Network Text Analysis to Summarize Online Conversations for Marketing Intelligence Efforts in Telecommunication Industry

Andry Alamsyah[1], Marisa Paryasto[2], Feriza J. Putra[3], Rizal Himmawan[4]
[1, 3, 4] School of Economics and Business, [2]School of Electrical Engineering
Telkom University, Bandung, Indonesia
andrya@telkomuniversity.ac.id

Abstract — **Market tight competition put pressure the companies to employ a new and faster way to support their marketing intelligence effort. The need of marketing intelligence includes gathering and analyzing data for confident decision making about market and its competition. Today, the abundant large scale data from online social network services has made possible to extract valuable information such as user opinions and sentiment from the conversations in the market. As the competition arise, new challenge emerged, which include faster data summarization. The common practice of summarize contents is using wordcloud or weighted list of appearance words. This approach is lack of sense and contextual relations between words in questions, because the words has no connection with other words that might construct an important phrase. With the help of graph formulation, we propose a methodology of network text analysis to summarize large conversation in online social network services. This proposed methodology capture complex relations between words, while still maintain fast summarization. In this paper, we compare three major telecommunication provider in Indonesia, which is *Telkomsel, XL* and *Indosat.* The conversations about those brands in online social network services *Twitter* is collected, Network text about each brands are constructed and analyzed.**

Keywords—network text analysis; marketing intelligence; sentiment analysis; social network analysis; information discovery; graph theory; data summarization

I. Introduction

Internet has literally changed the way people to communicate, especially in the subject of human social interaction. It is shown by the emerge of many new online social network services to facilitate conversation in different context, storytelling and objective. These platforms serve as a media for large number of people to interact, thus it generates a large volume data, which in the term of data analytics can be a benefit for us, if we can transform those data into information [1]. One application that can profit this information is Marketing Intelligence effort [2].

Given the large data available, the problem arises on how to summarize those data, in order to understand the conversation topics. The simplest way to do is to count the appearance frequency certain words, in the popular term it is called as *wordcloud* or weighted list of words. This practice has been used extensively by the companies for the reason of its simplicity to summarize large volume of contents, including market conversations. The drawback of this approach is to lose the sense or context of words in question, because the words has no connection with other words that might construct an important phrase. On the other side of the approach is sentiment analysis approach, which is able to extract information from conversation in detail [3], but it is a time consuming process when we face large volume data to analyzed.

Most major companies present their channel in social media through popular online social network services such as *Twitter, YouTube, Facebook* and some others [4]. The contents of conversations in interest can be in the form of products reviews, exchange experiences, expert questions answers, and other socializing-related activities. Thus information discovery regarding opinions and issues concerning the market can be done from the data available. One way to find information linked to a certain topic can be done by summarizing conversations [5].

We are entering the era where sentiment analysis or opinion mining become a high necessity. Other than summarizing data, we can find the complex relations between the appearance words, in order to know the issues or senses of the opinion. Unfortunately, this process is very expensive in term of resources and time, while business and industries need faster and preferably close to real time process for their marketing intelligence effort.

Network text analysis [6] similar to social network analysis which model human relations. Both approach borrow a formulation from graph theory [7]. Each node represents a word, and each edge represent relations between words, if they appear in the same phrase. The edge weight construction come from how many times a pair of words shows up together.

In this paper, we propose a methodology to summarize conversations data in form of network text for supporting marketing intelligence effort. This approach provides more detail result than the current practice of *wordcloud*, while it is faster than sentiment analysis approach. A case study between three major telecommunication company in Indonesia is presented. We compare and analyze the summarizing result of each respective market conversations in a form of network text.

II. Data Summarization using Network Text Analysis

Methodology to analyze large data generated by user conversations are included in Big Data problem. Big Data is large volume data that surpass the capabilities of conventional database systems to process [8]. The data characteristic from online social network services is mostly unstructured, thus it needs new approach to analyzed. Big Data is vast and active research in its own interest. We found *Social Network Analysis* (SNA) [9] [10] methodology is practical solution to construct pattern from unstructured data. Due to similarity of the unstructured data in text analysis, network text analysis can benefit from social network analysis constructions.

We can model the social network with the help of graph theory. A graph $G(N,E)$ comprise of n number N nodes and m number of edges E. The nodes correspond to actors and the edges correspond to relations between actors. The advantages of social network model are intuitive, easy to visualize and accommodate several metrics based on graph theory characteristics. The model and set of metrics are subset of social network analysis methodology. In this paper, we use centrality and modularity metric [11]. An example of a social network visualization can be shown in Figure 1 [12].

