2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)
April 09-11. 2021, ISTTS Surabaya, Indonesia

Customer Complaints Clusterization of Government Drinking Water Company on Social Media Twitter using Text Mining

1st Ajeng Dewinta
Faculty of Creative Design and Digital Business
Sepuluh Nopember
Institute of Technology
Surabaya, Indonesia
dewinta.ajeng@gmail.com

2nd Mohammad Isa Irawan
Faculty of Mathematics, Computation and Data Sciences
Sepuluh Nopember
Institute of Technology
Surabaya, Indonesia
mii@its.ac.id

2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT) | 978-1-6654-0514-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/EIConCIT50028.2021.9431931

***Abstract*—Social media is considered one of the most effective platforms to communicate between companies and customers. Frequently, the customer of a product or service sends complaints via social media. Customers' complaint data serve as a good suggestion for companies and organizations to improve their products and services. With the increasing number of customer complaints that have entered through social media accounts, government-owned drinking water companies need a more efficient way to extract information from complaint data. In this research, text mining is used to extract information about customer complaints against drinking water companies from social media Twitter. Latent Dirichlet Allocation (LDA) and self-organizing maps (SOM) approach is applied to model complaint topics and find out which are most frequently complained. The test results indicate grouping the data into five classes is the most appropriate model. Pipes leakage are the most frequently reported topics, 27.8% of total datasets.**

Keywords—clustering, LDA, SOM, social media, text mining

I. Introduction

The shifting process of how humans communicate and exchange information in the digital era has made social media an effective platform for companies to connect with their customers. Frequently, customers send the content of appreciation or complaints about products or services they use through social media. Whether about satisfaction or dissatisfaction written by social media users, the content of products and services can be seen by thousands or even millions of people worldwide, which will indirectly affect the image of a brand [1].

Private companies and government-owned companies use social media to communicate between service or product owners and customers. In the public sector, government water drinking companies use social media as a tool for announcing information about their products and services. Drinking water companies also use social media as a medium to receive customer complaints.

Nowadays, though companies use many social media, customers of products or services are more likely to convey their dissatisfaction through Twitter than Facebook [2]. Twitter is a microblogging site that allows users to exchange information. Text-based content from Twitter, also known as tweets, limited to 140 characters for each post [3] and often used in research [4]. Uploaded tweets have valuable information and can be used to analyze the meaning contained therein [5]. In the business world, content about products or services is beneficial for companies as additional input for their business development process. The company uses the contents of customers' disappointment of products or services to identify and manage dissatisfied customers [6]. Customers' complaints from social networks can be extracted and analyzed to determine what topics are complained of by customers and used for future product and service development.

The information extracting process is closely related to the use of data mining. Data mining is a method to dig up information on large data sets using a statistical approach to find patterns in the data. The data mining process has a crucial role and can help companies or organizations in industrial processes [7]. In its development, data mining is used to extract information on structured data and unstructured data such as text data [8], [9].

Text mining is a part of data mining that uses a statistical approach to extract information from text data by converting unstructured data into more structured data and facilitating accessing and managing data mining algorithms [10]. In extracting information from Twitter content, the text mining approach is often used to find the meaning of opinions issued in tweets, then analyze them and retrieve the contained information [1], [11], [12].

One of the objectives of extracting information on text content on Twitter is to determine the topics or problems from tweets. In companies, understanding the subjects about their products and services will benefit them in the business development. Topic modeling may be performed to gain information about the topics discussed in text data. The purpose of topic modeling is to determine the contained topics in a document with a statistical approach [13], [14].

Several methods can be used to perform topic modeling, one of which is by using Latent Dirichlet Allocation (LDA). LDA is often used as a topic modeling method [13]–[15]. The basic idea of LDA is to represent a document as a mixed model of the topics hidden within and to make the words characteristic of those topics. The application of LDA helps in more semantic and statistically better modeling of documents [16].

