World Academy of Science, Engineering and Technology
International Journal of Information and Communication Engineering
Vol:12, No:8, 2018

Lexical Based Method for Opinion Detection on Tripadvisor Collection

Faiza Belbachir, Thibault Schienhinski

Open Science Index, Information and Communication Engineering Vol:12, No:8, 2018 publications.waset.org/10009435.pdf

***Abstract*—**The massive development of online social networks allows users to post and share their opinions on various topics. With this huge volume of opinion, it is interesting to extract and interpret these information for different domains, e.g., product and service benchmarking, politic, system of recommendation. This is why opinion detection is one of the most important research tasks. It consists on differentiating between opinion data and factual data. The difficulty of this task is to determine an approach which returns opinionated document. Generally, there are two approaches used for opinion detection i.e. Lexical based approaches and Machine Learning based approaches. In Lexical based approaches, a dictionary of sentimental words is used, words are associated with weights. The opinion score of document is derived by the occurrence of words from this dictionary. In Machine learning approaches, usually a classifier is trained using a set of annotated document containing sentiment, and features such as n-grams of words, part-of-speech tags, and logical forms. Majority of these works are based on documents text to determine opinion score but dont take into account if these texts are really correct. Thus, it is interesting to exploit other information to improve opinion detection. In our work, we will develop a new way to consider the opinion score. We introduce the notion of trust score. We determine opinionated documents but also if these opinions are really trustable information in relation with topics. For that we use lexical SentiWordNet to calculate opinion and trust scores, we compute different features about users like (numbers of their comments, numbers of their useful comments, Average useful review). After that, we combine opinion score and trust score to obtain a final score. We applied our method to detect trust opinions in TRIPADVISOR collection. Our experimental results report that the combination between opinion score and trust score improves opinion detection.

***Keywords*—**Tripadvisor, Opinion detection, SentiWordNet, trust score.

I. Introduction

TODAY all social media is about notation or opinion. All the services or products today can be marked, judge and appreciate by all the user and reading online by everyone. When you want a restaurant, a hotel, when you booked a cab, when you buy some product on website, there are grade and opinion about it. It's very useful to decide between this product or service. Internet users have the habits to scroll down and check the commentary, what the previous user think about it. "This hotel get a 5 grade and all the comment about it are incredible is obviously the best that i can found, it's obviously better than this one who are only a grade of 1" Companies like amazon, airbnb, booking, tripadvisor mainly work on this system of mark that allow user to chose by experience of other, and if their experience doesn't match with their expectation the fault is all on the website: their system failed and the grade that produce have are wrong. So it's very important today for this website to develop tools to analyse and deal with all the comment that they have and it's very difficult. All users haven't the same to mark or the same taste about things: someone can put a good mark on an hotel because it is what he liked when he go in holidays but someone else can give a bad mark because according to him it was a really bad. So it's a bad hotel? the first client is it trustfully? How can we deal with the human nature and their different way of thinking? So the website try to develop way to go against all this problem. Commentary are take into account in a mass of comment not one by one, they can be sort out by topics. But how can we deal with problem of: "it's a great hotel i give 3/5" "it's a great hotel !!! deserve 5/5!" Same opinion but different mark because that don't have the same way to think: which grade is closer to the real notation? That why today many people work on the sentiment analysis: they looked after the comment only the text and try to give the mark that correspond best to the comment, but there so many variable to take into account that it's really difficult to find a way that worked in 100% of case.

Faiza Belbachir is Doctor at Institut Polytechnique des Sciences Avancees IPSA (e-mail: phdups@gmail.com).

Thibault Schienhinski is student at Institut Polytechnique des Sciences Avancees IPSA, France (e-mail: Thibault.schienhinski@ipsa.fr).

II. Related Work

Many studies are made in sentiment analysis and especially in field like hotel or restaurant by example a paper about the classification of customer reviews based on sentiment analysis put in place a specific process in order to extract a domain specific lexicon of semantically word based on a given corpus. In using three target level: good, neutral and bad they are able to extract a sentiment from a text and provide satisfying result for a specific corpus. Indeed a lot of case are exclude for the moment in order to quickly process a lot of information and to have good information's but moreover this paper shows that it seems to be necessary to considerably increase result to take into account neutral comments and not only positive or negative one [1], [3], [6], [12].