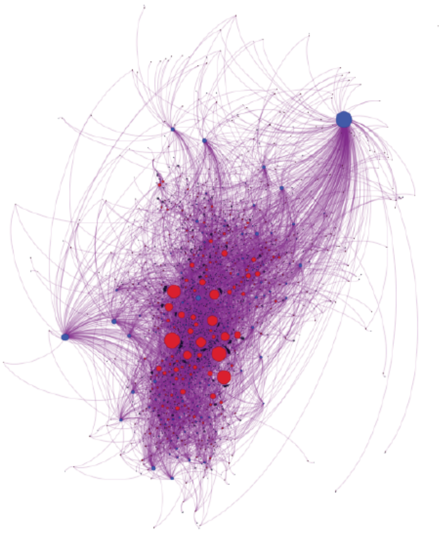

Fig. 1. An example visualization of a social network

Sentiment analysis, is a study of analyzing people opinions, sentiments, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. Automated sentiment analysis is needed because the average human reader will have difficulty identifying relevant sites, extracting and summarizing the opinions in them [13]. Conducting sentiment analysis procedure in large data is not practical and very expensive [14].

An important process to construct a network text is associative rules, which define as finding frequently appearance data together. The goal of this process is to detect relationships or associations between specific variable in large data sets [15]. In network text analysis, this process is used to calculate appearance a pair of words conversations phrase.

The combination of large data available, unstructured data, network model, and association rules leads to construction of network text analysis methodology to summarize large volume of conversation in online social network services.

III. Methodology and Experimentation

We crawl data concerning three major telecommunication company Indonesia by using keywords *"Telkomsel", "XL"* and *"Indosat"* from *Twitter*. The duration of data collection is in September 2015. Data or tweets is in Indonesian language. The data descriptions are as follows: Telkomsel is 46911 tweets, XL is 74771 tweets and Indosat is 14253 tweets. The workflow of network text analysis shown in Figure 2.

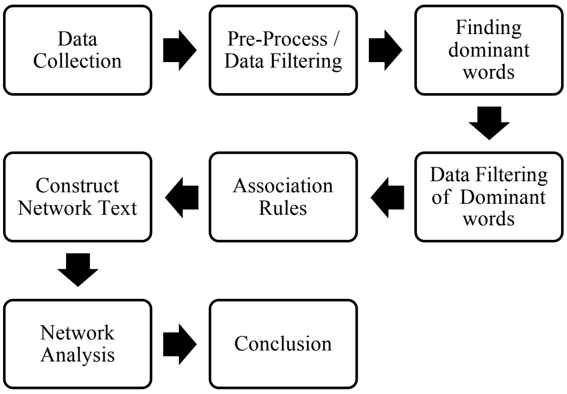

Fig. 2. Workflow of network text analysis

The first process is data collection; we crawl *Twitter* data directly using their application program interface. After we get large volume of tweets, we conduct pre-process, which include deleting irrelevant tweets to the topic we investigate. The third process is to find dominant words by using *wordcloud* applications. The fourth process is removing unsuitable dominant words to sentiments or opinions words, this include removing conjunctions, pronoun, numbers, date, promotions, advertising, spam, url links and others. The fifth process is to calculate association rules between dominant words. The sixth process is to construct network text of dominant word, which include weighted edge result for association rules process. At last, we analyze the data, by creating storytelling, context, and sense from network text. In network analysis, we employ centrality to find the most influential words in the networks and modularity to find words cluster / groups in network. The data profile is shown in Table 1 below.

TABLE I. Data Profile of each Telecommunication Company

	TELKOMSEL	XL	INDOSAT
NUMBER OF TWEETS	46911	74771	14253
AFTER SECOND PROCESS	7974	4521	2210
AFTER FOURTH PROCESS	7818	3743	1209

WORD CLUSTERS	3	2	5

The result from the fifth process which is association rules process for each company can be seen in Table ll, Table lll and Table IV in the form of words pair list. Due to space limitations, we only show the top 15 words pair for each company. A word pair is simplified representation, because in reality we have more than pair relationship, such as triangle, quadruple or more.

The centrality measures are shown in Table V to see which words are dominant in the network text. The centrality measure consists of two parts, which is degree centrality and weight degree. Degree centrality calculate the number of connection a word has. Weight degree calculate the number of connection of a word while it regards the weight summary of each connection to other words.

