Reviewers Classification in an Online Community of Romanian Tourists

Mihaela Colhon[1] and Costin Bădică[2]

[1]Department of Computer Science, University of Craiova
[2]Department of Computer and Information Technology, University of Craiova
Alexandru Ioan Cuza, 13, 200585, Craiova, Romania
[1]mcolhon@inf.ucv.ro, [2]cbadica@software.ucv.ro

Abstract. Our previous work addressed the computational analysis of communities of Romanian online users involved in tourism activities and interested in sharing their impressions and experiences, by focusing on touristic content sharing and review sites. Our studies comprise several tasks such as sentiment analysis, keywords extraction and graph-based structuring of the users community. In this paper we extend our previous analysis of the users community of AmFostAcolo tourism Web site, by proposing a supervised classification method of reviewers based on their portfolio data. The goal of our classification is to develop a method for automated labeling the users that post reviews as novice, common or experienced reviewers. The classification uses features derived from data extracted from users' portfolio. We have investigated several multi-class classifications and our results are encouraging.

Keywords: supervised classification, online community, social-media, user features, review features

1 Introduction

More and more online tourists prefer to write reviews on various social network platforms with the goal of sharing their opinions and experiences [1]. For example, tourist information is most often presented as reviews or comments expressed in unstructured natural language texts describing customer impressions about their visited tourist destinations [2]. The textual information available on the Web is of two types: facts and opinion statements [3]. All these textual generated contents are valuable for many Web applications that crawl and search the Web looking for meaningful information. Thus, there is a constant need for automatic tools that analyse and classify user generated data collected from the Web.

But not all the Web data is reliable and trustable. On various forums and blogs we deal with different users which assume different roles in their online communities. Some of them are contributing with their knowledge to the community, while others are only searching for recommendations and advices. Regarding the users of the first category, we must automatically differentiate trustworthy opinions from

untrustworthy opinions. Thus this issue raises the importance of defining a reliable user classification, so we setup this task of differentiating novice and less trusted opinions from the expert and more trusted opinions.

The users that are responsible for producing the largest proportion of high quality content in online communities are known as *experts*. Actually they are the real contributors in the Web communities where they post. A review written by an expert describes an expert opinion in sufficient detail basing on real experiences with several peer products or services in order to highlight which offers a better value or a better set of features. On the other hand, novices' reviews usually contain no or very few valuable or trustable information.

The quality of text reviews usually involves the computation of their sentiment score. This is usually achieved with the help of special words called *emotion triggers* that are stored in specialized lexicons called *opinion lexicons*. The subjectivity values of these triggers may be changed in context by the so-called *valence shifters* [4]. These are terms that can change or even cancel the semantic orientation of the term they modify. For example a valence shifter can make a positive term to become negative as in “bad” and “*not so* bad” example. The construction of valence shifters has been intensively studied, as they play an important semantic role in natural language descriptions. A comprehensive work dedicated to the analysis of valence shifters of Romanian language is [5]. As textual classification derives from the natural language studies, we can find a lot of research studies in the field of automated text classification [3, 6, 7], but not so many addressing the classification of the users behind those texts.

In this work we address the task of reviewers classification based on a combination of features related to the reviewers' personal data, as well as to the characteristics of the texts this particular type of website users write. Our goal is to derive a general purpose classification model of reviewers of content sharing networks that can be used to rank this type of users based on their most-common attributes.

The paper is organized as follows. In Section 2 we introduce the reader to the relevant works that have been done on user classification for web site media. In Section 3 we describe the characteristics of the data set that we have chosen for designing and evaluating our classification method, the features used and the statistical models we have applied on our data. Finally, in Section 4 we draw our conclusions and outline some paths for continuing this work.

2 Previous Work

Previous work has explored the impact of user profiles on the style, patterns and content of their communication streams [8]. Studies have been performed for exploring the impact of the users attributes on their possible classification. So far, the impact of the *user gender* [9], the *user location* [10, 11], the *user age* [11] or the *user political orientation* [12] was analysed. All these works address data collected from blogs or other informal large texts.

In what follows we will consider review data extracted from microblogs, where the text is usually not so large (in some cases it consists of only one or two phrases). As consequence, a micro-blog user is characterized by other data such as the *user writing behavior* or the *user social network information* [8] along with the *user profile information*. On Twitter, the user behavior is dictated not only by the tweet characteristics (presence of hash tags, URLs or other informal data in text), but also by numbers of followers [13]. In [14] the authors suggest that users who rarely post tweets but have many followers tend to be information seekers, while users who often post URLs in their tweets are most likely information providers. On contrary, in [15] the main idea suggests that tweeting behavior is not useful for most classification tasks being subsumed by linguistic features.

