Hybrid Sentiment and Network Analysis of Social Opinion Polarization

Andry Alamsyah, Fidocia Adityawarman
School of Economics and Business, Telkom University
Bandung, Indonesia
andrya@telkomuniversity.ac.id

Abstract— **The rapid growth of social media and user-generated contents (UGC) has provided a rich source of potentially relevant data. The problems arise on how to summarize those data to understand and transforming it into information. Twitter as one of the most popular social networking and micro-blogging service can be analyzed in terms of content produced with sentiment analysis. On the other hand, some types of networks can also be constructed to analyze the social network structure and network properties. This research intended to combine those content and structural approaches into hybrid approach for identifies social opinion polarization, this is in the form of conversation network. Sentiment analysis used to determine public sentiment, and social network analysis used to analyze the structure of the network, detecting communities and influential actors in the network. Using this hybrid approach, we have comprehensive understanding about social opinion polarization. As case study, we present real social opinion polarization about reclamation issue in Indonesia.**

Keywords—opinion polarization; social network analysis; sentiment analysis; twitter; reclamation

I. Introduction

The rapid growth of social media and user-generated contents (UGC) has provided a rich source of potentially relevant data. Social media has changed the way of dissemination of information. Social media allows internet users become contributors and active disseminators of information. The existence of social media has given a new way to effectively spread information and even in real-time.

Twitter is one of the most popular social networking and micro-blogging service, with over 270 million monthly active users generating over 500 million *tweets* per day. Social media is generally guided by the ethics of social networking and communication from many-to-many communication, which often leads to the emergence of space for discussion and polarization of opinion in social media [1]. The problems arise on how to summarize this space, to understand, and transforming it into information. The data characteristic from online social network services is mostly unstructured, thus it needs new approach to analyze [2]

Twitter can be analyzed in terms of content produced with a sentiment analysis or opinion mining algorithm. Sentiment analysis is a field study of analyzing someone's opinion, sentiment, evaluation, assessment, attitudes, and emotions to an entity of a product, service, organization, people, issues, events and topics [3].

Some types of network can also be constructed from the data in it, e.g., a network of followers, mention, or retweet. Social Network Analysis (SNA) trying to build a network structure in graph format and analyze the properties of the network [4]. SNA helps researcher to understand the ensemble of actors, dynamics of the network, network cohesion and several others measurement such as search for actors who were most responsible for the network.

This research intended to combine those content and structural approaches into hybrid approach in order to have network of opinion polarization. Sentiment analysis used to determine public sentiment, and social network analysis used to analyze the structure of the network, detecting communities and influential actors in the network. There are some studies that have tried to measure opinion polarization. First, a research on a climate change topics got an interesting result of detecting an open forums and echo chambers inside the networks, that explains how users interacts with one another [1]. Second, a research that managed to find differences among the major communities in their sentiment leanings towards various environmental issues [4].

As a case study, we use an event recently happened debates or opinion polarization in Indonesian language twitter, i.e., the reclamation issue. Reclamation is a process of creating a new land from ocean through land filling. Polarization of opinion occurred partly associated with impacts, both environmental, social, and economic that might arise when reclamation is done. Broadly speaking, users split into two, pro-reclamation and counter-reclamation. Reclamation issues considered important because it relates to government policy, business interests, and society at large. Thus, the debate over the issue of reclamation must be better understood comprehensively.

II. Methodology

In this research, we analyzed the collected data using sentiment analysis to get users sentiment towards reclamation issue, then we use SNA to analyze the structure of the network,

ISBN: 978-1-5090-4911-0 (c) 2017 IEEE

i.e., network properties, detecting communities and identifying the most influential users in the network. The methodology flow shown in Fig.1.

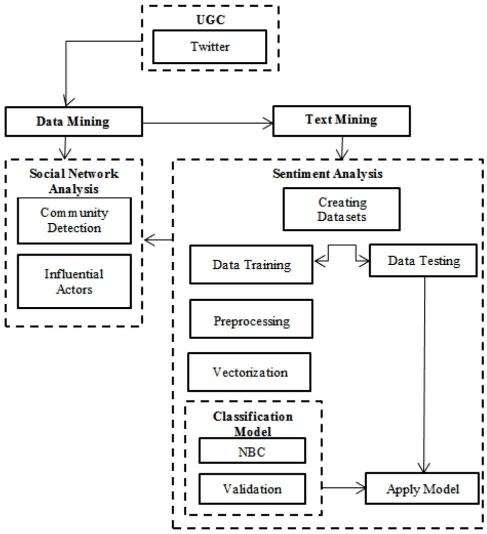

Fig. 1. Methodology

A. *Pre-processing*

First, we do pre-processing to clean the data from any noise as the preparation for the classification process. In the pre-processing, we perform these steps [5]:

1) Casefolding. Changes all the uppercase letter to lowercase.

