Combining Opinion Mining with Collaborative Filtering

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

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

Abstract: An experimental analysis of a combination of Opinion Mining and Collaborative Filtering algorithms is presented. The analysis used the Yelp dataset in order to have both the textual reviews and the star ratings provided by the users. The Opinion Mining algorithm was used to work on the textual reviews, while the Collaborative Filtering worked on the star ratings. The research activity carried out shows that most of the Yelp users provided star ratings corresponding to the related textual review, but in many cases an inconsistence was evident. A set of thresholds and coefficients were applied in order to test a hypothesis about the influence of restaurant popularity on the user ratings. Interesting results have been obtained in terms of Root Mean Squared Error (RMSE).

1 INTRODUCTION

Nowadays users have the possibility to express their opinions about products or services by a global rating, according to their experience. The overall rating is very important, as it represents the electronic 'word of mouth' that customers have about a product.

Moreover most websites, like Yelp, allow their customers to better explain their opinion of the product by more detailed textual reviews.

Many existing Recommender Systems are based only on users' overall ratings about items, but do not consider and do not work on the opinions expressed by the users about the different aspects of an item. As a result, the rate does not wholly summarize the opinion of the users, maybe ignoring important information.

From the point of view of the Opinion Mining the most recent studies focus on detailing such information in order to gain knowledge more closely reflecting the complexity of businesses, products and services contexts. While Recommendation Systems are currently mature technologies, the ones related to Opinion Mining are not yet able to provide reliable solutions beyond the research contexts.

In this paper, we propose an experimental analysis of a combination of Opinion Mining and Collaborative Filtering algorithms applied to the Yelp dataset of businesses. The analysis used this particular dataset in order to have both the textual reviews and the star ratings provided by the users. Opinion Mining was used to work on the textual reviews, while Collaborative Filtering worked on the star ratings.

As Pang and Lee affirm in (Pang and Lee, 2008) at least one related set of studies claims that "the text of the reviews contains information that influences the behaviour of the consumers, and that the numeric ratings alone cannot capture the information in the text" (Ghose and Ipeirotis, 2007).

A fundamental aspect in Yelp dataset is given by the fact that there is sometimes a discrepancy between the information written by a user in a textual review about a certain restaurant and the star rating.

This fact has been analysed by means of a manual evaluation of the reviews, as described later in Section 4.3.

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, the data extraction and the related issues. Section 4 describes the Opinion Mining analysis process while the prediction analysis methodology is described in Section 5. Section 6 describes the experimental setup and the results for the proposed approach. Lastly Section 7 reports conclusions and future

Angioni M., Clemente M. and Tuveri F..
Combining Opinion Mining with Collaborative Filtering.
DOI: 10.5220/0005412403700380
In *Proceedings of the 11th International Conference on Web Information Systems and Technologies* (WEBIST-2015), pages 370-380
ISBN: 978-989-758-106-9

works.

2 RELATED WORK

Several studies have been written describing the combination of Opinion Mining and Collaborative Filtering.

Collaborative Filtering techniques aim to predict the preferences of users providing suggestions of further resources or entities that could be of interest.

The most popular commercial services on the web have demonstrated that user profiling is able to improve the revenues. For this reason the research in the field of user profiling and recommender systems have been developed at a very high speed in the last ten years. The most effective algorithms used by commercial services are defined as Collaborative Filtering, which can take as input a simple matrix of recommendations given by the users (the rows of the matrix) to the items (the columns). As stated in (Koren, Bell & Volinsky, 2009) the most important families of Collaborative Filtering algorithms are the neighbourhood methods (deriving from k-Nearest Neighbour) and the latent factor models (which are based on the factorization of the matrix of recommendations).

Important results have been developed thanks to the 1 million dollars Netflix Prize, a competition started in 2006 by Netflix, the well-known dvd-rental company, for an algorithm able to increase by 10% the accuracy of Cinematch, the algorithm used at the time by Netflix for movie recommendation.

The effect of this competition was to multiply the number of researchers involved in the topic, the number of related conferences, and most importantly the quality of the collaborative algorithms used by recommender systems. The million was won in 2009 by a combination of three different teams and their algorithms: item-based (a kind of kNN) (Sarwar et al, 2001), Restricted Boltzmann Machine (RBM) (Hinton, 2012), and Biased Matrix Factorization (Koren, Bell & Volinsky, 2009).

While before the Netflix competition the item-based algorithms were considered the most effective for recommender systems, and in fact at the time they were used also by Amazon (Linden, Smith & York, 2003; Clemente, 2008), during the competition it has been demonstrated that the matrix factorization algorithms, working alone, were the most effective for this kind of problems (Koren, Bell & Volinsky, 2009; Tosher, Jahrer & Bell, 2009).

Although many types of algorithms can be used in the field of recommender systems, each of them has limitations, but these limitations change from one algorithm to another. It has been experimented that generally ensemble methodologies allow obtaining a blending prediction, which improve the ones coming from each of the algorithms singularly taken (Jahrer, Töscher & Legenstein, 2010).

