eJAICT
ISSN 2355 1771
JOURNAL OF APPLIED INFORMATION, COMMUNICATION AND TECHNOLOGY
Published by:
Research Center of Technology and Innovation
Academic Research and Community Services
Swiss German University

Study Of Automotive Brands Popularity In Indonesia Using Twitter Data

Stevent Efendi, Alva Erwin, Kho I Eng

Faculty of Engineering and Information Technology, Swiss German University

Article Information

Received: 12 January 2016
Accepted: 18 March 2016
Published: 25 April 2016
DOI:

Corresponding Author:

Stevent Efendi
Email: stevent.efendi@student.sgu.ac.id

ISSN 2355-1771

ABSTRACT

Social media has been a widespread phenomenon in the recent years. People shared a lot of thought in social media, and these data posted on the internet could be used for study and researches. As one of the fastest growing social network, Twitter is a particularly popular social media to be studied because it allows researchers to access their data. This research will look the correlation between Twitter chatter of a brand and the sales of brands in Indonesia. Factors such as sentiment and tweet rate are expected to be able to predict the popularity of a brand. Being one of the biggest industries in Indonesia, automotive industry is an interesting subject to study. A wide range of people buys vehicles, and even gather as communities based on their car or motorcycle brand preference. The Twitter results of sentiment analysis and tweet rate will be compared with real world sales results published by GAIKINDO and AISI.

Keywords: *Sentiment Analysis, Data Mining, Twitter Mining, Automotive Industry.*

1. Introduction

Social media has become a part of lifestyle in the community. Since the beginning of their existence, social medias have millions of contents created by people. Because of their ease of use, speed and reach, social medias become a mean of giving and receiving updates on a wide range of topics such as politics, technology and entertainment. These makes them a form of collective wisdom[1].

These collective wisdom, if can be analyzed to infer public attitude the same as traditional polls, will be a faster and less expensive alternative to collect public opinion[2]. It is also possible to aggregate the highly varied information from the large user communities to build models of public opinion and predict future trends. Moreover, these data could also be used to help designing marketing and advertising campaign[1], [3], [4].

This paper uses Twitter, one of the fastest growing social media, as a medium to predict the popularity of a brand. As a micro-blogging network, Twitter had a burst of popularity leading to tens of millions users actively creating contents. These contents are called tweets and have been used as data sources for social studies such as opinion mining and analyzing sentiment towards low cost green cars[5] and political parties[6] in Indonesia. Some researches have also utilized the prediction power of Twitter. It was proven possible to use Twitter to predict box office movies revenue [1], predict the political polarization of the people[7], and predicting the stock market performance[8].

In Indonesia, Twitter has become widely used with the number of users that reached twenty nine million users by March 2013 (Lukman 2013). These users post millions of tweets daily and these tweets could be used as a research data source. However, there haven't been many researches published that utilize Twitter in Indonesia.

In this research, the object of the study is the automotive industry in Indonesia. Automotive industry is one of the biggest industries in Indonesia. Figure 1 shows sales and growth of car sales in Indonesia up until May 2014. The blue bars show the number of unit sales in thousands, and the red line shows growth from previous year. Although there are some declining years, from the year 1993 to 2013, a significant growth in number of sales could be seen. Motorcycle sales also see the same growth as cars as seen in figure 2. There are even automotive section in almost all Indonesian news sites such as kompas.com, detik.com, and viva.com. This enthusiasm makes automotive industry an interesting subject to study. Moreover, there haven't been much researches utilizing Twitter to study the automotive industry.

Fig. 1. Indonesian Car Sales YOY Growth and Number by Mid 2014 (Indoanalisis N.D.)