But there is more study that want to show that we need to improve technique of sentiment analysis, a paper always about opinion mining for review from hotel go further. Their algorithm extracts a corpus of review from the website directly by itself and classify it in relation to positive, negative or neutral aspect. A lot of worked is make to extract comments carrying lot of opinion and to deal with it before to analyse it: POS tagging and word filter are applied before to determine the opinion. Moreover they used an interesting equation to determine the score of a review to know positivity or negativity, it's not just about polarity. And they obtain on several corpus a good ranking of the review [8], [9], [13].

Open Science Index, Information and Communication Engineering Vol:12, No:8, 2018 publications.waset.org/10009435.pdf

But it's not enough satisfying, indeed a recent paper propose another way to deal with review from hotel. Based on a TripAdvisor collection they propose to divide review into topics, in order to determine opinion for each topic. By example for an hotel we will have room, cleaning, bathroom, staff .. for topics. Their algorithm separate review into topics and like before with sentiment analysis base on language processing they are able to determine a grade for each topic for one comment. Thanks to this method they obtain a more precise score and for different aspects They obtain very good result with this method and can make better ranking in relation to the grade of each aspect of comments [10], [2], [4].

In parallel of all this study around the language processing there is more and more studies and papers about machine learning and neural network [5], [7]. A recent paper propose a general class of models based on neural network architecture and word embeddings. They test with several kind of neural network architecture and collect a lot of information in order to initialize their own method of learning which have provide good result and which is very flexible. This work can be successfully applied to fine-grained opinion mining tasks external effort. They applied their algorithm to a corpus of restaurant and the result overpass the result of the top performing system on the Laptop dataset.

III. Our Approach

In this paper, we want to improve opinion score to have a more right score. We use collection of Tripadvisor. We decide to add to this current opinion score a new variable:the trust score. This score represent at each level can we trust a comment. The problematic is to represent and calculate this new trust score? And is the result of our new trust score better than old score with only the score link to the lexical word?

We will develop which collection and why this one in the following sentences, then all the experimentation that we made on the subject. The steps for our approach are:

- Cite different collection contain opinion.
- Extract retrieval information.
- Calculate opinion score.
- Calculate Trust score.

A. Difference Collection Contain Opinion

To work we first need to select a collection to analyse comments on a subject. First we had define some features for our collection, it has to contain:

- Grade system: system who allow use to grade the object of the website.
- Lot of Commentary.
- Information about user (gender, ages, specific information in relation to the collection).

  From this criteria we thought about several solutions:
  Imdb:
  Imbd is a website about movies and series. You can check for all informations about everything who concern movies and you can give our opinion about all the movies that you've seen. You give a grade and leave a comment below movies that you want. Imbd has a very active community that allow us to have a lot of comment. Moreover user can create an account and add personal informations that we need for our work.
  Tripadvisor:
  One of the most popular website for the holidays. Indeed on tripadvisor you can check hotel, restaurant, flying, activities that you may want to go and booked directly on the website. You can prepare your vacation directly on the website. Again a lot of comment, grade and information about users.
  Amazon:
  Probably one of the most famous website in the world. If you need something try to buy it on amazon. It's interesting for us cause on every product can be marked and comment, moreover we have a lot of informations about all the users who comment.
  Ciao:
  One of The website of advice on consumer product. You can give grade and comment on every kind of consumer product. A lot of comment and a little information about user.
  So we had to decide which website we will use to create our collection. We first eliminate Ciao not enough client information or trustfully information. Moreover the was the less interested subject.Then we decide to choose between Imdb and tripadvisor, indeed there are no many work on for this subject on TripAdvisor moreover Tripadvisor were very interesting because its cover subject that very interested comment to analyze and there were a lot a user informations that we can exploit. So finally we decide to choose Tripadvisor for our project
  1) Tripadvisor Collection: We have to know how we will organize our collection. The first approach was to say we will make a collection of users: we will have users and all information about him and all the comment that he posted but it seem to be problematic to related the comment to the hotel. So we decided to organize our collection not by user but by hotel. Indeed we select several hotel there all comment and a user information
  We use the Tripadvisor collection that contains 200 000 hotel [11]. For each hotel we have a text file with all the information: Name of the hotel, Url, Global grade, comment, grade, name of user. Unfortunately this collection doesn't contain user information that we want. Indeed for our project we will need to have the maximal information about user who comment. In tripadvisor we will use this informations: Gender, Age, Address, Subscription date, Number of review, Number of useful review: other user can show "this review was useful to me", Number of points: points give by Tripadvisor, Number of badge: reward give by the website. We extract these information from 1800000 users.

