A Feature Selection Method for Twitter News Classification

Inoshika Dilrukshi and Kasun de Zoysa

***Abstract*—This paper presence a new feature selection method which can be used for creating data set in order to classify Twitter short messages. The Twitter short messages contain only 140 characters. Thus, the number of words per sentence is almost equal for all sentences. Once you pool the all text messages together, there can be number of words in the pool but, for a given sentence, there will be only few words included from the pool. This causes to have a sparse matrix as the feature vector. By removing the unrelated words from the feature space, the dimension can be reduced and therefore, the sparseness can be reduced. The unrelated words can be define as the common words (high frequent words) and noise words (low frequent words). Even though by removing these unrelated words, still it may contain some unrelated words. Thus, a feature selection technique was needed to apply in order to select the best feature set. The suggested new feature selection method was based on the Information Theory. It was named as Ratio Method. The calculated value increase when the word occurs frequently in a particular group and it decrease when the word occur in all groups. The best features can be choose by using a proper threshold. Some popular text classifiers such as SVM, Naïve Bayes and Decision Trees are used to evaluate the performance of the new feature selection method and to compare the new method with existing methods.**

***Index Terms*—Ratio method, information theory, term frequency, inverse document frequency.**

I. Introduction

Twitter is a Social Network which becomes popular daily. It is popular among many types of users from various countries. Thus, many organizations, news providers use Twitter to share their news and messages. Because of this reason, Twitter contains a large amount of hidden information. This information can be used for many purposes. Thus, by analyzing this news, one can generate a large database which contains real time information. With the development of machine learning techniques [1], now-a-days, many researchers tend to use machine learning techniques in text analyzing [2], [3]. However, the important fact is to choose the best feature set where the machine learning could properly apply.

The format of Twitter short messages is text. Therefore, by performing a suitable text mining technique, one can extract the unseen information about the short messages. However, analyzing text is not much easy as analyzing numerical values. In text mining, words are used as features. Text may contain stop words, ambiguous words, prefixes and postfixes. This will cause to increase the dimension of the feature set. A high A high dimensional feature set will cause to have a complex model, and therefore it can cause to over fitting. Thus, a proper feature selection method was needed to reduce the dimension as the performance of the classification will be highly depending on the feature selection technique.

Using Twitter Social Network, there was an advantage and a disadvantage as it restricts the character length into 140. The advantage is, it automatically restricts the number of words. Thus, the number of words for a sentence is almost same. Therefore, the feature vector does not need to normalize. The disadvantage is, this will cause to increase the sparseness of the matrix. For an example, considering 2 following Twitter message examples:

"President pays last respects to Jayalath Jayawardena"

"President returns to Sri Lanka"

Once we pool the 2 sentences in order to get the feature set, the features will be as follows.

"President", "pays", "last", "respects", "to", "Jayalath", "Jayawardena", "returns", "Sri", "Lanka" In the 2 sentence, only the words "President" and "to" will be the common word to both sentence. Thus, the instance regard to 1st sentence is,

1,1,1,1,1,1,1,0,0,0

And the instance regarding to 2nd sentence is,

1,0,0,0,1,0,0,1,1,1

Result is a sparse matrix. By removing unrelated words from the feature set, one can reduce the sparseness of the matrix.

According to the Ziph's Law [4], given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table which means, highly used words(common words) will provide very low information. Thus common words can be removed by removing most frequent words. According to Pang et al[3] and Joachims [5], the lowest frequency words does not contain much information and can be neglected as noise words. To remove the noise words and common words, a threshold was needed to be define. This can be easy in document as there are large number of paragraphs and words thus, difference between common words and features are highly noticeable. However, not like documents which contain paragraphs, in Twitter short messages, sometimes there won't be much difference in between the frequencies of most important features and stop words. Thus, by defining a threshold, these unwanted words can remove only for some extent because of the word restriction. Removing stop words manually may cause to lose some important information because sometimes, a stop word can be a feature which provide very useful information. Thus, a proper feature

Manuscript received February 18, 2014; revised May 10, 2014. This work was supported in part by the National Research Council under Grant 11-25.

