International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

DETECTING FAKE REVIEWS USING MULTI-DIMENSIONAL REPRESENTATION WITH FINE-GRAINED ASPECTS PLAN

T. JHANSI RENUKA[1], N.NAVEEN KUMAR[2]

[1]PG STUDENT,DEPARTMENT OF INFORMATION TECHNOLOGY,JNUTH,UNIVERSITY COLLEGE OF ENGINEERING,SCIENCE AND TECHNOLOGY

[2]ASSOCIATE PROFESSOR,DEPARTMENT OF INFORMATION TECHNOLOGY,JNUTH,UNIVERSITY COLLEGE OF ENGINEERING,SCIENCE AND TECHNOLOGY

ABSTRACT:

Online reviews play a vital role on E-commerce based websites. In general, customers always prefer to read reviews about the product before purchasing something. The reviews help the customer in decision-making. Positive reviews help the user to decide to buy or not also it increases an organization's business sales but negative reviews make a falling bad impression about the product on the customer's view and decrease the business sales of an organization. Negative reviews always, directly and indirectly, impact an organization socially and economically. Due to fake reviews, more than 45% of users discontinue their buying behavior and lose reliance on bands after discovering fake reviews of a product. Authenticity and reliability of network information have become a trending challenge. Several methods for fake review detection start with textual and behavioral features they are time taking and easily identified by fraudulent users. Fake reviews may not impact the entire online review system, but ultimately cause a loss of credibility. it indicates the importance of the identification of fake reviews online and provides users with more reliable information. Fake review detection was first introduced specifically as opinion spam detection by Jindal and Liu Due to the paramount research implications of this work, Nikhil Padhi was the associate editor coordinating the review of this manuscript and approving it for publication. In these recent years, fake review detection became a proposal. The early stage of the work was on manual design features in combination with machine learning methods likely semantic features of the text include the length of the review text its lexical features and its affective polarity Users' behavioral features include the number of positive or negative reviews that they publish and the frequency of these reviews Driven by prof- its, spammers are enhancing and disguising their schemes under the corresponding detection methods. In these years, along with the development of deep learning, several fake review detection methods

based on deep learning have been developed. Compared to feature-based methods, these methods have a greater ability to automatically capture semantic information implicit within the text without a manual design and have stronger domain adaptability and effectiveness.

INTRODUCTION

Fake news has been demonstrated to be problematic in multiple ways. It has been shown to have real influence on public perception and the ability to shape regional and national dialogue. It has harmed businesses and individuals and even resulted in death, when an individual responded to a hoax. It has caused some teenagers to reject the concept of media objectivity and many students can't reliably tell the difference between real and faked articles. It is even thought to have influenced the 2016 United States elections. Fake news can be spread deliberately by humans or indiscriminately by bot armies, with the latter giving a nefarious article significant reach. Not just articles are faked, in many cases fake, mislabeled or deceptive images are also used to maximize impact. Some contend that fake news is a "plague" on society's digital infrastructure. Many are working to combat it. Farajtabar, et al., for example, has proposed a system based on points, while Haigh, Haigh and Kozak have suggested the use of "peer-to-peer counter-propaganda.

Intentionally deceptive content presented under the guise of legitimate journalism (or 'fake news,' as it is commonly known) is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most fake news is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto mainstream media platforms such as traditional television and radio news. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as excessive use of unsubstantiated hyperbole and non-attributed quoted content. The results of a fake news identification study that documents the performance of a fake news classifier are presented and discussed in this paper.

In this paper, the research process, technical analysis, technical linguistics work, and classifier performance and results are presented. The paper concludes with a discussion of how the current system will evolve into an influence mining system. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as excessive use of unsubstantiated hyperbole and non-attributed quoted content. The results of a fake news identification study that documents the performance of a fake news classifier are presented and discussed in this paper.

LITERATURE SURVEY

Diagnosis of Thyroid disorder using Artificial Neural Networks. AUTHORS: Anupam Shukla, Ritu Tiwari, Prabhdeep Kaur, R.R. Janghel

A major problem in medical science is attaining the correct diagnosis of disease in precedence of its treatment. This paper presents the diagnosis of thyroid disorders using artificial neural networks (ANNs). The feed-forward neural network has been trained using three ANN algorithms; the Back propagation algorithm (BPA), the radial basis function (RBF) Networks and the learning vector quantization (LVQ) networks. The networks are simulated using MATLAB and their performance is assessed in terms of factors like accuracy of diagnosis and training time. The performance comparison helps to find out the best model for diagnosis of thyroid disorders.