In this study, we determine to gather information concerning customer complaints in government-water drinking companies and determine what topics are more frequently raised on Twitter. We propose to use LDA-based

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 23,2021 at 20:28:09 UTC from IEEE Xplore. Restrictions apply.

to detect customer complaint topics at government-owned water utilities. LDA will generate complaint topics based on semantic and statistical approaches. The use of LDA is based on the possibility of several topics contained in one complaint tweet. Besides that, LDA also allows us to group customer topics without providing training data.

The LDA process produces a matrix output which shows that a tweet can consist of several topics [15]. We use the self-organizing map to grasp topic distribution. SOM is proven to be useful for use in text data clustering [17]. SOM will cluster data from the LDA process into groups that have similar characteristics. From this process, we will discover what complaints are most discussed on Twitter.

II. Literature Review

Extracting information from textual data to find text data's essence has been done since 1999 [18]. One way to extract the essence of document data is by doing topic modeling. In this study, we use LDA to carry out the topic modeling process. To reduce the dimensions, we use TF-IDF in removing unnecessary words. TF-IDF and LDA in modeling topics give better results than when LDA is run singlehandedly [19]. The results of modeling topics with LDA will be used as SOM (Self-organizing map) input data; thus, the cluster formed represents the input data.

A. TF-IDF

TF-IDF is a method for extracting data features and converting data vectors. Changing unstructured data into data vectors will improve information extraction with the text mining approach. TF-IDF also reduces commonly used words, thereby helping to speed up processing [14].

The TF-IDF algorithm was used to convert the data into vectors. In TF-TDF, tweet data were processed through preprocessing and converted into vectors using algorithms. TF-IDF algorithm is a combination of two statistics, namely Termed Frequency and Inverse Document Frequency. Term Frequency (TF) defines as the frequency of a term or word that appears in a document. Meanwhile, Inverse Document Frequency (IDF) is a weight used to measure the term's importance in text document collection [20].

$$tfidf(w) = tf * log\frac{N}{df(w)'} \quad (1)$$

Equation (1) shows that is defined as the number of occurrences of words in text data, is the number of documents that contain certain words, and N shows the number of documents. The result of this process was the word matrix used for the topic modeling process.

B. LDA Topic Modeling

Latent Dirichlet Allocation (LDA) is a method commonly used for topic modeling. The use of the LDA method was conducted with the consideration of obtaining better cluster results. Also, this method was added based on research wherein one document could consist of several topics [20].

LDA was used to conduct topic modeling, which was needed to determine the right number of topics. It can be done by analyzing a coherent score. The number of topics entered was used as a reference for categorizing the words included in topic classes. The determination results of the number of topics were carried out by looking at the score coherence used in inputting the classification of topics. In the LDA process, each tweet's data was searched for the number of possible topics contained in a tweet. The results of this LDA matrix were used at the next stage, namely knowledge discovery using SOM.

C. SOM Clustering

Self-organizing map (SOM) or known as Kohonen network, was introduced by Teuvo Kohonen as unsupervised learning in 1996 [21]. SOM has the purpose of transforming high dimensional data to low dimensional data and visualize it, hence able to be understood by human beings [22]. It also aims to maintain topological features of the input space pattern. The self-organizing map (SOM) principle has been widely used as an analysis tool and visualization in data analysis [23]. Learning using SOM for each input is conducted by identifying each processing element by applying the shortest Euclidean distance (Best Matching Unit, or BMU).

$$w_i(t+1) = w_i(t) + a(t)h_{ji}(t)(x_m - w_j) \quad (2)$$

Prototype vector (2) for this element and other elements in an environment are updated based on as prototype vector related to processing element jth. α (t) is the learning rate which decreases monotonically, h_ji (t). SOM algorithm can be divided into six processes, specifically [24]:

- Initialize the weight of each node;
- A randomly selected vector from datasets, then put it into the network;
- Examine each node in the network, calculate its weight, compared with the input vector, and change the node as part of the best matching unit (BMU). The vector is randomly selected from the training of datasets and presented to the network;
- Calculate the radius of BMU scope start from the largest radius;
- Each node within the BMU radius will be adjusted to the input vector, and if it is closer to the BMU, then the weight will change a lot;
- Repeat step two to obtain the smallest error value or until the weight has not changed.