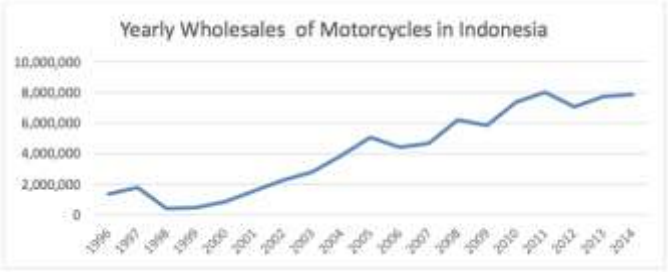

Fig. 2. Year to Year Sales of Motorcycle in Indonesia from AISI (AISI n.d.)

The objective of this research is to prove whether it is possible to predict the popularity of automotive brands in Indonesia using Twitter data. The study will involve both cars and motorcyles industry. After determining the popularity scores of the brands, we compare them with the sales report released by Gabungan Industri Kendaraan Bermotor Indonesia (GAIKINDO) and Asosiasi Industri Sepeda Motor Indonesia (AISI) which are posted on kontan.co.id to see if there are a correlation between the popularity and sales of the brand. These sales data are posted in Kontan.com.

There will be three categories to be researched on in this paper, which are:
- Motorcycles
- Japanese cars
- Luxury cars

The cars are divided into two categories because in the Japanese cars compete in different market than the luxury cars. Most Japanese cars have price range that is reachable by a lot of people and targets as many people as possible, and the luxury cars target executives in which consist of less number of person.

2. Related Work

Twitter has been used by several studies to conduct a research on social media. It has used been as sentiment analysis medium. Kasim [5] has performed sentiment analysis on automotive industry, specifically on Low Cost Green Cars in Indonesia. Gemilang [6] did the analysis on political party in Indonesia. In these researches, Twitter was used to get publics' opinions on the topics, and the results are classified in positive, negative or neutral sentiments. These works only analyze the polarization of the tweets and made no further analysis on them.

Twitter has also been used as a mean of prediction, such as political election results [9], stock market performance [8] and box office movie revenue [1]. In these studies, tweet rate with certain pre- determined keywords are used to predict the popularity of an object, with addition of sentiment analysis to support the results. According to Asur [1], positive sentiments only act as a support for the prediction and doesn't impact the result much compared to tweet rate.

In this research, Asur's [1] method of predicting movie revenue will be applied to study the Indonesian automotive market and see if it also applicable and able to see the relationship between automotive sales and Twitter. Although the data is not enough to determine the prediction power, this study could see a preliminary result whether the tweets have a relation to real world data.

3. Twitter

Twitter is a very popular micro-blogging website that was launched on July 13th, 2006. It has a very large user base, and by March 2015, it has 288 million monthly active user [10].

Each user in Twitter creates 140 characters maximum posts, which typically consist of personal opinion, news, and links to pictures, videos and articles. These posts are called tweets and are displayed on the user profile page as well as his/her follower's timeline. It is also possible to direct the posts to other users, either privately using direct message, or publicly by mentioning the targeted user using the format '*@userid'* in a tweet. Tweets could also be forwarded by a user to his/her follower in which the forwarded tweet is called a retweet.

Twitter allows developers to create application connected to it, and it provide an Application Programming Interface (API). The two main API of Twitter are Search API and Stream API. Search API allows developer to search through up to 7 days old tweets, whereas Stream API listens to new tweets in the public timeline. Both API enable filter through keywords, location, language, and username so the developer could retrieve the desired tweets. These retrieved tweets are returned in JSON Format.

4. Dataset Characteristic

The dataset we used was obtained by crawling Twitter using the Stream API. The retrieved data was in JSON format, which then converted and stored in MySQL database. The crawling was done by filtering through language and keywords to ensure the data retrieved was from Indonesia and in accordance to the study case, which is about brands in Indonesia. We stored the content of the tweets, timestamp and their author.

We collected data containing keywords of car brands in Indonesian language that are posted in April 2015 and compare it to the sales report of the month. In each category, we select brands that are popular and have sold many cars during the previous months. Table 1 shows the keywords we used to filter through tweets mentioning the brands in each category.

