mention (@), hashtag (#), and RT (retweet) from the tweets.
- *Stemming*, that is taking root words by eliminating the word affixes and transforming them into their simplest form.
- *Case folding*, that is changing all uppercases to lowercases in a document and vice versa. This step is also to ensure that only letters a through z are contained in the document.
- *Matrix*, which is a vector representation of word tokens based on the occurrences of words in the document [13]. In general, there will be three different matrix generated, i.e., the term frequency (tf), inverse document frequency (idf), and singular value decomposition (SVD). Those three were created using the following mathematical equations.

$$tf \cdot idf(t,d) = tf(t,d) \cdot idf(t) \quad (1)$$

$$tf(t,d) = \sum_{i \epsilon d}^{|d|} 1\{d_i = t\} \quad (2)$$

$$idf\ (t) = log \frac{|D|}{\sum_{d \epsilon D}^{|D|} |t \epsilon d} \quad (3)$$

Equations 1, 2 and 3 produce a high dimension of *term frequency-inverse document frequency*. To that end, *Singular Value Decomposition* (SVD) is used to reduce the number of variables by transforming correlated variables into a set of uncorrelated variables, which will reveal relationships contained in the original data [13], [14]. These uncorrelated variables will later be referred as *concepts*.

TABLE II. ILLUSTRATION OF PRE-PROCESSING STEPS

Before pre-processing	@mszami I look forward to read the study about Text Mining in the area of staple foods price changes topic
Tokenization	@mszami - I - look - forward - to - read - the - study - about - Text - Mining - in – the – area – of - staple - foods - price - changes – topic
Filtering	I - look - forward - to - read - the - study - about - Text - Mining - in – the – area – of - staple - foods - price - changes – topic
Stemming	I - look - forward - to - read - the - study - about - Text - Mine - in – the – area – of - staple - food - price - change – topic
Case folding	i - look - forward - to - read - the - study - about - text - mine - in – the – area – of - staple - food - price - change – topic
Matrix	***Terms*** / ***documents*** / i / look / . . . / topic; 1st Tweet / 1 / 1 / . . . / 1

B. Sentiment Analysis and Classification

Analyzing mood on opinions is called *Sentiment Analysis*, which refers to the automated valuation of text sentiment by observing their sentence polarity [15]. Document classification will separate positive from negative sentiments by building word occurrence probability model based on pre-classified documents [16].

Classification is done by first collecting *training data*, i.e., those tweets that have been labeled positive or negative. Training data will be analyzed using *Machine Learning* algorithms in which they will learn the patterns of the training model using predetermined algorithms. The next step is to classify the unlabeled documents using the trained classification model. Some algorithms that commonly used in sentiment classification are Support Vector Machine (SVM), Na ïve Bayes and Decision Tree.

- *Support Vector Machine* (SVM) algorithm is a technique for regression and classification. It is

categorized as supervised learning, which requires pre-classified document as training model. SVM is geometrically described as a hyperplane that separates document into two groups of data, i.e., positive and negative opinion. Hyperplane chosen is the one with maximal distance between the hyperplane and the two groups' nearest point [5]. SVM algorithm will search for the optimum function (i.e., the hyperplane) to separate those two sets of data [17]. Hyperplane with the maximum margin of SVM is shown in Fig. 2 [12].

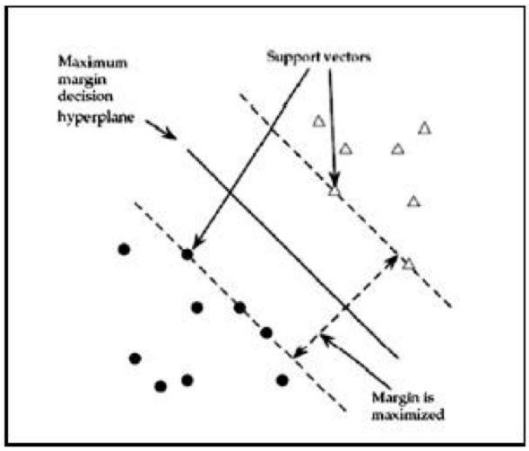

Figure 2. Hyperplane SVM

- *Naïve Bayes*. It is based on Bayesian theorem where class proportions in a data represents the probability of new object being categorized to certain class. It is assumed that all objects are classified independently to each other. This simple algorithm is proven to give an efficient and accurate result, especially in the case of high number of variables [12].
- *Decision Tree*. It is known as a symbolic algorithm, which is easy to interpret, by human [5]. Using a series of test questions in the form of tree, this algorithm will classify objects into categories. The decision tree structure will consist of: root, internal, and terminal nodes, as can be seen in Fig. 3 [18].

