Ontology-Based Social Media Talks Topic Classification (Twitter Case)

Fransisca Julia Kusuma Deviyanti[1], Sri Suning Kusumawardani[2], P. Insap Santosa[3]

***Abstract*— In the era of digital communication, the use of Twitter as a customer service has been widely encountered. Companies have started to develop strategies around effective use of Twitter, one of which was to identify problems that customers frequently complain about. Twitter, with its straightforward tweet characteristics, will certainly contain sentences with very specific and easily recognizable keywords. These characteristics can be used as a basis for classifying tweets into certain topics. With a help of ontology, classification with keywords can be done automatically. The purpose of this paper is to design an ontology used as a basis for classifying tweets into certain topics related to the 4G telecommunications network in Indonesia and to evaluate performance of proposed classifier model.**

***Keywords*— Topic Classification, Twitter, Look-up Ontology**

I. Introduction

Telecommunications technology in Indonesia continues to grow, including telecommunications technology in cellular sector which now has entered the 4G LTE era. The growth of telecommunication technology has made it easier for people to access the internet and enter the digital communication era. In this digital communication era, Web 2.0 is a stimulant for the emergence of new media for social interaction. Twitter is one of media to do social interaction with a fairly high growth rate. This figure is a main magnet for companies that prioritize a customer engagement concept, namely the approach to customers through interaction in addition to purchasing activities [1]. This has made the company started to develop strategies around effective use of social media, including the 4G LTE telecommunications network provider companies.

The new 4G network ecosystem is certainly not separated from various opinions from customers, both directly on customer service and through Twitter accounts owned by the company. Twitter with its ease of use allows users to quickly write down the things that have just happened, including the services quality from a company. This makes the amount of information about the services quality of various 4G service provider companies rapidly increase. The customer's opinion on the company's services that continues to grow from time to time is a challenge for service providers to find information relevant to customer needs. The information is then used as a reference for telecommunication companies to continue to improve and develop as best as they can to create an increasingly mature ecosystem.

Topic classification using machine learning methods has been widely carried out. The limitation of this machine learning method, especially in the supervised learning method, is that it still needs training data [2], [3]. Twitter, with its straightforward tweet characteristics, contains sentences with very specific and easily recognizable keywords. This can be used as an alternative to classify topics without training data. Classification with these keywords can be carried out with ontology. This becomes a main foundation of selecting ontology as topic classifiers in Twitter talks.

Based on this background, this paper develops ontologies that can be used to process talks topic classification from Twitter data. Data gathering was carried out by conducting Twitter data crawling using Python library for Twitter, i.e., Tweepy. To recognize the problems frequently complained about by users, firstly, a data collection of relevant keywords for each topic was carried out. Afterwards, ontology was then constructed based on determined keywords using top-down paradigm. This was done because ontology construction departed from a predetermined domain, namely the 4G telecommunications network company in Indonesia.

II. Classifications and Ontology

A. Topic Classification with Ontology

Researches related to the topics classification using ontology have been carried out [2]. Method used in the study was based on similarity measurement on semantic thematic graph made from text document and ontology subgraphics resulted from projections of defined contexts. The language used as research object was English, so Wikipedia was used as a reference for building ontology. Classification method novelty in this research was that no document needed for training. This research has proved that the application of ontology could be used for text classification methods.

Similar research has also been carried out [4]. In contrast to previous research, this study used Spanish, especially those found on Twitter as research objects. This study aimed to show that programming language and ontology could help the analyzing process of tweets expressed by users. The developed ontology contains a collection of corpus representing negative and postive expressions. This study presented a way to exploit information found in tweets from Twitter using the API from Python and Twitter to obtain information from Twitter and ontology to classify tweets into two categories, namely negative and positive.