While algorithms based on user ratings produce interesting results, they do not consider qualitative information, like the actual opinion of a user about a resource and whether or not he/she actually would propose it to other users (Koukourikos et al., 2012). Moreover, explicitly given user ratings do not consider the different features of a resource and the weight that the users give to each of them, more or less unknowingly.

On the other hand, feature-based Opinion Mining can be a very valuable resource to improve Collaborative Filtering performances, by adding qualitative information to explicit user ratings (Quadrana, 2013).

(Levi et al., 2012) proposed an interesting context-aware recommender system that uses Opinion Mining in order to analyse hotel reviews and to organize user tastes according to some users' preferences and to provide better recommendations in the cold-start phase.

Another study related to a particular combination of Opinion Mining and Collaborative Filtering is (Wu & Ester, 2015), where the textual reviews are analysed in order to be able to predict the level of interest of each user about different aspects of an item (representing a more detailed prediction than the single number of the predicted rating).

A unusual combination of Collaborative Filtering and Opinion Mining is described in (Singh et al, 2011), where the output of an item-based collaborative filtering is further filtered by two different OM approaches.

A common problem to the user-generated reviews is usually related to the inconsistency in terms of length, content, treated aspects and usefulness because not every user writes about all the relevant aspects which characterize a business activity. For this reason relevant information would be disregarded, causing a lack of useful data in the input of the Opinion Mining algorithm.

3 YELP DATASET

The dataset chosen for the presented activity is the one made available by the Yelp social network (http://www.yelp.com) for the RecSys Challenge 2013 "Yelp business rating prediction". An

important feature of this particular data set is that it provides not only the star ratings (from 1 to 5 stars) assigned by the users to the business located in the Phoenix (AZ) metropolitan area, but also a textual review (along with many more information about users and business). This feature makes the Yelp dataset suitable for research in the fields of machine learning algorithms, which work on these two different types of information (Huang, Rogers & Joo, 2014).

Only the training set for the competition was considered because in the test set the actual ratings were obviously missing.

The original training set was made of the following information:

- 229,907 reviews
- 43,873 users
- 11,537 business

For the aim of the presented activity the Restaurant category was chosen, which was the most represented in the original dataset. Restaurants represent one of the most considered items in recommender systems (Burke, 2002; Ganu et al., 2012; Ganu et al., 2009) and also the Restaurant category in Yelp dataset (Trevisiol et al., 2014; Huang et al., 2014; Govindarashan, 2014). It must be specified that more than one Yelp available dataset exists because there have been more than one competition providing each time a different version.

Actually each business in the data set has a list of categories, but in the one used for our activity, the word Restaurant is always present.

Another interesting aspect is the distribution of the different number of stars (value of the ratings):

- five stars count = 14831
- four stars count = 24600
- three stars count = 12464
- two stars count = 6275
- one stars count = 2938

3.1 Data Extraction

We collected our data from the Yelp Dataset, considering only the users giving a number of reviews greater than 9, as more reliable. We target the most famous category in the set, Restaurants, and extracted 67,451 text reviews.

We did a spell check on the obtained reviews and then a transformation of the contracted forms of verbs in order to avoid introducing errors and to facilitate the syntactic parser activities.

In fact, in the case of a sentence such as "We didn't have a fridge in our room", the parser was not able to correctly identify the contracted verb form *didn't*. So, before parsing the text, some pre-processing steps related to the verbs were necessary, replacing the contracted forms into the long forms: *didn't* became *did not, I've* became *I have*, *I'll* became *I will*, and so on.

The reviews have been divided into sentences, obtaining a number of 953,314 of them.

The phrase parser chunking process has been carried out by TreeTagger (Schmid, 1994), annotating the sentences with part-of-speech tags and lemma information and identifying in each sentence its sub-constituents.

A Java class wraps the evaluation provided by TreeTagger and, analyzing the parts of speech, identifies the associations between nouns and their related information.

The "sentence analysis" (see Figure 1) includes the result of the previous syntactic analysis, manages the feature extraction, and then uses the linguistic resources in order to calculate the polarity values of each sentence. As results, the Sentence Analysis provides the categorization of each sentence of the reviews in order to distinguish between subjective and objective sentences, with or without orientation, and in particular in order to detect factual sentences having polarity value. In such a way, we consider only subjective sentences or factual sentences having polarity valence. The set of 953,314 sentences has been so reduced to about 394,000 subjective sentences bringing the entire set to the number of 50,705 reviews.

To achieve this task, we made use of SentiWordNet (Baccianella et al., 2010) a lexical resource that assigns to each synset of WordNet (Miller, 1998) three sentiment scores: positivity, negativity, objectivity, and we considered only the sentences containing adjectives with a polarity valence.

Opinion Mining analysis, as better described in Section 4, produced a set of rating predictions about the business activities to be compared with the Yelp ratings. Two researchers have manually evaluated a collection of 200 reviews to check the validity of the related ratings, using a common evaluation criterion. The comparison with the ratings manually assigned by the researchers, and described in detail in Section 4.3, allowed the evaluation of the performance of the Opinion Mining methodology.