The authors are with the University of Colombo School of Computing, Colombo 7, Sri Lanka (e-mail: inoshi@scorelab.org, kasun@ucsc.lk).

selection method was needed for further dimension reduction.

Forward selection and backward elimination [6] are two popular statistical techniques which can be use for feature selection. This technique does not performs well when the features are correlated. However, words are very correlated. There are some other methods which was based on Claude Shannon's Information Theory. These methods choose the attributes which provide the highest information. The main focus of these methods is to minimize the randomness of the attribute.

The suggested approach was developed based on Claude Shannon's Information Theory. The main focus of the suggested system is to measure the relevance of the attribute towards the category. Both the frequency within category and frequency of words will be considered for calculation. The calculated value will be increase when it occurs frequently in a given group and the value get decrease when it occurs commonly in all groups. Using a threshold, one can separate relevant attributes from irrelevant attributes.

The new feature selection method was compared with other feature selection methods which are commonly use in text classification. These feature selection methods are applied into selected classifiers in order to measure the performance. SVM, Naive Bayes classifier and Decision Trees are the common techniques which often use for text classifications. Thus, in this paper the existing methods and new method was evaluated using these techniques. The content of the paper was ordered as follows. Section II will describe the state of the art for feature selection. Section III will brief out the suggested feature selection method. Section IV will describe the evaluation of the suggested method and conclusion will brief out in Section V.

II. State of the Art for Feature Selection

A. Sequential Forward Selection

Sequential Forward Selection starts with the empty set and sequentially adds one feature at a time. A problem with these Sequential Forward Selection techniques is that when a feature added in Sequential Forward Selection cannot be deleted once selected [6].

B. Sequential Backward Elimination

In Backward feature elimination, it starts with all the features and sequentially eliminates one feature at a time (eliminating the feature that contributes least to the criterion function). A problem with this Sequential Backward Elimination techniques is that when a feature is deleted, it cannot be re-selected [6].

The issue which researchers faced when using statistical feature selection methods is it is not applicable when the variables are correlated. Thus, they discovered methods which are based on Shannon's Information Theory. Following are the methods which are usually used in text classification.

C. Information Gain

Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A) [7]. That is,

$$Gain_A = Info(D) - Info_A(D) \quad (1)$$

where

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \quad (2)$$

and

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{D} \times Info(D_j) \quad (3)$$

D. Inverse Document Frequency

Inverse Document Frequency (IDF) [8] represents the scaling factor, or the importance, of a term t. If a term t occurs in many documents, its importance will be scaled down due to its reduced discriminative power [7]. The equation of IDF is,

$$IDF(t) = \log \frac{1+|d|}{|d_t|} \quad (4)$$

where d is the document collection, and d_t is the set of documents containing term t. If $|d| << |d_j|$ the term t will have a large IDF scaling factor and vice versa.

E. Term Frequency

The term frequency be the number of occurrences of term t in the document d, that is, $freq_{(d,t)}$. The (weighted) term-frequency matrix $TF_{(d,t)}$ measures the association of a term t with respect to the given document d: it is generally defined as 0 if the document does not contain the term, and nonzero otherwise [7].

$$TF_{(d,t)} = \begin{cases} 0 & \text{.....iff} freq_{(d,t)}=0 \\ 1+\log(1+\log(freq_{(d,t)}=0 & \text{.....otherwise} \end{cases} \quad (5)$$

III. The Suggested Feature Selection Method

The issue of previous define methods is, most of them deals with the frequency of a word occurrence. But in Twitter, the number of words are restricted. This will bound the users to create the text message as short as possible. Thus, when creating the shortest sentence, some keywords maybe use frequently, more than usual common word and some stop words may not get used frequently. Thus, the number of occurrence may not provide a better idea about the importance of the word. Thus, the suggested method was developed in order to identify the importance of the word by compare the occurrence with respect to the group (class label).

