Understanding Machine vs Human Generated Text in News Articles

Mertay Dayanc, Computer Science
Mentor: Huan Liu, Professor - Ahmadreza Mosallanezhad, Ph.D
School of Computing, Informatics, and Decision System Engineering

Introduction

With the current advancement in NLP, machine-generated fake news, like fake news articles that attempt to influence us by appearing to be credible or deep fakes that try to influence us by generating fake videos of influential entities. It can change our political views and our basic understanding of reality if misused, for, how will we truly be able to tell the difference if we were to be exposed to fake news continuously.

After analyzing the trends in Machine generated and human generated text, a model that can take text as an input and outputs Human Generated or Machine generated will be implemented. Impact of this research would be on our society, because if this Machine Generated news take adversarial point to manipulate societies thoughts it will cause huge ethical problems. This research can help preventing the spread of fake news articles.

Dataset

Human Written articles are collected from Politifact Dataset, around 32,000 entries. First 10 words of the human written articles are fed into GPT-2 - 1558M. Outcome is labeled as Machine Generated Article. Only 2,000 human written and 2,000 machine generated articles used to train dataset. Then it is split into 3,200 training and 800 validation.

Model

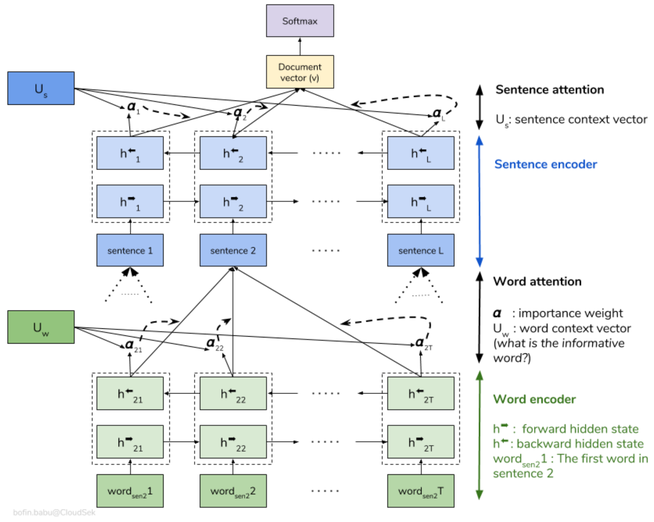
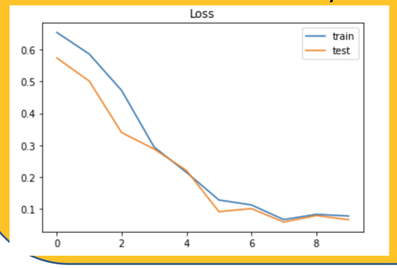


Figure 1. Colored Hierarchical Attention Networks Model for Document Classification.

Implementation Details

After having the model implemented, text preprocessing and tokenization done over the dataset. Then, using Stanford word2vec weights, word embeddings of 50 dimensions created. We converted the preprocessed text into numeric data and this become the input to the model. Model returns posterior probability of 2 classes, word importance weights and sentence importance weights. Which is then used for data visualization by creating heatmap of text input.



According to the Loss figure it can be seen that model is trained properly.

Figure 2. Validation and Train loss figure

Results

Model	Accuracy
LSTM - Baseline	%83.7
HAN	**%97.5**

Figure 3. LSTM and HAN performance on the same dataset

LSTM can be used as a text classifier however HAN uses several LSTM architecture to create more comprehensive contextual information. Hence, HAN does significantly a better job then LSTM.

Data Visualization

Human Written Text

some stuff with josh under a blanket at josh 's mom 's house , but she did n't get naked or anything and they just rubbed against each other with all their clothes had no idea she was ready to start spreading for guys , ' added . 'm really glad I held out for her . jessica is really frickin ' hot . ' milly 's level of willingness to put out has reportedly been monitored closely by the student body of dewey

Machine Generated Text

time to arts and culture . his wife , kristin , 28 , loves making enamel jewelry with icelandic bronzes and labradorite . the whaling village community is , says greg starr , 29 , a creative pastoral robotics spec designer . everyone here seems to be in cutthroat competition with all the other arts organizations in town . it 's huge and very vibrant and local and unusual . besides yellowhammer 's the zoo and kenai museum (limited admissions everyone !) , the most popular events this fall go to

---

Discussion

In the human written text phrase “is really fricking hot” is highlighted. It makes sense because this is an abstract phrase. According to the experiment made that would not be generated by GPT with high probability. In the Machine Generated one obsolete words are highlighted like, “enamel, labradorite, kenai”. Also, there are more words highlighted in Machine Generated one with less weight. Which might be due to machine generated ones having a monotone language. So, words equally effect the outcome.

FURI

Ira A. Fulton Schools of Engineering
Arizona State University