Finally, the ratings coming from the Opinion Mining, combined with the user ratings (Yelp ratings), have been used by the ensemble algorithms.

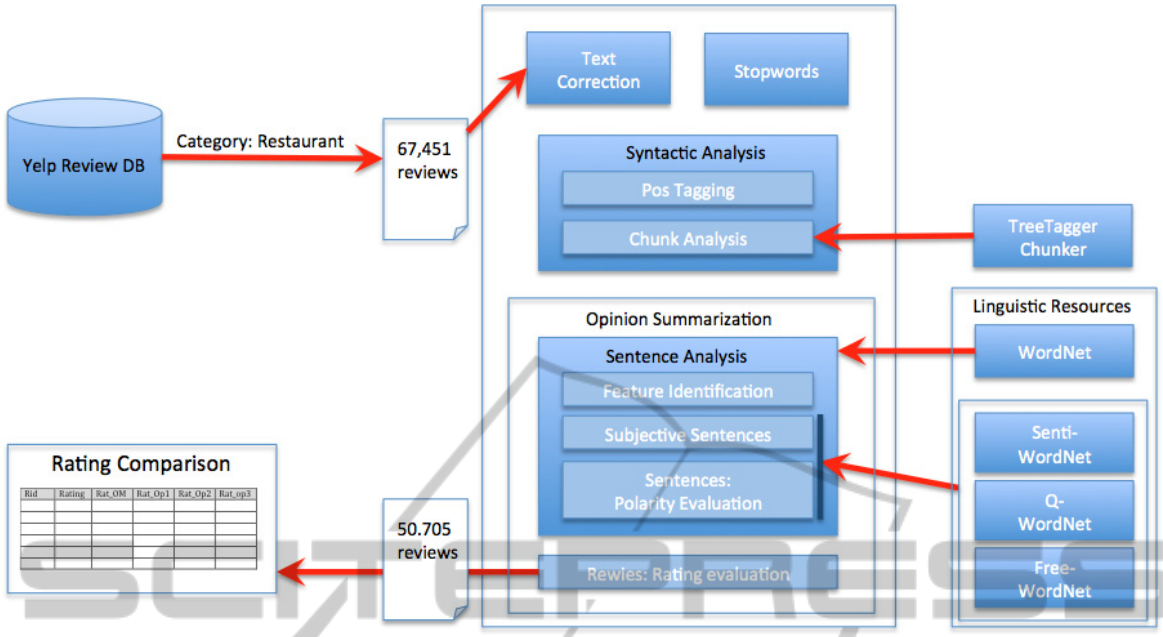

Figure 1: The Opinion Mining analysis.

3.2 Issues Related to the Dataset

It is important to consider that since the textual reviews have been written without restrictions, they resulted affected by some limitations. Here after some examples of such limitations are explained.

The users could choose to talk about any of the aspects related to the restaurants (restaurant location, interior design, parking area, quality of service, quality/variety/amount of food, quality/variety of wine, prices, entertainment/live music, and intention to come back). This caused that some users talked about almost all of them, while many others limited their review to the quality of food.

Another problem is related to the fact that in the same review about a restaurant, some users describe dissenting opinions on different occasions (for example they were enthusiastic after the first time they went to a particular restaurant, but a more recent experience has made them change their opinion), or make comparisons between different restaurants (something like: "in this place the enchilada is not as good as the one available at …").

Although Yelp applies an algorithm in order to filter out all the reviews posted by people related in some way to the business referenced by the review (such as the owner of the business, a relative of the owner or a person working there), it must be assumed that not all the reviews are spontaneous. This problem has caused also some lawsuits (Clark, 2013), and obviously Yelp will always be affected by phony reviews.

Jong in (Jong, 2011) faces with the problem of the Yelp dataset in which the star ratings rarely provide the most objective or the fairest rating. In fact, most of the stars range from 3.5 to 4.5 stars with very few ratings below or above, resulting meaningless. In their study, (Mingming et al., 2014) put in evidence the differences of evaluation of distinct users (Michelle and Clif) who wrote about "Providence", a restaurant in LA area. Both users described their experience as very good using multiple positive words such as "perfection", "must go", "great treat", "tasted great", etc. However, the first user gave five stars to the restaurant whereas the second user gave only three stars.

Nevertheless most of the reviews can be considered reliable and this is the reason why Yelp has become so popular during the years.

4 OPINION MINING

Opinion Mining has been introduced in the presented activity to predict a business' rating based on textual reviews to be compared to the Yelp ratings.

As in (Benamara et al., 2007), we propose a linguistic approach to Opinion Mining and, more in details, to the automatic extraction of feature terms by means of the syntactic and semantic analysis of textual resources. We focus on the analysis of the

opinions through the processing of textual resources, the information extraction by means of the syntactic chunk analysis, and the evaluation of a semantic orientation.

