An Intelligent Approach to Review Filtering and Review Quality Improvement

Iain Lee, Yu Sun
Department of Computer Science
California Polytechnic State University, Pomona
Pomona, CA
{iainlee, yusun}@cpp.edu

YueXin (Sophia) Li

Branksome Hall
Toronto, Canada
sli2@branksome.on.ca

***Abstract*—This paper presents a new approach to review filtering in order to bring a more transparent solution to generate more trust between the user base and review-based sites. Instead of removing reviews based on authenticity, users decide on the level of filtering that is provided on the reviews. Each review is given a score and this scoring system is based on three categories: location based check-in (LBS), receipt authenticity, and sentiment analysis on the actual review. Once all three categories are factored, an algorithm will be used to return a score on the review. This score determines the supposed authenticity (confidence that the review is a legitimate review) of the review and this score is what will be used as the filtering mechanism for the user.**

I. INTRODUCTION

Online user reviews are so important to modern day purchasing as it provides a means for users to get to know and understand more on the product they are purchasing or the location they want to visit. However, it is really hard to differentiate between false data and true data which also applies to user reviews and it can be argued that this problem is much larger for user reviews since so many businesses and consumers rely heavily on this information.

While some may not view false reviews as a huge negative there are some serious repercussions to false reviews. While it may not affect major brands, businesses, or services that have already established a name for themselves small companies and independent companies that do not have a strong backing will find it difficult to overcome the hurdles of false negative reviews [1]. Many review sites acknowledge the problem already of small businesses attracting customers as many users tend to brush off any businesses, services, or products with low amounts of reviews. This leads to a vicious cycle of lacking customers due to low review count, but without any customers there will be no reviews to begin with on the product.

Not only do these businesses have to overcome this obstacle but to add to the burden could be the potentially harmful negative fake reviews. With a low count of reviews to begin with, a single fake review could lead a serious lack of interests from potential customers and even avoid that institution. Having to overcome both these problems is tough on small time businesses so many user review sites try and address this problem with their own solutions.

On the flipside there are business services that can be employed to perform non-appropriate business practices and bring up the rankings of certain institutions with massive amounts of positive fake reviews. This problem is just as large as it causes distrust between the consumer and the business even if a majority of them do not employ such tactics. Just like the other two problems this one also affects the truthful smalltime institutions as they may be overshadowed by another competitor employing such under handed moves to gain the advantage of better and more reviews leading to more customers. Let us look further into one of these major sites that employ user reviews, Yelp and how they try to combat these problems.

The main use of Yelp is to search for businesses based on filters that users apply and find ones that are highly rated [2]. The impact of the rating system on Yelp has made it so that many small businesses depend their livelihood on good ratings and thus having fake reviews that slander their business could be extremely troubling. While Yelp does provide filter algorithms some believe that they are a bit too draconic and the filtering is not reliable enough, especially when people's businesses are on the line. Due to this form of filtering it has caused upset and has caused Yelp to face certain litigations based on their tactics [1]. To better the relation between the users and the site this project will focus on a different approach to filtering.

The current approach that many review sites do is to remove all reviews that are considered fake based on their algorithms. Yelp also employs this method and will by default hide any reviews it deems as fake. The user has the ability to show these reviews but they will never be factored into the average reviews. These algorithms tend to be non-disclosed to prevent gaming of the system but also leads to distrust as it is not easy to trust something that is not understood or known. The approach presented in this paper instead does not filter out any reviews. Each review will be scored based on three criteria: location based check in, image upload of a receipt, and an analysis of the review content itself.

The first criteria is as simple as it sounds in which the reviewer can check in by pressing a button with their Global Positioning Satellite on to show that they have been within a certain radius of the address of the business being reviewed. Once this button has been checked and the location has been

verified when the reviewer writes a review within a certain time frame it will give full points towards the user for the first criteria.

The second criterion involves uploading the image of the receipt of the establishment that they are reviewing. Establishments tend to have the name of their business on the receipt and what the image reader tries to do is to see if a legitimate receipt has been uploaded from said establishment and gives the full amount of points accordingly. A better method would be to not only read the receipt's title only but to also check the date, and the items ordered to ensure validity of the review itself. An example would be a user reviewing a restaurant's burger and when uploading the receipt the image processor should be able to read the establishment's name, the date, and the order of a burger.

The main problem with this implementation is that current image processing is very weak and also many receipts label the item differently and thus it would be very hard to pick out. As for now this project will mainly focus on just the title of the receipt and matching that to the establishment but this part can definitely be improved upon once better image processors are available.

