Dynamic Large Scale Data on Twitter using Sentiment Analysis and Topic Modeling

Case Study: Uber

Andry Alamsyah[1] , Wirawan Rizkika[2], Ditya Dwi Adhi Nugroho[3], Farhan Renaldi[4], Siti Saadah[5]
[1,2,3,4]School of Economics and Business, Telkom University
[5]School of Computing, Telkom University
Bandung, Indonesia
[1]andrya@telkomuniversity.ac.id, [2]wirawanrhrp@yahoo.com, [3]dityaadhi@outlook.com,
[4]farenaldi@yahoo.com, [5]sitisaadah@telkomuniversity.ac.id

***Abstract*— Digital flows now exert a larger impact, the world is now more connected than ever, the amount of cross-border bandwidth that used has grown 45 times larger since 2005. With the massive amount of data spreading in the net, including social media, speed is one most essential factor in business. companies can take advantage of social media as a source to analyze and extract the customer's opinion, and therefore the company can have quick response towards the condition. The main purpose of this research is content analysis, to obtain the goal, we need to extract the information as well as summarize the topic inside it. However, in order to analyze the content quickly, there are varies choice of tools with its specific output that creates challenges in the process. We use Naïve Bayes Sentiment Analysis based on time-series, specifically on daily basis and topic modeling based on Latent Dirichlet Allocation (LDA) to evaluate the sentiment of the topic as well as the model of the topics discussed. This research may help both companies and individuals to map the public opinion towards certain topic by analyzing the sentiment of the text and create a topic model. Therefore, a real-time information for determining the consumer opinion become a crucial part. Twitter can serve the purpose as one source of real-time information from user-generated content. We pick Uber as the case study, viewed as one of the most favored transportation methods in most part of the world. Data collection period is from 10th February 2017 until 28th February 2017 with 1.048.576 tweets collected.**

Keywords—Sentiment Analysis; Topic Modeling; Latent Dirichlet Allocation; Big Data; Summarization; Content Analysis.

I. INTRODUCTION

User opinion in social media is always changing. Currently, there are many companies that use public opinion to be able to achieve the goals of the company. 20-30 companies in the United States of America offer sentiment analysis as one of the tools to help corporate decision-making [1]. Therefore, we use dynamic sentiment to discover additional information. This research uses dynamic sentiment because it can gather more precise and detailed result.

Along with the increasing number of internet, social media users and mobile devices will certainly impact on the increasing amount of data or user-generated content [2]. The simple forms to collect opinions are and retrieve data from social media [3]. With a massive information flows from social media, a highly effective approach is needed to summarize and retrieve information in a real-time situation.

Several classification methods are suitable to analyze the data, such as Support Vector Machine (SVM), Naïve Bayes (NB), Nearest Neighbors (NN), and Decision Tree. The model used to summarize is Naïve Bayes classification method. Considering several classifiers that we have known, Naïve Bayes has been widely used because of its simplicity in both training and classifying stage [4]. Naïve Bayes method allows each attribute towards the final decision equally and independently from the other attributes. The Naïve Bayes classification methodology is applied for the reason of its high-level accuracy and support large data processing [5]. Hence, it is more efficient and more accurate compared to other classifiers.

Despite their Naïve design and apparently oversimplified assumptions, Naïve Bayes have worked relatively well in many complex real-world situations [6]. Naïve Bayes also have advantages in term of training data, Naïve Bayes only requires a small number of training data to estimate the parameters necessary for classification.

We use topic modeling to determine topics that contained in the data. Numerous way of topic distribution is applicable such as Clustering, Feature Generation, and Dimensionality Reduction. Dimensionality reduction has an advantage compared to the other because each document's distribution over topics gives a summary of the document. Compare them in this reduce feature space can be more meaningful than comparing in the original feature space [7].

Twitter is one of such social media that has become a prominent source to exchange the online text, providing a vast platform for sentiment analysis and topic modeling. Twitter is a very popular social networking website that allows registered user to post short messages, also called tweets. Twitter database is one of the largest database having 200 million users who post 400 million tweets in a day [8]. Uber is one of the largest and greatest innovation in transport with its fast development, the interaction among user in the platform is high.

