EVALUATING POTENTIAL IMPROVEMENTS OF COLLABORATIVE FILTERING WITH OPINION MINING

Manuela Angioni, Maria Laura Clemente and Franco Tuveri
CRS4, Center of Advanced Studies, Research and Development in Sardinia
Parco Scientifico e Tecnologico, Ed. 1, 09010 Pula (CA), Italy
{angioni, clem, tuveri}@crs4.it

Keywords: Opinion Mining, Natural Language Processing, Collaborative Filtering, Matrix Factorization, Ensemble methods.

Abstract: An integration of an Opinion Mining approach with a Collaborative Filtering algorithm has been applied to the Yelp dataset to improve the predictions through the information provided by the user-generated textual reviews. The research, still in progress, based the Opinion Mining approach on the syntactic analysis of textual reviews and on a beginning polarity evaluation of the sentences. The predictions produced in this way was blended with the predictions coming from a Biased Matrix Factorization algorithm obtaining interesting results in terms of Root Mean Squared Error (RMSE), with potential enhancements. We intend to improve these results in a further phase of activity by including in the Opinion Mining approach the semantic disambiguation and by using better criteria of evaluation of the reviews taking into account a set of 12 business aspects. The Opinion Mining approach will be evaluated comparing the output in terms of predictions with the values manually assigned by a small group of people to a sample of the same reviews.

1 INTRODUCTION

The spontaneous textual reviews written by the users through online services represent interesting information for Opinion Mining methodologies, although the total absence of any kind of rules implies new difficulties to be dealt with. A dataset, such as the Yelp business recommender service, provides the researchers with a valuable source of material enriched by the star ratings. In fact, this combination is ever more common in the available datasets and gives the possibility to analyse the textual reviews through Opinion Mining methodologies while the star ratings can be used by Collaborative Filtering algorithms in order to verify the level of reliability.

The rest of the paper is structured as follows: we provide related work about the combination of Opinion Mining and Collaborative Filtering in Section 2. Section 3 describes the Yelp dataset and its issues. Section 4 describes the Opinion Mining analysis process and the algorithm evaluation while the prediction analysis methodology is described in Section 5. Lastly, Section 6 reports conclusions and future works.

2 RELATED WORK

Since the Yelp dataset has been provided for the RecSys2013 competition, there are a great number of studies related to it, which describe research activities with various purposes and involving different aspects.

Many of these studies have the aim of proposing effective recommender systems based on the textual reviews (Trevisiol et al., 2014; Fan et al., 2014), some of them combining Opinion Mining methodologies with Collaborative Filtering algorithms (Govindarajan, 2014). Again in Trevisiol et al., (2014) the users' preferences about food are used to produce personalized menu recommendations.

Several types of algorithms can be used to deal with predictions for recommender systems, but many researchers agree that the most effective algorithm working alone for this kind of problems is based on matrix factorization (Koren, et al. 2009; Tosher et al. 2009).

It is important to note that although algorithms based on the ratings produce winning results in terms of Root Mean Squared Error (RMSE), they do not consider the content provided in a textual review, which can be very effective in order to make a deeper analysis of the actual opinion of users about the different features which characterize a resource and, most importantly, if a user would really recommend that resource to other people (Koukourikos et al., 2012).

These kind of limitations can be overcome by Opinion Mining methodologies which could be used to enrich the output of Collaborative Filtering algorithms (Quadrana, 2013), including but not limited to the well-known *cold start* problem (Levi et al. 2012).

It must be said that user generated textual reviews without rules bring to another class of issues, as described in Section 3.

3 THE YELP DATASET AND ITS ISSUES

The Yelp dataset was chosen for the presented study because it provides both textual reviews and star ratings assigned by the users to the businesses (http://www.yelp.com).

The credibility of the Yelp textual reviews has been sometimes put in doubt because the business-owner of a weak activity could be tempted to write positive reviews for his/her business, or worse than this, fake negative reviews for some of its competitors. To avoid this, Yelp filters out these kind of false opinions and in any case the dataset represents a valid source of information for the research activities because most of the reviews can be considered reliable.