III. Methods

A. Data Cleaning

This study acquired data on customers' complaints of government-owned drinking water supply company of the City Z, posted from June 2019 to March 2020. The data cleaning process was carried out; therefore, irrelevant and redundant data were deleted. Text content on Twitter has a limit of 140 characters which makes users use non-standard language and abbreviations. To overcome that problem, we propose a data cleaning process and preprocessing data, as illustrated in Fig. 1.

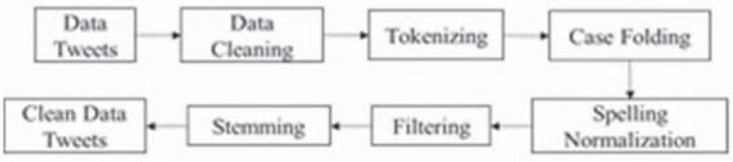

Fig. 1. Data Cleaning Workflow.

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 23,2021 at 20:28:09 UTC from IEEE Xplore. Restrictions apply.

Fig. 2. Data Processing Workflow

These steps were taken to change the textual data; thus, it has the same spelling. Also, meaningless symbols, numbers, and words were deleted, then other words are converted into basic words; this was done to reduce the number of index words in a document.

B. Data Processing

Fig. 2 presents the data processing workflow. First, the data processing stage was done by extracting the features from the text data. The TF-IDF algorithm was used to convert the data into vectors. Tweet data was processed through preprocessing stages then converted into vectors by weighting the words based on their appearance. The weighting results would form a corpus matrix used for the next process, namely modeling the topic with LDA.

The first process carried out in topic modeling with LDA was determining the right number of topics. The determination of the number of topics in the LDA was obtain the right iteration. The perplexity value was recorded, and its trend was analyzed. The analysis results in the perplexity test would be the input in the coherent test. In the coherent test, we tested several models with different k. The models used are the models with the high coherent score, or when viewed by the graph, the model with the coherent score before the graph tends to slope and dive down the selected ones. The iterative and coherent test results will be used as input in modeling topics with LDA.

The LDA process produces a matrix output which shows that a tweet can consist of several topics. Determining the distribution of topics can be seen by clustering tweets. In this study, the clustering algorithm, namely the self-organizing map, was used to classify tweets' tendencies on a topic. Tweets with a tendency of similar topics will be grouped into clusters according to their characteristics. The results of this grouping would show what complaints were most often submitted.

C. Cluster Validation

Measuring the quality of a model in the grouping can be done by performing cluster validation. In this study, cluster validation was carried out by analyzing the Silhouette score. The models tested were the models suggested from the coherent test process. In this study, the model with the highest silhouette score would be chosen as the best model.

IV. Results and Discussions

A. Data Cleaning

We collect tweets of customer complaints were posted through the City Z of the government's water drinking company's twitter account from June 2019 until March 2020. We removed the same and irrelevant data from the cleaning process; thus, we obtained 500 pieces of data used in the next process. We matched the form and spelling and removed punctuation, numbers, prepositions; therefore, the document's words became more standardized. The cleaned words would be further processed into the data processing stage to extract their information.

B. Data Processing

The results of preprocessing data were used as input in the word weighting process using TF-IDF. This weighting process calculated the word weight based on the occurrence frequency of words. This feature extraction stage produced a TF-IDF corpus containing word vectors. Thus, 500 tweets data in feature extraction produced 243 unique tokens.

1) Topic Modeling: The first step we took was to perform a perplexity test to determine the right number of iterations. The perplexity test was done by testing the model with the number of k 5, 10, 15, and 20 topics. Each topic will be tested 50 times from 1 iteration to 50 iterations. The results of the perplexity test are presented in Fig. 3.

In Fig. 3, we can see that from four experiments with 5, 10, 15, and 20 topics, the graph starts to look stable when iterated 19 to 21 times. In this experiment, it was concluded that the number of iterations used was 21 iterations. The next process was determining the number of topics. In this process, a coherent value was taken. The number of k used was one topics to 15 topics. The value of the coherence will be noted and presented in Fig. 4.

