Fake News Detection using LIAR-Dataset

A) INTRODUCTION

1.1) Fact-checking is the process of assessing the veracity of claims. It requires identifying evidence from trusted sources, understanding the context, and reasoning about what can be inferred from the evidence. Several organizations such as FACTCHECK.org, POLITIFACT.com and FULLFACT.org are devoted to such activities, and the final verdict can reflect varying degrees of truth (e.g., POLITIFACT labels claims as true, mostly true, half true, mostly false, false and pants on fire).

1.2) **There are two Tasks**

- Binary Classification
- Six-Way Classification

1 Binary Classification

It is the task of classifying the elements of a given set into two groups (predicting which group each one belongs to) on the basis of a classification rule.) There are many classifiers algorithms to use for the different number of predictions, but here we use **Random Forest.** and another one is being used is **LSTM Classification Keras.**
Random Forest: Random Forest Classifier is ensemble algorithm. In next one or two posts we shall explore such algorithms. Ensembled algorithms are those which combines more than one algorithms of same or different kind for classifying objects. For example, running prediction over Naive Bayes, SVM and Decision Tree and then taking vote for final consideration of class for test object.

- **Here by using Bags of word Technique we are achieving the Accuracy: 63 percent**

LSTM Classification Keras:
How it Work: A sequence is a set of values where each value corresponds to an observation at a specific point in time. Sequence prediction involves using historical sequential data to predict the next value or values. Machine learning models that successfully deal with sequential data are RNNâs (Recurrent Neural Networks).

First, to give some context, recall that LSTM are used as Recurrent Neural Networks (RNN).

RNNs are neural networks that used previous output as inputs. We consider that RNNs has a kind of internal dimension, that will be the dimension of hi vectors. This dimension is constant over time.

After one fed all inputs xi into the RNN, the last output, ht, is supposed to carry information about the whole sentence. This is theoretically true, but it shows weakness for long sequences. Thatâs what LSTMs tries to solve.

- **Here by using RNN we are achieving the Accuracy: 58 percent**

2 Six-Way Classification

With continuous increase in available data, there is a pressing need to organize it and modern classification problems often involve the prediction of multiple labels simultaneously associated with a single instance. Known as Multi-Label Classification, it is one such task which is omnipresent in many real world problems.Here we are classifying the data by using again **Random Forest** as random forest classifier is so much flexible to train multiple label datasets.

- **Here by using Bag of words we are achieving the Accuracy: 28 percent**

B) Dataset

2 Dataset The LIAR dataset introduced by (Wang, 2017) consists of 12,836 short statements taken from POLITIFACT and labeled by humans for truthfulness, subject, context/venue, speaker, state, party, and prior history. For truthfulness, the LIAR dataset has six labels: pants-fire, false, mostlyfalse, half-true, mostly-true, and true. These six label sets are relatively balanced in size. The statements were collected from a variety of broadcasting mediums, like TV interviews, speeches, tweets, debates, and they cover a broad range of topics such as the economy, health care, taxes and election. We extend the LIAR dataset to the LIAR-PLUS dataset by automatically extracting for each claim the justification that humans have provided in the fact-checking article associated with the claim. Most of the articles end with a summary that has a headline âour rulingâ or âsumming upâ. This summary usually has several justification sentences that are related to the statement. We extract all sentences in these summary sections, or the last five sentences in the fact-checking article when no summary exists. We filter out the sentence that has the verdict and related words. These extracted sentences can support or contradict the statement, which is expected to enhance the accuracy of the classification approaches.

LINK: https://www.cs.ucsb.edu/ william/data/liar$_d ataset.zip$

Data-Preprocessing

About NLTK: The Natural Language Toolkit is an open source library for the Python programming language originally written by Steven Bird, Edward Loper and Ewan Klein for use in development and education. It comes with a hands-on guide that introduces topics in computational linguistics as well as programming fundamentals for Python which makes it suitable for linguists who have no deep knowledge in programming, engineers and researchers that need to delve into computational linguistics, students and educators.

NLTK includes more than 50 corpora and lexical sources such as the Penn Treebank Corpus, Open Multilingual Wordnet, Problem Report Corpus, and Linâs Dependency Thesaurus.

About Bags Of Word technique: A bag-of-words model, or BoW for short, is a way of extracting features from text for use in modeling, such as with machine learning algorithms. ... A bag-of-words is a representation of text that describes the occurrence of words within a document.

About Tokenization: Tokenization is the process of splitting the given text into smaller pieces called tokens. Words, numbers, punctuation marks, and others can be considered as tokens. In this table (âTokenizationâ sheet) several tools for implementing tokenization are described.

C) Conclusion and Future Scope

Here Show casing the model of the human-provided justification form the factchecking article associated with a claim is important leading to significant improvements when compared to modeling just the claim/statement and metadata for all the machine learning models both in a binary and a six-way classification task. We released LIAR-PLUS, the extended LIAR dataset that contains the automatically extracted justification. We also provided an error analysis and discussion of per-class performance. In addition, we plan to develop methods for evidence extraction from the web (similar to the goals of the FEVER shared task (Thorne et al., 2018)) and compare the results of the automatically extracted evidence with the human-provided justifications for factchecking the claims.

D) Installations and Instructions

1) Import all the required python libraries and packages
2) Data Cleaning and pre-processing
3) Perform task 1 with the following algorithm Random forest and LsTM
4) Perform task 2 with Random Forest classification
5) Saving of model and verify the prediction.