IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

An Overview of Opinion Mining in Spatial Database for Web Based Application

Janani. S[1], Dr. R. Manickachezian[2]

Research Scholar, Dr. Mahalingam Centre for Research and Development, NGM College, Pollachi, India[1]

Department of Computer Science, N G M College (Autonomous), Pollachi, Coimbatore, India[2]

Abstract: The Opinion mining is an ongoing field of research and development in web text mining and Data Engineering domain. It is the computational treatment of opinions and subjectivity of text. This survey paper mainly focus on a comprehensive overview of the Opinion mining algorithms and the different classification with their field of applications. Day to day proliferation of the current digital based economy a large amount of information is available in the form of textual data and user behaviours model which can often be used more easily if it is categorized or classified into some predefined classes. In case of social Networks, Ecommerce Architecture or a people hub, there are numerous varieties of people will be involved for their purchase, suggestions, posts, reviews, blogs and etc. The primary impression on the network is people suggestion. In Most of the networks, fake users revolving for two major purposes. 1) For increasing the rating of their own company. 2) To Decrease the rating of their competitor company. This survey is about how the opinion mining is used to find out fake users in spatial database.

Keywords: Opinion Mining, Spatial Database, Data engineering, User behaviour model, Fake users.

I. INTRODUCTION

In the present scenario, customers are more dependent on making decisions to buy products either on ecommerce sites or offline retail stores. Since these reviews are game changers for success or failure in sales of a product, reviews are being manipulated for positive or negative opinions. Manipulated reviews can also be referred to as fake/fraudulent reviews or opinion spam or untruthful reviews. In today's digital world deceptive opinion spam has become a threat to both customers and companies. Distinguishing these fake reviews is an important and difficult task. If the deceptive reviewers are often paid to write these reviews. As a result, it is a herculean task for an ordinary customer to differentiate fraudulent reviews from genuine ones, by looking at each review. There have been serious allegations about multi-national companies that are indulging in defaming competitor's products in the same sector. A recent investigation conducted by Taiwan's Fair Trade Commission revealed that Samsung's Taiwan unit called Open tide had hired people to write online reviews against HTC and recommending Samsung smart phones. The people who wrote the reviews, fore grounded what they outlined as flaws in the HTC gadgets and restrained any negative features about Samsung products [12].

Recently ecommerce giant amazon.com had admitted that it had fake reviews on its site and sued three websites accusing them of providing fake reviews [13], stipulating that they stop the practice. Fakespot.com has taken a lead in detecting fake reviews of products listed on amazon.com and its subsidiary ecommerce sites by providing percentage of fake reviews and grade. Reviews and ratings can directly influence customer purchase decisions. They are substantial to the success of businesses. While positive reviews with good ratings can provide financial improvements, negative reviews can harm the reputation and cause economic loss. Fake reviews and ratings can defile a business. It can affect how others view or purchase a product or service. So it is critical to determine fake/ fraudulent reviews. Traditional methods of data analysis have long been used to detect fake/fraudulent reviews. Early data analysis techniques were oriented toward extracting quantitative and statistical data characteristics. Some of these techniques facilitate useful data interpretations and can help to get better insights into the process behind data. To go beyond a traditional system, a data analysis system has to be equipped with considerable amount of background data, and be able to perform reasoning tasks involving that data. In effort to meet this goal researchers have turned to the fields of machine learning and artificial intelligence. A review can be classified as either fake or genuine either by using supervised and/or unsupervised learning techniques. These methods seek reviewer's profile, review data and activity of the reviewer on the Internet mostly using cookies by generating user profiles. Using either supervised or unsupervised method gives us only an indication of fraud probability. No stand alone statistical analysis can assure that a particular review is fraudulent one. It can only indicate that this review is more likely to be suspicious. Detection and filtering of genuine reviews is an interesting problem for the researchers and e-commerce sites. One such review site that filters fake reviews is yelp.com. The filter used in yelp.com to hide fake reviews from public is a trade secret. In this work we try to analyze Yelp Academic Challenge Dataset [4] and determine whether a review is genuine or fake.