Table 1. Categories and Keywords Used

Categories	Keywords
Motorcycle	honda, kawasaki, suzuki, tvs, yamaha
Japanese Car	toyota, mitsubishi, daihatsu, nissan, datsun, honda, suzuki
Luxury Car	audi, bmw, bimmer, mercedes, benz, mercy, lexus

Since the brand 'BMW' and 'Mercedes-Benz' are often abbreviated as 'Bimmer' and 'Mercy' respectively, we used those keywords as filters to have more accurate number of tweets mentioning those brands.

5. Methodology

A. Analysis Construction

We first look for papers as bases and apply the methods to the research. We will define keywords and collect data from Twitter public timeline. After the data is collected, they will be classified and found out the tweet rate for each brands in the month. With the support of sentiment analysis, the popularity of the brands will be determined.

B. Data Collection

We use Twitter4j library to collect data. We created a program based on the library and run it for two months. Twitter4j is a third party library based on Twitter API version 1.1, which to be able to be run, needs authentication tokens from Twitter. These tokens were obtained from Twitter's developer website and consist of four tokens:

- Consumer Key (API Key)
- Consumer Secret (API Key)
- Access Token
- Access Tiken Secret

The tokens were put in properties file of Twitter4j. The crawler runs on the stream library of Twitter4j which was based on Twitter Stream API. The crawler were set to retrieve the following information of a tweet:

- Tweet ID
- Username of the owner
- Tweet content
- Time posted

We also add filters in the code to get the tweets relevant to the research. We made the keywords determined before, and Indonesian language as the filter. This will make Twitter4j to retrieve tweets containing one of the keywords, and in Indonesian language. The crawler could then run and start collecting data from Twitter. The crawler run for two months, April and May 2015.

C. Data Pre-Processing

The data collected will first need to be cleaned up before being processed for it to provide a more accurate results. In this step we remove duplicate tweets, classify the tweets into car and motorcycle tweets, and check for unrelated tweets. We don't remove retweets and advertising tweets because retweets and advertisement do affect the possibility of people to buy[1], [11]. Since there are brands that produce both car and motorcycle, the keywords used twice in the two instance of the application that run. Those brands are Honda and Suzuki. This makes the same tweets to be inputted twice and therefore there are duplicates. To remove the duplicates, we will discern the tweets by their ID and add 'unique' property to the ID column while removing the duplicate entries.

The brands that produce multiple categories of automotive will need to be separated, whether the tweet is talking about the car or motorcycle brand. To do this, we used

products of each brands that were mentioned with the brand in a tweet. Table 3.2 shows the list of products launched in Indonesia taken from Honda and Suzuki's official Indonesian websites [12]–[15].

Table 2. Products of Honda and Suzuki

Brand	Car	Motorcycle
Honda	Accord, City, Civic, Brio, CR-V, Jazz, Freed, Odyssey, CR-Z, Mobilio, HR-V	Blade, Revo, Supra X, Beat, PCX, Scoopy, Vario, CBR, MegaPro, Verza, CB150R, Spacy
Suzuki	Ertiga, Splash, APV, Vitara, Swift, Karimun, Carry	Nex, Hayate, Shooter, Axelo, Satria, Address, Inazuma, Thunder, Hayabusa, Burgman, GSR, V-Storm

Tweets mentioning Honda and Suzuki are labeled with NEU or neutral because it cannot be determined whether the tweets are talking about the car or motorcycle part of the brands. We consider these tweets contribute for popularity in both part of the brands. After labeling all as neutral, we start classifying the tweets to car and motorcycle. We set the label for tweets mentioning the word Honda and Suzuki, and one of their product as CAR and MTR based on the category of the products. We further separate the tweets between cars and motorcycles by labeling the brands that was mentioned with another car brands or motorcycle brands. If the Honda or Suzuki is mentioned with for example Toyota, then we label the tweet as talking about Honda's or Suzuki's car, and we applied the same to the motorcycle counterpart. The tweets are separated with mentions of words 'mobil' and 'motor' which are Indonesian words for car and motorcycle respectively. Since Honda participate in formula-1 race with the name McLaren-Honda, we look for the word 'mclaren- honda' and label them as 'car'. Honda also participate in MotoGP as Repsol Honda and Suzuki as Suzuki Ecstar, therefore we also look for the mentions of 'repsol', 'ecstar' and 'motogp' in the tweets and classify them as tweets for motorcycle.