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

Many approaches to Opinion Mining are based on linguistic resources, lexicons or lists of words, used to express sentiments or opinions and are used for the identification of the polarity of words and their disambiguated meanings.

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

4.1 Feature Extraction

The feature extraction 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.

We extracted the features by the textual reviews expressed by the users.

Considering that the domain is well known, the identification of the features for the Yelp reviews has been performed evaluating the nouns frequency in the text through a word counter. We first removed the stop words and then the cleaned text was tokenized obtaining as a result a collection of about 4000 words, including individual and compound words.

We condensed this set by only considering words with a frequency greater than 100, in order to test the potential of the proposed approach, to be extended in a future work.

Finally, we identified the nouns as candidate features. The features were then manually validated and separated into six aspects: Food, Service, Staff, Ambience, Location and Price. As a result we obtained about 935 features. In a further development of this study we will also consider the verbs.

Although the reviews have been analyzed through the features they come with, we did not consider any criterion to evaluate them, putting at the same level each of the six aspects of the business. The evaluation of the reviews instead relies on the simple sum of the values of polarity associated with the terms they contain and on the identification of chunks, such as adverb + adjective, negations, and superlatives in their sentences.

For example, chunks can be considered as "not bad" or "very very good".

4.2 Feature Evaluation

Each sentence of the corpus of reviews was analysed and the association between features with adjectives and adverbs was found:

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

feature	reviewSid	attribute	pos	Card
Staff	id112795s40	great	JJ	2

In the above example, the adjective great (JJ) is associated twice (cardinality = 2) to the feature staff belonging to the fortieth sentence of the review identified by the id 112795. The polarity of each attribute is calculated evaluating for all the synsets related to the term in WordNet the polarity associated to the synset in three different lexical resources: SentiWordNet, Q-WordNet and FreeWordNet.

SentiWordNet expands WordNet 2.0 and associates to each synset three numerical scores describing how much Objective, Positive and Negative are the terms related to that synset. This means that a synset may have nonzero scores for all the three categories.

Q-WordNet (Agerri and Garcia-Serrano, 2010) is a lexical resource consisting of WordNet senses automatically classified by Positive and Negative polarity. Polarity evaluation has been used to decide whether a textual content is associated to a positive or negative connotation.

FreeWordNet (Tuveri and Angioni, 2012) is another lexical database of synsets defined as extension of a subset of adjectives and adverbs of WordNet. Each synset has been enriched with a set of properties concerning the polarity and other properties according to a set of attributes identified by their association with nouns and verbs and chosen on the basis of their frequency of use in the language.

The Opinion Mining system has produced a set of rating predictions affected by the choice of the lexical resource. In some cases the three resources have produced discrepancies of polarity related to the same synset. For this reason we chose to consider the average of the three values obtained from the assessment made by the Opinion Mining system with the three resources.

4.3 Reviews Analysis and Algorithm Evaluation

Regarding the analysis of the reviews we faced with the issue related to the representation of the rating values given by the Opinion Mining system in order to compare them with the ratings of Yelp.

In fact, the values produced by the Opinion Mining system were not directly comparable with the ratings of Yelp, because they were distributed in a range between -25 and 36. Also the distribution of the ratings obtained was totally different if compared with the Yelp one. Several transformations were possible, here after we describe the one we chose. The values have been initially linearly scaled on a rating system that ranges between 0 and 5. As a first step, we chose the transformation, which produced a distribution of the Opinion Mining values similar to the distribution of the Yelp ratings. The introduction of the thresholds, shown in Table 2, gave us the opportunity to assign to the rating classes a number of reviews similar to the Yelp ones.

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

These values produced the distribution depicted in Figure 2.

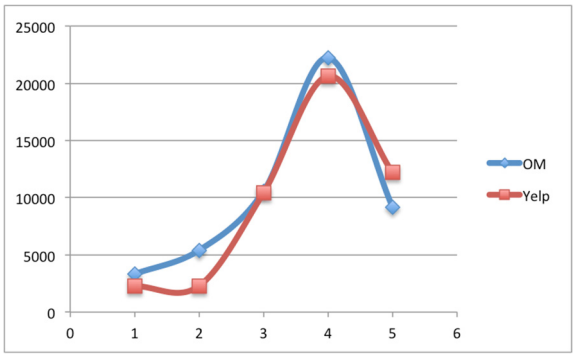

Figure 2: Ratings distribution.

As already mentioned in Section 1, in order to evaluate the performance of the Opinion Mining algorithm, two researchers have manually evaluated a collection of 200 reviews. A preliminary tuning phase was carried out on a limited number of reviews in order to agree on a common evaluation criterion.

The choice was based on the length of the text, assuming that longer reviews contain more information.