A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis. AUTHORS: Kemal Polat,Seral Özşen,Salih Güneş

Proper interpretation of the thyroid gland functional data is an important issue in the diagnosis of thyroid disease. The primary role of the thyroid gland is to help regulation of the body's metabolism. Thyroid hormone produced by the thyroid gland provides this. Production of too little thyroid hormone (hypothyroidism) or production of too much thyroid hormone (hyperthyroidism) defines the type of thyroid disease. Artificial immune systems (AISs) are a new but effective branch of artificial intelligence. Among the systems proposed in Page | 5 this field so far, artificial immune recognition system (AIRS), which was proposed by A. Watkins, has shown an effective and intriguing performance on the problems it was applied. This study aims at diagnosing thyroid disease with a new hybrid machine learning method including this classification system. By hybridizing AIRS with a developed Fuzzy weighted pre-processing, a method is obtained to solve this diagnosis problem via classifying. The robustness of this method with regard to sampling variations is examined using a cross-validation method. We used thyroid disease dataset which is taken from UCI machine learning respiratory. We obtained a classification accuracy of 85%, which is the highest one reached so far. The classification accuracy was obtained via a 10- fold cross-validation.

A survey on applying machine learning techniques for management of diseases. AUTHORS: Enas M.F. El Houby

During the past years, the increase in scientific knowledge and the massive data production have caused an exponential growth in databases and repositories. Biomedical domain represents one of the rich data domains. An extensive amount of biomedical data is currently available, ranging from details of clinical symptoms to various types of biochemical data and outputs of imaging devices. Manually extracting biomedical patterns from data and transforming them into machineunderstandable knowledge is a difficult task because biomedical domain comprises huge, dynamic, and complicated knowledge. Data mining is capable of improving the quality of extracting biomedical patterns. In this research, an overview of the applications of data mining on the management of diseases is presented. The main focus is to investigate machine learning techniques (MLT) which are widely used Page | 6 to predict, prognose and treat important frequent diseases such as cancers, hepatitis and heart diseases. The techniques namely Artificial Neural Network, K-Nearest Neighbour, Decision Tree, and Associative Classification are illustrated and analyzed. This survey provides a general analysis of the current status of management of diseases using MLT. The achieved accuracy of the various applications ranged from 70% to 100% according to the disease, the solved problem, and the used data and the technique.

EXISTING SYSTEM

Since the review spam detection task was proposed, the early research, based on feature engineering and machine learning, has mainly focused on the analysis of user behavioural features, structural features, and text semantic features. After analyzing reviews and users on Amazon.com, Jindal and Liu classified spam reviews into three categories: untruthful opinions, reviews of brands only, and non-reviews such as advertisements. Additionally, they proposed a total of 36 text-centric, user-centric, and product-centric features that could be combined with logistic regression methods to identify spam reviews. Li et al combined semi-supervised machine learning methods to identify fake reviews based on multiple texts- and user-related features and analyzed the impact of each feature. Li et al identified the defectiveness of language usage between truthful and fake reviews. Wang et al. performed taens or customers always prefer to read reviews about the product before purchasing something viewed Classified them according to the SVM model.

PROPOSED SYSTEM

In this paper author is describing concept to detect fake news from social media or document corpus using Natural Language Processing and attribution supervised learning estimator. News documents or articles will be uploaded to application and then by using Natural Language Processing to extract quotes, verbs and name entity recognition (extracting organizations or person names) from documents to compute score, verbs, quotes and name entity also called as attribution. Using supervised learning estimator we will calculate score between sum of verbs, sum of name entity and sum of quotes divided by total sentence length. If score greater than 0 then news will be considered as REAL and if less than 0 then new will be consider as FAKE.

IMPLEMENTATION

Decision tree classifiers

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision-making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi, we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

Gradient boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

K-Nearest Neighbors (KNN)

Simple, but a very powerful classification algorithm

Classifies based on a similarity measure

Non-parametric

Lazy learning

Does not "learn" until the test example is given

Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

Training dataset consists of k-closest examples in feature space

Feature space means, space with categorization variables (non-metric variables)

Learning based on instances, and thus also works lazily because instance close to the input vector for test or prediction may take time to occur in the training dataset

Logistic regression Classifiers

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name multinomial logistic regression is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST
ISSN: 2457-0362

regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

To run the project double, click on the 'run.bat' file to get the below screen

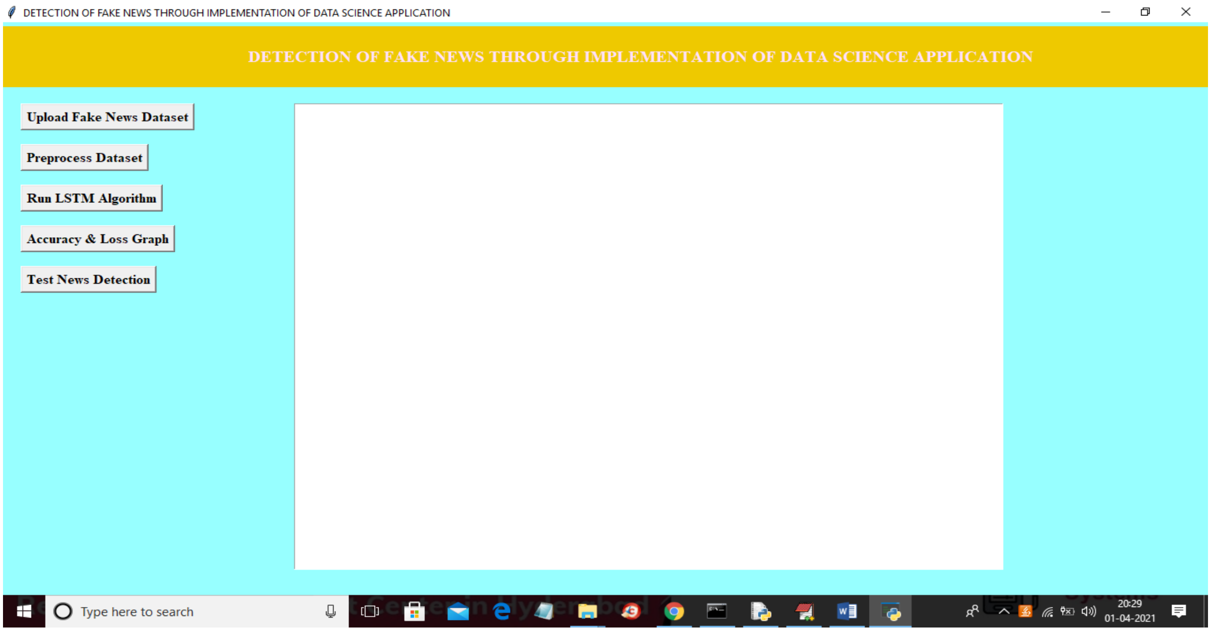


Fig 9.1. Upload Fake News Datasets

IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

In above screen click on 'Upload Fake News Dataset' button to upload dataset

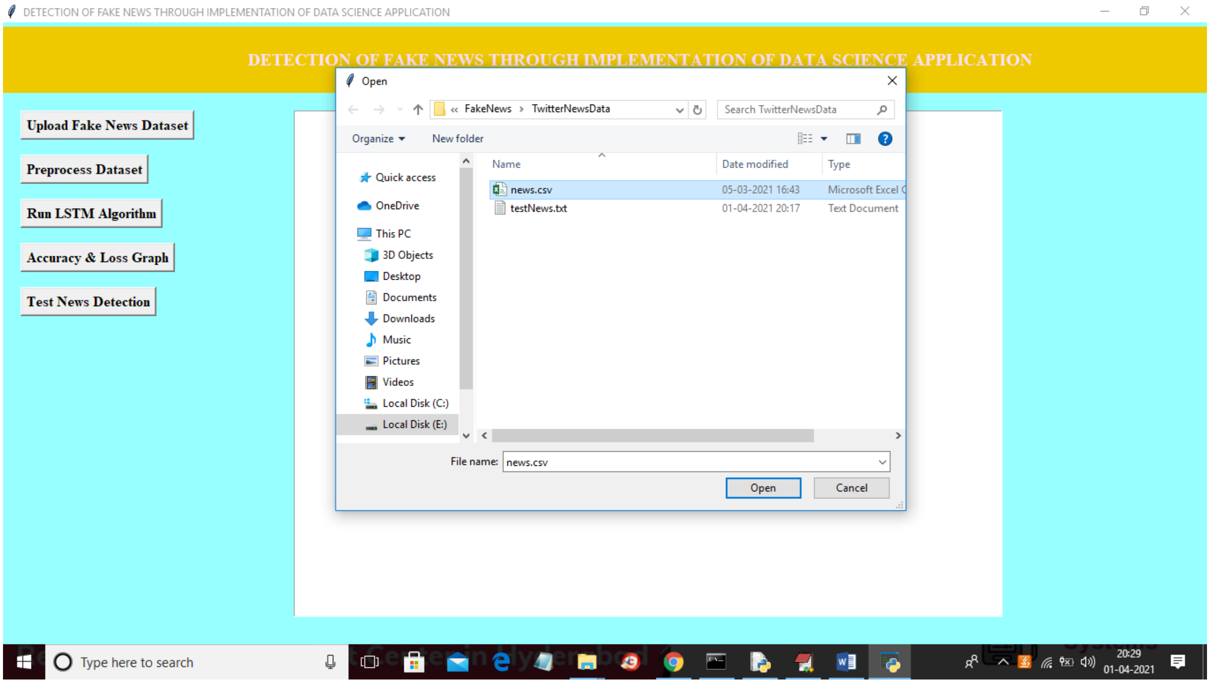


Fig 9.2. Open the button to load the dataset

In above screen selecting and uploading 'news.csv' file and then click on 'Open' button to load dataset and to get below screen