DOI10.17148/IJARCCE.2017.6828

II. BACKGROUND STUDY

A number of studies have been conducted which focused on spam detection in e-mail and on the web, however, only recently have any studies been conducted on opinion spam. [Jindal and Liu (2008)][5] have worked on “Opinion Spam and Analysis” and have found that opinion spam is widespread and different in nature from either e-mail or Web spam. They have classified spam reviews into 3 types: Type 1, Type 2 and Type 3. Here Type 1 spam reviews are untruthful opinions that try to mislead readers or opinion mining systems by giving untruthful reviews to some target objects for their own gains. Type 2 spam reviews are brand only reviews, those that comment only on the brand and not on the products. Type 3 spam reviews are not actually reviews, they are mainly either advertisements or irrelevant reviews which do not contain any opinions about the target object or brand. Although humans detect this kind of opinion spam they need to be filtered, as it is a nuisance for the end user. Their investigation was based on 5.8 million reviews and 2.14 million reviewers (members who wrote at least one review) crawled from amazon.com and they have discovered that spam activities are widespread. They have regarded spam detection as a classification problem with two classes, spam and non-spam. And have built machine-learning models to classify a review as either spam or non-spam. They have detected type 2 and type 3 spam reviews by using supervised learning with manually labeled training examples and found that the highly effective model is logistic regression model. However, to detect type 1 opinion spam, they would have had to manually label training examples.

Thus they had to use duplicate spam reviews as positive training examples and other reviews as negative examples to build a model. In the paper "Finding Deceptive Opinion Spam by Any Stretch of the Imagination" by [Ott, et al. (2011)[10]], they have given focus to the deceptive opinion spam i.e. the fictitious opinions which are deliberately written to sound authentic so as to deceive the user. The user cannot easily identify this kind of opinion spam. They have mined all 5-star truthful reviews for 20 most famous hotels in Chicago area from trip advisor and deceptive opinions were gathered for the same hotels using amazon mechanical trunk (AMT). They first asked human judges to evaluate the review and then they have automated the task for the same set of reviews, and they found that automated classifiers outperform humans for each metric. The task was viewed as standard text categorization task, psycholinguistic deceptive detection and genre identification. The performance from each approach was compared and they found that the psycholinguistic deceptive detection and genre identification approach was outperformed by n-gram based text categorization, but a combined classifier of n-gram and psychological deception features achieved nearly 90% cross-validated accuracy. Finally they came into a conclusion that detecting deceptive opinions is well beyond the capabilities of humans. Since then, various dimensions have been explored: detecting individual [Lim et al., 2010][6] and group spammers [Mukherjee et al., 2012][7], time-series [Xie et al., 2012][8] and distributional analysis [Feng et al., 2012a][9] . [Yoo and Gretzel (2009)][15] gather 40 truthful and 42 deceptive hotel reviews and, using a standard statistical test, they have manually compared the psychologically relevant linguistic differences between them. In (Mukherjee, et al., 2013)[11], authors have briefly analyzed “What yelp filter might be doing?” by working with different combination of linguistic features like unigram, bigram, distribution of parts of speech tags and yielding detection accuracy. Authors have found that a combination of linguistic and behavioural features comparatively yielded more accuracy.

III. OVERVIEW OF OPINION MINING

a). News filtering and Organization:
Most of the news services today are electronic in nature in which a large volume of news articles are created every single day by the organizations. In such cases, it is difficult to organize the news articles manually. Therefore, automated methods can be very useful for news categorization in a variety of web portals [8]. This application is also referred to as text filtering.

b). Document Organization and Retrieval:
The above application is generally useful for many applications beyond news filtering and organization. A variety of supervised methods may be used for document organization in many domains. These include large digital libraries of documents, web collections, scientific literature, or even social feeds. Hierarchically organized document collections can be particularly useful for browsing and retrieval [9]. It is defined as matching some of stated user query against a set of free text records and full description of an information. A document retrieval system consists of a database of documents.

c). Email Classification and Spam Filtering:
It is often desirable to classify email [13] in order to determine either the subject or to determine junk email [13] in an automated way. This is also referred to as spam filtering or email filtering.