On AmFostAcolo site[1] we can find a large and valuable touristic information given in the form of textual touristic impressions referred in what follows as *reviews*. On this site users post their touristic impressions or comment on others impressions. The users that wrote reviews, called *reviewers*, post on this site semi-structured reviews about a large variety of tourist destinations covering specific aspects of accommodation units, as well as general impressions about tourist geographical places, regions and attractions [16]. The task of collecting, aggregating and presenting the reviews' content in a meaningful way can be very difficult by the many cognitive challenges implied through the process.

In the last three years we have setup a research project focused on the management of information and knowledge extracted from reviews and opinions about tourist destinations that can be publicly found on the Web [2,16, 17,18, 5]. All these studies are based on real data extracted from AmFostAcolo Web site. Using data collected from AmFostAcolo Web site we have developed several applications. In [2] we presented an unsupervised sentiment classification method of tourist reviews that was built primarily on dependency links between the words of a text. In [2] we introduced the concept of *seed* which is used in the classification task in order to determine a sentiment towards a specific facet of the target entity by finding correlations between the corresponding facet and its textual realization in the review (its seeds). For this reason, we built several sets of seeds, each set corresponding to a certain entity's facet.

The works reported in [17, 18] are based on the results that we obtained on the same data set by employing graph based representations of reviews, as well as a set of analysis metrics designed for complex networks community. The graph-based representations we proposed are intended for analysing the reviewers' community and also for applying a keywords extraction method on the touristic reviews.

In this paper we address the task of reviewers' classification: we attempt to automatically infer the reputation of reviewers based only on the most common features of their profile and activity. We have designed and evaluated a classification method using data from the same repository - the AmFostAcolo web site. The input data represent attributes extracted from the reviewers' personal profile (their age) and from their entirely activity on the site (number of reviews, their quality) while the

[1] http://amfostacolo.ro/

evaluation of the proposed classification method is done using the reviewers classes posted on this site and determined in a semi-automatic manner using several criteria[2].

Because our intended scope is to build a method for reviewers classification that can be further applied on any reviews' site data, from the set of features used to define the reputation rules of AmFostAcolo community we kept only the features that can be encountered on any site within the same domain (see Section 3.2). Even if, in this way, we did not use all the data involved in defining the reputation rules, the obtained results are quite promising as it will be shown in Section 3.3.

3 The Classification Model

The research presented in this paper is focused on the task of identifying and further using the most relevant features from the user participation in an online community provided by a reviews site in order to define a user classification method depending on his/her role in the process of generating the community content. Based on the features extracted from the online community the reviewer profile is defined. The set of the resulted profiles will constitute the input data for a supervised classification algorithm designed for determining the reviewers' reputation classes.

A classification process is divided into two main steps (phases): *training* and *testing* [19]. Training phase gets a set of labelled examples as input and produces the classification model as output. The testing phase uses additional labelled examples to evaluate the classification model produced by the training phase.

Every user classification can be approached as a typical classification problem, so the user classification model can be trained using a set of labeled examples. Each example is structured into a feature template defined by a given set of features. A machine learning classification task involves the following three basic factors: *feature template* (what type of features are used by the model), *feature function* (the function that maps each feature of a given example into a special value of that feature) and the *classification algorithm* (that maps examples to classes using a specific classification model, like for example Naïve Bayes, Support Vector Machine or Maximum Entropy) [1].

In the next section we present the sample data set extracted from the AmFostAcolo community, including the features that we have chosen for developing our proposed classification model.

3.1 The AmFostAcolo Data Set

In view of developing a user classification method we firstly need to understand the users' behavior. When analysing the website of a users' community, it is important to identify all the different components of this online community.

[2] The set of rules applied for determining the so-called PMA points, based on which the class of an user is further established is given at the address http://amfostacolo.ro/pma_explic.php

The main components of every online community are: the *actors/users* who participate in the community, the *user features* (some of them are personal, some can results from the participation in the online community), and the *principles of interactions* in community.

Most services (such as Twitter) publicly show a user profile information which includes the user name, the age, the location and a short bio [8]. Still, on most of the sites, the profile fields do not contain enough good-quality information, thus these data cannot be used as a reliable data.

