Opinion Spam: Issues and Techniques

Tarek Amr (@gr33ndata)

Abstract

With the rise of online business, consumers nowadays are not only able do their shopping online, but also they can leave reviews on their purchased products for other potential users to see. There is also potential for vendors trying to influence their consumers' decisions, by injecting deceptive product reviews online. Many efforts have been done recently to develop algorithms to detect such deceptive opinion spam. Throughout this report, we will try to shed the light on some of the work done in this area.

1 Introduction

More than fifty years ago and Word-of-Mouth (WoM) has been studied as a way of influencing the consumers' purchase decisions. In the early 1950's, [Whyte Jr, 1954] wrote that "word-of-mouth occurred mostly among neighbours exchanging news on what was being offered by neighbourhood stores". Back then [Katz and Lazarsfeld, 1955] found that the effect of WoM on consumers' decisions was nearly five times more than advertisements, 15 years later [Day, 1971] estimated its effect to be up to seven times more than advertisements.

Today, online product reviews are becoming part of WoM information [Chen and Xie, 2008], sometimes they are referred to as "Word of Mouse" [Skrob, 2005]. Hence they have a good share in shaping the purchasers' decisions. [Skrob, 2005] highlighted three factors leading to WoM information cascade: "the information a receiver needs about a firm, product or service; coincidental conversations between a current customer and a potential one; and high levels of either satisfaction or dissatisfaction". He also noted in his experiment on Expedia [1] that 69% of those who had bad experience with the website, compared to their expectations, communicated this to other potential customers.

With such importance of online product reviews, comes "Review Spam" or "Opinion Spam". "Opinion spam can range from annoying self-promotion of an un-related website or blog to deliberate review fraud" [Ott et al., 2011]. According to [Ott et al., 2011] the former is disruptive and can easily be detected by humans, as well as computers, compared to the latter, which is more deceptive.

Throughout this report we will try to to shed the light on some of the efforts done to detect such kind of spam. The detection methods reviewed here will be divided into "Content-based Analysis" and "Time Series Analysis". We will focus more on the former, which [Ott et al., 2011] sub-divided into "Text Categorization", "Genre Identification" and "Psycholinguistic Deception Detection".

2 Background

2.1 Definitions

Deceptive Opinion Spam: "Fictitious opinions that have been deliberately written to sound authentic, in order to deceive the reader" [Ott et al., 2011].

Disruptive Opinion Spam: Obvious kinds of spam that are not reviews, such as links to other websites, advertisements, random text and irrelevant reviews containing no opinions [Jindal and Liu, 2008].

[1] http://www.expedia.com

3 Review Method

We will try to analyze the work done in each one of detection techniques, and try to answer the following questions:

- How different is it from previous work?
- Which methods achieved better classification scores?
- What is the quality of the training data used?
- Are there other notes or comments worth mentioning here?

4 Analysis

4.1 Text Categorization

[Ott et al., 2011] based their work in this area on that of [Joachims, 1998]. Thorsten Joachims's goal was to classify documents into fixed number of predefined categories using *Support Vector Machines (SVM)*. He first built his feature-set on stemmed words in the given documents, ignoring stop words and words/stems that occur less than 3 times. He then scaled his feature vector using the *inverse document frequency (idf)*. He justified by experiment that SVM is more suitable for text classification compared to *Naive Bayes (NB) Classifier*, since there is no risk of over-fitting, i.e. not selecting the best features. He also stated that SVM deals best with high dimensional feature space.

In their work, [Ott et al., 2011] experimented both SVN and NB. They used the *Kneser-Ney* smoothing method, as described by [Chen and Goodman, 1996], for their NB Classifier. Sommothing was used instead of feature selection as done earlier by [Joachims, 1998]. Their NB results were surprisingly almost as accurate as those of the SVM.

NB Classifiers have been used with email spam detection. [Sahami et al., 1998] presented this approach using words as well as domain specific terms [2] as the feature-set. Open-source email filtering packages were used by [Thomason, 2007] to detect blog comment spam, which is more similar to disruptive opinion spam, with false-positive and false-negative detection rates in the range of 1-2% . [Jindal and Liu, 2008] on the other hand stated that "opinion spam is both widespread and different in nature from either e-mail or Web spam".

