Topic Identification and Categorization of Public Information in Community-Based Social Media

RP Kusumawardani and MH Basri
Department of Information Systems, Institut Teknologi Sepuluh Nopember (ITS)
Kampus ITS, Jl. Raya ITS, Keputih – Sukolilo
Surabaya, Indonesia

E-mail: renny.pradina@gmail.com

Abstract. This paper presents a work on a semi-supervised method for topic identification and classification of short texts in the social media, and its application on tweets containing dialogues in a large community of dwellers in a city, written mostly in Indonesian. These dialogues comprise a wealth of information about the city, shared in real-time. We found that despite the high irregularity of the language used, and the scarcity of suitable linguistic resources, a meaningful identification of topics could be performed by clustering the tweets using the K-Means algorithm. The resulting clusters are found to be robust enough to be the basis of a classification. On three grouping schemes derived from the clusters, we get accuracy of 95.52%, 95.51%, and 96.7 using linear SVMs, reflecting the applicability of applying this method for generating topic identification and classification on such data.

1. Introduction

An implication of the proliferation of the social media is its increasingly pertinent role in the society. Nowadays, not only is it used simply as a medium of self-expression and communicating with the members of one’s circle of family and friends as it traditionally is, the social media has roles as divergent as a vehicle for the propagation of news, opinion, and lifestyles, coordinating efforts in disaster relief, to acting as means for organizing social and political uprisings.

This breadth of role of the social media is basically enabled due to its ability in assisting people in doing two things: to seek for information, and to share it. Due to its real-time nature and expansive network of members, people turn to social media to get and share information. As a result, the dialogues exchanged in the social media contain a wealth of information that is often the most up-to-date compared to more traditional information outlets.

However, the timeline format of the social media is often not the best for conveying and enabling its audience to get a better insight of the information content. In the case of Twitter as studied in this work, in order to get a thorough understanding on a topic, one must read all tweets in the stream of the account or search term of his interest, while it is very possible that not all of these are relevant, or that the tweet containing an important information occurring just before or shortly after the length of the stream that the person read.

Therefore, a further processing and organization of these social media posts is necessary. However, especially for Twitter, users often adopt linguistic styles that are not grammatically correct – rife with misspellings or deliberate omission of characters or words, and use of incorrect structures due to the
restriction of only 140 characters allowed per tweet, and the informal nature of communication in Twitter. For these reasons, natural language processing techniques developed under the assumption of some regularity in the corpus language become unsuitable.

A method with minimal reliance to linguistic resources is thus the solution for this problem. The aim of this paper is to explore the use of such method by combining clustering for topic identification and classification techniques for organizing the tweets. In between these two stages, we perform an analysis to determine whether the clustering results in groups of tweets with meaningful content themes. To see if the groups are robust and retrievable, we then use these groups to generate labels and subsequently use these to create labeled data for the classification process.

In this paper, we are interested in applying such method to a dataset of tweets comprising dialogues in a community of city dwellers, sharing real-time information on topics such as traffic condition, latest news, and events in the city. Specifically, we use a set of tweets collected from a large community of citizens in the city of Surabaya, written mostly in Bahasa Indonesia. Using K-Means for the clustering, we develop three labeling schemes which consequently fed into Linear SVMs, resulting in accuracy of 95.52%, 95.51%, and 96.7%, suggesting that this minimal language-specific resource method could be applied successfully to organize information in public tweets, especially with regard to a specific community and regardless of the language. We also found that despite the high language irregularity, a relatively small number of tokens can be used as cue to the category of the tweets.

2. Related Work

Twitter as a means for sharing information has been widely studied from a diverge point of view; for example, for sharing health-related information [1, 2], its use in catastrophic disasters [3], and its roles in social unrest [4], among others. Its use as a probe of activities in an urban context has been discussed, e.g. by [5]. [6] discuss the use of Twitter for information sharing by and related to the police department of a big city. [7] works in the domain of e-Government, extracting complaints on public services.

