Sentiment Classification of Hotel Reviews in Social Media with Decision Tree Learning

Stanimira Yordanova
Ph.D. Student at
University of National and
World economy
Studentski grad,
Sofia 1000, Bulgaria

Dorina Kabakchieva
Assist. Professor, Ph.D. at
University of National and
World economy
Studentski grad,
Sofia1000, Bulgaria

ABSTRACT
The aim of this paper is to present an approach for prediction of customer opinion, using supervised machine learning approach and Decision tree method for classification of online hotel reviews as positive or negative. The preliminary extraction and preparation of the data used in the research are described. Three classification models are generated for three different data sets - balanced and unbalanced training sets with two schemes of filtering frequent and infrequent words in the attribute list. The results from the classifier evaluation are compared and discussed. The three classification models are also applied on new unseen data for predicting opinion of hotel guests. The achieved results reveal that the most accurate prediction is achieved when applying the model generated from the balanced training set with filtering rare words.

General Terms
Sentiment classification, Hotel Industry, Online reviews

Keywords
Sentiment classification, supervised machine learning, decision tree

1. INTRODUCTION
At present, one of the main challenges a business organization is facing is to gather and use, in cost-effective and timely manner, all relevant information in order to acquire reliable and meaningful insights to support effective decision-making process. Business Intelligence (BI) Systems provide tools, methods, and technologies, and are a reliable instrument to respond to such challenges, therefore more businesses realize the value and the indispensability to use them in their decision making. Traditional BI systems process structured data, coming from various sources; apply advanced analytical tools and visualize the results interactively to help business users in discovering new beneficial business knowledge. Advanced BI systems also process unstructured data which not only come from organizational inner sources (emails, reports, etc.) but from social media as well. Very popular definition of a social media is "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content" [1] provided by Kaplan and Haenlein in 2010. Reviews, comments, blogs, microblogs, and forum posts are user generated content in the form of unstructured text data, published on Social media and expressing opinions on topics, products, services, people or organizations. Sharing experience on using products or services in the Social media sites increases the volume of unstructured data from which new business knowledge can be extracted. For most of the industries which are offering products or services, understanding customer experience becomes crucial for improving corporate performance and remaining competitive on the market. Reviews are very popular among hotel customers and extremely important for the hotel industry. On one hand, hotel guests share their experience of using hotel services on review sites like TripAdvisor and Booking.com, thus influencing both, booking decisions of future hotel guests and the online hotel reputation. On the other hand, negative social media feedback is a valuable source for guiding improvements in the provision of hotel services while maintaining positive online hotel reputation has direct impact on decision for purchasing hotel services. Management of online reputation implies monitoring of positive and negative reviews, published on different social media sources. Some of the review sites like Booking.com contain positive and negative feedback labeled by the authors' review while others like TripAdvisor.com do not provide such option.

The first challenge when analyzing hotel guest responses is to predict the opinion of an author, expressed in the hotel review, by classifying it as positive or negative feedback. It can be addressed by application of sentiment analysis. The second challenge is to visualize the results in order to extract business knowledge, achieved by using Business Intelligence tools.

This paper focuses on the implementation of a methodology for sentiment classification and prediction of opinion of hotel guests, published in the review sections of hotel travel and accommodation sites. The generated models for prediction of online hotel reviews are presented and compared. Conclusions from the experimental cases are also provided at the end, as well as outlines of the future research activities that will be performed.

2. PROBLEM DEFINITION
Discovering valuable knowledge from reviews requires, as a first step, to structure the unstructured user generated content, to analyze the data and to visualize the results in a way to be understood and used by business users.

Text mining includes methods and tools for structuring and analyzing unstructured text content generated by hotel reviewers. The process of knowledge discovery from text content of hotel reviews covers (1) gathering and organizing text documents in a corpus; (2) using different techniques for text preprocessing, aimed at structuring the data and extracting key representative features and (3) extracting knowledge using data mining algorithms. In case of using classification algorithms for sentiment prediction expressed in a document, sentiment analysis is applied.

According to B. Liu (2012) "sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" [2]. A main task of sentiment analysis is to classify text or part of the text as positive, negative (subjective classification) or neutral. Classification can be performed at document, sentence or aspect levels.

User generated text content in hotel reviews, published on review sites, can be presented as pros (+) and cons (-), e.g.:

Pros (+): Excellent location, good reception

Cons (-): Uncomfortable bed, awful breakfast

Or as free text, e.g.:

Very helpful staff. Lounge with refreshments made the trip very relaxing. Free wifi also a plus. Would definitely stay there again.