[1] *Department of Electrical and Information Engineering, Faculty of Engineering, Universitas Gadjah Mada, Jln Grafika 2, UGM Campus Yogyakarta 55281 INDONESIA (phone: 0274-552305; fax: 0274-552305; e-mail: fransisca.julia.k@mail.ugm.ac.id)*

[2,3] *Lecturer, Department of Electrical and Information Engineering, Faculty of Engineering, Universitas Gadjah Mada, Jln Grafika 2, UGM Campus Yogyakarta 55281 INDONESIA (phone: 0274-552305; fax: 0274-552305; e-mail[2]: suning@ugm.ac.id; e-mail[3]: insap@ugm.ac.id)*

B. Look-up Ontology

According to [5], look-up ontology is a process of finding phrases in a dataset based on phrases that have been stored in ontology. Research using the ontology look-up method has also been carried out [6]. Look-up ontology method was used to extract information in website content. In the research, information extraction process was carried out using General Architecture for Text Engineering (GATE). The conducted research shows that the use of ontology to extract information is an approach which performance is still possible to be improved in future studies.

Similar research was also carried out in [5]. In a research entitled a development of ontology-based information extraction method for Named Entities Recognition (NER), look-up ontology was used to anotate and label phrases representing one instance in ontology that was built based on tourism domain. Result of the study showed that system design to handle NER in tourism domain could be implemented without involving machine learning because knowledge represented by ontology could be understood by non-technical people.

C. Definition and Component of Ontology

The most popular definition of ontology is presented in [7]. According to the study, ontology is a description of the concepts and relations that might be formed within the concept. Whereas according to [8], ontology aims to aprehend knowledge that is conceptual in a general way, so that it can be reused by other groups and applications. In general, ontology can be interpreted as a theory of object's meaning, object's property, and the relation of the objects that might occur in a knowledge domain [9].

There are several main components constructing an ontology, namely concepts, relations, instances, and axioms [10].

$$O = (C, R, I, A^{\circ}) \tag{1}$$

Concept represents a set of classes and entities in related ontology domain. Concepts can be organized into a hierarchy. Relation describes an interaction between concepts or properties of a concept. Relations can be divided into two types, namely taxonomies types and associative types. Taxonomies type relations play a role in organizing concepts into sub-concept or super-concept, while associative type relations play a role in connecting concepts outside the hierarchy. Like concept, relations can be arranged into a hierarchy. Relation can also have properties that can describe relations' characteristics, such as cardinality and relations nature that it forms. Instances are components represented by the concept. An ontology must have an instance because an instance is a domain conceptualization. Combination of an ontology and interconnected instances can be called as a knowledge base. An axiom is used to limit concept value or instance. Property of relations can also be called as an axiom.

D. Protégé

Protégé is an open-source platform that provides tools for building domain models and knowledge-based applications with ontology [11]. With Protégé, ontology of relations for each subclass can be modeled and visualized in a form of a knowledge tree. Protégé has a Graphical User Interface (GUI) that makes it easy for users to use various tools embodied in it.

E. Classifier Model Testing

Testing is carried out by calculating value of accuracy, precision, recall, and f-score.

Accuracy is a measurement value of quantity measurement proximity level to the actual value, by showing how much data is accurately predicted. Accuracy can be calculated with (2).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

Precision is used to show the accuracy level of a classification system by showing the amount of correct data from data predicted to enter into class. Precision can be calculated by (3).

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

Recall is used to determine how well a system can return relevant results by showing the amount of entered data into a category that is correctly predicted. Recall can be calculated by (4).

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

F-score is a value frequently used in information retrieval to measure accuracy based on the value of precision and recall. F-scores can be calculated by (5).

$$F\ score\ = \frac{2\ x\ precision\ x\ recall}{precision + recall} \tag{5}$$

III. Method

Flow chart of the research is shown in Fig. 1. The research began with problems identification by conducting literature studies related to raised problems and needs identification including analysis of required equipment during carrying out the research, both hardware and software.

The next step was system design. At this stage, a design of method used in data collection on Twitter, ontology development method, and tweet classification method was carried out. Data was obtained by utilizing API provided by Twitter. The API could be accessed using a library owned by Python, i.e., Tweepy.

After that, it was proceeded with data collection and preparation phase. The utilized data was in a form of Tweet data obtained by conducting data collection (crawling) on Twitter. Crawling was carried out twice. The data on the first crawling was used as a reference for making a keywords list that will be included in ontology, while the second crawling was used as test data to compare the ontology performance. Data preparation stage carried out in this study was case folding, website address deletion, non-alphanumeric characters deletion, stopwords removal, stemming and calculation of terms occurrence frequency along with the instances determination for ontology.