The methodology of evaluation of the reviews was based on a set of sub-aspects of the original 6 aspects, previously introduced in Section 4.1, plus the "intention to come back" to the restaurant. Each sub-aspect was independently evaluated by the two researchers with values ranging between 0 and 5, while the aspects disregarded by the author of the review were penalized by a value of 0.2. We wanted to penalize the sub-aspects not covered in the review because, although they were not negatively considered, their absence from the description meant that they had not positively impressed the customer anyway. The average rating provided by the reviewers was finally used in order to evaluate the algorithm in terms of Precision (P), Recall (R) and F1-score (see Figure 3).

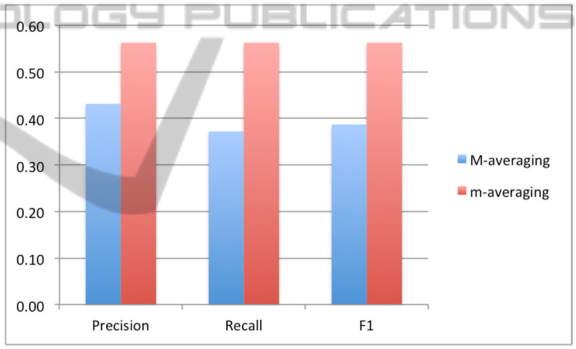

Figure 3: Opinion Mining system evaluation.

Figure 3 illustrates the evaluation of the Opinion Mining system according to the values of threshold shown in Table 2 in terms of micro and macro averaging.

During the Opinion Mining algorithm evaluation, it was noticed that the star rating not always appeared in line with the content of the review. These inconsistencies were shown up throughout the manual analysis of the reviews, and evidenced by the discrepancies between the star ratings assigned by the Yelp users and the manual rating given by the researchers.

The nature itself of the star rating does not depict a detailed experience or does not express emotions and feelings, which are instead described by the several aspects covered by textual reviews.

Let us use an example to illustrate this concept considering a specific review having 4 as star rating, while the two researchers gave respectively a rating

of 2,58 and 2,27, with an average value of 2,43 as shown in Table 3.

The following text has been extracted from the review and analysed according to the aforementioned 11 aspects and sub-aspects: food-quality, food-quantity, food-variety, food-beverages, food-desserts, service, staff, ambience (atmosphere), location-parking, location-bar, price, plus the intention to come back.

Here after we discuss some of them in detail.

Table 3: The ratings, as evaluated by the two researchers.

	Aspects	Res. 1	Res. 2	Avg.Rate
1	Food Quality	3,0	1,0	2,0
2	Food Quantity	4,0	4,0	4,0
3	Food Variety	3,0	3,0	3,0
4	Food Beverages	-0,2	-0,2	-0,2
5	Food Desserts	-0,2	-0,2	-0,2
6	Service	2,4	2,8	2,6
7	Staff	2,5	2,0	2,25
8	Ambience	0,2	0,1	0,15
9	Location-Bar	-0,2	-0,2	-0,2
10	Location-Parking	-0,2	-0,2	-0,2
11	Price	-0,2	-0,2	-0,2
12	Come Back	4,0	4,0	4,0
	Total	**2,58**	**2,27**	**2,43**

Food: Quality, Quantity, Variety

- *"I had the Chicken Tikka Masala and my friend had the Chicken Pot Pie - both were delicious! I was super impressed with the breadth of the pasty 'stuffings' and got very excited to see they had over 40 options! I want to go back and try them all. The yogurt served was perfect.. the chicken was plentiful and well seasoned…their fillings are both inventive, somewhat unique and really entice me to come back."*
- *"I appreciated that their pastys are a good size…their pastys were yummy"*
- *"I found a piece of red thick rubber-band in my soup."*

Service

- *"Yes, this was not a short lunch... but it was also not a long lunch. Food took a bit long but we also were expecting 'custom-baked' pasty."*
- *"their service was excellent"*

Staff

- *"Our server was quick enough to bring us drinks and my soup order, etc., to hold us over."*
- *"Our server guy was outstanding - friendly in a genuine way, prompt, kept checking on us, and had very good customer service skills."*
- *"I did note one of the food preppers talking on his cell phone while he was handling food. I thought that was gross…"*
- (about the "red thick rubber-band in my soup") *"Yeah, not excited about that find, but I did appreciate that our server immediately apologized... fixing a problem immediately and correctly - kudos for him for a great response and showing good customer service."*
- *"Can understand that 'shit happens' sometimes but it's the aftermath of how you treat the customer that found the rubber band in their soup that matters"*

Ambience (Atmosphere)

- *"Smells: Yes, but we deduced it was the cabbage."*
- *"It needs a good cleaning … there is a distinctly strong smell"*
- *"Those pew cushions were nasty and the wax should be scraped off and menus cleaned up"*
- *"The pews had dirty cushions on them…the candle wax was all over the table/menus, the menus had other grime on there…this place REALLY needs to quit the slacking and clean this place up."*

Intention to come back

- *"I'll be back!"*

Not mentioned:

- Location-Bar;
- Location-Parking;
- Price;
- Food-Beverages;
- Food-Desserts;

The obtained ratings are the algebraic sums of each sub-aspect divided by the number of mentioned sub-aspects. The average rating is used, as said, in order to evaluate the algorithm.