Many of the previous work is related to the identification of generally trending topics in Twitter, for example works by [8, 9, 10, 11]. Others are aimed at specific topics, for example traffic information. [12] studied the extraction of traffic information in tweets in the Thailand language, with the first stage of the system filtering on words such as “traffic congestion” and “accident”. Similar keyword-based filtering approach was taken by [13] for Japanese tweets. [14, 15, 16] uses hashtags or twitter accounts oriented for sharing traffic information as their data source, extracting traffic-related information using various methods, working on tweets in Bahasa Indonesia.

However, we note that in the case of urban dialogues in Twitter, the public information contained not only trending topics, but also seems to consist of multiple themes, which is quite stable from time to time. In this work we aim to investigate what these themes are, and whether they could be the basis of a classification system, which could then be used to organize these information, ultimately assisting people in getting the information relevant to them in real-time.

3. Data

The data used in this research is a five months’ worth of tweets either mentioning or by the accounts @e100ss and @SapawargaSby. The first is the account of Suara Surabaya Radio, a radio station in Surabaya with a strong community actively engaging in the sharing of information on the city of Surabaya and its surrounding regions. The second account is the official account of the Government of the City of Surabaya. By using these two accounts as seeds, [17] collected a total of 30,339 messages, spanning from the beginning of September 2015 until early February 2016. Table 1 shows some examples of these tweets.
Table 1. Example tweets

| No. | Message |
|-----|---------|
| 1.  | @e100ss mohon pihak terkait mengenai jalan mastrip yang banyak berlubang dan sangat berbahaya bagi roda dua pada saat musim hujan. |
| 2.  | @e100ss jln.girilaya ke arah dukuh kupang macet poooll...silahkan ambil jalan lain |
| 3.  | @e100ss agak kecewa sekali terhadap pegawai kecamatan thdp pelayanan.mengurus surat pengantar e-ktp |
| 4.  | RT @satpolppbsby: Tim Odong Odong melepas baner yg salah penempatan di jl. Ngagel Jaya Selatan @e100ss @SapawargaSby |

None of the messages in Table 1 can be considered to contain entirely correct Bahasa Indonesia. Each message contains multiple syntactic errors and variations in spelling. Especially, many abbreviations are employed on the basis of the assumed understanding of the reader. For example, the word ‘jalan’ – road – correctly spelled in tweet no. 1 is written as ‘jln’ in tweets no. 2, and as ‘jl.’ in tweet no. 4. Only in these four tweets we can find ‘silakan’ – please – written as ‘silahkan’, ‘terhadap’ – to – written as ‘thdp’, and ‘banner’ – banner – not even an Indonesian word and written as ‘baner’. We can also find ‘poooll’ – not a word, just an emphasis, and employing repetition of characters, where there should have been none. Consequently, the number of token types become very large. More importantly, since many tokens appears very rarely or even only once, these could cause failure in finding a good generalization in the classification stage of this method, resulting in overfitting.

4. Token Variability and Minimum and Maximum Document Frequency

Tokenizing all 30,339 tweets in the dataset results in 28,201 distinct token types with a total number of tokens of 371,182. This results in a comparable ratio of distinct token type to number of tokens of 0.076 when compared to other corpora of Bahasa Indonesia. For example, the ratio of distinct token type to number of token in the Indonesian classic literature Sengsara Membawa Nikmat [18] is 0.09. It should be remembered, however, that considering that due to its shortness, the language in tweets are usually quite simple, such high variation of token types is more likely due to artefacts such as variations in spellings, both inadvertently or deliberately to reduce the number of characters necessary, rather than being actually informative. Therefore, it is important to choose boundaries for the minimum document frequency appearance of tokens. If a token occurs only rarely, then it would not be something that reflects a category. However, if a tweet contains only rare tokens, it would have an empty representation. Obviously, the number of ‘empty’ tweets need to be limited. Using unigram + bigram as features, a min_df value of 20 was chosen, reducing the number of terms to 2052, while retaining 95% of the tweets; upon closer inspection, tweets with empty representation typically consist of hyperlinks, which are unique. Conversely, we also might need to set a maximum limit for document frequency, since if a token appears in too many tweets, then we would not be able to infer the topic from it, thus rendering it useless in the topic identification in the clustering stage. However, the dataset shows that the highest document frequency for any token in the data is 3800, less than 13% of the number of documents, thus in this work we do not set a limit on maximum document frequency.