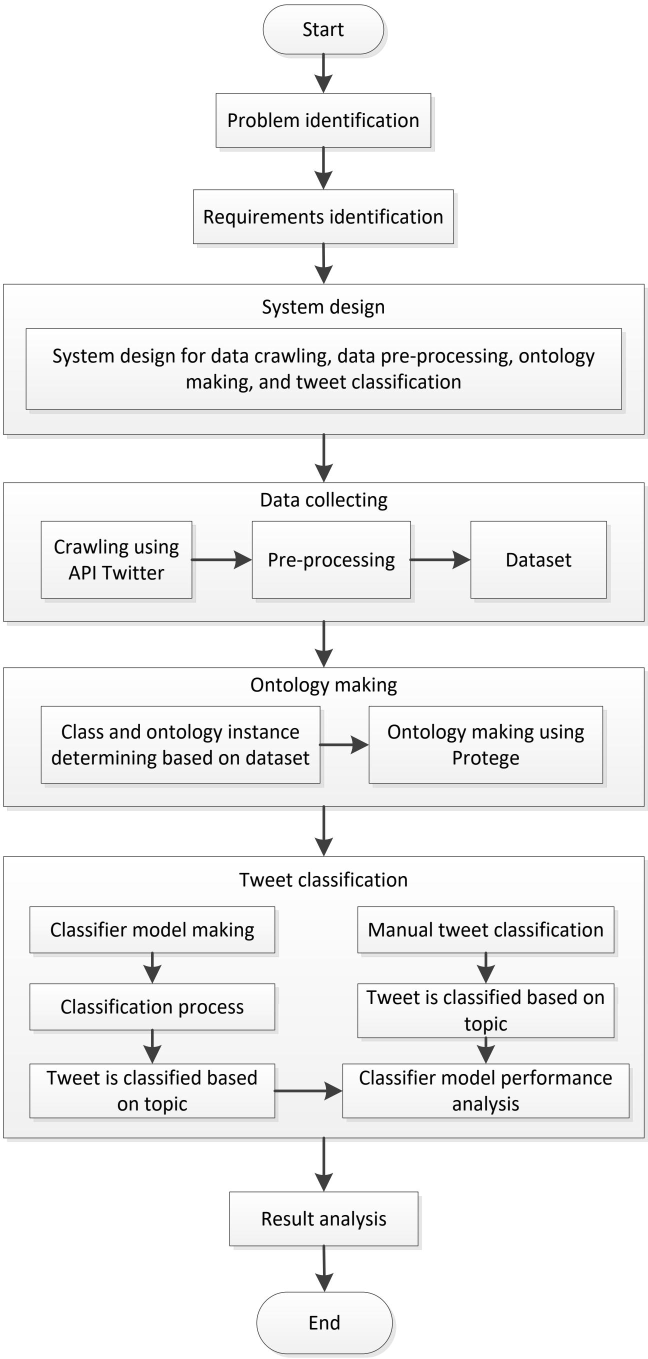
Start
Problem identification
Requirements identification
System design
System design for data crawling, data pre-processing, ontology making, and tweet classification
Data collecting
Crawling using API Twitter
Pre-processing
Dataset
Ontology making
Class and ontology instance determining based on dataset
Ontology making using Protege
Tweet classification
Classifier model making
Classification process
Tweet is classified based on topic
Manual tweet classification
Tweet is classified based on topic
Classifier model performance analysis
Result analysis
End

Fig. 1 Research flow chart.

Ontology creation stages began with grouping keywords that had been collected in the previous stage. This keyword grouping was carried out to determine classification reference class. A taxonomy overview of built ontology is illustrated in Fig. 2.

Instances of each class were determined from a list of selected keywords based on the domain and word occurence frequency. Instances on ontologies were built based on the keywords contained in tweets within a certain time period. In this paper, keyword collection was based on two considerations, namely the frequency of occurrence and the word domain.

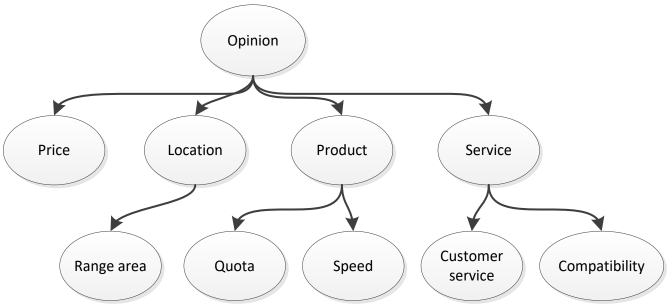

Fig. 2 Overview of ontology taxonomy in the study.