B. Extract Retrieval Information

All word doesn't not helping for opinion, many research [6] have been done to determine which POS is carrying most part of opinion. Take an example: "This hotel has a great entrance

Open Science Index, Information and Communication Engineering Vol:12, No:8, 2018 publications.waset.org/10009435.pdf

but no parking, moreover the room were not clean when we arrived If we highlight only the word with a real impact on opinion we obtain this: "This hotel has a great entrance but no parking, moreover the room were not clean when we arrived" Only 4 word on 18 are useful (22%) so is it interesting to analyze every word and loss time and resource for word who doesn't help? Clearly not: so we have to select what world we can choose. In our work we take superlative, adjective and adverb.

C. Opinion Score Using SentiWordNet

1) SentiWordNet: Now we can start talking about analysing opinion. The aim of this kind of analysis it's to look word after word and try to know if the comment is more positive or more negative. The human can easily make the difference, he not only understand the meaning of word one by one but understand the context when the word are put together. Currently computer doesn't know how doing that. If I said "this hotel is great !" instantly you said positive opinion, same argument "this hotel is the worst that i have ever seen" negative opinion. But how the computer can make the difference

So the main principle is to take word one by one: analyse it. By example i take "great" positive word, "bad" negative word, To doing that we need a kind of dictionary that contain all informations about the words and who can give if positive or negative it is. This kind of lexicon exists, for our project we will use SentiWordNet (SWN). For each word SWN has an positive score and a negative score. But it's not that simple. SWN give us a score not only for a word but for a word, his type (noun, adj, adv) and a context. SWN is a huge text file who contain all informations about all words organize as follow (see Fig. 1).

POS	Offset	PosScore	NegScore	SynsetTerms
a	1000003	0.0	0.125	form-only#a#1
a	1000159	0.25	0.0	dres#a#1 full-dress#a#
a	1000307	0.0	0.125	titular#a#5 nominal#a#6
a	1000440	0.0	0.125	prescribed#a#4 positive#a#5

Fig. 1 SentiWordNet

2) Exploiting SWN Score: For each comment we calculate a positive score: posScore and a negative score: negScore using flow equation.

$$Opinionscore = (posScore + negScore)/N \quad (1)$$

where posScore and negScore are extract on SWN. N is number of word analyse in a comment.

In our work we determine opinion of comment but we precise if it is positive or negative. We use two equations.

By polarity if result is positive so we have positive opinion and vice versa see folow equation.

$$polarity = posScore - negScore \quad (2)$$

By maximum:

$$max(posScore, negScore)/N \quad (3)$$

The goal is to determine the best score of opinion mining

D. Trust Coefficient

In this work we won't just take into account the content of the comments. We want to go further and take into account another variable. In Tripadvisor anyone who will use the website to booked something can comment what he booked. But how can we trust this person? We don't know her, we don't know if he is trustful. There is a lot of people in the community of Tripadvisor: people who booked once and comment just one time and important people in the community we can have make a hundred of comment. So it will be logical that the second user is more truthful but in the website notation this trust parameter is not consider whether it can make a difference and allow to improve grade system. Before we talk about informations about user that will be useful. We have several trust parameter that we will use to determine the level of trust user:

1) Number of review: more a user make review more he has experience, he has visited many hotel or restaurant, he can make many comparison.
2) Number of useful review: when you read comment you can agree with one and can judge it useful or right and you can notify it. So more a user has review judge useful means that this comment are right and trust-able. So it's an important criteria.
3) Average useful review: it may be interesting to know how many review are judge useful.
4) Indeed is a user who comment a lot but who haven't useful review more truthful than user who comment less but with a lot of review judge useful by others. So we will consider ration NumberUsefulreview/numberReview to have the average of review judge useful per review.
5) Number of badge: you can achieve success by doing stuff in Tripadvisor website. Like you have visited 6 different country, so more you had badge more you have trip experience so more trustful you are
6) Length of service: the number of year of subscribe in the website. If you have subscribe in 2007 you have, in theory, more experience than someone who do it in 2017.

