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

# THE EXAMINATION AND INVESTIGATION OF DECEPTIVE EMPLOYMENT POSTS BY UTILIZATION OF DIVERSE DATA MINING METHODOLOGIES

**Ramesh Babu Jonnalgadda[1], Dilip Solanki[2], Naga Manwita Reddy Vuribindi[3], Kommula Ajay Reddy[4], Amjan Shaik[5]**

*Associate Professor, Dept. of Computer Science Engineering, St. Peter's Engineering College, Telangana, India,*

*Professor, Dept. of Computer Science Engineering, St. Peter's Engineering College, Telangana, India,*
dilipsolanki13427@gmail.com

**ABSTRACT –** The widespread utilization of modern technology and social communication in recent years has resulted in a prevalent concern in today's society: the frequent advertisement of new job vacancies. As a result, the identification of deceptive job postings has become a significant focal point. Similar to other classification tasks, effectively predicting counterfeit job advertisements poses numerous challenges. This research proposes employing diverse data mining techniques and classification algorithms, such as KNN, decision tree, support vector machine, naive Bayes classifier, random forest classifier, multilayer perceptron, and deep neural network, to authenticate job listings. The experiments were conducted on the Employment Scam Aegean Dataset (EMSCAD), which encompasses 18,000 samples. Notably, the deep neural network classifier exhibited exceptional performance in this classification task by employing a three-layered dense architecture. The trained classifier achieved an impressive accuracy rate of approximately 98% (DNN) in accurately discerning fraudulent job postings.

## 1. INTRODUCTION

The current business and technological advancements have provided job seekers with abundant opportunities to explore diverse employment options. Job advertisements play a crucial role in helping job seekers assess their suitability, availability, qualifications, and experience for different positions. The growing impact of social media and the internet has revolutionized the hiring process. Effective recruitment heavily relies on well-executed advertising strategies, with social media playing a pivotal role in this aspect. The advent of social media and digital marketing has opened up new avenues for disseminating job information rapidly. Unfortunately, this increased accessibility has also led to a rise in fraudulent job postings, causing frustration among job seekers. Concerns about safeguarding personal, educational, and professional information deter individuals from engaging with new job openings. Establishing trust and credibility becomes a daunting challenge for legitimate job advertisements through social and digital media platforms. It is essential to leverage technology to enhance our lives and promote secure working environments rather than creating unsafe conditions. Accurate filtering of job postings to detect fake ones would be a significant advancement in streamlining the hiring process and alleviating the burden on job seekers. False job postings not only waste valuable time for job seekers but also hinder their ability to find desired positions. By implementing an automated system that predicts fake job postings, organizations can tackle the challenges in human resource management more effectively.

## 2. LITERATURE SURVEY

### 1.Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset

Fraudulent job postings create obstacles for job seekers in their pursuit of desired positions, resulting in a significant time waste. Introducing an automated system capable of predicting fake job postings opens up new possibilities for addressing challenges in the field of human resource management. These conventional business practices have introduced vulnerabilities that can jeopardize applicants' privacy and damage organizational reputations. Among various online fraud concerns, employment scams currently represent the most prevalent form of online recruitment fraud (ORF). Unfortunately, ORF has not received adequate attention compared to other forms of internet fraud, remaining largely overlooked until now.

### 2.Spotting fake reviews via collective positive-unlabeled learning

1. Online reviews have become an increasingly vital asset for decision-making and product creation. However, identifying opinion spam remains a prevalent issue for

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

review systems. Despite researchers employing supervised learning to explore the detection of fraudulent reviews over a prolonged period, the true nature of extensive datasets remains undisclosed. Furthermore, the majority of existing supervised learning techniques rely on counterfeit reviews that do not genuinely represent fake reviews. In this study, we present the initial examination of counterfeit review identification in the Chinese language, utilizing filtered reviews obtained from Dianping's system for counterfeit review detection. Dianping1, the largest hosting platform for Chinese reviews, serves as the primary data source. While determining the recall of Dianping's method proves challenging, the precision is remarkably high. Consequently, all deceptive reviews are successfully recognized.