Twitter Stream API will capture any tweets containing the keyword we specified. Therefore, some unrelated mentions might be included. This cannot be checked automatically so we have to perform manual check. While checking manually, we found that using the keyword 'mercy' where we intended for it to be abbreviation of Mercedes-Benz, there are tweets that use it as the real English meaning 'to forgive'. The word usually comes with 'God' or 'have', therefore we remove tweets that have these words alongside the word 'mercy'. We also found some reference to people's name such as the football player 'Keisuke Honda'. Other names that were found were 'Tsubasa Honda', 'Suzuki Emi', 'Suzuki Konomi', 'Suzuki Tatsuhisa', 'Yu Suzuki', and 'Moe Toyota'. There are also names from a popular comic book character named 'Sonoko Suzuki' found. Since Yamaha have music products, we looked for tweets with keywords related to music with Yamaha, such as 'music', 'drum', 'guitar', 'piano', and 'keyboard'.

D. Data Processing

Asur mentioned that the key factor to predicting from Twitter is to get the tweet rate, the number of tweets mentioning a keyword in a period of time, of the keywords we want to predict [1]. Therefore we process the data to show the number of tweets each brand got during the crawling period. The next step is to perform sentiment analysis. Using tweet

rate, Asur had already obtained coefficient of determination of 0.79. The number increased to 0.92 by adding sentiment analysis. We will find out how the coefficient will be in Indonesian automotive industry. The sentiment analysis will be done using a classifier application provided by Akon Teknologi.
Before being able to classify sentiments, the application will need a model as a base for classification. In creating the training set, we exported the tweets in the database into csv format, and manually label it with positive, negative, or neutral sentiment. In this case we took 1000 tweets to be the training set. We tried to choose tweets that are as varied as possible and that aren't a retweet. The classifier application was written in Java on Maven. This makes it require some configuration. Using the Netbeans IDE, the configuration will automatically be done. The application will look for a file as a training set, and classify the text based on it. The tweets that will be classified is placed on a file named listText.txt. The output of this application is supposed to be a text file, but we modified the program to directly update the table containing tweets in the database.

E. Evaluation

The data will be evaluated and compared to the real world sales published by GAIKINDO and AISI in news sites. In this research, the data is taken from www.kontan.co.id. The sales result will be compared with tweet rate to see if it is enough to see the correlation between Twitter popularity and the real world. Sentiment analysis will then be added to improve the accuracy. To do this, a score will be created by adding tweet rate and positive – negative ratio which will be calculated monthly.

To get the tweet rate score, the number of tweets obtained will be divided by the time it is collected, in this case one month which is 720 hours in April and 744 hours in May. The positive – negative ratio could be calculated by dividing the number of positive tweets with negative tweets.

$$Tweet-rate = \frac{|tweets|}{|Time\ (in\ hours)|} \qquad PNRatio = \frac{|Tweets\ with\ positive\ sentiments|}{|Tweets\ with\ negative\ sentiments|}$$

After finding out the scores and compare them with the real world sales data, the correlation measure will be found. The measure goes from -1 to 1, from negative correlation to positive correlation (MathisFun 2014).

$$r_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{n\Sigma x_i^2 - (\Sigma x_i)^2}\ \sqrt{n\Sigma y_i^2 - (\Sigma y_i)^2}}$$

Where:
r = Pearson's Correlation
n = Total individual sample or the sample size
x = total score from variable x
y = total score from variable y

Lastly, we will do content analysis. The content of the tweets will be analyzed to gain more insight. What people talked about a brand will be categorized and analyzed to see the popular traits of each brands. To ensure positive sentiments, words that contradicts the traits are excluded such as 'tidak', 'kurang' and 'tidak'. The classified conversation will be presented in pie charts to compare what people talked about most.

