International Journal of Computational Linguistics and Applications vol. 7, no. 2, 2016, pp. 129–143
Received 05/02/2016, accepted 07/03/2016, final 16/09/2016
ISSN 0976-0962, http://ijcla.bahripublications.com

Short Text Classification on Complaint Documents

SHIRLEY ANUGRAH HAYATI, ALFAN FARIZKI WICAKSONO, AND MIRNA ADRIANI

Universitas Indonesia, Indonesia

ABSTRACT

Indonesian government has developed a system for citizens to voice their aspirations and complaints, which are then stored in the form of short documents. Unfortunately, the existing system employs human annotators to manually categorize the short documents, which is very expensive and time-consuming. As a result, automatically classifying the short documents into their correct topics will reduce manual works and obviously increase the efficiency of the task itself. In this paper, we propose several approaches to automatically classify these short documents using various features, such as unigrams, bigrams, and their combination. Moreover, we also demonstrate the use of information gain and Latent Dirichlet Allocation (LDA) for selecting discriminative features.

1 INTRODUCTION

Short Message Service (SMS) and the Internet have become important and powerful communication media for people. Some countries take advantage of this advancement in information technology to develop website as a medium for their citizens to give feedback or report problems related to government policies. Public feedback

and suggestions play an important role to improve public services, resolve national problems, and support open government.

We partnered with Indonesian Government which provides online aspiration and complaint system, namely LAPOR![1], for Indonesian citizens to communicate with the government. This system allows Indonesian citizens to voice their aspirations or problems by directly writing in the LAPOR! website, sending text messages to LAPOR!, or using LAPOR! mobile application. LAPOR! itself stores their complaints in the form of short electronic documents.

Currently, LAPOR! system employs several human annotators to categorize new complaint documents. As the incoming documents arrive in a massive number, manually categorizing a large number of documents will be very expensive and time-consuming. Furthermore, manual classification is also prone to human inconsistency. Therefore, an automatic text categorization which can help human annotators identify the right topic of those documents becomes necessary to help reduce the expensive and tedious manual work, which is still done today.

In recent years, many approaches on text classification have been proposed by several researchers [1–4]. Most of them focused on automatic text classification for English documents. Evaluation of these approaches for Indonesian documents has been limited, especially for Indonesian short documents.

Furthermore, short text is different from long text, especially because of its shortness. As a result, classifying short text presents many challenges. Short documents often do not provide enough word co-occurrence for learning. In addition to that, they are often misspelled and ambiguous. Some of them also contain inconsistency in terms of abbreviations and informal words.

New techniques for short text classification have been introduced in the last few years. In order to overcome data sparseness and shortness, Phan et al. [5] used external data to obtain more knowledge for their classifier. Considering the significance of se-

[1] https://www.lapor.go.id/

mantic terms, Wang et al. [6] proposed a collection of representative terms and a new term weighting model to improve the accuracy of short text classification. Both of approaches proposed by Phan et al. [5] and Wang et al. [6] were detecting hidden topic using Latent Dirichlet Allocation (LDA) model. Furthermore, Wang et al. [7] defined a new framework, namely *bag-of-concepts*, to group words with similar domain for short text classification. None of them addressed short text classification in Indonesian language, especially for government short complaint documents mainly from text messages.

The main contribution of this paper is that we are the first to research on Indonesian short text classification from LAPOR! system. We propose several different features for building automatic Indonesian short document categorization for government online aspiration and complaint system. We also adopt Wang's method in [6] in utilizing LDA for selecting words as features.

2 Related Work

Techniques for classifying long documents may fail when they are applied for classifying short documents since short documents, especially those which come from text messages (SMS), are usually noisier and contain less topic-related words. Many approaches to overcome these challenges have been addressed in the previous research.

Gabrilovich and Markovitch [8] used additional knowledge from Wikipedia to expand bag of words and generate new features. Their work showed improvement in terms of classification accuracy. Furthermore, Zelikovitz [9] combined training data with unlabeled testing data and applied Latent Semantic Indexing (LSI) for short text classification. However, both approaches relied heavily on using external data to improve accuracy. Their approaches can not be directly applied on Indonesia data since Indonesian language itself is still lack of resources and tools for language technologies.

In the work conducted by Wang et al. [6], LDA model was used to extract 10 hidden topics from each category and collect top

20 semantic-oriented terms with maximum probability from each LDA-generated topic. They did not consider terms which appear in different topics. Wang et al. [6] then proposed a calculation method to measure the contribution of category (COC) of a specific representative term as the ratio of its information gain to the sum of information gain of all representative terms. Then, they defined a new weighting method for a feature value as term frequency of a strong term added by COC of that term. They added these new terms and the weights to bag-of-words features. In this research, we also use LDA for feature extraction. Unlike the method proposed by Wang et al. [6], we choose the number of LDA topic to be the same with the number of real existing topics in hope of correspondence of this topic. We give a threshold probability score in extracting words as features and uses information gain score to limit the number of unigram features to reduce the computational complexity.

