A Big Data Architecture to Support Bank Digital Campaign

[1]Irfan Wahyudin, [2]Salmah
[1] Department of Computer Science, Universitas Pakuan Bogor, West Java, Indonesia
[2] Department of Economic, Universitas Pakuan Bogor, West Java, Indonesia
[1]irfan.wahyudin@unpak.ac.id, [2]salmahazzubaidi@gmail.com

Abstract

Bank marketers still have difficulties to find the best implementation for credit card promotion using above the line, particularly based on customers preferences in point of interest (POI) locations such as mall and shopping center. On the other hand, customers on those POIs are keen to have recommendation on what is being offered by the bank. On this paper we propose a design architecture and implementation of big data platform to support bank's credit card's program campaign that generating data and extracting topics from Twitter. We built a data pipeline that consist of a Twitter streamer, a text preprocessor, a topic extractor using Latent Dirichlet Allocation, and a dashboard that visualize the recommendation. As a result, we successfully generate topics that related to specific location in Jakarta during some time windows, that can be used as a recommendation for bank marketers to create promotion program for their customers. We also present the analysis of computing power usages that indicates the strategy is well implemented on the big data platform.

Keywords*: big data, Hadoop, topic modeling, LDA, bank marketing, product campaign*

Introduction

As of February 2018, based on Indonesian Credit Card Association (AKKI), the number of credit card holders in Indonesia are constantly grow since 2009. There were 17.079.966 card holders in Indonesia untill November 2017, with 290,701,721 transactions have been recorded from January to November 2017. These numbers shown that the credit card are one of the most potential product for a bank. On the other hand, the goverment also encourages Indonesian citizen to start using non-cash transaction in daily activities.

To keep this positive trend, a bank must have promotion strategy. Various kind of conveniences and services must be offered to attract not only new customers to apply for credit card facility, but also their existing customers to increase their transactions by using credit card. However, the offering will be useless if there are no awareness from the customers. Thus, the promotion program must be conducted by the bank and ensure that the promotion has a good reception from the customers.

There are two types of promotion strategy, namely below the line and above the line. The first mentioned is the conventional method that commonly used, typically done by spreading brochures, sending product catalogues, advertising promotion on billboards etc. While the second one is a kind of promotion using digital platform such as email and social media. This method is also oftenly called as digital campaign by the bank marketers.

Based on our survey of 100 bank customers, we found that about 27% of them are occasionally satisfied with the bank's campaign strategy, and 45% of them are still doubtful with the bank's campaign strategy it means the campaign that have been offered is beyond their expectation and their needs . There are only 37% of them feel that the bank's campaign strategy is already good in terms of that the offering meets the customers needs and help them to found the items they need to buy especially when they are in POI places. On the other side, we also conducted an interview with a bank marketers, and revealed that mostly they met difficulties in finding how to offer a promotion that meets with the customers needs such as their hobbies, and their weekly/daily spendings and activities. Even more, they also have to define where are the most attractive merchants to promote.

Digital campaign is an effort from an organization to promote their products and increase the customers loyalty by using digital technology such as digital advertisement on social media platform such as Twitter, Facebook, and Instagram (Jayaram et al, 2015). Some of them also used the social media platfom to quantify the sentiment from their customers against the products they have (Kaushik et al, 2013). Research on content recommendation based on location has been conducted by Khusnul et al, that successfully develop a model based on information posted from social media and processed into a recommendation using vector space model (Khotimah et al, 2014).

Big data has been leveraged by many organizations, including banks to retrieve useful insights from many data sources. Some of the use cases are to predict regional economic by evaluating the individual credit card usages (Sobolevsky et al, 2015) and to analyze credit card usage by implementing Hadoop platform (Rodrigues et al, 2017). Both previously mentioned researches shown that although big data already proven to be implemented in the banking area, none of them utilized external data source such as from social media platform. Other use case is to quantify risks from SME business done by (Wahyudin et al, 2016) that successfully quantify business risks from narative text written by the credit analyst. As we can see, although

there were some use cases that discussed about how the big data implementation in banking sector, it is still difficult to find use case that specifically discuss about the implementation of banking marketing on top of big data platform.