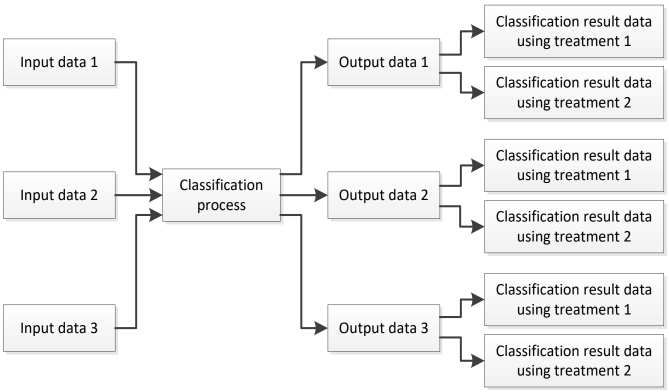
Input data 1
Input data 2
Input data 3
Classification process
Output data 1
Output data 2
Output data 3
Classification result data using treatment 1
Classification result data using treatment 2
Classification result data using treatment 1
Classification result data using treatment 2
Classification result data using treatment 1
Classification result data using treatment 2

Fig. 3 Testing scenario.

This tweet classification process was carried out in two ways, namely automatic tweet classification with proposed classifiers model and manual tweet classification for validation process. On manual classification, tweets were classified into “Kuota”, “Biaya”, “Area Jangkauan”, “Kompatibilitas”, “*Customer Service*”, and “*None*” classes with directions from Indonesian Literature students. Results of this manual classification were then used for validation process. The next classification process was the classification of tweets automatically. This process was carried out using ontology lookup method. Classification was done by ontology lookup method. Ontology lookup was utilized to provide a class and to classify a sentence according to the word class if the word represented an instance in ontology. Classifier model in this research was a basic prototype. Inside this model, there was no appropriate mechanism to handle tweet with more than one keyword which was also an instance of ontology. To overcome this issue, classification process was conducted with two treatments, namely Treatment 1 and Treatment 2. At treatment 1, senteces were labeled with class name of keyword found in the first tweet, while in Treatment 2, sentences were labeled with class name from last found keyword in a tweet. The conducted testing scenario is shown in Fig. 3.

To find out how good this classifier model was, it needed to be evaluated. Evaluation was carried out by mapping classification data into several results categories that could be described in a confusion matrix. After that, an analysis was carried out by calculating values of accuracy, precision, recall, and f-score based on mapping in confusion matrix. Then proceeded with an analysis of occuring classification errors causes.

IV. RESULTS AND DISCUSSION

A. Accuracy

Accuracy value shows a comparison between number of tweet correctly classified with total tweet number. Obtained average accuracy value is shown in Fig. 4.

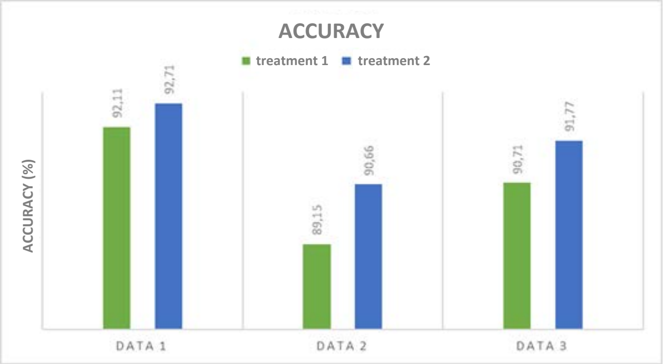

Fig. 4 Average accuracy value.

Based on the graph in Fig. 4, Output data 1 with Treatment 2 has the highest accuracy value. From these data it is known that in two periods of data retrieval, Treatment 2, namely classification of sentences labeled with class name of the last found word in a tweet, resulted in a higher accuracy value than tweet classified using Treatment 1, while accuracy value obtained from Output Data 3 is equal to average accuracy value of Output Data 3 because this classifier determines the class only with keywords appearing in a sentence. This certainly makes the classification results always the same.