You can notice that, even if the user said that he would go back there, the lack of cleanness, the strong smell, the slow service, the presence of a rubber in the food, are so negative elements that it is incredible that he/she could assign a high rating.

5 PREDICTION ANALYSIS

Predictions for the test set were computed by means of three different algorithms singularly run:

1. The Baseline algorithm made of average ratings, described in Section 5.2
2. Opinion Mining, described in Section 4
3. Biased Matrix Factorization, described in

Section 5.3

Then a RMSE was calculated for each different set of predictions singularly taken and compared with an ensemble of algorithms 2 and 3, as described later, in Section 5.4.

5.1 Thresholds

As already mentioned, during the experimental activity, it was evident that the output in terms of predictions coming from the Opinion Mining algorithm were not always aligned with the star ratings. In particular while working at the activity related to the manual check of the Opinion Mining predictions (described in Section 4.3) against the actual ratings, most of the times the star rating overestimated what the same user expressed in words.

This inconsistency between the textual review and the star rating appeared to be an interesting behaviour and a possible explanation brought to think that maybe many users were influenced by the average rating of the business (which in the Yelp web site is obviously well shown).

In order to test this hypothesis, an experimental analysis was carried out applying some coefficients to the predictions obtained by the Opinion Mining; during this activity the thresholds already used during the calculation of the predictions were applied (see Table 2): T0 = 1.2, T1 = 2.2, T2 = 3.2, and T3 = 4.2. In particular, the predictions have been multiplied by a coefficient, but only under the condition that the average business rating (BRT) was greater than a certain value depending on the threshold.

A schema of the thresholds is shown in Figure 5.

5.2 Baseline Algorithm

The baseline algorithm chosen for the activity was run using a 5 fold cross-validation method and based on the following averages calculated for each user and business related to a rating:

- average rating of user i (avg_u_i), the average of all the ratings in the training set given by user i (u_i)
- business average (avg_b_j), the average of all the ratings in the training set received by the business j (b_j)
- global average (global_avg), the average of all the ratings in the training set (3.6891).

In particular, when both the user and the business were present in the training set, the prediction $p(u_i,b_j)$ of the star rating that the user i (u_i) could give to a business j (b_j) was calculated as the following weighted average:

$$p(u_i, b_j)=(avg\_u_i * w_1+avg\_b_j * w_2)/2 \quad (1)$$

where w_1 and w_2 are the weights, described in Section 6.1.

When only the user was known in the training set the prediction was calculated as:

$$p(u_i, b_j)=(avg\_u_i * w_1+global\_avg * w_2)/2 \quad (2)$$

When only the item was known in the training set:

$$p(u_i, b_j)=(avg\_b_j * w_1+global\_avg * w_2)/2 \quad (3)$$

And lastly, when both the user and the business were not present in the training data (the famous *cold start* problem), the prediction was set equal to the global average.

5.3 Biased Matrix Factorization

As already recalled in the Introduction, an effective latent factor model is represented by the biased matrix factorization, which is based on the fact that each review and each rate is influenced by a certain number of latent factors not known. These factors can be inferred by the algorithm, although the scope of the presented activity was limited to improve the quality of the predictions, this analysis would be very interested and it is our intention to develop it in a future work.

During the presented experimental activity the Mahout Taste library (Owen et al., 2011) was used; in particular, the Stochastic Gradient Descent Factorizer as learning algorithm, which is an implementation of the Biased Matrix Factorization algorithm described in (Y. Koren et al. 2009; Paterek, 2007), while the Singular Value Decomposition was used as Recommender.

The research activity considered different values of the following parameters through a 5-fold cross validation: number of features, number of iterations, learning rate, regularization constant, random noise, and learning rate decay.

The cases of unknown restaurants or unknown users were dealt with averages analogue to the ones used in the *baseline* algorithm already described in section 5.2.

5.4 Ensemble Methods

As already mentioned in section 2, ensemble methodologies allow improving the results coming by multiple algorithms because typically they are weak in different ways and their combinations produce a more robust and generalizable solution.

For this reason, in the presented activity some ensemble methods were applied in order to combine the output in terms of predictions coming from the Opinion Mining and the Biased Matrix Factorization algorithms. For this task the *sklearn* python library was chosen (scikit-learn.org), which is particularly easy to use and well documented. Moreover it provides the same API for all the classes performing regressions. The regressions applied in the presented activity are the followings:

- Linear Regression, which finds the best fitting line minimizing the sum of the squared errors of the predictions;
- Ridge regression, which differs from a linear regression because it applies a 'ridge' penalty to reduce the variance of the values;
- Gradient Boosting Regression Trees (GBRT), which uses decision trees as weak learners and is extensively used also for its predictive power.

In figure 4 the scheme of the ensemble methodology is shown. The training dataset has been split into 5 folds to be used as input of both Opinion Mining and Collaborative Filtering algorithms.