Before applying the suggested method, one can remove the noise words and common words as primary feature reduction. hus, the low frequent words and high frequent words can be removed. Then, the further feature selection method can be

applied. The suggested method was based on Information Theory. The main idea of the method is, to eliminate the stop words which did not captured when removing high frequent words and to eliminate other words which do not provide much information to identify the category. A term called Frequency Ratio will be calculated for the selection process.

Assume that our classifier needs to classify Twitter short messages into *n* number of pre-defined groups. The data which will be used to feature selection should be tagged into groups. Then, the calculation is as follows.

Let $F_{(i,j)}$ be the frequency of the i^{th} word in j^{th} group. Then the term frequency $tf_{(i)}$ of i^{th} word can be calculated by,

$$tf_i = \sum_{j=1}^{n} F_{(i,j)} \tag{6}$$

Then, the Frequency Ratio, $FR_{(i,j)}$ for i^{th} word given the group *j* can be calculated as,

$$FR_{(i,j)} = \frac{F_{(i,j)}}{tf_{(i)}} \tag{7}$$

This will provide us an idea about how important the word *i* could be to the selected group *j*.

Then, the Maximum Frequency Ratio for given word *i* will be calculated as,

$$MRF_i = \max\left\{FR_{(i,j)}\right\} \tag{8}$$

Now, by providing a threshold value, one can filter the keywords from unrelated words. The concept lies as follows. If the i^{th} word is a keyword, the frequency, $F_{(i,j)}$ for a selected group, say j^{th} should be a large value. For other groups, the frequency will be a very low value. Therefore the term frequency $tf_{(i)}$ won't be much larger value compared to $F_{(i,j)}$. Therefore, the ratio $F_{(i,j)}$will be large for j^{th} group.

If the i^{th} word is a stop word or a non-related word, the frequency, $F_{(i,j)}$for a selected group, say j^{th} may be high or low. However, even for other groups, the distribution of the frequency won't be change with respective to the selected group. Thus, there is no significant difference of the $FR_{(i,j)}$ for a particular group with respect to other groups. Therefore the term frequency $tf_{(i)}$ won't be closer to any $F_{(i,j)}$value. Therefore, the ratio, $FR_{(i,j)}$will be small for all groups. In short, the distribution for a keyword is not common to all groups but the distribution for a stop word is almost common for all groups.

In more descriptive way, if a given word i is a keyword, there should be a group *j** where $F_{(i;j^*)}$ is significantly large than other group's $F_{(i,j)}$. Therefore, the Frequency Ratio $FR_{(i;j^*)}$ will be closer to 1. By getting MFR_i value, we can obtain the maximum $FR_{(i;j^*)}$ value for a given word. Thus, large MFR_i values will obtain if the i^{th} word is a keyword for any group, and small MFR_i values will be obtain if i^{th} word is not a keyword for any group. By applying a threshold, we can separate the keywords from non-related words.

Five active Twitter local news providers were chosen to evaluate the method. The reason for choosing news providers is, they provide news in standard English without unstructured words and shorten words. The news providers are,

- Ada Derana
- Ceylon Today
- ITN
- Lanka Breaking News
- News First

The news was classified into 12 categories manually. The 12 categories were defined according to popular newspaper articles and news websites [9]. The 12 groups are,

- Economic-Business
- War-terrorist-crime
- Health
- Sports
- Development-government
- Politics
- Accident
- Entertainment
- Disaster-Climate
- Education
- Society
- International

An amount of 3336 news records were taken from these 5 news providers. Preprocessing was done in order to remove the traceable noise words and common words. The noise words were remover by removing low frequent words and the common words were removed by removing high frequent words.

