IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 11, May 2023

Detecting and Analyzing Spam Reviews Using ML

Arati Avhad[1], Navnath Avhad[2], Geeta Nagare[3], Kalyani Sanap[4], Prof. Dr. R. S. Khule[5]

Students, Department of Information Technology[1,2,3,4]

Professor. Department of Information Technology[5]

Matoshri Collage of Engineering, Nashik, India

Abstract: *Now a days with the increasing popularity of internet, online marketing is going to become more and more popular. This is because, a lot of products and services are easily available online. Hence, reviews about these all products and services are very important for customers as well as organizations. 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.*

Keywords: Machine Learning, BERT Algorithm, Spam Detection, Reviews Spam

I. INTRODUCTION

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.

II. LITERATURE SURVEY

Literature survey is gathering the information of previous work done related to your project. It contains the research study year, researchers name, technologies used and drawback of the system. Detection of opinion spam was first introduced by Jindal & Liu in 2008. They categorized the review spam into 3 categories: Untruthful opinions (if fraudsters write positive fake opinions to promote some targets is called as hyper spam and if fraudsters write negative fake opinions to damage the reputation of some targets is called as defaming spam), reviews on brands only (fraudsters write only about the brand, i.e. the manufacturers of the products rather than the products) and non-reviews (fraudsters write something that is totally unrelated to the products, this may be either advertisements or irrelevant opinion). Authors introduced three types of featurein their proposed work i.e., review centric features, reviewer centric features and product centric features. Lim et al. proposed a model that is based on behavior of spammers. They used to assign a rank to spammer on the basis of behavior scoring method and they detect spammers according to that rank. Authors collected data set from amazon.com and applied the concept of both behavior scoring method and supervised learning technique to detect review spammers.

III. MODELLING AND DESIGN

3.1 SDLC–Waterfall Model

The Waterfall Model was the first Process Model to be introduced. It is also referred to asa **linear-sequential life cycle model**. It is very simple to understand and use. In a waterfall model, each phase must be completed before the next phase can begin and there is no overlapping in the phases Level-2 Heading: A level-2 heading must be in Italic, left-justified and numbered using an uppercase alphabetic letter followed by a period. For example, see heading "C. Section Headings" above .The Waterfall model is the earliest SDLC approach that was used for software development .The waterfall Model illustrates the software development process in a linear sequential flow.This means that any phase in the development process begins only if the previous phase is complete. In this waterfall model, the phases do not overlap.

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 11, May 2023

Waterfall Model–Design

Waterfall approach was first SDLC Model to be used widely in Software Engineering to ensure success of the project. In "The Waterfall" approach, the whole process of software development is divided into separate phases .

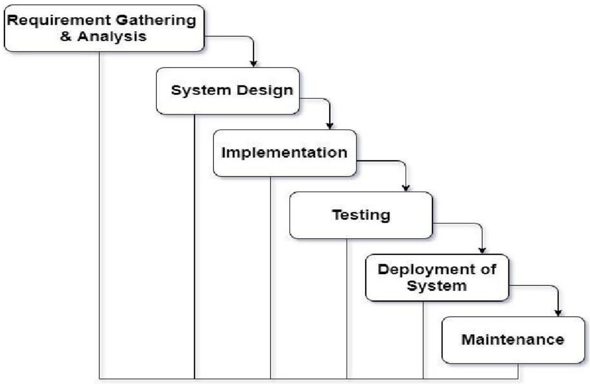


Fig 1.0. Waterfall Model

3.2 Block Diagram

As shown in above architecture a dataset has been collected and a classification model has been developed and evaluated. The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling. A dataset used for machine learning should be partitioned into three subsets training, test, and validation sets. Training set. A data scientist uses a training set to train a model and define its optimal parameters it has to learn from data. Test set.

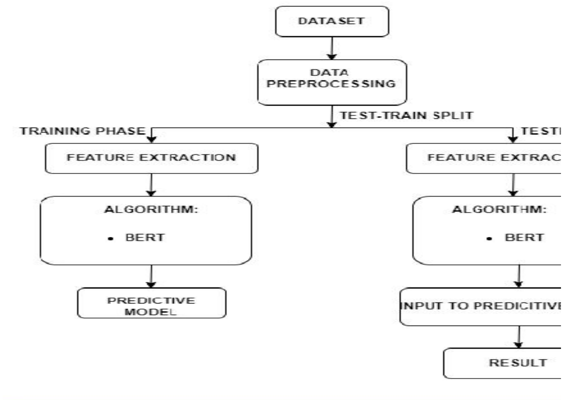


Fig 2.0. Block diagram

2.3 System Architecture

in above architecture a dataset has been collected and a classification model has been developed and evaluated. The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling. A dataset used for machine learning should be partitioned into three subsets training, test, and validation sets. Training set.