B. Precision

Precision values indicate system preciseness to classify tweets into its class from all tweets. The average obtained precision value is shown in Fig. 5.

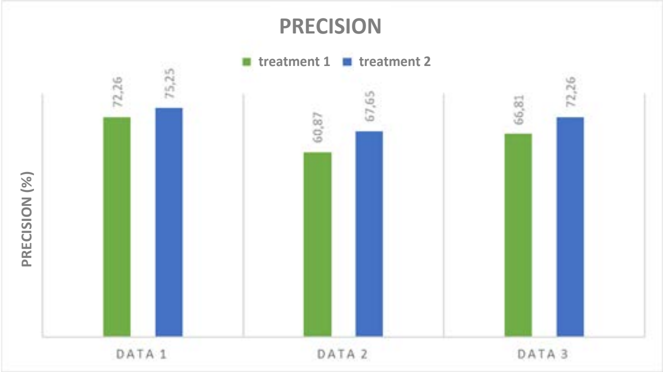

Fig. 5 Average precision value.

Based on the graph in Fig. 5, Output data 1 with Treatment 2 has the highest accuracy value. From these data it is known that in two periods of data retrieval, Treatment 2, namely classification of sentences labeled with class name of the last found word in a tweet, resulted in a higher precision value than tweet classified using Treatment 1, while accuracy value obtained from Output Data 3 is equal to average precision value of Output Data 3 because this classifier determines the class only with keywords appearing in a sentence. This certainly makes the classification results always the same as what happened in accuracy calculation.

C. Recall

The recall value indicates the success rate of the classifier in finding information. The average obtained recall value is presented in Fig. 6.

Based on Fig. 6, Output Data 1 with Treatment 2 has the highest recall value. From these data, it is known that in two periods of data retrieval, Treatment 2, namely classification of sentences labeled with class name of the last found word in a tweet, resulted in a higher recall value than tweet classified using Treatment 1. The obtained recall value from Output Data 3 is equal to the average recall value of data 1 and data 2 because this model classifier determines class only with keywords appearing in a sentence. This certainly makes the classification results always the same.

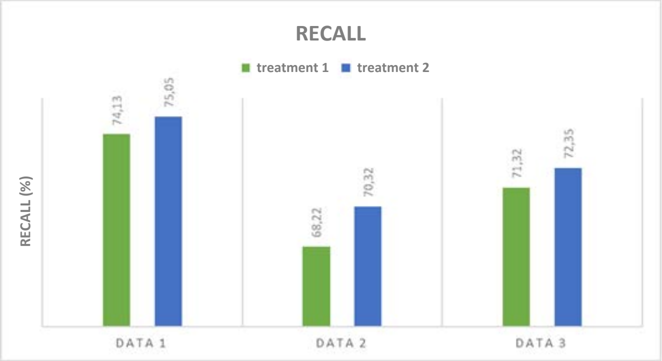

Fig. 6 Average recall value.

D. F-score

F-score is a value frequently used in information retrieval to measure accuracy based on the value of precision and recall. The average obtained f-score is shown in Fig. 7.

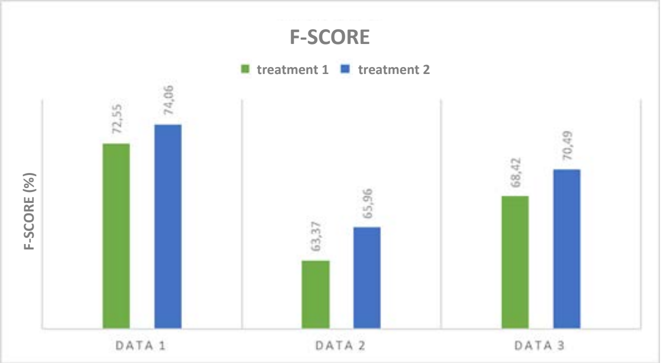

Fig. 7 Average f-score value.

Based on graph in Fig. 7, Output Data 1 with Treatment 2 has the highest f-score value. From these data it is known that in two periods of data retrieval, Treatment 2, namely classification of sentences labeled with class name of the last found word in a tweet, resulted in a higher f-score than tweet classified using Treatment 1 The obtained f-score value from

Output Data 3 is equal to the average recall value of data 1 and data 2 because this model classifier determines class only with keywords appearing in a sentence. This certainly makes the classification results always the same.