Limited research has been conducted for short document classification in Indonesian language, especially for government complaint system. The closest research to ours were conducted by Laksana and Purwarianti [10] as well as Fauzan and Khodra [11]. Laksana and Purwarianti [10] performed multi-label classification from microblog in Indonesian language for Bandung complaint management system. Complaints from Bandung citizens' tweets are classified to relevant government agencies. Fauzan and Khodra [11] also worked on multi-label authority classification for Bandung complaint management system using learning to rank framework to determine priorities between agencies. Our work is different from their works. Both of them focused only on Bandung government complaint system whose complaints are from local system and Twitter. On the other hand, we work on Indonesian government feedback and aspiration documents which are delivered by text messages (SMS), mobile application, or website, but never tweets. In fact, we address a different task. Instead of classifying each text to multiple Bandung government agencies [10, 11], we study on categorizing each short document to a single topic.

3 Proposed Method

3.1 *Task Definition*

We define short text classification as a task of assigning short documents to a predefined class based on their contents [12]. We are given a set of short documents $X = \{d_1, d_2, ..., d_m\}$ and a fixed set of classes $C = \{c_1, c_2, ..., c_n\}$. A training set D which consists of labeled documents $\langle d_i, c_j \rangle$ is used to generate a classification function $\gamma: X \rightarrow C$ that will accurately classify a unlabeled test document from X. The case in which a document d_i assigns to exactly one class is called the single-label text classification [13]. The case in which a document d_i assigns to any number from *1* to n classes is called the multilabel text classification. In our research, we deal with the single-label case.

3.2 *Features*

Since our work is a supervised learning problem, documents need to be converted from a collection of words into feature vectors which are more suitable for the learning algorithm and classification task. A text document is represented as a vector of binary features with boolean representation. A feature will have the value of 1 if it appears in the document, and 0 if otherwise happens. In this paper, we propose several features: combination of unigrams, bigrams, and LDA-based unigrams. Information gain scoring is applied to select features with the best features, that is the most relevant features. Hence, the dimensional space can be reduced.

Unigram Unigram is a 1-gram sequence of word from each document. In unigram approach, a document is represented as bag-of-words (BOW), that is, features are identified with words occuring in the document. An example of unigrams for the Indonesian sentence *"saya melakukan penelitian dalam klasifikasi teks"* ("I do a research in text classification") is *"saya"* ("I"), *"melakukan"* ("do"), *"penelitian"* ("a research"), *"dalam"* ("in"), *"klasifikasi"*

("classification"), *"teks"* ("text"). Furthermore, Bekkerman [14] mentioned that in spite of its simplicity, bag-of-words approach has shown very powerful for text classification.

BIGRAM Unigram feature is unable to capture significant information on a particular phrases. For example, a phrase, like *"klasifikasi teks"* ("text classification"), will be broken down as separated words: *"klasifikasi"* ("classification") and *"teks"* ("text"). As a result, its real meaning as a phrase will be lost. As a 2-gram sequence of word, bigram feature can handle phrases containing two words. Tan et al. [15] argued that previous works on using phrases as features degraded the performance of text classification. Therefore, we propose to incorporate bigrams and unigrams for the features. Unlike the previous study, our experiment shows that the combination of bigrams and unigrams perform the best in our case.

LDA-BASED UNIGRAM LDA is a generative probabilistic model which represent documents as random mixtures of hidden topics[16]. Each LDA-generated topic contains bag of words and word distribution. As a generative model, LDA works as follows. First, we assume that there are *K* topic distributions for our data. For each document, LDA randomly chooses a distribution over topics and select one word probabilistically given the topic. The topic distribution and word distribution are modelled as multinomial distribution with parameters following the dirichlet distributions. The parameters of those two dirichlet random variables are denoted as α and β, respectively. For each word in the document, LDA assigns a specific topic z_{ij} for the j^{th} word in document *i*.

We employed Latent Dirichlet Allocation (LDA) technique to extract words with high probability in each topic as the features. Given 9 sets of manually labeled data $D_i = \{doc_{i1}, doc_{i2}, ..., doc_{in}\}$, where $i \in \{1, 2, ..., 9\}$ and doc_{ij} is the j^{th} document of class *i*, we specified 9 LDA topic distributions in hope that each topic will correspond with the real class. Words related to each LDA topic are ranked according to term frequency. We set 0.025% as

the minimum frequency percentage threshold in selecting words as features. For example, these are some LDA-extracted terms for **Health** category: "*rs*" ("hospital"), "*sakit*" ("sick"), "*dokter*" ("doctor"), and "*obat*" ("medicine").

