Generating of Automatic Disaster Hashtag Based on OCHA Standard

Kharisma Wiati Gusti
School of Electrical and Informatics Engineering
Bandung Institute of Technology
Bandung, Indonesia
23515041@std.stei.itb.ac.id

Rila Mandala
School of Electrical and Informatics Engineering
Bandung Institute of Technology
Bandung, Indonesia
rila@informatika.org

***Abstract*— The use of disaster hashtags especially Twitter does not have a standard format, this makes it difficult to search and collect data for emergency response. OCHA (Office for Coordination of Humanitarian Affairs) proposes to standardize the hashtag for emergency response. This study was conducted to generate automatic disaster hashtags in accordance with the OCHA standards. This research used word representation method with Skip Gram model, and SMOTE filter for handling imbalanced datasets. Then the classification of tweets and identifiers of named entities was carried out using the Conditional Random Field (CRF) model. The results showed that the use of Skip-Gram models was able to improve accuracy. The highest average accuracy of 83.9695 % was obtained by using instance based learning with k 15 and the introduction of an entity named with *recall* 70.3%, *precission* 89.4% dan *f-measure* 77.1%. Automatic hashtag generation has good results with an average of 61.2% recall, 87.4% precision and 66.9% f-measure.**

Keywords-component; generation, hashtag, disaster, automatic, skip gram

I. INTRODUCTION

Twitter has been widely used as a communication tool for emergency response when disasters occur in various countries. Emergency response teams or researchers use hashtags for emergency disaster searches. The use of disaster hashtags in Twitter does not have a standard format, this makes it difficult to search and collect data for emergency response. OCHA (Office for Coordination of Humanitarian Affairs) proposes to standardize hashtags for emergency response [13].

Information on disaster names and locations was used to make disaster hashtags automatically in accordance with the standards proposed by OCHA. This research was made to facilitate the emergency response team and researchers in getting disaster information in Indonesia from Twitter. In addition, it is easier for users to report events and needs in an emergency, so users do not need to think about hashtag rules.

The next section will discuss related studies, and the experimental design is discussed in section 3. The results of the experiment and discussion are discussed in section 4. At the end of the paper, the conclusions of the research are discussed.

II. RELATED STUDIES

This section will discuss several related fields of study, e.g. hashtag recommendations, machine learning, and named entity recognition.

A. Hashtag Recommendation

Several studies have been conducted to resolve the issue of hashtag recommendations using various methods. Hashtag recommendations using collaborative filter filtering by combining tweets and similar users [8]. Xiao et al, recommended hashtag-oriented topics based on the detection of co-occurrence word characteristics [20]. While Godin et al, designed and developed a binary classifier to tweet based on the Naive Bayes and Expectation-Maximization methods. Furthermore, the Latent Dirichlet Allocation (LDA) model was applied in the context of a tweet hashtag recommendation [5].

Gong & Zhang combined Convolution Neural Network (CNN) with attention mechanism architecture to avoid hand-craft features in solving hashtag recommendation problems [6]. The use of attention mechanism was also used by Li et al, in this case the attention mechanism used was based on topics combined with Long Short-Term Memory (LSTM) for hashtag recommendations [10]. Unlike Li et al, who used LSTM and RNN (Recurrent Neural Network) to study vector-based tweet representations for hashtag recommendations [9]. Tomar et al, used distributed word representation (Word Embedding) and Feed Forward Neural Network (FFNN) [17]. This hashtag recommendation targets English tweets on twitter. Distributed word representation was generated from the continuous Skip Gram model.

Twitter automation has been done by Tran and Popwich for traffic information. This approach was based on the technique of making natural language data (Natural Language Generation) using couplings that contain examples of natural language tweets [18].

ISBN 978-602-53524-0-9

B. Machine Learning

Machine learning is used for the classification of Twitter disasters. Each machine learning algorithm has its own characteristics in each model formed from learning data. This section describes four machine learning algorithms, namely Naïve Bayes, Instance-Based Learning, Support Vector Machine, and Logistic Regression.