From the above mentioned problems, we propose an approach that could help the marketers to ensure that the promotion they blast to the customers are indeed based on the customers preferences and the merchants whose the bank have chosen are also attract the customers attention to have the transactions onto. We started the approach by collecting data from tweets were posted from several malls in Great Jakarta region. We use Twitter since this platform is the only platform that allowed to crawl the posted information not only based on keyword, but also by geolocation as well. The process followed by cleansing and transforming the words collected into a bag of words, and eventually we described the most discussed by using Latent Dirichlet Allocation (LDA) technique. Recently, alot of research in topic modeling used LDA to find the hidden meanings in the document collection (Jelodar et al, 2017). Introduced by Blei, Ng, and Jordan in 2003, LDA uses the probabilistic approach to determine the appropriate topic in a document, by compute the probability the words contained in a document compared to the others. There are various style of LDA implementation, and we choose the implementation in Apache Spark MLLib library, since we are focusing on how to implement LDA so that it can executed in parallel manner (Dai, 2017). To stores the data, we utilized a Hadoop based platform as a proof of concept that the proposed approach is feasible to be implemented in a big data platform.

Research Method

The process started with collecting the posted information from Twitter, we use a daily schedule that works during weekdays and weekends as well as we want to compare the customer behaviour during those time windows. The method of this research takes these following steps as seen in the Fig. 1 below.

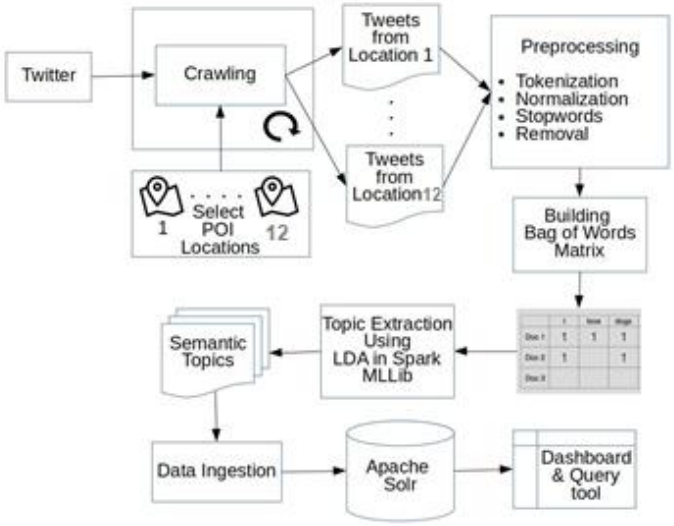

Fig. 1 Start to end diagram process to fetch insights from Twitter on top of proposed Big data platform

In this research we design a workflow model that can simultaneously retrieve tweets, cleanse, preprocess, and refine the topic model in realtime manner as described in detail as follows:

a) The process begin with the selection of POI location in greater Jakarta. As a trial we focusing on malls where most of the bank merchants located. With the help of Google Maps, we mark the location of mall as reference for the Twitter API.
b) The second step is define the crawling job on top of Hadoop platform. For this task we use custom Python script that deploy on Oozie workflow scheduler. Later, this job will executed to stream the tweets posted by the people on the selected locations. All of the results from this streaming process stored into a flat text file for each location.
c) Next, the preprocessing task is performed by tokenizing the words. We use whitespace to define the word feature from each tweets. Moreover, we also normalize the words by convert all the letters into lowercase.
d) Indonesian Twitter users oftenly used slangs, non-formal terms and abbriviations. That caused us to transform those terms into formal terms. Hence, we create a library as a (Purwarianti & Lunando, 2013). To preserve the important terms only, we also need to remove the stopwords. For this case, we

also add a criteria: terms that have number of appearance more than the threshold that we have defined, will be categorized as stopwords.

e) Once the terms are already preprocessed, we build a bag of words that used for the computation later on. The bag of words itself is a two dimensional matrix with size of m x n, where m is the number of documents, and n is the number of terms collected.

f) We continue the process with the topic extraction using Latent Dirichlet Allocation that from this phase, some topics will be yielded. The basic idea of LDA is to set each documents with a set of topics, whereas each topic has a set of probabilistic mixtures, composes of set of terms distribution (Blei et al, 2003). There are several parameters that has to be set in this algorithm as seen in Fig. 1.

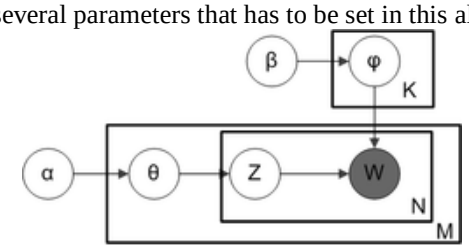

Fig. 1 Latent Dirichlet Allocation architecture

Given the parameters α, a Dirichlet prior per document topic distributions and β, Dirichlet prior per topic distribution, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is denoted as Eq. 1 below (Blei et al, 2003):

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \quad (1)$$

g) Then, the topics will be ingested by the data pipeline and the data will be forwarded into the end point repository.

h) To ease the user to monitor the results, we provide a dashboard to visualize the end result for the user.