4.2 Genre Identification

[Rayson et al., 2001] examined the relationship between part-of-speech frequencies and document categories using the British National Corpus Sampler [3]. They calculated the *Part-of-Speech (PoS)* frequency list for each sub-corpus in the BNC sampler. For each tag, the *log-likelihood (LL)* was calculated as suggested by [Dunning, 1993]. Dunning promised better results using LL, since the assumption of normal distribution is not suitable for text analysis as rare events form a large portion of text data. [Rayson et al., 2001] then sorted the LL values to express the most indicative characteristics for each sub-corpus category.

Two of the different categories [Rayson et al., 2001] studied are important for our analysis here, as we will see later on: "Informative" and "Imaginative" writing. Among Rayson's findings,

[2] HTML tags, SMTP fields, etc...

[3] Part of the British National Corpus (BNC, http://www.natcorp.ox.ac.uk/), where PoS tags have been manually checked

coordinating conjunctions[4] - except 'but' - were found to be less frequent in imaginative writing compared to informative one. Nouns also were less in imaginative writing, while adverbs were less in the informative writing, with the exception of comparatives, superlatives and appositives [5].

The good thing about PoS detection is that the size of the training corpus doesn't have to be as big is in text categorization. [Biber, 1993] noted that a sup-corpus is representative of the corpus as long as the frequency of the items under examination is matching. [Rayson et al., 2001] added that PoS categories are much more frequent than terms, since each PoS category contains multiple terms, and that's why he didn't consider working with BNC Sampler a problem compared to the BMC, which is 50 times as big as the sampler.

In their study, [Ott et al., 2011] found relationship between Truthful and Deceptive Opinions on the one hand, versus Informative and Imaginative Writing on the other hand. But the detection accuracy achieved with Genre Identification using PoS was the least compared to Text Categorization (section 4.1) and Psycholinguistic Deception Detection (section 4.3)

4.3 Psycholinguistic Deception Detection

[Yoo and Gretzel, 2009] tried seven different measurements. Numbers such as minimum/maximum word-count in a review didn't give enough indication about how deceptive or truthful it is. They then measured the sentences complexity[6], and it was noticed that deceptive reviews tend to be more complex. They also tried to build relation between deceptive reviews and the use of *brand-names*, first person pronouns and positive/negative sentiments[Pang and Lee, 2008]. [Harris, 2012], who came after [Yoo and Gretzel, 2009], also experimented with their seven measurements. He used ARI formula[7] for measuring sentences complexity. When it comes to sentiment analysis[8], he tried to differentiate between the polarity of the sentiment and not just whether it is positive or negative. As opposed to [Ott et al., 2011], [Harris, 2012] categorized the reviews according to ratings given to products in the reviews, and then he tried to find out the truthfulness or deception level for reviews in each category. They also found a relationship between sentiment polarity, and how low/high is the review's rating. The sentiment scores and the sentence complexity, as well as some other measures were feed to a SVMlight[Joachims, 1999]. [Ott et al., 2011] were able to outperform [Harris, 2012] using Linguistic Inquiry and Word Count (LIWC) [Pennebaker et al., 2007], however their *Psycholinguistic Deception Detection* couldn't match that of their *Text-Categorization*.

4.4 Time-Based Detection

Instead of looking into the content of the reviews, [Xie et al., 2012] tried to detect opinion spam using *time-series* pattern discovery. They calculated the average rating reviewers give in an online store, number of reviews and that of "Singleton reviews"[9], all within a certain window.

[4]Coordinating conjunctions can be used to join two complete thoughts: for example, 'For', 'And', 'Nor', 'But', 'Or', 'Yet' and 'So'. In a complex sentence, we have one complete thought and at least one incomplete thought. The incomplete thought begins with a subordinate conjunction. There are many subordinating conjunctions, but the most common include the following: 'When', 'While', 'After', 'Although', 'Before', 'Because', 'If', 'Though' and 'Since'. [Jackson,]

[5]Aappositives: a noun, noun phrase, or series of nouns placed next to another word or phrase to identify or rename it. For example: "The king, my brother, has been murdered". http://grammar.about.com/od/ab/g/apposterm.htm

[6]There was no mention for how complexity is measured in their analysis

[7]"The ARI decomposes text into its structural elements to provide the minimum reading level needed to understand a snippet of text, based on United States grade levels". $ARI = 4.71(c/w) + 0.5(w/s) - 21.43$: where c/w is number of characters/word while w/s is the number of words/sentence

[8]The API provided by text-processing.com was used here.

