Identifying Psycho-Social Fingerprints in Medical and Engineering Documents

MATEJA VERLIC, GREGOR STIGLIC, PETER KOKOL
Faculty of Electrical Engineering and Computer Science
University of Maribor
Smetanova 17, 2000 Maribor
SLOVENIA

Abstract: Sentiment analysis enables new possibilities for classification of text documents by considering different aspects of sentiment, appraisal and other linguistic elements. Our psycho-social fingerprint approach to the text-classification is based on the General Inquirer dictionary of psycho-social categories and provides instant solution for text classification in quickly changing target domains, such as newsgroups.

Key-Words: text mining, text analysis, sentiment analysis, sentiment classification, psycho-social categories, machine learning

1 Introduction
Text-mining has been used to analyse texts from different sources, including business documents, web pages, news, discussion groups, blogs, e-mails, literature, scientific papers and many more. Automatic analysis and classification of documents has gained its popularity also due to its potential in business intelligence.

On one hand, conventional text analysis includes classification of documents into topics, document clustering, building dictionaries and thesauri for specific domains, and automatic summarization. On the other hand, recent text analysis techniques are more qualitatively oriented, trying to identify several non-topic attributes of the analysed texts. In contrast to quantitative attributes like word frequencies or part of sentence for specific word, qualitative attributes are, for example, used for describing the sentiment in the text.

Sentiment analysis is used to identify writer's sentiment towards the subject and to classify documents into positive or negative [2]. Sentiment classification is usually approached from two different ways. In the first approach, also called a bag-of-words approach, document classifiers are based on the frequencies of the various words in the document. Different methods can be used to select different parts of analysed text for the classification. In the second approach, words are classified into two opposite classes - in positive and negative or in 'good' or 'bad'. After that, overall class score for the document is computed. However, these two approaches are just the most frequent ones. Several other approaches include additional aspects of sentiment analysis by introducing appraisal modifiers [3] combinations of adverbs and adjectives for scoring the strength of sentiment on a topic [4], three dimensions of Osgood et al. [5] or a whole range of other psycho-social factors.

We believe that text documents from specific problem domains display a variety of distinctive features regarding psycho-social factors, which can be thought of as a psychosocial fingerprint.

2 Problem
The primary goal of this study was not detecting the sentiment of analysed documents, but rather to identify the psycho-social fingerprint of documents from two very specific domains – electronics and medicine, and to find out if such psycho-social fingerprint can be used for text classification. We analysed medical and engineering newsgroup documents from English 20 Newsgroup dataset [1]

2.1 Psycho-social fingerprint
Psycho-social fingerprint is defined as a set of psycho-social word categories that differentiates between documents from two different domains. We used General Inquirer's psycho-social dictionary [6] with 182 categories as a pool of categories, from which the final fingerprint was extracted.

The psycho-social dictionary is based on categories from Harvard IV-4 dictionary, Lasswell value dictionary, and some other constructed categories and markers. Each category of the dictionary is a list of words and word senses. For example, the biggest valence categories are Positiv and Negativ. Category Positiv includes 1,915 words of positive outlook without the words for yes,
while category *Negativ* contains 2,291 words for negative outlook not including words for no in the sense of refusal.

### 2.2 20 Newsgroups dataset

The 20 Newsgroups dataset [1] is a collection of approximately 20,000 newsgroup messages. A thousand messages from each of twenty groups were randomly chosen and partitioned by the newsgroup name. This document collection has been widely used as a reference dataset in applications of text classification and text clustering for topic detection research [7-9].

In our experiment we used documents from only two newsgroups: *sci.med* and *sci.electronics*. We used these two groups because they represent very specific interest groups.

### 3 Methodology

We implemented a simple tool in java for classification of messages in two selected groups based on word counts. First, sentences of each message were tokenized. Then tokens presence was checked against the psycho-social dictionary.

For each document a vector of 182 values was stored as an entry for a new dataset. If token was found in one or more categories of the dictionary, word counters for respective categories were incremented. Because messages vary in length, we decided to store calculated relative frequencies (number of words in a category divided by number of all words) instead of actual frequencies.

Finally, when all messages were processed, new dataset of vectors with relative frequencies and additional attribute representing their origin (*sci.med* or *sci.electronics*) was analysed with the Weka toolkit [15].

After classification of this new dataset with different classification methods, we preformed feature selection techniques to obtain psycho-social fingerprint.

#### 3.1 Classification

Weka supports numerous different machine learning methods, but in our case these four different classification methods were used: Naive Bayes, Decision Trees (J48 in Weka), an optimized version of Support Vector Machines (SMO) and Random Forests.