In the first case the text is classified by the author, while in the second, the text is not classified as positive or negative.

The proposed model for classification aims to predict the positive or negative sentiment of hotel reviews which are not previously classified, thus determining the author opinion. The classification is performed at document-level, using supervised machine learning and a decision tree method. The data used for the classification model generation is extracted from Booking.com and TripAdvisor.com. The following research activities are aimed at improving prediction by experimenting with the classification model parameter values. Improving the classification is important for both, the negative reviews as main indicators for identification of problems in hotel service delivery, and the positive, as a tool for maintaining high positive online reputation.

3. RELATED WORKS

Machine learning approach requires extracting the best attributes for classification and applying algorithms for classification. Most of the research literature focus on the application of Naïve Bayes, Support Vector Machines (SVM), Decision Tree, k-NN, Neural Networks and etc in the classification. In relation to attribute extraction, the research focuses on frequency based extraction, application of unigrams, N-grams, POS tagging or combination of all these techniques, Information gain, and CHI statistics. Most of the experiments are provided by using movie or product reviews. There is limited research applying classification of online hotel reviews.

Gautami Tripathi and Naganna S (2015) proposed a model for sentiment analysis of movie reviews testing four different feature selection schemes, using Naïve Bayes and Linear SVM. The results showed that linear SVM gives a maximum accuracy of 84.75% for TF-IDF scheme. [4]

M. Bilal et al (2016) conducted on Roman Urdu data set by using Naïve Bayes, Decision Tree, and k-NN. The results showed that Naïve Bayes algorithm performed best with 97.33% accuracy, compared to the Decision Tree (94.67%) and k-NN (86.67%). [5]

V. Elango et al (2014) used hotel review data from TripAdvosor.com to explore performance Naive Bayes, Support vector machine, Laplace smoothing and Semantic in the classification of hotel reviews. To extract the frequent words from the reviews Term Frequency and Inverse Document Frequency are used. The results showed that Naïve Bayes model performed better with achieved higher accuracy of 79.12% compared to SVM with achieved accuracy of 75.29%. [6]

P. Kalaivani et.al (2013) applied sentiment classification techniques on movie reviews and compared SVM, Naive Bayes, and k-NN. SVM performed better than Naive Bayes and k-NN with accuracy of at least 80%. [7]

Sharma and Dey (2012) investigated five feature selection methods - Document Frequency, Information Gain, Gain Ratio, Chi Squared, and Relief-F) and sentiment feature lexicons on classification of movie reviews, using SVM. Best results were achieved using Gain Ratio for a large number of sentimental features selection (more than 5000 features). [8]

H. Sui, et al (2003) used SVM and Decision Tree to classify product reviews as positive and negative. They investigated five different approaches - unigrams, part-of-speech tagging, association rules, use of negation, and use of WordNet synsets – in syntactic and semantic processing of text. The data set contained 1,200 product reviews for training, and 600 for validation. SVN with unigrams approach reached an accuracy rate of 81.3%. The use of WordNet synsets obtained better result of 81.7%. Decision Tree induction was used to generate a list of indicative words that can identify the polarity of articles. [9]

Pang, Lee and Vaithyanathan (2002) conducted sentiment classification using Naïve Bayes, Support Vector machine, and Maximum Entropy, augmented with using also n-grams. The results revealed that the SVM performed better as compared to others [10].

The most applied algorithms for classification are Naïve Bayes and Support Vector Machines (SVM). Beside them, other classification algorithms are also used like Decision Tree, k-NN, Neural Networks.

This paper investigates the performance of three Decision tree classification models using unbalanced and balanced training set with two schemes filtering most frequent words and rarely used words.

4. METHODOLOGY

The proposed model for classifying and predicting opinion of hotel guests, published on the review sites is implemented by following the Cross Industry Standard Process for Data Mining (CRISP – DM), one of the most popular and widely used framework for the implementation of data mining projects. The CRISP-DM approach includes six main steps - Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment [11]. For the purposes of the research, the CRISP – DM framework is adapted to the processing of unstructured text data, which requires structuring by applying text preprocessing techniques during data preparation [12]. The resulting methodology is presented in Figure 1.

The data mining instrument used in the research for is RapidMiner.