Naïve Bayes classifier is a classification method that uses probability calculations. The naive Bayes classification is assumed that the presence or absence of certain characteristics of a class has nothing to do with the characteristics of other classes.

$$p\ (C\ |D) = \frac{p(D\ |C)\ p(C)}{p(D)} \quad (1)$$

Where D is the class data that is not yet known, H is the D hypothesis on a particular label, P (C | D) is the probability C based on condition D (posteriori), P (C) is the probability C (prior), P (D | C) is the probability of D in hypothesis C, P (D) is probability D [16].

Instance-Based Learning in the Weka application named IBk is an algorithm that implements k-Nearest Neighbors (KNN). K-Nearest Neighbor (KNN) is a method commonly used in data classification, using a supervised learning approach so that it requires training data that has been labeled. This algorithm is an instance-based learning or lazy learning type so it does not produce learning models, but keeps all existing learning data so that all calculations are postponed to the time of classification [4]. KNN is done by looking for groups of objects in the training data that are closest (similar) to objects in new data or data testing. This algorithm is used to classify objects based on learning data that are the closest neighbors or have a small difference with the object.

Support vector machine (SVM) is a classification technique that is included in a supervised learning class. When training, certain learning targets are needed. SVM finds optimal global solutions. SVM is a machine learning method that works on the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates two classes in input space. The way the SVM algorithm works is basically to find an optimal hyperplane that can separate the data precisely so that the data is divided according to the class. The optimal hyperplane can be searched by measuring the hyperplane margin and finding the maximum point between the two. The SVM algorithm will look for hyperplanes that have a maximum margin where the margin is the distance between the hyperplane and the closest data from each class. The closest data to the hyperplane is called support vector [4].

Logistic regression is one form of non-linear regression that has a discrete dependent variable and has a binomial distribution, while the independent variable can consist of continuous variables, discrete, dichotomous, or a combination thereof. Parameter estimates from logistic regression using the maximum likelihood method. The probability of each value of the dependent variable has its own value for each line.

Logistic regression is a regression model that is used when the response variable is qualitative [7]. Logit transformation is applied to logistic regression models,

$$\boldsymbol{Logit}\ (\boldsymbol{p}(\boldsymbol{x})) = \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{ln}\left[\frac{\boldsymbol{p}(\boldsymbol{x})}{\boldsymbol{1}-\boldsymbol{p}(\boldsymbol{x})}\right] = \boldsymbol{\beta_0} + \sum_{k=1}^{p} \boldsymbol{\beta_k x_k} \quad (2)$$

Logit transformation aims to create a linear function of its parameters. The g (x) function is linear to the parameters and has a range (-∞, ∞), depending on the range of X predictor variables.

To handle imbalanced datasets, the SMOTE (Synthetic Minority Over-sampling Technique) method was first introduced by Nithes V. Chawla [2]. SMOTE created an example of a synthetic minority class that operates in a feature space rather than a data space [15] [19]. This approach works by making replication of minority data. Replication is known as synthetic data (synthetic data). The generation of synthetic or artificial data is made to increase the number of minority classes to equal the majority class. The artificial or synthetic data is made based on the nearest neighbor (k-nearest neighbor) [1]. The synthesis sample is made by calculating the value of the difference (reduction) between the selected vectors attributes and neighboring vectors located close together. Then the subtraction value is multiplied by the random number between 0 and 1, and then added to the vector attribute value that was previously selected. This process describes the selection of points randomly on a segment line between two specific attributes [2] [19].

C. Named Entity Recognition

Named Entity Recognition (NER) is a subproblem of information extraction and involves structured and unstructured document processing and handles the problem of identification (detection) and type classification of pre-defined named entities, such as organizations, people, place names, temporal expressions, expressions numerical and currency, etc. [11] [12] [14]. NER is a fundamental task and is at the core of a natural language processing system (NLP). The NER task can also include describing descriptive information from the text about entities detected by filling out small-scale templates. For example, in the case of people, it might include extracting the title, position, nationality, gender, and other people's attributes. NER also involves the lemmatization (normalization) of the entity called, which is very important in highly inflective language [14].