5. Topic Identification based on Clustering

The remaining 28,837 tweets are then clustered using the unigram and bigrams as features. Each feature is represented by its Tf-Idf value, the combination of term frequency (tf) and inverse document frequency (idf) [19, 20]. The tweets are primarily in Indonesian, but English words are occasionally also used. Therefore, the stopwords used are a mix of Indonesian [21] and English from the NLTK [22] used in this work. For clustering the tweets, we use the K-means algorithm [23], a very popular algorithm for clustering due its speed and simplicity [24, 25]. Basically, it has a single parameter to set: k, the number of
clusters to find. Using the elbow method [26] we plot the total sum of squared error (SSE) of the clustering against the k value used to create the clusters. However, as shown in Figure 1, the value of SSE is consistently gradually decreasing with the increase of k, thus indicating that the data does not have a natural set of grouping.

![Figure 1. SSE of the clusters on different values of k](image)

Therefore, we take the approach of creating a large number of clusters, then based on the most frequent words and the semantic content of each cluster, group together clusters that are deemed to be most similar, in accordance the merge-two-clusters post-processing approach of [27]. An initial k value of 45 was chosen and the result of topic analysis of these clusters is shown in Table 2.

**Table 2. Resulting clusters**

| Clust # | Topic                  | # of members | Clust # | Topic                  | # of members | Clust # | Topic                  | # of members |
|---------|------------------------|--------------|---------|------------------------|--------------|---------|------------------------|--------------|
| 1       | traffic                | 362          | 16      | (spam)                | 138          | 31      | traffic jam            | 339          |
| 2       | traffic light          | 798          | 17      | demonstration         | 276          | 32      | loss                   | 335          |
| 3       | electricity            | 1041         | 18      | to Surabaya           | 353          | 33      | car                    | 547          |
| 4       | toll road              | 916          | 19      | weather               | 780          | 34      | Sidoarjo               | 532          |
| 5       | weather                | 230          | 20      | water                 | 478          | 35      | clear roads            | 600          |
| 6       | greetings              | 409          | 21      | clear roads           | 739          | 36      | accident               | 256          |
| 7       | Ahmad Yani             | 184          | 22      | electricity           | 694          | 37      | police                 | 370          |
| 8       | Surabaya               | 7210         | 23      | (spam)                | 143          | 38      | weather                | 198          |
| 9       | Malang                 | 821          | 24      | lane                  | 420          | 39      | truck                  | 708          |
| 10      | electricity            | 446          | 25      | government performance| 433          | 40      | clear roads            | 194          |
| 11      | crowded                | 962          | 26      | fire                  | 314          | 41      | congested              | 564          |
| 12      | weather                | 374          | 27      | bridge                | 231          | 42      | child issues           | 218          |
| 13      | motorcycle             | 460          | 28      | Krian                 | 212          | 43      | weather                | 1268         |
| 14      | congested              | 520          | 29      | Gresik                | 315          | 44      | economy                | 171          |
| 15      | weather                | 394          | 30      | traffic jam           | 1438         | 45      | toll road              | 386          |

Note in Table 2 that some clusters have the same topic. We merge these and arrive at 31 topics. Looking at the content of the tweets belonging to each topic, it is obvious that some are the subtopics of larger topics. Therefore, we create three-layer hierarchy to group these topics from the most general to the more specific. This hierarchy is shown in Figure 2.