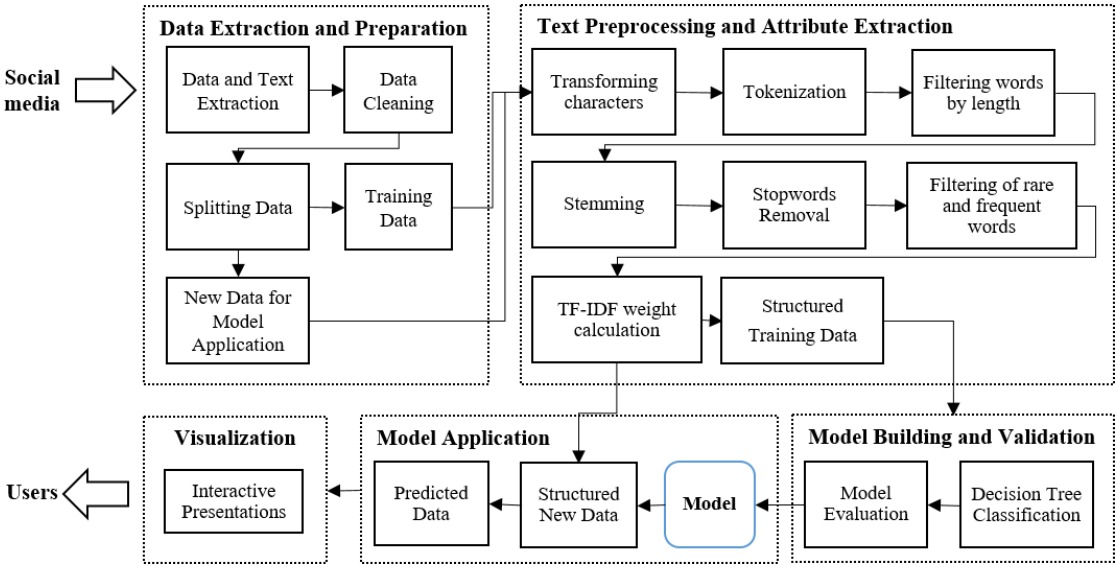

Fig 1: Methodology for Sentiment Classification of Hotel Reviews

4.1 Data Extraction and Preparation

Booking.com and Tripadvisor.com are the Internet sources from which the research data is extracted, using import.io. It is a web-based platform for extracting data from websites by navigating a website and teaching the application Extractor to extract data by training it which data to be extracted from the selected page links. During classification, only user-generated content is used where a particular opinion about hotel services is expressed. It is unstructured text data that needs to be prepared for classification purposes.

During data preparation, the text data is preprocessed and unstructured text is turned into a structured format. Firstly, all characters in the example set are transformed to lower case. The text is split into a sequence of tokens, consisting of one single word. Then all tokens which equal a Stopwords from the English built-in Stopwords list are removed from the text. Stopwords are noise words that increase the classification error on new data. The tokens are filtered based on their length, with minimum 3 and maximum 99 characters. Finally, stemming is performed by Porter stemming algorithm applying an iterative, rule-based replacement of word suffixes intending to reduce the length of the words until a minimum length is reached [13]. The tokens are used to generate word vectors numerically representing each example and TF-IDF score of each available word is calculated. The results from preprocessing are in the form of a term document matrix, where each token is now an attribute in a column and each review is an example in a row. The values in the cells are the calculated TF–IDF scores for each word in the word vector creation process. The generated word attributes and their TF–IDF scores are used by the classifier.

4.2 Attribute Extraction

The aim of attribute extraction and selection is to select a subset of words occurring in the training set and using only this subset as attributes in text classification. Attribute selection decreases the vocabulary size by eliminating noise or irrelevant words and increases classification accuracy. There are various attribute selection methods based on mutual information, chi-square, Information gain or Gain ratio, frequency-based feature selection. The decision tree algorithm incorporates attribute selection by using Information gain as a criterion for evaluation of attribute importance; during preprocessing, frequency-based feature selection by calculation of Term Frequency-Inverse Document Frequency (TF-IDF) is applied to both, training and testing data sets. It diminishes the weight of terms that occur very frequently in the data set and increases the weight of terms that occur rarely. TF-IDF is calculated through the following formula:

$$TF - IDF = TF_{t,d} * IDF_t$$

where *t* is a term(attribute) in a document(example) and *d* is given document(example), where *t* appears.

Term Frequency (TF) is the ratio between the number of times a term *t* appears in a given document *d* (n_t) and the total number of terms in the document (*n*).

$$TF - IDF = \frac{n_t}{n} * \log_2 \frac{N_d}{N_t}$$

Inverse Document Frequency is the ratio between the total number of documents in the corpus (N_d) and the number of documents that contain the term *t* (N_t).

The result from applying only TF is that frequent words have higher TF score and infrequent words - lower TF score. TF-IDF takes into account not only the importance of a word in a given document but also its importance in the entire corpus. This technique decreases the weight of frequent words and increases that of rare words in a corpus. The TF-IDF score of a word increases when the number of times the word appears in a document (TF) increases. If the number of documents that contain a word is increased (the word appears more frequently in the corpus), the TF-IDF score of the word decreases, otherwise increases.