**2.An Intelligent Model for Online Recruitment Fraud Detection**

By developing a trustworthy model that can identify fraud exposure in online recruitment environments, this study research aims to prevent privacy violations and financial losses for people and businesses. This study offers a significant contribution in the form of an ensemble approach based Random Forest classifier detection model for Online Recruitment Fraud (ORF). The current nature and dearth of studies on this idea distinguish online recruitment fraud detection from other sorts of electronic fraud detection. To meet the goals of this study, the researcher suggested the detection model. Support vector machine is utilized for feature selection, while ensemble classifiers employing Random Forest are used for classification and detection.

**Job Prediction: From Deep Neural Network Models to Applications**

Based on their job descriptions, which often include tough knowledge and abilities, one can decide whether a student or someone searching for employment would be a good fit for the position. Employers must also figure out how to select individuals who are a good fit for the position. With various pre-trained word embeddings on the IT job dataset, we use various deep neural network models, such as Text CNN, Bi-GRU-LSTM-CNN, and Bi-GRU-CNN, to examine job prediction. They also suggested an easy-to-use ensemble model that combines many deep neural network models.

## 3. EXISTING SYSTEM

Detecting fake job postings has been a significant focus in numerous studies, with a particular emphasis on identifying fraudulent accounts. To achieve this, researchers have employed various techniques such as analyzing user demographics, social graphs of followers and followings, content of postings, and temporal aspects of user behavior. In the realm of social network detection, the utilization of a user's posts and social graph has proved effective in identifying counterfeit accounts. To tackle the rapidly evolving false news industry, online learning has been utilized, incorporating both content and network data to continuously enhance the fake detection model. Tan proposed an automated mechanism for detecting phoney users on social networks, leveraging the behavior of genuine users and highlighting the distinctive patterns exhibited in synthetic social network dynamics. By examining user connection graphs and social networks, authentic user patterns have been used to identify fake patterns. Gao et al. observed significant clusters characterized by sporadic posting patterns when applying clustering techniques based on text and URL similarity. Incremental clustering has also been employed on Twitter to identify counterfeit campaigns.

### 3.1 DISADVANTAGES OF EXISTING SYSTEM

• The absence of semi-supervised learning is noted.

• The lack of an option to identify different types of job postings is evident.

## 4. METHODOLOGY

### 4.1 AIM OF THE PROJECT

The primary goal of this undertaking is to identify fraudulent job predictions. We utilize a blacklisted domain detector to classify posts that include URLs from blacklisted sources.

### 4.2 PROPOSED SYSTEM

The method being considered proposes a semi-supervised framework for identifying counterfeit posts. The proposed system primarily comprises two key modules: 1) the module for detecting fraudulent posts, which employs four lightweight detectors to promptly identify deceptive posts in real-time, and 2) the module for updating, which regularly updates the detection models by utilizing posts from the previous time window that have been confidently labeled. These detectors are computationally efficient, suitable for real-time detection, and have been designed based on our observations from a dataset consisting of 14 million posts.

Significantly, both the cluster-level and post-level of our detectors employ categorization techniques. A

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

cluster, in this context, refers to a collection of posts with related characteristics. This flexible architecture allows for easy integration of any features that can be valuable in false detection within the detection framework. The system initiates with a limited set of labeled samples and, through semi-supervised learning, updates the detection models by leveraging confidently labeled posts from the preceding time window.

### 4.3 ADVANTAGES OF PROPOSEDSYSTEM

**Confidently Labeled Posts:** Posts that are labeled by the first three detectors (i.e., blacklisted domain, near duplicate and reliable ham posts) are considered as confidently labeled posts.