The last criterion will factor the length of the review and use sentiment analysis [3] on the review and compare it to the score given by the user. To further clarify based on the score given by the user the algorithm will rate the review differently. This will be further explained later on and go further in depth of how the scoring works.

The algorithm will factor in all three and grade the review and the user can then choose to apply review filters of their own showing what scores they would want factored into the average review of each restaurant.

The rest of the paper is organized as follows: we explain the major challenges of the problem in Section 2; Section 3 and Section 4 presents solution and the implementation techniques, followed by showing the experimental results in Section 5; Section 6 talks about related work and Section 7 gives conclusion remarks.

II. CHALLENGES

A. Challenge 1: The Inconsistency of Reviews and Ratings

The first of many challenges to address is the inconsistency between user reviews and the rating given. User reviews usually come with two parts: the rating or number of stars and the written review. The inconsistency comes into play when occasionally there will be a mismatch between the rating and the review given. For this paper we will use the one-through-five start rating system that is most common where one star is the worst and five start being the best.

An example of this inconsistency: let's say a user will give out the following review of a restaurant "The restaurant had a nice décor and food was delicious. I ordered the specialty burger and fries and they were both delicious. The meat was cooked well, packed full of flavor, and tender. However, the family sitting across from me was too noisy and so I decided to subtract some points". And the reviewer leaves a rating of two stars all for a problem that the business has no real control over other than declining the noisy family a sitting. This simple action could cause a huge problem for small business owners and is a problem that this project will try and address through the use of sentiment analysis [4]. A review like this could fare to be a three or four star review.

Figure. 1 An inconsistent review that rates a place as above average but gives it an average rating.

B. Challenge 2: The Extremity of Reviews

Another similar challenge to the previous one is also the extremity of reviews that usually occur. Most people will only leave a review when either they had the best experience or the worst. Middle reviews tend to be nonexistent for the most part and many do not factor in these reviews. Not only that but when users of a website look for certain products or business services they will only look at the extreme ratings. Average or middle of the line is generally tend to be ignored which this project will also try to address by having it harder to have a good scoring review that is either one star or five star but much easier to write a three star review with a good score.

C. Challenge 3: The Insufficiency of Review Comments

Continuing on with this trend of reviews another challenge to address would be the length of the review itself. While businesses and products given a mediocre rating could have reviews that are also mediocre it is for the intent of improving upon the quality of reviews to instead have certain reviews require more depth. Many times people will either give a location or product one or five stars which means either this place or product was the best or the worst. There should be a good lengthy reasoning to behind why this is considered the best or the worst that the user has experienced. This is definitely another area that could be improved upon which will be covered more chapter four.

D. Challenge 4: The Lack of Motivation to Review

Adding in more requirements to user-based reviews will put off many people from creating reviews to begin with. Only a small percentage will of customers or consumers will ever leave a review (less than 10%). So this paper will also address this challenge of getting more people to write reviews while improving the quality of the review.

III. THE SOLUTION

In order to address the challenges above, we have developed a web-based platform BUTRPlease (Bring up the

Review Please), where users can leave reviews for restaurants and also give them a rating.

A. Main Page

The main page will show a list of restaurants that have been input by the user as shown below.

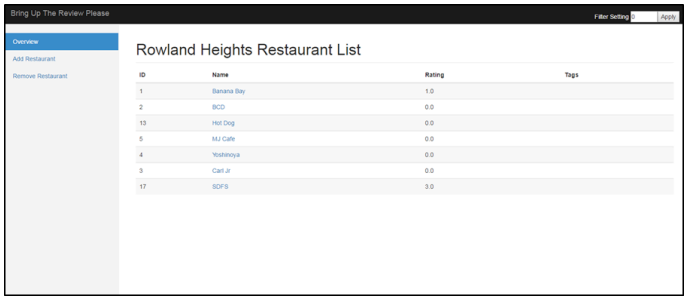

Fig 2. The main page of the site

The left column shows three functions available:

- Overview – allows the user to refresh the page or brings the user back to the home page.
- Add Restaurant – allows the user to add restaurants to the list by inputting an ID number, name of the restaurant, longitude coordinates, and latitude coordinates
- Remove Restaurant – A restaurant can be removed by inputting the ID of the restaurant

The reason as for why an ID was given was just in case the name would not be unique as there are always multiple restaurants in the same city with the same name. Of course the address or longitude and latitude of the location could be also set as the key to ensure uniqueness.