From Fig. 4, topic modeling would be carried out on the model with 5,9,11, and 13 topics. These models were chosen since the coherent value was higher than other models. The topic modeling process with LDA produced an LDA matrix that became the input for the clustering process using SOM.

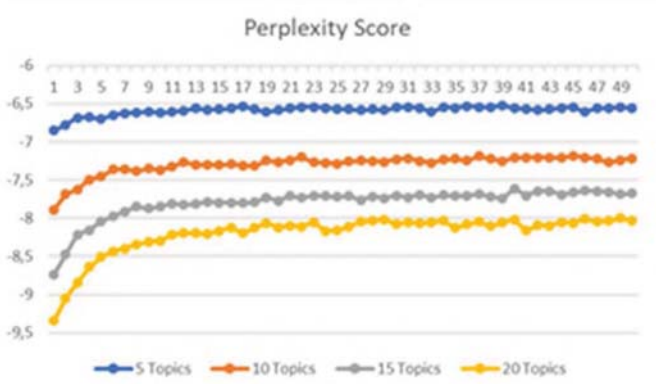

Fig. 3. Graph of Perplexity Test

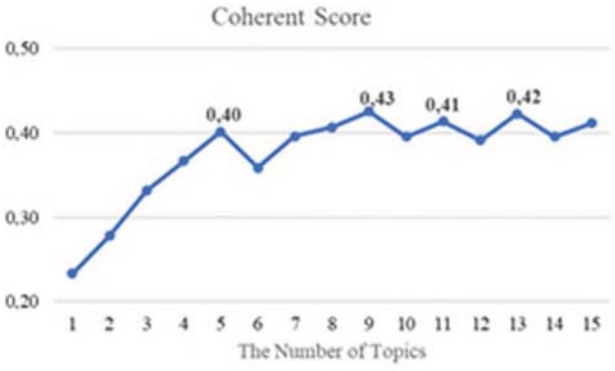

Fig. 4. Coherent Score

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 23,2021 at 20:28:09 UTC from IEEE Xplore. Restrictions apply.

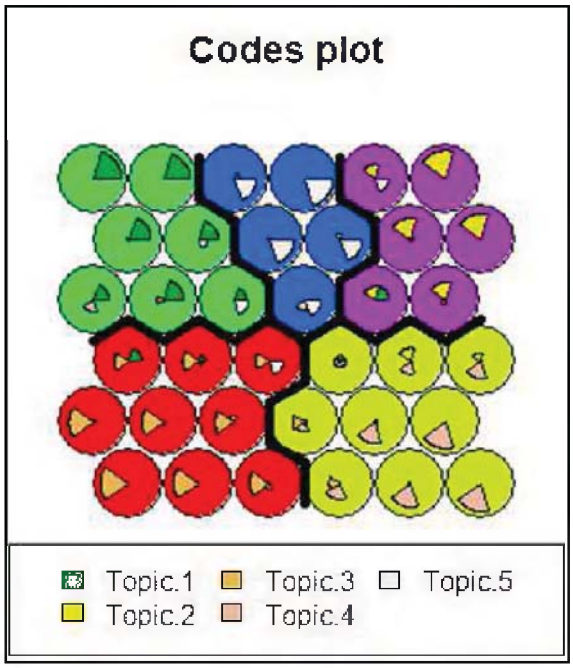

Fig. 5. Codes Plot Map Customers' Complaints' Data

2) Clustering: Topic modeling using LDA made it possible to recognize that it can contain several topics in one tweet. In determining the tendency of data on a certain topic, it was necessary to group them into clusters.

In the clustering process using the self-organizing map method, the data were transformed into a two-dimensional form. Five hundred tweet data were transformed into SOM map with dimension data 6x6 as shown in Fig. 5. The total of nodes in Fig. 5. are 36 nodes. The tweets were included in the most appropriate node for its characters.