Table 3. Category of Conversation

Conversation Category	Explanation	Keyword
Performance	How the cars perform. Fuel economy, Speed, and Power.	Irit, tenaga, kencang, cepat
Price	Price of the vehichles. Cheap, and reachable.	Murah, sanggup, terjangkau, turun harga
Interest	Desire to own the vehicles.	Idaman, incar, ingin, pakai, punya, mau
Feature and Safety	Comfort and safety technology included in the vehicles	Airbag, aman, fitur, canggih
Design	Preference of design and looks of the vehicles	Bagus, keren, suka, cool, menyukai

6. Results

During our crawling in April and May 2015, we have found 931927 tweets using the keywords we determined. After we cleaned the data, the number got down to 700987 tweets and the table is 183.9 MB in size. We also found 150190 unique users in our database. This means that user to tweet ratio is 1:5.

We counted the tweet rates and the positive-negative ratios for each month and sum the results to obtain the popularity scores, and compare them to the real world sales.

Our finding shows that Twitter data has inaccuracies in reflecting the real world. In the Japanese car category for example, despite selling the second most number of cars in April, the score of Daihatsu is smaller in rank. In May, it was Honda that was out of place in the score. The same also happened to the luxury car category in which Lexus got the bottom rank in popularity score where in real world, it sold more than Audi. In the motorcycle category, although in April the rank was in order, the result was inconsistent in May, with TVS having a very big popularity score although having the lowest sales.

Table 4. Japanese Car Score and Sales

Brand	April				May			
	Tweet Rate	PN Ratio	Score	Sales	Tweet Rate	PN Ratio	Score	Sales
Toyota	68.20	5.31	73.52	30053	59.43	7.70	67.13	23223
Mitsubishi	36.82	16.86	53.68	9662	9.72	26.19	35.91	9126
Daihatsu	17.65	38.78	56.43	14855	22.42	19.60	42.01	14486
Nissan	38.97	2.55	41.52	1321	25.96	1.72	27.68	2091
Datsun	5.08	39.64	44.71	1711	21.79	9.55	31.34	2874
Honda	64.29	4.58	68.87	10583	67.92	20.06	87.98	11301
Suzuki	41.06	3.79	44.85	8019	33.27	6.12	39.39	10017

Table 5. Luxury Car Score and Sales

Brand	April				May			
	Tweet Rate	PN Ratio	Score	Sales	Tweet Rate	PN Ratio	Score	Sales
Audi	12.88	10.84	23.73	17	20.30	15.99	36.29	16
BMW	29.44	8.14	37.58	130	21.76	110.73	132.49	200
Mercedes Benz	38.23	4.23	42.46	236	56.07	11.61	67.68	566
Lexus	7.93	1.43	9.36	22	5.71	20.22	25.93	47

Table 6. Motorcycle Score and Sales

Brand	April				May			
	Tweet Rate	PN Ratio	Score	Sales	Tweet Rate	PN Ratio	Score	Sales
Honda	72.41	10.22	82.63	371584	83.07	22.88	105.94	304900
Kawasaki	14.43	2.10	16.53	5679	17.43	35.49	52.91	6330
Suzuki	42.86	7.07	49.93	11754	31.82	73.26	105.08	7355
TVS	2.31	0.23	2.54	796	6.13	503.40	509.53	300
Yamaha	62.42	4.26	66.68	158958	84.17	42.43	126.60	150745

On contrary to what Asur said, the PN-ratio in this research actually reduced the correlation between Twitter and real world, instead of supporting it. Table 4.13 shows the correlation between real world sales and tweets using mentions only and popularity score. Although by glance the relation improved with the addition of sentiment analysis, the correlation score actually less than comparing tweets and real world using only mentions. In a scale of 0 to 1 with 1 being perfectly related, in April the correlation was 0,61 without PN-ratio, and 0,58 with PN-ratio, and in May it was 0,7 and 0,28 respectively. While in April the difference is not much, in May the score had a really big difference. This means, in Indonesia, it is likely to be better to use only mentions to see the correlation between Twitter and sales data of automotive industry.