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 11, May 2023

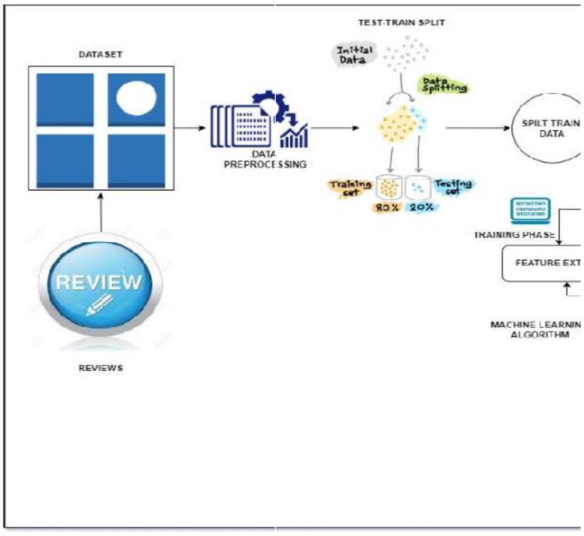


Fig. system architecture

2.4 DFD Diagram

Data objects represented by labeled arrows and transformation are represented by circles also called as bubbles. DFD is presented in a hierarchical fashion i.e, the first data flow model represents the system as a whole. Subsequent DFD refine the context diagram (level 0 DFD), providing increasing details with each subsequent level. The DFD enables the software engineer to develop models of the information domain & functional domain at the same time. As the DFD is refined into greater levels of details, the analyst perform an implicit functional 25 decomposition of the system.

