Evaluating Credibility of Web-Based News Articles by using Natural Language Processing and Deep Learning

Virtee Parekh (vvp2639@rit.edu) | Advisor: Dr. Carol Romanowski (cjrcms@rit.edu)

Rochester Institute of Technology

Computer Science @ RIT
B. Thomas Golisano College of Computing & Information Sciences

INTRODUCTION

Fake news articles became rampant during the 2016 US presidential elections when hoax news was used to attempt to influence voters. Fake news consist of religious propaganda, scientifically dubious claims, clickbait, articles promoting racial discrimination and conspiracy theories. Such falsified articles result into chaos and unwise decision making that proves to be harmful to mankind.

DATA

- Pre-existing lists of credible and non-credible news sources are obtained from OpenSources and MediaBiasFactCheck.
- Each source was manually scraped for the news headline and its text body.
- A target label, 0 for fake news and 1 for genuine news item is added.
- Total number of records collected was 45,000.

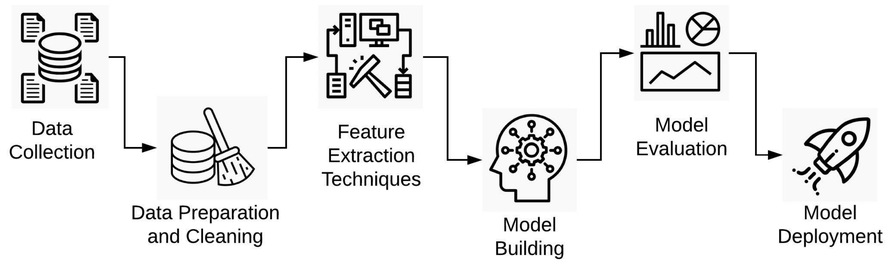
FEATURE EXTRACTION

Raw text needs to be converted to a numeric format for processing by using the following methods:

- Bag of Words with *N*-grams
- Term Frequency - Inverse Document Frequency (TF-IDF)
- Hashing Vectorizer
- Custom Word Embedding in Keras
- Global Vectors for Word Representation (GloVe)

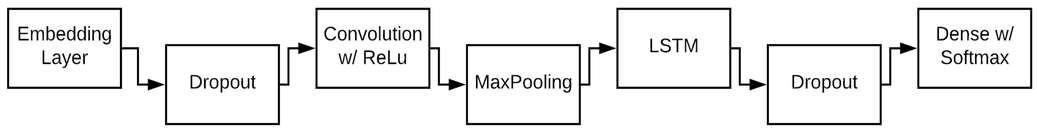
IMPLEMENTATION

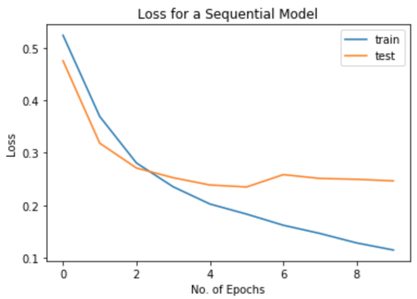
General Pipeline of the System

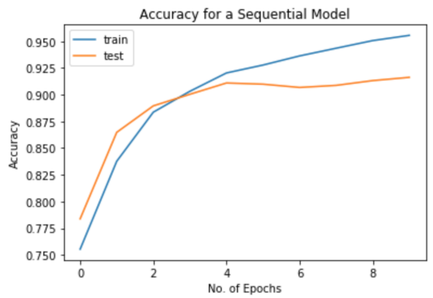


Machine Learning Approach	Deep Learning Approach
• **Data Cleaning** Stop-words removal, lemmatization, removal of extra whitespaces & special characters, expanding contractions • **Models** Naïve Bayes, Stochastic Gradient Descent w/ Support Vector Machines (SGDC w/SVM), Logistic Regression	• **Data Cleaning** Removal of extra whitespaces & special characters, expanding contractions • **Models** Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), Convolutional Neural Network w/ LSTM (CNN-LSTM)

Best Performing Deep Learning Architecture for this Data







TOP RESULTS

Top results consist of models that have high test accuracies with minimum over-fitting.

Model	Feature Extraction	Train Accuracy (%)	Test Accuracy (%)
CNN-LSTM	GloVe	93.3	91.1
GRU	GloVe	92.2	90.0
SGDC - SVM	TF-IDF	94.1	89.4
Logistic Regression	TF-IDF	96.3	89.2
SGDC - SVM	Hashing	93.5	88.0
LSTM	GloVe	92.8	87.8
Naïve Bayes	TF-IDF	94.7	86.3
LSTM	Keras Embedding	93.2	85.5

REVIEW AND FUTURE WORK

- From all the numerous combinations of feature extraction methods and models developed, the model using CNN and LSTM along with GloVe embedding works best for this data.
- Future work would involve training deeper networks and deploying it as a web application.

REFERENCES

[1] Wang, William Yang. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection.” ACL (2017).
[2] Estela, Zach. “Check the Bias of a News Source.” News Bias Classifier, www.areyoufakenews.com
[3] System pipeline icons from The Noun Project.