CONTENT MINING OF MICROBLOGS

M.Özgür Cingiz

Computer Engineering Department
Yıldız Teknik Üniversitesi
mozgur@ce.yildiz.edu.tr

Banu Diri

Computer Engineering Department
Yıldız Teknik Üniversitesi
banu@ce.yildiz.edu.tr

ABSTRACT

Emergence of Web 2.0, internet users can share their contents with other users using social networks. In this paper microbloggers' contents are evaluated with respect to how they reflect their categories. Migrobloggers' category information, which is one of the four categories that are economy sport, entertainment or technology, is taken from wefollow.com application. 3337 RSS news feeds, whose category labels are same with microbloggers' contributions, are used as training data for classification. Unlike the similar studies if a feature of microblog doesn't appear in RSS news feeds as a feature, this feature is omitted so abbreviations and nonsense words in microblogs can be eliminated. In this study two types of users' contributions are taken as test data. These users are normal microbloggers and bots. Classification results show that bots provide more categorical content than normal users.

1. INTRODUCTION

Recent advancements in Web 2.0, people can't be regarded as simple content reader they can also contribute content as writers. Web 2.0 introduces concepts like social network, blog and microblogs with internet users. Users share their opinions, feelings, images, favorite videos and other user's contributions as microblog content.

Microblogs differ from blogs. Microblogs have size limitation for content. Twitter is one of the most popular microblog applications because of its easy sign up process, easy to use and mobile access. It has limitation of 140 characters for content. User contribution is called as tweet in Twitter.

In Twitter users can follow other users with respect to their field of interest. Followers expect users who are followed in terms of their categorical information, to share content about their field of interests. This study aims to evaluate two types of users' contents according to specifying they reflect their category or not.

Users can find out microblogger's category information with using some applications such as wefollow.com. Users can describe their field of interests and category which they provide tweet about. Because of the character limitations microbloggers' contents can consist of abbreviations and nonsense words so this decrease success of classification. In this paper, we aim to eliminate this kind of features that lead to decrease success of classification.

Similar studies can be separated into two area. First one is to find out user who shares similar interest. Second is to obtain patterns from microblogs. Degirmencioglu [1] extracts word-hashtag, word-user and hashtag-user pairs from tweets to discover users' common interest areas. Yurtsever [2] classifies microbloggers according to their contents with using semantic resources. Akman [3] extract categorical features from 150 microbloggers' contents. Aslan [4] uses news pattern similarity for discovering microbloggers who broadcast news content. Pilavcilar [5] classify texts with using text mining techniques that some of them are used in this study. Güc [6] uses microbloggers' contents and text classification techniques to measure convenience of users' categories.

In this study in part two we examine data sets and their features. In third part analyzing prepared model and model steps to find out users whose contents are more valuable for its related category. In last part, we refer classification results and feature work.

2. DATA SETS

This study consists of two parts. First part is training part and second is test part. Training part data consists of RSS news feeds. Content suppliers like BBC, CNN, SKYNews provide their subscribers news with RSS format. RSS is a kind of web feed format like atom. Users can follow news with using web browsers or aggregators. We use RSS4j java library for getting RSS news feeds. 3337 RSS news feeds whose category is one of the four categories which are economy sport, entertainment or technology. 924 entertainment, 738 technology, 721 economy and 954 sport RSS news feeds are taken for build training model.

In test part 10630 tweets are obtained form 32 bot users and 30 normal users. Category information of bot users and normal users are taken from wefollow.com application. Category labels are the same with training case. We obtain users' tweets with using Twitter4j java library.

3. PROPOSED SYSTEM

Figure 1 shows the steps of the proposed system. Proposed system consists of two phase. These phases are training and testing phases.

In training phase RSS news feeds are used for building training model. Content distributors supply categorical information of RSS news feeds so we can obtain category of training data. However, summaries of news also consist of valuable features so taking RSS news feeds as training data reduce feature of training data set. RSS news feeds and microbloggers' contents are taken in same time period for checking up-to-dateness of microblogers' contents.