In this research, we aim to dig further information from the dataset. We see the topics from the public opinion perspective in Twitter social media, especially their opinion regarding Uber, seen from the dynamic of the social media, the sentiment is surely changing every day then which topic has

positive or negative sentiment. This is useful for data processing to be more effective and fast.

II. Content Analysis

A. Sentiment Analysis

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinion, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. Automated sentiment analysis is mandatory because the average human reader has difficulty in identifying relevant sites, extracting and summarizing the opinions in them [8]. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks. Sentiment analysis is being applied in most business and social domain because opinions are central to almost all human activities and are key influencers of our behaviors. Naïve Bayes models allow each attribute to contribute towards the final decision equally and independently from other attributes, in which it is more computationally efficient when compared with other text classifiers. [4].

B. Topic Modeling

Text documents composed of words, a topic that is mentioned in multiple documents can be expressed by a combination of strongly related words. Each document consists of multiple topics. Topic Modeling is a technique used to infer hidden topics in text documents. Topic modeling represents each document as a complex combination of multiple topics and each topic as a complex combination of multiple words, it is also used as text mining tool to classify documents based on topic inference results [10]. Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [11].

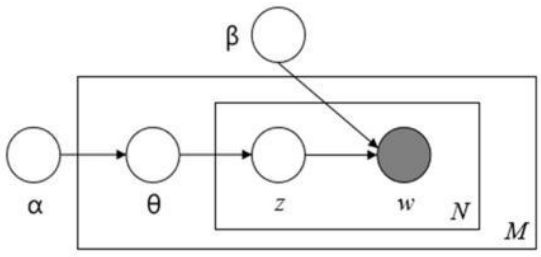

Fig. 1. Graphical model representation of LDA

N represents the number of words in the document, *M* is the number of documents to analyze, α stands for the Dirichlet-prior concentration parameter of the per-document topic distribution. β is the same parameter of the per-topic word distribution. θ is the topic distribution for the document. *z* is the topic assignment for *w*, *w* is the x-th word in the y-th document. Boxes are the "plates" that represents replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document [11].

III. Methodology and Experiment

We break down the research into 4 stages, it begins with data collection, preprocessing, main process and summarization. Data collection describe the process of data crawling. Preprocess describe the process of preparing data. The main process is where the data extraction occurs. The last stage is summarization process, the analysis of the extraction result. In this research, we pick one day that has the highest negative sentiment. We consider it interesting to explore the topics in the highest point of negative sentiment. Compare the result with positive sentiment in the same day, to map out the issue or topics that evolve that day.

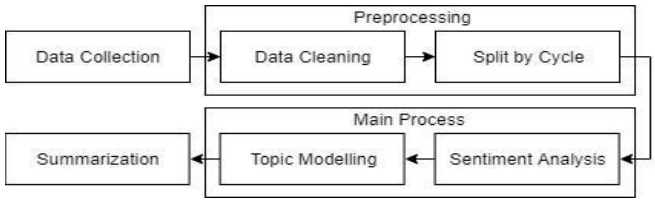

Fig. 2. Research workflow

A. Data Collection Process

Data Collection process uses Twitter Data Streaming API. First, we define the keyword related to the topics. We use keyword "Uber" to collect the data from 10th February 2017 until 28th February 2017 with 1,048,576 tweets collected in JSON format.

B. Data Preprocessing

1) Data Cleaning

Data Cleaning is the most important yet time-consuming step during the process. Pre-processed data is mandatory in order to have them ready to be processed and reducing data noises. It affects the end-result of this analysis. Improper data cleaning might cause multi-interpretation that impact on lack of accuracy. The first process to clean data begins with manually remove all text that is non-English text, remove non-ASCII characters. Then it continues to transform cases that intend to lowercase all characters in the data. Tokenize to separate sentences into a word. Stem is done to refer the word to its original form. English stopword is used to remove common words such as I, You, We, They, etc.

2) Split by Cycle

Depending on the time, every Tweet can have different topic and sentiment. In order to extract the detailed information on sentiment in the data, we split the data into a daily cycle or per day (24 hours). The purpose of this process is to analyze the topic (and or opinion) to be compared if there are any changes in tweets sentiment and topics regarding Uber.