The list that is portrayed in the center of the main page listing all the restaurants are separated into four columns:

- ID – a unique identifier for the restaurant to differentiate between multiple restaurants with the same name
- Name – the name of the restaurant
- Rating – The average rating of all the reviews that exist with this restaurant
- Tags – tags given to the restaurant so that search filters can more easily identify appropriately i.e. Thai, Indian, sandwich, etc.

At the top right of the site is the most important part of the website that differentiates it from other user review sites: the filter setting.

Each review given to a restaurant will also be scored by an algorithm. This score ranges from zero to ninety-nine and the filter is defaulted to zero. This means that any review with a score zero or above will be applied to the average review of the restaurant. If the user decides to change this filter setting to say fifty then any review scored higher than fifty will be applied to the average review.

A different approach to this could also be to add functionality where instead of a scoring system for the reviews the user can check which types of review filtering there should be. To further clarify the user can say that he or she wants only reviews that have LBS check-in to be factored and no other reviews. So the average rating of each restaurant will be based on only reviews that had the LBS check-in. The user could also check another filter method such as only factoring in reviews with both LBS check-in and an image upload of the receipt. It is up to the user to decide which type of filters he or she would want. For instance, users can set the filter at 0 with the average review of a restaurant being 2.0. This is because there are currently four reviews with 2 being a rating of 1 and another 2 with a rating of 3. The two reviews with a rating of 1 have a much higher confidence rating of 66 and 99 while the two other reviews have a confidence of 0. When the filter setting is set to 50 it will change the average rating of Banana Bay to 1.0 because the other two reviews with a confidence score of below 50 will be factored out.

This approach would bring more trust to the average user as giving them more control is always a better approach then to remove control from the user. Many will probably never bother with this system and just set it to default anyways but it will appease those who would rather have the control on their end.

B. Review Page

Diving into the review section of the site we will click on a restaurant link underneath the Name column. The figure below shows the same overview button which will go back to the home page and two other buttons to add and remove reviews.

Figure 3. Review page of Banana Bay.

In the center we see the name displayed and the table is also split into four columns for User ID, Rating (Rating of 1 to 5), Confidence (Level of confidence from 0 to 99 that the algorithm believes the review to be authentic) and Review.

The important part to note is the confidence column which determines whether a review will be filtered out or not during the average review calculation depending on the setting the user has applied. How this score is exactly created will be mentioned in this next section.

Clicking on add review we get a popup window to input User ID, Rating, and the Review text. The Rate Review button rates the text review portion of the submission and adds on a specific amount of points towards the confidence based on the length and sentiment of the review in reference to the rating. The Receipt button will ask for an image upload of the receipt and if the image processor can detect the name of the restaurant being reviewed in the receipt it will tack on more points to the confidence score.

IV. DEVELOPMENT TECHNIQUES

The project was developed using four languages. Two client side languages: Javascript and HTML and two server

side languages: PHP and MySQL. The main page of the site was developed using the Bootstrap Framework and mainly uses two libraries jQuery and Ocrad. jQuery is used to retrieve html elements to use with Javascript and Ocrad is used for image text recognition for the receipt. An API is used for sentiment analysis called Vivekn Sentiment Analysis. The storing and retrieval of restaurant data and review data involves the usage of PHP and MySQL.

A. Location Based Check-In

As explained in the previous chapter this will be just a simple check-in system that will allow the user to check-in with their GPS on to show that they've been around the location that they will be reviewing. The best method would be to allow the user to check-in and then write the review later with the check-in already established. However, for simplicity sake the user was required to be at the location that they are reviewing.

Much of this project was developed and tested using non mobile devices, mainly desktop pcs or laptops which generally do not have a GPS adapter attached. This problem was circumvented by using Google Chrome's sensor device in its console. It allows for an override and a faking of GPS coordinates which was used to test out the LBS Check-in.

B. Text Image Processing

There are many text image processors in the current market such as Mathworks, Imagemagick, and ProjectNaptha. Because there were so many different processors a list of criteria was created to zero in the choice. Two of the main things that were looked for in this were easy implementation and no form of subscription or payment (free). Preferably open source also.

Ocrad was picked out of the many different image processors out there because it matched all three criteria. Most image processors that are available require a form of membership or registration to access their api and while a free trial is given for a period of a month this length was not enough for the project at hand. Since there was an open source library readily available that required no signup or membership for access Ocrad.js was used instead.