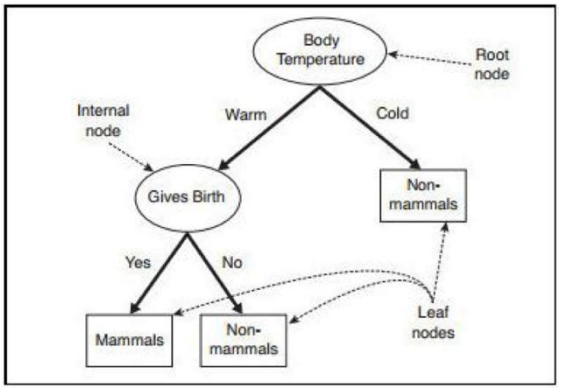

Figure 3. Example of decision tree in mammalia case

To quantify the accuracy of the trained models, the models were used to perform classification on testing data, i.e., pre-classified data. The accuracy is determined by how many sentiment classifications were correctly obtained from the trained model by referring to the pre-classified testing data. Accuracy calculation of the model is based on two types of testing data, the first use data that has been through the stemming process, and second uses data without stemming. The reason is to find out whether or not stemming has an effect on improving the accuracy of the classification models.

C. K-Proportion Test

The goal of this phase is to find out the amount of messages for each sentiment group on each type of staple foods. Tweets with negative sentiment will be counted and its proportion to the total number of messages for each staple food will be calculated.

Chi Square test was then conducted to determine whether there is a significant difference between negative sentiment proportions of each staple food. If there is a significant difference, then Marascuilo procedure will be performed to check which staple food causes these differences.

In *Marascuilo* procedure, the absolute difference between combinations of proportion that may occur between populations are being compared, as formulated in Equation 4 with the critical range obtained from Equation 5 [17]. Significant difference of proportions is indicated by the greater absolute difference than the value of r_{ij}.

$$Absolute\ difference = |p_1 - p_2| \tag{4}$$

$$r_{ij} = \sqrt{{\chi^2}_{1-\alpha,k-1}} \sqrt{\frac{p_i(1-p_i)}{n_i} + \frac{p_j(1-p_j)}{n_j}} \tag{5}$$

III. Result and Analysis

Classifications are done using three different algorithms (i.e., Naïve Bayes, Support Vector Machine (SVM), and Decision Tree), so that their accuracies can be compared. Table III shows that SVM algorithm gives the highest accuracy rate, while classification model with stemming is found to be more accurate than that of without stemming. Hence, SVM algorithm is used as a classifier algorithm to classify the unlabelled tweets using trained model with stemming process (i.e., data deployment) as shown in Fig. 4.

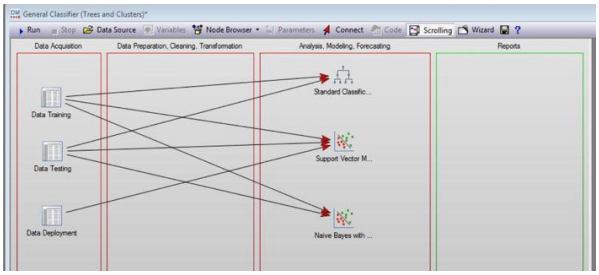

Figure 4. Process classification with the three algorithms

As all tweets have been classified to sentiment classes, the total amount of tweets in each class can be discovered

as shown in Table IV. Generally, people tend to upload their negative posts rather than their positive ones for all kind of staple foods. Negative post describes the increase in price, price volatility or the inability to purchase certain kind of staple food with the current price, while positive posts illustrate the opposite situations.