DOI10.17148/IJARCCE.2017.6828

d).Opinion Mining:
Customer reviews or opinions are often short text documents which can be mined to determine useful information from the review. Details on how classification can be used in order to perform opinion mining are discussed in [12].

IV. METHODOLOGIES

Each of the features discussed below are only for reviews of product/business

a) .Review length (RL)
Review length is the average number of words present in a review [11]. Usually the length of fake review will be on the lesser side because of the following reasons Reviewer will not be having much knowledge about the product/businessReviewer tries to achieve the objective with as few words as possible.

b) n-gramFrequency:
An n-gram [2] is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. These n-gram's typically are collected from a text or speech corpus. In this project we use unigram and bigram as important features for detection of fake reviews. Unigram is an n-gram of size 1 and Bigram is an n-gram of size 2. An n-gram model sequences notably natural languages, using statistical properties of n- gram phonemes and sequences of phonemes and modelled using n-gram distribution.

c) Unigram Frequency
Unigram frequency is a feature that deals with number of times each word unigram has occurred in a particular review. Note: The unigram is conditional on document length; the above gives the conditional likelihood of generating a particular set of frequencies given that their sum is l. The {wi} are the normalized word occurrence probabilities

d) Unigram Presence
Unigram presence is a feature that mainly finds out if a particular word unigram is present in a review. Subsequently upon receiving a query, a set of features corresponding to the query, such as the length and/or frequency of the query, **unigram** probabilities of respective words and/or groups of words in the query, presence of pre-designated words or phrases in the query, or the like, can be generated.

e) Bigram Frequency
Bigram frequency is a feature that deals with number of times each word bigram has occurred in a particular review. Frequency analysis is the practice of counting the number of occurrences of different cipher text characters in the hope that the information can be used to break ciphers. Frequency analysis is not only for single characters

f) Bigram Presence
Bigram presence is a feature that mainly finds out if a particular word bigram is present in a review. Bigram counts maintain the same principle as monogram counts, but instead of counting occurrences of single characters, bigram counts count the frequency of pairs of characters. It is one approach to statistical language identification often pair of characters occurs the text measures

f) Database Architecture
With reference to object-oriented thinking, each surface features can be abstracted as a class object with public properties, such as point , line , area and so on. Specific surface features are an instance of the object. It also has its own attributes and manages various objects hierarchically. It is good at describing the complex data types. Its shortcomings are lack of OODBS standard, development tools and defence mechanism. Its model is complex. ORDBMS (Object -Relational Spatial Database) has the features inherited from both of SQL of relation world and object world in essence. It also adds flexibility in data server. It supports complex "user-defined" application object and logic. It uses abstract data type which can hide any complex internal structure and properties to express spatial object. It also adds that type's operation in user-defined data types.

- Spatial databases in ORDBMS must support (at minimum)
- Complex (Ecommerce) data types
- Spatial data within related tables – feature classes, feature
- datasets Validation rules - subtypes and domains
- Spatial metadata

DOI10.17148/IJARCCE.2017.6828

V. RESULTS

Since the detection accuracy percentage varies with different sets of test reviews, we have used 5-fold cross validation technique by considering folds of trained dataset and test dataset in the ratio of 80:20. Test frequency accuracy obtained for unigram presence, unigram frequency, bigram presence, bigram frequency and review lengths are tabulated in table. In table 1.1 to represent the accuracy detection. If fig 1.1 comparison of frequency shown in graph chart.

Table 1.1 Accuracy detection

Classifiers/ Features	Applications	Logistic Regression	Decision Tree	Support Vector Machine
Spatial Database	E-Commerce application	85	70	60
Relational Database	Social Networks & Social Media	84	70	59
Operational Database	Secured Banking application	84	65	59
Distributed Database	Client server applications	82	62	58
End user database	Billing sotware, Tally and etc	79	62	53