2) Removing duplicated tweets. Removing duplicated tweets generated by users and/or bots, and also removing repeated tweets that returned by Twitter API.

3) URL and username removal.

4) Remove punctuation.

5) Replacing slang words with formal words listed in Indonesian dictionary (KBBI).

6) Tokenize.

7) Stopword removal.

B. *Vectorization*

To process the data in text mining algorithm, we need to change it into tabular data. The process of changing non-tabular data into tabular data is called vectorization. Vectorization method that is widely used is the TF-IDF vector space model [6].

TF-IDF will give a higher weight to word with low frequency of occurrence in some documents, and at the same time, have a high frequency of occurrence in a document. This ensures that the word with high TF-IDF value can be used as a representative example of a document in which class it is originated. So that words like "*ini*" (Indonesian language for "this") or "*itu*" (Indonesian language for "that"), which often appear in documents, will be given lower weights.

C. *Sentiment Analysis*

Sentiment analysis or opinion mining is a study of opinions, sentiments, evaluation, compliment, attitudes, and emotions towards an entity or object, i.e., product, service, individual, organization, event, or a topic [3]. Sentiment analysis is widely used to classify textual information with categorizing it as opinions and facts, positive and negative, or tiers in the category of emotions, e.g. angry, sad, happy [7].

Classification is a form of data analysis by creating a model that describes a class of important data [6]. The model, called a classifier, categorically predict class labels. Classification of data consists of two processes; learning step, in which the classification model is built and the classification step, using classification model to predict the class label of new data provided. In learning step, which is also called the phase of training, a training set is made with a class that has been labeled manually and used as data learning for later use in the classification model. In the classification step, the test data is used to estimate the accuracy of classification model that have been created. If the accuracy is considered appropriate and adequate, then the model created can be used to classify the test data.

This study uses Naïve Bayes classification method because it has been shown to have a high accuracy values [8]. Naïve Bayes Classifier (NBC) is a method of classification that can be used to predict the probability of class membership, such as the probability that a given tuple belongs to a certain class [6]. NBC assumes that the effect of a certain class attribute value independent from the values of other attributes. This assumption is called class-conditional independence. This was done to simplify the calculations involved and, in this sense, it considered "naive".

K-fold cross validation is used to reduce bias from training data when comparing the accuracy of predictions [9]. To find out how well or how accurate classification model that has been created, it can be measured by some calculations, namely accuracy, recall and precision. The measuring instrument used is the confusion matrix [6].

D. *Social Network Analysis*

The interactions network constructed from the twitter data by combining mention and retweet data into one network. We exclude tweets from user that was not making any interactions with another user. In the network, username represented as nodes, and interactions between users will be represented as an edge.

We calculate network properties such as density, modularity, diameter, average degree, average path length, clustering coefficient, and connected components.

Density is a statistical measure on the structure of the graph as measured by the proportion of edge connections relative to the total available connections, which will return a value between 0 to 1 [10]. A network with a density value closer to 1 normally considered as a solid network. Average path length is a measure of the efficiency of the whole network, by measuring the shortest path possible between all nodes in the network. Connected components used to determine how much the components in the network which is connected to the other components. So, this measurement can be used to determine if the network is fully connected. Clustering coefficient is a measure of the rate at which nodes were grouped together, as opposed to being equally or randomly connected across the network [10].

We use Louvain method [11] to detect communities in the network. In this method, the community as a group of users is assumed which the frequency of interaction is higher than others. Each user may only belong to one community.

III. Experiment and Dataset

We use Twitter Search API to collect Indonesian language messages shared on Twitter between 23rd June 2016 and 23rd July 2016 related to the reclamation issue, including personal *tweets*, *mention*, *reply*, and *retweet*.

We use several key words in the data mining process. Keywords are based on several location of reclamation, they are "*reklamasi*", "*reklamasi jakarta*", "*teluk benoa*", "*reklamasi teluk benoa*", "*reklamasi makassar*", and "*pulau G*". We also use hashtags, a mechanism that are utilised by twitter to reference a particular topic or event, thus enabling users to locate and contribute to related topics or discussions. The hashtags we use are *#reklamasi, #reklamasijakarta, #reklamasijkt, #telukjakarta, #telukbenoa, #reklamasibali,* and *#reklamasitelukbenoa.* In this research, we also use *#reklamasiuntukjakarta* and *#dukungreklamasi* that being used by pro-reclamation twitter users, and also *#tolakreklamasi* that used by counter-reclamation users.