TABLE II. Top 15 Telkomsel Words Pair

Words pair	Weight
Internet - Paket	166
Paket - Tau	109
Paket - Flash	101
Loop - Simpati	96
Paket - Data	94
Paket - Kuota	91
Loop - Paket	89
Paket - Simpati	74
Internet - JAringan	64
Paket - Tarif	62
Internet - Kuota	54
Internet - Lambat	54
Paket - Pulsa	53
Paket – 4G	53
Murah - Paket	50

TABLE III. Top 15 XL Words Pair

Words pair	Weight
Sinyal – 3G	81
Sinyal - Internet	72
Paket - Internet	48
Paket – 3G	48
Internet - Jaringan	48
Edge - SOS	47
Internet – 3G	47
Edge - Keluhan	47
4G - Jaringan	46
Sinyal – Tidak Stabil	42
3G - Hotroad	41
Sinyal - Paket	40
Paket - Kuota	37
Edge – Tidak Stabil	37
Sinyal - Jelek	35

TABLE IV. Top 15 Indosat Words Pair

Words pair	Weight
Internet - Paket	37
Sinyal - Jelek	27
Kuota - Paket	24
SMS - Paket	22
Internet- Jaringan	22
SuperInternet -Internet	21
Data - Paket	20
Kuota - Internet	17
IM3 - Paket	15
Cepat - Kecepatan	15
Gangguan - JAringan	13
Handphone - Malam	13
Superinternet - PAket	12
Pagi - Keluhan	12
Sinyal - Internet	11

TABLE V. Degree Centrality and Weight Degree Measurement

Telkomsel		XL		Indosat	
Word	Degree Centrality / Weight Degree	Word	Degree Centrality / Weight Degree	Word	Degree Centrality / Weight Degree
Internet	59 / 1040	Sinyal	36 / 568	Internet	40 / 281
Paket	58 / 1892	Internet	36 / 519	Sinyal	37 / 144
Simpati	56 / 712	3G	33 / 432	Gangguan	33 / 100
Lama	55 / 431	Kuota	33 / 297	SMS	33 / 140
Data	54 / 476	Lambat	32 / 244	Paket	31 / 197

The network text constructed from words pair and word cluster with regard to degree centrality and weight degree measurement can be seen in Figure 3, Figure 4 and Figure 5 for each telecommunication company. In the picture, different node and edge colors means different words cluster. Thicker edge means more weight relations between nodes. For each company, we have dominant word pairs, word cluster and network text.