C. Main Process

There is 2 main process in this research. It consists of sentiment analysis and topic modeling. Each process has its specific function. Sentiment Analysis produces the predicted sentiment; topic modeling produces the mapping of the topic. Sequential process is mandatory to map the topic based on sentiment

1) Sentiment Analysis

The results represent the sentiment of the topic e.g. attitude, thought, judgment or a specific view towards an opinion. We split the main data into 2 parts, which is training data and testing data, with the percentage of 30% and 70% [4] of the total dataset amount respectively. The training data consist of 2 labels, positive and negative. Yet, the tool classifies the result into 3 categories. Several texts that is not suitable for a positive or negative category. The unclassified data are considered as neutral.

TABLE I. SENTIMENT SAMPLE

Sentiment	Text
Positive	I like using @uber, the driver was nice to me :D
Negative	I will never use @uber again for my whole life
Neutral	http://e2hh376 @uber

We transform the output into numerical form, the computer refers at the amount of the confidence score, to define the prediction of the sentiment, which the score of each text maximum 1 or 100% confidence.

TABLE II. SENTIMENT RESULTS SAMPLE

Prediction	Confidence (Positive)	Confidence (Negative)	Confidence (Neutral)	Text
Positive	1	0	0	wow service is great
Neutral	0.001	0	0.999	rt alex uber is online transportation
Negative	0	1	0	The driver is drunk

2) Topic Modeling

To map the topic, we separate sentiment analysis result into 2 files (positive and negative). The neutral classified text is not used since it contains a lot of data noises. Then list the most probable terms within topics to summarize the topic. From the result, we use saliency (term's frequency) measurement method, these quantities measure on how much information a term conveys about topics by computing the Kullback-Liebler divergence between the distribution of topics given the term and optionally weighted by the term's overall frequency [12]. We use frequency parameter to count probability the word relation on a topic that noted with a weight of probability λ. The λ score is between $0 \leq \lambda \leq 1$ [12].

$$r(w,k \mid \lambda \log(\emptyset_{kw}) + (1 - \lambda) \log(\emptyset kw/Pw) \quad (1)$$

Where λ determines the weight given to the probability of term under topic relative to its lift (measuring both on the log scale). Setting $\lambda = 1$ results in the familiar ranking of terms in decreasing order of their topic-specific probability and setting $\lambda = 0$ ranks terms solely by their lift.

IV. RESULTS AND ANALYSIS

From the 19-day observation period, we produce 19 results. We present the result into graphical visualization to track daily movement dynamic sentiments. As shown in Fig. 3, it facilitates bigger picture on how the sentiments vary each day.

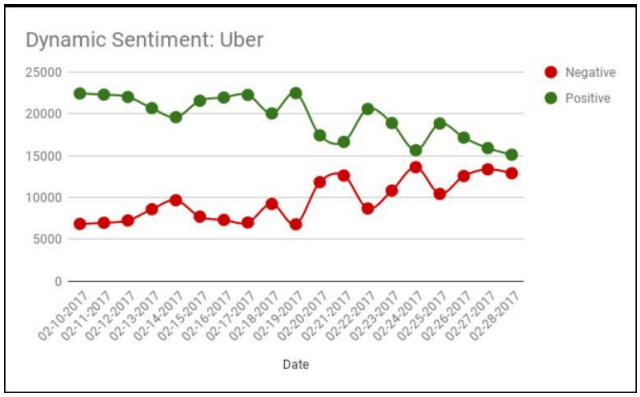

Fig. 3. Dynamic sentiment analysis result

We have 371,754 positive tweet sentiments and 184,443 negative tweet sentiments from the total of 1,048,576 tweets throughout the data period. From the visualization, we see the sentiment fluctuates each day that can be analyzed independently. For example, in 24th February 2017, the chart displays both positive and negative results relatively have significant changes than other days. A business organization needs to know the reason behind this phenomenon, thus, they can take corrective action or resolve problems earlier.

In topic modeling, we pick two highest topics to be discussed, as it is the most frequent/salient word. Light blue color represents the overall term frequency, while the red one represents estimated term frequency within the selected topic.