After retrieval of RSS news feeds, RSS news feeds are processed for text classification. In text classification area vector space model is used as representing text documents as vectors in vector space. In training phase steps of text preprocessing, feature weighting, dimension reduction, specifying term count threshold are applied to RSS news feeds respectively. After preprocessing steps, Support Vector Machines and Multinominal Naive Bayes are used separately as classifier for training phase.

In test phase tweets of 30 normal microbloggers and tweets of 32 bots, which their categorical information is obtained from wefollow.com application, are used. Microbloggers' categories are sport, economy, entertainment and technology. Before the selection of features of microbloggers' contents, removing punctuations and tokenization steps are applied. Microbloggers' tweets split into their tokens (features, words). If any word that is part of microbloggers' tweet doesn't be in training feature set, this word is omitted. Features are only taken from training set and search these features in microbloggers' contents because of abbreviations and nonsense words in microblogs. If these words are regard as features, classification success rate is decreasing and testing phase results are specious. After feature specifying steps, features are weighted.

Contents of tweets can be hyperlink of picture or video so in testing phase we eliminate hyperlinks. After selection of only training features and removal of hyperlinks, some tweets become featureless. Featureless tweets are meaningless as test data so we specify 3 different term count threshold values. Tweets must include at least three, four or five words as test data. Testing is implemented with these 3 different threshold values. Test data is given to training model for classification that is formed by Support Vector Machines classifier and Multinominal Naive Bayes classifier. In this section we explain all training and testing steps clearly.

3.1. Preprocessing

Removal of punctuations, tokenization, and selection of features in terms of their linguistic information, stemming and elimination of stop words are preprocessing steps that are used in text mining area. According to selection of linguistic features only nouns and verbs are used as features.

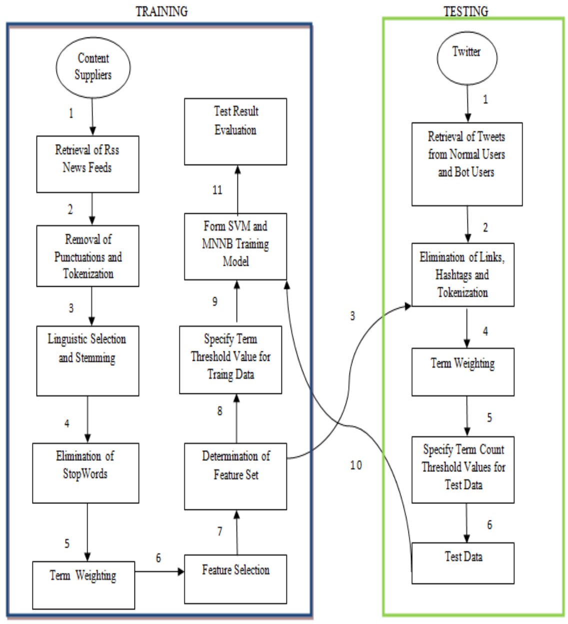
TRAINING
TESTING
Content Suppliers
1
Retrieval of Rss News Feeds
2
Removal of Punctuations and Tokenization
3
Linguistic Selection and Stemming
4
Elimination of StopWords
5
Term Weighting
6
Feature Selection
7
Determination of Feature Set
8
Specify Term Threshold Value for Traing Data
9
Form SVM and MNNB Training Model
11
Test Result Evaluation
10
3
Twitter
1
Retrieval of Tweets from Normal Users and Bot Users
2
Elimination of Links, Hashtags and Tokenization
4
Term Weighting
5
Specify Term Count Threshold Values for Test Data
6
Test Data

Figure 1. Proposed System Structure

3.1.1. Removal of Punctuations and Tokenization

First preprocessing step is removal of punctuations of RSS news feeds. After removal of punctuations, word tokenization is applied to train data. Word tokenization splits RSS news feeds into their words. Tokens also can be n-grams or collocations but in this study words are taken as features of text. In vector space RSS news feeds are shown as vectors, tokens of RSS news feeds are dimension of concerned vector.