Figure 2 shows that a major theme of the dialogues around @e100ss and @SapawargaSby is about traffic. This is indeed the case, especially for @e100ss, since this is the kind of information most often shared by the community built around this account. In the topic of traffic, people may report about
traffic condition in a particular area; whether there is traffic jam, congestion, or clear roads. @e100ss and @SapawargaSby are also often used by the public to voice their complaints on public services, especially on electricity and water. People also reports on general news and events, such as demonstrations, fire, and regional news.

6. Classifying the Tweets

The topics identified in Figure 2 gives us an idea of the types of public information shared in the tweets. We use the clustering results to create a dataset for training and testing classifiers based on the topics and topic hierarchy. This dataset is created by attaching labels to each tweet in a cluster with the topic category of that dataset.

Figure 2. Topic hierarchy identified from the analysis of the clusters

As the topics form a hierarchy of different specificities, we created three schemes for classifying the data. The first scheme consists of all the topics found, thus there are 31 labels. In Figure 2, these are the result of following each topic in the first layer to the leaf in the last layer in the topic tree. In the second scheme, the labels are the topics up to the second layer of the tree, thus there are 21. Lastly, the
third scheme consist of 14 topics in the first layer in Figure 3. These three schemes are used to create three datasets, as summarized in Table 3.

Table 3. Summary of the labelling schemes derived from the identified topics hierarchy

| Data set | Topic Specificity | # of Topics | Example Labels |
|----------|------------------|-------------|----------------|
| 1        | Up to layer 3    | 31          | traffic_accident, traffic_vehicle_car, traffic_condition_congested, weather |
| 2        | Up to layer 2    | 21          | traffic_accident, traffic_vehicle, traffic_condition, weather |
| 3        | Layer 1 only     | 14          | traffic, weather, greetings, economy |

For the classification, we explore the use of several algorithms popularly used for text classification: Naïve Bayes (see [28] for a thorough discussion), logistic regression [29], and SVMs [30] with RBF and linear kernels. The parameters used are the default parameters of the scikit-learn [31] implementation we use. In particular, the gamma value of the RBF kernel of the SVM is $1/(\text{number of features})$, which in this case is $1/2052 = 4.8733 \times 10^{-4}$. The default value for the $C$ parameter of the soft-margin SVM classifier [30] is 1.0.

Applying these algorithms to dataset 1 results in accuracy values shown in Figure 3. It shows that the best accuracy is 95.20%, given by SVM-linear, a stark contrast to Naïve Bayes with 32.13% and SVM-RBF with 24.48%. The performance of logistic regression of 94.10% is actually not too far behind. However, we choose SVM-linear because it has the soft-margin parameter $C$ that could still be optimized. For datasets 2 and 3, we get similar accuracy profiles, thus we also choose SVM-linear for these datasets.

Figure 3. Accuracy of classifiers from dataset 1
automatically derived from topic labels on different algorithms

Subsequently, for each dataset, we optimize the $C$ parameter of soft-margin SVM-linear to seek the best performance value possible. Figures 4, 5, and 6 plots the accuracy obtained by systematically varying the value of $C$ using the grid search method [32]; each for datasets 1, 2, and 3, respectively. These figures shows that the categorization schemes, derived from topics identified from the clusters has successfully captured the themes in the public information shared by the community Twitter users revolving around the accounts of @e100ss and @SapawargaSby. Specifically, the highest performance of 96.97% is obtained with the third labeling scheme, occurring at the $C$ value of 2; this is likely to do with the fact that this scheme has the least number of classes. The second best performance is from dataset 1, which is derived from the original clusters found: 95.92% at the $C$ value of 5. Lastly, the second labeling scheme yields the lowest maximum accuracy of 95.51% at the $C$ value of 2; a very good performance nevertheless.
In addition to accuracy, we calculated the precision, recall, and create the confusion matrix for the classifier, and these confirms that the performance of the classifier is indeed very satisfactory. Since all three levels categorizing data works very well, in practice, for a particular application, we could choose any labeling scheme out of the three best on the level of granularity of topic desired. We would also like to point out that since the method we use needs a minimal level of language-specific resources, it could easily be used for other languages as well.