4.3 Classification using Decision Tree

The main goal of the classification model for hotel reviews is to classify them as positive or negative, thus determining the opinion of the hotel guests expressed on the websites. Decision tree algorithm is used to generate a classification model for predicting the values of a target attribute (class or label) based on the values of several input attributes in the training data, used for classification of reviews. In RapidMiner, the decision tree algorithm is similar to Quinlan's C4.5 or CART [12]. The attribute with the Label role is the target variable for prediction. It is used for classification of nominal and numeric data types. As a tree-like model, it has root at the top and it grows downwards. In

each node, an attribute is tested. The nodes split the data into subsets based on data homogeneity. The Decision tree or subtree ends with leaf where a prediction about the target variable is made based on the conditions set forth. The decision tree is generated by recursive partitioning. In general, the recursion stops when all the examples have the same label value, i.e. the subset is pure, or if most of the examples are of the same label value, or when a certain condition is reached. Information gain is selected for measuring the entropy and as a criterion for selecting the attribute for splitting the data. The attribute with the minimum entropy (the highest Information Gain) is selected for each split. The size of the decision tree is restricted to 5 nodes and the recursion stops when a maximum depth of these nodes is reached.

4.4 Model Training and Application Evaluation

The prediction accuracy of the model is evaluated using k-fold Cross-validation. Cross-validation divides the training data set into 10 equally sized, non-overlapping subsets and the model is trained on the first nine sets and tested on the tenth remaining set. The process is repeated 10 times and each time different subset for training is used. The performance values of the model for each iteration are measured and the average is returned. Stratified sampling is applied when building the subsets, in order to build random subsets and ensure that the class distribution in the subsets is the same as in the whole example data set. A confusion matrix is produced as a result of the cross-validation, showing the number of correct and incorrect predictions made by the classification model, compared to the actual target value in the data. The performance values used for comparison are accuracy, positive and negative precision and recall, f-measure, and AUC.

Accuracy is the proportion of the total number of predictions that are correctly predicted. Positive Class Precision is the proportion of positive cases that are correctly identified. Negative Class Precision is the proportion of negative cases that are correctly identified. Positive Class Recall is the proportion of actual positive cases which are correctly identified. Negative Class Recall is the proportion of actual negative cases which are correctly identified. F-measure is a harmonic mean of precision and recall. The area under ROC curve is used as a measure of the quality of classification models. A random classifier that cannot distinguish between the two classes has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1.

5. EXPERIMENTAL SET UP

Following the methodology described in section 4, three experiments are conducted to evaluate the overall accuracy of the classifier and to improve the negative class prediction when applying the model on new unseen data. The details of the experiments are:

1. Unbalanced training set, 4-600: u*nbalanced* training set is used by classifier with filtering words that occur in less than 4 and more than 600 examples.
2. Balanced training set, 4-600: *balanced* training set is used by classifier with filtering words that occur in less than 4 and more than 600 examples.
3. Balanced training set, 20-200: *balanced* training set is used by classifier with filtering words that occur in less than 20 and more than 200 examples.

5.1 Unstructured Data Extraction

The experimental data is extracted from the review section of Booking.com and Tripadvisor.com for a particular 4-star Hotel. Reviews published on both sites are filtered to show only reviews in English. 586 examples from Booking.com and 347 examples from TripAdvisor.com are extracted. The sets contain data about the author, country, reason for stay, rating, room type, date of publication, etc. in the form of user-generated unstructured text content. The data set for model building is an extract from these two sets where user-generated text content from both sets is combined. The data is reorganized into a single spreadsheet consisting of 1519 rows (586 positive and 586 negative from Booking.com, labeled by authors and 347 from TripAdvisor.com, manually labeled) and 3 columns. The first column contains the ID identifying the author. The second column contains the Label and takes two possible values – positive and negative. The third column contains review text, which has been labeled as positive or negative (second column Label). Spell check is performed to improve data quality in the third column. During data examination, different interchangeable variants of some words are discovered, e.g. “wifi, wi-fi, wi fi, internet”, “air-conditioning, A/C, air con, air conditioner“, “tv, television”. One variant of each word has been chosen - "internet”, “air-conditioning“ and “television”.