A part from these aspects, there are also some further elements, which make the Yelp dataset more complicated to be analysed. In fact, there are no rules about what aspects of a business should be considered in a review and while some people describe almost everything (location, presence of a parking area, interior decoration, furnishing, size of the sitting area, quality of service, variety of food and wine, quality and amount of food, prices, etc.), some others limit their exposition to one or few sentences about the quality of food.

There are also some reviews made of a mixing of positive and negative different experiences in the same restaurant, which took place during a range of many years.

A further issue arises when a description is related to the comparison between two or more restaurants.

4 OPINION MINING

In the presented activity the Opinion Mining methodology has been considered to predict business ratings from the user-generated textual reviews, and the effect of this approach is expected to reduce the bias implicitly introduced by the various personalities of the reviewers in the star ratings.

As in Benamara et al., (2007), we propose a *linguistic approach* to Opinion Mining and, more in details, to the extraction of feature terms by means of the syntactic and semantic analysis of textual resources. In particular, we focus on the analysis of the opinions through the processing of textual resources, the information extraction by means of the syntactic chunk analysis, the semantic disambiguation of terms, and the evaluation of the semantic orientation.

The identification of adjectives and adverbs and the use of subjective lexical resources have a relevant role in this phase.

The main tasks of the Opinion Mining process are described in more detail in the following Section.

4.1 Data Processing

The reviews considered for the research activity are the ones related to the business belonging to the "Restaurant" category. We collected our data considering only the users giving a number of reviews greater than 9. As shown in Figure 1, a total of 67,451 text reviews have been extracted.

A spell check on the obtained reviews and then a replacement of the verbal short forms with the long forms were made in order to reduce the amount of errors and to facilitate the syntactic parsing activities. Dividing the reviews in sentences, a total of 953,314 sentences has been obtained.

Then the TreeTagger (Schmid, 1994) phrase parser was run for the chunking, along with the sentence annotation with Part Of Speech (POS) tags and lemma information, and in each sentence the

sub-constituents were identified. Subsequently an analysis of the parts of speech was carried out in order to associate the nouns with their related information.

Through the semantic disambiguation task, it is possible to reduce the number of synsets activated by the syntactic analysis. Calculating the synset density in a document, we can take advantage of the semantic relations available from WordNet. Moreover, it is possible to refer sentences to domain topics during the semantic disambiguation phase, identifying all the synsets referred to a textual content, and evaluating its most probable sense (Angioni et al. 2008). The "sentence analysis" performs the distinction between subjective and objective sentences, with or without orientation, to detect factual sentences, which have a polarity value. This is a very important step because only the subjective sentences and the factual sentences having a polarity valence have been considered.

This task was performed with SentiWordNet (Esuli and Sebastiani, 2007), which is a lexical resource able to assign the following three sentiment scores to each synset of WordNet (Miller, 1998): positivity, negativity, and objectivity.

This kind of analysis will produce a set of predictions to be compared with Yelp ratings and with a sample set of ratings partially discussed by two researchers and based on a common evaluation criterion, as described in Section 4.4.

4.2 Feature Identification

The feature identification is a relevant task of the process. The term *feature* is used with the same sense given by Ding et al. (2008) in their approach to Opinion Mining. Given an object, that could be a service, a person, an event or an organization, the term feature is used to represent a component or an attribute describing that object. Considering that the domain is well known, it was not necessary to perform a process of contextualization of the information by means of the categorization of the set of reviews (Angioni and Tuveri, 2011). The identification of the features for the Yelp reviews has been therefore performed evaluating the nouns frequency in the text through a word counter.

The stop words have been removed and then the cleaned text has been tokenized obtaining as a result a collection of individual and compound words. The semantic disambiguation provides us with the corresponding synsets associated to the features terms and to the related adjectives and adverbs, semantically referred to the domain. The domain associations among the word-senses were based on the mapping of the synsets on WordNet Domains categories (Magnini et al., 2002) (Magnini et al., 2004). The semantic disambiguation of terms allows grouping them according to their properties of synonymy, hyponymy and meronymy.

Once manually validated, disambiguated and grouped, the set of terms obtained will be organized in 12 collections of features according to the 12 business aspects used in the evaluation criterion. An example could be the aspect "Staff" that collects features like waitress, chef, bartender, employee, etc., whereas examples of the considered aspects are: the quality of the food, the ambience, the presence of a parking area, etc. This step is still in progress.

