Opinion Mining With Spam Detection Using Real Time Data on Cloud

Salma Tarannum
M.tech CSE (3rd yr)
Integral University
Lucknow, India
salma.tarannum25@gmail.com

Mr. Shahid Hussain
M.tech CSE
Integral University
Lucknow, India
shahid@iul.ac.in

Mr. Jameel Ahmad
M.tech CSE
Integral University
Lucknow, India
jameel@reddifmail.com

ABSTRACT

In recent year, online reviews have become the most important resource of customer opinion. Now-a-days consumer can obtain information for products and service from online review resources, which can help them make decision. The social tools allow online user to interact, to express their opinions and to read opinions from other users. But the spammers provide comments which are written intentionally to mislead users by redirecting them to web sites to increase their rating and to promote products less known on the market. Reading spam comments is a bad experience and a waste of time for most of the online users but can also be harming and cause damage to the individual or an organization. This paper focuses on opinion spam detection methods which aims to find irregular or discontinuous text flow, vulgar language or not related to specific context and check similarity between the comments. The results from this experiment show that the methods provided herein can achieve the purpose of preliminary comment spam detection.

With the increased amount of data collection taking place as a result of social media interaction, scientific experiments, and even e-commerce applications, the nature of data as we know it has been evolving. As a result of this data generation from many different sources, “new generation” data, presents challenges as it is not all relational and lacks predefined structures. In this project we try to sort these issues and provide a way for better acquisition and processing of this type of data. We will be Analysing the real time social network data and try to eliminate the Fake reviews.

Keywords

Distributed Computing, Cloud Computing, Server, Sentiment Analysis, Python.

1. INTRODUCTION

Online social media is a popular platform where millions of people can communicate with each other in real time. These are the dynamic data sources where the users can create their own profiles and communicate with each other independent of geographical location. It provides communication platform with large scale and large extent. Furthermore these tools are beyond the boundaries of the physical world in studying human relationship and behaviours. As these social Medias are becoming more popular, cybercriminals have utilize these as a new platform for communicating different types of cybercrimes.

Nowadays, online social networking sites are becoming popular. This network has created social communication among the users and this leads to generation of huge amount of user generated data. In recent year, cyberbullying has become a major problem in these social networks. Cyberbullying is considered as a national health issue among online social network users. In our project we will develop a unique feature derived from Twitter such as network, activity, user and tweet contents. Depending on the feature we will develop a supervised machine learning solution for detection of cyberbullying in twitter.

Nowadays, different cybercrimes are happening such as phishing, spamming, spread of malware and cyberbullying is considered as a major problem along with the recent development of social media .It is a technique by which users get harass from other individual user of the group of user. It is also considered as a notational health problem in which there is high risk of suicidal ideation. Cyberbullying is a type of bullying with negative effects on the victim. Before an entire online community, a cyberbully can harass his/her victim. Online social media such as Facebook, twitter have become integral component of a user’s life. Because of this, these websites have become the most common platform for cyberbullying victimization. The characteristics of online social network have expanded the reach of cyberbullies to unreachable locations of countries.

Twitter is one of the common online social network services that provide the facility of communication. It enables users to read and send message of length 140 character. There are about 500 millions of users out of which 288 millions of users are active. Our study determined that twitter is becoming a cyberbullying playground.

2. SPAM DETECTION

Detecting review spam is challenging task as no one knows exactly the amount of spam in existence. Due to the openness of product review sites, spammers pose as different users contributing spammed reviews making them harder so eradicate completely. Spam reviews usually looking perfectly normal until one can compares them with other reviews of same products so as to identify that the review comments not consistent with latter. The efforts of additional comparisons by the users make the detection task tedious and non-trivial. One approach taken of review site such on Amazon.com is to allow users to label or vote the reviews so as helpful or not. Unfortunately, this still demands to user efforts and is subject to abuse of spammers. The state-of-the-art approach to review spam detection is to treat the reviews as the target of detection. This approach represents review by review-, reviewer- and product- level features, and trains a classifier so as to distinguish spam reviews from non-spam ones. However, these features may provide direct evidence against the spammed review. Both are behaviours of reviewer that to deviate from normal practice and highly suspicious of review manipulation. This suggests that the one should focus on detecting spammers based on their spamming, instead of detecting spam reviews. In fact, the more spamming behaviours we can detect for a reviewer, the more likely the reviewer is a spammer. Subsequently, the reviews to this reviewer can be removed so to protect the interests of other review users. Without doing this the customer is never going to get the quality reviews and thus the decision making will not be an easy task.