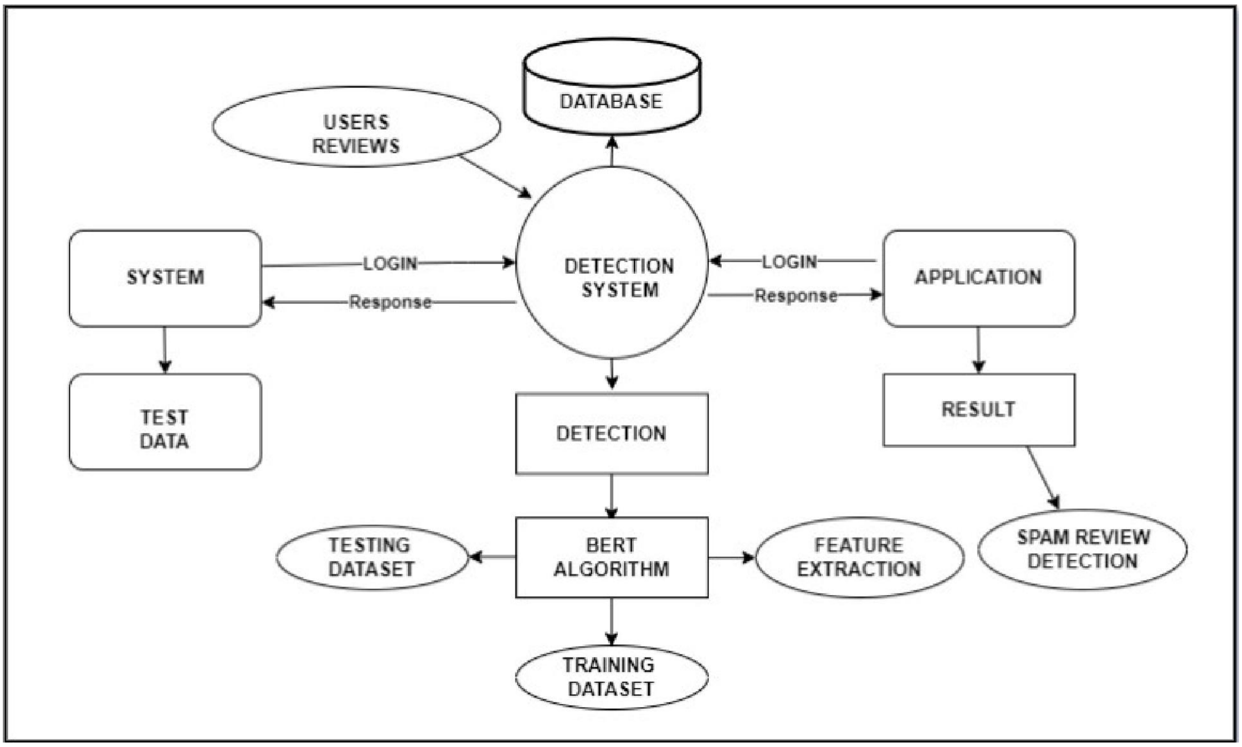
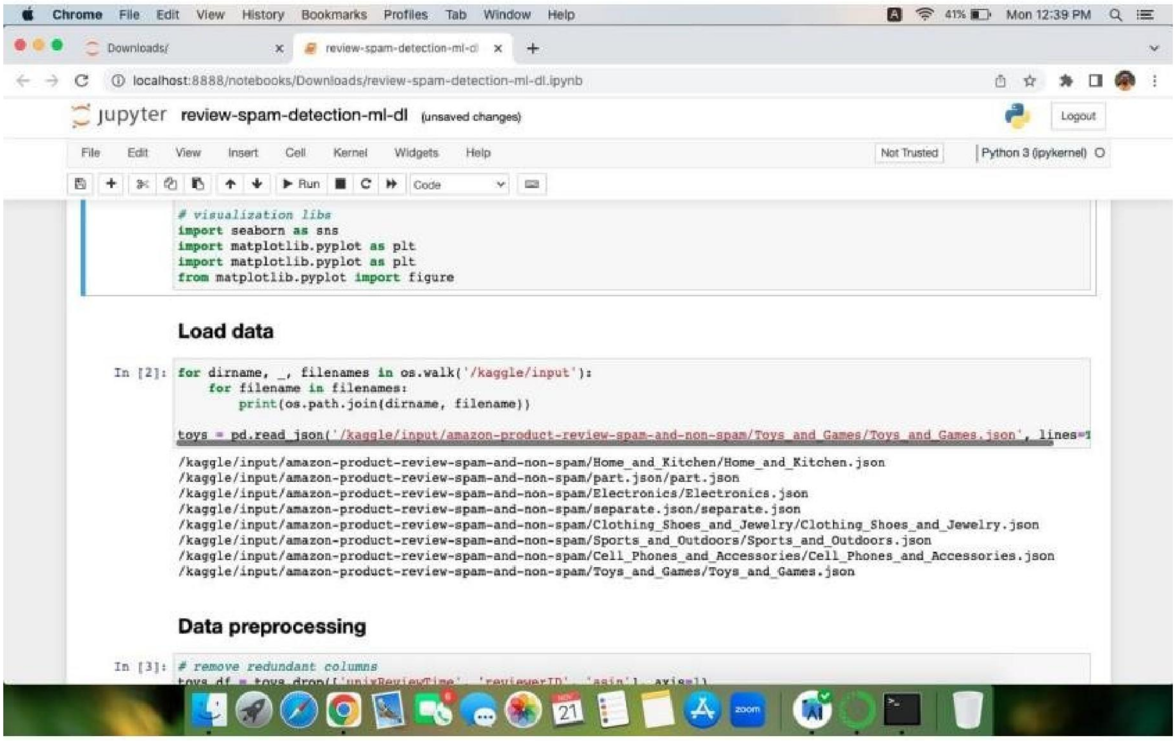
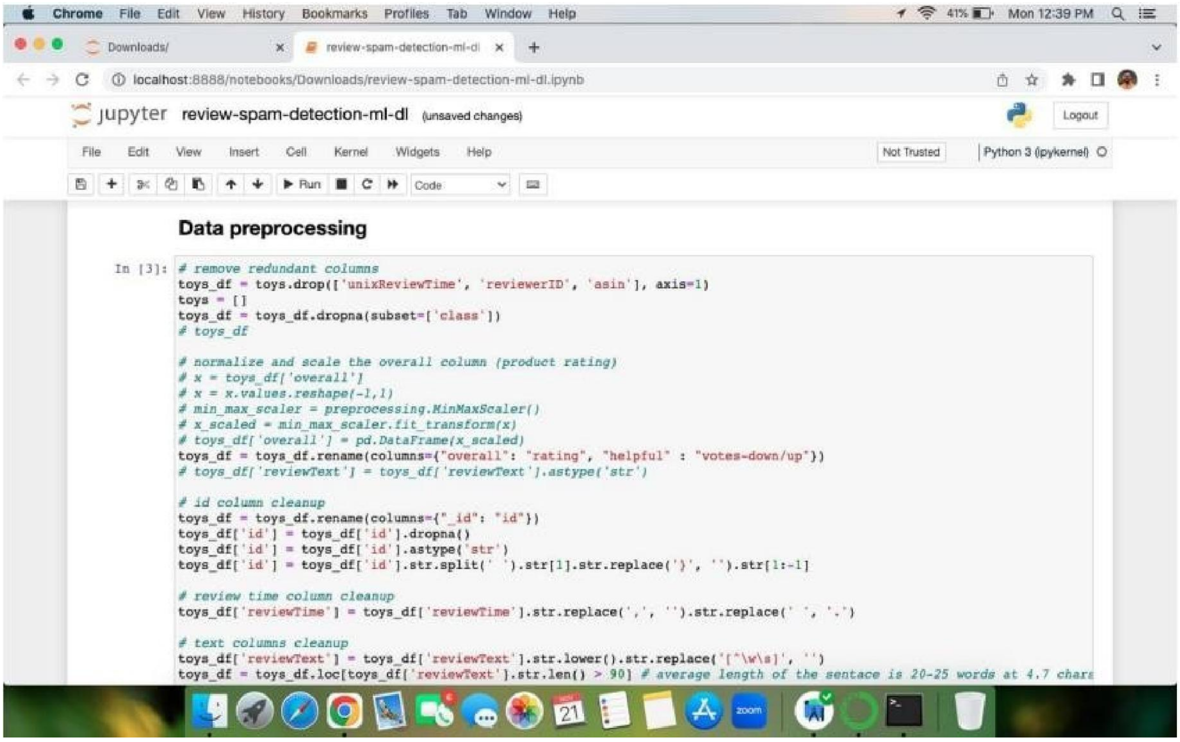


