Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 135 (2018) 5–14

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Computer Science and Computational Intelligence 2018

Business Intelligence Model to Analyze Social Media Information

Parama Fadli Kurnia, Suharjito *

Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
**corresponding Author, e-mail: parama.kurnia@binus.ac.id; suharjito@binus.edu*

Abstract

Social media is a platform to share information that is very liked by everyone nowadays because some of the facilities that make it easier for us to communicate with each other, share documents, chat and even create a community. In addition, we can also analyze the content of social media by using several methods in data mining, so that we can get new the information to support decision making that can bring benefits to individuals and companies. The purpose of this research, to create a business intelligence dashboard to observe the performance of each Topic or channel of news posted to social media accounts such as Facebook and Twitter. Topical performance in social media is the number of Topics in articles posted to social media getting like, share, comment etc. To be able to know the Topic of a news post in social media, used some text classification techniques such as Naive Bayes, SVM and Decision Tree. The comparative results of the algorithms are taken which has the best accuracy of SVM for subsequent implementation in the data warehouse. Meanwhile, the business intelligence dashboard data source will be sourced from the data warehouses that have been made before.

Keywords: Social Media, Naive Bayes, SVM, Decision Tree, Business Intelligence

1. Introduction

Nowadays, social media is a natural thing that is owned by all circles either children, teenagers to adults. Not only Facebook and Twitter, social media now also enlivened with the presence of LinkedIn, Instagram, Path and

* Corresponding author. Tel.: 08128400536; fax: +0-000-000-0000 .
E-mail address: suharjito@binus.edu

10.1016/j.procs.2018.08.144

Pinterest. Their role that makes it easy for its users to easily participate, share, and create content including blogs, social networks, wikis, forums and virtual worlds, making social media more and more attractive to all users. This is reinforced by statistics obtained smartinsights.com, where the number of active Facebook users reached 1.59 trillion, Twitter reached 320 million, Instagram reached 400 million, and so on [1]. The results of these statistics show that Facebook ranks first as a social media with the most active users [2].

Given the large enough active users described above have a significant impact on the usefulness of the social media itself. Here the impact is the use of social media that previously only serves as a place to participate, share, and create content, can now also be used as a medium to promote, campaign, and communicate with consumers easily with the social media. We also realize that social media marketing requires monitoring to prevent misleading strategies (promotion, campaign and communication with consumers), so it should be precisely what areas should be monitored in order to do marketing in social media, either increase brand awareness or create traffic to the website [3]. Strategic areas that need to be monitored in social media are brand terms, brand-adjacent terms, customer needs, customer sentiment, and competitors [4]. This monitoring can be done directly on social media like Facebook as the most popular social media platform for marketing activities, followed by LinkedIn, Twitter, YouTube and Instagram[5].

In this research will be designed and implemented data warehouse model and software for business intelligence system. Data sources will be extracted from account data on Facebook and Twitter. Both will be classified into several topics that exist in the coil of entertainment, economy, health, culinary, lifestyle, automotive, politics, soccer, technology and travel. The text classification method for social media will be tested on 3 algorithms namely Naive Bayes, Decision Tree and SVM [6]. The best results of the comparison between the three algorithms will be selected to classify text which will be very useful information for the implementation of data warehouse [7]. Meanwhile, the design method to be used for data warehouses is the method of Kimball [8]. In this method there are four stages that must be passed in the design of data warehouse, which is select the business process, declare the grain, identify the dimensions, and identify the facts. In the ETL process (Extraction, Transformation, Loading) is a process that must be passed in the formation of data warehouse [9]. And the last one in business intelligence design using Carlo Vercelli's method, in this method there are four main stages, namely analysis, design, planning, and implementation and control [10].

