Fake Review Detection using Opinion Mining

Dhairya Patel[1], Aishwerya Kapoor[2], Sameet Sonawane[3]

[1,2,3]SRKAY Consulting Group, Surat

Abstract - *As e-commerce is growing and becoming popular day-by-day, the number of reviews received from customer about any product grows rapidly. People nowadays heavily rely on reviews before buying anything. This instigates many people to write fraud and useless reviews about other related products or service. Even there are some organizations in the market who are hiring professionals to write fake reviews and promote their products or defame its competitors' product. Hence, we aim to develop a method which will detect fake reviews and annotate them. The proposed method automatically classifies users' reviews into "suspicious", "clear" and "hazy" categories by phase-wise processing. The hazy category recursively eliminates elements into suspicious or clear. This results into richer detection and be useful to business organization as well as to customers. Business organization can monitor their product selling by analyzing and understanding what the customers are saying about products. This can help customers to purchase valuable product and spend their money on quality products. Finally end users sees each individual review with polarity scores and credibility score annotated on it.*

1. INTRODUCTION

1.1 PROBLEM DESCRIPTION

Online reviews are often the primary factor in a customer's decision to purchase a product or service, and are a valuable source of information that can be used to determine public opinion on these products or services.

Reliance on online reviews gives rise to the potential concern that wrongdoers may create false reviews to artificially promote or devalue products and services. This practice is known as Opinion (Review) Spam, where spammers manipulate and poison reviews (i.e., making fake, untruthful, or deceptive reviews).

Many giant companies are trying to combat with opinion spam: Eg. Amazon, Yelp, TripAdvisor.

1.2 AIM

To detect and flag fake reviews posted on ecommerce websites.

1.3 OBJECTIVES

- To implement best approach available for detection of fake reviews using opinion mining (sentiment analysis) techniques.
- To let users, know if each individual review is trustworthy or not for efficient use of money from user's side.

1.4 FACTORS CONTRIBUTING TO OPINION SPAM

- Most problems coincide with problems occurring during sentiment analysis, since it is a basic building block and an inevitable part of fake review detection.
- User prefers product of liking but depends heavily on the reviews :
  - A survey on yotpo.com suggests that decision of buyers are affected by other user reviews by as much as 77%, which is a considerable figure.
  - Thus it becomes utmost necessary to filter fake reviews, whether positive or negative, from the system as it may have a great impact on the buyer.
- Conference resolution :
  - Conference resolution means identifying which exact object is referred to, in case of continuation of sentences and using pronouns to address objects.
  - For eg: "We watched the movie and went to dinner; it was awful." What does "It" refer to?
  - Coreference resolution improves accuracy rate for sentiment analysis.
- Temporal Relations :
  - Timestamp is a major entity when reviews are being analysed. This is because after some amount of time, the review may not hold valid.
  - A great example of this case can be with electronic items.
  - For eg., a 1GB RAM mobile phone may be performing greatly in 2013, but can be irrelevant in 2017.
- Use of Sarcasm for expression in review:
  - Sarcasm is where positive words are used to denote negative meanings.
  - For example, "*What a great car, it stopped working in the second day*."
  - It is very difficult to process sarcastic tone when it comes to sentiment analysis.

- Requirement of World Knowledge:
  - Knowledge about worlds' facts, events, people are often required to correctly classify the text. Consider the following example [9],
  - "*Casablanca and a lunch comprising of rice and fish: a good Sunday*"
  - The system without world knowledge classifies above sentence as positive due to the word "good", but it is an objective sentence because Casablanca is the name of the famous movie.

- Reference to domain may create Ambiguity:
  - Same review can act as positive or negative when the domain of analysis is changed.
  - For example, "*You should definitely read the book*" holds positive sense for the book, while it is negative for a movie review.

- Level of Sentiment Analysis Used:
  - For example, "*although the service is not that great, I still love this restaurant*" clearly has a positive tone, we cannot say that this sentence is entirely positive. In fact, the sentence is positive about the restaurant (emphasized), but negative about its service (not emphasized).

- Marinating a group of synonyms as per product category
  - Many times text contains different words having same meaning. So such word should be identified and group together for accurate classification.
  - It is a difficult task to identify these words, as human vocabulary tends to differ.