INFORMATION GAIN Feature selection is applied to all features in order to reduce noise as well as the computational complexity of learning. We first removed features with very few word frequency counts. After that, we selected highest-scored features using information gain measure. To calculate information gain of a feature, we need to understand the notion of *entropy*. Given *S* as a set of data and *n* different target feature values, *Entropy(S)* is defined as

$$Entropy(S) = H(S) = \sum_{i=1}^{n} -p_i \; log_2 \; p_i$$

The information gain of a feature *F* related to *S* is defined as

$$InfoGain(S, F) = H(S) - H(S \mid F)$$

where

$$H(S \mid F) = \sum_{v \in values(F)} \frac{|S_v|}{|S|} Entropy(S_v)$$

The information gain itself represents how informative a particular feature is. In other words, a feature with high score of information gain serves as a discriminative feature.

4 EXPERIMENT

4.1 *Data Collection and Experimental Settings*

We used corpus from LAPOR! database. It consists of 50000 short documents about public aspiration and feedback. LAPOR! management team has defined more than 170 hierarchical categories[2]. However, based on our observation, the documents are classified to only 57 categories. The number of documents in each category

[2] In text classification, *categories* are often used to refer to the same entity as *classes* or *topics*. In this paper, we use them interchangeably.

vary very widely. 11% of the topics have more than 1000 documents assigned, whereas 37% of the topics have less than 10 documents assigned. LAPOR! management team also describe that these topics can be changed or removed at anytime, except for the root topics, so they request that the system to detect only root topics for the documents.

First, we excluded 5 topics because they are no longer in use. Then we assigned documents with non-root topics to their root topics following the request of LAPOR! management team. After this new assignment, we observed that the data distribution was still very imbalanced. Therefore, we excluded 4 more topics with very low F1-measure.

After removing stopwords and other noises, we filtered out duplicate documents and documents which contain less than 3 words. The corpus size was reduced to about 17000 short documents with 9 topics. Table 1 shows the detail of our data collection.

Table 1. The statistics of document and topic distribution

No	Topic	#Documents
1	Reformation and Governance	4432
2	Education	2466
3	Health	1578
4	Infrastructure	4264
5	Energy and Natural Resources	578
6	Environment and Disaster Management	1998
7	Politics, Law, and Security	659
8	Economics	339
9	People's Welfare	703
	Total	**17017**

We utilized free machine learning tool Weka[3] for classification learning with 5-fold cross validation. For extracting terms using LDA model, we used Mallet[4].

[3] http://www.cs.waikato.ac.nz/ml/weka/
[4] http://mallet.cs.umass.edu/

We choose Naive Bayes and Support Vector Machine (SVM) learning methods for our experiment due to the following reasons. Naive Bayes is very fast and efficient in terms of storage usage. We must consider efficiency due to the future practical usage, that is, we plan to implement our research results on the real LAPOR! system. Meanwhile, the popular SVM is robust and powerful for text classification, yet a huge amount of data slows down its performance.

4.2 *Experiment Scenario*

Four sets of experiments were conducted. We selected features with the highest information gain for all sets. Our first experiment scenario used unigram features. The number of features varies from 100, 200, 300, ..., 900, 1000, 1500, and 2000. For the second experiment, we used LDA-based unigram as our features. The number of features used follows the first experiment. In our third experiment, we added bigram phrases as new features to the basic unigram features. We varied the composition of both unigram and bigram features. In the last experiment, we combined LDA-based unigram terms with bigram as our features. For all unigram experiments, we used Naive Bayes and SVM learning algorithm. Meanwhile, for the third and fourth experiment scenario, we employed Naive Bayes for all experiments. SVM was only used for unigram-bigram experiments with the best result.

4.3 *Results*

Table 2 shows accuracy result of experiment 1 and experiment 2. Trained by using Naive Bayes learning algorithm, LDA-based unigram model performance is quite comparable to basic unigram model when the number of features ranges from 200 to 1000. However, with 2000 features, LDA-Based unigram accuracy drops due to the information gain scoring, i.e. we observed that the 1378rd word (and below) have zero information gain score.