3.1.2. Linguistic Selection and Stemming

In previous text classification works features are evaluated separately according to their linguistic labels. Classification results which are obtained using different features according to their linguistic information shows that nouns and verbs are more valuable features than other types [7, 8]. In this study only nouns and verbs are used as features. Pronouns, adjective, conjunctions are eliminated. Stanford[1] part-of-speech tagger is used for getting linguistic information of words.

Words can be in different formats in texts. Such as "*children, child"* is *different* forms of noun *child* or "*drink, drank, drunk"* is different forms of verb *drink*. Stemming is necessary for successful classification and feature reduction.

3.1.3. Elimination of Stop Words

Stop words are commonly used in every category so they don't have any differentiation impact. Stop words decrease classification success. We eliminate stop words from feature set in preprocessing phase.

[1] http://nlp.stanford.edu/software/tagger.shtml

3.2. Term Weighting

In vector space model a text document is symbolized as vectors, words (features, terms) in this text document are symbolized as dimensions of vector. In vector space model every word has a weight value if it is in text document. In this paper, term frequency-inverse document frequency (tf-idf) is used for weighting for features. In related works show that selection of weighting approach is more important than selection of kernel function for Support Vector Machines classifier [9, 10].

In tf-idf weighting term frequency, tf, gives number of times a term occurs in a text document. Inverse document frequency, idf, gives number of times a term occurs in whole text documents. If any terms occur in every document, it is worthless feature for classification. Valuable features for classification have high term frequency score and low inverse document score. Equation 1 and equation 2 show calculation of term frequency and inverse document frequency and equation 3 shows tf-idf weighting calculation. In this study tf-idf weighting is used as term weighting.

$$\mathrm{tf(d,f)} = \begin{cases} 0, \text{ if } \mathrm{(d,t)}=0 \\ 1+\log(1+\log\big(\mathrm{frequency(d,t)}\big)\,) \end{cases} \quad (1)$$

$$\mathrm{idf(t)} = \log\ \frac{\mathrm{n}}{|\mathrm{dft}|} \quad (2)$$

$$\mathrm{tf\text{-}idf(d,t)} = \mathrm{tf(d,f)} * \mathrm{idf(t)} \quad (3)$$

3.3. Feature Selection and Term Count Threshold

In vector space model contains high dimensional sparse feature vectors. In text mining works every term is represented as feature so this makes vector high dimensional. A RSS news feed is summary of news content so it doesn't have many features but 3337 RSS news feeds generate high dimensional feature space. Classification is hard, ineffective and time consuming to implement in high dimensional feature space so dimension reduction is necessity. We eliminate stop words and taking only nouns and pronouns therefore we put term count threshold value for training data.

Two common approaches are used in dimension reduction. These are feature selection methods and feature extraction methods. Feature extraction methods combine different features to make new and low dimension feature set. Feature selection methods try to select the best subset of feature set. Two different types of feature selection approach are used in literature. These are wrapper methods and filtering methods. Wrapper methods find best subset with testing all subset combinations. Filtering methods order all features according to filtering approaches. Chi square statistics, document frequency, information gain, mutual information is well known feature selection filtering methods. In similar works [11, 12] chi square statistics and information gain methods give better result than other filtering methods so these two methods are applied separately as feature selection methods in this study.

Chi square statistics and information gain methods are applied to RSS news feeds for dimension reduction. The most successful classification result ,which is equal to 92,2% F_1-Measure, are taken with using Multinominal Naive Bayes classifier and chi square statistics. 9477 features are reduced to 1296 features with chi square statistics method.

