Online Post Fake News Detector

by Gradient Ascent: Ariel Rotter-Aboyoun (arottera), Daniel Kostovetsky (dkostove), Julius Sun (jsun6), Raymond Cao (rcao6)

1. Introduction

The proliferation of fake news on social media has become a well-publicized problem in recent years. Currently the biggest issue is the difficulty of verifying the authenticity and accuracy of content shared online. Since relatively few people take the time to extensively vet the news they share, developing an automated process to detect false content could significantly curb the spread of dangerous misinformation. We build a model that will classify public statements, such as social media posts, on the basis of whether they are real or fake, without prior knowledge of the subject domain. The model is based on the paper *Fake News Identification on Twitter with Hybrid CNN and RNN Models* [1], and, as the paper's title suggests, involves LSTMs and CNNs.

1.1. Related Work

Fake news detection is a popular problem in natural language processing, with many papers released on the subject in recent years, such as *Effective Fake News Detection with Deep Diffusive Neural Network* [2], and *Event Adversarial Neural Networks for Multi-Modal Fake News Detection* [4]. We chose our paper because it used techniques that we had already encountered in class, as well as a relatively structured dataset, while being complex enough to provide a challenge.

2. Methodology

The model paper suggested three different architectures: a LSTM RNN, a LSTM with dropout regularization, and a LSTM with a 1D CNN. Few further details were given, other than a suggested dropout rate of 20% [1]. We implemented all three architectures. We used a word embedding layer; then, either a dropout layer, a 1D convolution and pooling layer, or none of the above; then, an LSTM, and finally, a dense output layer with softmax activation. As suggested by the paper, we used a trial-and-error grid search to tune our hyperparameters, such as the learning rate, embedding size, and the size of the LSTM. General observation was also used to select the number of epochs for training.

2.1. Data

The dataset used in the model paper consisted of 5,800 tweets that reported on a news event, which were classified as either true or false. This dataset, however, is not publicly available. Instead, we used a similar dataset, known as LIAR, introduced in *"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection* [3]. LIAR consists of 13,000 public statements made by prominent political and media figures, including a validation set (10%) and test set (10%). They are classified into one of six truth categories (taken from the fact-checking website PolitiFact): true, mostly true, half true, barely true, false, or pants on fire. The dataset also contains additional information, such as the speaker of the statement and the occasion, although we ignored this information to maintain consistency with our model paper.

Statement	**Truth Value**
Over the past five years the federal government has paid out $601 million in retirement and disability benefits to deceased former federal employees.	True
Suzanne Bonamici supports a plan that will cut choice for Medicare Advantage seniors.	Half-True
In the case of a catastrophic event, the Atlanta-area offices of the Centers for Disease Control and Prevention will self-destruct.	Pants on Fire

Figure 1. Sample statements and truth labels from the LIAR dataset.

2.2. Metrics

The model paper achieved 82% accuracy on its two-class dataset [1]. However, it is difficult to compare our results to this benchmark, since our dataset has six classes. The LIAR paper tested its dataset on a variety of machine learning models, incorporating both the statements and their contextual information, and none of their models obtained higher than 28% accuracy [3]. We will be aiming to achieve close to this accuracy, and at the very least, to do better than a random classifier (16.7%) and a majority classifier (20.8%).

3. Challenges

A large challenge with beginning this project was finding a suitable dataset to use. Ideally, the paper's models could be reimplemented with the dataset originally used or one very similar to it. However, after a lot of searching the group was unable to find such a public dataset. Time was spent attempting to get access to and compile datasets from other papers and Amazon Turk, but all of these ultimately had a missing component, had unreliable truth ratings, or were unable to be obtained.

With the required change in dataset and the change in creating a binary classifier to a multi-class classifier, it was more difficult to draw hard performance distinctions between the three models we reimplemented, unlike the original paper. On the LIAR dataset, as mentioned above, a simple majority classifier achieves an accuracy of 20.8%. The best performing classifier achieved an accuracy of 27.8%, and many fell around 25% accuracy. Furthermore, between training instances of the models, we observed around 1% variations of the test accuracy. With these factors, any differences in percentages were very small and hard to make any conclusions from.

4. Results

Our best accuracy of 25.78% was achieved by the plain LSTM model with optimally tuned hyperparameters. All results are given below. We also assigned the truth classes numerical labels (0-5) and calculated the mean squared error for each model.

Model	**Accuracy**	**Mean Squared Error**
LSTM	25.78%	3.34296875
LSTM with Dropout	25.16%	3.3671875
LSTM with CNN	25.16%	3.3984375

Figure 2. Sample correct and incorrect model predictions.