The AmFostAcolo community is the unique source of information for our data set as the data extracted from this site was appropriate not only for defining the classifier but also for evaluating it. Nevertheless, a similar methodology can be used to extract data from any other tourist Web sites.

AmFostAcolo provides a large semi-structured database with information describing post-visit tourist reviews about an important variety of tourist destinations, as well as geographical places and regions [2]. The reviews posted on the AmFostAcolo site are hierarchically organized, such that the top level corresponds to *destinations*, each destination is composed of several *regions*, each region is composed of *sections* and the last level corresponds to the *locations* representing the leafs of this hierarchy. Locations do not have any other inner structure as they represent a specific touristic item.

As an overview of the AmFostAcolo online community of reviewers we graphically represent them in Fig. 1 grouped upon the destinations described in their reviews, that is grouped upon their touristic interests or touristic favorite destinations. As it can be seen in Fig. 1 most of them wrote about Romanian destinations. Also a lot of users wrote about Austria, Egypt, Montenegro or Greece destinations.

Our dataset contains 2521 reviews and 1085 reviewers.

Each review is characterized by a series of elements, including: its textual content - an unstructured natural language text in which the user expresses his/her touristic (as well as other more general, for example history-related) impressions about the described location and the review title, as well as other attributes such as: the location referred in the review, the number of echoes (comments) the review received, and the id of the user that wrote the review. Also, each review has a *set of aspects* that are usually (but not mandatory) debated in reviews such as accommodation, kitchen, services, etc. These aspects are accompanied by positiveness scores based on which the overall positiveness score of the review is calculated (0 positiveness score means that a specific aspect is purely negative or is missing at all, while 100 positiveness score means that the corresponding aspect got the maximum appreciation from that user).

All the information extracted from the AmFostAcolo site is stored in three dedicated XML files: *reviews.xml*, *reviewers.xml* and *destinations.xml*.

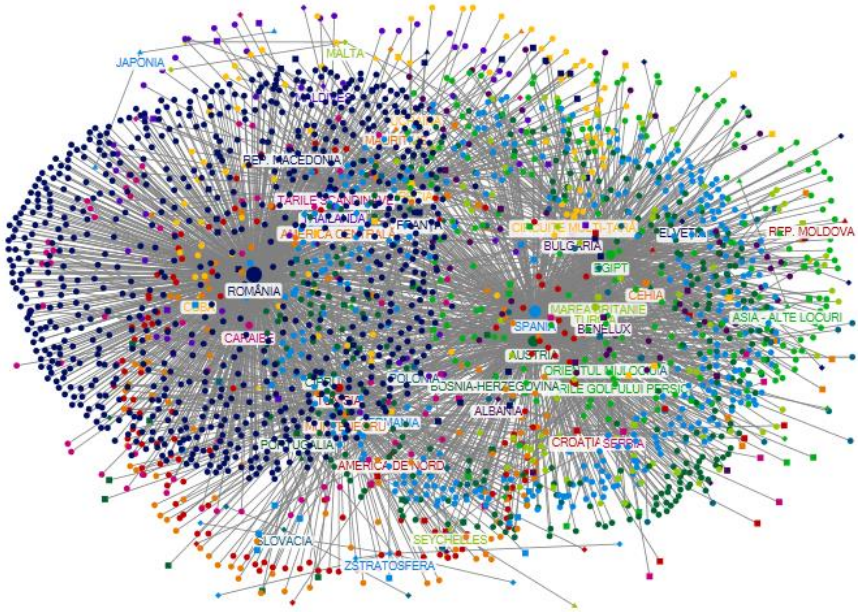

Fig. 1. Communities of reviewers based on their reviews' destination topic.

More precisely, all the reviews extracted from the AmFostAcolo site are stored in an XML file consisting of *review* XML elements. In order to illustrate the attributes of this element we give an example of a review about the location "Ioana Pension" having the title "A clean and beautiful pension":

```
<review id="f71f841a-938d-4db1-a460-5ebff1961037"
   location_id="PENSIUNEA IOANA" no_echo="6"
   text="Am dorit sa raman pentru o noapte (...)"
   title="O pensiune curata si frumoasa"
   user_id="Gabi Troia">
   <SS_services>100</SS_services>
   <SS_accomodation>100</SS_accomodation>
   <SS_kitchen>0</SS_kitchen>
   <SS_landscape>100</SS_landscape>
   <SS_entertainment>0</SS_entertainment>
</review>
```

The reviews from our data set were written by 423 male users and 662 female users. Our data set contains information about the users that posted reviews on the AmFostAcolo Web site. These information are: the user identifier, the age interval (one of 20-30 years, 30-40 years, 40-50 years, and 50-60 years), sex, geo-location, and user class (also called "statut" in Romanian) [2].