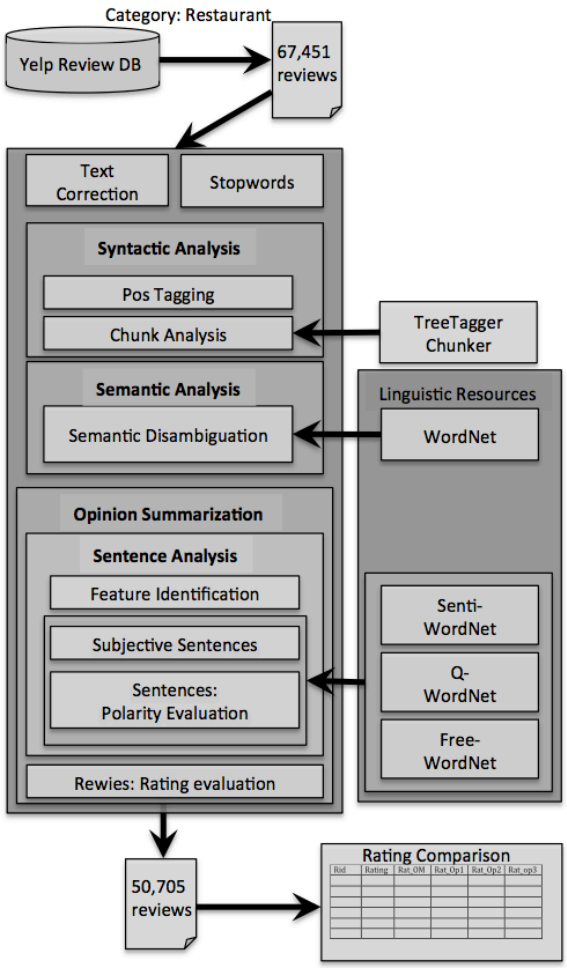

Figure 1: The Opinion Mining analysis

4.3 Feature Evaluation

Each sentence of the corpus of reviews has been analysed in order to find the association between features, adjectives and adverbs as shown in Table 1.

Table 1: Sample of feature, attribute and review relation.

Feature	ReviewSid	Attribute	POS	Card
staff	id11279s40	great	JJ	2

In the above example, the adjective *great*, tagged as JJ by the parser, is associated twice (cardinality = 2) to the feature *staff* belonging to the fortieth sentence of the review identified by the id 11279.

In this starting phase, all the synsets related to the features have been considered in the evaluation of the predictions. The overall value has been calculated as the mean of the three different lexical resources: SentiWordNet (Esuli and Sebastiani, 2006) Q-WordNet (Agerri and Garcia-Serrano, 2010) and FreeWordNet (Tuveri and Angioni, 2012).

At this point, a comparison between the polarities related to the same synsets in the three resources highlighted some discrepancies. For this reason we chose to consider the average of the three values.

In order to produce a more accurate evaluation of the polarity, during a further phase of the research activities, we intend to perform the semantic disambiguation. We will exclude the "not relevant" synsets, retaining only the meanings that are semantically related to the domain.

4.4 Reviews Analysis and Algorithm Evaluation

The Opinion Mining algorithm has been implemented considering the syntactic analysis in order to identify the parts of speech. The identification of the chunk structure for the sentences allowed the association of the feature terms with the related adjectives and adverbs. In this first step, no semantic disambiguation has been performed and a set of ratings has been produced.

We faced with the issue related to the representation of the rating values given by the Opinion Mining system, distributed in a range between -25 and 36, in order to compare them with the ratings of Yelp, ranged between 0 and 5. Also the two cumulative distributions were totally different.

Initially the values were linearly scaled on a rating system that ranges between 0 and 5 and then they have been associated to a cumulative distribution of the ratings similar to that of Yelp. This task has been performed acting on the threshold values shown in Table 2.

In order to assess the performance of the Opinion Mining algorithm, two researchers evaluated a collection of 200 reviews, after a previous tuning phase.

Table 2: The thresholds applied.

Thresholds	
Id	**Range**
T_0	x<1.2
T_2	1.2<=x<2.2
T_3	2.2<=x<3.2
T_4	3.2<=x<4.2
T_5	x>4.2

The choice of the data sample was based on the length of the content, assuming that longer reviews contain more information about the business considered.