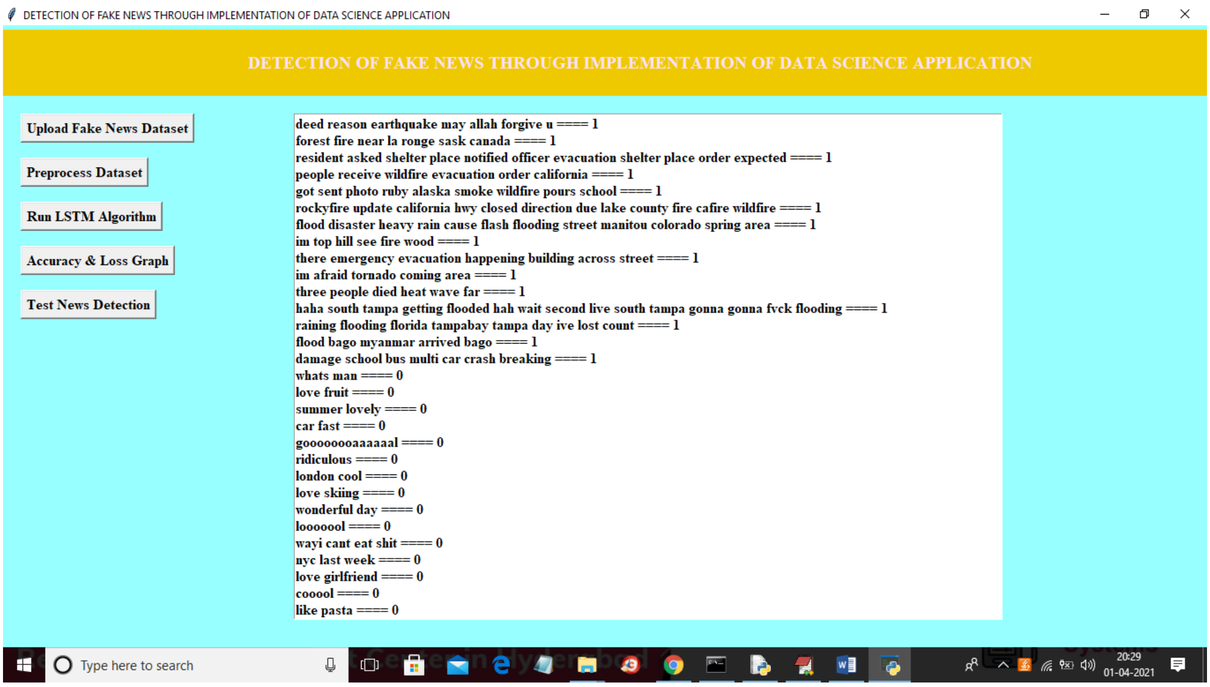


Fig 9.3. Text with the class label as 0 or 1

IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

In above screen dataset loaded and then in text area we can see all news text with the class label as 0 or 1 and now click on 'Pre-process Dataset & Apply NGram' button to convert above string data to numeric vector and to get below screen

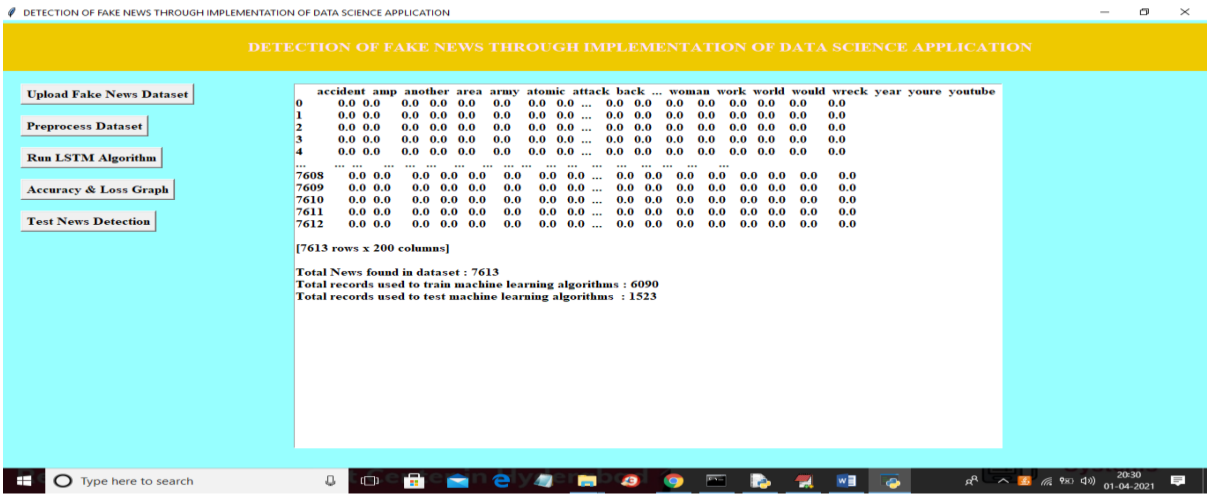


Fig 9.4. column header

In above screen all news words put in column header and if that word appears in any row then that rows column will be change with word count and if not appear then 0 will be put in column. In above screen showing some records from total 7612 news records and in bottom lines we can see dataset contains total 7613 records and then application using 80% (6090 news records) for training and then using 20% (1523 news records) for testing and now dataset is ready with numeric record and now click on 'Run LSTM Algorithm' button to train above dataset with LSTM and then build LSTM model and then calculate accuracy and error rate