#### 3.2 Naive Bayes

The Naive Bayes classifier is the simplest of probabilistic models based on the Bayes' theorem. This model works under the naive assumption that all attributes of the samples are independent of each other given the class [10, 11], which simplifies the computation.

Although this assumption is false in most real-world tasks, Naive Bayes can outperform more sophisticated classification methods. Other than simplicity, Naive Bayes has additional advantages; it is very efficient, in general robust and has a high accuracy. From this reasons it is often used as a referential method for testing prototypes of new datasets to obtain reference classification accuracy.

#### 3.3 Decision trees

J48 Decision Tree is Weka toolkit implementation of C4.5 decision tree algorithm proposed by Quinlan [16]. It is used for inducing classification rules in the form of decision trees from a set of given examples. We are given a set of records where each record has the same structure, consisting of a number of attribute/value pairs. One of these attributes represents the category of the record. The problem is to determine a decision tree that on the basis of answers to questions about the non-category attributes predicts correctly the value of the category attribute also called class attribute.

#### 3.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) [12] is more advanced classification technique. SVMs are classifying by constructing a hyperplane, which would have maximal Euclidean distance to the closest training samples. If the training data is linearly separable, it is possible to achieve the maximal distance and thus the best generalization of the classifier can be obtained. In other case, when data is not linearly separable, a hyperplane that splits the data with the least error possible is chosen. In our experiment sequential minimal optimization (SMO) algorithm for training a support vector classifier was used [13, 14], which is faster and has better scaling properties than standard SVM.

#### 3.5 Random forests

Random Forests classifier originates from one of the first ensembles of classifiers called bagging [17]. To increase the diversity of classifiers in bagging, Breiman upgraded the basic idea of bagging by combining bootstrapping with random feature selection for decision tree building. Random decision trees created this way are grown by selecting the feature to split on at each node from randomly selected number of nodes. Number of chosen features is set to \(\log_2(k+1)\) as in [18], where \(k\) is the total number of features. Random Forests is ensemble
building technique that works well even with noisy content in training dataset and is considered as one of the most competitive and robust methods that can be compared to bagging or boosting [19]. One of the advantages in random forests is also their simplicity and speed. Computational complexity can be even lower when parallelization of the ensemble building process is used that makes them even more suitable for complex classification problems.

3.2 Identifying psycho-social fingerprint

Classification results provided a set of potential category candidates for the psycho-social fingerprint. Unfortunately a direct extraction of important features is only possible in case of Decision Trees, but this method did not perform good enough. Therefore we identified categories for the fingerprint by using feature selection. A filter based feature selection process was used. This means that all features were ranked according to their discernability regarding to decision class and feature selection technique.

Two different feature selection techniques were used to identify features that contribute the most to the classification: InfoGain feature selection (Information gain based method) from Weka toolkit and SVM based Recursive Feature Elimination (SVM-RFE) [20]. Features, that were common to both feature selection techniques, represent the most common components of psycho-social fingerprint for classification of our newsgroup subset.

4 Results and discussion

In our first experiment, we measured classification accuracy using psycho-social attributes from General Inquirer word groups. There were four different types of classifiers used, representing two classical (J48, Naïve Bayes) and two advanced classification techniques (Random Forests, SMO).

Table 1 shows the results of classification accuracy for four different classification techniques, based on four data sampling techniques.

Table 1 Classification accuracy (%) of four classifiers used

<table>
<thead>
<tr>
<th></th>
<th>J48</th>
<th>Naïve Bayes</th>
<th>Random Forests</th>
<th>SMO</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-fold CV</td>
<td>75.40</td>
<td>78.30</td>
<td>86.30</td>
<td>86.65</td>
</tr>
<tr>
<td>20-fold CV</td>
<td>75.55</td>
<td>78.70</td>
<td>86.00</td>
<td>87.00</td>
</tr>
<tr>
<td>50-fold CV</td>
<td>75.95</td>
<td>78.70</td>
<td>86.25</td>
<td>86.75</td>
</tr>
<tr>
<td>100-fold CV</td>
<td>74.75</td>
<td>78.55</td>
<td>86.75</td>
<td>86.50</td>
</tr>
</tbody>
</table>

Sufficient number of samples in the dataset enables experimenting with different settings when n-fold cross-validation is used. We used cross-validations with 10, 20, 50 and 100 folds to show the stability of classification results when the training to test set ratio changes. All decision trees were built using default Weka settings – i.e. using pruning.

In the second part of our experiment we tried to explain classification results from Table 1. Since there is only one classifier that can be directly used for interpretation, a Decision Tree originating from our Electro-Med dataset is presented in Figure 1. To rank features according to their relevance, a feature selection technique based on information gain was used on the whole dataset. Ten top ranked features from InfoGain feature selection process can be seen in Table 2.

