Journal of Theoretical and Applied Information Technology
15th June 2017. Vol.95. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MUSIC INTEREST CLASSIFICATION OF TWITTER USERS USING SUPPORT VECTOR MACHINE

[1]YUSRA, [1]MUHAMMAD FIKRY, [2]BAMBANG RIYANTO TRILAKSONO, [3]RADO YENDRA, [4]AHMAD FUDHOLI

[1]Department of Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim (UIN Suska) 28293, Pekanbaru, Riau, INDONESIA
[2]School of Electrical and Informatics Engineering, Bandung Institute of Technology Bandung, Indonesia, INDONESIA
[3]Department of Mathematics, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim (UIN Suska) 28293, Pekanbaru, Riau, INDONESIA
[4]Solar Energy Research Institute, Universiti Kebangsaan Malaysia, 43600 Bangi Selangor, MALAYSIA
E-mail: [1] usera84@yahoo.com, [1] mfikry1980@yahoo.com, [2] briyanto@lskk.ee.itb.ac.id, [3] rado.yendra@uin-suska.ac.id, [4] a.fudholi@ukm.edu.my

ABSTRACT

Interest determination in music is very beneficial for business communities such as social network advertiser, music studio rental, musical instrument sales, and music concert promoter. This research discusses the classification of music interest of Twitter users based on their tweets in Bahasa Indonesia (Indonesian language). We classify tweets into three music genre categories (jazz, pop, or rock) and three sentiments (positive, negative, or neutral) using Support Vector Machine (SVM). Tweet text classification includes user text, retweet text, mention, hashtag, emoticon, and link (URL). Preprocessing is initiated with word segmentation, removal of symbol and numeric character codes, stemming, word normalization, removal of stopwords, and searching DBpedia for some important words that does not have basic words. This research use dataset of 450 tweets. By generating SVM model on training process that use 360 tweets, Gaussian RBF kernel, and 10-fold cross validation, a pair of parameter (C=0.7, γ=0.9) for music genre category and a pair of parameter (C=0.7, γ=0.8) for sentiment were obtained. The testing process use 90 tweets and resulting the best accuracy for music genre category (96.67%) and the best accuracy for sentiment (86.67%).

Keywords: *Music Genre Category, Sentiment, SVM, Tweet*

1. INTRODUCTION

Twitter as micro-blogging is a blog writing service with a limited size of up to 140 characters message called tweet. People use Twitter to obtain and share information. As social media, Twitter has become a written media to express interests in various domains using unstructured language. Tweet consists of one or more elements: user text, retweet text, and entities, such as mention, hashtag, emoticon, and link. All links posted in tweet are shortened. This research consider link as an important entity in order to understand the full text of the tweet. The link is visited to get its content.

Training and testing is done in two steps. First, the classifier is trained for music genre category, and then the classifier is trained for music genre sentiment. A model was built for each training step and used in the testing process respectively. The results of this research are classification of Twitter users' interests in music genre categories (jazz, pop, or rock) and their sentiments (positive, negative, or neutral). Various benefits are expected from this research, for example in the corporate world of advertising in the social network (such as music studio rental and musical instrument sales). Based on this classification, business owners are expected to monitor and take strategic steps in their business development efforts.

2. RELATED WORK

Many researches have been conducted in text classification, in particular tweet text classification. Thongsuk et.al. [10] has classified tweet into three business types, i.e., a airlines, food, and computer technology. Asur and Huberman [1] used Twitter to predict box-office revenues for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

movies and achieved 97% accuracy. Twitter was used to monitor the U.S. presidential debate in 2008, Diakopoulos [3]. Some text classification researches for sentiment analysis in Indonesian language have also been conducted. Nur and Santika [6] used tweets as dataset and SVM as classification method, obtained 73.07% accuracy for mobile phone brand. Naradhipa and Purwarianti [5] conducted sentiment classification for product or service companies based on 180 data taken from Facebook, and achieved 86.66% accuracy using SVM method. Tweet classification for traffic jam in Bandung was conducted by Rodiansyah, with SVM method, which achieved 92% accuracy from 100 tweets [9]. Classification process from some related work tends to perform classification limited to a particular tweet entity, as Go [4] and Rao [8] who used emoticon for sentiment classification. Pang used hashtag for sentiment classification and conducted a comparison of several classifier methods [7]. Bifet removes mentions, URLs and emoticons from tweets [2]. Research by Nur [6] and Rodiansyah [9] also reduced tweet entities by removing URLs, mentions, and hashtags from tweets. Technically, tweet has multiple entities, i.e. media, URL, mention, hashtag, emoticon and retweet. Each entity can be used for text classification. For example, to post longer than 140 characters on Twitter called extended tweet, various tools can be used to include a link so people can continue reading the full message.