Equation 4 shows chi square statistics of t- th feature in class *c*. Chi square statistics value is calculated with occurrence of term in class and absence of the same term in the same class. Chi square statistics give good results in high dimensional sparse feature set.

$$X^2(t,c) = N\left[\frac{\mathrm{P(t,c).P(t^{-},c^{-}) - P(t,c^{-}).P(t^{-},c)}}{\mathrm{P(t).P(t^{-}).P(c).P(c^{-})}}\right]^2 \quad (4)$$

3.4. Retrieval of Test Data

This study aims to evaluate contents of microblogs. 32 bot users' tweets and 30 normal users' tweets are taken as test data. Study also makes comparison between bots and normal users according to their contributions. Categorical information of bots and normal users are taken from wefollow.com. After retrieval of tweets from these two different user types, removal of punctuations and tokenization, which are preprocessing steps, are implemented to test data. Links of images and videos are omitted from tweets. Hashtags are also omitted from tweets. Some tweets consist of only links so after elimination of links make tweets featureless. Featureless tweets or tweets that have one or two features decrease classification success rate. So in test phase description of term count threshold is necessity. We specify three term count threshold values for tweets.

Table 1. User Tweets and Term Count Threshold Values

	Term Count Threshold Values		
	>2	>3	>4
Number of Normal Users' Tweets	921	416	162
Number of Bots' Tweets	2115	1039	435

Tweets that are used as test data are arranged according to its term count in testing. Tweets which have more than two terms, three terms and four terms are only evaluated as test data. Table 1 shows that if term counts in tweets and number of tweets which have more than specified term count threshold value is inversely proportional.

This study use only training feature set in training phase and testing phase. After tokenization steps of tweets, features are obtained. If a feature of tweet doesn't occur in RSS news feeds then feature is eliminated from test data set. Tweets have 140 character limitations so microbloggers use abbreviations and nonsense words that belong to social networks. Elimination of words which don't occur in training feature set provides to omit abbreviations and nonsense words so this process enables to make correct classification.

Tweet of normal users and bots are taken in the same time period with RSS news feeds for checking users' up-to-dateness.

3.5. Classification

Multinominal Naive Bayes (MNNB) and Support Vector Machines (SVM) are used as classifier in training and test phases. SVM is popular classifier in text classification area. SVM outperforms k-Nearest Neighbor, Linear Least Square, Naive Bayes, Neural Networks and Decision Methods in terms of classification results [13, 14]. SVM is also good classifier in many other classification areas whose dimensions are high. Multinominal Naive Bayes are especially used in information retrieval and text mining works. It gives good results in these work areas.

3.5.1. Support Vector Machines

Support Vector Machines try to determine the most suitable decision boundary which separate data into their correct classes. The decision boundary must be as far away from data of all class as possible.

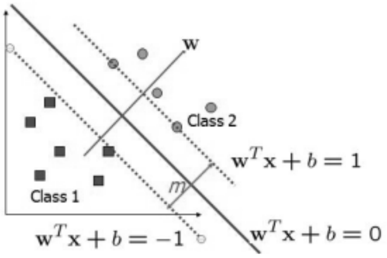
w
Class 2
$\mathbf{w}^T\mathbf{x} + b = 1$
m
Class 1
$\mathbf{w}^T\mathbf{x} + b = -1$
$\mathbf{w}^T\mathbf{x} + b = 0$

Figure 2. SVM and Decision Boundary

Dashed lines show boundaries of each class. Thick line indicates the decision boundary. Samples that are on the dashed lines are called as support vectors. Decision boundaries are determined by support vectors. Data except support vectors has no weights for determination of decision boundaries. Equation 5 shows that for an optimal decision boundary margin (*m*) must be maximized.

$$m=\frac{2}{||w||} \qquad (5)$$

In other words distance between decision boundary and boundaries of each class is aimed to maximize with minimizing ||w||. Assume that class labels of x_i are y_i € {1,-1} and {x_1, x_2, x_3,.., x_n} is data set. The decision boundary should classify all data correctly with following constraint that is described in equation 6. Equation 6 tries to prevent data from falling into margin.

$$yi(w^tx+b)\geq 1 \qquad (6)$$

In some cases data can't be separated like Figure 2 with linear boundary. Data are transformed to higher dimensional space for linear separation with kernel functions. Polynomial, hyperbolic tangent and radial bases functions are used as kernel functions. In this study we prefer linear classifier that generally gives good results.