![Figure 4](image1.png)  
**Figure 4.** Accuracy of SVM-linear on dataset 1

![Figure 5](image2.png)  
**Figure 5.** Accuracy of SVM-linear on dataset 2

![Figure 6](image3.png)  
**Figure 6.** Accuracy of SVM-linear on dataset 3

7. Conclusion

In this paper, we described a work on a semi-supervised method for identifying topic and classifying tweets containing public information shared by citizens. We first cluster these tweets using the K-means algorithm using a large number of k, since the data does not seem to naturally have a value for k for which the clustering is optimal. We subsequently analyze the clusters to find the topics of each cluster. Clusters with similar topics are merged; furthermore, we found that some topics have hierarchical relations to each other. Based on this, we create three datasets which corresponds to three levels of granularity on the specificity of the topics, by automatically labeling each tweet with the word descriptive of the category of the cluster. Among the algorithms we explored for creating the classifier, SVM-linear gives the best performance. Optimizing for the soft-margin constant c, we get accuracy of more than 95% for all of the three levels of topic granularity, which shows that our method works very well in finding the topic and categorizing tweets containing public information shared over the social media.

The work in this paper is primarily aimed is to study whether it is possible to find a meaningful identification of topics of public information in the tweets and to retrieve an automatic categorization
based on those topics. Having demonstrated that the tweets do have some identifiable internal organization of topics within them, it is interesting to explore the use of topic modelling with Latent Dirichlet Allocation [33]. LDA is a method which in recent years has been very popular for learning the latent topics in a corpus, using the assumption that the generation of words could be modelled by a Dirichlet distribution. One variation of this algorithm that might be relevant is its use with latent feature word representations [34], since it offers improvement in topic identification in cases where the documents are short.

References

[1] D. Scanfeld, V. Scanfeld and E. L. Larson, "Dissemination of health information through social networks: Twitter and antibiotics," *American journal of infection control*, vol. 38, no. 3, pp. 182-188, 2010.

[2] S. J. Sullivan, A. G. Schneiders, C.-W. Cheang, E. Kitto, H. Lee, J. Redhead, S. Ward, O. H. Ahmed and P. R. McCrory, "What's happening? A content analysis of concussion-related traffic on Twitter," *British Journal of Sports Medition*, 2011.

[3] F. Toriumi, T. Sakaki, K. Shinoda, K. Kazama, S. Kurihara and I. Noda, "Information Sharing on Twitter During the 2011 Catastrophic Earthquake," in *22nd International Conference on World Wide Web*, Rio de Janeiro, 2013.

[4] E. Tonkin, H. D. Pfeiffer and G. Tourte, "Twitter, Information Sharing and the London Riots?," *Bulletin of the American Society for Information Science and Technology*, pp. 49-57, 2012.

[5] F. Kling and A. Pozdnoukhov, "When a city tells a story: urban topic analysis," in *20th International Conference on Advances in Geographic Information Systems*, Redondo Beach, 2012.

[6] T. Heverin and L. Zach, "Twitter for City Police Department Information Sharing," *ASIST 2010*, 2010.

[7] D. Anggareska and A. Purwarianti, "Information Extraction of Public Complaints on Twitter Text For Bandung Government," *2014 International Conference on Data and Software Engineering (ICODSE)*, pp. 80-85, 2014.

[8] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, Y. Kompatsiaris and A. Jaimes, "Sensing Trending Topics in Twitter," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1268 - 1282, 2013.

[9] J. Benhardus and J. Kalita, "Streaming trend detection in Twitter," *International Journal of Web Based Communities*, vol. 9, no. 1, pp. 122-139, 2013.