In this research, tweet entities: URL, mention, hashtag, emoticon and retweet are taken into consideration to improve classification accuracy. The shortened URL will be expanded and visited to acquire additional text, which can be either the full message of tweet, or web page's title, description and content. The classification of Twitter user using SVM is based on music-related tweets and its sentiments.

3. MUSIC INTEREST CLASSIFICATION

Figure 1 shows the steps for music interest classification of Twitter users using SVM.

3.1. Preprocessing

Efforts in labeling a music genre category refer to music ontology research. To ensure that a tweet is labelled correctly, we used keywords related to music genre (jazz, pop, and rock).

Unlike category, sentiment was labelled automatically using existing unigram word features of positive and negative sentiment. Automatic labeling is based on rules of addition and subs traction of numbers. Each tweet initially considered as neutral sentiment with zero value (0). If the tweet contains a positive sentiment word or emoticon, then the value plus one (+1), or if the tweet contains a negative sentiment word or emoticon then the value is minus one (-1). If the final value is positive, then the tweet is labeled a positive sentiment. If the final value is negative, then the tweet is labeled a negative sentiment.

The labeling also influenced by negation words, such as 'tidak', 'belum', and 'jangan'. If the positive sentiment word is preceded by a negation word, then the word becomes negative sentiment word, for example 'tidak baik' will have a negative sentiment meaning.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