For a perfect evaluation of a feature selection, the dataset which was used for feature selection should be independent from the training dataset and testing dataset. Thus, from the first thousand records (record number 1-1000) was used for feature selection, next 1500 records (1001-2500) was used to train the system and the rest was used for test the system. For the suggested feature selection method, the value 0.5 was used as the threshold.

IV. Evaluation

The evaluation was carried out in order to measure the effectiveness of the suggested method. Effectiveness is purely a measure of the ability of the system to satisfy the user in terms of the relevance of short messages retrieved [10]. It is assumed that the more effective the system, the more it will satisfy the user.

The effectiveness of the retrieval system was measured using precision, given in equation 9, and recall, given in equation 10, values [10]. Precision is the fraction of retrieved short messages that are relevant. Recall is the fraction of relevant short messages that are retrieved [11].

$$precision = \frac{tp}{tp + fp} \tag{10}$$

$$recall = \frac{tp}{tp + fn} \tag{11}$$

The new method, Ration method, was compared with 4 popular feature selection methods. They are,

- Sequential Forward Selection
- Sequential Backward Elimination
- Information Gain
- Chi square

Table I shows the Precision and the recall values for the group accident. The feature selection methods are tested using 5 classification methods. Those are,

- J48
- Support Vector Machine
- Random Forest
- Random Tree
- Simple CART

TABLE I: EVALUATION OF FEATURE SELECTION METHODS FOR THE GROUP ACCIDENT

Classification method	Feature selection method	No of features	Precision	Recall
J48 decision tree	Sequential Forward Selection	13	0.714	0.321
	Sequential Backward Elimination	13	0.714	0.321
	Information Gain	49	0.844	0.346
	Chi square	49	0.844	0.346
	Ratio method	270	0.868	0.387
SVM	Sequential Forward Selection	13	0.826	0.487
	Sequential Backward Elimination	13	0.826	0.487
	Information Gain	49	0.969	0.397
	Chi square	49	0.969	0.397
	Ratio method	270	0.977	0.361
Random Forest	Sequential Forward Selection	13	0.792	0.487
	Sequential Backward Elimination	13	0.792	0.487
	Information Gain	49	0.939	0.59
	Chi square	49	0.939	0.59
	Ratio method	270	0.935	0.689
Random Trees	Sequential Forward Selection	13	0.97	0.410
	Sequential Backward Elimination	13	0.714	0.321
	Information Gain	49	0.885	0.59
	Chi square	49	0.844	0.346
	Ratio method	270	0.988	0.689
Simple CART	Sequential Forward Selection	13	0.792	0.487
	Sequential Backward Elimination	13	0.714	0.321
	Information Gain	49	0.897	0.449
	Chi square	49	0.844	0.346
	Ratio method	270	0.822	0.504

It also displays the number of features which was selected by each method. However, it is not easy to compare 2 values to identify the best feature selection method. Thus, we use F -measure to get a single value instead of Precision and Recall. The F-measure can be calculated as follows.

$$F_\beta = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R} \quad (11)$$

For β equals 1, one can calculate the balance F measure (Harmonic mean). Many alternative methods were proposed over the years and Harmonic mean had identified as the best single value summaries. The weighted harmonic mean can be calculated using (11) by changing the β value. Values of β less than 1 emphasize the precision whereas values of β greater than 1 emphasizes recall [11].

The Table II shows the F values which was calculated for the 12 groups. According to Table II, it is clear that except for Develop and Government, for all other groups, the highest F value had obtained for the dataset where the features were selected using Ratio method. The highest accurate classification technique was changed subject to the group as the distribution of the data in each group is subjective. In each group, when considering one classification method per time, some classification methods provide the best result for the dataset which were selected using Feedfarward method and Information gain method. But, when considering the maximum accuracy for each group regardless the classification method, the ratio method had performed very well.