2. METHODS USED AND IMPLEMENTATION

2.1 SENTIMENT ANALYSIS AND VADER

Sentiment analysis is simply the process of working out whether a piece of text is positive, negative or neutral. The majority of sentiment analysis approaches take one of two forms: **polarity-based**, where pieces of texts are classified as either positive or negative, or **valence-based**, where the intensity of the sentiment is considered. For example, the words 'good' and 'excellent' would be treated as the same in a polarity-based approach, whereas 'excellent' would be treated as more positive than 'good' in a valence-based approach.

VADER belongs to a type of sentiment analysis that is based on lexicons of sentiment-related words. In this approach, each of the words in the lexicon is rated as to whether it is positive or negative, and in many cases, **how** positive or negative. Here is an excerpt from VADER's lexicon, where more positive words have higher positive ratings and more negative words have lower negative ratings.

Word	Sentiment rating
tragedy	-3.4
rejoiced	2.0
insane	-1.7
disaster	-3.1
great	3.1

Figure 1.1 Sample sentiment ratings for words

VADER has been built and validated across 4 datasets:

1. Social Media Text
2. Movie Reviews
3. Amazon Product Reviews
4. NY Times Editorial

For the building of this dictionary of 7500+ words, VADER used Amazon's Mechanical Turk to find out the polarity and intensity scores for each word using a likert scale type approach. This resulted in a quick and cheap way of getting results. Below is a snapshot of vader_lexicon.txt

```
447  abandons      -1.3    0.9      [-2, -1, -1, -2, -1, -2, -1, -2, 1, -2]
448  abducted      -2.3    1.18743  [-3, -1, 0, -3, -1, -3, -4, -2, -3, -3]
449  abduction     -2.8    0.87178  [-4, -3, -3, -4, -1, -3, -2, -2, -3, -3]
450  abductions    -2.0    1.41421  [-3, -4, -1, -3, -1, -3, 1, -2, -1, -3]
451  abhor  -2.0   1.09545  [-3, -3, -1, -1, -2, -1, -3, -3, 0, -3]
452  abhorred      -2.4    1.49666  [-4, -4, 0, -3, -2, -1, -4, -3, -3, 0]
453  abhorrent     -3.1    1.3      [-4, -4, -4, -2, 0, -4, -2, -3, -4, -4]
454  abhors  -2.9  1.51327  [0, -4, -3, -3, -4, -4, 0, -4, -3, -4]
455  abilities     1.0     0.63246  [1, 2, 0, 1, 0, 1, 1, 1, 1, 2]
456  ability 1.3   0.64031  [1, 1, 1, 0, 1, 2, 2, 2, 2, 1]
457  aboard  0.1   0.3      [0, 0, 0, 0, 1, 0, 0, 0, 0, 0]
458  absentee      -1.1    0.53852  [-1, -1, 0, -2, -1, -1, -2, -1, -1, -1]
459  absentees     -0.8    0.6      [-1, 0, 0, -1, -1, 0, -2, -1, -1, -1]
460  absolve 1.2   1.46969  [2, -3, 2, 2, 1, 1, 2, 1, 2, 2]
461  absolved      1.5     0.92195  [3, 1, 2, 1, 0, 2, 3, 1, 1, 1]
462  absolves      1.3     1.00499  [3, 1, 1, 0, 0, 2, 3, 1, 1, 1]
```

Figure 1.2 Screenshot of vader lexicon

The first column consists the word. The second column is the polarity score (sentiment rating). The third column is the intensity score. The fourth column is an array of 10 intensity scores given by "10 independent human raters" using a below given format:

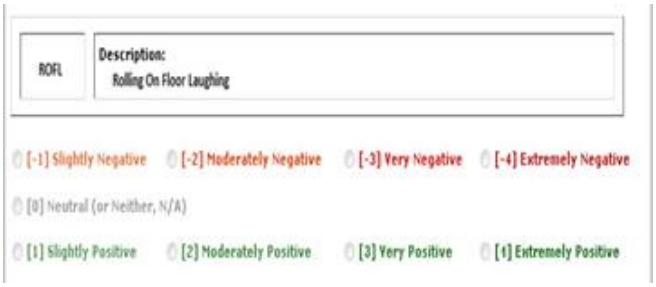

Figure 1.3 UI used to rate words in terms of intensity and polarity

VADER produces output in the form of four different sentiment metrics:

1. positive
2. negative
3. neutral
4. compound

The range of compound score lies between -1 and 1. It is calculated by the following formula:

norm_score = score/math.sqrt((score*score) + alpha)

Where default alpha value is 15. Alpha points to the maximum value that can be achieved. Score is the total summation of polarity words (pos+neg+neu) and normalised score is the one that is achieved in the desired range.