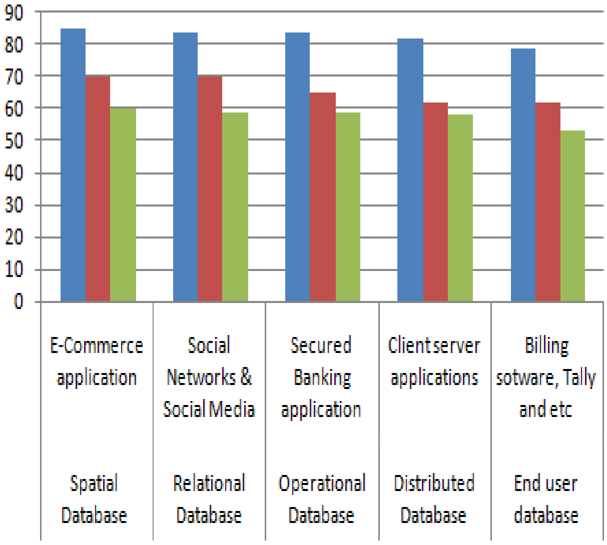

Figure 1.1 comparison of frequency

VI .CONCLUSION

Determining and classifying a review into a fake or truthful one is an important and challenging problem. In this paper, we have used linguistic features like unigram presence, unigram frequency, bigram presence, bigram frequency and review length to build a model and find fake reviews. After implementing the above model we have come to the conclusion that, detecting fake reviews requires both linguistic features and behavioural features.

REFERENCES

[1] Wikipedia- Supervised Learning http://en.wikipedia.org/wiki/Supervised_learning
[2] Wikipedia- n-gram http://en.wikipedia.org/wiki/N-gram
[3] Wikipedia- SVM (Support Vector Machine) http://en.wikipedia.org/wiki/Support_vector_machine
[4] Yelp Challenge Dataset http://www.yelp.com/dataset_challenge
[5] "Opinion Spam and Analysis" by Nitin Jindal and Bing Liu. ACM-2008.
[6] Lim, E., Nguyen, V., Jindal, N., Liu, B. Lauw, H. 2010. Detecting product review spammers using rating behavior. CIKM.
[7] Mukherjee, A., Liu, B. and Glance, N. 2012. Spotting fake reviewer groups in consumer reviews.
[8] Xie, S., Wang, G., Lin, S., and Yu, P.S. 2012. Review spam detection via temporal pattern discovery. KDD.
[9] Distributional Footprints of Deceptive Product Reviews by Feng, S., Xing, L., Gogar, A., and Choi, Y. 2012a. ICWSM.
[10] Ott, Myle, et al. "Finding deceptive opinion spam by any stretch of the imagination." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies- Volume 1. Association for Computational Linguistics, 2011
[11] Mukherjee, et al. "What Yelp Fake Review Filter Might Be Doing?" ICWSM. 2013.
[12] “Samsung probed in Taiwan over fake web reviews” –BBC News http://www.bbc.com/news/technology-22166606
[13] V. R. de Carvalho, W. Cohen. On the collective classification of email ”speech acts”, ACM SIGIR Conference, 2005.

DOI10.17148/IJARCCE.2017.6828

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

[14] "Amazon's Had Enough of Fake Reviews on Its Site, Files Lawsuit"- Yahoo Tech News https://www.yahoo.com/tech/amazonshad-enough-of-fake-reviews-on-its-site-116028350069.html
[15] "Comparison of deceptive and truthful reviews" by Yoo and Gretzel (2009).

BIOGRAPHIES

Janani.S received her B.Sc (Computer Science) from PSG College of Arts and Science, Coimbatore, India. She completed her Master of Computer Science (MSc) from Sri Ramakrishna College Of Arts and Science for Women, Coimbatore, India. Currently, she is a Research Scholar at Department of Computer Science, NGM College, Pollachi, India. She participated in a International Conference. Her area of interest includes Data mining, web content mining, Opinion mining.

Dr. R. ManickaChezian received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph.D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor of Computer Science in NGM College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published one-fifty papers in international/national journal and conferences: He is a recipient of many awards like Desha Mithra Award and Best Paper Award. Recently he received the award "Best Computer Science Faculty of the Year 2015" from Association of Scientists, Developers and Faculties. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.

DOI10.17148/IJARCCE.2017.6828