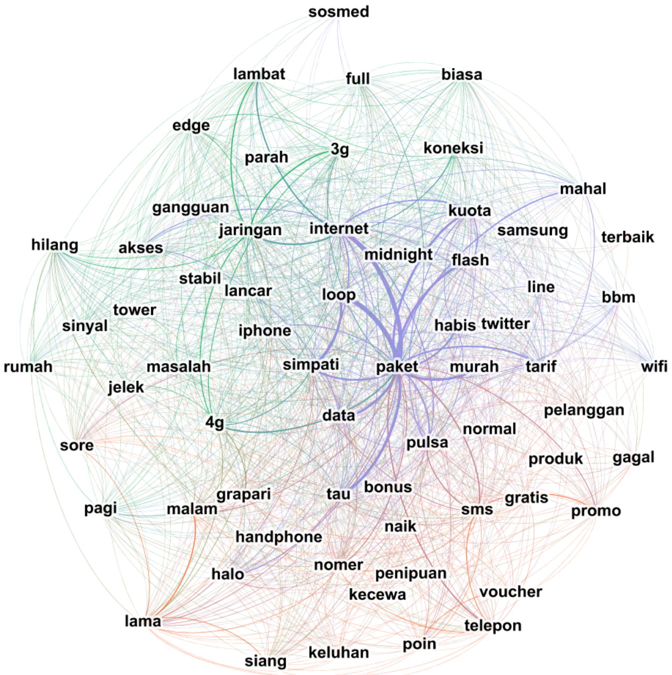

Fig. 3. Network text based on word pair and word cluster of Telkomsel

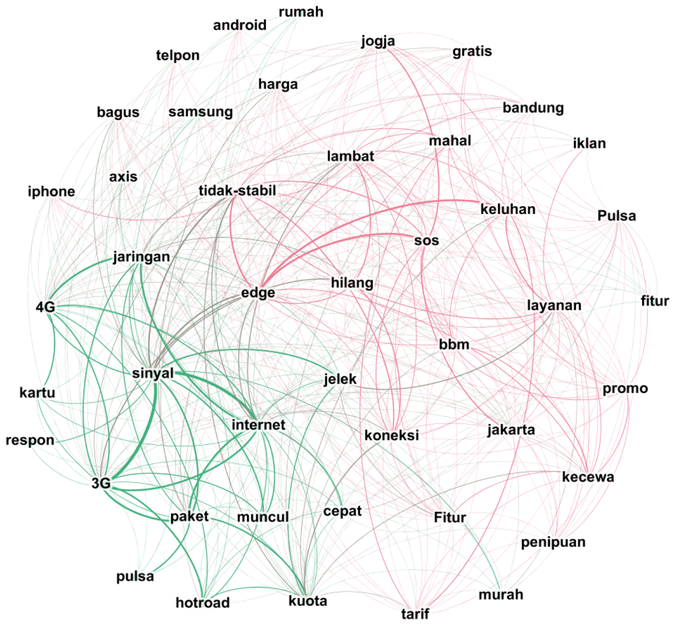

Fig. 4. Network text based on word pair and word cluster of XL

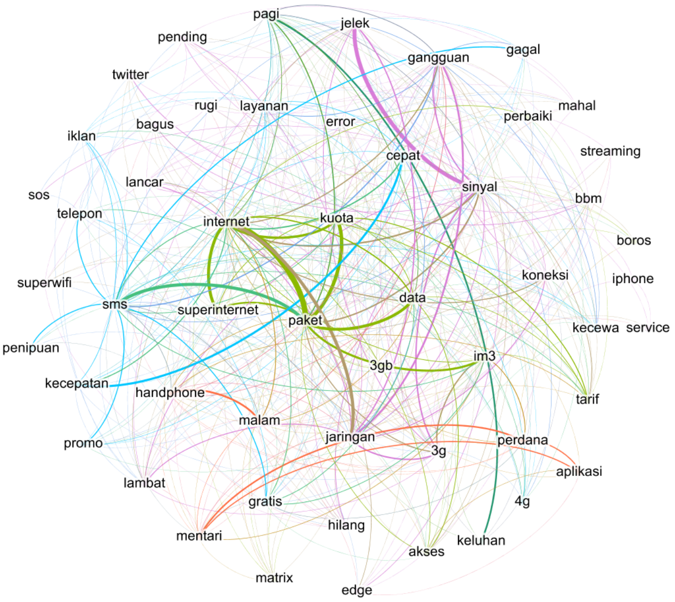

Fig. 5. Network text based on word pair and word cluster of Indosat

IV. Result and Analysis

For the case of *Telkomsel* network text shown in Figure 3. Words are connected in more centered fashion in which *"Paket"* and *"Internet"* words are in the center of the network. In the words pair, we have found that *"Paket, "Internet"* and *"Tau"* are the most weighted connection, this refer to the most dominant consumer expression is about their opinion in *Telkomsel* internet package product called *"Tau"*, which they consider it cheap and valuable product. We also have another major opinion which is concerning slow internet connection that customer experienced. The positive sentiment of *Telkomsel* network dominate the network text more than negative sentiment. From word cluster, we can see how words groups together construct different issue of *Telkomsel* products.

For *XL* network text in Figure 4, we found that words *"Sinyal", "3G",* and *"Paket"* dominate the networks text. Word cluster process found only two text groups. First group about *XL* internet package product information, with words such as *"Internet", "HotRoad", "Kartu", "Paket"* and some others. Second group dominated by complaints words such as *"Kecewa", "Penipuan", "Keluhan", Lambat", "Mahal"* and some others. We can check word cluster accuracy with word pair list where it also polarize the network into two main issues.

For *Indosat* network text in Figure 5, we found that *"Internet", "Sinyal"* and *"Gangguan"* are dominant words, this signify consumers have negative sentiment towards *Indosat* services, such as lost or bad network signal for data and voice, slow speed connection, and mostly they filed their complaint in the morning. Another dominant topics is information about *Indosat* internet product called *SuperInternet.* From word cluster process, we group the issues into several cluster that construct *Indosat* network text. Overall, we can see that Indosat network text contains more negative sentiments than positive sentiments.