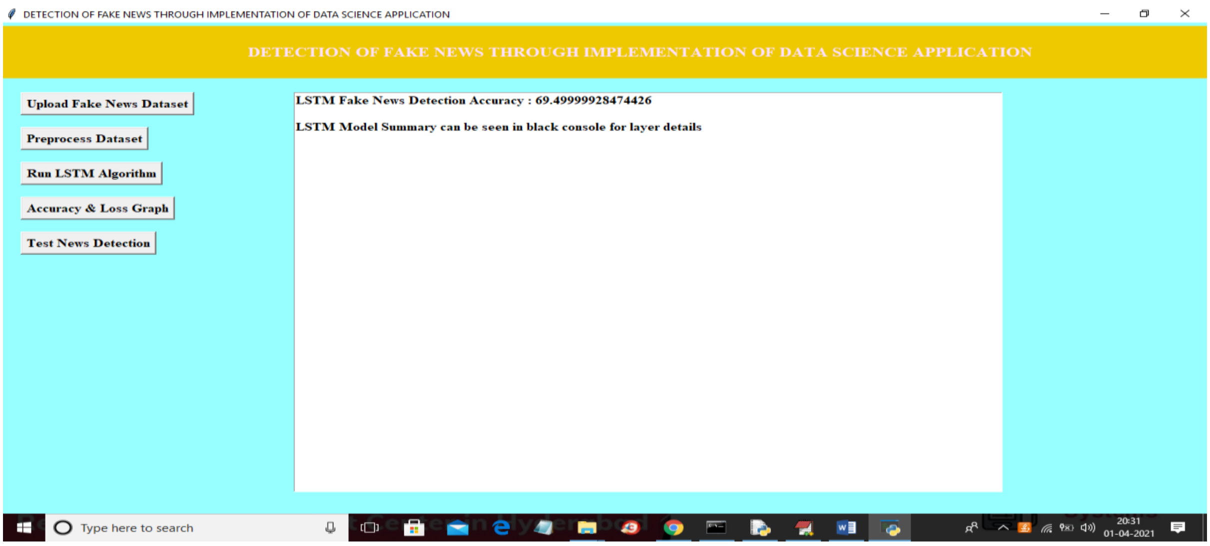


Fig 9.5. LSTM model

IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

In above screen LSTM model is generated and we got its prediction accuracy as 69.49% and we can see below console to see LSTM layer details

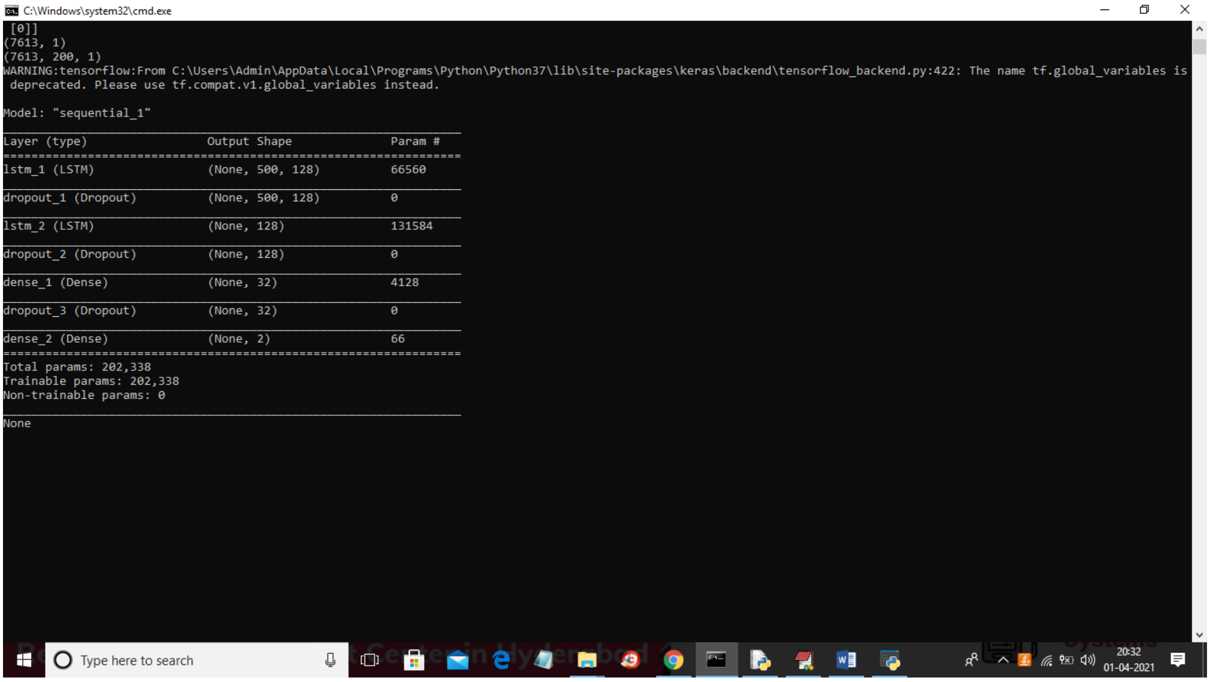


Fig 9.6 LSTM layers filter input data

In above screen different LSTM layers are created to filter input data to get efficient features for prediction. Now click on 'Accuracy & Loss Graph' button to get LSTM graph