Due to it being a Javascript library installation was simple with just a include statement and the use of it just required some unique function calls. However, with the positives comes with its negatives. Its main weakness being that due to it being a lightweight and free processor it does not handle all situations very well. Looking at Figure 14 below shows an example of the weakness of the image processor. Due to the fact that the labeling of the restaurant is not neatly formatted and the letters connect together the image processor can't distinguish the words appropriately and show the letters "whzlel _ds" and not "whole foods".

For the intent and purpose of getting this project running Ocrad.js was still used due to its simplicity and so requirements were a little bit more relaxed for sample images. Instead of using receipt images for testing purposes instead custom text images were created for testing and presentation. For future work a better and more robust image processor will be used such as Mathworks which will clean the image before text recognition is run first.

C. Vivekn Sentiment Analysis

For the sentiment analysis section the same criteria which suited the image processor suited the analysis processor also. This is why Vivekn Sentiment Analysis [5] was chosen as it was an open source api available for free usage. The API itself was a simple HTTP POST call by sending the site a text and it will return a JSON with the sentiment value of positive, negative, or neutral and the percentage strength of that sentiment. For this project the strength of the sentiment was ignored and only the positive, negative, and neutral scores were looked at.

In most cases even though this was a very lightweight sentiment analysis it functioned really well when reading the sentiment of a review text. Its main weakness was also seen in many other sentiment analysis processors in that the detection of sarcasm was not possible. Sarcastic and joking reviews will be given a wrong sentiment. This problem was noticed in all the sentiment analysis processors that were tested and thus the choice of using the free one was an easy choice to make.

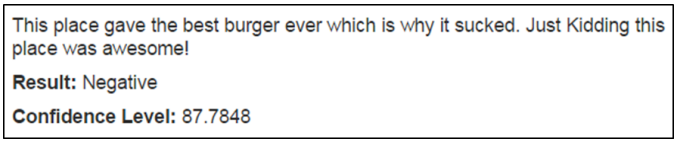

Figure 5. A sarcastic positive remark given the wrong value of negative.

For future work hopefully sentiment analysis gets better and sarcasm and joke detection but some of these things are even hard to detect when a person reads them so the difficulty of these challenge can be daunting and should not expect to ever become perfect.

D. Confidence Scoring

The score of 0 to 99 is split into three different parts. The first part was the LBS check-in explained earlier. If the user is within 100 meters of the location of where restaurant he or she is reviewing then the review will be allotted 33 points to the confidence score.

The second part involves the image receipt upload. If the receipt image uploaded contains the name of the restaurant being reviewed it will give another 33 points to the score however, if the user is somewhat close to the name of the restaurant it will give 10 points. This part of the algorithm can definitely be more varied in terms of scoring and things can change around for a better scoring system.

The last 33 points is determined by the sentiment analysis. The first 11 points of the 33 will be given if the sentiment of the review matches the rating. If the rating is 5 or 4 a positive sentiment will give 11 while if 1 or 2 a negative sentiment will give 11. The other 22 points are given when the length of the review matches the rating. Ratings that are more extreme are expected to give more in-depth answers as to why they have given such an extreme review. Ratings that are more in the middle can give mediocre reviews since the emotion or sentiment is not as extreme.