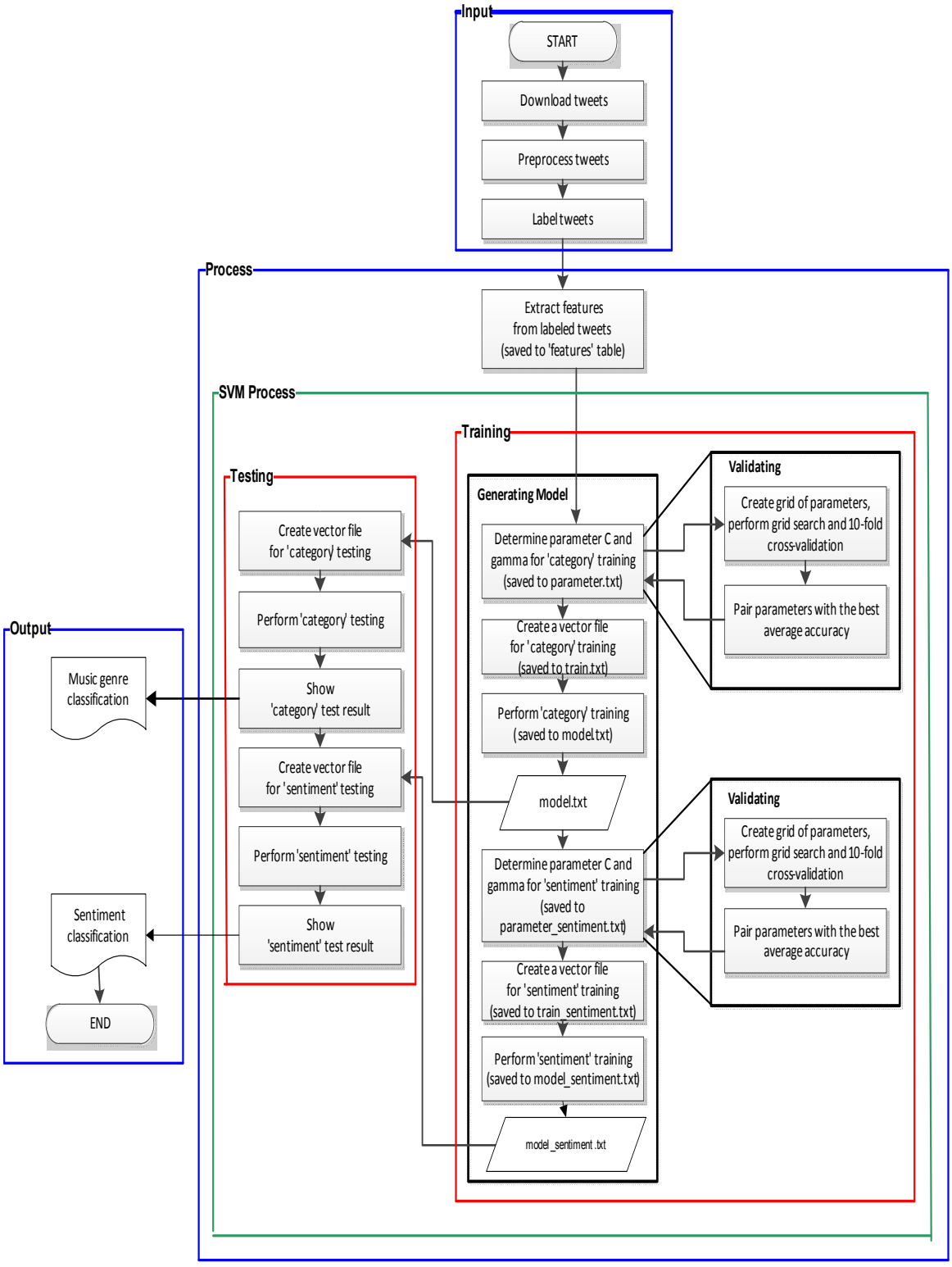

Figure 1: Steps for Music Interest Classification of Twitter Users

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Preprocessing steps for each tweet are as follows:

1. Find links. They can be grouped into two types.
   a. A link points to web pages which contain the full text of the tweet. The full text is used as the tweet instead of the shorter one.
   b. A link points to a web page which contain additional information. Extracted text from title, description and summary content of the web page are used as a replacement for the link.
2. Perform word segmentation by using regular expression.
3. Find words starting with: '@', 'RT', and '#'.
   a. Characters allowed after '#' are upper and lower case, hyphen and underscore.
   b. Remove characters: '@', 'RT', and '#'.
4. Find emoticons. Each emoticon is translated into words that represent emotion and its sentiment (positive, negative, or neutral), then the emoticon is removed.
5. Remove symbols or numeric character codes.
6. Perform word normalization by changing common abbreviations, such as 'bgmn' to 'bagaimana'.
7. Remove stopwords, such as: 'akan', 'antara', 'kemudian', and 'sebagai'.
8. Perform stemming based on the official dictionary of the Indonesian language, called KBBI.
9. If the word does not have a basic word, then check whether the word is part of a resource, such as music concert location 'Bali', Indonesian singer 'Agnes Monica', category name 'Jazz'. We facilitated data retrieval by querying resources from external web sites, such as DBpedia.

Tweet sample before preprocessing:

```
Hooray!! :)
@SahabatSetiaFC: Gemilang- ka
@andienaisyah posisi 1 di @987genfm
sekarang. Makasih gen fm :) #jazz
http://t.co/M9u8yh6Lg4
```

Exspansion result from http://t.co/M9u8yh6Lg4 can be seen as follows :

```
Tampil di Java Jazz, Andien Luncurkan
Album Baru - Berita Liputan6
Setelah dua tahun tak merilis album,
Andien akhirnya kembali. Penyanyi yang
sudah belajar musik sejak umur tiga
tahun ini, menelurkan album ke-5 nya
yang bertajuk
```

Tweet result after preprocessing:

```
gembira gemilang posisi gen gembira
sahabatsetiafc andienaisyah 987genfm
jazz
tampil java jazz andien luncur album
berita rilis album andien nyanyi musik
umur album tajuk
```

3.2. Feature Extraction

Feature extraction for music genre category was done automatically and its sentiment was done manually by using existing positive and negative sentiment words.

As a guide for labeling music genre category, we prepared features manually involving entities from music ontology research: `mo:Genre`, `mo:MusicArtis`, `geo:SpatialThing`, `time:TemporalEntity`. Feature list of 4 ontology entities was compared to feature list of 5 feature extraction methods (Document Frequency, Inverse Document Frequency, Mutual Information, Information Gain, CHI2). Based on feature extraction results, DF method with the highest matches number was obtained. Here is table (Table 1) of its feature distribution.

Table 1: Feature Distribution with DF Method

Occurrence Frequency	1-5	6-10	11-15	16-20	21-25	36-40	41-45	46-50	51-55	>=56
Jazz feature	281	15	3	1	2	2	0	0	0	2
Pop feature	305	23	5	2	1	0	0	0	1	0
Rock feature	391	11	5	1	1	0	0	0	0	1

According to Table 1, it can be seen that the graph starts to look relatively constant at the intervals 11 up to 20. Hence in this research, feature extraction by DF method will be performed, with a threshold between 11 and 20. From this feature extraction results, there were 12 features obtained for music genre category and 320 features for its sentiment.