Some approaches used in the NER include Rule-Based, Machine Learning Based which utilizes Hidden Markov Models, Maximum Entropy, Decision Trees, Support Vector Machines, Conditional Random Fields, and Hybrid approaches [11] [12].

ISBN 978-602-53524-0-9

Generous classifications for tweeting and identifying entities named in disaster tweets using Support Vector Machine (SVM) and TF IDF classification methods for weighting. Whereas for the introduction of entities named in disaster tweets using the BIO and Support Vector Machine (SVM) methods [3].

III. EXPERIMENT

A. Labeling

There were two corpus tweets needed, the corpus for the classification of tweets and corpus for Named Entity Recognition. So at the initial stage, the tweet data obtained were annotated manually by annotator using GATE. Each tweet was represented as one document, and annotations were conducted per tweet. So far the tweet was obtained from crawling from 2017 September 19 to February 20, 2018, as many as 2,685 tweets with various names of disasters. After annotation tweets consisted of 183 emergency tweets, 346 non-emergency tweets, and 2156 tweets that were not relevant to the disaster. Tweets obtained were used as 70% training data and 30% test data.

Figure 1. Example of Tweet Labeling

B. Preprocessing

At this stage tokenization was carried out to separate by token sentences, normalization to change non-standard words into standard words so that there was uniformity of words. Stop word removal was used to delete words that were general and not too important in information extraction, and the filtering process to select a complete tweet in accordance with the category. In addition, case folding was conducted to turn all text into lowercase letters. URL removal was conducted to eliminate the URL. Convert emoticons to change the emoticon symbol, and punctuation removal to delete all non-alphabet characters. After pre-processing and filtering, out of 2,685 tweets obtained a clean dataset of 1309 unique tweets.

C. Tweet classification

The features used for tweet classification are bag of words. Tweets are divided into three categories:

1) Emergency
2) Non-Emergency
3) Other

Only emergency and non-emergency tweets were used for the next process. Classification was conducted by comparing the Naïve Bayes method, Instance-Based Learning, Support Vector Machine, and Logistic Regression.

From 1309 disaster tweets, there were 89 emergency tweets, 168 non-emergency tweets, and 1052 irrelevant tweets. Data was divided into 70% training data and 30% test data. The SMOTE method was used to overcome imbalanced datasets.

TABLE I. TRAINING AND TEST DATA

Category	Training	Testing
Emergency	62	27
Non-Emergency	118	50
Other	736	316
Total	**916**	**393**

After using the SMOTE filter:

TABLE II. STRAINING AND TEST DATA AFTER SMOTE

Category	Training	Testing
Emergency	725	27
Non-Emergency	731	50
Other	736	316
Total	**2.192**	**393**

TABLE III. LIST OF WORD LEVEL FEATURES

Fiture Name	Description	*Feature Type*
Current word	What is the current word?	Nominal
Named Entity	What are named entities currently?	Nominal
NE before 1 word	What is named entity 1 word before this word?	Nominal
NE after 1 word	What is named entity 1 word after this word?	Nominal
before 1 word	What is named entity 1 word before this word?	Nominal
before 2 word	What are 2 words before this word?	Nominal
after 1 word	What are 1 words after this word?	Nominal
after 2 word	What are 2 words after this word?	Nominal

IV. RESULTS

The study was conducted by comparing the use of Skip Gram and SMOTE. The following are the results of using Skip Gram on the original data and data that has been SMOTE:

TABLE IV. RESULT OF CLASSIFICATION

No	*Classifier*	Data	SMOTE
1	***Naïve Bayes***	71.7557 %	72.7735 %
2	***SVM***	80.4071 %	80.4071 %
3	***IBK 1***	81.4249 %	81.1705 %
4	***IBK 3***	83.715 %	83.2061 %
5	***IBK 5***	82.1883 %	81.9338 %

ISBN 978-602-53524-0-9

No	*Classifier*	Data	SMOTE
6	***IBK 15***	83.9695 %	81.9338 %
7	***Logistic Regression***	74.3003 %	72.2646 %

Based on the research, the use of Skip-Gram can improve the accuracy results for the IBK classifier. While the use of skip grams with SMOTE filters results in an increase in accuracy in almost all classifiers. The best results for skip grams are obtained using IBK 15 with an accuracy value of 83.9695%.