5.2 Training Data set and Data set for Model Application on Unseen Data

In Booking.com, text content is classified by an author as positive or negative, because the authors have the possibility to express their experience about positive and negative sides of hotel services. Therefore, all of the available 1172 examples are included in the training data set. In TripAdvisor.com, user-generated content is not previously classified by review authors, thus it is not possible to know what is the author’s opinion expressed in the text. One-third of this data (120 records) is included in the training data set. The remaining two-thirds of data (227 records) is used as a data set for model application on unseen data.

In the training set (1292 records in total), 17 examples from TripAdvisor.com contain positive and negative opinion in one example. A decision is taken to divide such reviews into two records – one with the positive comment and one with the negative comment, labeled respectively as positive and negative, thus resulting in increasing the number of training examples to 1309.

5.3 Text Preprocessing

The training data set is loaded into the local repository in RapidMiner. The data set has three attributes – ID, Label, and Review_text. The data type of the ID attribute is numeric, integer, identifying the author of each review, the Label attribute is binominal with two values “positive” and “negative” and in the Review_text attribute is text. Missing values are discovered in column Review_text. The reasons for missing values could be not sharing out negative (in case of missing negative example) or positive (in case of missing positive example) aspects of hotel services because the authors did not have such experience when using the services. The records with the missing values are filtered. As a result, the final training data set contains 1058 examples of which 678 (64%) are labeled as positive and 380 (36%) are labeled as negative, making it an unbalanced training data set. A second training set is also constructed – it is a balanced training set containing 760 examples (380, labeled as positive

and 380 labeled as negative). The classification algorithm is applied to both, the unbalanced and the balanced training sets.

During preprocessing, text data in the Review_text attribute is structured, applying transformation of capital in small letters, tokenization, filtering of Stopwords, filtering of words containing less than 3 letters and stemming. In addition, TF-IDF technique for attribute frequency is applied, as well as filtering of words that occur in less than 4 and more than 600 examples.

When using the *unbalanced* training data set, in the generated attribute list of 481 words, *locat* (stemming form of location/s and located) occurs in 424 examples. It is also the most frequent word which occurs 427 times in examples labeled as positive and only 9 times in examples labeled as negative. This could make it a proper attribute that can be chosen by the decision tree algorithm for splitting and for classifying positive examples. The next most frequent words are *room*, *breakfast*, and *staff*.

When using the *balanced* training data set, the attribute list contains 440 attributes, *room* occurs in 298 examples and 400 times in them, followed by *breakfast* and *locat.*

When using the *balanced* training set with filtering words that occur in less than 20 and more than 200 examples, the attribute list contains 62 attributes. *Locat*, *breakfast*, and *room*, are ignored as these words are the most frequent. In the attribute list, the most frequent word is *good*, occurring in 172 examples, then *help* (stem form of helpful) in 133 examples, followed by *stai* (stay), *comfort* (comfortable, comfort), *clean* and etc.

5.4 Generated Decision Tree Rules

During the training, decision tree rules are generated. The generated decision tree rules resulting from the three classification models – using unbalanced training set, 4-600; balanced training set, 4-600 and balanced training set, 20-200, are presented on Fig.2.

Generated Decision Tree Rules, using unbalanced training set, 4-600

1. if locat > 0.032 and term > 0.094 and great > 0.121 then positive (1/0)
2. if locat > 0.032 and term > 0.094 and great ≤ 0.121 then negative (0/2)
3. if locat > 0.032 and term ≤ 0.094 and view > 0.249 and breakfast > 0.065 then positive (3/0)
4. if locat > 0.032 and term ≤ 0.094 and view > 0.249 and breakfast ≤ 0.065 then negative (0/2)
5. if locat > 0.032 and term ≤ 0.094 and view ≤ 0.249 then positive (406/0)
6. if locat ≤ 0.032 and staff > 0.019 then positive (139/27)
7. if locat ≤ 0.032 and staff ≤ 0.019 and good > 0.086 then positive (31/12)
8. if locat ≤ 0.032 and staff ≤ 0.019 and good ≤ 0.086 and excel > 0.221 then positive (13/2)
9. if locat ≤ 0.032 and staff ≤ 0.019 and good ≤ 0.086 and excel ≤ 0.221 then negative (85/335)