[10] H. Rosa, J. P. Carvalho and F. Batista, "Detecting a Tweet’s Topic within a Large Number of Portuguese Twitter Trends," *3rd Symposium on Languages, Applications and Technologies*, vol. 38, pp. 185-199, 201.

[11] E. Miller, K. Vodrahalli and A. Lee, "Estimating Trending Topics on Twitter with Small Subsets of the Total Data," 2015.

[12] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom and P. Chaovalit, "Social-based traffic information extraction and classification," in *11th International Conference on ITS Telecommunications*, Petersburg, 2011.

[13] T. Sakaki, Y. Matsuo, T. Yanagihara, N. P. Chandrasiri and K. Nawa, "Real-time event extraction for driving information from social sensors," *Proceedings of the 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems*, 2012.

[14] S. K. Endarnoto, S. Pradipa, A. S. Nugroho and J. Purnama, "Traffic Condition Information Extraction & Visualization from Social Media Twitter for Android Mobile Application," *2011 International Conference on Electrical Engineering and Information*, 2011.
[15] M. Hasby and M. L. Khodra, "Optimal path finding based on traffic information extraction from Twitter," In ICT for Smart Society (ICISS), 2013 International Conference, pp. 1-5, 2013.

[16] A. Wibisono, I. Sina, M. A. Ihsannuddin, A. Hafizh, B. Hardjono, A. Nurhadiyatna, W. Jatmiko and P. Mursanto, "traffic intelligence system architecture based on social media information," International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 25-30, 2012.

[17] D. F. Priyambada, N. Aini and R. Prasetianto, "Rancang Bangun Aplikasi Backend dan Web Services Penyampaian Laporan Masyarakat Berdasarkan Crowdsourcing melalui Jejaring Sosial Twitter," Juli 2016.

[18] T. S. Sati, Sengsara Membawa Nikmat, Jakarta: Balai Pustaka, 1929.

[19] H. P. Luhn, "a statistical approach to mechanize encoding and searching of literary information," IBM Journal, 1957.

[20] K. Spärck Jones, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," Journal of Documentation, vol. 28, pp. 11-21, 1972.

[21] D. H. Wahid, Februari 2016. [Online]. Available: https://github.com/masdevid/ID-Stopwords/blob/master/id.stopwords.02.01.2016.txt. [Accessed 18 August 2016].

[22] Bird, Steven and E. K. Edward Loper, Natural Language Processing with Python, O'Reilly Media Inc., 2009.

[23] S. P. Lloyd, "Least Squares Quantization in PCm," IEEE Transactions on Information Theory, vol. 28, no. 2, p. 129, 1982.

[24] D. Arthur and S. Vassilvitski, "How Slow is the k-Means Method?," in Proceedings of the twenty-second annual symposium on Computational geometry, Sedona, 2016.

[25] J. Issa, "Performance characterization and analysis for Hadoop K-means iteration," Issa Journal of Cloud Computing, 2016.

[26] R. L. Thorndike, "Who Belongs in The Family?," Psychometrika, vol. 18, no. 4, p. 267, 1953.

[27] G. Ball and D. Hall, "A Clustering Technique for Summarizing Multivariate Data," Behavior Science, vol. 12, pp. 123-155, 1967.

[28] D. D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval," in 10th European Conference on Machine Learning, London, 1998.

[29] D. R. Cox, "The Regression Analysis of Binary Sequences," Journal of the Royal Statistical Society. Series B (Methodological), vol. 20, no. 2, pp. 215-242, 1958.

[30] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011.

[32] C.-W. Hsu, C.-C. Chang and a. C.-J. Lin, "A Practical Guide to Support Vector Classification," Taipei, 2016.

[33] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research 3, pp. 993-1022, 2003.

[34] D. Q. Nguyen, R. Billingsley, L. Du and M. Johnson, "Improving Topic Models with Latent Feature Word Representations," Transactions of the Association for Computational Linguistics, vol. 3, p. 299–313, 2015.