2. Related Work

Smita and Seema conducted a survey on the use of data mining on business intelligence for social networking. Background of their research that is, the data mining makes it easier to do data analysis on social networking, even can take hidden knowledge that exist in data. This can happen if we add data mining processes to business intelligence. With the existence of data mining used to perform the process with large amounts of data, can dig new information from a data such as class, pattern to predict something based on existing data [11]. In data mining there are several techniques that are divided into 2 namely the predictive task and descriptive task. In the predictive task there are techniques classification, regression and deviation detection. Meanwhile, the descriptive task is subdivided into clustering, association, sequenced pattern and summarization techniques. Currently, some data mining techniques have been used in traditional customer data, for example: regression analysis, Naïve Bayes (NB), Support Vector Machine (SVM), and Neural Network (NN) [12]. Regression analysis is the most popular technique for predicting customer satisfaction [13]. This data mining technique can be used to solve some problems such as identifying sentiments in user comments on social media, detecting habits and performing demographics on users in social media. Performing data mining on social media is very potential and useful for extracting more information and gaining deeper insights about customers so it is necessary to support activities such as customer interaction and analysis, information systems development, marketing, and business intelligence analysis [14]. Also development and analysis of individual and group strategies for problem solving in the community. To be able to support the solution of some of the problems described previously, data mining has several parts in order to handle electronic business data. These sections begin with data collection, which then performs data analysis process including preprocessing and then modeling data with data mining algorithm to get information from the data extraction result that is useful for decision making, predict user behavior, determine business strategy [15].

Sangameshwari and Uma conducted research on the importance of the role of data mining on business intelligence systems. Data mining is the extraction of concealed prescient data from vast databases; it is a compelling innovation with extraordinary potential to help associations concentrate on the most vital data in their data stockrooms. Data mining devices foresee future patterns and practices, helps associations to make proactive information driven choices. The mechanized, prospective examines offered by data mining move past the breaks down of past occasions gave by prospective apparatuses normal of choice help supportive networks. Data mining apparatuses can address the inquiries that customarily were excessively time intensive to purpose. They get ready databases for discovering concealed examples, discovering prescient data that specialists may miss in light of the fact that it lies outside their desires. With the existence of data mining on business intelligence, it is possible to be used to create applications such as Fraud Detection, Financial Analysis, Customer Behavior Analysis, Product Analysis and Sales Analysis [16].

Vitri Tundjungsari made business intelligence systems to support customer satisfaction in telecommunication industry. In order to perform customer satisfaction analysis, it is necessary to know the characteristics of users whose data can be derived from social media. Mining social media is very potential and useful for extracting more information and gaining deeper insight about customer [17]. For example, it can be used to identify the influential customer in a social networking site, detect implicit or hidden groups in a social networking site, perceive customers opinion related to their product or service's satisfaction for proactive planning, develop recommendation systems to maintain existing customers and gaining new ones, or build and strengthen trust among customers or between customers and other stakeholders. In short, mining social media is a promising multidisciplinary area, thus researchers of different backgrounds can make important contributions that matter for social media research and development [18]. To help the development of business intelligence systems know customer satisfaction, she uses data mining that can perform several process of discovering useful or actionable knowledge in large-scale data [19]. Detailed methods used to build business intelligence systems such as text mining, clustering that divides data into meaningful or useful groups [19] and visualization as an interface for displaying data to end-user. Some types of representation used in visualization are computationally expensive to generate, so that one challenge for the visualization community is to respond to the rapidly expanding data volumes seen in many application areas. Another obvious barrier to the ready use of visualization tools in science is the wide variety of data formats used in different scientific disciplines, which often necessitates the translation of data before it can be visualized [20].

3. Methodology

In general, the steps involved in this research are data collection, data analysis, data warehouse design, business intelligence design, text classification and evaluation methods. Data collection was carried out through interviews and literature studies. The interviews were used to determine the issues faced in relation to the need for business intelligence design from social media data to improve company performance. a literature study was conducted to derive methods that could be used in business intelligence design of social media with information from books, current journals and the Internet. Data analysis using the business intelligence analysis method [21] to identify the needs of business information that will bring maximum profit for the company. The analysis of business intelligence consists of: analysis of business drivers, analysis of business strategy, objectives and objectives, analysis of value disciplines, analysis of business core processes and analysis of company values. Data warehouse design using the Kimball [8] method in which there are 4 phases in data warehouse design, ie selecting the business process, declaring grain, identifying dimensions and identifying facts. The design of business intelligence with the Carlo Vercellis [10] method, where in this method there are four main phases, namely analysis, planning, planning, implementation and control. To obtain the best text classification method, the comparison of the performance of different text classification methods are Naive Bayes, SVM and Decision Tree [22].