All this parameter composed the trust coefficient TC. We compute an equation using these coefficient see folowing equation.

$$TC = NR + NUR + AN + TP + NB + D \quad (4)$$

With NR is a number of review, NUR: number of useful review, AN: average useful review by review, TP: tripadvisor point, NB: number of badges, D: number of year since subscribe But is it logical that number of review has the same weight that number of badge? To have a better trust coefficient we can pondered the formula in relation to the importance of each criteria:

$$TC = \alpha * NR + \beta * NUR + \gamma * AN + \eta * TP + \varphi * NB + \delta * D \quad (5)$$

Considering this coefficient we take into account all criteria with their respective importance.

Open Science Index, Information and Communication Engineering Vol:12, No:8, 2018 publications.waset.org/10009435.pdf

1) Trust Score: We combine SWN score and the trust coefficient. We will able to calculate our new opinion score that we call the trust score:

$$TS = \lambda * SWNScore + (1 - \lambda) * TC \quad (6)$$

After some test we have find ad equation value of lambda: for positive comment is equal to $= 0.5$ for negative comment is equal to $= 0.8$

The trust coefficient has a real impact on the trust score. Indeed more the user will be trustful more his TC score will be important, more the TC score will be important less the SWN score will be decrease. Indeed more the TC will be important more the TS will be important, if we trust the user who has posted this comment so we increased my trust score to show that this comment can be more considerate. Contrary if we have a user not very trustful the trust score will be less important.

	Ordre by Tripadvisor	Trust coefficient	New classement	Trust score
1	Great hotel and location a great hotel with a fantastic location. Stayed on the 9th floor with a space needle view. Clean spacious rooms with a superb view. Staff were very friendly and chatty. Breakfast in the restaurant was very good as was happy hour from 4pm to 7pm in the bar.10 mins walk to space needle and 5 mins walk to and the shops...i cannot fault the hotel over my4 night stay.....excellent.....	0.0259	They made it a special memory! am amazed at how special they made my mom feel. When I booked this hotel, I told them it was for my mom's 60th birthday, and asked simply that someone wish her a happy birthday. Much to my surprise and delight, they did so much more. When we arrived, I was told that our room had been upgraded to a room with a view of the Space Needle - it brought tears to my eyes. Last time we came to Seattle, we stayed at The Fairmont - and while it was nice	6.521
2	Good hotel, good value, great location. Just got back from a weekend in Seattle at the Warwick. Took the grayling airport shuttle and it dropped us off right in front of the hotel. Also picked us up for our return to SeaTac. Our room had an incredible view of the city skyline, was a fairly good size, with a nice size closet and plenty of drawers for your clothes. Also had a good size desk to work, but the wireless internet was not free and the signal was very weak...	0.0158	Nice Surprise! We selected the hotel based on the criteria that I was traveling with 2 12-year-old-girls who wanted an indoor pool and to be close to Seattle Center and Pike Market. This narrowed the field, so we chose Warwick. Based on other reviews, I wasn't sure how the hotel would hold up to our expectations; happily, it did. I gave it a 5 because	5.325
3	Great stay at the Warwick We, along with another couple, stayed at the Warwick for 4 nights from 18th May. We had a room with a city view which was very comfortable. The staff were helpful and check-in was easy. The	0.637	The Warwick in Seattle adventure Great location which was close to everything. The room was very nice, well maintained and had a wonderful view of the Space Needle. The Business Center was very convenient. The Staff was	3.549

Fig. 2 Result of Trust Score

E. Experimentation and Validation

We choose randomly 1000 comments of 5 stars and 1000 comments of 1 stars. This collection will be our reference collection. We calculate our opinion score positive and negative. The results show that we detect more than 95% of positive opinion comments and for the negative opinion comments 85%.