Table 2. Accuracy (%) Result of Experiment 1 and Experiment 2

Number of Features	Basic Unigram		LDA-Based Unigram	
	NB	SVM	NB	SVM
100	70.13	71.71	67.95	71.98
200	73.52	77.39	73.62	77.10
300	75.48	78.89	75.75	78.97
400	76.68	79.54	76.69	79.77
500	77.74	80.10	77.42	80.34
600	78.02	80.60	77.96	80.44
700	78.39	80.64	78.42	80.73
800	78.85	80.54	78.61	80.62
900	79.01	**80.65**	79.02	80.76
1000	79.23	80.61	79.26	**80.89**
1500	80.21	80.35	**79.64**	80.40
2000	**80.96**	80.42	79.37	80.00

When using SVM, the LDA-based unigram model (1000 features) surpasses basic unigram model (900 features). However, overall it turns out that basic unigram model trained with Naive Bayes achieve the highest accuracy. Thus, it becomes our baseline.

After further inspection, we note that the removal of features which appear in different topics in LDA-based unigram model does not provide enough discriminative words for our data. We hypothesize that the accuracy can be improved by adding bigrams because unigrams fail to capture phrase with important meaning. For example, the phrase "***bahan bakar***" ("fuel") in the following document is a discriminative phrase to classify this document to **Energy and Natural Resources** category.

"... mengadu mengenai persoalan berlarut larutnya kelangkaan ***bahan bakar*** *gas elpiji 3 kg ..."*

As mentioned before, unigram model is incapable of capturing the true meaning of ***"bahan bakar"***. Unigram separates the phrase to be ***"bahan"*** ("material") and ***"bakar"*** ("burn") and misleads the learning algorithm to classify the document to **Environment and Disaster Management** class.

As we add bigrams, the accuracy for experiments which employed 1500 or more unigrams combined with 2000 bigrams increases compared to the baseline. Table 3 presents the accuracy result of unigram-bigram experiment compared with the accuracy result of LDA-based unigram-bigram experiment. The term ”basic” refers to employing basic unigram combined with bigram while the ”LDA-based” means employing LDA-based unigram with bigram. The highest accuracy is reached when using 2000 basic unigrams and 2000 bigrams with 81.69% accuracy. For the LDA-based unigram and bigram model, the highest accuracy is 81.02% using 1500 LDA-based unigram features and 2000 bigram features. We can see that when using 1000 unigrams and 1000 to 2000 bigrams, the accuracy of LDA-based unigram-bigram model is slightly better than the basic unigram-bigram model. This shows that LDA can be an alternative in extracting unigram features for short text classification.

Table 3. Accuracy Result of Unigram-Bigram Models using Naive Bayes and Comparison to Baseline

Baseline: 80.96%			
Number of Features		**Accuracy Difference (%)**	
Unigram	Bigram	Basic	LDA-Based
500	500	77.71 (−3.25)	77.55 (−3.41)
	1000	78.57 (−2.39)	78.79 (−2.17)
	1500	79.07 (−1.89)	79.34 (−1.62)
	2000	79.76 (−1.20)	79.80 (−1.16)
1000	500	79.14 (−1.82)	78.93 (−2.03)
	1000	79.64 (−1.32)	79.74 (−1.22)
	1500	80.04 (−0.92)	80.08 (−0.88)
	2000	80.49 (−0.47)	80.59 (−0.37)
1500	500	79.86 (−1.10)	79.6 (−1.36)
	1000	80.33 (−0.63)	80.18 (−0.78)
	1500	80.68 (−0.28)	80.50 (−0.46)
	2000	81.03 (+0.07)	**81.02** (+0.06)
2000	500	80.49 (−0.47)	79.41 (−1.55)
	1000	80.95 (−0.01)	79.86 (−1.10)
	1500	81.32 (+0.36)	80.27 (−0.69)
	2000	**81.69** (+0.73)	80.74 (−0.22)

Details of precision, recall, and F1-measure of classes from the experiment with the best accuracy result is presented in Table 4. **Health** topic has the highest precision (95.3 %), meanwhile **Education** has the highest recall (94.6%) and F1-measure (93.9 %). **Economics** has the lowest precision (53.2%), recall (59.6%), and F1-measure (56.2%). Upon further inspection of the gold standard, we note that there are many human inconsistencies when classifying these documents to **Economics** and many of those documents are ambiguous. An example of ambiguous document is a text which discusses about legal building be classified. In the gold standard, this kind of text is categorized into **Economics** topic as well as **Reformation and Governance**.