3.5.2. Multinominal Naive Bayes

Naive Bayes assumes that occurrence of terms are independent from each other. Multinominal Naive Bayes (MNNB) differs from Naive Bayes according to count of term occurrences in text document. Count of term occurrences is used for calculating probability which shows occurrence of term in related class. Equation 7 shows that multiplication of the conditional probabilities for all terms which occurs in the same class gives probability of related class. After probabilities of all classes are calculated, the class which has the highest probability value is selected as correct class among all the probable classes. Equation 8 aims to eliminate zero probability for class so Laplace smoothing is used for it. Class label is given as *c* and term in a text document is given as *t.*

$$P(c|d)=\arg\max_{c \,€\, C}\ P(c)\prod_{1\leq k\leq V}P(t_k|c) \qquad (7)$$

$$P(t|c)=\frac{T_{ct} \quad +1}{\sum t'€V\ ^{T_{ct}}+B'} \qquad (8)$$

4. EXPERIMENT RESULTS AND DISCUSSION

Training model is formed by 3337 RSS news feeds. Categorical information of RSS news feeds is given by content suppliers. Four categories are used in this study. These categories are sport, economy, entertainment and technology. RSS news feeds which are preprocessed for text classification are used to form training models by Multinominal Naive Bayes and Support Vector Machines separately.

In testing phase 32 bot users' tweets and 30 normal users' tweets are taken as test data. Categorical information of users is taken from wefollow.com application. Table 2 shows categorical information of user. Number of bot users whose category is sport is higher than other users who have different category. However, numbers of tweets which are taken from bot users whose category is sport are less than other users.

Table 2. Categorical information of users

	Number of Normal Users	**Number of Bot Users**
Sport	8	12
Entertainment	7	7
Technology	7	6
Economy	8	7

Tweets of two different types of users are given as test data to the training model which is formed by RSS news feeds. Three different term count threshold values are used for test data. These threshold values are more than two, more than three and more than four. More than two means tweets must contain more than two terms, more than three means tweets must contain three terms and more than four means tweets must contain more than four terms. Table 1 shows the number of tweet in terms of threshold values.

F-measure measures for evaluating the performance of classification. F-measure is weighted harmonic mean of precision and recall. Precision and recall weights are taken equal to each other. This is also know F_1 measure. Table 3 gives F-measure values of classification results. First value that is given under the threshold value indicates F-measure of SVM and second indicates F-measure of MNNB. Table 3 shows performance of classification.

Bot users' tweets demonstrate more successful results than normal user's tweets by using tweets that have more than two and three terms. However, F-measure values of bot users' tweets are higher than F-measure value of normal users, classification result of normal users' tweets is higher than classification result of bot users' tweets where SVM and tweets that have more than four terms are used for classification.

Table 3 shows that bot users' tweets are more valuable than normal users' tweets in terms of their categorical information. Contents of bot users reflect their own category more than contents of normal users.

Table 3 Classification Results, F-measure Values (%)

SVM|| MNNB	Term Count Threshold Values					
	>2		>3		>4	
Bot Users	82.8	95.2	87.2	96.9	89,9	97.9
Normal Users	78.4	86.7	84.2	92.8	91.6	95.7

According to the classification performance results, Multinominal Naive Bayes outperforms than Support Vector Machines with any given threshold value and user type. Figure 3 shows the results of Table 3.

Figure 3 and Table 3 shows that using Multinominal Naive Bayes as classifier and tweets of bot users as test data gives the best classification results. Choosing of a classifier affects classification results more than choosing of different types of users' content. Classification performance is also increased by term count threshold value. Selecting of tweets which has more than four terms gives the best classification results with any given classifier. It proves that if a tweet consists of more terms, this makes tweet valuable as test data.