Along with the simultaneous process, we also design the workflow to executes each task above, per day and per location. We break down the process with the consideration that we want to minimize server load and to make it easier to track each task. All of those tasks are executed in a cluster that consist three machine with an architecture as follow:

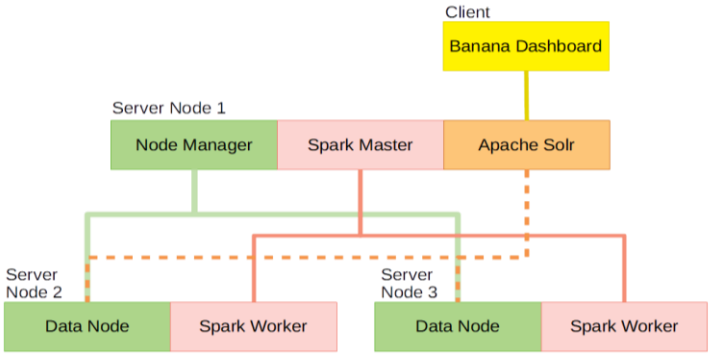

Fig. 2 The interaction between three main big data programs that deployed in the proposed platform

In this research, we use three machines, each have roles as depicted in Fig. 2. The data are distributed across two data node so we can have the availability features and parallel processing that performed by two spark workers. Apache Solr adds the term indexing and search engine functionality to the system. We also used a dashboard software that we discussed later on the next section, as an interface for the bank marketers to query the analysis results.

Result and Discussion

We have collected 234,885 tweets from April 12 2018 to May 12 2018 from twelve big malls in **Great Jakarta Region**, namely Pacific Place Plaza, Mall Taman Anggrek, Plaza Senayan, Senayan City, Kota Kasablanka, Central Park, Bintaro XChange, Pondok Indah Mall, Tangerang City Mall, Plaza Semanggi, and Botani Square Mall. We run our crawler program in a virtual machine with 1 Gigabyte RAM and 1 Core processor, and the job splitted into twelve different threads respectively, to get the all tweets from twelve different location in parallel manner.

To cleanse the unnecessary text, we use a stopword list and some criteria to filter and extract the important terms. There are 357 stopwords taken from a research conducted by Tala in 2013, that we used as a

reference to exclude the unnecessary terms. We did not remove the original text, otherwise, we keep those text so that the bank marketing able to posts a query against the original tweets without having to know what is the stopwords and what is not. After removing some unnecessary terms, we have 136,373 tweets. Most of the tweets removed are tweets that only contain emoticons and slang words such as ‘okay’, ‘yoiii’, ‘wkwkwk’ and ‘hahaha’.

The task followed with topic extraction using LDA that performed in parallel manner by Spark workers. The strategy we used for LDA is still the same with the preprocessing task, which the job executed per day and per location. This strategy is intended to help the bank marketing to perform pattern analysis, to know what is currently being popular among the mall visitors, in particular day. Another benefit of grouping the process by the ate and location is we do not have to use the entire computing power to process whole data. Otherwise, we only uses partial computing power by selecting partial information of tweets data by date andlocation parameter.

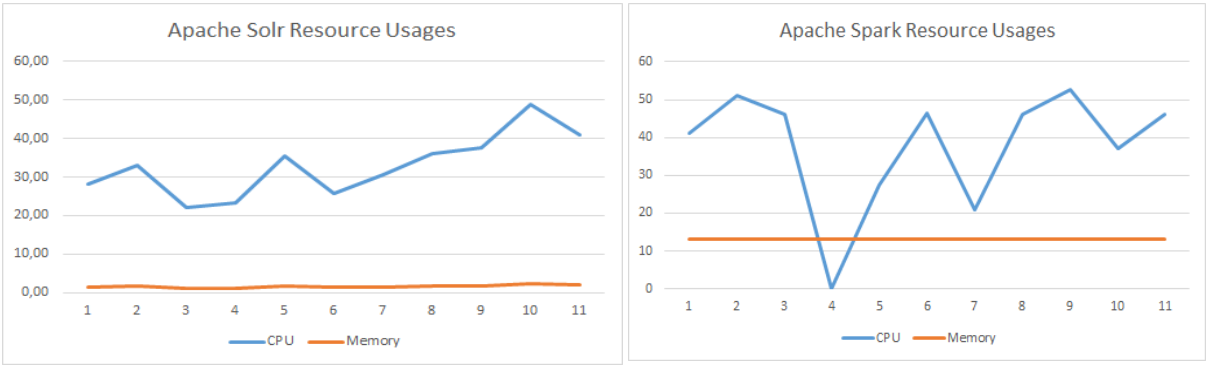

Fig. 3 Graphical resource (CPU and Memory) usages by Apache Solr and Apache Spark