From the data retrieval process during the time range, we collected 60,828 tweets. Then we remove the duplicated tweets, bot-generated tweets, spam, and irrelevant tweets from the dataset that resulting 23,115 tweets. For the purpose of social network analysis, to determine the sentiment of the users, we chose and take latest tweets from each user in the datasets and resulting 7,345 tweets.

IV. Results and Discusions

A. Sentiment Analysis

Sentiment classification divides the text data into two sentiments, the positive sentiment and negative sentiment. The result of the classification is influenced by the learning model created from training data. The better models are made, the better the classification which done by machines.

Training data used in the classification sentiment in this study consisted of manually labeled 1,974 negative data (counter-reclamation) and 1,667 text data labeled positive (pro-reclamation), thus totaling 3,641 labeled text data.

In the TF-IDF weighting process, we prune the attributes which occurrence is less than 0.83% and more than 90% to reduce complexity and improve accuracy. Training data generates 94 attributes of words used as a learning model. The confusion matrix is shown in Table. 1.

TABLE I. CONFUSION MATRIX

	True Positif	True Negatif	Class Precision
False Positif	1,631	58	96.57%
False Negatif	36	1,916	98.16%
Class Recall	97.84%	97.06%	

From the validation process using the k-fold cross validation and evaluation, we obtain 97.42% accuracy. From confusion matrix, we obtained 97.84% positive class recall value, and 97.06 % for negative class recall value. We also obtain 96.57% positive class precision value and of 98.16% negative class precision value.

Those values indicate that the classification model can classify the data very well. Achieve 97.42% accuracy means the possibility of a model to classify the test data correctly is high. From these values, can also be known that the ability of the classification model in measuring levels of predictive accuracy in a class (precision) and the success rate of the system in rediscovering a data deemed relevant to the class (recall) is high.

The test data consisting of 23,115 tweets was and successfully classify 14,981 or 65% of the text data as negative and 8,134 or 35% of text data as positive.

This means, in Indonesian language twitter, tweets with negative sentiment (counter-reclamation) are more dominant than the positive (pro-reclamation) tweets.

B. Social Network Analysis

In social network analysis step, to determine the sentiment of the users, we chose last tweets from each user during the time range of data collection in the study and get 7,345 tweets. The nodes sentiment was derived from the sentiment analysis that has been done before.

Network that is built consists of 4,832 nodes and 5,152 edges. Density calculation result is 0.0004308, which means the network is a very sparse. The calculation of modularity with 5 resolution value produce 4.564, and able to detects 777 communities. We add more resolution value instead of the standard value (1) so that communities that have very few members did not participate in the calculation of modularity detected. The diameter of the network is 5, which means the farthest distance between two nodes in the network is as much as five nodes.

Average degree calculation shows that the average nodes connected to each other is equal to 2.132. The average path length of the network is 1.3433031, which means the average number of nodes that must be passed by a node in order to reach a specific node is equal to the number. Clustering coefficient illustrates how a node connected to the node nearby. Clustering coefficient of the network amounted to 0,040. Value of connected components is 763, indicates that weakly connected component in the networks are as many as 763 components.

We build a graph where the red nodes are users with a negative sentiment (counter-reclamation), blue nodes are the users with a positive sentiment (pro-reclamation), and the gray colored nodes are nodes whose sentiment is not defined. Users with undefined sentiment are users who do not produce a tweet but the username was mentioned by another user, or a user that is active as a news portal that produce tweets such as news and therefore it tweets an objective sentence.

The proportion of users based on sentiment is 48.42% of users with negative sentiment (red), 22.81% of users with positive sentiment (blue) and 28.37% of users with undefined or neutral sentiment (grey). The sentiment network is shown in Fig. 2.