TABLE II: F-MEASURE FOR FEATURE SELECTION METHODS

Group	Classifier	Feedfarward and feedback	Information Gain and Chi square	Ratio Method
Accident	J48	0.442	0.491	0.535
	SVM	0.613	0.564	0.528
	Random Forest	0.603	0.724	0.800
	Random trees	0.577	0.708	0.812
	Simple CART	0.603	0.598	0.625
Development and Government	J48	0.034	0.344	0.223
	SVM	0.332	0.321	0.210
	Random Forest	0.342	0.554	0.508
	Random trees	0.342	0.548	0.509
	Simple CART	0.330	0.463	0.327
Disaster and Climate	J48	0.341	0.327	0.526
	SVM	0.465	0.543	0.519
	Random Forest	0.542	0.600	0.681
	Random trees	0.542	0.524	0.681
	Simple CART	0.435	0.459	0.542
Economy and Business	J48	0.254	0.257	0.352
	SVM	0.523	0.366	0.546
	Random Forest	0.554	0.653	0.681
	Random trees	0.550	0.649	0.681
	Simple CART	0.532	0.543	0.545
Education	J48	0.512	0.573	0.461
	SVM	0.817	0.784	0.566
	Random Forest	0.810	0.888	0.875
	Random trees	0.827	0.889	0.890
	Simple CART	0.823	0.817	0.778
Entertain	J48	0.545	0.549	0.645
	SVM	0.428	0.428	0.432
	Random Forest	0.657	0.687	0.734
	Random trees	0.541	0.548	0.563

	Simple CART	0.431	0.631	0.425
Health	J48	0.578	0.596	0.658
	SVM	0.647	0.574	0.687
	Random Forest	0.689	0.657	0.727
	Random trees	0.573	0.659	0.729
	Simple CART	0.289	0.382	0.402
International	J48	0.526	0.517	0.538
	SVM	0.415	0.538	0.472
	Random Forest	0.592	0.649	0.727
	Random trees	0.628	0.616	0.729
	Simple CART	0.437	0.528	0.631
Politics	J48	0.371	0.527	0.551
	SVM	0.627	0.649	0.742
	Random Forest	0.759	0.742	0.961
	Random trees	0.684	0.785	0.993
	Simple CART	0.478	0.528	0.689
Society	J48	0.511	0.493	0.512
	SVM	0.511	0.493	0.523
	Random Forest	0.673	0.694	0.946
	Random trees	0.551	0.674	0.993
	Simple CART	0.562	0.549	0.626
Sports	J48	0.752	0.798	0.804
	SVM	0.401	0.412	0.429
	Random Forest	0.853	0.873	0.987
	Random trees	0.812	0.864	0.993
	Simple CART	0.511	0.518	0.538
War Terrorist and Crime	J48	0.697	0.751	0.804
	SVM	0.513	0.483	0.429
	Random Forest	0.794	0.874	0.987
	Random trees	0.829	0.863	0.999
	Simple CART	0.640	0.679	0.740

V. Conclusion

The research was done in order to introduce a new feature selection method. The feature selection method was applied for Twitter short messages. Considering about Twitter short messages, it had restricted the character length up to 140 characters. This will cause to have large number of words with small frequencies. To create the feature set, the idea is to pool the words together. Once pool these short messages in order to extract the features, this restriction will cause to have large number of features and only few of them will occur in a given sentence. Thus the result will be a sparse matrix. Therefore it is essential to have a proper feature selection in order to reduce the sparseness.

Most of feature selection methods remove the stopwords by providing a stopwords list. There can be some words where one can define it as a stopword and one can define it as features. Therefore, removing the stopwords directly will make harm for the classification accuracy. The technique which was used for the current research is, to remove common words and noise words. However, most common words and noise words does not provide a valuable information. High frequent words can be identified as common words and low frequent words can be identified as noise words. Thus, high frequent words and low frequent words were removed from the feature set. However, to reduce the sparseness further, the researcher had introduced a new feature selection method.