E. Causes of Classification Errors

With this ontology lookup classification method, a tweet containing one or more instance names under a same class will be correctly classified. However, errors still occur when dealing with tweets that have more than one instance that has different classes. Prediction errors occuring when using Treatment 1 are shown in Table I.

Keywords which ares also as instances in ontology are indicated by bold words. In this example, the keywords found are "kartu", "paket", and "stabil". "Kartu" is an instance of “Kompatibilitas” class, "paket" is an instance of “Kuota” class, and "stabil" is an instance of “Kecepatan” class. Because applied treatment in this test is Treatment 1, with the ontology lookup process stopped when in a tweet a word representing an instance in ontology has been found, then label given for this tweet is “Kompatibilitas” class, even though the sentence discusses about 4G internet speed provided by one network provider. When classified using Treatment 2, tweet will be predicted as same topic as class name of "stabil" instance, namely the “Kecepatan” class.

TABLE I
EXAMPLE OF PREDICTION ERROR IN TREATMENT 1

Tweet	Predicted Class	Actual Class
Parah ini **kartu paket** warna biru xl, gak pernah **stabil** jaringan 4G seperti 2G	[opinion. “Kompatibilitas”]	[opinion. “Kecepatan”]

As happened in the classification process with Treatment 1, Treatment 2 also has similar limitations. Some examples of topic prediction errors when classification is done with Treatment 2 are shown in Table II.

Words written in bold in Table II are words that are ontology instances were built in this study. In this example, the written keywords are "lemot" and "kuota". After passing through the data preparation stage, the word "lemot" was found to have changed to "lot" because of overstemming. Overstemming occurs because the word "lemot" contains "-em" which is one of inserts in Indonesian. Therefore, the word "lemot" on a tweet is not detected as an instance in ontology, so tweets are classified based on other keywords in the tweet, namely the word "kuota". In ontology, word "kuota" is an instance of “Kuota” class so that tweet is put into “Kuota” class. Because the applied treatment applied in this test is Treatment 2, with the ontology lookup process stopped when keyword/instance is the last instance found, then the label given for this second tweet is “Kompatibilitas” class, even if classification is done with Treatment 1, obtained prediction class will be the same as the actual class, namely the “Kecepatan” class.

Prediction errors can also occur when keywords expected to represent topic of a tweet are among incorrect keywords. Some examples of errors are shown in Table III.

TABLE II
EXAMPLE OF PREDICTION ERROR IN TREATMENT 1

Tweet	Predicted Class	Actual Class
msih belum ada solusi nih ya sinyal 4g **lemot** di cibungur purwakarta. nunggu sabar sampe **kuota** ilang gak kepake, gmna kelnjutannya?	[opinion. “Kuota”]	[opinion. “Kecepatan”]

In Table III, keywords written in example are "daerah", "hilang", and "paket". "Daerah" is an instance of “Area Jangkauan” class, "hilang" is an instance of “Kecepatan” class and "paket" is an instance of “Kuota” class. When classified with Treatment 1, tweet is classified as “Area Jangkauan” topic, while when classified with Treatment 2, tweet falls into “Kuota” topic. With these two possibilities, tweets cannot be classified correctly, either with Treatment 1 or Treatment 2, because the instance that actually represents tweet's topic is after the first instance, namely "daerah", and before the last instance, namely "paket". Tweet in this example should be included in “Kecepatan” class which is an integrated class instance, which is "hilang".

TABLE III
EXAMPLE OF PREDICTION ERRORS BECAUSE RIGHT KEYWORDS ARE IN THE MIDDLE

Tweet	Predicted Class		Actual Class
	Treatment 1	Treatment 2	
Di**daerah**ku jaringan 4G **hilang** timbul bahkan sering ilang jadinya **paket** data 4G gk trpakai sampai masa aktif habis.	[opinion. “Area Jangkauan”]	[opinion. “Kuota”]	[opinion. “Kecepatan”]

In addition to a number causes that have been described, there are some stemming error as mentioned earlier. Stemming errors are found to occur in words that are one of built ontology instances, namely the word "lemot". "Lemot" in ontology belongs to “Kecepatan” class. Stemmer considers that "-em" in the word "lemot" is an insertion, so that "-em" is omitted from the word "lemot". The word "lemot" which loses "-em" changes to "lot". This makes all tweets that contain the word "lemot" cannot be classified into the “Kecepatan” class unless there are other keywords that are “Kecepatan” class instances.