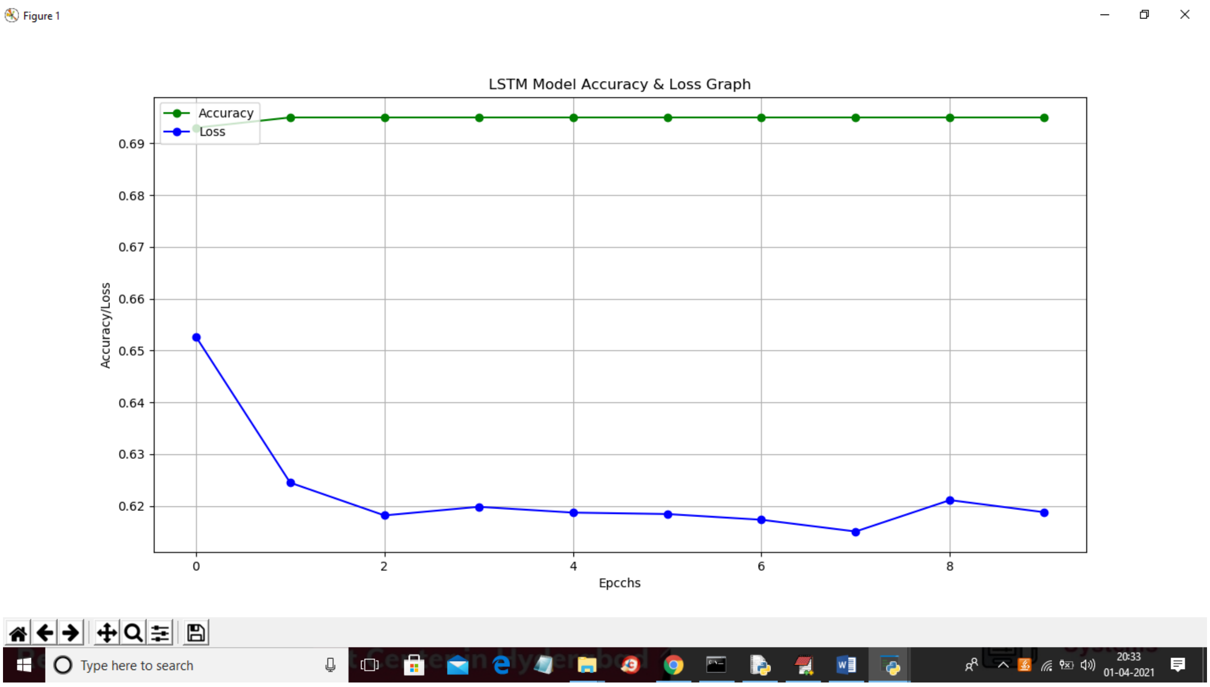


Fig 9.7 LSTM graph

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST
ISSN: 2457-0362

In above graph x-axis represents epoch/iterations and y-axis represents accuracy and loss value and green line represents accuracy and blue line represents loss value and at each increasing epoch loss values get decrease and accuracy reached to 70%. Now click on 'Test News Detection' button to upload some test news sentences and then application predict whether that news is genuine or fake. In below test news dataset, we can see only TEXT data no class label and LSTM will predict class label for that test news