The codes plot in Fig. 5 shows that 500 scattered customer complaints were filled the nodes with neighbors with characters similar to one another. The results of grouping using SOM were transformed into tabular form, as shown in Fig. 5.

.In Table 1, topic 3 and topic 4 have the same number of nodes. However, topic 1, which is only represented by seven nodes, has a larger amount of data in each node, specifically as many as 139 complaint data.

TABLE I. NUMBER OF NODES PER TOPIC

Topic	Number of Nodes	Number of Data
Topic 1	7	139
Topic 3	9	118
Topic 4	9	106
Topic 2	6	69
Topic 5	5	68
Total	36	500

TABLE II. SILHOUETTE SCORE

Number of Topics	Coherent Score	Silhouette Score
5	0.4	0.5314
9	0.43	0.5143
11	0.41	0.5190
13	0.42	0.5056

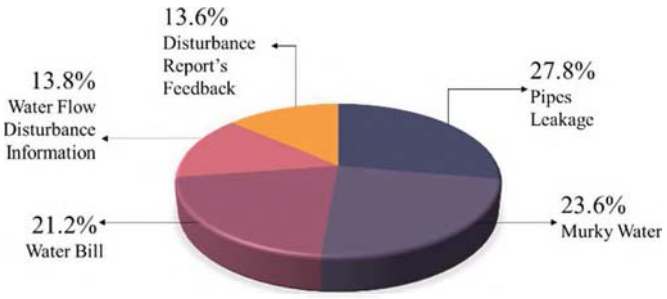

Fig. 6. Results of Customer Complaint Clustering

C. Cluster Validation

Based on the results of the coherent value analysis visualized in Fig. 3, the Silhouette taking process was carried out on four models with 5 topics, 9 topics, 11 topics, and 13 topics.

From Table 2, it can be seen that the number of topics of five has the highest score for the Silhouette, 0.5314. In performing cluster validation using the Silhouette score, the model with the highest score is considered the best model. Therefore, in modeling topics and grouping tweet data, a model with grouping tweets into five topics is the most suitable for this research.

The pie chart in Fig. 6 is the result of data transformation from code plot in Fig. 5. From there, we can see that Topic 1 regarding leakage pipes is the most frequently discussed complaint, amounting to 27.8% or 139 complaints from 500 clean data.

Topic modeling using LDA was used to determine what topics appear in a text data corpus. The SOM approach was carried out to determine the tendency of documents on a topic. SOM was helping to determine what topics were discussed most often. In the coherent value analysis in Fig. 3, we see the more k is entered, the trend of the coherent score increases [25]. Further analysis is needed to find the right model. In this research, we conducted cluster validation with silhouettes to determine the most suitable model. In Fig. 7, we can see that models with high coherent values do not always have high silhouette scores. A high silhouette score that represents the data was well segregated. We can conclude that a model with five topics is the most appropriate for this study from the silhouette test results.

V. CONCLUSION

In this study, we conducted a process of grouping customer complaints against City Z Government Drinking Water Company based on complaints posted on social media. TF-IDF method helps simplify the form of data into word vectors. It makes the data more structured and easier to extract information with text mining. Determining the number of topics in LDA topic modeling was the most appropriate model in determining model stage. The highest coherent scores model is not mean the best model. A silhouette validity test must be added to determine the most appropriate model. Silhouette test is done using the result of topic modeling matrix with LDA based on the coherent test results. The result of silhouette test show that grouping into five topic categories that have the highest coherent score of the recommendations are the most appropriate model for this study. Furthermore, the result of grouping customer complaints can be seen that pipes leakages have the most frequent complaints. It is

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 23,2021 at 20:28:09 UTC from IEEE Xplore. Restrictions apply.

necessary to improve in pipe maintenance and guarantee the products quality distributed to customers.

REFERENCES

[1] N. F. Ibrahim and X. Wang, "A text analytics approach for online retailing service improvement: Evidence from Twitter," *Decis. Support Syst.*, vol. 121, pp. 37–50, Jun. 2019.