3. OPINION MINING

The increase in the data rates generated on the digital universe is escalating exponentially. With a view in employing current tools and technologies to analyse and store, a massive volume of data are not up to the mark, since they are unable to extract required sample data sets. Therefore, we must design an architectural platform for analysing both remote access real time and offline data. When a business enterprise can pull-out all the useful information obtainable in the Big Data rather than a sample of its data set, in that case, it has an influential benefit over the market competitors. Big Data analytics helps us to gain insight and make better decisions. Therefore, with the intentions of using Big Data, modifications in paradigms are at utmost.

To support our motivations, we have described some areas where Big Data can play an important role. In healthcare scenarios, medical practitioners gather massive volume of data about patients, medical history, medications, and other details. The above-mentioned data are accumulated in drug-manufacturing companies. The nature of these data is very complex, and sometimes the practitioners are unable to show a relationship with other information, which results in missing of important information. With a view in employing advance analytic techniques for organizing and extracting useful information from Big Data results in personalized medication, the advance Big Data analytic techniques give insight into her editorially causes of the disease. In the Same way data is also generated for the reviews of the product across various services but sometimes we have to differentiate between fake reviews and Genuine Reviews for the input of our decision making process in business.

4. RELATED WORK

[1] Accessing the trustworthiness of reviews is a key issue for the maintainers of opinion sites such as TripAdvisor, given the rewards that can be derived from posting false or biased reviews. In this paper we present a number of criteria that might be indicative of suspicious reviews and evaluate alternative methods for integrating these criteria to produce a unified "suspiciousness" ranking. The criteria derive from characteristics of the network of reviewers and also from analysis of the content and impact of reviews and ratings. The integration methods that are evaluated are singular value decomposition and the unsupervised hedge algorithm. These alternatives are evaluated in a user study on TripAdvisor reviews, where volunteers were asked to rate the suspiciousness of reviews that have been highlighted by the criteria.

[2]Social networks are useful for judging the trustworthiness of outsiders. An automated antispam tool exploits the properties of social networks to distinguish between unsolicited commercial e-mail spam and messages associated with people the user knows.

[3] We study how an online community perceives the relative quality of its own user-contributed content, which has important implications for the successful self-regulation and growth of the Social Web in the presence of increasing spam and a flood of Social Web metadata. We propose and evaluate a machine learning-based approach for ranking comments on the Social Web based on the community's expressed preferences, which can be used to promote high-quality comments and filter out low-quality comments. We study several factors impacting community preference, including the contributor's reputation and community activity level, as well as the complexity and richness of the comment. Through experiments, we find that the proposed approach results in significant improvement in ranking quality versus alternative approaches.

[4] In the research to date, the performance of recommender systems has been extensively evaluated across various dimensions. Increasingly, the issue of robustness against malicious attack is receiving attention from the research community. In previous work, we have shown that knowledge of certain domain statistics is sufficient to allow successful attacks to be mounted against recommender systems. In this paper, we examine the extent of domain knowledge that is actually required and find that, even when little such knowledge is known, it remains possible to mount successful attacks.

[5] User-generated reviews are a common and valuable source of product information, yet little attention has been paid as to how best to present them to end-users. In this paper, we describe a classification-based recommender system that is designed to recommend the most helpful reviews for a given product. We present a large-scale evaluation of our approach using TripAdvisor hotel reviews, and we show that our approach is capable of suggesting superior reviews compared to a number of alternative recommendation benchmarks.

[6] Pattern ordering is an important task in data mining because the number of patterns extracted by standard data mining algorithms often exceeds our capacity to manually analyze them. In this paper, we present an effective approach to address the pattern ordering problem by combining the rank information gathered from disparate sources. Although rank aggregation techniques have been developed for applications such as meta-search engines, they are not directly applicable to pattern ordering for two reasons. First, the techniques are mostly supervised, i.e., they require a sufficient amount of

IJSER

labeled data. Second, the objects to be ranked are assumed to be independent and identically distributed (i.i.d), an assumption that seldom holds in pattern ordering. The method proposed in this paper is an adaptation of the original Hedge algorithm, modified to work in an unsupervised learning setting. Techniques for addressing the i.i.d. violation in pattern ordering are also presented. Experimental results demonstrate that our unsupervised Hedge algorithm outperforms many alternative techniques such as those based on weighted average ranking and singular value decomposition.