Generated Decision Tree Rules, using balanced training set, 4-600

1. if locat > 0.041 and star > 0.232 and comfort > 0.110 then positive (1/0)
2. if locat > 0.041 and star > 0.232 and comfort ≤ 0.110 then negative (0/2)
3. if locat > 0.041 and star ≤ 0.232 and view > 0.253 and breakfast > 0.076 then positive (3/0)
4. if locat > 0.041 and star ≤ 0.232 and view > 0.253 and breakfast ≤ 0.076 then negative (0/2)
5. if locat > 0.041 and star ≤ 0.232 and view ≤ 0.253 then positive (240/0)
6. if locat ≤ 0.041 and staff > 0.024 and good > 0.017 then positive (32/0)
7. if locat ≤ 0.041 and staff > 0.024 and good ≤ 0.017 and help > 0.024 then positive (24/2)
8. if locat ≤ 0.041 and staff > 0.024 and good ≤ 0.017 and help ≤ 0.024 then negative (24/25)
9. if locat ≤ 0.041 and staff ≤ 0.024 and good > 0.101 and hotel > 0.074 then positive (10/1)
10. if locat ≤ 0.041 and staff ≤ 0.024 and good > 0.101 and hotel ≤ 0.074 then negative (6/11)
11. if locat ≤ 0.041 and staff ≤ 0.024 and good ≤ 0.101 and shop > 0.039 then positive (4/0)
12. if locat ≤ 0.041 and staff ≤ 0.024 and good ≤ 0.101 and shop ≤ 0.039 then negative (36/337)

Generated Decision Tree Rules, using balanced training set, 20-200

1. if good > 0.130 and internet > 0.589 then negative (0/4)
2. if good > 0.130 and internet ≤ 0.589 and coffe > 0.676 then negative (0/2)
3. if good > 0.130 and internet ≤ 0.589 and coffe ≤ 0.676 then positive (146/11)
4. if good ≤ 0.130 and friendli > 0.063 and morn > 0.539 then negative (0/1)
5. if good ≤ 0.130 and friendli > 0.063 and morn ≤ 0.539 and servic > 0.672 then negative (0/1)
6. if good ≤ 0.130 and friendli > 0.063 and morn ≤ 0.539 and servic ≤ 0.672 then positive (59/0)
7. if good ≤ 0.130 and friendli ≤ 0.063 and great > 0.054 and work > 0.424 then negative (0/2)
8. if good ≤ 0.130 and friendli ≤ 0.063 and great > 0.054 and work ≤ 0.424 then positive (48/2)
9. if good ≤ 0.130 and friendli ≤ 0.063 and great ≤ 0.054 and excel > 0.081 then positive (30/7)
10. if good ≤ 0.130 and friendli ≤ 0.063 and great ≤ 0.054 and excel ≤ 0.081 then negative (97/350)

Fig 2: Generated Decision Tree Rules in the three classification models

In the first case, when using the unbalanced training data set, filtering scheme 4-600, the attribute selected as the tree root is *locat* and it divides examples into two sub-trees – 414 examples with *locat*>0.032 and 644 examples with *locat*≤0.032. Nine decision tree rules are generated – six rules produce positive results and three rules – negative results. Rule 5, including locat>0.032 and term≤0.094 and view≤0.249, classifies best the positive examples since the leaf contains 406 positive examples and no negative examples. If the *locat* TF-IDF is >0.032, 100% of the examples in this branch are classified as positive, which means that the word *locat* has a significant role in classifying positive examples. The increasing of the *locat* TF-IDF score above 0.032 means that the word *locat* appears more often in a given document. If *locat*≤0.032 which means that the word *locat* appears rarely in examples, the attributes *staff>0.019* in rule 6, *good>0.086* in rule 7 and *excel (excellent)>0.221* in rule 8, are of significance in positive classification of 188 (82%) examples out of 224. Three rules classify the negative examples – rules 2, 4 and 9. However, the most important is rule 9, if locat≤0.032 and staff≤0.019 and good≤0.086 and excel≤0.221 then negative (85/335). When the TF-IDF values

of *locat, staff, good, excel* are low, then 80% (335) out of 410 examples in this branch are classified as negative. Based on that, we can conclude that if the words *locat, staff, good, excel* are not mentioned in the reviews, it is more likely the comments to be negative.

In the second case, when using the balanced training set, filtering scheme 4-600, the attribute selected for the root is again *locat* and it divides 760 examples into two subtrees – 248 examples if *locat*>0.041 and 512 examples if *locat*≤0.041. Twelve decision tree rules are generated – seven rules produce positive results and five rules produce negative results. In the first subtree, the most important rule is 5. Rule 5 including *locat*>0.041 and *star*≤0.232 and *view*≤0.253, classifies best the positive examples since the leaf contains 240 positive examples and no negative examples. If the *locat* TF-IDF is>0.041, 100% of the examples in this branch are classified as positive, which means that word *locat* again has a significant role in classifying positive examples. In the second subtree when *locat*≤0.041, *staff*>0.024 and *good*>0.017 in rule 6, *staff*>0.024 and *help*>0.024 in rule 7, *good*>0.101 and *hotel*>0.074 in rule 9 have classified 66 examples out of 69 as positive. In rule 12 when TF-IDF values of *locat, staff, good, shop* are low, then 90% (337) out of 373 examples in this branch are classified as negative. Therefore, it can be concluded that if the words *location, staff, good and shop* are not mentioned in the reviews, it is more likely the comments to be negative.