3.3. Discussion

This research emphasizes on the SVM application for tweet text classification based on music genre and its sentiment. The number of test data is 360 tweets with balanced composition (120 tweets for each music genre).

Test data was divided into 10 subsets, where each subset has the same number of 36 data with a balanced data for each music genre. These subsets would be used as training and testing data according to the 10-fold cross validation method.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SVM training process using Gaussian RBF function requires parameters C and γ. To find the best value for C and γ, 10-fold cross validation method was used. We manually gave some values to determine the best C and γ, which are C (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1) and γ (0.7, 0.8, 0.9).

With 10-fold cross validation method, training and testing process for dataset were conducted. The process aims to construct a classification model and calculate the SVM accuracy to predict labeled test data. The best classification model is if it reaches the highest average accuracy when applied to test the data with the best value of C and γ. The best values of C and γ are when it reached the highest average accuracy when applied to SVM classification.

Accuracy is the comparison of the number of test data already correctly predicted and the number of the entire test data. The average of the accuracy is the average values of the accuracy in each pair of C and γ. Accuracy average as the testing result of each pair of parameter value C (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1) and γ (0.7 ,0.8, 0.9) can be shown in the table below:

Table 2: Accuracy Average for Test Category (C,γ)

C \ γ	0.7	0.8	0.9
0.1	54.44%	63.89%	70.28%
0.2	54.44%	63.89%	70.28%
0.3	54.44%	63.89%	70.28%
0.4	58.89%	63.89%	70.28%
0.5	86.39%	86.67%	86.67%
0.6	86.39%	86.67%	86.67%
0.7	86.39%	86.39%	86.67%
0.8	86.39%	86.39%	86.67%
0.9	86.39%	86.67%	86.67%
1	86.39%	86.39%	86.67%

*Table **Error! No text of specified style in document.**: Accuracy Average for Test Sentiment (C, γ)*

γ \ C	0.7	0.8	0.9
0.1	69.44%	80.56%	80.56%
0.2	68.33%	80.83%	80.56%
0.3	68.33%	80.83%	80.56%
0.4	68.33%	80.83%	80.56%
0.5	68.33%	80.83%	80.56%
0.6	68.33%	80.83%	80.56%
0.7	68.33%	80.83%	80.56%
0.8	68.33%	80.83%	80.56%
0.9	68.33%	80.83%	80.56%
1	68.33%	80.83%	80.56%

Based on Table 2, the highest average accuracy for category obtained was 86.67% of several pairs (C, γ). According to Table 3, the highest average accuracy for sentiment obtained was 80.83% of several pairs (C, γ). If there are several parameters (C, γ) with the same highest accuracy values, the program will randomly determine parameters (C, γ) with the highest accuracy. In this case, we get the best output value of parameter (C, γ) for category, which is (C=0.7, γ =0.9) and for sentiment (C=0.7, γ =0.8).

Accuracy for each 10-fold cross validation on the category test with (C=0.7, γ =0.9) and for sentiment test with (C=0.7, γ =0.8) can be shown in Table 4 below:

Table 3: Category and Sentiment accuracy (10-fold cval)

Fold	Category accuracy (C=0.7, γ =0.9)	Fold	Sentiment accuracy (C=0.7, γ =0.8)
1	86.11%	1	77.78%
2	91.67%	2	66.67%
3	88.89%	3	88.89%
4	80.56%	4	83.33%
5	75.00%	5	80.56%
6	86.11%	6	86.11%
7	75.00%	7	77.78%
8	94.44%	8	80.56%
9	94.44%	9	75.00%
10	94.44%	10	91.67%
Average	86.67%	Average	80.83%

Referring to the research background that involved hashtag and link, we implemented classification comparison without involving the two entities. This accuracy comparison also involved some feature weighting methods. Accuracy comparison for testing the category and sentiment involving 5 tweet entities is shown in Table 5. Accuracy comparison for testing the category and sentiment involving 3 tweet entities is shown in Table 6.