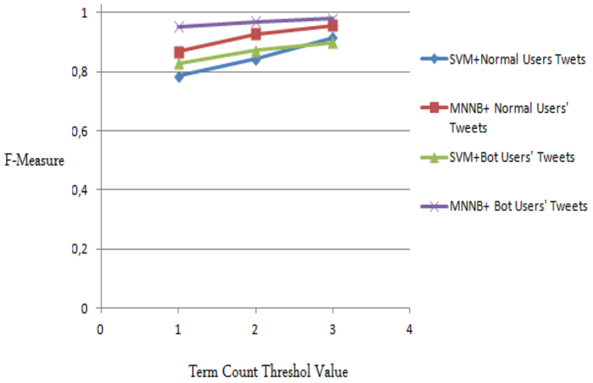

Figure 3. Classification Results

Rate of correctly classified data is changeable from class to class. Normal users whose categorical information is sport supply more categorical tweets. Rate of correctly classified data is higher than data of other normal users whose categorical information isn't sport. Bot users whose categorical information is economy supply more categorical tweets. Rate of correctly classified data is higher than data of other bots users whose categorical information isn't economy. Both normal users and bot users whose categorical information is technology has the lowest rate of correctly classified data.

5. CONCLUSION

In our study, we want to check up-to-dateness of users' tweets with using RSS feeds and it is also intended to measure how users reflect their categories.

Bot users' content is more categorical than normal users' content. Classification performance of bot users' tweets are higher than normal users' tweets. 97,9% F-measure value can be assumed as good result for microblogs. Microblogs consists of abbreviations and nonsense words because of character limitation so we use only training feature set as complete feature set. In the future, we can collect more than 3337 RSS news feeds for training for precision of classification. After putting term count threshold for tweets it decreases number of tweets which has more terms than threshold values so we can get more tweet for precision of classification.

Working on content mining of microblogs is popular recently. Microblogs reflects microbloggers' thoughts and field of interests. Companies observe content of microbloggers for marketing. Police department also follows contents of microbloggers. Police department observe microbloggers' thoughts and action with using contents of microblogs. Activities of terrorism or crime can be distinguished by this observation. To sum up content mining of microblogs can be used in different areas. Popularity of microblogs is increasing rapidly so works on content mining of microblogs are important for all these different areas.

6. REFERENCES

[1] Degirmencioglu, E. A., *Exploring Area-Specific Microblogger Social Networks*, Master's Thesis, Bogazici University, 2010

[2] Yurtsever, E., *Sweettweet:A Semantic Analysis For Microblogging Environments*, Master's Thesis, Bogazici University, 2010

[3] Akman, D. S., *Revealing Microblogger Interests by Analyzing Contributions*, Master's thesis, Bogazici University, 2010

[4] Aslan, O., *An Analysis of News On Microblogging Systems*, Master's Thesis, Bogazici University, 2010

[5] Pilavcılar, I. F., *Metin Madenciliğiyle Metin Sınıflandırma*, Master's Thesis, Yıldız Teknik Universitesi, 2007

[6] Güç, B., *Information Filtering on Micro-Blogging Services*, Master's Thesis, Swiss Federal Institute of Technology, Zurich, 2010

[7] Chua, S. "The Role of Parts-of-Speech in Feature Selection", *The International Conference on Data Mining and Applications-IAENG*, 2008

[8] Masuyama, T. and Nakagawa, H., "Applying Cascaded Feature Selection to SVM Text Categorization", *The DEXA Workshops*, pp. 241-245, 2002

[9] Lan, M., Sung, S. and Low, H., "A comparative study on term weighting schemes for text categorization", *Neural Networks-IJCCN'05*,.IEEE International Conference, 2005

[10] Leopold, E. and Kindermann, J., Text categorization with Support Vector Machines. How to represent texts in input space?, Machine Learning, 2002

[11] Yang, Y. and Pedersen, J., "A Compartive Study on Feature Selection in Text Categorization", *The Proceedings of ICML-97*, 1997

[12] Zheng, Z., and Srihai, R., "Optimally Combining Positive and Negative Features for Text Categorization", *ICML Workshop*, 2003

[13] Yang, Y. and Liu, X., "A re-examination of text categorization methods", *The Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval* (Berkeley, US), pp. 42–49, 1999

[14] Joachims, T., "Text Categorization with Support Vector Machiness: Learning with Many Relevant Features", *The European Conference on Machine Learning (ECML*), 1998