In the third case, when using balanced training set, filtering scheme 20-200, the attribute selected for the root is *good* and it divides 760 examples into two subtrees – 163 examples if *good*>0.130 and 597 examples if *good*≤0.130. Ten decision tree rules are generated – four rules produce positive results and six rules produce negative results. Rule 3 when *good*>0.130 and *internet*≤0.589 and *coffe*≤0.676 then classified positive examples are 146 and negatively classified are 11. *Good* is the most important word that classifies positive examples in the subtree. In rule 6 when *good*≤0.130 and *friendli*>0.063 and *morn*≤0.539 and *servic*≤0.672 then positively classified examples are 59 and negatively classified examples are none. *Friendli* (a stemmed form of friendly) is also significant in positive classification since it classified 100% of positive examples in the branch. In rule 8 the word *great* and in rule 9 the word *excel* are also of importance in the positive classification of examples. Based on rules 3, 6, 8, 9, the important words for positive classification are *good, friendli, great* and *excel*. Rule 10 when *good*≤0.130 and *friendli*≤0.063 and *great*≤0.054 and *excel*≤0.081, which means that *good, friendli, great* and *excel* are rarely or not present in an example, the classified examples as negative are 350 and 97 are classified as positive.

5.5 Results from the Comparison of the Trained Classifiers

The results from the evaluation of the trained classifiers are presented in Table 1. The overall accuracy of the three classification models is between 80.79 and 86.67 which is very high. The highest accuracy is achieved for the unbalance training data set, 4-600 and the lowest – for the balanced training data set, 20-200. The proportion of actual positive examples which are correctly identified (Positive Class Recall) is lower than the proportion of actual negative examples which are correctly identified (Negative Class Recall) for the three classification models. The model with the best Positive Class Recall (86.58) is using the unbalanced training data set,4-600, and the model with the best Negative Class Recall (90.53) is using balance training data set, 20-200.

Table 1. Results from the Evaluation of the Trained Classifiers

Results (%)	Unbalanced 4-600	Balanced 4-600	Balanced 20-200
Accuracy	86.67	85.53	80.79
Positive Class Recall	86.58	81.32	71.05
Negative Class Recall	86.84	89.74	90.53
F-measure	82.37	86.11	82.53
Positive Class Precision	92.15	88.79	88.24
Negative Class Precision	78.38	82.77	75.77
AUC	0.901	0.892	0.814

When using the balance training set with filtering scheme 20-200, the negative class recall is the highest, while the positive class recall is the lowest. The reason for the low positive class recall in this case is the filtering of the most frequently used words – *locat, room, breakfast, staff and hotel*, and especially the missing of the word *locat*, which turns out to be the most significant in positive classification. This classifier works best for negative example prediction. Identifying negative reviews is important for hotel managers in order to explore the possible problems shared by the guests in hotel service delivery.

The best classifier is using unbalanced training set and filtering scheme 4-600. It has the highest overall accuracy of .86.67% and the best positive class recall of 86.58% and still high negative class recall of 86.84%. The values of positive and negative class recall are almost equal, which means that the model equally well classifies positive and negative examples. The AUC of 0.901 is the highest and above 0.9, which means that it is a very good prediction model.

6. OPINION PREDICTION OF NEW HOTEL REVIEWS

The main purpose of this part of the research is to check the validity of the generated classification models. The data set for model application contains 227 example from TripAdvisor.com. The three classification models are applied on the data set to predict opinion of hotel guests. The text data is preprocessed and structured by applying subsequently the same techniques as applied in the training set. The results from the evaluation of the application of the trained classifiers on unseen data are presented in Table 2.

Table 2. Results from the Evaluation of the Application of the Trained Classifiers on Unseen Data

Results (%)	Unbalanced 4-600	Balanced 4-600	Balanced 20-200
Accuracy	85.46	80.62	85.90
Positive Class Recall	88.26	81.69	86.85
Negative Class Recall	42.86	64.29	71.43
Positive Class Precision	95.92	97.21	97.88
Negative Class Precision	19.35	18.75	26.32

Before prediction, the data set contains 212 (97%) examples manually labeled as positive and 14 (6%) manually labeled as negative. The overall accuracy, achieved in application of the three classification models on unseen data is between 80.62 and 85.90 which is very high again. The highest accuracy is achieved for the balanced training set, 20-200 and the lowest - for the balanced data set, 4-600. The highest positive class recall of 88.26 is achieved for the unbalanced data set, 4-600, and the highest negative class recall of 71.43 is achieved for the balanced training set, 20-200.