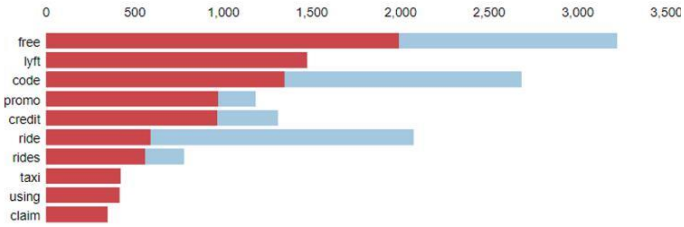

Fig. 4. February 24th Positive sentiment analysis result

Uber currently focus on increasing the number of users around the world. Uber formulates a strategy to attract people on using Uber to commute, one of the strategies is to give a discount to the consumer in the form of promo-code. Interestingly, the strategy is relatively successful, it is proved by the topic modeling output shows that on 24th February 2017, positively talked topics are around free-code or promo code and taxi rides.

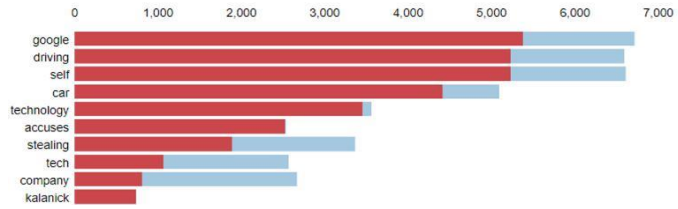

Fig. 5. February 24th Negative sentiment analysis result

On the other hand, there is an increasing number of negative sentiment, the most dominant negative topics on 24th February 2017 and internal issues that appear on the news, Uber fires executive accused of stealing Google's self-driving car secrets. For a thorough overview of the dataset, we conduct topic modeling for the whole dataset to capture the whole topic in the dataset.

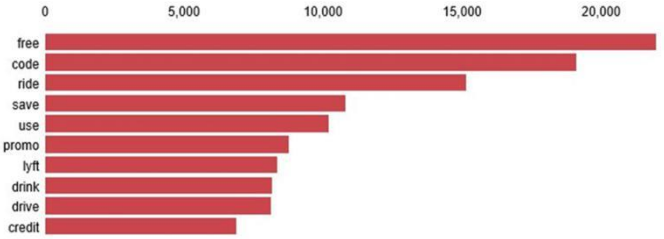

Fig. 6. Largest positive topic modelling sentiment analysis result

The first dominant positive topics in the dataset are about the free rides, promo. We interpret this topic as an information in which most of the users tweeted about the promotion of Uber, the public is reacting positively toward the promotion.

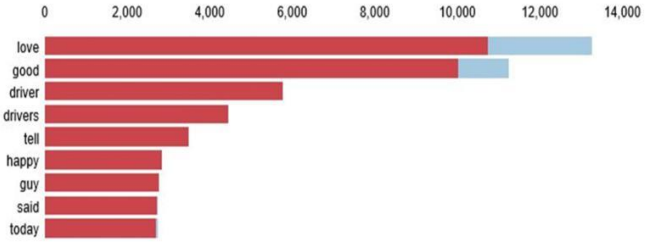

Fig. 7. 2nd Largest positive topic modelling sentiment analysis result

Second dominant positive topic in the dataset is about the driver compliment, the users express their feelings on the usage of the services. Most of the customer is satisfied with the service, this can be interpreted from the word love, good, drivers.

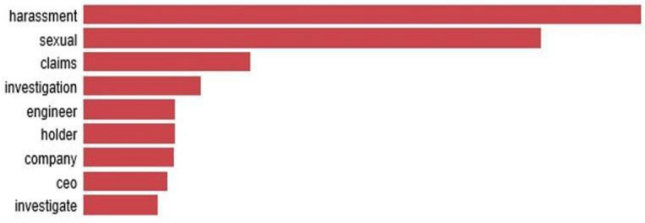

Fig. 8. Largest negative topic modelling sentiment analysis result

There is a negative sentiment that influences the customer opinion. Based on the visualization above, the first dominant negative topic is about sexual-harassment is probably occurred by the news that is stating that there is a sexual-harassment towards Uber's employee. These issues are revealed by former Uber engineer. then it is followed by the claim of the Uber CEO who conduct the investigation.