[9]Users with only one published review

They then created a template representing bursty patterns then try to fit it to the curves of the pre-defined measures. Using *longest common sub-string (LCS)* they calculated the number of matching points, hence the intensity of the burst. They then reported time windows where bursts are matched in the three metrics. They were able to detect dishonest stores by precision of 75.86%. It worth mentioning that this approach cannot detect specific deceptive reviews, but it rather detects deceptive stores and time-ranges.

5 Discussion

It seems from [Ott et al., 2011], and also by reviewing other papers that *text-categorization* tend to give better results than other approaches. [Ott et al., 2011] suggesting combining *text-categorization* with *Psycholinguistic Deception Detection* for even better results. I think *Time-Based Detection* can also be combined with *content-based analysis*, as well as other meta-data surrounding the review, such as ratings as suggested by [Harris, 2012]. The reviews user-name[10] and other domain specific data can also be taken into out analysis.

The small size of the corpus used for testing in most of the above papers seems to be a limiting factor here due to the scarcity of review spam data, compared to email spam data for example. However, it was strange that *text-categorization* - the one where smaller corpus tend to be less representative to real data due to frequency differences between them [Biber, 1993] - was still the one with superior results, compared to PoS which seems to be more imune to frequency changes due to corpus size.

References

[Biber, 1993] Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, 8(4):243–257.

[Chen and Goodman, 1996] Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. *CoRR*, cmp-lg/9606011.

[Chen and Xie, 2008] Chen, Y. and Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science*, 54(3):477–491.

[Day, 1971] Day, G. (1971). Attitude change, media and word of mouth. *Journal of Advertising Research*.

[Dunning, 1993] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

[Harris, 2012] Harris, C. (2012). Detecting deceptive opinion spam using human computation. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[Jackson,] Jackson, W. Sentence structure.

[Jindal and Liu, 2008] Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *WSDM*, pages 219–230.

[Joachims, 1998] Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. In *ECML*, pages 137–142.

[Joachims, 1999] Joachims, T. (1999). Making large scale svm learning practical.

[10] spammers names are created in masses, and sometimes they have fixed patters such as numbers at their end.

[Katz and Lazarsfeld, 1955] Katz, E. and Lazarsfeld, P. (1955). *Personal influence: The part played by people in the flow of mass communications.* Transaction Pub.

[Ott et al., 2011] Ott, M., Choi, Y., Cardie, C., and Hancock, J. (2011). Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557.*

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). *Opinion mining and sentiment analysis.* Now Pub.

[Pennebaker et al., 2007] Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., and Booth, R. (2007). The development and psychometric properties of liwc2007. *Austin, TX, LIWC. Net.*

[Rayson et al., 2001] Rayson, P., Wilson, A., and Leech, G. (2001). Grammatical word class variation within the british national corpus sampler. *Language and Computers*, 36(1):295–306.

[Sahami et al., 1998] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998). A bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*, volume 62, pages 98–105. Madison, Wisconsin: AAAI Technical Report WS-98-05.

[Skrob, 2005] Skrob, J. (2005). Open source and viral marketing. *University of Applied Science Kufstein, Austria.*

[Thomason, 2007] Thomason, A. (2007). Blog spam: A review. In *CEAS.*

[Whyte Jr, 1954] Whyte Jr, W. (1954). The web of word of mouth. *Fortune*, 50(1954):140–143.

[Xie et al., 2012] Xie, S., Wang, G., Lin, S., and Yu, P. S. (2012). Review spam detection via time series pattern discovery. In *WWW (Companion Volume)*, pages 635–636.

[Yoo and Gretzel, 2009] Yoo, K. and Gretzel, U. (2009). Comparison of deceptive and truthful travel reviews. In *Information and Communication Technologies in Tourism 2009: Proceedings of the International Conference in Amsterdam, The Netherlands, 2009*, page 37. Springer.