The training process on each text classification method uses the data obtained from the joining results of several internal company data tables, e.g. Stories table, topics table and stories_Topics table. The join results of these three tables generate the lead_text and Topic_name data. The reason for choosing lead_text as a modeling component in the training process because on every post of social media accounts such as Facebook and Twitter always use lead_text as a description used on the internal platform on the company website. The classification test is performed using the k-cross validation method with value k = 10, this test aims to know the accuracy of the Naïve Bayes, SVM

And Decision Tree methods in the classification of the text using different data training and data testing. The evaluation of business intelligence system development using User Acceptance Test (UAT) method, which is done until the system is really in accordance with the expectations of User.

4. Proposed Model

In this research, it has been built Business Intelligence System which has the architecture that has shown in Fig. 1.

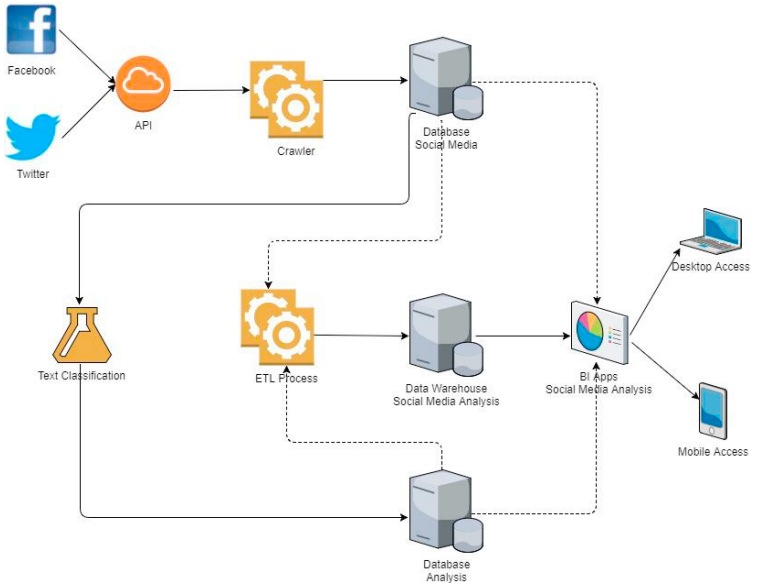

Fig. 1. Business Intelligence Architecture.

From the Fig. 1, it can be seen that to make the system required several important components such as data collection, content analysis, data warehouse process and business intelligence system are detailed in detail below.

4.1. Data Collection

At the data collection stage there is a process like first, data retrieval is done from Social Media Platform on Facebook and Twitter through Social Media API available on each platform. Second, data retrieval will be done periodically by crawlers that have been created using the Social Media Token API that is useful for authentication and authorization during data retrieval. And finally, the results will be stored in the database as raw data.

4.2. Content Analysis

At this stage social media content analysis uses text classification. Classification or categorization of text is the process of placing a document into a category or class according to the characteristics of the document. In text mining, the classification refers to the activity of analyzing or studying the collection of pre-classified text documents to derive a model or function that can be used to group other unknown class documents into one or more pre-defined classes [23].

At this stage of the analysis, there are several such processes first, retrieving data from each data in each social media platform. Next stage on the data loaded earlier, will be processed on text classification. In this research the text classification algorithm used is Naive Bayes, Decision Tree and SVM. But before going into the classification algorithm processing, the data will go through the preprocessing stage such as case folding, tokenizing, filtering, and stemming to eliminate data noises.

Each text classification algorithm here is evaluated to determine its performance. The evaluation method that will be used for the text classification algorithm is by applying the test scheme with 5 different data compositions and also for each of the test compositions will be done by applying the 10-fold cross-validation method. In addition, in the classification there are several ways of measuring the performance of classification as with the calculation of accuracy, precision, recall, and f-measure [24]. The Confusion Matrix is also used for performance evaluation of classification models based on the predictive accuracy of a model. Then, the results of the text processing will be stored in the database as the data analysis.

4.3. Data Warehouse Process

The Data warehousing design method used is Kimball method which there are 4 stages that must be passed in the design of data warehouse that is select the business process, declare the grain, identify the dimensions, and identify the facts [8]. In addition, there is also a process ETL (Extraction, Transformation, Loading) is a process that must be passed in the formation of data warehouse [9].