As comparison between analyzed providers, there is always be one particular product that stands out as the most mentioned item from the group of information which has been collected. For example, in *Indosat* case, the most mentioned product is *"SuperInternet",* as for *Telkomsel* is *"Loop",* and *XL* is *"Hotroad".* This indicate that for all three analyzed companies, there is always be the most popular product among consumer. Furthermore it also shows the interest of consumer in general, thus we can modify information to target specific market segment. It is also possible to identify and provide the consumer need toward certain service.

The analysis of three telecommunication companies above give us insight about similarity and difference issues in telecommunication industry. We can classify the similarity issues are about product information and dissatisfaction about the services. The difference between those companies is from which sentiments is dominants, some have more positive sentiments than other, and so otherwise.

V. Conclusion

By heuristic, our proposed methodology will be more efficient in summarizing large volume conversation data from online social network services with more detail result, such as sense and contextual than *wordcloud* methods. This methodology is also faster than using sentiment analysis approach, because it did not employ learning procedures. In general marketing intelligence context, the usage of this methodology can benefit each company to measure the strength and weakness of themselves and the competitors.

Social network analysis method can help network text analysis methodology construction. It is explained that this method is useful in summarizing large number of data. Some advantage of this method are the capability to indicate negative and positive sentiment by the help of centrality and modularity metric, the simplicity to show the sentiments for both centered and scattered dominant connection, and as a side finding, this method also exceptionally useful in spotting the most popular product among consumer as well as the tendency it created.

As conclusion, this method proven more than relevant in summarizing large volume of data conversation in more effective and efficient way. Furthermore, examining popularity of a particular product, the tendency of product in question, and thus provide enormous aid to boost market analysis and brand evaluation.

REFERENCES

[1] C. C. Aggarwal. *Social Network Data Analytics*. Springer Science + Business Media, LLC. 2011

[2] H. Hedin, I. Hirvensalo, M. Vaarnas. *The Handbook of Market Intelligence: Understand, Compete and Grow in Global Market*. John Wiley and Sons. 2014

[3] B.Liu. *Sentiment Analysis: Mining Opinions, Sentiment and Emotions*. Cambridge University Press. 2015

[4] S. Stieglitz, L. Dang-Xuan, A. Bruns, C. Neuberger. *Social Media Analytics: An Interdisciplinary Approach and Its Implications for Informations Systems.* Business & Information Systems Engineering Journal, Volume 6, Issue 2, pp 89-96. April 2014

[5] H. Jee-Uk, J. Jin-Woo, I. Qasim, J. Young-Doo, C. Joon-Myun, L. Dong-Ho. *Multi-Document Summarization Exploiting Semantic Analysis Based on Tag Cluster*. Advances in Multimedia Modelling, volume 7733 of the series Lecture Notes in Computer Science, pp 479-489. 2013

[6] S. Hunter. *A Novel Method of Network Text Analysis. Open Journal of Modern Linguistics*, 4, 350-366. 2014

[7] R. Diestel. *Graph Theory: Electronic Edition 2005*. Springer-Verlag Heidelberg, New York 1997, 2000, 2005

[8] E. Dumbill. *Big Data Now: What is Big Data?*. O'Reilly, USA: O'Reilly Media, Inc. 2012

[9] D. Birke. *Social Network and their Economics : Influence Consumer Choice.* John Wiley & Sons, 2013

[10] A. Alamsyah, F. Putri, O. O. Sharif. Social Network Modelling Aprroach for Brand Awareness. *2nd International Conference on Information and Communication Technology, 2014*

[11] M.E.J. Newman. *Network: An Introduction.* University of Michigan and Santa Fe Institute. Oxford University Press, 2010

[12] Y. Peranginangin, A.Alamsyah. Social Engagement Analysis in Online Conversation In Indonesia Higher Education: Case Study : Telkom University. *3rd International Conference on Information and Communication Technology*. 2015

[13] A.Alamsyah, W. Rahmah, H. Irawan. *Sentiment Analysis Based on Appraisal Theory for Marketing Intelligence in Indonesia's Mobile Phone Market*. Journal of Theoritical and Applied Information Technology, Volume 82, No.2. 2015

[14] R.S. Ramanujam, R. Nancyamala, J. Nivedha, J. Kokila. Sentiment Analysis using Big Data. *International Conference on Computation of Power, energy Information and Communication*. 2015

[15] Y. Xiaoqing, L. Huanhuan, S. Jianhua, H. Jenq-Neng, W. WangGen, L. Jing. Association Rules Mining of Personal Hobbies in Social Networks. *IEEE International Congress on Big Data.* 2014