We also saved all these data about the profiles of this type of users into an XML file consisting of *reviewer* XML elements. Here is an example of the XML structure corresponding to the user id "Vasile S":

```
<reviewer age="40-50 ani" location="Alba Iulia" sex="M"
  user_id="Vasile S" user_class="GOLD" />
```

The geographical information about the locations described in the reviews of our data set is stored in an XML file consisting of all the destinations listed on the site. Here is an example of a location named "Discovering Mexico" nested by the section "#Traveling in Mexico" which is included in the destination named "Central America":

```
<destination name="AMERICA CENTRALĂ">
   <region name="MEXIC">
    <section name="#CALATORII MEXIC">
          <location name="DESCOPERIND MEXICUL" />
      </section>
    </region>
</destination>
```

As we have already specified, from the available set of users we have extracted only those users that have reviews on this site, finally obtaining a subset of relevant reviewers of the AmFostAcolo site.

The AmFostAcolo Web site provides a quite significant set of public characteristics of its reviewers, either directly input by the user, or derived from the user activity on the site:

- the *reviewer identifier*, which is a name that uniquely identifies the user in the community;
- the personal data containing demographic information such as the *age interval*, *sex* and the *user geo-location*. Fig. 2 illustrates graphically the statistical distribution of user ages;
- the *reviewer class*, called "statut" in Romanian. This data is semi-automatically generated based on the user activity portfolio measured by the so-called **PMA** (*Points of Contentment and Appreciation*, in Romanian ***P**uncte de **M**ulţumire şi **A**preciere*) points that the user received so far. The reviewer class is very important, as our classification is based on labelled examples starting from this data.

As we have already pointed out, the most interesting attribute based on which the proposed classifier has been evaluated is the *reviewer class*. This data actually represents a qualitative score that characterizes the reviewer experience as a traveler as well as in writing touristic reviews. The reviewer class is determined taking into account the user activity portfolio which includes the user reviews, the replies and the answers that the review receives, the photos that can accompany the reviews and the possible question-answering chains that can follow the review (so-called *echoes*).

Based on the points gained for its portfolio, the reviewer is assigned to one of the following user classes: UCENIC, JUNIOR, SILVER, GOLD, PREMIUM, SENIOR, PARTENER, SENATOR, PRETOR. These classes are given in ascending order upon the number of PMA points that the reviewers of the corresponding class acquired. Furthermore, based on the data extracted from the Web site, we also identified another reviewer class called CHIBIT which, we suspect, is a transition type, i.e. it describes a reviewer having the lowest rank or the smallest number of PMA points (i.e. smaller than UCENIC users) [2].

In order to create a reviewer classification method that can be applied on any reviews site data, from the set of features extracted from the AmFostAcolo site we have considered only the most common ones. More precisely, we have chosen only those features that are not specific to a particular online community of reviewers being widespread in this community type.

In this manner, the classification method we propose in this paper can be applied even on a site which does not generate or show the reputation classes of the users that wrote its reviews, being helpful for everyone who wants to obtain information about how much trustworthy is someone's review.

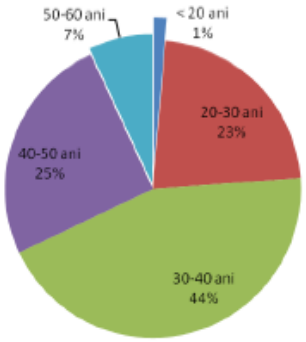

Fig. 2. The reviewers from our data set classified upon their age

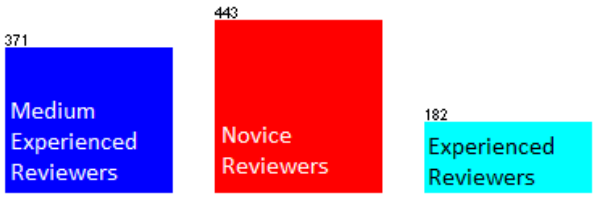

Fig. 3. The reviewers from our data set classified upon their age

3.2 The Reviewers Classification Problem

As everyone can notice, on the Web there is a large amount of user generated reviews, from the contemplative literary critiques such as GoodReads to the impressions about

hotels on TripAdvisor. The constant growing of the volume of the online reviews asks for development of automatic tools that can classify the reviews upon their helpfulness. One way to address this problem is by training classifiers using general review features including: the *readability* of their textual content, the *star rating* of the review (if the Web site provides such information) or the *reputation* of the reviewers [20]. The last approach is addresses in this paper but is it developed as a final scope and not within a reviews classification task as we consider that the reviewer reputation is itself a very important trustworthiness indicator for his/her opinions.