Table 4. Details of Precision, Recall, and F1-measure

Topic	Precision	Recall	F1
Reformation & Governance	0.854	0.712	0.777
Education	0.932	0.946	0.939
Health	0.953	0.838	0.892
Infrastructure	0.828	0.85	0.839
Energy and Natural Resources	0.856	0.957	0.904
Environment	0.648	0.877	0.746
Politics, Law, and Security	0.648	0.727	0.685
Economics	0.532	0.596	0.562
People's Welfare	0.814	0.683	0.742
Weighted Average	0.828	0.817	0.818

We also trained the best unigram-bigram scenarios (2000 unigram + 200 bigram and 1500 LDA-based unigram + 2000 bigram) using SVM to compare the Naive Bayes model with SVM model. We find that Naive Bayes outperforms SVM when the number of features is huge, at least in our case as shown in Table 5.

5 Conclusion and Future Work

In this paper, we have proposed several methods to classify Indonesian short documents for the LAPOR! system. Based on our

Table 5. Unigram and Bigram Models Using SVM

Feature	**Accuracy** (%)
2000 Basic Unigram + 2000 Bigram	80.89
1500 LDA-Based Unigram + 2000 Bigram	81.16

experiments, the best scenario was obtained when both unigrams and bigrams were used as our features (with accuracy of 81.69%). In addition to that, we have shown that the performance of LDA-based unigram-bigram model is comparable with the performance of basic unigram-bigram model. Our work found that the information of bigram is important to identify the class in which a document belongs to. For our problem, our proposed method has been proven to be more effective than the baseline method (unigram) since many phrases in the short documents have strong characteristics of their class. We realize that our work still has limitation in terms of data collection. We really need to increase the size of our data as well as make it more balanced. In the future, we will further find pattern for capturing co-occurring terms related to a topic.

ACKNOWLEDGMENTS The authors would like to thank Ferdy Alfarizka from Kantor Staf Presiden Indonesia (Indonesian Presidential Staff Office) for assistance in providing LAPOR! data and category information.

REFERENCES

1. Kalt, T.: A new probabilistic model of text classification and retrieval title2:. Technical report, Amherst, MA, USA (1998)
2. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. In: IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION, AAAI Press (1998) 41–48
3. Ogura, Y., Kobayashi, I.: Text classification based on the latent topics of important sentences extracted by the pagerank algorithm. In: ACL (Student Research Workshop), The Association for Computer Linguistics (2013) 46–51

4. Johnson, R., Zhang, T.: Effective use of word order for text categorization with convolutional neural networks. In: NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015. (2015) 103–112
5. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th International Conference on World Wide Web. WWW '08, New York, NY, USA, ACM (2008) 91–100
6. Wang, B.k., Huang, Y.f., Yang, W.x., Li, X.: Short text classification based on strong feature thesaurus. Journal of Zhejiang University SCIENCE C **13**(9) (2012) 649–659
7. Wang, F., Wang, Z., Li, Z., Wen, J.R.: Concept-based short text classification and ranking. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. CIKM '14, New York, NY, USA, ACM (2014) 1069–1078
8. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In: Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2. AAAI'06, AAAI Press (2006) 1301–1306
9. Zelikovitz, S.: Transductive lsi for short text classification problems. In Barr, V., Markov, Z., eds.: FLAIRS Conference, AAAI Press (2004) 556–561
10. Laksana, J., Purwarianti, A.: Indonesian twitter text authority classification for government in bandung. In: Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014 International Conference of. (Aug 2014) 129–134
11. Fauzan, A., Khodra, M.: Automatic multilabel categorization using learning to rank framework for complaint text on bandung government. In: Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014 International Conference of. (Aug 2014) 28–33
12. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
13. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1) (March 2002) 1–47
14. Bekkerman, R., Allan, J.: Using bigrams in text categorization (2003)
15. Tan, C.M., Wang, Y.F., Lee, C.D.: The use of bigrams to enhance text categorization. Inf. Process. Manage. **38**(4) (July 2002) 529–546
16. Blei, D.M., Ng, A.Y., Jordan, M.I., Lafferty, J.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 2003

SHIRLEY ANUGRAH HAYATI
FACULTY OF COMPUTER SCIENCE,
UNIVERSITAS INDONESIA,
DEPOK, WEST JAVA, INDONESIA
E-MAIL: <SHIRLEY.ANUGRAH@UI.AC.ID>

ALFAN FARIZKI WICAKSONO
FACULTY OF COMPUTER SCIENCE,
UNIVERSITAS INDONESIA,
DEPOK, WEST JAVA, INDONESIA
E-MAIL: <ALFAN@CS.UI.AC.ID>

MIRNA ADRIANI
FACULTY OF COMPUTER SCIENCE,
UNIVERSITAS INDONESIA,
DEPOK, WEST JAVA, INDONESIA
E-MAIL: <MIRNA@CS.UI.AC.ID>