## 5. SYSTEM ARCHITECTURE

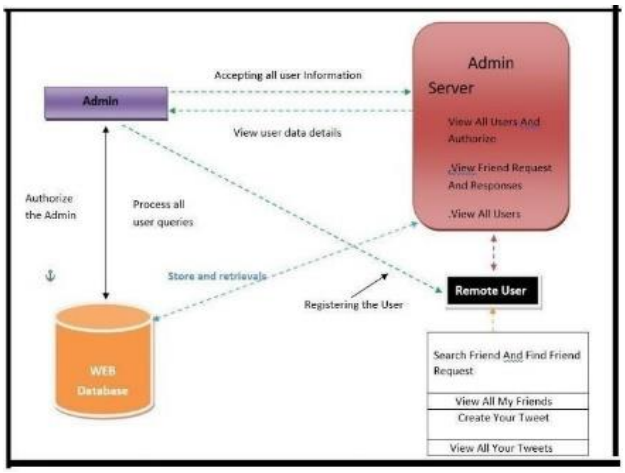


*Fig 5.1 –* ***System Architecture***

This shows that first the data is collected from where the training can be done, after that the data analysis is done on the data. The main step is that the data preprocessing which removes the unwanted tags of HTML or keeps only the important information. Then the model is trained and then the testing is performed. Then finally the classifier is trained and tested for its results.

## 6. ALGORITHMS

When conducting a comparative examination on predicting counterfeit job postings using various data mining techniques, multiple algorithms can be employed to analyze and categorize job postings as authentic or deceptive. The selection of algorithms depends on specific requirements, the available dataset, and the nature of the problem. The following are commonly utilized algorithms for detecting fake job postings:

1. Decision Trees: Decision tree algorithms, such as C4.5 or CART (Classification and Regression Trees), can be utilized to construct a tree-like model based on extracted features from job postings. These algorithms recursively partition the dataset using different attributes to form a decision tree that can classify job postings as genuine or fake.

2. Random Forest: Random Forest is an ensemble learning algorithm that combines multiple decision trees. It creates a collection of decision trees, each trained on a subset of the data and utilizing a random subset of features. The final prediction is made by aggregating the individual tree predictions.

3. Support Vector Machines (SVM): SVM is a widely used algorithm for binary classification tasks. It separates data points by identifying an optimal hyperplane that maximally separates different classes. SVM can effectively differentiate between authentic and fraudulent job postings by mapping the data into a high-dimensional feature space.

4. Logistic Regression: Logistic regression is a commonly employed algorithm for binary classification. It models the relationship between input variables and the probability of a specific class. Logistic regression can be trained on features extracted from job postings to predict the probability of a job posting being counterfeit or genuine.

5. Naive Bayes: Naive Bayes is a probabilistic algorithm based on Bayes' theorem. It assumes independence between features, which contributes to its computational efficiency. Naive Bayes can be used to calculate the probability of a job posting being authentic or deceptive based on the occurrence of specific features.

6. Neural Networks: Neural networks, particularly deep learning models, have shown promising outcomes in various classification tasks. They have the ability to learn intricate patterns and relationships within the data. Techniques such as feedforward neural networks or convolutional neural networks (CNNs) can be employed to classify job postings as genuine or fake.

It is important to consider the specific characteristics of the dataset and the desired performance metrics when selecting the most appropriate algorithm. It is customary to compare multiple algorithms and evaluate their performance using metrics such as accuracy, precision, recall, F1-score, or area under the receiver operating characteristic curve (AUC-ROC) to determine the optimal algorithm for the given task.

## 7. IMPLEMENTATION SCREENSHOTS

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

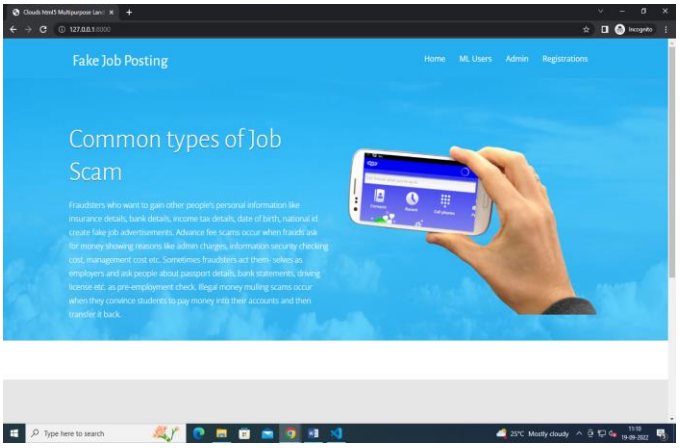


***Fig 7.1** –Homepage*

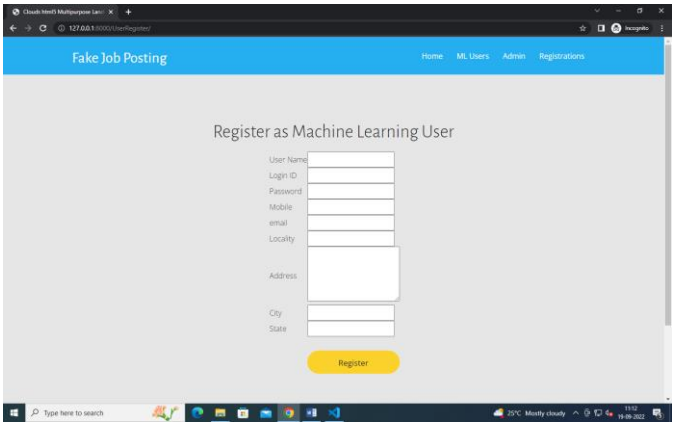


***Fig 7.2** –Register page*

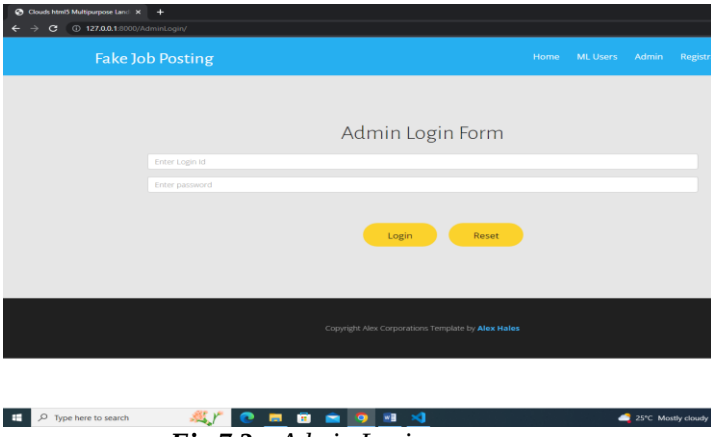


***Fig 7.3** – Admin Login page*

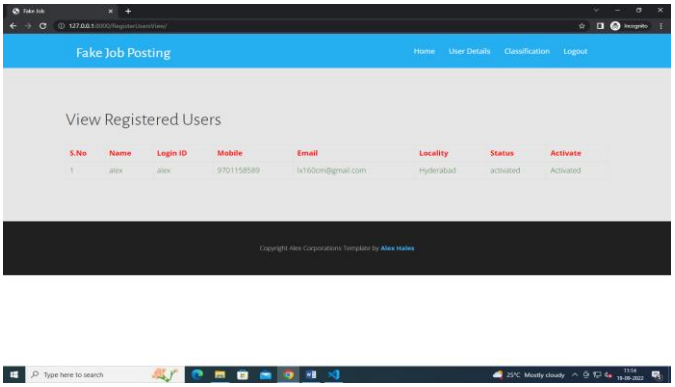


***Fig 7.4** –Registered Users*

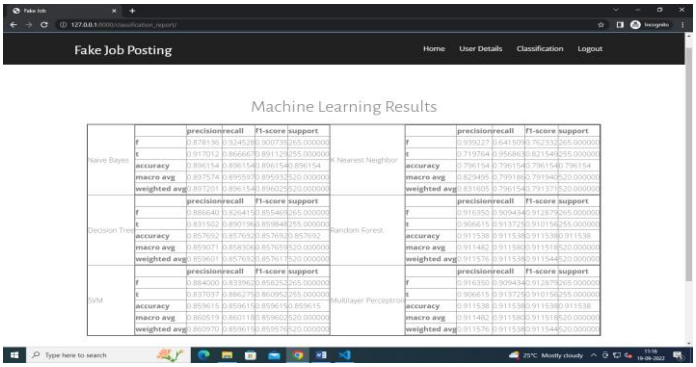


***Fig 7.5** –Machine Learning Results*

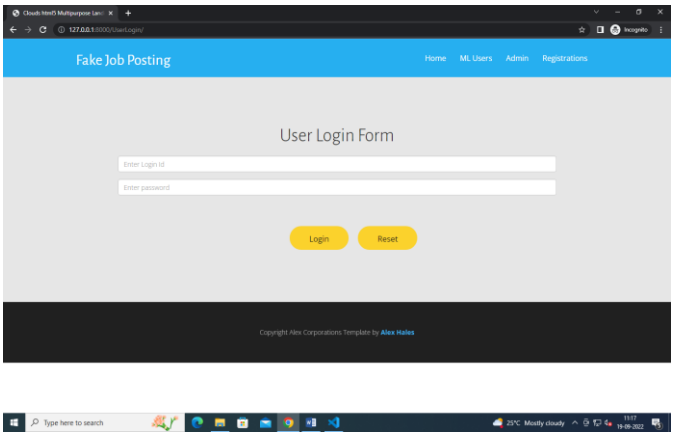


***Fig 7.6** – User Login Page*

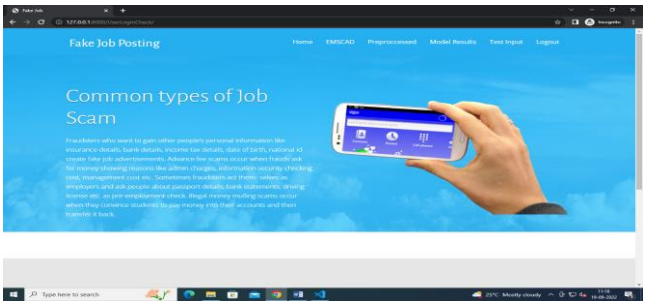


***Fig 7.7** – User Home page*

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

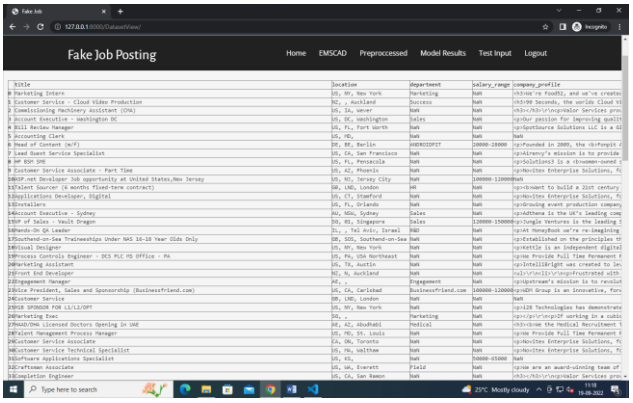


***Fig 7.8** – Fake Job Links*

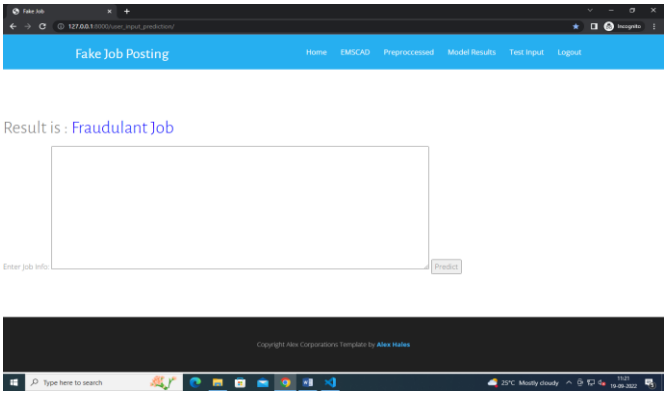


***Fig 7.9** – Output for Fake Jobs*

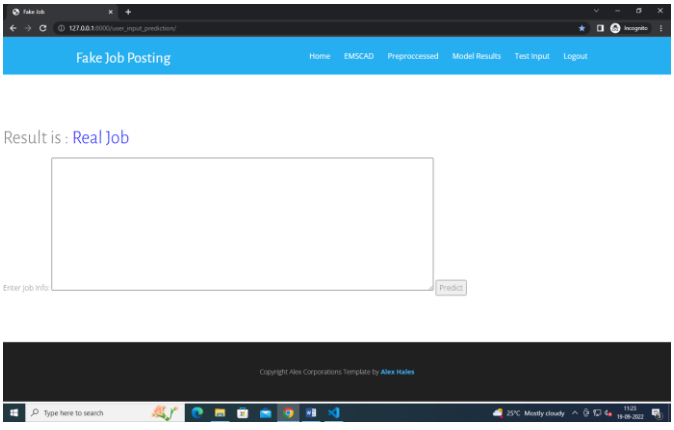


***Fig 7.10** –Output for Real Jobs*

## 8. CONCLUSIONS

The detection of fraudulent employment schemes has become a growing concern worldwide. We have conducted a comprehensive examination of the impact of job scams, an area of study that presents numerous challenges in identifying counterfeit job postings while offering significant research opportunities. We provide an overview of the dataset's users, highlighting distinctive characteristics that differentiate between legitimate and deceptive jobs. Utilizing the insights gained from this characterization study, we develop a robust system for detecting spammers. By implementing a classification technique, we achieve a notable accuracy in identifying a substantial portion of fake job postings, with only a minimal misclassification rate of genuine users. Additionally, we explore the impact of different attribute sets on our classification strategy, considering various trade-offs. Our findings demonstrate that our approach can effectively identify counterfeit job postings with accuracy, even when utilizing diverse subsets of variables.

## 9. REFERENCES

[1] S. Vidros, C. Kolilias , G. Kambourakis ,and L. Akoglu, "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset", Future Internet 2017, 9, 6; doi:10.3390/fi9010006.

[2] B. Alghamdi, F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection", Journal of Information Security, 2019, Vol 10, pp. 155 176, https://doi.org/10.4236/iis.2019.103009 .

[3] Tin Van Huynh1, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen1, and Anh Gia- Tuan Nguyen, "Job Prediction: From Deep Neural Network Models to Applications", RIVF International Conference on Computing and Communication Technologies (RIVF), 2020.

**1.** Jiawei Zhang, Bowen Dong, Philip S. Yu, "FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network", IEEE 36th International Conference on Data Engineering (ICDE), 2020.

**2.** P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," Neurocomputing, vol. 174, pp. 806 814,2016.

**3.** C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved BiLSTM-CNN," in 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018, pp. 890-893.

[10] K. R. Remya and J. S. Ramya, "Using weighted majority voting classifier combination for relation classification inbiomedical texts," International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 1205- 1209.

**4.** Yasin, A. and Abuhasan, A. (2016) An Intelligent Classification Model for Phishing Email Detection. International Journal of Network Security& Its Applications, 8, 55-72. https://doi.org/10.5121/imsa.2016.8405

[7] Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi- Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan LuuThuy Nguyen."Emotion Recognition for Vietnamese Social Media Text", arXiv Prepr. arXiv:1911.09339, 2019. [13] Thin Van Dang, Vu Duc Nguyen, Kiet Van Nguyen and Ngan LuuThuy Nguyen, "Deep learning for aspect detection on vietnamese reviews" in In Proceeding of the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), 2018, pp. 104-109.

collective positive-unlabeled learning. In Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14-17 December 2014; pp. 899-90

[8] Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan LuuThuy Nguyen."Emotion Recognition for Vietnamese Social Media Text", arXiv Prepr. arXiv:1911.09339, 2019.

[9] Thin Van Dang, Vu Duc Nguyen, Kiet Van Nguyen and Ngan LuuThuy Nguyen, "Deep learning for aspect detection on vietnamese reviews" in In Proceeding of the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), 2018, pp. 104-109.

[10] Li, H.; Chen, Z.; Liu, B.; Wei, X.; Shao, J. Spotting fake reviews via collective positive-unlabeled learning. In Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14-17 December 2014; pp. 899-904.