The manual average rating provided by the researchers were considered as the most reliable. Hence, they were used to evaluate the Opinion Mining algorithm in terms of Precision (P), Recall (R) and F1-score (see Figure 2).

During a further phase of the experiment the features will be grouped in the 12 business aspects, as already anticipated in Section 4.2. The polarity evaluation of the synsets related to each feature will permit to calculate a rate for each aspect, as manually done by the two researchers.

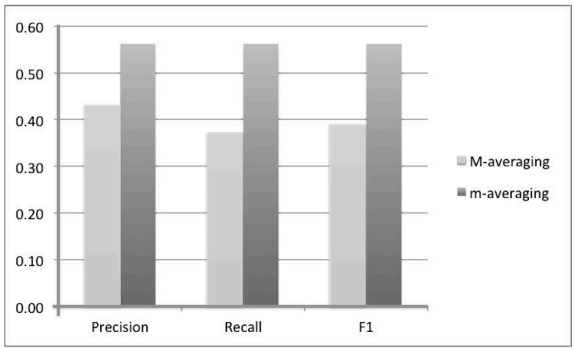

Figure 2: Performances of the Opinion Mining algorithm.

The rating of each review has been calculated as the sum of m partial ratings r_i, corresponding to the business features present in the user's textual review. This rating was penalized by a constant value of -0,2 for each of the n features not present, and divided by m (see equation 1), where $m+n = 12$.

In case the resulting total rating R is less than zero it will be set equal to 0.

$$R = \frac{\sum_{i=1}^{m} r_i - 0.2*n}{m} \qquad 0 < r_i \leq 5 \qquad (1)$$

The evaluation will be expressed again in terms of Precision, Recall, and F1-score in order to compare the new values with the previous ones (see Figure 2).

5 PREDICTION ANALYSIS

The prediction analysis was carried out through a 5-fold cross validation using three different algorithms independently: a Baseline made of average values, the Opinion Mining already described in section 4, and a Biased Matrix Factorization (BMF).

To evaluate the predictions coming from these algorithms the Root Mean Squared Error (RMSE) was used, calculated with the following well known formula:

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^{N}(P_i - r_i)^2}{N}} \qquad (2)$$

where N is the total number of reviews in the test set, P_i is the prediction for each of them, and r_i is the actual rating (which is known because all the data come from the training set of the original Yelp dataset).

Here after, the Baseline algorithm, the BMF algorithm, and the final ensemble with the results will be explained more in detail.

In the original dataset most of the star ratings given by the users were included in a small range across the average rating of all the reviews (3.68). For this reason, a good prediction to be used as baseline of each rating appeared to be made of the mean of the following two averages: average rating given by the user and average rating of the business. When one of the two average values was not available, the global average was used to replace it.

As already mentioned in Section 2, at present the Biased Matrix Factorization working alone is considered as the collaborative filtering algorithm able to produce the best predictions. For this reason it was applied in this study, using in particular a learning model based on the Stochastic Gradient Descendent (Koren et al, 2009).

The ensemble of the predictions output of Opinion Mining and BMF was performed by means of Linear Regression (LR).

At the conclusion of the first phase of work, the Opinion Mining produced a set of predictions, which in terms of RMSE resulted worse than the one from the Baseline algorithm (see Table 3). In any case the ensemble of the same predictions with the predictions coming from the BMF gave the best value of RMSE of the presented activity. In fact, ensemble methodologies are commonly applied for system recommendations because they are able to improve the results coming from the same algorithms working alone (Jahrer et al., 2010).

Nevertheless an opportunity for improvements in the predictions obtainable by the OM (and consequently also by its combination with the BMF) will be offered by the planned activity described in Section 4.4.

All the resulting RMSEs are shown in Table 3, with the best RMSE value produced through an ensemble of OM and BMF.

Table 3: Summary of the results obtained

Alg.	Baseline	BMF	OM	Ens. LR
RMSE	1.02593	1.00859	1.25011	0.99401

6 CONCLUSIONS AND FUTURE WORKS

The Yelp dataset providing both user-generated ratings and textual reviews allows interesting research activities related to the combination of Opinion Mining and Collaborative Filtering.