In the implementation, at this stage of the data warehouse process there are several processes such as the first, taking data derived from the analysis of content (database analysis) and the results of the crawler (social media database). Furthermore, the data already loaded earlier, then enter the ETL process, such as counting comments, post, sentiment, and others. ETL process data is stored in the data warehouse.

4.4. Business Intelligence System

On the client side is a business intelligence system that can be accessed by users. According to Vercellis [10], in the process of developing a Business Intelligence in a company there are 4 phases namely analysis, design, planning, implementation and control [10]. In its implementation, there are several important components of the business intelligence system such as APIs used as a bridge between applications and data (whether raw data or data warehouse). In addition, the application will request data to the API with Token API input as the authorization and authentication, so the existence of this API can improve data access security.

5. Results and Discussion

In this research, the evaluation will be used for the classification algorithm by applying the test scheme with 5 different data compositions. For each test composition was performed by applying a 10-fold cross validation method on the Naive Bayes, Decision Tree and SVM algorithms. The following test results are presented in the table 1.

Table 1. Naive Bayes Algorithm Test Results using 10-Fold Cross Validation

Iteration	**Test Case**				
	1st	2nd	3rd	4th	5th
K-1	73.076%	72.448%	73.426%	74.479%	72.033%
K-2	60.000%	65.979%	69.718%	73.936%	75.213%
K-3	71.428%	76.842%	80.141%	74.731%	70.689%
K-4	64.583%	70.967%	71.428%	76.086%	75.324%
K-5	65.217%	68.817%	71.942%	73.224%	74.458%
K-6	60.869%	77.173%	75.362%	73.770%	75.217%
K-7	59.090%	61.111%	78.676%	75.824%	71.491%
K-8	66.667%	79.545%	68.382%	75.274%	75.000%
K-9	61.904%	70.454%	77.037%	75.274%	74.561%
K-10	42.857%	70.454%	72.180%	74.274%	74.889%
Average	62.569%	71.379%	73.829%	74.787%	73.888%

Table 1 describes that the Naive Bayes algorithm test with cross validation with the 3rd test data on the 3rd iteration has the best accuracy with an accuracy value of 80.141%. However, if seen from the average of all cross validation test on each test data got the best accuracy result in 4th test data with accuracy value equal to 74.787%.

Table 2. Decision Tree Algorithm Test Results using 10-Fold Cross Validation

Iteration	Test Case				
	1st	2nd	3rd	4th	5th
K-1	52.941%	51.546%	54.545%	61.170%	57.203%
K-2	46.938%	46.391%	63.380%	59.893%	62.127%
K-3	51.020%	56.842%	54.285%	59.893%	56.896%
K-4	46.808%	48.936%	55.395%	59.893%	61.471%
K-5	51.111%	47.872%	56.521%	56.756%	62.173%
K-6	55.556%	58.064%	48.175%	55.434%	59.388%
K-7	45.454%	46.667%	57.664%	60.109%	55.021%
K-8	56.818%	53.409%	50.735%	56.830%	51.091%
K-9	36.363%	54.022%	66.911%	57.222%	54.385%
K-10	46.511%	62.068%	56.296%	49.444%	56.637%
Average	48.952%	52.582%	56.391%	57.664%	57.639%

Table 2 shows us that the Decision Tree algorithm test with cross validation with the 3rd test data on the 9th iteration has the best accuracy with an accuracy value of 66.911%. However, when viewed from the average of all cross validation tests on each test data obtained the best accuracy results in the 4th test data with an accuracy of 57.664%.

