INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 5 ISSUE 2 (2019) PAGES 4374 - 4378
RECEIVED : 22.03.2019 PUBLISHED : 27.04.2019
April 27,2019

Review Spam detection using machine learning

Karthik D R,Heena Khanum ,Shreethrisha[1]
[1]UG scholor, department of information science and Engineering
[2]Mr Yogaprakash MG assist prof, Department of information science andengineering
[1,2]BGS Institute of Technology, B G Nagar mandya Karnataka.

Abstract:
Prior to buying a product, people usually inform themselves by reading online reviews. To make more profit sellers often try to fake user experience. As customers are being deceived this way, recognizing and removing fake reviews is of great importance. This paper analyzes spam detection methods, based on machine learning, and presents their overview and results.

1. INTRODUCTION

Machine learning is a field of computer science that allows computers to learn from data without being explicitly programmed. Supervised learning, a subfield of machine learning, needs labeled data to be able to learn. Data is labeled by human experts or some system whose behavior should be mimicked. During the training process, algorithm tries to find relationship between input (data) and output (labels). After the training, system can be used on unlabeled data. Algorithms used by methods in this paper belong to supervised learning algorithms. As Internet continues to grow, online reviews are becoming more relevant source of information. Knowing that products' success depends on customer reviews; sellers often try to deceive buyers by posting fake comments. Sellers can post reviews themselves or pay other individuals to do it for them. This practice of posting fraudulent reviews is known as opinion or review spam. Spammers can be hired to post positive reviews, or to write bad reviews to damage competitors' business. Canadian Competition Burau issued an official warning to their citizens in 2014, stating that they should be aware of fraudulent reviews and estimating that third of reviews found online are fake. Poll taken on over 25 000 participants in 2009, says that over 70% consumers believe online reviews. This shows that spam reviews present a major concern today. To tackle this problem many methods have been proposed during the last decade.

2. RELATED WORK

The problem as a het-erogeneous network where nodes are either real components in a dataset (such as reviews, users and products) or spam features. To better understand the proposed framework we first present an overview of some of the concepts and definitions in heterogeneous information networks. Analysis is a detailed study of the various operations performed by a system and their relationships within and outside of the system. One aspect of analysis is defining the boundaries of the system and determining whether or not a candidate system should consider other related systems. During analysis data are collected on the available files decision points and transactions handled by the present system. This involves gathering information and using structured tools for analysis. In the Existing system the problem as a het-erogeneous network where nodes are either real components in a dataset (such as reviews, users and products) or spam features. To better understand the proposed framework we first present an overview of some of the concepts

and definitions in heterogeneous information networks. A new weighting method for spam features is proposed to determine the relative importance of each feature and shows how effective each of features are in identifying spams from normal reviews. Previous works , also aimed to address the importance of features mainly in term of obtained accuracy, but not as a build-in function in their framework (i.e., their approach is dependent to ground truth for determining each feature importance). As we explain in our unsupervised approach, model is able to find features Feasibility is the determination of whether or not a project is worth doing. The process followed in making this determination is called feasibility Study. This type of study if a project can and should be taken. In the conduct of the feasibility study, the analyst will usually consider seven distinct, but inter-related types of feasibility.

