

Python & Deep Learning

Professor: Dr.Lee

Instructors: Saria /Ashish

Project Team: Haritha Bheemineni– 8, Satya Anusha Prattipati- 18, Yamini Dontireddy – 28, Vamshi Jakkula-26

Fake Job Description Prediction (Real Or Fake)

**MOTIVATION**

The aim is to provide a clear picture of the Employment Scam problem to the academic & research community.

**DESCRIPTION OF THE PROBLEM**

Fake news has evolved into a sizeable industry all over the world and 35% of news read online are fake - its cuts across politics, entertainment, finance and social matters which has high public attention. Websites which deliberately published hoaxes and misleading information popped up across the internet and were often shared on social media to increase their reach. Job and employment is sensitive topic where fake advertisements are misleading naive students and gullible job seekers to share their confidential information (PII) and money and also sizeable economic loss to the human society.

**SCOPE**

Scope of this project is to identify few real time scenarios for predictive analytics and understand the reality of the data and will cover the below topics:

* Summarize & Classify Dataset
* Visualizing the Dataset
* Evaluate algorithms like BERT
* Derive predictable insights

**DATA SET DESCRIPTION**

The University of the Aegean published the Employment Scam Aegean Dataset. The data contains about 18K real-life job advertisements. The dataset consists of both of real and fake advertisements. This data consists of both textual and meta-information about the jobs such as company profile, logo, industry etc. Using the classification models available in Python, job posting would be filtered based on the 10 job related meta-data provided for each of the job postings.

**Summarize & Classify the Dataset**: Data would be summarized based on the dimensions and the statistical summary of the attributes and classified based on the class variables.

**Visualizing the Dataset**: Identified 2 different types of plots: Univariate plots to better understand all the attributes and Multivariate plots to understand the relationship between the attributes.

**Evaluate Algorithms:** Will try to set-up the test harness and build multiple different models in classification like Logistic Regression, Naïve Bayes Classifier, Random Forest, Decision tress, Adaptive Boosting to compare along with BERT. BERT stands for Bidirectional Encoder Representations from Transformers. BERT works by randomly masking word tokens and representing each masked word with a vector based on its context. The two applications of BERT are “pre-training” and “fine-tuning”.

**Derive predictable insights:** Will fit the model on the entire training dataset and make predictions on the validation dataset. And finally evaluate the predictions by comparing them to the expected results in the validation set, then calculate classification accuracy.

**References:**

<https://en.wikipedia.org/wiki/Fake_news>

<https://www.statista.com/topics/3251/fake-news/>

<http://emscad.samos.aegean.gr/>

<https://arxiv.org/pdf/1810.04805.pdf>