The ensemble methods were used to merge these 5 output and so 5 different RMSE values have been obtained. Then, the average of these 5 values has been considered as the final result.

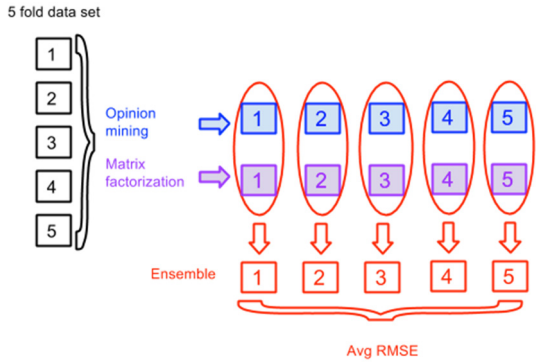

Figure 4: Scheme of cross validation and ensemble.

6 RESULTS IN TERMS OF RMSE

The Root Mean Squared Error (RMSE) is a very common way to evaluate the quality of predictions for recommender systems and in fact it is greatly used in competitions and related leader boards (such as Netflix prize, available at www.netflixprize.com/leaderboard, RecSys2013 available at www.kaggle.com/c/yelp-recsys-2013/leaderboard, etc.). It amplifies large errors and provides the advantage of concentrating the result in a single parameter.

Since the errors are squared, negative and positive errors do not cancel each other. Smaller RMSE values correspond to better results.

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^{N}(P_i - r_i)^2}{N}} \quad (4)$$

In the above formula, P_i is the prediction for each of the N reviews in the data used as test set, while r_i is the actual rating (since the data set used was a training set, the actual ratings were known).

In the presented study the RMSE was used to evaluate the quality of the predictions coming from Baseline, Opinion Mining, and Biased Matrix Factorization algorithms.

The same evaluation was used to analyse the ensemble Opinion Mining with Biased Matrix Factorization as well.

6.1 Baseline

The algorithm used as Baseline, described in section 5.2, was initially run giving the same weights to the user average (avg_u_i), the business average (avg_b_j) and the global average (global_avg). Further experimental analysis brought to the choice of penalizing the user average contribution, and in the end the best values were $w_1 = 0.4$ and $w_2 = 0.6$.

Since most of the actual ratings of the dataset are included in the range between 3.5 and 4.5 stars, the baseline could produce a RMSE value of 1.0259, which was hard to be outperformed for this particular dataset and for this reason represented a good reference point for the study.

6.2 Opinion Mining

In terms of RMSE the predictions, which were originally output of the Opinion Mining methodology, did not outperform the Baseline predictions, giving a value of 1.25011. But as already stated, a more attentive analysis of this result induced to work with the set of thresholds in order to apply some coefficients to take into account the influence under which most users had expressed the star rating, due to its aforementioned inconsistency with the content of the textual reviews. The set of thresholds applied and their values have been schemed in Figure 5.

The application of these thresholds to the Opinion Mining (OM+T) caused a change in almost the 50% of the original predictions. The new value of RMSE was 1.00548, which greatly outperformed the baseline, but unexpectedly did slightly better

than the BMF algorithm as well.

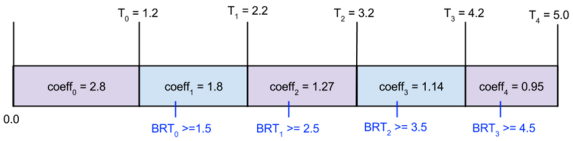

Figure 5: The thresholds used on OM predictions.

6.3 Biased Matrix Factorization

As already explained in section 5.3, the experimental analysis on the Biased Matrix Factorization was carried out through a 5-fold cross validation and involved many different configurations depending on the values given to the parameters taken as input by the factorizer (the RatingSGDFactorizer).

The two matrixes output of the factorizer, in a future work, will be the subject of a further analysis in order to look for correlations with the aspects and sub-aspects explained in Section 4.3.

The parameters which provided best results in terms of RMSE are the followings:

- Number of features = 14
- Number of iterations = 75
- Learning rate = 0.0025
- Regularization constant = 0.02
- Random noise = 0.01 (the default value)
- Learning rate decay = 1.0 (the default value)

With this configuration the resulting RMSE was 1.00859.

6.4 Ensemble

The Gradient Boosting Regression Tree, the Linear Regression, and the Ridge Regression produced better RMSEs than each of the predictive algorithms, singularly taken.

In particular, the best value of RMSE with the different ensemble algorithms was obtained by the GBRT that, as expected, was also the better result of all the presented experimental analysis, as summarized in Table 3.

Table 4: Summary of the best RMSEs obtained.

Alg.	Baseline	BMF	OM	OM+T	Ens. GBRT
RMSE	1.02593	1.00859	1.25011	1.00548	**0.98874**

7 CONCLUSIONS AND FUTURE WORK

Most existing Recommender Systems are based only on users' overall ratings about items, but do not consider and do not work on the opinions expressed by the users about the different aspects of an item. As a result, the rate does not wholly summarize the opinion of the users, maybe ignoring important information.