TABLE III. CLASSIFICATION MODEL ACCURACY

Algorithm	Model Accuracy without Stemming (%)	Model Accuracy with Stemming (%)
Naïve Bayes	65.76	72.23
SVM	75.19	80.35
Decision Tree	53.99	54.22

TABLE IV. TWEET CLASSIFICATION BASED ON SENTIMENTS AND TYPE OF STAPLE FOODS

No	Commodities	Positive	Negative
1	Chicken	555	1882
2	Orange	212	561
3	Cooking oil	111	104
4	Kerosene	5	46
5	Salt	57	81
6	Corn	436	480
7	Liquefied Petroleum Gas (LPG)	196	784
8	Sugar	394	689
9	Cayenne (red chili)	1088	526
10	Rice	645	803
11	Red onion	84	1305
12	Beef	267	3714
13	Milk	24	411
14	Egg	161	2727
15	Butter	0	0
	Total	4235	14113

TABLE V. CHI SQUARE TEST

Chi-square (Observed value)	4125.624
Chi-square (Critical value)	22.362
DF	13
p-value	< 0.0001
Alpha	0.05

There was neither negative nor positive tweet about butter price within the period of observation. This could describe that there is no significant change in the price of butter reported by public through twitter. Thus, butter price will no longer be included in the next comparative analysis of proportion because this group of tweets doesn't have any proportion to be compared with other staple food's proportion.

As the amount of tweets for each sentiment class has been known, then Chi square test can be performed, as shown in Table V. It shows that there is at least one proportion that is different with that of other proportions based on *significance level* of 5%. In this case, the null hypothesis is rejected when p-value is smaller than 0.05.

Marascuilo procedure is then performed after the Chi Square test has been conducted. This procedure is done by doing pairwise comparison to see if there is any difference between any particular combinations as illustrated in Table VI.

TABEL VI. MARASCUILO PROCEDURE

Comparison	Value	Critical Value	Significance
|p(1) - p(2)|	0.047	0.086	No
|p(1) - p(3)|	0.289	0.166	Yes
|p(1) - p(4)|	0.130	0.201	No
:	:	:	:
|p(13) - p(14)|	0.001	0.056	No

The significance difference will be used as a reference in classifying the staple foods. Classification is based on a comparison of the effect on sentiment of the tweets and also on the order of the group's influence on the emergence of negative sentiment of a type of staple foods. Table VII shows the result of classification results.

TABLE VII. CLASSIFICATION BASED ON VALUE OF SIGNIFICANCE AND PROPORTION

No	Commodities	Proportion	Group				
9	Cayenne	0.326	A				
3	Cooking oil	0.484	A	B			
6	Corn	0.524		B			
10	Rice	0.555		B			
5	Salt	0.587		B	C		
8	Sugar	0.636		B	C		
2	Orange	0.726			C	D	
1	Chicken	0.772			C	D	
7	LPG	0.800				D	
4	Kerosene	0.902				D	E
12	Beef	0.933					E
11	Red onion	0.940					E
14	Egg	0.944					E
13	Milk	0.945					E

Group A indicates commodities that have least influence on the emergence of negative public sentiment, followed by Group B, C, D and E that have the most influence.

Grouping a staple food in two different groups (e.g. cooking oil in both Group A and Group B) shows that there is no significant difference between the proportion of corn from Group B and cooking oil (i.e., |p(3) - p(6)|). So does the proportion of cooking oil compared to cayenne from Group A (i.e., |p(3) - p(9)|). Therefore, cooking oil can be classified in both groups. However, the proportion of corn from Group B compared to cayenne from Group A indicates a significance different (i.e., |p(6) - p(9)|), so that cayenne will remain in a different group with corn.

Table VII also shows that price of milk, egg and red onion have the highest effect on the occurrence of negative tweets, which means that people are sensitive to the price changes of these commodities. This is understandable since these commodities have relatively high price elasticity for Indonesian people.

IV. Conclusions

The results of this study show classification of staple foods according to their influence on public sentiment that is reflected through tweets that they upload. Prices changes of milk, egg and red onion are the three commodities that have the highest contribution in the emergence of negative sentiment of tweets.