To evaluate the performance, we capture one of the process cycle (preprocess-topic extraction-term indexing). As we can see in Fig. 3 above, for eleven iteration (100 data per iteration), both Apache Spark that performs topic extraction using LDA and Apache Solr that performs term indexing, were only used up to 60% of available resources. This performance were never can be achieved when all of the data are being processed at once, due to the lack of computing power (100% failed).

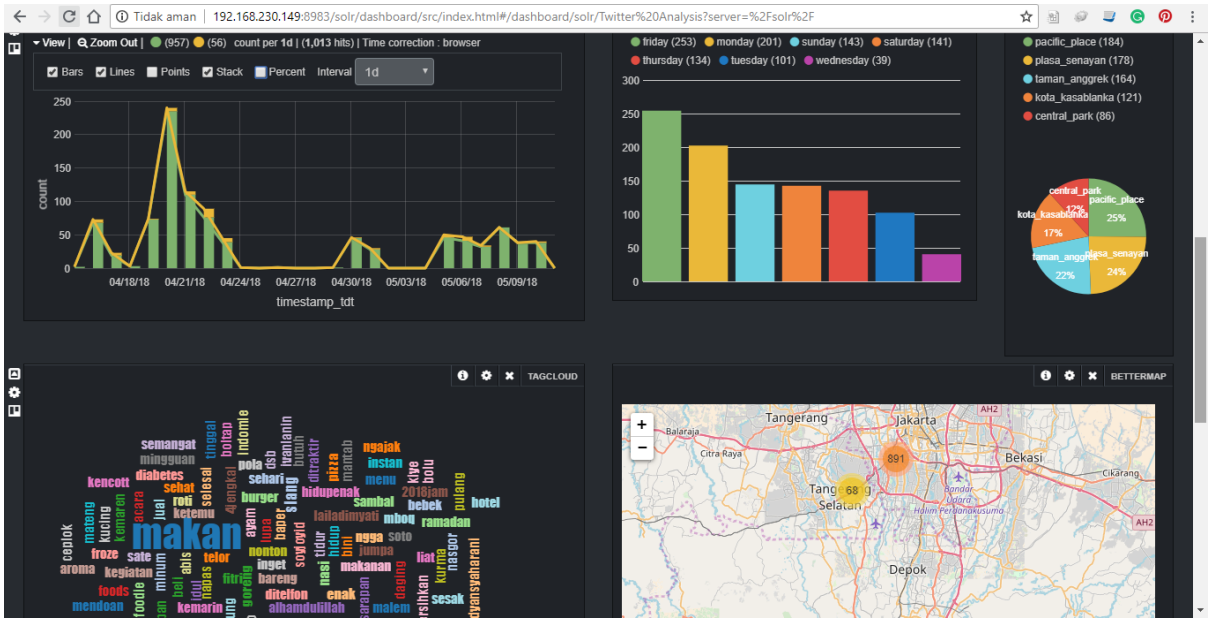

Fig. 4 The graphical presentation to analyze the models result, there are various type of chart and map presentation

To examine the result we use Banana dashboard to visualize the topic distribution across the dates, and location as seen in Fig. 4. The dashboard provide a dynamic layout, and interactive user interface that enables users to design the desired layout and choose the chart type that suitable to represent the data. First examination is by select the tweets by the day name. Here we use the regular query feature in Apache Solr to perform the search. The top 3 result as seen in Table 1 as follows. It is obvious that during weekend days (Friday to Sunday), people in large city are mostly spend their time on recreational area such as malls.

Table 1 Days with most number of Tweets

No	Date	Number of Tweets	Peak Hour
1	20-April-2018 (Friday)	20,212	06:00 PM (278)
2	21-April-2018 (Saturday)	19,811	10:00 AM (382)
3	22-April-2018 (Sunday)	14,075	08:00 PM (308)

Other examination that we perform is topic distribution and it's correlation with the day name, the location, and the terms that mostly occur. We picked up five different activities and topics in Bahasa Indonesia, that most people are often do and discuss in malls: shopping (belanja), watch movie (nonton film), eat pizza (makan pizza), buy new phone (beli handphone baru), looking for discount (cari diskon). The result as can be seen below in Table 2. The query execution time is relatively same, regardless the number of returned rows. From the given result, we expect that the marketing is able to figure on what day and where the location most people are attracted to some activity or a product. For instance, going to theaters to watch movie are mostly do by people in Saturday to Monday, and the most interested place to watch movie in greater Jakarta is Taman Anggrek, Plasa Senayan, and Pacific Place.