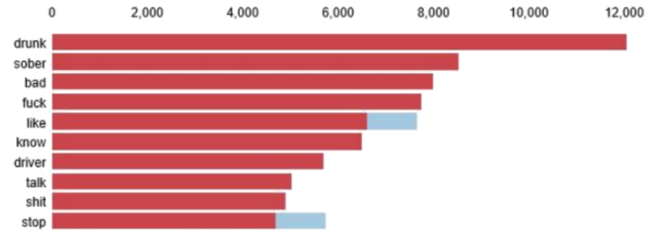

Fig. 9. 2nd Largest negative topic modelling sentiment analysis result

The second negative sentiment on the dataset is related to the driver. Even though some people satisfied with the service, there are several complaints on its service, some of the drivers have been reported driving while he/she is drunk. The customer is unsatisfied with the fact that the driver tends to talk with the customer they expect to remain silent or calm, this is referred to driver talk shit stop.

V. CONCLUSION

We have successfully implemented the method to properly analyze, summarize and extract massive scale tweets. In our large-scale case study regarding Uber. Naïve Bayes method is suitable for sentiment analysis as well as the LDA method that has proven capability in extracting in-depth insight about the topics discussed in the large-scale dataset. In our opinion, the method that we use in this research are better in the term of real-time processing capability compared to the traditional way.

The result complete within less than 5-10 minutes (only processing time – data already preprocessed). Compared to conventional method of insight extraction, dynamic sentiment analysis (sentiment based on a daily basis) is capable in the extraction of opinion in massive data, this creates more effective and efficient process. As the globalization era evolves, speed is one most essential factor in business, a business organization has to quickly react to their customer's opinion.

As for the shortcomings of this research, it may need more research. Therefore, we can find a method to calculate the accuracy of the Topic Modeling and the method to clean data effectively. Our suggestion for future research, the stop word dictionary must be enriched, reduce the ambiguity of the text in order to increase the accuracy, using t-SNE to model topic with even more detailed results, which can have more customizable parameter (e.g. Multidimensional Scaling Method, Topical Distance Calculation, Number of Clusters, Number of Terms). Furthermore, further researcher need to have in-depth discussion on how this method can be implemented in business analytics.

REFERENCES

[1] Liu,Bing., NLP Handbook, Illinois, Chicago: University of Illinois at Chicago, 2009.

[2] Alamsyah, Andry et al. "Top Brand Alternative Measurement Based on Consumer Network Activity." Advanced Science Letters 23.4 3813–3816.

[3] Alamsyah, Andry., Zuhri, Faishal Nuruz, "Measuring Public Sentiment Towards Services Level in Online Forum using Naïve Bayes Classifier and Word Cloud", CRS-ForMIND International Conference and Workshop, 2017.

[4] Ting, S.L., W.H. Ip., Albert H.C., Tsang., "Is Naive Bayes a Good Classifier for Document Classification?" International Journal of Software Engineering and Its Applications, vol.5, No.3, pp. 37-46, July 2011.

[5] Alamsyah, Andry. "Measuring e-commerce Service Quality from Online Customer Review using Sentiment Analysis (Case Study: Tokopedia)", The international Conference on Data and Information Science, 2017.

[6] Daniela X, Christopher J, Roger G, "Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages", International Journal of Computer Science Issues, Vol.4 No.1, 2009.

[7] Paquali, Arian Rodrigo., "Automatic Coherence Evaluation Applied to Topic Models" Portugal, Universidade Do Porto, 2016.

[8] Kaur, Gurpret., Kaur, Manpreet., "Case Study of Color Model of Image. Processing", International Journal of Computer Engineering and Technology, 6(12), 2015, pp. 65-71.

[9] Liu, Bing., Zhang, Lei., "A Survey of Opinion Mining and Sentiment Analysis", Illinois, Chicago: University of Illinois at Chicago, 2010

[10] Liu, Bing., Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[11] Blei, Ng, Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research 3, pp. 993-1022, 2003.

[12] Sievert, Carson., Kenneth E.Shirley, "LDAvis: A Method for Visualizing and Interpreting Topics," Stanford University, Natural Language Processing Group.