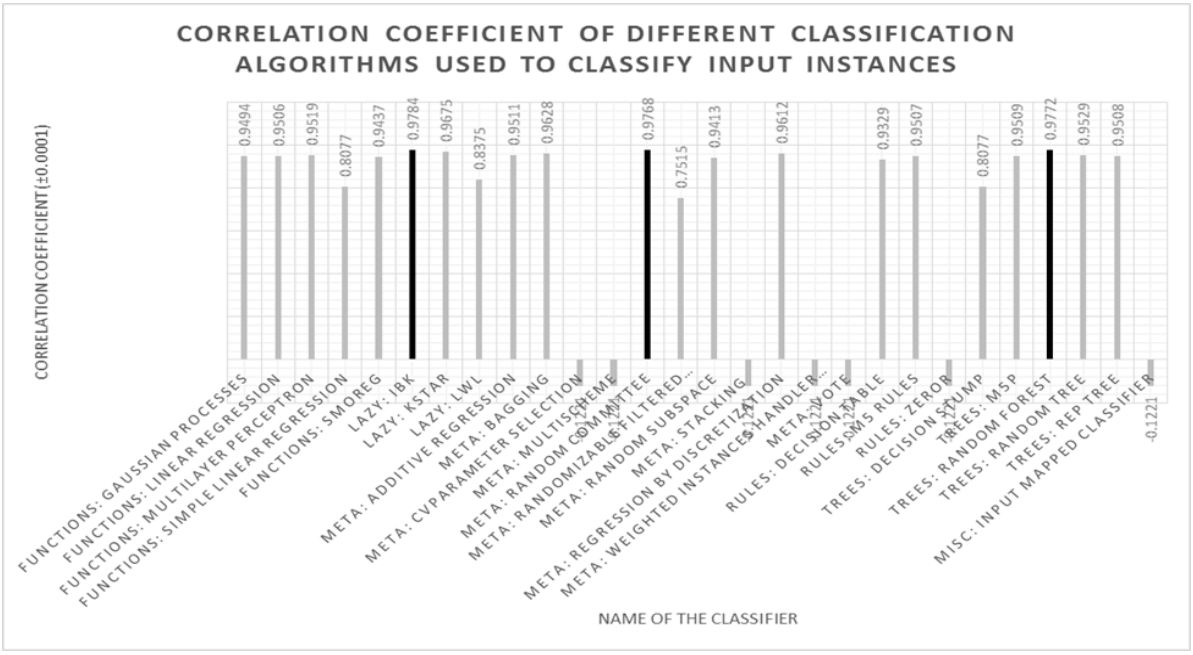

Figure 8. Experimental Data using Weka

V. EXPERIMENTATION AND ALGORIHTM IMPROVEMENT

In this section, we verify the accuracy of the scoring algorithm with a machine learning based simulation. In addition, we investigate how to use machine learning to handle the situation where certain review inputs are missing. Based on the result, we propose a new algorithm schema to improve the accuracy.

A. Algorithm Accuracy Experiment with Machine Learning

To build an algorithm model that would calculate the confidence score from input information, a group of sample data containing 1000 instances is processed in the software Weka. The first simulation is done based on the assumption that all components of a qualified review including rating, GPS, receipt, and comments are provided by the customer. In total, 28 different classifiers are tested under the same qualification of 10-fold cross-validation, and the correlation coefficient is recorded to evaluate the quality of the least squares fitting to the original data provided by the algorithm.

In the diagram above, the correlation coefficient value given by each classifier is displayed. As defined, the correlation coefficient should have an absolute value lying in the range from 0 to 1, in which larger values indicate stronger linear correlations that better represent the set of data. While a positive correlation coefficient indicates a positive linear correlation, a negative coefficient indicates a negative linear correlation. In this case, three classifiers produce relatively high correlation coefficients compared to others, which are Random Forest (0.9772), IBk (0.9784), and Random Committee (0.9768). The results show that the algorithms behind these three classifiers are able to evaluate a highly accurate confidence score from the input information.

To have a comprehensive assessment of the classifiers, an additional parameter is investigated. In Weka, the relative absolute error of an algorithm is provided as a percentage, which measures the discrepancy between the exact value of the data and some approximation to it. Therefore, it serves as an indicator of the numerical stability of an algorithm. Evident from their relative absolute errors of 20.2355%, 16.3975%, and 18.6500%, Random Forest, IBk and Random Committee can potentially be three appropriate algorithms to filter and evaluate the reviews.

B. Score Prediciton and Algorithm Improvement

In reality, it is highly possible that one or more of the multiple components of a qualified review is missing or unreported. Due to this weakness, further experimentation is completed specifically to three classifiers: Random Forest, IBk, and Random Committee. Four trials are done to simulate the real-life situation where some of the information is missing. To reflect that situation in Weka, one column of data at a time is removed, and classifiers are then applied to the rest of the data. The new correlation coefficients and relative absolute errors are compared to the original values. The results are presented in the data table below.

Table 1: Comparison of Correlation Coefficients and Relative Absolute Errors under different simulations

Classifier name	Original Correlation Coefficient (±0.0001)	Correlation Coefficient with missing information (±0.0001)				Original Relative Absolute Error (±0.0001%)	Relative Absolute Error with missing information (±0.0001%)			
		Missing rating	Missing GPS	Missing review	Missing receipt		Missing rating	Missing GPS	Missing Review	Missing Receipt
Random Forest	0.9772	0.9533	0.4520	0.9205	0.8316	20.2355	29.4356	97.1406	35.9482	55.2395
IBk	0.9784	0.9259	0.3072	0.9111	0.7711	16.3975	32.4225	104.6286	36.9579	65.0519
Random Committee	0.9768	0.9398	0.3736	0.9113	0.7836	18.6500	31.0657	100.2301	37.1730	62.3637

The overall trend applicable to all three classifiers is that the correlation coefficient will decrease and the relative absolute error will increase if there is one piece of information missing from the review. Among three classifiers, random

forest experiences the smallest decrease in its correlation coefficient and the smallest increase in relative absolute error in all four conditions, while random committee is the most susceptible to missing information. Therefore, it can be concluded that random forest is the algorithm that can best resist the negative effects of missing information and model the limited available data in the most accurate fashion.