Table 7. Correlation Score

Month	Mentions (Monthly Tweet Rate)	Score (Tweet Rate + PN Ratio)
April	0,61	0,58
May	0,7	0,28
Average	0,65	0,43

When we tried to see the content, we found that people's interest towards a brand makes more of a deciding factor for them to buy a Japanese car and motorcycle. Whereas features come second. In the luxury car category though, brand that have more talks about feature and safety, and performance tend to sell more in real world. Participation in Formula-1 and MotoGP seems to also affect sales since brands that participate like Mercedes in Formula-1, and Honda and Kawasaki in MotoGP tend to sell more. In the word cloud of Mercedes and Honda motor, the biggest words were mostly about the Formula-1 and MotoGP.

Fig. 3. Word Cloud for Mercedes-Benz

Fig. 4. Word Cloud for Honda Motorcycle

7. Conclusion

In our research, we found that Twitter doesn't reflect real world sales of automotive industry in Indonesia. The result of the analysis is also inconsistent, with some categories being more related than others. Sentiment analysis also affected the result greatly and in negative way. Sentiment analysis actually reduced the Pearson correlation score.

Each category of vehicle has preference of factors in popularity. In the Japanese car and motorcycle category, customers usually already have interest in a brand when buying a car, whereas in Luxury car category, people look for cars that have a lot of feature and safety measure.

This shows that just by the surface, Twitter data cannot reflect popularity of Indonesian's automotive brands in real world, but there is a possibility to find more correlation when the contents are analyzed deeper.

References

Asur S., Huberman, B.A. (2010). *"Predicting the Future with Social Media,"*. *Appl. Energy*.

Bollen, J., Mao, H., Zeng, X.-J. (2010). *"Twitter mood predicts the stock market," J. Comput. Sci.* pp. 1–8.

Conover, M., Ratkiewicz, J., Francisco, M. (2011). *"Political polarization on twitter"*. *ICWSM*.

De Vries, L., Gensler, S., Leeflang, P.S.H. (2012). *"Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing"*. J. Interact. Mark., vol. 26, no. 2. pp. 83–91.

Gemilang, H.T., Erwin, A., I Eng, K. (2014). *"Indonesian Political Parties Sentiment Analysis By Using Twitter Data"*.

Honda, A. *"Produk"*. (2015). Online Available: http://www.astra-honda.com/index.php/produk/.

Honda. (2015). Indonesia. *"Product"*. Online Available: http://www.honda-indonesia.com/product.htm.

Jansen, B.J., Zhang M., Sobel, K., Chowdury, A. (2009). *"Twitter Power: Tweets as Electronic Word of Mouth". ASIS&T.*

Kasim, A.M. (2014). *"Sentiment Analysis of Indonesian Low Cost Creen Car with Twitter Data"*.

Leskovec, J., Adamic, L.A. (2008). *"The Dynamics of Viral Marketing," vol. 1, no. May 2007"*. pp. 1–46.

O'Connor, B., Balasubramanyan R. (2010). *"From tweets to polls: Linking text sentiment to public opinion time series". ICWSM.*

Statista. (2014). *"Leading social networks worldwide as of March 2015, ranked by number of active users (in millions)"*. Available: http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Suzuki. (2015). *"Automobile"*. Online Available: http://www.suzuki.co.id/automobile.

Suzuki. (2015). *"Motorcycle."*. Online Available: http://www.suzuki.co.id/motorcycle.

Tumasjan, A., Sprenger, T.O., Sandner, P.G., and Welpe, I.M. (2010). *"Predicting Elections with Twitter : What 140 Characters Reveal about Political Sentiment"*. ICWSM. pp. 178–185.