The result of this study can be used by the food industry and government in evaluating the effect of price changes of food commodities in the society, as well as a basis for decision-making or policies related to the staple food. The voice of people in society as the primary customers of staple foods is the ultimate tool to find out the possibility of demand in the future, the purchasing trend of certain types of commodities, as well as price and supply crisis that possibly happen in the community.

References

[1] W. Fan and A. Bifet, "Mining big data: Current status, and forecast to the future," *SIGKDD Expolorations*, vol. 14, pp. 1-5, 2012..

[2] I. H. Witten, "Text mining," in *The Practical Handbook of Internet Computing*, M. P. Singh, Ed. Danvers, MA: Chapman and Hall/CRC, 2005, ch. 14, pp. 314-341.

[3] B. Amang and M. Sawit, *Kebijakan Beras dan Pangan Nasional: Pelajaran Orde Baru dan Orde Reformasi*, 2nd ed. Bogor, Indonesia: IPB Press, 2001.

[4] Badan Pusat Statistik. Persentase Pengeluaran Rata-rata per Kapita Sebulan Menurut Kelompok Barang, Indonesia, 1999, 2002-2013. [Online]. Available: http://www.bps.go.id/tab_sub/view.php?tabel=1&daftar=1&id_subyek=05&notab=7

[5] R. Feldman and J. Sanger, *The Text Mining Handbook: Advances Approaches in Analyzing Unstructured Data*, New York, NY: Cambridge University Press, 2007.

[6] Y. A. Pudyastomo, *Microblogging Paling Populer! (Gaya Hidup Orang Modern)*, Yogyakarta, Indonesia: Penerbit MediaKom, 2009.

[7] Invonesia. 10 Media Sosial Paling Populer di Indonesia. (March 1, 2014). [Online]. Available: http://www.invonesia.com/10-media-sosial-paling-populer-di-indonesia.html

[8] Wisegeek. What is a Web Crawler? [Online]. Available: http://www.wisegeek.org/what-is-a-web-crawler.htm

[9] A. Spillane. (2010). Online Tool Helps To Create Greater Public Data Transparency. [Online]. Available: Politico. http://politico.ie/index.php?option=com_content&view=article&id=6906:online-tool-helps-to-create-greater-public-data-transparency&catid=193:science-tech&Itemid=880

[10] Scraperwiki. Help. [Online]. Available: https://scraperwiki.com/help/twitter-search/

[11] Bank Indonesia, *Laporan Tahunan Perekonomian*, Jakarta: Bank Indonesia, 2013.

[12] G. Miner, *et al.*, *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*, Waltham, MA: Academic Press, 2012.

[13] K. Baker, "Singular value decomposition tutorial," Dept. Lingusitic, Ohio State University, OH, 2013.

[14] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.

[15] Mashape. List of 20+ Sentiment Analysis APIs. (2013). [Online]. Available: http://blog.mashape.com/post/48757031167/list-of-20-sentiment-analysis-apis

[16] NIST/SEMANTECH. *e-Handbook of Statistical Methods*. [Online]. Available: http://www.itl.nist.gov/div898/handbook/prc/section4/prc474.htm

[17] B. Santosa, "Tutorial support vector machine," Institut Teknologi Sepuluh Nopember - Open Content, 2013.

[18] Tan, *et al*., *Introduction to Data Mining.* Boston, MA: Pearson Addison Wesley, 2004.

Isti Surjandari is a Professor and Head of Statistics and Quality Engineering Laboratory in the Department of Industrial Engineering, Faculty of Engineering, Universitas Indonesia. She holds a bachelor degree in industrial engineering from Universitas Indonesia and a Ph.D degree from the Ohio State University. Her interests are in the area of industrial management, quality engineering, statistics analysis and data mining.

Muthia S. Naffisah is a research associate at the Statistics and Quality Engineering Laboratory, Industrial Engineering Department, Universitas Indonesia. She holds a bachelor degree in industrial engineering from Universitas Indonesia. Her interest is in statistics and data mining.

Irfan Prawiradinata is now pursuing his bachelor degree in Economics from Universitas of Indonesia. His interest is in poverty and public policy.