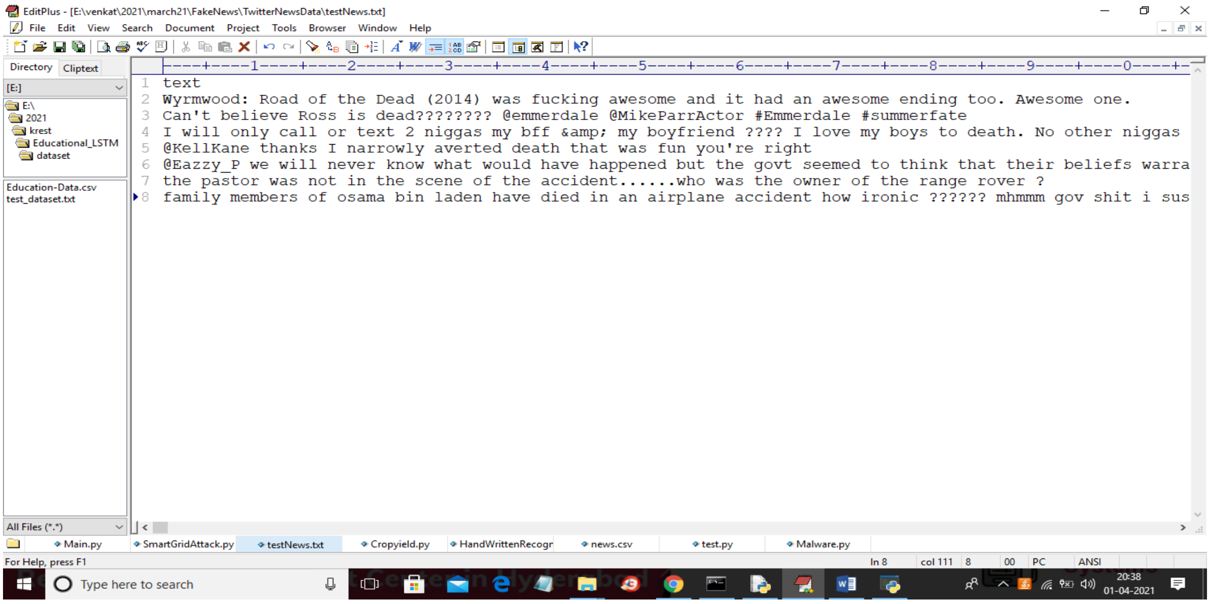
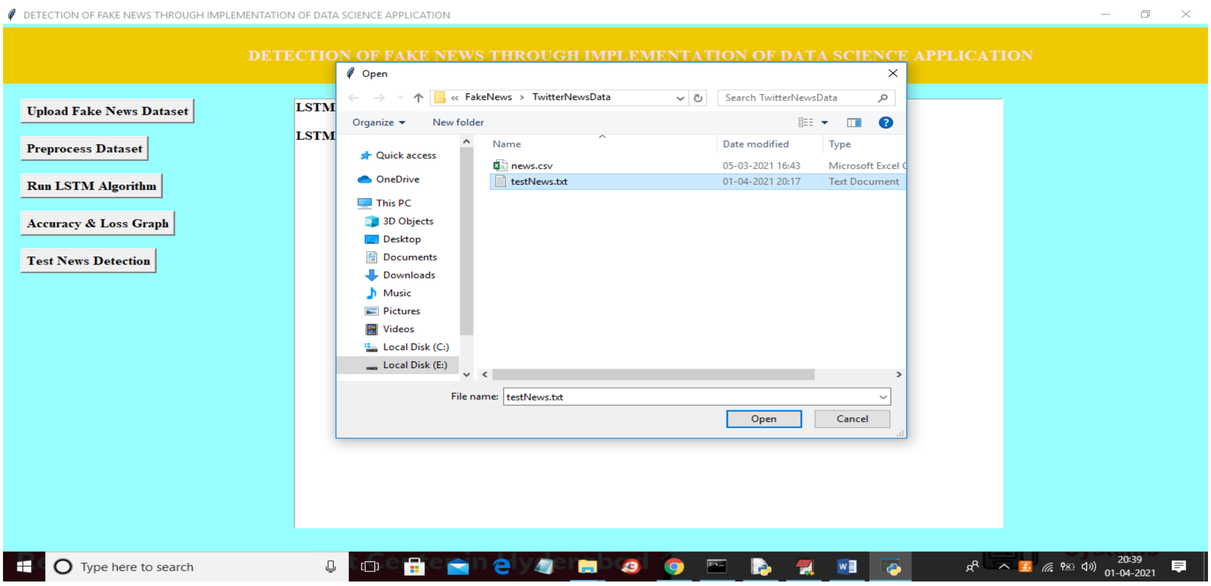


Fig 9.8 TEXT representation

In above screen in test news we have only one column which contains only news 'TEXT' and after applying above test news we will get prediction result



9.9 Open button to load data

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST
ISSN: 2457-0362

In above screen selecting and uploading ‘testNews.txt’ file and then click on ‘Open’ button to load data and to get below prediction result