Table 3. SVM Algorithm Test Results using 10-Fold Cross Validation

Iteration	Test Case				
	1st	2nd	3rd	4th	5th
K-1	76.470%	79.381%	81.560%	73.684%	77.966%
K-2	78.431%	74.736%	80.000%	75.661%	76.694%
K-3	81.250%	87.234%	79.285%	72.727%	74.678%
K-4	76.595%	74.468%	74.100%	77.419%	73.706%
K-5	71.739%	82.795%	78.417%	74.054%	74.347%
K-6	84.782%	77.173%	71.223%	80.874%	77.826%
K-7	84.090%	77.778%	77.697%	74.585%	77.092%
K-8	76.744%	73.333%	76.470%	76.243%	75.770%
K-9	86.046%	81.111%	78.676%	78.453%	75.330%
K-10	73.809%	71.264%	68.656%	77.900%	74.889%
Average	78.996%	77.927%	76.608%	76.160%	75.830%

Table 3 describes that the SVM algorithm test with cross validation with the 3rd test data on the 3rd iteration has the best accuracy with 87.234% accuracy. However, if seen from the average of all cross validation test on each test data got the best accuracy result on 1st test data with accuracy value equal to 78.996%.

The previously tested classification algorithm is compared against the best accuracy results again to get which algorithm has the best accuracy. Figure 2 below is a summary of the best test results performed on each algorithm in the previous explanation.

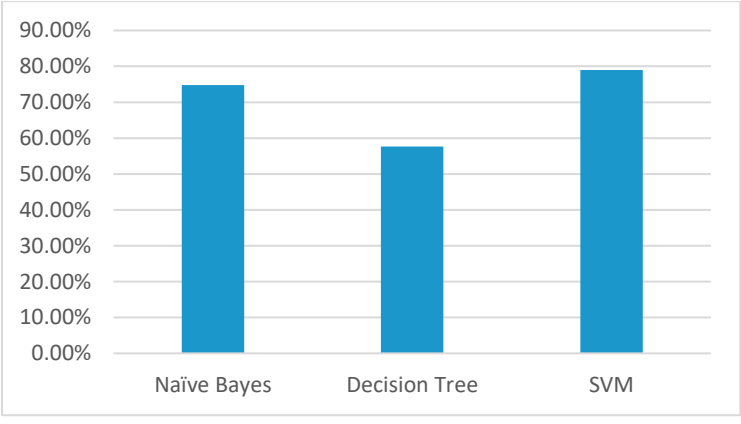

Fig. 2. The Comparison of Accuracy between Naive Bayes, Decision Tree, and SVM Algorithm.

Fig. 2 explains that the SVM algorithm has the best accuracy than the Naive Bayes and Decision Tree algorithm, so if it is ranked based on the SVM accuracy value at the first rank with an accuracy of 78.99%, the second rank is the Naive Bayes algorithm with an accuracy of 74.78% and finally the Decision Tree algorithm which has an accuracy of 57.66%. From this comparison it is concluded that SVM algorithm which will be used for text classification process in business intelligence system.

5.1. Data Warehouse

Data warehouse is made by using star schema model which consists of dimension table and fact table. The dimension tables and facts tables will be built in accordance with the data sources that have been stored on the Facebook and Twitter databases.

Star schema on Facebook data consists of fact_page_topic as fact tables and dim_page, dim_topics, dim_time as dimension tables. In this scheme can display the number of posts, comments, share, and like on articles that are grouped again based on topics related to the articles posted on Facebook. In addition, in this scheme also displays the number of types of article posts, such as video, link, status, offer and photo. For each record of this scheme is saved daily in the dim_time table. Fig. 3 shows the star schema of Facebook datawarehouse.