F. Strengths and Limitations

Ontology built on this research is used to classify topic of a sentence using ontology lookup method. Ontology lookup method for classifying topics in a tweet in Indonesian has never been found in previous studies. In addition, even though performance value cannot be said to be high, topic classification using ontology can be done without using training data so that it can simplify steps in topic classification process.

This classifier model has limitations. The limitations due to its inability to classify tweets in real time and display visualization of their classification results. To be able to classify tweets in realtime, the system must be able to do automatic queries on Twitter. Another limitation can be identified with the decreasing number of classifier model testing result in data 1 and data 2. The decrease is caused by a change in keywords used by user in expressing a conversation topic. It shows that static ontology will not produce a better performance. In addition, in the designed classifier model there is no appropriate mechanism to deal with tweets with more than one keyword which is also an ontology instance.

V. Conclusion

Conclusion that can be drawn from the conducted research is that from the tested data side, classifier method tested in Output Data 2 produces a lower value than the value obtained from Output Data 1. Performance reduction occurs because the built ontology is still static so it has not been able to handle differences in keywords written by customers from different periods.

On the other hand, in terms of treatment, testing with Treatment 1 produces a lower value than Treatment 2. It shows that dataset resulted from crawling in this study, correct keywords are mostly located at the end of sentences.

References

[1] E. Rahmawati and Sanaji, "Pengaruh Customer Engagement Terhadap Kepuasan Pelanggan dan Kepercayaan Merek serta Dampaknya pada Loyalitas Merek," *J. Res. Econ. Manag.*, Vol. 15, pp. 246–261, 2015.

[2] M. Allahyari, K.J. Kochut, and M. Janik, "Ontology-based Text Classification into Dynamically Defined Topics," *IEEE Int. Conf. Semant. Comput.*, 2014, pp. 273–278.

[3] P.W. Basnur and D. Indra, "Pengklasifikasian Otomatis Berbasis Ontologi untuk Artikel Berita Berbahasa Indonesia," Vol. 14, No. 1, pp. 29–35, 2010.

[4] D.L.C.S. Carlos and C.H. Paola, "Extraction and Classification of Twitter Messages to Apply in Business Intelligence," *Lect. Notes Softw. Eng.*, Vol. 1, No. 2, pp. 126–130, 2013.

[5] K.E. Saputro, S.S. Kusumawardani, and S. Fauziati, "Pengembangan Metode Ekstraksi Informasi Berbasis Ontologi untuk Pengenalan Named-Entities dari Halaman Web Tidak Terstruktur," Thesis, Universitas Gadjah Mada, Yogyakarta, Indonesia, 2017.

[6] E. Susanti, "Ekstraksi Informasi Konten Web Menggunakan Pendekatan Berbasis Ontologi," *Jurnal Teknologi Technoscientia*, Vol. 7, No. 2, pp. 128-136, 2015.

[7] T.R. Gruber, "A Translation Approach to Portable Ontology Specifications," *Knowl. Acquis.*, Vol. 5, No. 2, pp. 199–220, Jun. 1993.

[8] S.S. Kusumawardani, R.S. Prakoso, and P.I. Santosa, "Using Ontology for Providing Content Recommendation Based on Learning Styles inside E-learning," *Proc. - 2nd Int. Conf. Artif. Intell. Model. Simulation (AIMS 2014)*, 2014, pp. 276–281.

[9] H. Jayadianti, L.E. Nugroho, P.I. Santosa, W. Widayat, and C.S. Pinto, "Ontology sebagai Solusi Pencarian Makna Ambigu dalam Sistem yang Heterogen," *Telematika*, Vol. 10, No. 1, pp. 63–70, 2013.

[10] Z. Ma and H. Wang, *The Semantic Web for Knowledge and Data Management: Technologies and Practices*. Hong Kong: IGI Global, 2008.

[11] (2016) "Protégé." [Online], http://protege.stanford.edu/products.php, access date: 14-Feb-2017.