In order to overcome this problem a research activity about possible combinations of Opinion Mining and Collaborative Filtering has been carried out.

The encouraging results obtained in terms of RMSE seem to confirm the hypothesized influence of the average business rates on the users in the choice of the number of stars to set as rate.

We would like to further develop this study in many ways: regarding the evaluation of the textual reviews, we would like to apply an algorithm able to calculate the ratings through the estimation of the aspects described in Section 4.3; regarding the Opinion Mining, we would like to improve the syntactic and semantic analysis; in relation to the Collaborative Filtering, we are interested to carry out an analysis of the latent factors in order to find their correlations with the aspects and sub-aspects characterizing the businesses.

ACKNOWLEDGEMENTS

This study is part of a POR FESR 2007-2013 project co-funded by the Autonomous Region of Sardinia: *Comunimatica* (P.I.A. n. 205 co-funded according to the DGR 39/3 of 10/11/2012).

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.

Baccianella, S., Esuli, A., Sebastiani, F., 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC 2010, *7th International Conference on Language Resources and Evaluation*, Malta, pp. 2200-2204.

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.

Burke, R., 2002. Hybrid Recommender Systems: Survey and Experiments, User Modeling and User-Adapted Interaction, vol. 12, n. 3, pp. 331-370.

Clark, P., 2013. Yelp's Newest Weapon Against Fake Reviews: Lawsuits, http://www.businessweek.com/articles/2013-09-09/yelps-newest-weapon-against-fake-reviews-lawsuits.

Clemente, M. L., 2008. Experimental Results on Item-Based Algorithms for Independent Domain Collaborative Filtering, *Proceedings of AXMEDIS '08*, IEEE Computer Society, pp. 87-92.

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, NY, USA.

Ganu, G., Elhadad, N., Marian, A., 2009. Beyond the Stars: Improving Rating Predictions using Review Text Content, *Twelfth International Workshop on the Web and Databases (WebDB 2009)*, Providence, Rhode Island, USA.

Ganu, G., Kakodkar, Y., Marian, A., 2012, Improving the quality of predictions using textual information in online user reviews", Information Systems, http://dx.doi.org/10.1016/j.is.2012.03.001.

Ghose, A., Ipeirotis, P. G., 2007. Designing novel review ranking systems: Pre- dicting usefulness and impact of reviews, in *Proceedings of the International Conference on Electronic Commerce (ICEC)*.

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

Hinton, G., 2012, A Practical Guide to Training Restricted Boltzmann Machines, Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science Volume 7700, pp 599-619.

Huang, J., Rogers, S., Joo, E., 2014. Improving Restaurants by Extracting Subtopics from Yelp Reviews, SOCIAL MEDIA EXPO, https://www.ideals.illinois.edu/bitstream/handle/2142/48832/Huang-iConference2014-SocialMediaExpo.pdf ?sequence=2.

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.

Jong, J., 2011. Predicting Rating with Sentiment Analysis http://cs229.stanford.edu/proj2011/Jong-%20PredictingRatingwithSentimentAnalysis.pdf.

Koren, Y., 2009. The bellkor solution to the netflix granprize, http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf.

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.

Lee, D., Jeong, O.R., Lee, S., 2008. Opinion Mining of customer feedback data on the web. In ICUIMC '08 *Proceedings of the 2nd international conference on Ubiquitous information management communication*.

Levi, A., Mokryn, O., Diot, C., Taft, N., 2012. Find-ing 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.

Linden, G., Smith, B., York, J., 2003. Amazon.com Recommendations, *IEEE Internet Computing*, vol. 07, n. 1, pp. 76-80.

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

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

Owen, S., Anil, R., Dunning, T., Friedman E., 2011. Mahout in Action, Manning Publications Co., ISBN: 9781935182689.

Pang B., Lee L., 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135. DOI: 10.1561/1500000011.

Paterek, A., 2007. Improving regularized singular value decomposition for collaborative filtering, Proc. KDDCup and Workshop, ACM Press, pp. 39-42.

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

Sarwar, B., Karypis, G., Konstan, J., J. Riedl, J., 2001. Item-Based Collaborative Filtering Recommendation Algorithms, in *Proc. IEEE Internet Computing*, 10th International World Wide Web Conference.

Shelter, S. & Owen, S., 2012, Collaborative Filtering with Apache Mahout, RecSysChallenge'12.

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.

Singh, V. K., Mukherjee, M., Mehta, G. K., 2011. Combining Collaborative Filtering and Sentiment Classification for Improved Movie Recommendations. In Sombattheera et al. (Eds.): Multi-disciplinary Thrends in Artificial Intelligence, LNAI 7080, *Springer-Verlag*, Berlin Heidelberg, pp. 38-50.

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.

Wu, Y., Ester, M., 2015, FLAME: A Probabilistic Model Combining Aspect Based Opinion Mining and Collaborative Filtering, *WSDM'2015*, Shanghai, China.