Each Web site has its specific data model with respect to the information displayed and the way the data is organized. Besides the AmFostAcolo site there are many others Romanian (on available in Romanian language) review Web sites, some of them also in the touristic field such as Booking.com[3] (Romanian version), others in the IT domain such as ComputerGames[4], or online shop sites with an important product reviewing aspect. There are several online shops for Romanian customers, like for example Emag[5]- an online shopping site that sells products from various categories and which encourages users to write reviews about the products they bought because it has been proved that such opinions help buyers in decision making.

In order to use the data extracted from the AmFostAcolo site for designing a reviewers classification technique that can be applied to other touristic review sites, we selected from the reviewers and their reviews features set only those features that can be found on any other similar site. Our intended purpose is to create a reviewers classification method that can be applied on as many reviewing sites as possible.

As consequence, we selected three features for the input examples (Fig. 4) of the proposed classifier. These features were chosen in order to describe the two main characteristics of the reviewer role: “who is the reviewer” (who is the person under the reviewer id) and “how the reviewer posts” (how is the reviewer's activity in the online community):

- the “who is the reviewer" data: demographic information that personally describe the user. From the AmFostAcolo reviewer profile set we chose a single data:
  - the user age
- the “how the reviewer posts” data: information about the user activity on the site, that is how is he/she writing reviews. For each reviewer we have determined two pieces of information in order to be used in the classification:
  - the total number of reviews of the user
  - the average length of the user reviews (given as number of words)

Classification is only possible if class information is available in the given examples. For that purpose we had to use information about the user class, available in AmFostAcolo Web site. Initially we tried to consider all the available reviewer

[3] www.booking.com
[4] http://computergames.ro
[5] http://emag.ro

classes[6] from AmFostAcolo online community in order to train our classifiers and to evaluate our classification. But, because these classes greatly differ upon their size (understood as number of users in the class), we have decided to group them in three larger meta-classes: *Novice Reviewers*, *Medium Experienced Reviewers* and *Experienced Reviewers*. These three meta-classes will represent the class labels of our classification method (Fig. 4).

In order to group as smooth as possible the reviewers AmFostAcolo classes in three balanced meta-classes, in terms of size and reviewer ranks we defined a mapping, as follows:

- the SENIOR, PARTENER, SENATOR, PRETOR and PREMIUM AmFostAcolo classes include the reviewers with the highest PMA points. In consequence, we have grouped these reviewers into the *Experienced Reviewers* meta-class;
- the CHIBIT, UCENIC and JUNIOR AmFostAcolo classes correspond to the reviewers with the lowest PMA points. In our approach, these reviewers are considered to be part of the *Novice Reviewers* meta-class;
- the GOLD and SILVER classes represent the reviewers with medium portfolio PMA points. These users were assigned to the *Medium Experienced Reviewers* meta-class.

Fig. 4. The input attributes and the output class of the classifier for the corresponding AmFostAcolo data

As we have already pointed out, this mapping was designed in order to obtain meta-classes as balanced as possible taking into account their sizes measured in number of users (see Fig. 3 and the users' ranks. The resulted meta-classes are briefly described as follows:

[6] Here we consider also the CHIBIT class along with the other nine reviewer classes declared on the AmFostAcolo site.