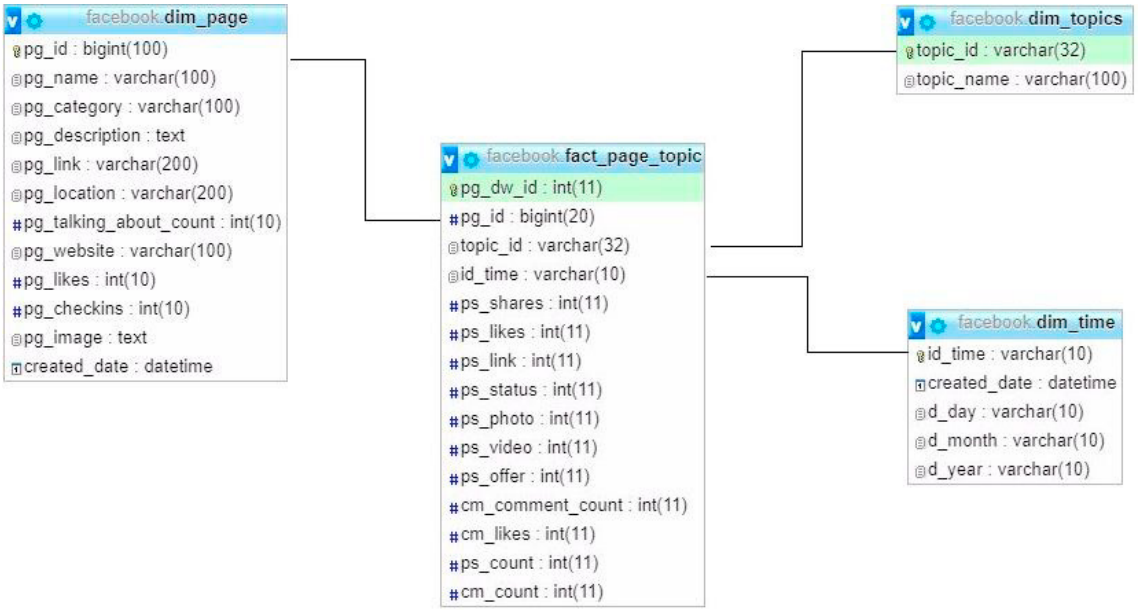

Fig. 3 Star Schema on Facebook Data Warehouse

Star schema on Twitter data consists of fact_account_topic as fact tables and dim_account, dim_topics, dim_time as dimension tables. In this scheme can display the number of tweets, retweets, and likes on articles that are grouped again based on topics related to articles posted on Twitter. In addition, in this scheme also displays the amount of retweet and like details either from tweet, retweet or mention an account on Twitter. For each record of this scheme is saved daily in the dim_time table. Figure 4 shows the star schema on Twitter data.

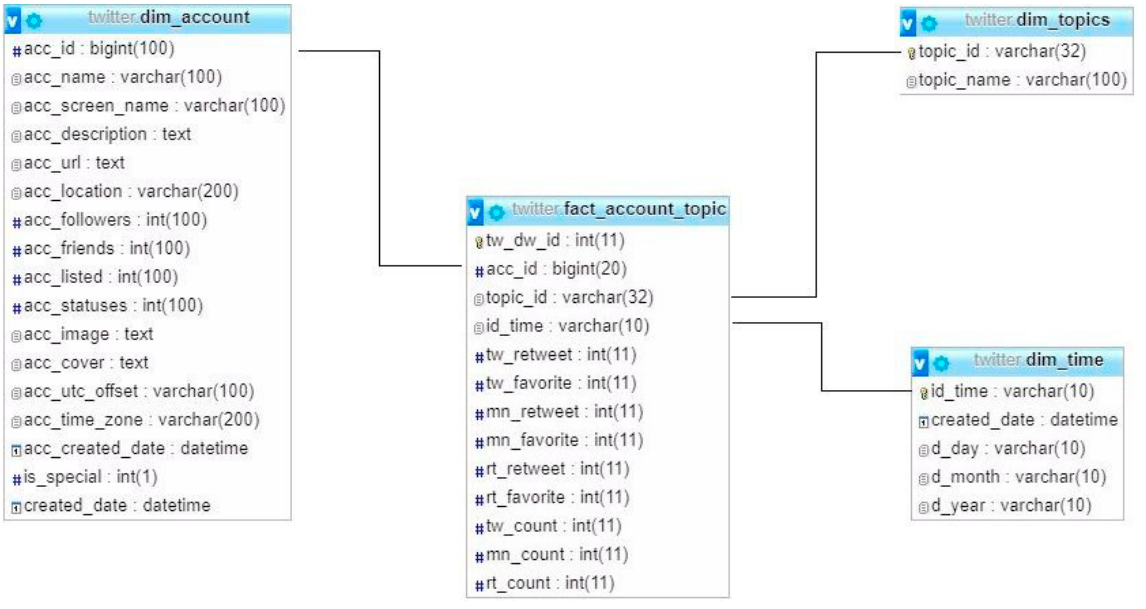

Figure 4 Star Schema on Twitter Data Warehouse

5.2. Business Intelligence

The Business Intelligence Dashboard is created using the CodeIgniter framework that contains the PHP programming language. The dashboard interface uses the template from AdminLTE v2 to speed up the development phase so there is no need to create the interface from scratch. The following are some of the interface of the business intelligence system dashboard displayed on the web portal.