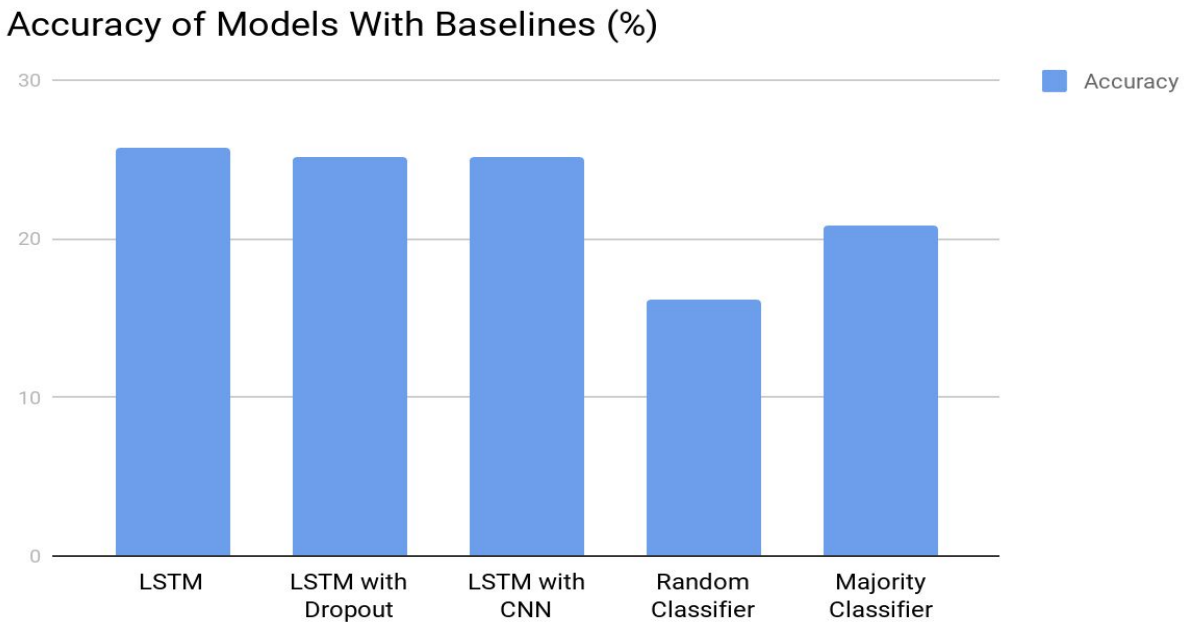

Figure 3. Test accuracy for the three models with two baseline classifiers for context.

Statement	**Prediction**	**Actual**
Rebuilding three high schools will benefit 40 percent of Portland Public School students.	mostly-true	pants-fire
My home state since June of 2009 created 40 percent of the new jobs in America.	mostly-true	mostly-true
Victory! Republicans by 2 to 1 vote to endorse Mark Neumann on first ballot at GOP convention.	half-true	false
If you are a member of union, your median weekly income is roughly $200 more than if you are a nonunion member, and that doesn't include benefits.	half-true	true

Figure 4. Sample correct and incorrect model predictions.

5. Reflection and Discussion

Determining the truthfulness of a statement without context is a difficult problem. We were only able to achieve a 25.78% accuracy on our dataset, which may seem low, but is close to the paper's 27% accuracy. For reference, we attempted to classify training examples by hand and achieved an accuracy that was far less than 25.78%, and in fact, no better than random guessing. Many statements in our dataset, such as *"The economy bled $24 billion due to the government shutdown,"* are virtually impossible to factually evaluate without an external reference. Others, such as *"We have a federal government that thinks they have the authority to regulate our toilet seats,"* are subjective, opinionated, or don't have a well-defined truth value. Some statements did contain certain words that made them identifiable as true, or more frequently, false. For instance, the sentence *"In the case of a catastrophic event, the Atlanta-area offices of the Centers for Disease Control and Prevention will self-destruct,"* includes the word 'self-destruct' that is associated with fictional stories, implying that it is false. Perhaps our model learned to make classifications by looking for such words.

Even though it was expected that the three models would have very similar performance, it surprised us just how close, even identical, the accuracy of the models and their mean squared error were. LSTM with Dropout and LSTM with CNN both had 25.16% testing accuracy with a difference in mean squared error around 0.03. It's clear that this dataset was not the best to compare and contrast these models.

Future work could involve taking the additional context information in the LIAR dataset into account. Additionally, models may be given access to some factual database, such as pages of Wikipedia, for assistance. Perhaps a new dataset could be gathered where sentences are labeled with an ‘unknown’ truth value, giving models an escape hatch for opinionated or subjected inputs. It may also be useful to analyze the model during execution in order to more precisely determine what it looks for when making classifications.

6. Code Repository

Our codebase is accessible on GitHub at: https://github.com/CaoRuiming/gradient-ascent-project

7. References

[1] Ajao, Oluwaseun & Bhowmik, Deepayan & Zargari, Shahrzad. (2018). Fake News Identification on Twitter with Hybrid CNN and RNN Models.

[2] Jiawei, Zhang & Cui, Limeng & Fu, Yanjie & Gouza, Fisher. (2018). Fake News Detection with Deep Diffusive Network Model.

[3] Wang, William. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. 422-426. 10.18653/v1/P17-2067.

[4] Wang, Yaqing & Ma, Fenglong & Jin, Zhiwei & Yuan, Ye & Xun, Guangxu & Jha, Kishlay & Su, Lu & Gao, Jing. (2018). EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. 849-857. 10.1145/3219819.3219903.