The presented Opinion Mining approach was used to analyse the textual reviews and to produce predictions to be compared with the manual evaluations made by a small group of people. In order to improve the results in terms of RMSE, in a further work a deeper syntactic analysis will be carried out along with the semantic disambiguation of the textual reviews. Better criteria of evaluation through the introduction of a set of 12 business aspects are expected to provide an important improvement of the results. The Opinion Mining approach will be again evaluated comparing the output in terms of predictions with the values manually assigned.

ACKNOWLEDGEMENTS

This study is part of a POR-FESR 2007–2013 project co-funded by the Autonomous Region of

Sardinia: Comunimatica (PIA n. 205 co-funded according to the DGR 39/3 del 10.11.2010).

REFERENCES

Agerri, R., Garcia-Serrano A., 2010. Q-WordNet: Extracting polarity from WordNet senses. In LREC 2010, 7th International Conference on Language Resources and Evaluation, Malta.

Angioni, M., Demontis, R., Tuveri, F., 2008, A Semantic Approach for Resource Cataloguing and Query Resolution. Communications of SIWN. Special Issue on Distributed Agent-based Retrieval Tools.

Angioni, M., Tuveri F., 2011, A Semantic Approach to the Extraction of Feature Terms. ICSOFT 2011, 6th International Conference on Software and Data Technologies. SciTePress.

Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D., Venkatramana S. Subrahmanian, 2007. Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone. Proceedings of ICWSM 07, International Conference on Weblogs and Social Media, pp. 203-206.

Ding, X., Liu, B., Yu, P.S., 2008, A Holistic Lexicon-Based Approach to Opinion Mining. WSDM '08 Proceedings of the international conference on Web search and web data mining, ACM New York, USA

Esuli, A., Sebastiani, F., 2006, SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), p. 417–422, Genova, Italy.

Fan, Mingming; Khademi, Maryam, 2014. Predicting a Business Star in Yelp from Its Reviews Text Alone. ArXiv e-prints: 1401.0864.

Govindarajan, M., 2014, Sentiment Analysis of Restaurant Reviews Using Hybrid Classification Method, International Journal of Soft Computing and Artificial Intelligence, Vol. 2, Issue 1.

Jahrer, M., Töscher, A., Legenstein, R., 2010, Combining Predictions for Accurate Recommender Systems, Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 693-702, ACM, 2010.

Koren, Y., Bell, R., Volinsky, C., 2009, Matrix Factorization Techniques for Recommender Systems, Computer, IEEE Computer Society, v. 42, n. 8.

Koukourikos, A., Stoisis, G., Karampiperis, P., 2012. Sentiment Analysis: A tool for Rating Attribution to Content in Recommender Systems. Presented at the 2nd Workshop on Recommender Systems for Technology Enhances Learning (RecSysTEL 2012), 18-19/09/2012, Saarbrucken, Germany

Levi, A., Mokryn, O., Diot, C., Taft, N., 2012. Finding a needle in a haystack of reviews: cold start context-based hotel recommender system. In Proceedings of the sixth ACM conference on Recommender systems, pages 115–122. ACM, 2012.

Magnini, B., Strapparava, C., 2004, User Modelling for News Web Sites with Word Sense Based Techniques. User Modeling and User-Adapted Interaction 14(2), pp. 239-257.

Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A., 2002. The Role of Domain Information in Word Sense Disambiguation. Natural Language Engineering, special issue on Word Sense Disambiguation, 8(4), pp. 359-373, Cambridge University Press.

Miller, G., 1998. WordNet: An Electronic Lexical Database, Bradford Books

Quadrana, M., 2013. E-tourism recommender systems http://hdl.handle.net/10589/84901

Schmid, H., 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of the International Conference on New Methods in Language Processing, pp. 44-49.

Tosher, A., Jahrer, M., Bell, R. M., 2009, The BigChaos solution to the Netflix grand prize, Netflix Prize Documentation.

Trevisiol, M., Chiarandini, L., Baeza-Yates, R., 2014, Buon Appetito - Recommending Personalized menus.

Tuveri, F., Angioni, M., 2012. A Linguistic Approach to Feature Extraction Based on a Lexical Database of the Properties of Adjectives and Adverbs, Global WordNet Conference (GWC2012), Matsue, Japan.