- *Novice Reviewers* (443 users): CHIBIT (14 users), UCENIC (202 users), JUNIOR (227 users)
- *Medium Experienced Reviewers* (371 users): SILVER (172 users), GOLD (199 users)
- *Experienced Reviewers* (182 users): PREMIUM (95 users), SENIOR (39 users), PARTENER (31 users), SENATOR (15 users), PRETOR (2 users).

```
Correctly Classified Instances         654               65.6627 %
Incorrectly Classified Instances       342               34.3373 %
Kappa statistic                          0.4441
Mean absolute error                      0.3074
Root mean squared error                  0.3943
Relative absolute error                 73.1821 %
Root relative squared error             86.0336 %
Total Number of Instances              996

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.539    0.245    0.567      0.539    0.552      0.297    0.676     0.561     MEDIUM
               0.795    0.273    0.700      0.795    0.744      0.518    0.806     0.702     NOVICE
               0.560    0.047    0.729      0.560    0.634      0.571    0.869     0.671     EXPERIENCED
Weighted Avg.  0.657    0.221    0.655      0.657    0.653      0.446    0.769     0.644

=== Confusion Matrix ===

   a   b   c   <-- classified as
 200 136  35 |   a = MEDIUM
  88 352   3 |   b = NOVICE
  65  15 102 |   c = EXPERIENCED
```

Fig. 5. The accuracy of the reviewers' classification method

3.3 Performance Evaluation

Extracting user characteristics constitutes an important step towards user classification. Another important step is selecting the set of classification algorithms. We have tested several classifiers: Naïve Bayes, Support Vector Machine or Multi-Class Classifier.

We have used WEKA tool to conduct our experiments. WEKA [21] is an open source software package supporting a collection of machine learning algorithms that can be either applied directly to a dataset or programmatically called from Java code. For the purpose of this evaluation we used three commonly used classifiers available in Weka: Naïve Bayes, Support Vector Machine and Multi-Class Classifier. The evaluation is performed in terms of the standard measures such as *Precision* (*P*), *Recall* (*R*) and *F-measure*. Based on the results, the performance of Naïve Bayes is the best, Support Vector Machine - second and Multi-Class Classifier's performance is the minimum. In Fig. 5 we show the results obtained with the best classification method for our data, i.e. with the Naïve Bayes classifier. The dataset we have used contains 2521 reviews each accompanied by its title, the reviewer id and its positiveness scores.

The WEKA's Naïve Bayes classifier [22] is evaluated in a 10-fold cross-validation which splits the dataset into 90% of training set and 10% of test set. Naïve Bayes is a probabilistic classifier, based on the Bayes theorem. Assuming that is a reviewer and

c is a target class, the probability that a reviewer (an instance) rev belongs to class c is:

$$P(c|rev) = P(rev|c) \times \frac{P(c)}{P(rev)}$$

Different types of errors performed by a classifier can be summarized in a *confusion matrix*. For a multi-class problem with *n* classes, the confusion matrix will have n^2 entries. The correct classifications lie on the diagonal line, and the off-diagonal entries contain the various cross-classification errors [23]. Weka evaluation output includes also the confusion matrix values. Using this matrix, the number of correctly/incorrectly classified instances can be seen in help to explain the classification accuracy of an algorithm.

In terms of *TP* (rate of true positive, i.e. instances correctly classified per class), *FP* (rate of false positive, i.e. instances wrongly classified per class), and *FN* (rate of false negatives, i.e. non-instances wrongly classified per non-class), *Precision* and *Recall* measures are defined as follows:

- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$

In Fig. 5 we give the detailed accuracy by class as it is determined using Weka, which includes *TP* rate and *FP* rate.

We consider that the obtained scores are very promising considering the small number of features that we have used in our classifier (only three features). For all the three meta-classes considered in the classification model, the precision is above 50% with 70% percentage for *Novice* and *Experienced reviewers*.

4 Conclusions and Perspectives

In this paper we reported our first results of a proposed reviewers' classification method applied to an online community of Romanian tourist reviews site. The intended scope is to create a reviewer classification method that can be applied on any reviews site data. For this reason in the created data set we have considered only the most common features. More precisely, we have chosen only those features that are not specific to a particular online community of reviewers being widespread in this community type.

In this manner, the classification method we propose in this paper can be applied even on a site which does not generate or show the reputation classes of the users that wrote its reviews, being helpful for everyone who wants to obtain information about the trustworthiness of someone's review.

References

1. Wang, X., Lin, Y., Zhou, A. Opinion Analysis for Online Reviews. World Scientific Publishing Co., Inc., River Edge, NJ, USA (2016).

2. Colhon, M., Bădică, C., Şendre, A. Relating the Opinion Holder and the Review Accuracy in Sentiment Analysis of Tourist Reviews. In Proceedings of 7th International Conference on Knowledge Science, Engineering and Management, KSEM 2014, pp. 246–257 (2014).
3. Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., Hussain A. Concept-Level Sentiment Analysis with Dependency-Based Semantic Parsing: A Novel Approach. Cognitive Computation 7, 4, pp. 487-499 (2015).
4. Polanyi, L. Zaenen, A. Contextual Valence Shifters. In Computing Attitude and Affect in Text: Theory and Applications, James G. Shanahan, Yan Qu, and Janyce Wiebe (Eds.). Springer Netherlands, pp. 1–10 (2006).
5. Colhon, M., Cerban, M., Becheru, A., Teodorescu, M. Polarity shifting for Romanian sentiment classification. In Proceedings of the International Symposium on INnovations in Intelligent SysTems and Applications (INISTA 2016). pp. 1–6 (2016).
6. Gîfu, D. Cristea, D. Public Text Categorization. In Proceedings of the 8th International Conference "Linguistic Resources and Tools for processing of the Romanian language", Mihai Alex Moruz, Dan Cristea, Dan Tufiş, Adrian Iftene, Horia-Nicolai Teodorescu (eds.), "Alexandru Ioan Cuza" University Publishing House, Iaşi (CONSILR '12). pp. 75–84 (2012).
7. Negru, V., Grigoraş, G., Dănciulescu, D. Natural Language Agreement in the Generation Mechanism based on Stratified Graphs, Proceedings of the 7th Balkan Conference in Informatics (BCI 2015), pp. 36:1-36:8 (2015).
8. Pennacchiotti, M. Popescu, A.M. A Machine Learning Approach to Twitter User Classification. In Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, (2011).
9. Herring, S.C., Paolillo, J.C. Gender and genre variation in weblogs. Journal of Sociolinguistics 10, 4, pp. 439–459 (2006).
10. Cheng, Z., Caverlee, J., Lee, K. You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10). ACM, pp. 759–768 (2010).
11. Jones, R., Kumar, R., Pang, B., Tomkins, A. I Know What You Did Last Summer: Query Logs and User Privacy. In Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM '07). ACM, pp. 909–914 (2007).
12. Thomas, M., Pang, B., Lee, L. Get out the Vote: Determining Support or Opposition from Congressional Floor-debate Transcripts. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06). Association for Computational Linguistics, pp. 327–335 (2006).
13. Campbell, W., Baseman, E., Greenfield, K. Content + Context Networks for User Classification in Twitter. In Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP) (2014).
14. Java, A., Song, X., Finin, T., Tseng, B. Why We Twitter: Understanding Microblogging Usage and Communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. ACM, pp. 56–65 (2007).
15. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M. (2010). Classifying Latent User Attributes in Twitter. In Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents. ACM, pp. 37–44 (2010).
16. Bădică, C., Colhon, M., Şendre, A. Sentiment analysis of tourist reviews: Data preparation and preliminary results. In Proceedings of the 10th International Conference Linguistic Resources And Tools For Processing the Romanian Language, ConsILR 2014. pp. 135–142 (2014).

17. Becheru, A., Buşe, F., Colhon, M., Bădică. C. Tourist review analytics using complex networks. In Proceedings of the 7th Balkan Conference on Informatics Conference, BCI 2015. 25:1–25:8 (2015).
18. Becheru, A., Bădică, C. A deeper perspective of online tourism reviews analysis using natural language processing and complex networks techniques. In Proceedings of the 12th International Conference Linguistic Resources And Tools For Processing the Romanian Language, ConsILR 2016. 189–192 (2016).
19. Jia, S., Liang, J., Xie, Y., Deng, L. A novel feature voting model for text classification. In 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). pp. 306–311 (2014).
20. Dong, R., Schaal, M., O'Mahony, M.P., Smyth, B. Topic Extraction from Online Reviews for Classification and Recommendation. In Proceedings of the Twenty-third International Joint Conference on Artificial Intelligence (IJCAI '13). pp. 1310–1316 (2013).
21. Weka 3. Data Mining with Open Source Machine Learning Software in Java, Machine Learning Group at the University of Waikato. (2017). Retrieved March, 2017 from http://www.cs.waikato.ac.nz/ml/weka/index.html (2017).
22. Meena, M.J., Chandran, K.R. Naïve Bayes text classification with positive features selected by statistical method. In First International Conference on Advanced Computing. pp. 28–33 (2009).
23. Monard, M.C., Batista, G. Graphical Methods for Classifier Performance Evaluation. In Proceedings of Advances in Logic, Artificial Intelligence and Robotics (LAPTEC'2003). pp. 59–67 (2003)