[7] The sentiment detection of texts has been witnessed a booming interest in recent years, due to the increased availability of online reviews in digital form and the ensuing need to organize them. Till to now, there are mainly four different problems predominating in this research community, namely, subjectivity classification, word sentiment classification, document sentiment classification and opinion extraction. In fact, there are inherent relations between them. Subjectivity classification can prevent the sentiment classifier from considering irrelevant or even potentially misleading text. Document sentiment classification and opinion extraction have often involved word sentiment classification techniques. This survey discusses related issues and main approaches to these problems.

[8] Collaborative filters help people make choices based on the opinions of other people. Group Lens is a system for collaborative filtering of net news, to help people find articles they will like in the huge stream of available articles. News reader clients display predicted scores and make it easy for users to rate articles after they read them. Rating servers, called Better Bit Bureaus, gather and disseminate the ratings. The rating servers predict scores based on the heuristic that people who agreed in the past will probably agree again. Users can protect their privacy by entering ratings under pseudonyms, without reducing the effectiveness of the score prediction. The entire architecture is open: alternative software for news clients and Better Bit Bureaus can be developed independently and can interoperate with the components we have developed.

[9] Collaborative filtering or recommender systems use a database about user preferences to predict additional topics or products a new user might like. In this paper we describe several algorithms designed for this task, including techniques based on correlation coefficients, vector-based similarity calculations, and statistical Bayesian methods. We compare the predictive accuracy of the various methods in a set of representative problem domains. We use two basic classes of evaluation metrics. The first characterizes accuracy over a set of individual predictions in terms of average absolute deviation. The second estimates the utility of a ranked list of suggested items. This metric uses an estimate of the probability that a user will see a recommendation in an ordered list. Experiments were run for datasets associated with 3 application areas, 4 experimental protocols, and the 2 evaluation metrics for the various algorithms. Results indicate that for a wide range of conditions, Bayesian networks with decision trees at each node and correlation methods outperform Bayesian-clustering and vector-similarity methods. Between correlation and Bayesian networks, the preferred method depends on the nature of the dataset, nature of the application (ranked versus one-by-one presentation), and the availability of votes with which to make predictions. Other considerations include the size of database, speed of predictions, and learning time.

[10] One of the important types of information on the Web is the opinions expressed in the user generated content, e.g., customer reviews of products, forum posts, and blogs. In this paper, we focus on customer reviews of products. In particular, we study the problem of determining the semantic orientations (positive, negative or neutral) of opinions expressed on product features in reviews. This problem has many applications, e.g., opinion mining, summarization and search. Most existing techniques utilize a list of opinion (bearing) words (also called opinion lexicon) for the purpose. Opinion words are words that express desirable (e.g., great, amazing, etc.) or undesirable (e.g., bad, poor, etc) states. These approaches, however, all have some major shortcomings. In this paper, we propose a holistic lexicon-based approach to solving the problem by exploiting external evidences and linguistic conventions of natural language expressions. This approach allows the system to handle opinion words that are context dependent, which cause major difficulties for existing algorithms. It also deals with many special words, phrases and language constructs which have impacts on opinions based on their linguistic patterns. It also has an effective function for aggregating multiple conflicting opinion words in a sentence. A system, called Opinion Observer, based on the proposed technique has been implemented. Experimental results using a benchmark product review data set and some additional reviews show that the proposed technique is highly effective. It outperforms existing methods significantly.

5. COMPARATIVE STUDY

Method used	Description	Advantages
Integration methods using singular value decompositi on and unsupervise d hedge algorithm	Integration was done based on the criteria. Criteria was derived from the identifying nature of the reviewers and also from classifying the content and effect of reviews and rating.	This alternative is used for user analysis on website like trip advisor reviews and also highlighting the suspiciousness of the reviews.

IJSER

Anti-spam tool for verifying the consistency of social networks.	This tool uses the social network characteristics to discriminate between the spam, commercial emails and original messages associated with the users.	This approach leads to trusted binding and allows user to analyses between actual messages and other messages. This increases trustworthiness.
Machine learning based methods	These approaches are used to discriminate the quality of comments using various factors that has a huge effect on community and contributors.	This approach improves the quality and authenticity of the comments on social web.
Classification based recommender	. This method designed a system which provides most quality reviews of any given product based on evaluation.	This approach increases User data understanding related to any product.
Hedge algorithm in Unsupervised learning	It has been designed to address the shortcomings of pattern Ordering. This technique uses singular value decomposition and weighted average ranking.	This scheme increases the accuracy of Pattern ordering.
Sentiment classifier	Sentiment classifier are used to identify the opinion extraction and document sentiment classification.	Sentiment analysis increases the accuracy of online reviews by organizing them.
Heuristic approach	This method uses the past opinion of the users to predict the opinions the new user might like. It allows user to get the best output.	Increases the users understanding which benefits user in taking decision.
Algorithms like Bayesian method, correlation coefficient, vector based similarity calculations	These methods make predictions on evaluation metrics, which include individual prediction and the ranked list of selected items.	It increases the speed of prediction and learning time.
Lexicon-based approach using natural language processing.	This method observes the semantic orientation of opinions regarding any product features expressed I reviews.	Highly effective technique and significant performance.