Table 2 Top 10 categories selected by InfoGain feature selection technique

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Feel</td>
<td>49 words describing particular feelings, including gratitude, apathy, and optimism, not those of pain or pleasure.</td>
</tr>
<tr>
<td>2. NonAdlt</td>
<td>25 words associated with infants through adolescents.</td>
</tr>
<tr>
<td>3. WlbPt</td>
<td>27 roles that evoke a concern for well-being, including infants, doctors, and vacationers.</td>
</tr>
<tr>
<td>4. Female</td>
<td>43 words referring to women and social roles associated with women.</td>
</tr>
<tr>
<td>5. IPadj</td>
<td>117 adjectives referring to relations between people, such as &quot;unkind, aloof, supportive&quot;.</td>
</tr>
<tr>
<td>6. WltTran</td>
<td>Wealth transaction, 53 words for pursuit of wealth, such as buying and selling.</td>
</tr>
<tr>
<td>7. WlbTot</td>
<td>487 words in well-being domain.</td>
</tr>
<tr>
<td>8. Think</td>
<td>81 words referring to the presence or absence of rational thought processes.</td>
</tr>
<tr>
<td>9. WlbPhys</td>
<td>226 words connoting the physical aspects of well being, including its absence.</td>
</tr>
<tr>
<td>10. Food</td>
<td>80 words referencing food.</td>
</tr>
</tbody>
</table>

It can be observed that most of the features used in the top nodes of the decision tree were also selected by the InfoGain feature selection algorithm. This is not entirely unexpected as information gain is also a splitting criterion in decision tree that was used to obtain classification accuracies in the range from 74.75% to 75.95%. However, it would be more interesting to get a list of features that contributed the most to the classification abilities of SMO classifier. Weka offers a reliable and well proven feature selection method that is also based on SVM – Recursive Feature Elimination (SVM-RFE). Top categories according to results of SVM-RFE based feature selection can be seen in Table 3. It can be observed that there are significant differences in both lists. Only five categories can be found in both
lists (Feel, WlbTot, Female, Food, WlbPhys). Most of categories seem to offer a sensible explanation of their meaning and discernability power for observed classification problem.

Table 3: Top 10 categories selected by SVM-RFE feature selection technique

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. WlbTot</td>
<td>487 words in well-being domain.</td>
</tr>
<tr>
<td>2. Econ@</td>
<td>510 words of an economic, commercial, industrial, or business orientation,</td>
</tr>
<tr>
<td></td>
<td>including roles, collectivities, acts, abstract ideas, and symbols, including references to money. Includes names of common commodities in business.</td>
</tr>
<tr>
<td>3. Female</td>
<td>43 words referring to women and social roles associated with women.</td>
</tr>
<tr>
<td>4. BodyPt</td>
<td>A list of 80 parts of the body.</td>
</tr>
<tr>
<td>5. Doctrin</td>
<td>217 words referring to organized systems of belief or knowledge, including those of applied knowledge, mystical beliefs, and arts that academics study.</td>
</tr>
<tr>
<td>6. Food</td>
<td>80 words referencing food.</td>
</tr>
<tr>
<td>7. MALE</td>
<td>56 words referring to men and social roles associated with men.</td>
</tr>
<tr>
<td>8. Feel</td>
<td>49 words describing particular feelings, including gratitude, apathy, and optimism, not those of pain or pleasure.</td>
</tr>
<tr>
<td>10. WlbPhys</td>
<td>226 words connoting the physical aspects of well being, including its absence.</td>
</tr>
</tbody>
</table>

Another interesting comparison of both feature selection methods is shown in Fig.2 and Fig.3, where histograms of two top ranked features according to two observed feature selection methods are shown. Histograms clearly show that SVM-RFE based features are widely dispersed and more useful for classification in comparison to InfoGain based selection, where values lie tightly together and defining a classification line can be very sensitive. Observing the histograms, we can notice that samples from both classes cannot be separated by using a simple linear discrimination.
6 Conclusion and Future Work

This paper shows that a proposed classification procedure would be suitable as an instant solution to text classification in quickly changing target domains, such as newsgroups. Because only small number of classification and feature selection techniques was used; there is certainly a room for experimenting with more complex classifiers or better feature selection methods. Another idea would be to use feature selection prior to classification as a pre-processing step and execute it during cross-validations. This could improve classification accuracy because unnecessary features would be removed before the classification process.

In our future work we plan to test identified psychosocial fingerprint approach with datasets from different sources but similar domains, for example with web pages with content from medical and engineering domain.

It would also be interesting to compare fingerprints of other newsgroups from 20 Newsgroups dataset with the fingerprint identified during our experiment.

References: