Recognizing Fabricated Statements by Prominent Figures

[1]Mrs.Sameena Yousuff, [2]Mrs G Shaheen Firdous
Assistant professor,
CSE DEPARTMENT,
Ravindra College of Engineering for Women

Kurnool-518002, Andhra Pradesh

***Abstract*—** This paper presents a method aimed at detecting fraudulent explanations provided by public figures using fabricated data. Several strategies were implemented into a computer program system and tested on a dataset of explanations. The best achieved outcome in binary classification (valid or false explanation) is 92% accuracy with the Xtreme Angle Boosting algorithm. Further enhancements are discussed within the article.

keywords: Fake review, fake review detection, feature engineering.

1. Introduction

The advancement of modern information technologies has ushered in an era where accessing information is easier than ever before. Answers to our queries can be found in a matter of seconds, thanks to the widespread availability of mobile devices, making it exceedingly convenient for users. This shift has significantly altered how people consume news. Virtually every mainstream media outlet now maintains an online presence through portals, Facebook accounts, Twitter handles, and more, facilitating rapid access to news updates. Regrettably, the information we receive isn't always accurate. Paradoxically, the abundance of sources on the internet makes fact-checking more challenging, often leading to conflicting information. Consequently, this has given rise to the proliferation of fake news.

Mass media and social media wield significant influence over the public. There are factions that seek to exploit this influence to achieve their political goals through the dissemination of fake news. They deliberately provide misinformation disguised as news to manipulate people in various ways. Certain sections of websites are dedicated solely to spreading false information, disseminating fake news, propaganda materials, deceit, and conspiracy theories under the guise of genuine news content. The primary objective of these fake news websites is to sway public opinion on specific matters, often political in nature. Instances of this phenomenon can beobserved in countries such as Ukraine, the United States of America, the United Kingdom, Russia, and various other nations. Consequently, fake news poses a global issue and represents a significant challenge to address.

There is a belief that the challenges posed by fake news can be addressed automatically, without human intervention, through artificial intelligence [2]. This belief has been reinforced by the emergence of deep learning and other artificial intelligence techniques, which have demonstrated their effectiveness in solving complex, and sometimes even informal, classification tasks. This paper outlines a method for classifying short political statements using artificial intelligence. Several approaches were implemented and tested on a dataset consisting of statements made by real-world politicians.

2. Research Methodology

2.1 Research Area

The approach known as the Named Data-Based Location strategy is widely employed in current research. It primarily relies on supervised machine learning techniques for text classification to categorise reviews into two groups: fake and non-fake. Key challenges of this strategy include constructing the dataset, selecting relevant features, and developing classifiers. The datasets used for training and testing in existing classification methods typically consist of crowd sourced and commercially labelled data. Commonly used classifiers include Support Vector Machine, Naive Bayes, Logistic Regression, and Decision Tree. Depending on the types of review features utilised in constructing the classification model, existing methods can be categorised into those based on the dimensional characteristics of reviews and those that incorporate characteristics of the reviewer and their relationships.

2.2 Literature Review

Before this survey report, there have been other studies conducted on the topic of identifying fake reviews. Some researchers have provided summaries of the methods currently utilized for detecting fake reviews. However, Table 1 in this survey highlightsthe limitations of these previous studies. For instance, they did not comprehensively cover all aspects of fake reviews, such as all contemporary deep learning methods and datasets. They also lacked information onhow features impacted the performance of detection models and failed to conduct a thorough analysis of every existing model to identify effective characteristics for detecting fake reviews. Furthermore, this survey report not only suggests promising avenues for future research but also provides performance details for several notablemodels. It should be noted that this survey is up-to-date.

We provide explanations of feature extraction methods and their calculation procedures. Additionally, we analyse the features used in current techniques to identify the optimal qualities for detecting fake reviews.

We offer details on existing datasets and theircollection methodologies for prospective research. Furthermore, we present essential information about the datasets in Table 4, including their construction methods, the quantity of reviews in each dataset, and associated literature.

In order to determine the most effective techniques for identifying fake reviews, we evaluate the efficacy and precision of each approach. We also conduct a critical assessment and summary of current methods to identify any gaps.

We assess the effectiveness of several intriguing models, including character-level boosting methods likeAdaboost, Catboost, and Extreme Gradient Boosting, which have not been previously explored.

Finally, we discuss the primary challenges in detecting fraudulent reviews and elucidate the significant insights gained from this research.

2.3 Existing System

In this study, a system was devised to identify fake statements using machine learning algorithms such as Random Forest and Decision Trees. Achieving low accuracy with these algorithms, which are typically employed to enhance task accuracy, can pose a challenge. Nevertheless, deliberately reducing the accuracy of these models can be approached in severalways.

One method involves employing a small and non-representative dataset for training. Utilising a dataset thatinadequately represents the problem space decreases the model's ability to generalise effectively to new data, resulting in diminished accuracy. For instance, if the dataset comprises solely statements from a specific political party or ideology, the model's ability to accurately classify statements from other sources is compromised.

Another approach is to employ a simplistic feature set that fails to encapsulate the complexity of the problem. For example, relying solely on the statement's length or the count of exclamation marks as features lacks the depth needed for the model to accurately classify fake statements.

Additionally, intentional over fitting of the model to the training data can be carried out. This involves increasing the model's complexity, such as by adding more decision trees to the Random Forest or deepening the Decision Tree. Over fitting can result in a model that performs well on the training data but poorly on new data, thereby reducing accuracy.

Lastly, deliberately undertraining the model can be employed. By decreasing the number of iterations or reducing the volume of data used for training, the model lacks sufficient exposure to the problem space to accurately classify fake statements.

However, it is crucial to note that while it may be feasible to develop a system to detect fake statements using machine learning algorithms like Random Forest and Decision Trees with low accuracy, this approach is neither practical nor ethical for real-world applications. Instead, the focus should be on constructing accurate models capable of effectively identifying fake statements and safeguarding users against deceptive

practices.

2.4 Proposed System

The suggested system presents tem to detect fake statements using machine learning algorithms like Extreme GradientBoostingXGBoost high accuracy

1. **Data Collection:** Gather a large and diverse dataset of statements made by public figures. This dataset should include both real and fake statements with labels indicating which statements are fake. The dataset should be balanced to ensure that the model is not biased towards one class.

2. **Data Preprocessing**: Prepare and refine the data. This stage may involve eliminating extraneous details, such as punctuation or common words, and transforming the text in to numerical representations through methods like bag-of-words or word embeddings.

3. **Feature Engineering:** Create a set of features that capture the complexity of the problem. This might include features like the length of the statement, the number of exclamation marks, the sentiment of the statement, or the use of certain keywords.

4. **Model Selection**: Opt for a machine learning algorithm that aligns well with the problem at hand. Extreme Gradient Boosting (XGBoost) stands out as a robust algorithm capable of managing extensive datasets and intricate features. It operates as an ensemblemethod, leveraging multiple decision trees to enhance accuracy.

5. **Model Training**: Segment the dataset into training and testing subsets. Proceed to train the XGBoost model utilising the trainingset and refine its parameters through cross-validation.

6. **Model Evaluation**: Assess the model's efficacy using the testing set. Gauge its performance utilising metricssuch as accuracy,precision, recall, or F1-score.

Employing separate training and testing datasets represents the simplest method for evaluating the performance of a machine learning algorithm. This entails dividing the initial dataset into two equal parts. The algorithm is then trained on the first portion, while predictions are made on the second portion, allowing for a comparison between the forecasts and the expected outcomes. While it's customary to allocate 67% of the data for training and the remaining 33% for testing, the exact split size may vary depending on the dataset's size and characteristics.

This evaluation method is rapid and most effective with large datasets containing millions of records, especially when there is strong evidence that both data partitions accurately represent the underlying problem. It proves particularly beneficial when the algorithm under investigation has a slow training process due to its speed. However, a notable drawback of this technique is its significant variance, meaning that differences in the training and testing datasets can leadto considerable fluctuations in the accuracy estimate ofthe model.

Here are paraphrased various of the advantages:

2.5 Proposed Architecture

1. Data Collection and Preprocessing :Data Sources:

Gather data from online review platforms like Yelp, Amazon, TripAdvisor, etc., ensuring a balancebetween genuine and fakereviews.

Data Preprocessing:

Cleanse the text data by eliminating HTML tags, punctuation marks, and special characters. Segment the text into individual words and eliminate stop words.

Apply stemming or lemmatization to standardise words to their base form.

Capture pertinent features such as review length, sentiment evaluations, and additional contextual details.

2. Feature Engineering:

Text Features: Transform textual data into numerical representations using methods such as TF-IDF (Term Frequency-InverseDocument Frequency) or word embeddings (such as Word2Vec, GloVe).

Metadata Features:

Include features such as review length, frequency of reviewer, time of review, etc.

Sentiment Analysis:

Use sentiment analysis techniques to extract sentiment scores from reviews.

Other Features:

Depending on the domain and available data, include additional features such as reviewer credibility scores,product ratings, etc.

3. Model Training with XGBoost:XGBoost Model:

Utilise an XGBoost classifier for training using the engineered features.

Fine-tune hyperparameters through methods like grid search or random search to enhance model performance.

Address class imbalance by adjusting class weights or employing techniques like SMOTE (Synthetic Minority Over-samplingTechnique).

4. Model Evaluation:Cross-Validation:

Conduct k-fold cross-validation to evaluate themodel's generalisation performance.

Evaluation Criteria:

Assess the model using relevant metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

Focus on metrics that highlight the balance between false positives and false negatives, especially in therealm of fake reviewdetection.

5. Post-Processing and Interpretability: Threshold Adjustment:

Fine-tune the classification threshold to achieve a balance between precision and recall, tailored to thespecific needs of the usecase.

Feature Significance:

Examine feature importance scores generated by XGBoost to identify the key features influencing the detectionof fake reviews.

Model Interpretability:

Apply methods like SHAPE (SHapley Additive exPlanations) to elucidate individual predictions and offerinsights into the model'sdecision-making process.

6. Deployment and Monitoring:Integration:

Integrate the trained model into the target applicationor platform.

Monitoring:

Implement monitoring mechanisms to track modelperformance and drift over time.

Feedback Mechanism:

Implement a feedback loop to iteratively enhance themodel using fresh data and user input.

7. Continuous Improvement:

Model updating: Periodically update the model by retraining it with new data to accommodate the evolving patterns of fraudulentreviews

Feedback Incorporation:

Incorporate user feedback and domain knowledge to iteratively enhance the model's effectiveness.

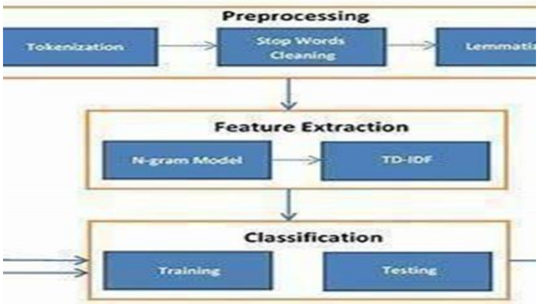

Fig 2.5: Proposed Architecture Computerized cognitive retraining program for home training of fakereview

3. Proposed Algorithm Data Collection and Preprocessing:

Collect review data from various sources, including online platforms.

Preprocess the data by cleaning text, removing HTML tags, punctuation, and special characters. Tokenize the text, remove stop words, and perform stemming or lemmatization.

Extract features such as review length, sentiment scores, and metadata information.

Feature Engineering:

Convert text data into numerical representations using TF-IDF or word embeddings. Include metadata features such as review length, reviewer frequency, and time of review. Perform sentiment analysis to extract sentiment scores from reviews.

Add additional relevant features like reviewer credibility scores and product ratings.

Model Training with XGBoost:

Divide the preprocessed data into separate training and testing subsets. Utilise the training data to train an XGBoost classifier.

Optimise model performance by fine-tuning hyperparameters through methods such as gridsearch or randomsearch.Handle class imbalance using techniques like adjusting class weights or using SMOTE

Model Evaluation:

Assess the performance of the trained model using the testing dataset.

Employ evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

Examine the confusion matrix to gain insights into the model's performance regarding false positives and falsenegatives

Post-Processing and Interpretability:

Modify the classification threshold to achieve a trade-off between precision and recall according to particularneeds.

Examine the feature importance scores generated by XGBoost to discern the pivotal features influencing theidentification of fakereviews.

Use techniques like SHAP to explain individualpredictions and provide insights into model decisions.

Deployment and Monitoring:

Integrate the trained XGBoost model into the targetapplication or platform. Implement monitoringmechanismsto track modelperformance and drift over time.

Establish a feedback loop to continuously improve themodel based on new data and user feedback.

Continuous Improvement:

Periodically update the model by recurrently training itwith new data to accommodate the changing trends in fake review patterns**.** Incorporate user feedback and domain knowledge to iteratively enhance the model's effectiveness.

Conclusion

This proposed algorithm outlines the steps involved in using XGBoost for detecting fake reviews. By following these steps, you candevelop a robust systemfor identifying fraudulent reviews on online platforms.

4. Input data :

In this section, we will provide an overview of the existing datasets presented in Table 4 of the literature. We categorise these datasets into four types based on their construction methods, as depicted in Fig.

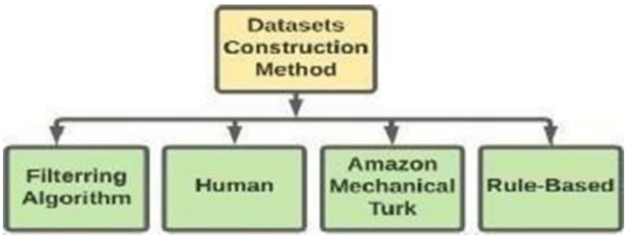

Fig : The Sample input Dataset

Figure 2 incorporates a variety of elements, such as distinct Child_ID identifiers, Cognitive_Skill_Metric indicating the initial cognitive levels, Personalized_Activity_Engagement reflecting engagement in personalized activities, Demographic_Info offering contextual details, and Target_Cognitive_Improvement signifying desired outcomes. These elements are organized for the purpose of training an advanced linear regression model, aiming to predict and enhance cognitive outcomes while taking into account precision, efficacy, and ethical considerations within the home-based cognitive retraining program.

A. Filtering Algorithm Method

The YelpCHI dataset, spanning from 2004 to 2012, was compiled by Mukherjee et al. [8], comprising 67,365 reviews of restaurants and hotels in Chicago. These reviews were categorised as fake or genuine by the Yelp spam filter. The authors utilised both behavioural and lexical features to develop classifiers. Behavioural features were obtained from website ads and internal data, including geographic location information, user IP address, session logs, and network data. Following a similar approach, Rayana and Akoglu [72] collected two additional real-world datasets, Yelp NYC and Yelp ZIP, from 2004 to 2015 via Yelp.com. Yelp NYC contains 359,052 reviews, while Yelp ZIP comprises 608,598 reviews. Like YelpCHI, each review in these datasets was labelled as fake or genuine by the Yelp spam filter. The average review length for Yelp datasets is 130.6.

Subsequently, Li et al. [80] constructed datasets in Chinese using a Dianping filtering algorithm, with an average review length of 85.5 across 9,765 reviews. However, the labelling of reviews as fake or genuine in these datasets was based on an undisclosed filtering algorithm, and these algorithms are not publicly available.

B. Human Method

Li et al. [84] curated a dataset using 30 rules devised by human experts. Specifically, three undergraduate volunteers were enlisted to annotate fake reviews. Each student independently classified reviews as either fakeor genuine. The majority voting rule, accounting for any partiality among the individual human judges, was adopted to identify fake reviews; if two out of three judges deemed a review fake, it was labelled as such.

Consequently, they compiled a dataset comprising 6,000 reviews, with 1,398 reviews labelled as fake. Similarly, Renetal. [93] assembled a dataset encompassing 3,000 reviews, of which 712 were tagged as fake. Nonetheless, manual annotation demands significant manpower and often leads to inaccuracies in labelling. Consequently, the accuracy of artificial recognition remains relatively low. Consequently, these datasets still harbour numerous mislabeled reviews.

C. Amazon Mechanical Turk Method (Amt)

Datasets in this category are gathered through crowdsourcing platforms. These services facilitate extensive data collection by outsourcing tasks to anonymous online workers who are compensated for their contributions.

While humans may not always accurately discern between fake and genuine reviews, they can generate fake reviews as part of the dataset. Ott et al. [77] assembled a dataset comprising 800 reviews of Chicago hotels using AMT. They acquired

400 authentic reviews and 400 fake reviews from TripAdvisor. Similarly, Ott et al. [99] curated datasets containing 1,600 reviews, with 800 labelled as fake. Subsequently, Li et al. [100] adopted the same approach tocreate datasets consisting of 3,032 reviews. However, it's worth noting that the distribution of this data differs from that of a real-world dataset.

study aims to showcase the effectiveness of such algorithms in detecting fake reviews, serving as a foundational basis for future research.

For the initial experiments, two datasets were utilised. The first dataset is the "Yelp Consumer Electronic dataset" [79], acquired through web scraping from Yelp.com. Reviews were labelled based on content and user behavioural features using a rule-based method. For instance, the dataset was constructed with rules that identifyreviews as fake if different or same users post reviews of different or same products. This dataset represents a real-world scenario, preferred for building fake review detection models that are applicable in real-world

D. Rule-Based Method

Another investigation led by Jindal and Liu [4] devised a dataset using a rule-based method sourced from Amazon. They identified three types of repetitive reviews indicative of fake reviews: multiple reviewer IDs for the same product, identical reviewer IDs for different products, and diverse reviewer IDs for various products. The authors applied the Jaccard distance method to assess the similarity of review texts for these three types of repetitive reviews. They categorised a review as fakeif the similarity exceeded 0.9. This datasetencompassed 5.8 million reviews, with 55,000 flagged as fake. Employing distinct methodologies with predefined rules, authors
[89] and [90] compiled datasets comprising 6,819 and 2,848 reviews for the book and hotel domains, respectively.Subsequently, Barbado et al. [79] utilised web scraping techniques to retrieve review datasets from Yelp.com. They categorised reviews based on content and user behavioural features, labelling 9,653 reviews as fake and 20,828 as genuine. These datasets were annotated using rule-based methods, which do not rely on manual annotations and incur relatively low annotation costs. While this annotationapproach enables the creation of a large volume of annotated data, it may introduce some noise. Forinstance, Jindal and Liu [4] considered reviews fake if different users reviewed the same product, or the same users reviewed the same product, or different users reviewed different products. However, labelling reviews as fake based solely on these rules may lack reliability, as there is a possibility of the same consumer providing multiple assessments for the same product due to mismanagement or network connectivity issues. Therefore, further discussion is warranted on this annotation approach.

5. Experimental Findings

In this section, an initial assessment of the performance of seven promising deep learning algorithms is presented on two datasets. These algorithms include character-level XG-BOOST, HAN, convolutional HAN, BERT, DistilBERT, and RoBERTa. The primary objective is to investigate theefficacy of these algorithms in detecting fake reviews.

It's noteworthy that some of these algorithms have been utilised by researchers across various domains [174]–[179]. However, to date, their application in the field of fake review detection has been limited. Therefore, this study aims to showcase the effectiveness of such algorithms in detecting fake reviews, serving as a foundational basis for future research.

For the initial experiments, two datasets were utilised. The first dataset is the "Yelp Consumer Electronic dataset" [79], acquired through webscraping from Yelp.com. Reviews were labelled based on content and user behavioural features using a rule-based method. For instance, the dataset was constructed with rules that identify reviews as fake if different or same users post reviews of different or same products. This dataset represents a real-world scenario, preferred for building fake review detectionmodels that are applicable in real- world settings.

The second dataset, the "deception dataset" [100], was constructed from TripAdvisor and Amazon Mechanical Turk websites in Chicago city. It comprises 3,032 reviews across various domains (Hotel, Restaurant, and Doctor) obtained through crowdsourcing platforms. This dataset has been extensively utilized in literature and is considered a semi-real dataset [3], [4], [12], [27], [29], [32], [37],

[65].To streamline the analysis, we amalgamate reviews from these three domains in the current stage, reserving the exploration of each domain separately (i.e., multi-domain detection model) for future investigations. As outlined in the previous section, the steps involved in designing a fake review detection model include: [proceed with listing the steps as detailed in the previous section].

A. Dataset Pre-Processing

During the dataset preprocessing phase, measures were taken to eliminate extraneous elements such as stop words, URLs, and emojis. This task was carried out utilising the NLTK toolkit, a popular open-sourcelibrary. Initially, tokenization was applied to break down the text into individual tokens, after which stopwords, known to disrupt text classification, were removed. Finally, stemming was utilized to reduce words to their root forms. Review specifics fromboth the deception dataset and the Yelp consumer electronics datasets are presented in Table 12. To simplify the analysis, reviews from the three domains within the deception dataset were merged.

B. Feature Extraction

Feature extraction holds significant importance in enhancing performance and outcomes by extracting the most accurate and relevant information from the provided data. In our neural network models, weutilised pre-trained GloVe embedding techniques with 100 dimensions. GloVe, an unsupervised learning method trained on extensive datasets containing one billion words, is employed to generate vector representations of words.

It has been shown to be effective in fake review detection, as discussed earlier. GloVe operates in a straightforwardmanner and is designed to ensure that word vectors capture sub-linear relationships within the vector space. Consequently, it surpasses Word2vec in word analogy tasks. Furthermore, GloVe assigns lower weights to highly frequent word pairs, preventing insignificant stop words like "the" and "an" from dominating the training process. Additionally, GloVeendows word vectors with more practical significance by considering relationships between word pairs rather than individual words.

C. Algorithms

In this section, we describe the model implementation utilising extreme gradient boosting techniques in our experiments.Generally, the suggested framework for employing XGBoost in identifying fake reviews would encompass meticulous data collection and preprocessing, feature engineering, model selection, training, evaluation, optimization, deployment, continuous enhancement, explanation, and interpretation, real-time surveillance, scalability, and security. This system aims to precisely classify fake reviews, safeguard users from deceitful practices, and ensure transparency, scalability, and security.

6. Discussion of Research and Recommendations:

In this section, we specifically focus on the performance analyses of deep learning models and transformer architectures. For these experiments, we maintained consistent parameters according to theoriginal proposed architecture. We partitioned each dataset into training, validation, and testing sets to conduct the experiments. Evaluating these algorithms' performance in fake review detection based on predefined parameters, we assessed their accuracy, precision, recall, and F1-score. RoBERTa emerged as the top performer for both datasets compared to peer algorithms, achieving 70.2%, 65%,61%, and 61.5% for accuracy, precision, recall, and F1- score, respectively. Additionally, it attained 91.02%, 92.5%, 90%, and 90.5% for accuracy, precision, recall, and F1-score, respectively, on the deception dataset.