Result of the named entity recognition, i-name class has good results because this class has specific characteristics. The name of a disaster that only consists of several types. Whereas i-location is usually preceded by certain lexical-lexical features such as lexical "on", and also the use of gazetteer which gives better results than before. As for the other class, most of them are dominated by words with very little frequency of occurrence.

TABLE V. RESULT OF NER

Kelas Named Entity	Recall	Prec	F-1
B-Name	0.5084	0.8695	0.6417
I-Name	0.8209	0.9277	0.8710
B-Location	0.4417	0.9024	0.5931
I-Location	0.7542	0.8804	0.8125
Other	0.9886	0.8916	0.9376
Average	0.7028	0.8943	0.7712

TABLE VI. RESULT OF GENERATING HASHTAG

Hashtag	Recall	Precission	F-1
#112Loc	0.3666	1	0.5365
#PublicRep	0.9850	0.7415	0.8461
#NameLoc	0.4857	0.8793	0.6257
Rata-rata	0.6124	0.8736	0.6695

Based on these results it was found that the F-Measure results for non-emergency hashtags were higher than others. This is because the accuracy of the classification of non-emergency classes is higher compared to emergency classes. Emergency classes are often classified into non-emergency classes.

The name of the disaster hashtag has a low F-measure because of the merger between entities named disasters and disaster locations. So, if one of the entities wrongly predicts the hashtag result will be wrong. More mistakes were seen in the introduction of disaster location entities with many location names, so that the system was not appropriate to determine the location of the disaster. This means that the system cannot handle multilabel problems correctly. The following are examples of disaster hashtag generation results:

TABLE VII. EXAMPLE RESULT OF GENERATING HASHTAG

Tweet	Result Hashtag	True Hashtag
Brebes arah jakarta kampung yg d sebelah kiri banjir gara2 kalinya meluap (atau tanggulnya ambrol ?) Sawah dan kampungnya kelelep	#PublicRep #banjirjakarta	#PublicRep #banjirBrebes
Dsn. dlopo desa kepel kec. Ngetos bahwa longsor susulan selebar 5 m , panjang sekitar 30 m yg mengakibatkan pralon ? https://t.co/HP00vYOjUv	#PublicRep #longsor	#112Ngetos #longsorNgetos
RT @infokaltim : Banjir Rendam 6 Desa di Kalimantan Timur , Perahu Karet Dibutuhkan - https://t.co/3gA2XQQtZf https://t.co/wmtaWyZnK0 #kutaibâ€¦	#112Kalimantan #BanjirKalimantan	#112Kalimantan #BanjirKalimantan
Banjir di Brebes akibat luapan Sungai Pemali akibatkan 2 orang meninggal dunia . Satu orang terbawa arus saat akan m ? https://t.co/0zLAAQWha3	#112Brebes #BanjirBrebes	#112Brebes #BanjirBrebes
RT @Uty_JKT48 : Ada gempa ??	#PublicRep #gempa	- -

V. CONCLUSION

Classification of categories of disaster tweets for Emergency, Non-emergency and Other was conducted using several classifiers. The use of Skip Gram were proved to improve accuracy. The best results were obtained using Instance-Based Learning k = 15, with an average accuracy of 83.9695%.

Identification of named entities was conducted using the Conditional Random Field model with 70.3% recall, 89.4% precision and 77.1% f-measure.

Automatic hashtag generation has good results with an average of 61.2% recall, 87.4% precision and 66.9% f-measure.

ISBN 978-602-53524-0-9

ACKNOWLEDGMENT

Thanks to the Ministry of Education and Culture for the assistance of research costs through the Social and Artists Activist Scholarship (BUPSS).

REFERENCES

[1] Barro, R. A., Sulvianti, I. D., & Afendi, F. M. (2013). PENERAPAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE. *Xplore, 1(1):e9(1-6)*.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002, Juni 1). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357. doi: https://doi.org/10.1613/jair.953

[3] Dermawan, R. (2016). Klasifikasi Tweet Dan Pengenalan Entitas Bernama Pada Tweet Bencana Dengan Support Vector Machine. Yogyakarta: Universitas Gadjah Mada.