One of the things that VADER recognises is capitalisation, which increases the intensity of words, both positive and negative. Another feature of VADER is that it increases the intensity of sentence sentiment when exclamation marks are detected, with up to 3 exclamation marks adding additional positive or negative intensity.

For example,

The food is good. ------- compound score:0.4404

The food is GOOD. ------- compound score:0.5622

For this project, we have used the Incremental Process Model.

2.2 FLOWCHART

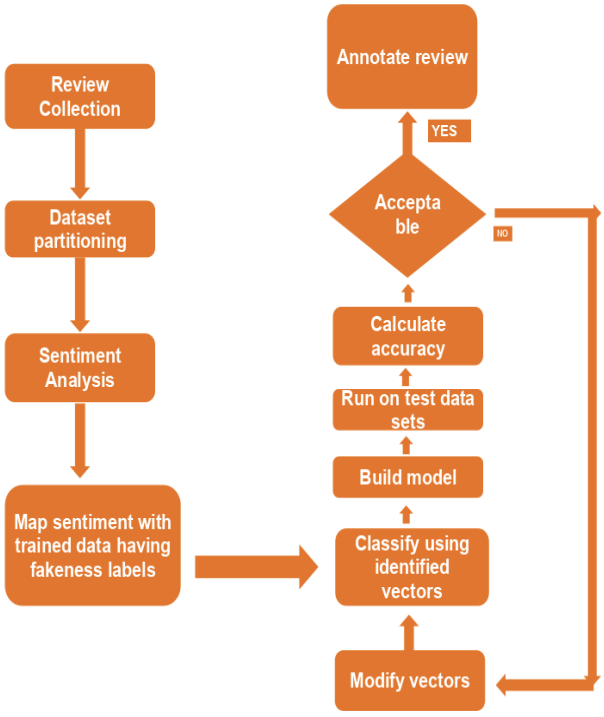

Figure 1.4 Work flow

2.3 ALGORITHM AND IMPLEMENTATION

The flow of our implementation is as follows:

1.Gathering Dataset(s):

"Labelled Dataset" applies in two situations –

(a) Sentiment Labelled
(b) Fakeness Labelled

To our knowledge, there is only one fully labelled dataset available, which is for hotel reviews. But since VADER was built across four datasets, half our battle was won. When it came to feature selection, we used to common vectors to start with. Thus, we had a lot of open options for test sets. We built a dataset manually. All our inputs were in. csv format.

Datasets we have used are:

1. Fully Labelled Hotel Review Dataset.
2. Our own dataset built by taking inputs from people around us.

3. Appending sentiment analysis results:

We run the given data sets through VADER where Dataset (1) showed 74% accuracy. We then appended the results in the dataset in form of negative, positive, neutral and compound score. In the compound score value, ranging from -1 to 1, we set thresholds through trial and error as follows:

-1.0 to -0.5	: *extremely negative*
-0.5 to -0.2	: *negative*
-0.2 to 0.2	: *neutral*
0.2 to 0.5	: *positive*
0.5 to 1.0	: *extremely positive*

4. Appending vector calculation results:

We then appended the length of review (in characters) to the dataset. Next, we divided the text in all reviews into trigrams, and constructed a frequency tally of it. From that, we picked top 10 most frequently occurring trigrams and annotated the reviews where they occurred. Our set of vectors comprises of 3 vectors chosen after much contemplation using knowledge from available literature –

X1 - *Difference between median and length of review*
Possible values: *positive or negative integer*

X2 - *Existence of most frequent trigrams*
Possible values: *0*, if false
1, if true

X3 - Is the review "extremely" positive or negative (check via threshold value)?
Possible values: *-1*, for extremely negative;

1, for extremely positive;
0 otherwise

Logic Table for Classification of Reviews into Suspicious, Clear and Hazy:

Table 1.1 Logic Table

X1	X2	X3	SUS	CLE	HAZ
+	0	0	0	1	0
+	0	-1	0	0	1
+	0	1	0	0	1
_	0	0	0	0	1
_	0	-1	0	0	1
_	0	1	0	0	1
+	1	0	0	1	0
+	1	-1	1	0	0
+	1	1	1	0	0
_	1	0	0	0	1
_	1	-1	1	0	0
_	1	1	1	0	0

As it is evident from the table, there is a 50% chance that a review will be placed in the hazy category first.