When selecting features, the idea is to consider about the probability of frequency which the given feature appear in the given group. In this case, we can use the frequencies instead of probability as the number of characters is restricted by Twitter. However, it is essential to find which features are specified to a given group. This means, it should appear in a given group frequently, and should not appear in other groups frequently. The suggested method was built on this concept.

In the suggested system, each frequencies $F_{(i,j)}$ for i^{th} word and j^{th} group was calculated. Then, the term frequency $tf_{(i)}$ for the i^{th} word was calculated by adding the $F_{(i,j)}$ frequencies. Then the frequency ratio $FR_{(i,j)}$ was calculated by dividing the $F_{(i,j)}$ from $tf_{(i)}$ for each i^{th} word and j^{th} group. Finally, the maximum frequency ratio MFR_i for i^{th} word was calculated by taking the maximum value of $FR_{(i,j)}$ for each i^{th} word.

The performance of the new system was compared with feedfarward selection, feedback selection, Information Gain and Chi square feature selection. The text was classified using J48, Support Vector Machine, Random Forest, Random Trees and Simple CART. The precision and recall values were calculated for each situation. Harmonic mean was calculated in order to get a single value for the comparison.

According to the results, it shows that the suggested method provides better results than existing methods. The classification method can be change depending on the distribution of the given dataset. However, the results shows that the feature selection method provide the best result for most of the situations, regardless of the distribution of the data. The accuracy can be increase by using bagging and boosting techniques.

References

[1] N. J. Nilsson, *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann, 1998.

[2] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." *LREC*, 2010.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proc. the ACL-02 conference on Empirical methods in natural language processing*, vol. 10, Association for Computational Linguistics, 2002, pp. 79–86.

[4] G. K. Zipf, *Human Behaviour and the Principle of Least-Effortg*, 1949.

[5] T. Joachims, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, 1998.

[6] H. Liu and H. Motoda, *Computational Methods of Feature Selection*, CRC Press, 2007.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 2nd ed., 2006.

[8] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[9] I. Dilrukshi, K. De Soysa, and A. Caldera, "Twitter news classification using SVM," in *Proc. 2013 8th International Conference on IEEE Computer Science & Education (ICCSE)*, 2013, pp. 287–291.

[10] C. van Rijsbergen, *Information Retrieval*, Butterworth, 1979.

[11] C. D. Manning, P. Raghavan, and H. Sch ̈utze, *Introduction to information retrieval*, Cambridge University Press Cambridge, 2008, vol. 1.

Inoshika Dilrukshi was born in 1987 at Colombo 4, Sri Lanka. She received BSc. (special) in industrial statistics in 2011 from University of Colombo, Department of Statistics. She had worked at ITI laboratory, Sri Lanka with Dr. Janaki Gonathilaka, Center of Digital Forensic, UCSC, Sri Lanka as a research assistant. At present she works as a research assistant at University of Colombo School of Computing. She was reading for a MPhil at University of Colombo, School of Computing, Sri Lanka, under the topic "Social Sensor Network for Opinion Analyzing". She was a member of Sri Lanka Association for the Advancement of Science (SLAAS). Her research interests includes data mining, neural network, pattern recognition, character recognition, image processing, text mining.

Kasun de Zoysa obtained his BSc in computer science from University of Colombo, Sri Lanka in 1998. He obtained his Ph.D. in the area of computer security from Stockholm University in 2003. After a brief period of post-doctoral work at the George Washington University, USA, he returned to the University of Colombo School of Computing(UCSC), Sri Lanka in 2003. At present he works as a senior lecturer and student counsellor at UCSC. In addition to that he works as an advisor for Sustainable Computing Research Group (SCORE Group) and Centre for Digital Forensic attached to UCSC. He also work as the head of the Computer Security and Forensics Special Interest Group at the Computer Society of Sri Lanka (CSSL). His research interests include secure multi-party electronic transactions and document protection, public key cryptography and certification infrastructures, hardware security tokens and devices, web security and privacy, security in wireless ad-hoc and sensor networks, and digital forensic.