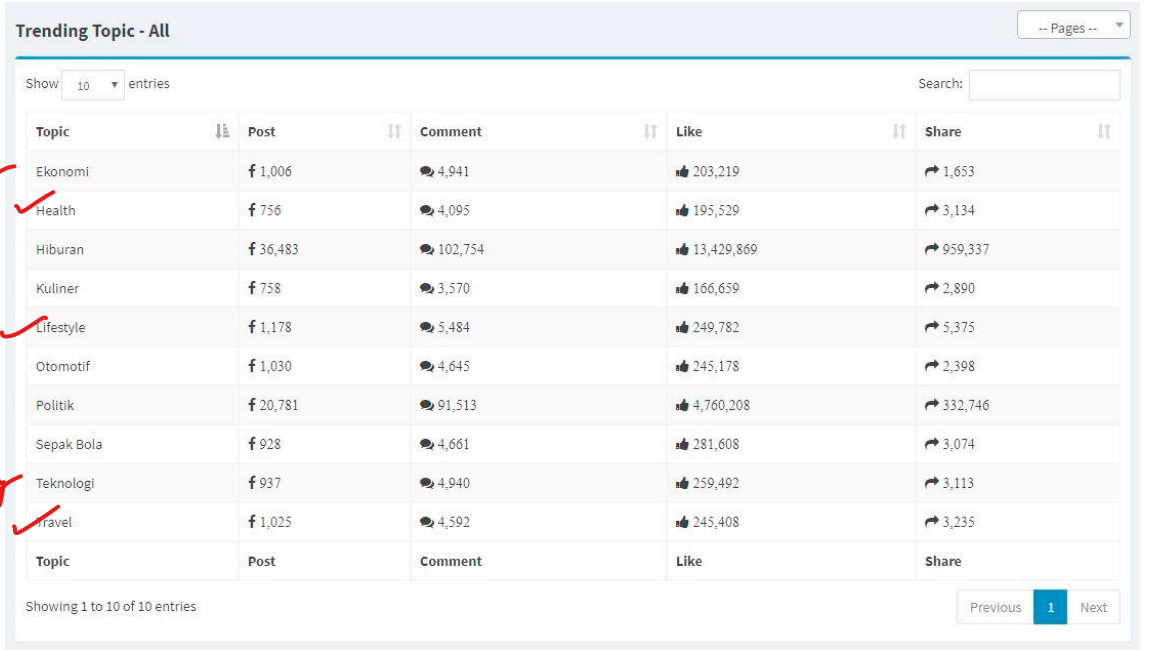

Figure 5 Trending Topic Page on Business Intelligence System

In the business intelligence system that has been created, can be seen several pages that can provide data insight, such as Trending Topic Page to know popular topics in social media; Most Active Users page, to show the most frequently commented accounts in social media that we can make them as an advertiser; Post Composition page, to display the number of posting types performed by an account on social media. Types of posts on social media are status, photo, link, video and offer on Facebook. Meanwhile on twitter, the post type is tweet, retweet and mention; Trending Hashtag Page, to display popular hashtags on Twitter; Popular Post Page, to show what kind of posts are popular on social media

6. Conclusion and Future Work

After evaluating the performance of the classification algorithm, the best performing algorithm is SVM with 78.99% accuracy, then the second rank is Naive Bayes has an accuracy of 74.67% and the last one is Decision Tree has an accuracy of 57.66%. Data Warehouse System created as a data source for Business Intelligence System can run the data calculations and summarization automatically, so no longer need to calculate manually. In addition, the business intelligence system created can be accessed by user anytime and anywhere.

The data warehouse system was created as a data source application for Business Intelligence that can automatically perform calculations and data summaries, so it is no longer necessary to calculate manually the number of comments, the number of Likes and the number of shares per article to be matched with the topics of article. The Business Intelligence application is very useful for monitoring the performance of news posted on social media both from the internal company account and from competitors in real time, so it is no longer necessary to visit one by one account in social media.