3. DESCRIPTION

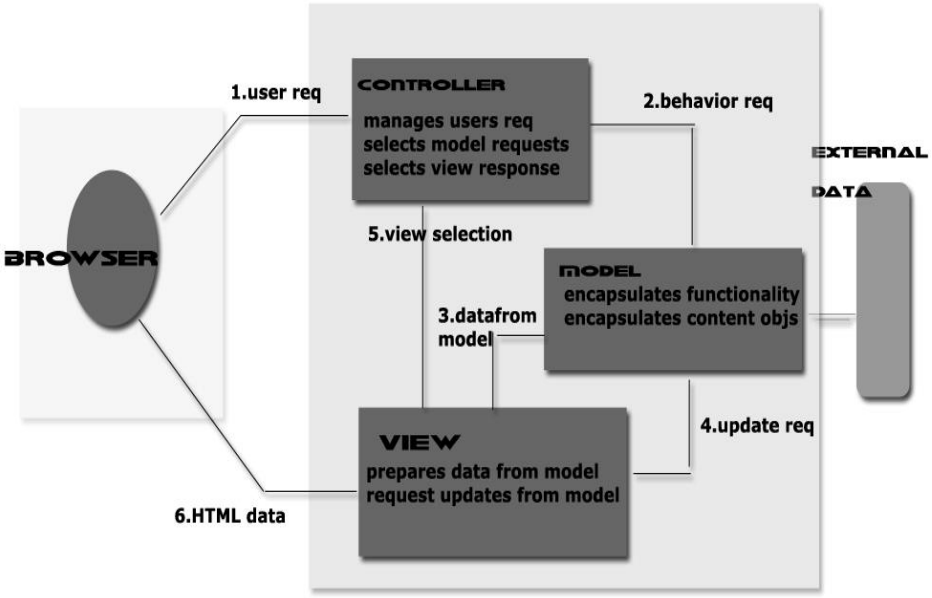
In the Upload Excel File Module, user has to select the file from the client machine and the file content will be sent to the server via URL in the form of multipart ,in the server side servlet receives the file content and write the file content in the folder of the application. From that folder it reads the file content and store the file content in to the database. In the Fake Review Detection 1 process , Data will be read from the database and checks whether the IP_Address and UserID is fake or not based on the meta data table and insert the fake reviews in to the fake review table . And also it checks whether the number of reviews from the IP_Address are exceeding the threshold limit with in the threshold time limit, if any IP_Address exceeds the threshold limit,then that reviews will be inserted to the fake review table and that IP_Address will be inserted to the Meta Fake IP_Address table and rest of the reviews will be inserted to the Real reviews table. In the Fake Review Detection 2 process , reviews will be read from the Real reviews table, considering each reviews , in the first level , unnecessary words and special characters will be removed, in the second level categorize each and every word is noun or adjective , in the third level paring the noun and adjacent adjective , in the fourth level checks whether the adjective which is paired with the noun is negative or positive , in the fifth level checks whether the maximum number of pairs are positive or negative , based on the maximum count of positive or negative, assign the review value as positive or negative ,in the sixth level calculate and insert the two gram and three gram pairs in to the database, in the seventh level calculate the count percentage, positive percentage and n-gram percentage of each user and add all the percentages and get total percentage threshold , if any user exceeds total percentage threshold, consider that user is fake and insert that user in to the meta fake user table.

4. SYSTEM ANALYSIS

Design leads to a model that contains the appropriate mix of aesthetics, content, and technology. The mix will vary depending upon the nature of the WebApp, and as a consequence the design activities that are emphasized will also vary. *Validation testing* is a concern which overlaps with integration testing. Ensuring that the application fulfils its specification is a major criterion for the construction of an integration test. Validation testing also overlaps to a large extent with *System Testing*, where the application is tested with respect to its typical working environment. Consequently for many processes no clear division between validation and system testing can be made. Specific tests which can be performed in either or both stages include the following. The final process should be a Software audit where the complete software project is checked to ensure that it meets production management requirements. This ensures that all required documentation has been produced, is in the correct format and is of acceptable quality. The purpose of this review is: firstly to assure the quality of the production process and by implication construction phase

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 5 ISSUE 2 (2019) PAGES 4374 - 4378
RECEIVED : 22.03.2019 PUBLISHED : 27.04.2019
April 27,2019

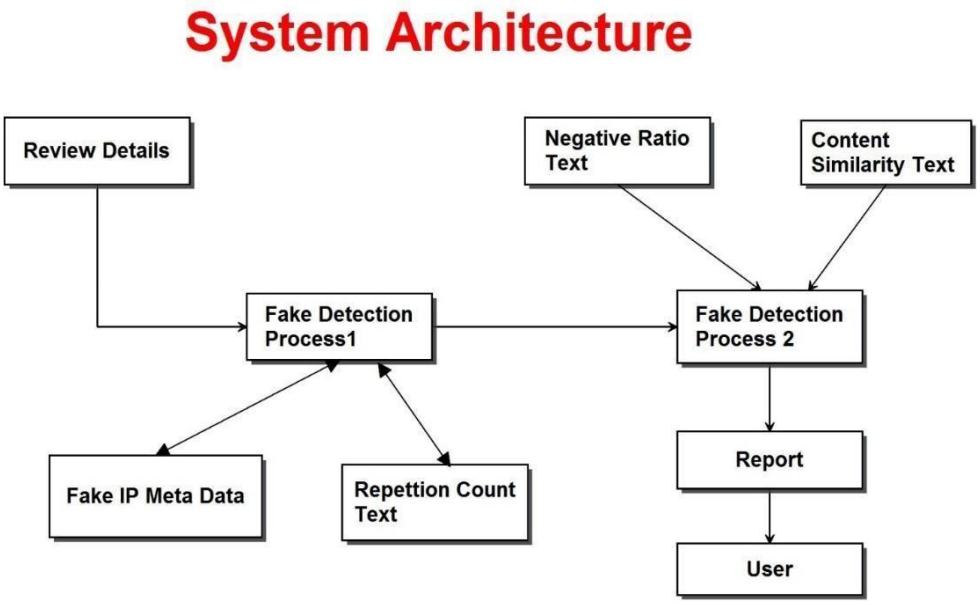
commences. A formal hand over from the development team at the end of the audit will mark the transition between the two phases.