Another interesting observation from the data is that the extent to which the correlation coefficient and relative absolute error change is different for every condition. The changes in these two parameter are relatively small when ratings and word reviews are missing, and are moderate when proofs of receipt are not provided. However, there is a steep decrease in correlation coefficient and a huge increase in relative absolute error when the information of location is missing. This fact shows that different components of a review actually have varying degrees of effects and weigh differently in calculating the confidence score.

The observation leads to an insightful re-consideration of the proportion occupied by each piece. In the previous sections of this paper, each of the components is proposed to be worth the same proportion that is one third of the confidence score. However, seeing the large discrepancy between the scales of impact of different pieces, some adjustments should be made to compromise for it. For example, the proportion of the information coming from GPS can be decreased so that the confidence score will not be distorted so much when lacking of it. Simultaneously, the proportions occupied by word reviews and receipts can be increased to balance the algorithm equations. Still, how much the proportion of each piece should be increased or decreased remains an unsolved problem for future investigation. A logical and rigorous approach is definitely required, and it is very likely that algorithm will become the ultimate solution to this problem again.

VI. Related Work

There have been many studies when it comes to spam filtering [4, 6, 7]. Some fake user reviews is considered a form of spam so some spam filtering systems could also be applied to user reviews. Spam filters such as semantic based approach could also be applied to user reviews however this system will really only catch reviews that are similar to forms of spam. This approach is not as effective as towards user review filtering due to the fact that the approach to faking reviews is different from creation of spam.

Spam in general wants to be sent and attract towards a different form of audience while as fake user reviews instead want to seem as real as possible while accomplishing their objective (destroying or creating a reputation). Because of these differences, spam filtering systems are not as effective towards fake user reviews and thus a different approach is needed.

It should be noted however, a form of machine learning algorithm to detect spam will still be effective as mentioned earlier of culling certain types of fake reviews but may create false positives especially when the difficulty is finding out which reviews are fake to begin with. Machine learning will really only be effective once a system is placed first in catching the fake reviews.

VII. Conclusion

In conclusion this project has used the three criteria to create a more transparent algorithm to be applied to user reviews to determine their authenticity. With a more transparent algorithm it could create a better bridge of trust towards the users and the reviews in the site that are being shown. The filtering of the reviews is given to the control of the users instead of being determined by the site creator that most other review sites employ. Giving more control to the users will also help in creating more trust and I believe to be a better direction as the Internet becomes more prominent in our lives.

More forms of ways to authenticate user reviews can definitely be implemented and while this project was supposed to create a transparent algorithm that was not very exploitable there are still ways to exploit the system. Hopefully these exploits can be fixed in the future such as faking of GPS locations or using fake receipt images but overall it will be impossible to create a 100% non-exploitable system but implementing ways to hamper them would be a step in the right direction.

References

[1] Clark, Patrick . "Yelp's Newest Weapon Against Fake Reviews : Lawsuits". Bloomberg Business Week, 2013.

[2] Graham, Jefferson. "'Yelpers' review local businesses". USA Today, 2007.

[3] Sun, Xiao, Chongyuan Sun, Fuji Ren, Fang Tian, and Kunxia Wang. "Emotional Element Detection and Tendency Judgment Based on Mixed Model with Deep Features." 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (2016): n. pag. Web.

[4] Hu, Wei, Jinglong Du, and Yongkang Xing. "Spam Filtering by Semantics-based Text Classification." 2016 Eighth International Conference on Advanced Computational Intelligence (ICACI) (2016): n. pag. Web.

[5] Narayanan, Vivek. "Sentiment Tool." Sentiment Tool. N.p., n.d. Web. 5 July 2016.

[6] Agrawal, Mohit, and R. Leela Velusamy. "R-SALSA: A Spam Filtering Technique for Social Networking Sites." 2016 IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (2016): n. pag. Web.

[7] M. Bosma E. Meij and W. Weerkamp "A framework for unsupervised spam detection in social networking sites" in Advances in Information Retrieval pp. 364-375 2012 Springer.