[4] Ilyas, R., & Khodra, M. L. (2015, Juni). Ekstraksi Informasi 5W1H pada Berita Online Bahasa Indonesia. *Jurnal Cybermatika, 3 No. 1*.

[5] Godin, F., Slavkovikj, V., De Neve, W., Schrauwen, B., & Van de Walle, R. (2013, May). Using topic models for Twitter hashtag recommendation. *Proceedings of the 22nd International Conference on World Wide Web* (pp. 593-596). ACM.

[6] Gong, Y., & Zhang, Q. (2016). Hashtag Recommendation Using Attention-Based Convolutional Neural Network. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 2782-2788). IJCAI.

[7] Hosmer, D.W. and Lemeshow, S. (1989) Applied Logistic Regression. John Wiley & Sons, Inc., New York.

[8] Kywe, S. M., Hoang, T.-A., Lim, E. P., & Zhu, F. (2012). On Recommending Hashtags in Twitter Networks. *Social Informatics : 4th International Conference* (pp. 337 - 350). Lausanne, Switzerland: Research Collection School Of Information Systems.

[9] Li, J., Xu, H., He, X., Deng, J., & Sun, X. (2016). Tweet modeling with LSTM recurrent neural networks for hashtag recommendation. *Neural Networks (IJCNN), 2016 International Joint Conference on* (pp. 1570-1577). IEEE.

[10] Li, Y., Liu, T., Jiang, J., & Zhang, L. (2016). Hashtag recommendation with topical attention-based LSTM. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3019–3029). Osaka, Japan: Coling.

[11] Mansouri, A., Affendey, L., & Mamat, A. (2008). Named Entity Recognition Approaches. *IJCSNS International Journal of Computer Science and Network Security, 8(2)*, pp. 339-344.

[12] Mohit, B. (2014). Named Entity Recognition. In I. Zitouni (Ed.), *Natural Language Processing of Semitic Languages* (pp. 221-245). Springer Berlin Heidelberg. doi:https://doi.org/10.1007/978-3-642-45358-8_7

[13] OCHA, U. (2014). *Hashtag Standars for Emergencies*. OCHA Policy and Studies series. United Nations Office for the Coordination of Humanitarian Affairs.

[14] Piskorski, J., & Yangarber, R. (2013). Information Extraction: Past, Present and Future. In T. S. Poibeau, *Multi-source, Multilingual Information Extraction and Summarization. Theory and Applications of Natural Language Processing.* (pp. 23-49). Springer, Berlin, Heidelberg. doi:https://doi.org/10.1007/978-3-642-28569-1_2

[15] Putri, S. A., & Wahono, R. S. (2015, December). Integrasi SMOTE dan Information Gain pada Naive Bayes untuk Prediksi Cacat Software. *Journal of Software Engineering, 1, No. 2*.

[16] Rasywir, E., & Purwarianti, A. (2015, Desember). Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin. *Jurnal Cybermatika, 3 NO. 2*.

[17] Tomar, A., Godin, F., Vandersmissen, B., De Neve, W., & Van de Walle, R. (2014). Towards Twitter hashtag recommendation using distributed word representations and a deep feed forward neural network. *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on* (pp. 362-368). New Delhi, India : IEEE. doi: 10.1109/ICACCI.2014.6968557

[18] Tran, K., & Popowich, F. (2016). Automatic Tweet Generation From Traffic Incident Data. *WebNLG 2016, 59*.

[19] Ulhaq, Z., & Adji, T. B. (2017, Juli 27). Integrasi Synthetic Minority Over-Sampling Technique (SMOTE) dengan Correlated Naïve Bayes Classifier (C-NBC) pada Klasifikasi Siswa Berkesulitan Belajar. *CITEE 2017*, 201-205.

[20] Xiao, F., Noro, T., & Tokuda, T. (2012, July). News-Topic Oriented Hashtag Recommendation in Twitter Based on Characteristic Co-occurrence Word Detection. *ICWE*, 16-30.

ISBN 978-602-53524-0-9