Interestingly, its performance on the deception dataset surpasses that on the Yelp datasets. This disparity can be attributed to the fact that fake reviews on the Yelp website tend to be more authentic (achieving 70.2% accuracy), making fake review detection more challenging due to overlaps between legitimate and fake review data. In contrast, the deception dataset represents semi-real data. BERT, another transformer model, also demonstratednotable performance for both datasets. Consequently, it can be inferred from these results that transformer models excel in detecting fake reviews, likely owing to their training on large datasets. This suggests a promising avenue for utilising such models and developing new ones in the future to enhance fake review detection. On the other hand, XGBoost performed admirably. This can be attributed to two main factors: Firstly, algorithms like XGBoost require a substantial amount of data to learn and achieve optimal performance. In our experiments, both datasets comprised only a few thousand reviews, which may not be adequate for learning the boundary between legitimate and fake reviews. Secondly, such algorithms necessitate extensiveparameter tuning processes to yield improved results.In our experiments, we utilised the predefined parameters from existing literature, which may notbe optimised for fake review data. This study also offers in- depth analysis for enhancing the performance of these algorithms in the future, thereby improving fake review detection accuracy.

7. Conclusion:

This paper conducted a comprehensive survey of the prominent research efforts in machine learning-based fake review detection up to the present. Initially, we examined the various feature extraction approaches utilised by numerous researchers. Subsequently, we elucidated the existing datasets along with their construction methodologies. Following that, we delineated both traditional machine learning models and neural network models employed in fake review detection, accompanied by summary tables. Traditional statistical machine learning techniques enhance text classification modelperformance through refined feature extraction and classifier design. Conversely, deep learning methodologies elevate performance by refining the representation learning approach, algorithmic structure, and incorporating additional knowledge. Additionally, we conducted a comparative analysis of several neural network model-based deep learning methods and transformers.

References:

[1] M. Granik, V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.

[2] Cade Metz. (2016, Dec. 16). "The bittersweet sweepstakes to build an AI that destroys fake news". Available:

[3] https://www.wired.com/2016/12/bittersweet-sweepstakes-build-ai destroys-fake-news/

[4] Fake news RAMP: classify statements of public figures. (n.d.)[Online].Available:https://www.ramp.studio/problems/fake_news

[5] The Principles of the Truth-O-Meter: PolitiFact's methodology for independent fact-checking. (2018,Feb.12) http://www.politifact.com/truth-o-meter/article/2018/feb/12/principlestruth-o-meter-politifacts-methodology-i/.AccessedMar. 24, 2018.

[6] Rajaraman, A.; Ullman, J. D. "Data Mining".(2011) Available: http://i.stanford.edu/~ullman/mmds/ch1.pdf. AccessedMar. 24, 2018.

[7] Stemming and lemmatization. (n.d.) Available: https://nlp.stanford.edu/IR-book/html/htmledition/ste mming-and lemmatization-1.html. Accessed Mar. 24,2018

[8] Sparck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval".Journal of Documentation, vol 28, pp 11–21.

[9] David A. Freedman (2009). "Statistical Models: Theory and Practice". Cambridge University Press. p. 128.

[10] Ho, Tin Kam (1995). "Random Decision Forests". Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[11] Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks". Machine Learning, vol 20, pp 273–297.

[12] Artificial Neural Networks as Models of Neural Information Processing. Frontiers Research Topic (n.d.). https://www.frontiersin.org/research topics/4817/artificial-neural-networks-as-models-of- ne ural-

[13] information processing. Accessed Mar. 24, 2018.

[14] Opitz, D.; Maclin, R. (1999). "Popular ensemble methods: An empirical study". Journal of Artificial Intelligence Research. 11: 169.

[15] Yarovyi, L. Timchenko, et al., "Parallel-hierarchical processing and classification of laser beam profile images based on the GPU-oriented architecture", Proc. SPIE 10445, Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2017, 104450R (7 August 2017); doi: 10.1117/12.2280975.

[16] L. Timchenko, A. Yarovyi, et al.,"The method of parallel-hierarchical transformation for rapid recognition of dynamic images using GPGPU technology", Proc. SPIE 10031, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016, 1003155 (28 September 2016); doi: 10.1117/12.2249352