Table 2 Query examination result

No	Query Terms	Terms Related	Top 3 POI	Query Time	Number of Records	Top 3 Day Name
1	belanja	belanja, menangkan, voucher, harga, tokopedia, bayar	Taman Anggrek, Pacific Place, Bintaro XChange	915ms	413	Tuesday, Monday, Saturday
2	nonton film	nonton, bioskop, film, imax, infinitywar, cinemaxxi, avengers	Taman Anggrek, Plasa Senayan, Pacific Place	409ms	75	Saturday, Sunday, Monday
3	makan pizza	makan, siang, burger,	Pacific Place, Plasa Senayan, Taman Anggrek	58 ms	57	Friday, Monday, Sunday
4	beli handphone	beli, aplikasi, promo	Pacific Place, Taman Anggrek, Kota Kasablanka	73ms	19	Saturday, Monday, Thursday
5	cari diskon	promo, karpet, islamicbookfair, karpet,karpetshaggy	Taman Anggrek, Pacific Place, Plasa Senayan	439ms	44	Monday, Saturday, Thursday

Above result also show that in average the response time are under one second. This result considered as good or at least tolerable if we refer to the standard of web application response time (Nah, 2004).

Conclusion

The design architecture, and development of big data platform to help the bank marketing were successfully done. The result shows that by distributing the process and split the process into several batch on top Apache Hadoop and Apache Spark technology, we effeciently reduce the computing power usage, by only use the available resources not more than sixty percent. From the collected tweets from twelve POIs in greater Jakarta area, we also managed to extract topics and let the user performs query and analysis by themself by using the dashboard. The dashboard that we provide also give flexibility for the user to interacts with the given result. Another achievement is that the query response time are also considered as tolerable as the average response time is under one second.

Acknowledgement

This research was supported by Ministry of Research, Technology and Higher Education Republic of Indonesia.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*.

Dai, C., Wang, Y., & Wang, Q. (2017). Topic model and similarity calculation of text on spark. *2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*.

Fui, F., & Nah, H. (2004). A study on tolerable waiting time: how long are Web users willing to wait? *Special Interest Group on Human Computer Interaction*.

Jayaram, D., Manrai, A. K., & Manrai, L. A. (2015). Effective Use of Marketing Technology in Eastern Europe: Web Analytics, Social Media, Customer Analytics, Digital Campaigns and Mobile Applications. *Journal of Economics, Finance & Administrative Science*, 118.

Jelodar, A., Wang, Y., Yuan, C., & Feng, X. (2017). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey.

Kaushik, A., Kaushik, A., & Naithani, S. (2013). A Study on Sentiment Analysis: Methods and Tools. *International Journal of Science and Research*.

Khotimah, H., Djatna, T., & Nurhadryani, Y. (2014). Tourism Recommendation Based on Vector Space Model Using Composite Social Media Extraction. *Advanced Computer Science and Information Systems*.

Purwarianti, A., & Lunando, E. (2013). Indonesian social media sentiment analysis with sarcasm detection. *Advanced Computer Science and Information Systems (ICACSIS)*.

Rodrigues, R. A., Filho, a. A., Gonçalves, G. S., & Mialaret, L. F. (2017). Integrating NoSQL, Relational Database, and the Hadoop Ecosystem in an Interdisciplinary Project involving Big Data and Credit Card Transactions. *Advances in Intelligent Systems and Computing* .

Sobolevsky, S., Massaro, E., Bojic, I., Arias, J. M., & Ratti, C. (2015). Predicting Regional Economic Indices using Big Data of Individual Bank Card Transactions. *2017 IEEE International Conference on Big Data (Big Data)*.

Tala, F. Z. (2003). *A Study of Stemming Effects on Information.* Amsterdam: Master of Logic Project Institute for Logic, Language and Computation Universiteit van Amsterdam.

Wahyudin, I., Djatna, T., & Kusuma, W. A. (2016). Cluster Analysis for SME Risk Analysis Documents Based on Pillar K-Means. *Telecommunication, Computing, Electronics and Control*.

.