Table 5: Accuracy comparison (5 entities)

Classi-fication	Feature Weighting Method						
	Binary	df	Idf	Tf	tf-idf	log tf	log tf-idf
Category	82.22%	96.67%	85.56%	81.11%	84.44%	63.33%	60.00%
Sentiment	78.89%	86.67%	70.00%	80.00%	71.11%	64.44%	64.44%

Table 6: Accuracy comparison (3 entities)

Classi-fication	Feature Weighting Method						
	Binary	df	If	Tf	tf-idf	log tf	log tf-idf
Category	82.22%	88.89%	85.56%	82.22%	85.56%	62.22%	85.56%
Sentiment	80.00%	86.60%	71.00%	81.11%	73.33%	65.56%	65.56%

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Based on the two comparisons above, it is shown that the 7 feature weighting method is used and then gives an accuracy average over 60%. This indicates that the model generated by training process has been working properly. From the accuracy comparison of 7 feature weighting methods, it can be seen that DF method has the highest accuracy value, involving either 3 or 5 tweet entities.

For category and sentiment classification test resulted, a constructed model can classify the category (jazz, pop, and rock) and sentiment (positive, negative, or neutral), but the feature sentiment still needs a combination with an automatic feature extraction as several features in the collection manually extracted cannot yet optimally represent the features and this makes classification different from the target class.

Summary of SVM classification result using DF method and involving user text, retweet, link, mention, emoticon, and hashtag, can be shown in the following Table 7.

Table4: SVM classification result

Category	Jazz	Pop	Rock
Negative	1.11%	0%	0%
Neutral	23.33%	24.44%	27.78%
Positive	7.78%	10%	5.56%

4. CONCLUSION AND FUTURE WORK

This paper presents music interest classification of Twitter user based on their tweets. A program for classification of Twitter users interest (music genre category and sentiment) has been developed by using SVM classification model based on dataset of 450 tweets. From training on 360 data for constructing the SVM model with Gaussian RBF kernel, the parameter values for category (C=0.7, γ=0.9) and sentiment (C=0.7, γ=0.8) were obtained. From testing on 90 data and comparison of 7 feature weighting methods, DF method gives the highest accuracy of 96.67% for music genre category and 86.67% for sentiment. It shows that testing involving link and hashtag can improve accuracy by 7.78% for music genre and 0.07% for sentiment. This accuracy improvement is still low, because the number of link expanded is less than 15% of the dataset. However, this research proves that using link and hashtag can improve classification accuracy.

Accuracy for music category is higher than that of sentiment may be caused by automatic feature extraction process and labeling process is guided by the ontology research in music, in order to obtain features whose appearance ranking is already measured.

Sentiment analysis results indicate that the number of neutral tweets is higher than positive and negative tweets, up to 75.56%. Out of the three classes of music genre (jazz, pop, rock) shows that the neutral tweets for Rock is higher than that of Jazz and Pop, amounting to 27.78%. This is caused by rock related tweets with many informative sentences or only news. This accuracy result will contribute to the business community, such as advertising on social network (music studio rental and sales of musical instrument), or for promoters who want to organize a music concert. Based on the classification of the music interest of Twitter users, business owner are expected to monitor and take strategic steps for business development.

Suggestions for further research:

1. In the case of multiclass, multiple domains can be added either horizontally or vertically by adding subgenre or class level.
2. Developing features and combining them with the feature of category and other sentiments.
3. Improving training data and making it possible to improve accuracy.

REFRENCES:

[1] S. Asur and B.A. Huberman, "Predicting the Future with Social Media", 2010, 1-8.

[2] A. Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", 2010,1-15.

[3] N.A. Diakopoulos and D.A. Shamma, "Characterizing Debate Performance via Aggregated Twitter Sentiment" *CHI*, 2010, 1-4.

[4] A. Go, R. Bhayani and L. Huang, "Twitter Sentiment Classification Using Distant Supervision" 2009, 1-6.

[5] A.R. Naradhipa and P. Purwarianti, "Sentiment Classification for Indonesian Message in Social Media", *2011 International Conference on Electrical Engineering and Informatics*, 2011, 1-4.

[6] M.Y. Nur and D.D. Santika, "Analisis Senntimen pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine" *Konferensi Nasional Sistem dan Informatika 2011*, 2011, 9-14.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[7] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning", *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002, 79-86.

[8] D. Rao, D. Yarowsky, A. Shreevats and M. Gupta, "Classifying Latent User Attributes in Twitter", 2009, 1-8.

[9] S.F. Rodiansyah and E. Winarko, "Klasifikasi Posting Twitter Kemacetan Lalu Lintas Menggunakan Naïve Bayesian Classification", *IJCCS*, 6(1), 2012, 91-100.

[10] C. Thongsuk, C. Haruechaiyasak and S. Saelee, "Improving Business Type Classification from Twitter Post Based on Topic", *World of Computer Science and Information Technology Journal (WCSIT)*, vol.1 No. 8, 2011, 333-338.