[2] A. N. Smith, E. Fischer, and C. Yongjian, "How Does Brand-related User-generated Content Differ across YouTube, Facebook, and Twitter?," *J. Interact. Mark.*, vol. 26, no. 2, pp. 102–113, May 2012.

[3] S. Das, A. Dutta, G. Medina, L. Minjares-Kyle, and Z. Elgart, "Extracting patterns from Twitter to promote biking," *IATSS Res.*, vol. 43, no. 1, pp. 51–59, Apr. 2019.

[4] A. Singh, N. Shukla, and N. Mishra, "Social media data analytics to improve supply chain management in food industries," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 114, pp. 398–415, Jun. 2018.

[5] H. M. Chen, P. C. Franks, and L. Evans, "Exploring Government Uses of Social Media through Twitter Sentiment Analysis," *J. Digit. Inf. Manag.*, vol. 14, no. 5, p. 290, Oct. 2016.

[6] W. Fan and M. D. Gordon, "The power of social media analytics," *Commun. ACM*, vol. 57, no. 6, pp. 74–81, Jun. 2014.

[7] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning," *IEEE Access*, vol. 5, pp. 20590–20616, 2017.

[8] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Syst. Appl.*, vol. 66, pp. 245–260, Dec. 2016.

[9] A. Zaini, M. A. Muslim, and W. Wijono, "Pengelompokan Artikel Berbahasa Indonesia Berdasarkan Struktur Laten Menggunakan Pendekatan Self Organizing Map," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 6, no. 3, Sep. 2017.

[10] N. Zanini and V. Dhawan, "Text Mining: An introduction to theory and some applications," *Res. Matters*, vol. 19, pp. 38–45, 2015.

[11] F. R. Lucini, L. M. Tonetto, F. S. Fogliatto, and M. J. Anzanello, "Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews," *J. Air Transp. Manag.*, vol. 83, p. 101760, Mar. 2020.

[12] A. S. Halibas, A. S. Shaffi, and M. A. K. V. Mohamed, "Application of text classification and clustering of Twitter data for business analytics," in *2018 Majan International Conference (MIC)*, 2018, pp. 1–7.

[13] E. S. Negara, D. Triadi, and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," in *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 2019, pp. 386–390.

[14] S.-C. Tseng, Y.-C. Lu, G. Chakraborty, and L.-S. Chen, "Comparison of Sentiment Analysis of Review Comments by Unsupervised Clustering of Features Using LSA and LDA," in *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, 2019, pp. 1–6.

[15] H. M. Alash and G. A. Al-Sultany, "Improve topic modeling algorithms based on Twitter hashtags," *J. Phys. Conf. Ser.*, vol. 1660, 2020.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[17] Y.-C. Liu, M. Liu, and X.-L. Wang, "Application of Self-Organizing Maps in Text Clustering: A Review," in *Applications of Self-Organizing Maps*, InTech, 2012.

[18] A.-H. Tan, "Text Mining: The state of the art and the challenges," in *Proceedings of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD*, 1999.

[19] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing techniques for text mining-an overview," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[20] J. R. Millar, G. L. Peterson, and M. J. Mendenhall, "Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps," in *FLAIRS Conference*, 2009, pp. 69–74.

[21] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas, "Engineering applications of the self-organizing map," *Proc. IEEE*, vol. 84, no. 10, pp. 1358–1384, 1996.

[22] U. Asan and S. Ercan, "An Introduction to Self-Organizing Maps," in *Computational Intelligence Systems in Industrial Engineering. Atlantis Computational Intelligence Systems*, C. Kahraman, Ed. Paris: Atlantis Press, 2012, pp. 295–315.

[23] T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, vol. 37, pp. 52–65, Jan. 2013.

[24] S. M. Guthikonda, "Kohonen Self-Organizing Maps," 2005.

[25] A. Fang, C. Macdonald, I. Ounis, and P. Habel, "Examining the Coherence of the Top Ranked Tweet Topics," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 825–828.

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 23,2021 at 20:28:09 UTC from IEEE Xplore. Restrictions apply.