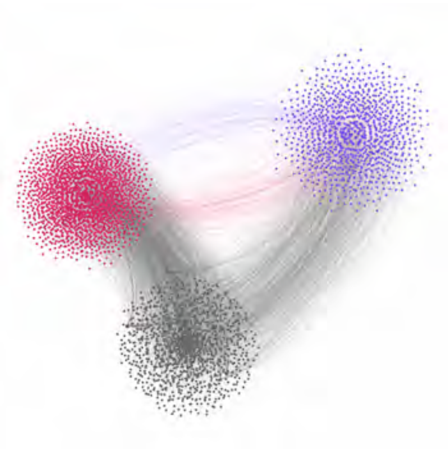

Fig. 2. Sentiment network

Social networks graph that has been constructed from Indonesian conversation on twitter related to reclamation issue showed that among the three partitions; i.e. pro-reclamation, counter-reclamation, and undefined sentiment. There are some interactions between those partitions. Most of the nodes with undefined sentiment are news portal twitter accounts. In the Fig.2., edges colored by sentiment targets. Thus, from the number of connections that are targeted at nodes with undefined sentiment, we see that interaction between nodes possibly occurs when:

1) Twitter users respond or expressing (in form of retweet, reply or mention) sentiments in it after news portal nodes producing a tweet.

2) Twitter users spread his/her views to individuals (nodes) that have not been clearly linked sentiments toward reclamation issues.

3) Twitter users expressing their grudges against username or account belonging to public figure that could be in a form of criticism or suggestions.

Furthermore, to analyze the polarization of opinion between pro-reclamation and counter-reclamation twitter users, can be seen in Fig. 3.

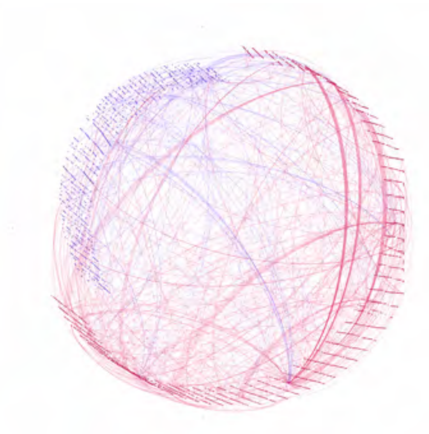

Fig. 3. Opinion polarization

To have a better understanding in observing the polarization of opinion, we filter the graph on nodes with positive and negative sentiment alone, and ignore the nodes with unidentified sentiment. Edges colored by sentiments of the source (blue means the interaction of positive nodes to nodes negative, and vice versa). From the graph, it can be seen there is a conflict of opinion between the two groups of opposing views (positive-negative, negative-positive). Although, it can also be seen that there is a tendency that nodes interact with other nodes with the same sentiment (positive-positive, negative-negative).

Overall, from graph analysis we know that the dominant interaction is nodes with sentiment expressing their sentiment toward nodes with undefined sentiment.

C. Community Detection

Fig. 4. Communities

Calculation result with Louvain method showed that there are 7 communities in 59.39% of the network. The remaining 40.61% consist of 770 small communities which only contains 2 to 22 nodes or 0.02% - 0.46%. Communities inside 59.39% network are presented in Table ll.

TABLE II. COMMUNITY DETECTION

Community	Size
1	22,56%
2	14,53%
3	5,84%
4	5,69%
5	4,51%
6	3,89%
7	3,37%

In the Fig.4. We give unique color for each community. Purple for community 1, light green for community 2, blue for community 3, yellow for community 4, orange for community 5, pink for community 6, and dark green for community 7. The communities with attachment sentiment shown in Fig.5.

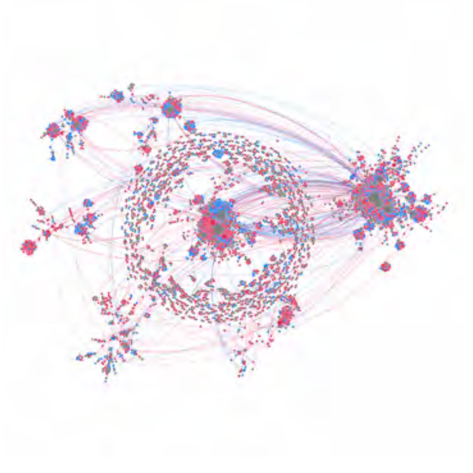

Fig. 5. Communities with sentiments

By taking the largest communities in the network (22.56% of the total size of the network), it is known that the community is composed of nodes that have different sentiments, it is shown in Fig.6. This means communities that exist in the network are a heterogeneous community.

Fig. 6. Largest community in the network

D. *Influential Actors*

To find influential actors in the network, we perform eigenvector centrality measurement and get 15 actors or nodes with the highest eigenvector centrality values presented in Table 3.