IJSER

6. CONCLUSION

Sentiment analysis methods till now have been used to detect the polarity in the thoughts and opinions of all the users that access social media. Researchers and Businesses are very interested to understand the thoughts of people and how they respond to everything happening around them. Companies use this to evaluate their advertisement campaigns and to improve their products. There is too much potential in machine learning, overtaking some of the manual labor of some lexicon based tasks that are labor intensive. For example, lexicon sentiment creation is labor intensive and there are already unsupervised methods to create them. This is where machine learning will play a crucial role. Such algorithms will also have to understand and analyze natural text concept-wise and context-wise. Time will also be a crucial element looking at the amount of data that is being generated on the Web today. Collecting opinions on the web will still requires processing that can filter out un-opinionated user-generated content and also to test the trustworthiness of the opinion and its source. There is a lot of scope in analyzing the video and images on the web. Now a days, with the advent of Facebook, Instagram and Video vines people are expressing their thoughts with pictures and videos along with text. Sentiment analysis will have to pace up with this change. Tools which are helping companies to change strategies based on Facebook and Twitter will also have to accommodate the number of likes and re-tweets that the thought is generating on the Social media. People follow and unfollow people and comments on Social Media but never comment so there is scope in analyzing these aspects of the Web as well.

7. FUTURE WORK

Although this project has lots of advantages but there are certain constraints attached with it like down time, cost, privacy, security, sustainability, version Control. These all constraints can be removed or improved in future work.

8. REFRENCE

[1] Merging multiple criteria to identify suspicious reviews Authors: Guangyu Wu University College Dublin, Dublin, Ireland Derek Greene University College Dublin, Dublin, IrelandPádraig Cunningham University College Dublin, Dublin, Ireland.

[2] Leveraging Social Networks to Fight Spam Authors: P. Oscar Boykin University of Florida Vwani P. Roychowdhury University of California, Los Angeles.

[3] Ranking Comments on the Social Web Authors: Chiao-Fang Hsu , Elham Khabiri ,James Caverle

[4] Recommender systems: attack types and strategies Authors: Michael P. O'Mahony University College Dublin, Belfield, Dublin 4, Ireland Neil J. Hurley University College Dublin, Belfield, Dublin 4, Ireland Guénolé C. M. Silvestre University College Dublin, Belfield, Dublin 4, Ireland.

[5] Learning to recommend helpful hotel reviews Full Text: PDFPDF Get this ArticleGet this Article Authors: Michael P. O'Mahony University College Dublin, Dublin, Ireland Barry Smyth University College Dublin, Dublin, Ireland

[6] Ordering patterns by combining opinions from multiple sources Full Text: PDFPDF Get this ArticleGet this ArticleAuthors: Pang-Ning Tan Michigan State University, Rong Jin Michigan State University.

[7] A survey on sentiment detection of reviews Authors: Huifeng Tang Information Security Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, PR China Songbo Tan Information Security Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, PR China Xueqi Cheng Information Security Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, PR China.

[8] GroupLens: an open architecture for collaborative filtering of netnews Authors: Paul Resnick MIT Center for Coordination Science, Room E53-325, 50 Memorial Drive, Cambridge, MA Neophytos Iacovou University of Minnesota, Department of Computer Science, Minneapolis, Minnesota Mitesh Suchak MIT Center for Coordination Science, Room E53-325, 50 Memorial Drive, Cambridge, MPeter Bergstrom University of Minnesota, Department of Computer Science, Minneapolis, Minnesota John Riedl University of Minnesota, Department of Computer Science, Minneapolis, Minnesota.

[9] Empirical analysis of predictive algorithms for collaborative filtering Authors: John S. Breese Microsoft Research, Redmond, WA David Heckerman Microsoft Research, Redmond, WA Carl Kadie Microsoft Research, Redmond, WA

[10] A holistic lexicon-based approach to opinion mining Authors: Xiaowen Ding University of Illinois at Chicago, Chicago, IL Bing Liu University of Illinois at Chicago, Chicago, IL Philip S. Yu University of Illinois at Chicago, Chicago, IL.

IJSER