The classification model using the unbalanced data set, 4-600, which achieves the best results in the training, does not achieve good results in classification of negative reviews during the model application phase. While the classification model using the balanced data set, 20-200, achieves the highest overall accuracy of 85.90%, very good positive class recall of 86.85% and the highest negative class recall of 71.43%, which makes it suitable for prediction of opinion about hotel services expressed on review sites.

7. CONCLUSIONS AND FUTURE WORK

The paper proposed a model for classification of online hotel reviews using machine learning approach, Decision tree as a classification algorithm and filtering frequent and rare words as an attribute selection. Three experiments are conducted to evaluate the overall accuracy of the classifier and to improve the negative class prediction when applying the model on new unseen data. Three classification Decision tree models are compared. The results reveal that in the training step the best accuracy is achieved in the classification model using the unbalanced training set. The lowest accuracy is achieved when using balanced training set with filtering rare words that appear in less than 20 examples and frequent words that appear in more than 200 examples. In the model application step, the best accuracy and negative class recall are achieved when using balanced training set with filtering rare words that appear in less than 20 examples and frequent words that appear in more than 200 examples. The reason for the best result in negative class recall is filtering the most frequent words – *locat, room, breakfast, staff* and *hotel*, and especially *locat* from the attribute list, which appears to be the word of most importance in classifying positive examples.

The next steps in future research will be to summarize the predicted positive and negative hotel reviews, and additional data extracted from the hotel review section on both sites, and to present them in a suitable format for end users, to support hotels in management of online hotel reputation. Business Intelligence tools will be used for data interactive presentation and exploration.

8. REFERENCES

[1] Andreas M. Kaplan, Michael Haenlein, 2010. Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons (2010) 53, 59—68.

[2] Bing Liu. 2012 Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[3] Huifeng Tang, Songbo Tan, Xueqi Cheng, 2009. A survey on sentiment detection of reviews, Expert System with Applications 36 (2009) 10760-10773

[4] Gautami Tripathi, Naganna S.2015. Feature Selection and Classification Approach for Sentiment Analysis, Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.2, June 2015

[5] M. Bilal, H. Israr, M. Shahid, A. Khan, 2016. Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques, Journal of King Saud University - Computer and Information Sciences, Volume 28, Issue 3, July 2016, Pages 330–344

[6] Vikram Elango and Govindrajan Narayanan. 2014. Sentiment Analysis for Hotel Reviews, http://cs229.stanford.edu/projects2014.html

[7] P.Kalaivani, K.L.Shunmuganathan, 2013. Sentiment Classification of Movie Reviews by Supervised Machine Learning Approaches. Vol. 4 No.4 Aug-Sep 2013. Indian Journal of Computer Science and Engineering (IJCSE)

[8] A. Sharma, S. Dey. 2012. Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis, Special Issue of International Journal of Computer Applications (0975 – 8887) on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, June 2012

[9] H. Sui, C. Khoo, S. Chan. 2003. Sentiment Classification of Product Reviews Using SVM and Decision Tree Induction, 14th ASIS SIG/CR Classification Research Workshop, 2003

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

[11] Rüdiger Wirth, Jochen Hipp, CRISP-DM: Towards a Standard Process Model for Data Mining, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf

[12] Kotu V., Deshpande B., 2015. Predictive Analytics and Data Mining. Concepts and Practice with RapidMiner, ISBN 978-0-12-801460-8

[13] M.F. Porter, 1980, An algorithm for suffix stripping, Program, 14(3) pp 130−137.https://tartarus.org/martin/PorterStemmer/

[14] Ian H. Witten, Eibe Frank, Mark A. Hall, 2011. Data Mining. Practical Machine Learning Tools and Techniques, Third Edition.

[15] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, 2008. Introduction to Information Retrieval, Cambridge University Press. 2008. http://www-nlp.stanford.edu/IR-book/

[16] Kelly A. McGuire, 2011. Do Guest Reviews Really Matter? Linking Social Media and Operations Data, Paper 381-2011, SAS Global Forum 2011 Travel, Hospitality and Entertainment

[17] Diane Korte, Thilini Ariyachandra, and Mark Frolick 2013. Business Intelligence in the Hospitality Industry, International Journal of Innovation, Management and Technology, Vol. 4, No. 4, August 2013