TABLE III. INFLUENTIAL ACTORS

No.	Username	Eigenvector Centrality
1	basuki_btp	1
2	susipudjiastuti	0,997694
3	jokowi	0,878031
4	ramlirizal	0,707369
5	kpk_ri	0,650982
6	temanahok	0,47271
7	kompascom	0,466773
8	rudolfdethu	0,43266
9	geetnotgood	0,418927
10	detikcom	0,362747
11	reiza_patters	0,345125
12	gendovara	0,329106
13	ilc_tvonenews	0,322044
14	metro_tv	0,268044
15	indrajpiliang	0,227413

Through eigenvector centrality calculations that can be seen in Table 3, it is known that twitter users with a username @basuki_btp is a node that has the highest value of eigenvector centrality among the other nodes.

The next rank is a Twitter user with the username @susipudjiastuti eigenvector centrality that has a value of 0.9976494, @jokowi with a value of 0.878031, @ramlirizal with a value of 0.707369 and 0.650982 @kpk_ri value. The first four nodes are a Twitter account belonging to public officials. This means that the opinions grudges directed by Twitter users to the accounts of the government, either in the form of criticism or suggestions.

News portal @kompascom become influential actors in seventh with eigenvector centrality value of 0.466773. Then @detikcom ranked 10th with a value of 0.362747, and @metro_tv ranked 14th with a value of 0.268044.

V. CONCLUSIONS

Based on the results of hybrid sentiment and network analysis of opinion polarization of reclamation issue, some conclusions can be drawn as follows:

Sentiment analysis performed on the data related to the reclamation issue in twitter using Naïve Bayes classification, we get 97.42% accuracy, so that it can be said that the classification model works well and reliably. Of 23,115 test data, the model successfully classify 14,981 or 65% of the text data as negative and 8,134 or 35% of text data as positive. Means in conversations related to the reclamation issue, counter-reclamation tweets are more dominant than positive or pro-reclamation.

From the network analysis, we know that the dominant interaction is nodes with sentiment expressing their sentiment toward nodes with undefined sentiment. By taking the largest communities in the network, it is known that the community is composed of nodes that have different sentiments. Influential actor detection was giving a result that the first four nodes is a Twitter account belonging to public officials. This means that the opinions grudges directed by Twitter users to the accounts of the government, either in the form of criticism or suggestions.

As conclusion, the hybrid method proven relevant in analyzing social opinion polarization in more comprehensive way than just using structural or content approach alone. Comprehensive in term of the information richness such as inter and intra group sentiment interactions, the heterogeneity or homogeneity group sentiment, and the possibility to measure the dynamics of network sentiments over time. This method can be used in any case that show sign of opinion polarization. Once we have enough evidence that there is conflicting interest, then we can apply the method.

REFERENCES

[1] Williams, Hywel T. P., James R. McMurray, Tim Kurz dan F. Hugo Lambert. "Network Analysis Reveals Open Forums and Echo Chambers in Social Media Discussions of Climate Change". Elsevier Journal of Global Environmental Change. 32, 126–138, 2015

[2] Alamsyah, Andry, Marisa P, Feriza J.P., Rizal H.. "Network Text Analysis to Summarize Online Conversations for Marketing Intelligence Efforts in Telecommunication Industry". The 4th International Conference on Information and Communication Technologies (ICoICT), 2016

[3] Liu, Bing, "Sentiment and Subjectivity Classification". Handbook of Natural Language Processing, 2, 1-38, 2010

[4] Sluban, B., Jasmina Smailovic, Matjaz Jursic, Igor Mozetic & Stefano Battiston, "Community Sentiment on Environmental Topics in Social Networks". The 10th International Conference on Signal-Image Technology & Internet-Based Systems. 377-382, 2016

[5] Aliandu, Paulina, "Sentiment Analysis on Indonesian Tweet". The Proceedings of The 7th ICTS, 203-208, 2013

[6] Han, J., Micheline K. dan Jian P, "Data Mining: Concepts and Techniques (3rd Ed.)". Waltham: Morgan Kauffman Publishers, 2012

[7] Moens, M., Juanzi L., Tat-Seng C., "Mining User Generated Content" . Chapman &. Hall/CRC, 2014

[8] Huang, J., Lu, J., Ling, C., "Comparing Naïve Bayes, Decision Trees, and SVM with AUC and Accuracy" . Proceeding of the 3rd IEEE International Conference on Data Mining, pp 553, 2003

[9] Olson, D. L. dan Dursun D. "Advanced Data Mining Techniques". Springer Science & Business Media, 2008

[10] Hanneman, R.A, Riddle, M. "Intoduction to Social Network Methods". Riverside, CA: University of California, Riverside, 2005

[11] Blondel, V. D., Jean-Loup G., Renaud L.3 and Etienne L,"Fast unfolding of communities in large networks, "Journal of Statistical Mechanics: Theory and Experiment, P10008, 2008