5. Output Snapshots

Here is a screenshot of the annotation on output website-

Figure 1.5 Sample Output UI

5.1 USEFULNESS OF THE PRODUCT TO THE USER

Basic application domain of this system consists of any website which includes review posting structures, viz. websites under domains of food(Zomato, Swiggy), e-commerce(Amazon), travel(TripAdvisor). Various e-commerce websites like Amazon, Flipkart and Snapdeal are the major key partners for our system, as they will share the necessary data for analysis. The system can include other websites to extend the review database.

The customer segment is divided into 2 parts: E-commerce website administrators as well as end users. E-commerce partners can use our system to eliminate major fake reviews, whereas the end user (online buyers) is alerted of suspicion wherever it exists.

This annotation brings to life a measurement of the intensity of the sentiment posted in a review, as well as its credibility. This ensures the end user to make an intelligent decision so as to whether the end user should trust that particular review or not.

6. SUMMARY AND FUTURE WORK

- Identifying fake reviews from a large dataset is challenging enough to become an important research problem. Business organizations, specialists and academics are battling to find the best system for opinion spam analysis. A single algorithm cannot solve all the problems' and challenges faced in today's generation with advancements in technologies, though a few are very efficient in analysis. More future work and knowledge is needed on further improving the performance of the opinion spam analysis, and developing one that is consistently efficient across all categories of data. For eg., we noticed that threshold value in 2.3(2) varied across product domains.
- As per our knowledge, there are currently no existing systems like ours is. Agencies like Fakespot provide a credibility score to the whole database, but no system individually annotates reviews, as it is done in our system. This combination of polarity and credibility has not to be found elsewhere as per our research.
- As we are also using sentiment analysis in our project, the scope for improvisation explodes. In future, we would try to update the dictionary containing sentiment words. Since VADER is an open source project, it is bound to get better with updates. As also mentioned in 2.1, another possible future work will consist of adding more words to the dictionary and updating the weights given to those words to get more accurate calculated score of the reviews.
- The logic table given in 2.3 (Table 1.1) has been built using our own intelligence based on available literature we gathered. It can further be improved to provide faster and more accurate results. The output generated using this table in phase one can

further be used to apply domain specific vectors over it to refine the results and eliminate hazy stack into clear and suspicious as much as possible.

REFERENCES

[1] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What Yelp Fake Review Filter Might Be Doing?" Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM), 2013.

[2] Spotting fake reviews using positive unlabeled learning (Huayi Li, Bing Liu, Arjun Mukherjee, Jidong Shao)

[3] Survey on Online Spam Review Detection Methods (Kalyani Adhav, Prof S.Z Gawali Prof Ravindra Murumkar

[4] Hutto, C.J. & Gilbert, Eric. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

[5] Shojaee, Somayeh & Azman, Azreen & Murad, Masrah & Sharef, Nurfadhlina & Sulaiman, Nasir. (2015). "A Framework for Fake Review Annotation".

[6] Exploiting Product Related Review Features for Fake Review Detection (Chengai Sun, Qiaolin Du and Gang Tian)

[7] M. Ott, C. Cardie, and J. T. Hancock, "Negative Deceptive Opinion Spam,"Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics.

[8] "Mining and Summarizing Customer Reviews" , Minqing Hu and Bing Liu

[9] C. Huang, Q. Jiang, and Y. Zhang, Detecting Comment Spam through Content Analysis , Lecture Notes in Computer Science.

[10] T. Ong, M. Mannino, and D. Gregg, "Linguistic Characteristics of Shill Reviews

[11] Survey of review spam detection using machine learning techniques, Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter and Hamzah Al Najada, Journal of Big Data

[12] "Research on Product Review Analysis and Spam Review Detection",Shashank Kumar Chauhan, Anupam Goel, Prafull Goel, Avishkar Chauhan and Mahendra K Gurve

[13] Opinion Spam Detection: A Review, Salma Farooq, Hilal Ahmad Khanday, Department of Computer Science, IUST, Kashmir, Assistant Professor, Department of Computer Science, University of Kashmir