Integration Testing can proceed in a number of different ways, which can be broadly characterized as top down or bottom up. In top down integration testing the high level control routines are tested first, possibly with the middle level control structures present only as stubs. Subprogram stubs were presented in section2 as incomplete subprograms which are only present to allow the higher. Level control routines to be tested. The other major category of integration testing is *Bottom Up Integration Testing* where an individual module is tested form a test harness. Once a set of individual module have been tested they are then combined into a collection of modules ,known as builds, which are then tested by a second test harness. This process can continue until the build consists of the entire application. In practice a combination of top down and bottom-up testing would be used. In a large software project being developed by a number of sub-teams, or a smaller project where different modules were built by individuals. The sub teams or individuals would conduct bottom-up testing of the modules which they were constructing before releasing them to an integration team which would assemble them together for top-down testing.

Internal and unit testing can be automated with the help of coverage tools. Analyzes the source code and generated a test that will execute every alternative thread of execution. Typically, the coverage tool is used in a slightly different way. First the coverage tool is used to augment the source by placing information prints after each line of code. Then the testing suite is executed generating an audit trail. This audit trail is analyzed and reports the percent of the total system code executed during the test suite. If the coverage is

high and the untested source lines are of low impact to the system's overall quality, then no more additional tests are required.



Internal and unit testing can be automated with the help of coverage tools. Analyzes the source code and generated a test that will execute every alternative thread of execution. Typically, the coverage tool is used in a slightly different way. First the coverage tool is used to augment the source by placing information prints after each line of code. Then the testing suite is executed generating an audit trail. This audit trail is analyzed and reports the percent of the total system code executed during the test suite. If the coverage is high and the untested source lines are of low impact to the system's overall quality, then no more additional tests are required.

CONCLUSION

In the construction industry, scaffolding is a temporary, easy to assemble and disassemble, frame placed around a building to facilitate the construction of the building. The construction workers first build the scaffolding and then the building. Later the scaffolding is removed, exposing the completed building. similarly, in software testing, one particular test may need some supporting software. This software establishes can a correct evaluation of the test take place. The scaffolding software may establish state and values for data structures as well as providing dummy external functions for the test. Different scaffolding software may be needed form one test to another test. Scaffolding software rarely is considered part of the system. Some times the scaffolding software becomes larger than the system software being tested. Usually the scaffolding software is not of the same quality as the system software and frequently is quite fragile. A small change in test may lead to much larger changes in the scaffolding.

REFERENCES

[1] J. Donfro, A whopping 20 % of yelp reviews are fake. http://www.businessinsider.com/20-percent-of

yelp-reviews-fake-2013-9. Accessed: 2015-07-30.
[2] M. Ott, C. Cardie, and J. T. Hancock. Estimating the prevalence of deception in online review communities. In ACM WWW, 2012.
[3] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination.In ACL, 2011.
[4] Ch. Xu and J. Zhang. Combating product review spam campaigns via multiple heterogeneous pairwise features. In SIAM International Conference on Data Mining, 2014.
[5] N. Jindal and B. Liu. Opinion spam and analysis. In WSDM, 2008.
[6] F. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. Proceedings of the 22nd International Joint Conference on Artificial Intelligence; IJCAI, 2011.
[7] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Exploiting burstiness in reviews for review spammer detection. In ICWSM, 2013.
[8] A. j. Minnich, N. Chavoshi, A. Mueen, S. Luan, and M. Faloutsos. Trueview: Harnessing the power of multiple review sites. In ACM WWW, 2015.