We conclude that our opinion score determine opinion document (positive and negative) our work does not just stop on opinion detection but re-rank opinion document according to the trust concept. For that

We have experiment with 10 comment of an hotel and we analyse the classement of Tripadvisor and our new classement with our new Trust score.

	nice grocery store across the road which is open till 2am and sells everything, even Australian wine. We ate one night in the hotel restaurant. Enjoyable meal. Would definitely stay at the Warwick again			
4	The Warwick in Seattle adventure Great location which was close to everything. The room was very nice, well maintained and had a wonderful view of the Space Needle. The Business Center was very convenient. The Staff was very courteous and made us feel welcome. Will stay there the next time that we are in town	0.003537546 1032410784	great hotel and location a great hotel with a fantastic location.stayed on the 9th floor with a space needle view.clean spacious rooms with a superb view.staff were very friendly and chatty.breakfast in the restaurant was very good as was happy hour from 4pm to 7pm in the bar.10 mins walk to space needle and 5 mins walk to macys and the shops...i cannot fault the hotel over my 4 night stay.....excellent.....	3.158021 3211999 64
5	Perfect I don't typically write reviews on sites, but I saw the last one and wanted to update it...I am physically writing from a room on the 16th floor of the Warwick hotel, visiting here from Hollywood, CA...	0.005507204 736901623	Good hotel, good value, great location. Just got back from a weekend in Seattle at the Warwick. Took the grayline airport shuttle and it dropped us off right in front of the hotel. Also picked us up for our return to seatac...	2.426112 1068741 52
6	They made it a special memory I am amazed at how special they made my mom feel. When I booked this hotel, I told them it was for my mom's 60th birthday, and asked simply that someone wish her a happy birthday. Much to my surprise and delight, they did so much more. When we arrived, I was told that our room had been upgraded to a room with a view of the Space Needle - it brought	0.012576417 365372855	Great stay at the Warwick We, along with another couple, stayed at the Warwick for 4 nights from 18th May. We had a room with a city view which was very comfortable .... at the Warwick again	2.137738 3379168 715

Fig. 3 Result of Trust Score

We can see in Figs. 2 an 3 that more the trust coefficient is important more the comment is better ranked in relation to is SWN score too.

IV. Conclusion

The result of the validation part show that the trust score give logical and satisfying result. The adding of this new variable: the trust, able to give a new dimension to the validity of a comment. The trust score determines the best comment between two with the same mark (for example number of star) based on user information. With the new trust score we are able to make a new appearance order on website to better help the user, he will directly read the comment who has the best trust score and who is the more trustful. In future works it would be interesting to validate our approach on a larger collection to validate match more the trust notion.

REFERENCES

[1] Haji Binali, Vidyasagar Potdar, Chen Wu, A State Of The Art Opinion Mining And Its Application Domains, 2014.
[2] Marco Guerini, Marco Turchi, Lorenzo Gatti, Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet, 2013.
[3] Julia Kreutzer, Neele Witte, Opinion Mining Using SentiWordNet, 2014.
[4] Bruno Ohana, Brendan Tierney, Sentiment Classification of Reviews Using SentiWordNet, 2009.
[5] Oscar Romero Llombart, Using Machine Learning Techniques for Sentiment Analysis, 2014.
[6] Guido Boella and Leonardo Lesmo, Automatic Refinement of Linguistic Rules for Tagging, 2012.
[7] Walaa Medhat Ahmed Hassan Hoda Korash, Sentiment analysis algorithms and applications: A survey, 2014.
[8] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani, SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.
[9] Bo Pang and Lillian Lee, Opinion mining and sentiment analysis, 2008.
[10] Jayashri Khairnar, Mayura Kinikar, Machine Learning Algorithms for Opinion Mining and Sentiment Classification, 2013.
[11] Hongning Wang, Yue Lu, Chengxiang Zha, Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach.
[12] Dietmar Grbner, Markus Zanker, Grnther Flied, Matthias Fuchs, Classification of customer reviews based on Sentiment analysis.
[13] Walter Kasper, Mihaela Vela, Sentiment Analysis for Hotel Reviews.

Open Science Index, Information and Communication Engineering Vol:12, No:8, 2018 publications.waset.org/10009435.pdf