Fig. Data-Flow Diagram

Load Data:



Data Preprocessing:



IJARSCT

ISSN (Online) 2581-9429

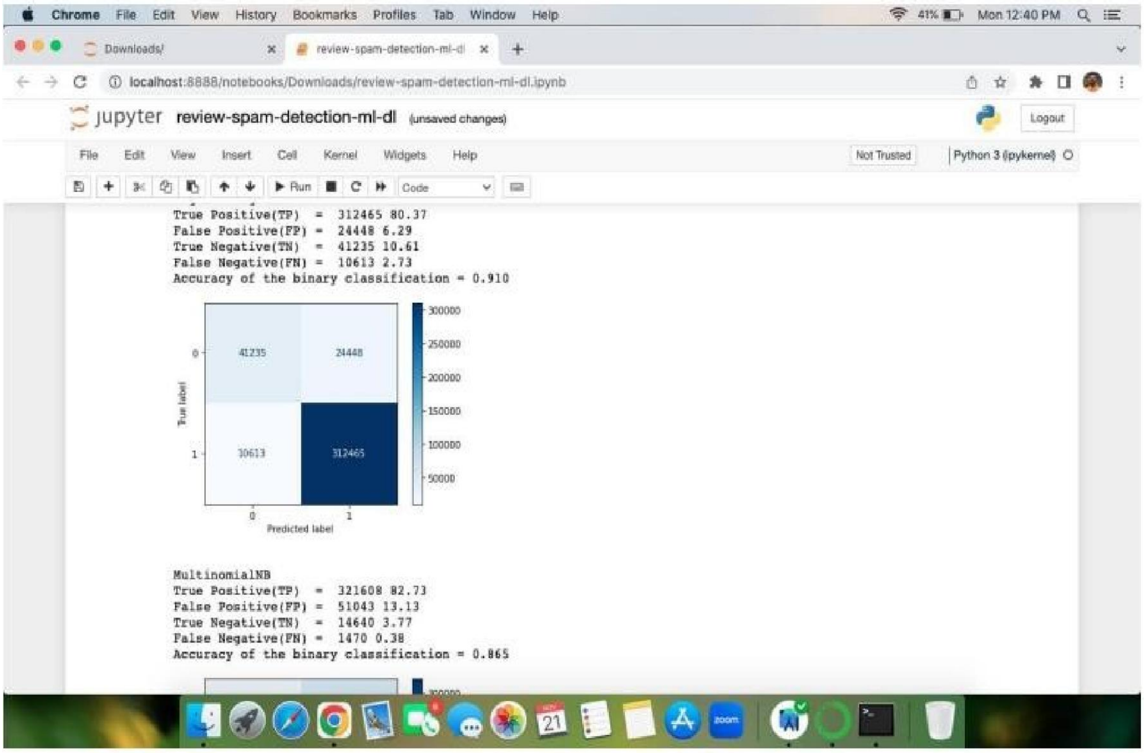
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