This study will be continued by developing Business Intelligence applications that can be implemented on other social media platforms such as Instagram, Linkedin, Path and others. In addition, the implementation of data warehouses and OLTP using Big Data technology allows faster processing of distributed data so that the data can be viewed on the dashboard in real time.

References

1. Chaffey D. Global social media research summary 2016: Smart Insight; 2016.
2. Salloum SA, Al-Emran , Shaalan. Mining Social Media Text: Extracting Knowledge from Facebook. International Journal of Computing and Digital Systems. 2017 Maret; 6(2).
3. Luke J, Suharjito. Data Mining of Automatically Promotion Tweet for Products and Services Using Naïve Bayes Algorithm to Increase Twitter Engagement Followers at PT. Bobobobo. Procedia Computer Science. 2015; 59(1).
4. Hasan , Moin , Karim , Shamshirband. Machine Learning-Based Sentiment Analysis for Twitter Accounts. Mathematical Computational Applications. 2018; 23(11).
5. Stelzner M. 2016 Social Media Marketing Industry Report: How Marketers Are Using Social Media to Grow Their Businesses: Social Media Examiner; 2016.
6. Tripathy , Agrawal , Rath. Classification of sentiment reviews using n-gram machine learning approach. Expert Systems With Applications. 2016; 57(1): p. 117–126.
7. Suharjito , Diana , Herianto. Implementation of classification technique in web usage mining of banking company. In International Seminar on Intelligent Technology and Its Application, ISITIA 2016; 2016; Lombok, Indonesia.
8. Kimball R, Ross M. The data warehouse toolkit: The definitive guide. 3rd ed Canada: John Wiley & Sons; 2013.
9. Kimball R, Caserta J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning Conforming, and Delivering Data: Wiley; 2004.
10. Vercellis C. Business Intelligence: Data Mining and Optimization For Decision Making Chennai: John Wiley & Sons; 2009.
11. Abdul-Aziz R. Data, Text, And Web Mining For Business Intelligence: A Survey. International Journal of Data Mining & Knowledge Management Process. 2013; 3(1).

12. Lu Y, Wang , Maciejewski. Business Intelligence from Social Media: A Study from the VAST Box Office Challenge. IEEE Computer Graphics and Applications. 2014 Sept.-Oct.; 34(5): p. 58 - 69.
13. John H, Ashutosh T, Rajkumar R, Dymitr R. Computer Assisted Customer Churn Management: State-Of-The-Art and Future Trends. Computers & Operations Research. 2007; 34(10): p. 2902-2917.
14. Moro , Rita , Vala. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. Journal of Business Research. 2016; 69(1): p. 3341–3351.
15. Bhanap S, Kawthekar S. Data Mining for Business Intelligence in Social Network: A survey. International Advanced Research Journal in Science, Engineering and Technology. 2015; 2(12): p. 129-131.
16. Sangameshwari B, Uma P. Survey on Data Mining Techniques In Business Intelligence. International Journal Of Engineering And Computer Science. 2014; 3(10): p. 8575-8582.
17. He , Wua H, Yan , Akula V, Shen. A novel social media competitive analytics framework with sentiment benchmarks. Information & Management. 2015; 52(1): p. 801–812.
18. Tundjungsari V. Business Intelligence with Social Media and Data Mining to Support Customer Satisfaction in Telecommunication Industry. International Journal of Computer Science and Electronics Engineering (IJCSEE). 2013; 1(1).
19. Tan P, Steinbach N, Kumar V. Introduction to Data Mining Boston: Pearson AddisonWesley; 2006.
20. Mann T. Visualization of Search Result from the WWW Konstanz: University of Konstanz; 2002.
21. Williams S, Williams N. The Profit Impact of Business Intelligence San Fransisco: Morgan Kaufmann; 2007.
22. Batrinca , Treleaven PC. Social media analytics: a survey of techniques, tools and platforms. Artificial Intelligence & Society. 2015; 30(1): p. 89–116.
23. Sebastiani F. Machine Learning in Automated Text Categorization. ACM Computing Surveys. 2002; 34(1): p. 1-47.
24. Sheu JJ. An Efficient Two-phase Spam Filtering Method Based on E-mails Categorization. International Journal of Network Security. 2009; 9(1): p. 34-43.