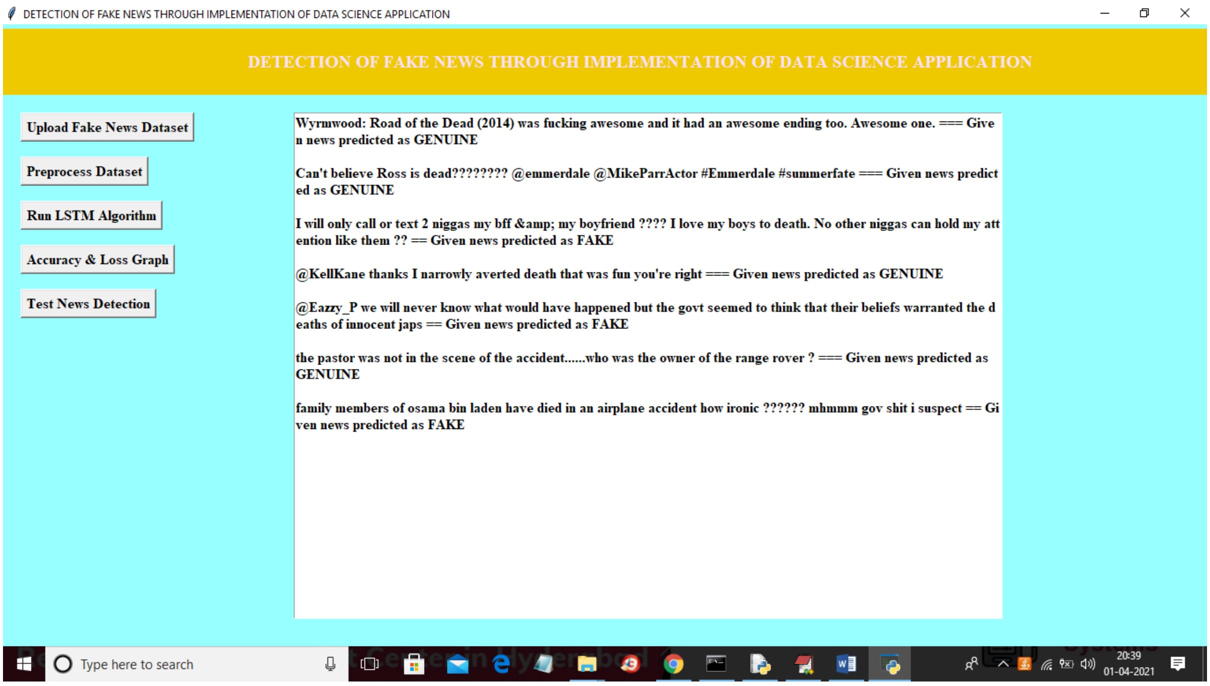


Fig 9.10 Representation of application for the class label

In above screen before dashed symbols we have news text and after dashed symbol application predict news as ‘FAKE or GENUINE’. After building model when we gave any news text then LSTM will check whether more words belongs to genuine or fake category and whatever category get more matching percentage then application will predict that class label.

CONCLUSION

This paper presented the results of a study that produced a limited fake news detection system. The work presented herein is novel in this topic domain in that it demonstrates the results of a full-spectrum research project that started with qualitative observations and resulted in a working quantitative model. The work presented in this paper is also promising, because it demonstrates a relatively effective level of machine learning classification for large fake news documents with only one extraction feature. Finally, additional research and work to identify and build additional fake news classification grammars is ongoing and should yield a more refined classification scheme for both fake news and direct quotes. In this study, we focused on the task of identifying spam reviews. After analyzing the reviews in the datasets, we propose a hypothesis that _ne-grained aspect information can be used as a new scheme for fake review detection and reconstructed the representation of reviews from four perspectives: users, products, reviews text, and _ne-grained aspects. We proposed a multilevel interactive attention neural network model with aspect plan; to optimize the model's objective function, we transformed the implicit relationship between users, reviews and products into a regularization term. To verify the effectiveness of the MIANA, we conducted extensive experiments on three public datasets. Our experiments showed that the classification effect has been significantly improved, that the MIANA outperforms the state-of-the-art methods for fake review detection tasks, and proved the effectiveness and feasibility of our proposed scheme. In this paper, the _ne-grained aspect terms are for restaurants and hotels. When it comes to cross-domain issues, you only need to further obtain _ne-grained aspects in the relevant domain. This is the current limitation of our proposed method, and it is also the content of our future research. Our further work includes: (a) validate the performance of our proposed method on cross-domain datasets, (b) build a joint model that can automatically extract _ne-grained aspects and identify fake reviews.

REFERANCES

[1] R. Filieri and F. McLeay, ``E-WOM and accommodation: An analysis of the factors that in_uence travelers' adoption of information from online reviews,'' J. Travel Res., vol. 53, no. 1, pp. 44_57, Jan. 2014.

[2] E. Kauffmann, J. Peral, D. Gil, A. Ferrández, R. Sellers, and H. Mora, ``A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making,'' Ind. Marketing Manage., vol. 90, pp. 523_537, Oct. 2020.

[3] N. Jindal and B. Liu, ``Review spam detection,'' in Proc. 16th Int. Conf. World Wide Web, 2007, pp. 1189_1190

[4] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, ``What yelp fake review _lter might be doing,'' in Proc. ICWSM, 2013, pp. 409_418.

[5] S. Rayana and L. Akoglu, ``Collective opinion spam detection: Bridging review networks and metadata,'' in Proc. 21th ACM

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2015, pp. 985_994.

[6] F. Li, M. Huang, Y. Yang, and X. Zhu, ``Learning to identify review spam," in Proc. IJCAI 22nd Int. Joint Conf. Artif. Intell., vol. 3, 2011, pp. 2488_2493.

[7] X. Hu, J. Tang, H. Gao, and H. Liu, ``Social spammer detection with sentiment information," in Proc. IEEE Int. Conf. Data Mining, Dec. 2014, pp. 180_189.

[8] S. Kc and A. Mukherjee, ``On the temporal dynamics of opinion spamming: Case studies on Yelp," in Proc. 25th Int. Conf. World Wide Web, Apr. 2016, pp. 369_379.

[9] Y. Ren and Y. Zhang, ``Deceptive opinion spam detection using neural network," in Proc. 26th Int. Conf. Computer. Linguistics, Tech. Papers COLING, Dec. 2016, pp. 140_150.

[10] X. Wang, K. Liu, and J. Zhao, ``Handling cold-start problem in review spam detection by jointly embedding texts and behaviors," in Proc. 55th Annu. Meeting Assoc. Computer. Linguistics (Long Papers), vol. 1, 2017, pp. 366_376. [Online]. Available: https://www.aclweb.org/anthology/P17 - 1034.pdf