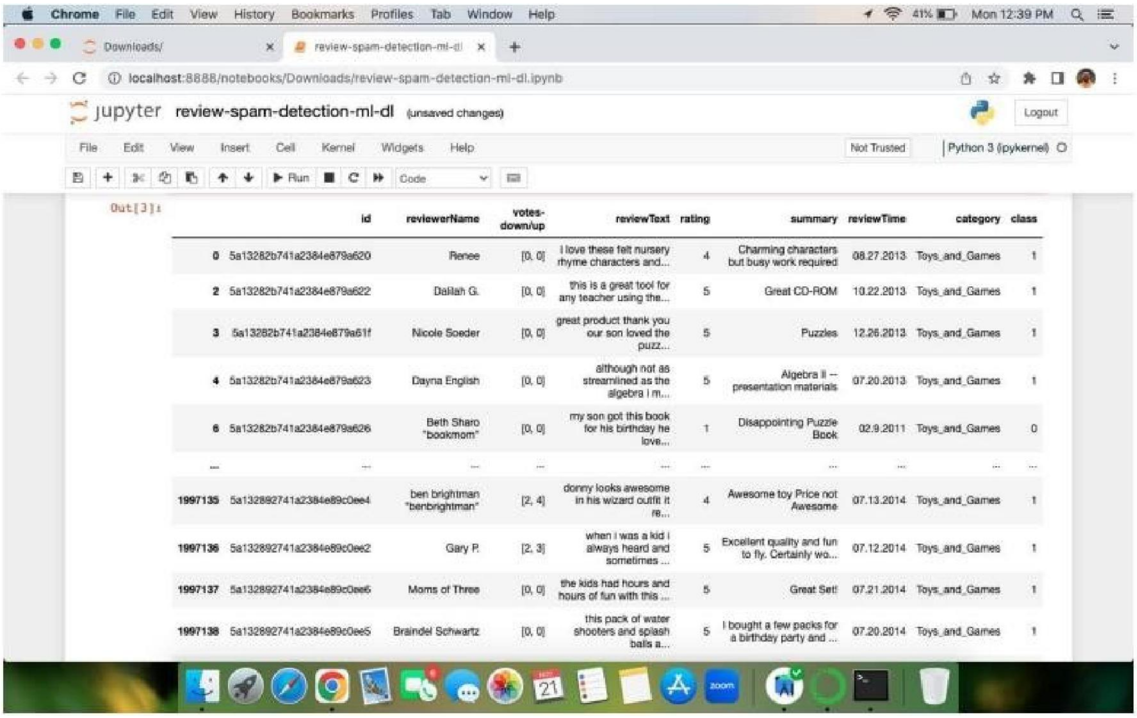
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 11, May 2023

Review-Spam detection:





IV. CONCLUSION

This paper presented an extensive survey of the most notable works to date on machine learning-based fake review detection. The spam review detection using ML is designed for filtering the fake reviews. People write unworthy positive reviews about products to promote them. In some cases malicious negative reviews to other (competitive) products are given in order to damage their reputation. Some of these consists of non-reviews (e.g., ads and promotions) which contain no opinions about the product We detecting the reviews that are not genuine or which are used to deviate the consumers opinion in a certain direction becomes even more difficult. Opinion spamming or fake review detection is thus significant problem for ecommerce sites and other service providers as the consumer these days rely highly on such opinions or reviews.

REFERENCES

[1]. Jindal, Nitin, and Bing Liu. "Opinion spam and analysis." Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008.

[2]. Lim, Ee-Peng, et al. "Detecting product review spammers using rating behaviors." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[3]. Algur, Siddu P., et al. "Conceptual level similarity measure based review spam detection." Signal and Image Processing (ICSIP), 2010 International Conference on. IEEE, 2010.

[4]. Feng, Song, et al. "Distributional Footprints of Deceptive Product Reviews." ICWSM12 (2012): 98-105.

[5]. Li, Wenbin, Ning Zhong, and Chunnian Liu. "Combining multiple email filters basedon multivariate statistical analysis." Foundations of Intelligent Systems. Springer Berlin Heidelberg, 2006. 729-738.

[6]. Liu, Bing, et al. "Partially supervised classi_cation of text documents." ICML. Vol. 2.2002.

[7]. Karimpour, Jaber, Ali A. Noroozi, and Somayeh Alizadeh. "Web Spam Detection by Learning from Small Labeled Samples." International Journal of Computer Applications 50.21, 2012.

[8]. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.

[9]. Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 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, HLT, 2011.

[10]. Duyu Tang, Bing Qin, and Ting Liu, Document modeling with gated recurrent neural network for sentiment classification. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 14221432, Lisbon, Portugal, 17-21 September, 2015.