Detect a News is Fake or Not

Abstract
With the development of social media, people can get and send more and more information from internet. People get news from internet for their daily life. However, many news in the internet including bias and other bad impact on people. When people read this bias news online, they don't realize it and will spread this bias news to others. Our project may help people to detect the reality of the news. Detection fake news is a hard topic. We plant to use LSTM and tensor flow to build the model to do detection.

Literature reviews

"I trained fake news detection Al with > 95% accuracy, and almost went crazy." *Toward Data Science,* 11/01 https://towardsdatascience.com/i-trained-fake-news-detection-ai-with-95-accuracy-and-almost-went-crazy-d10589aa57c

"Keras." *TensorFlow,* 8/08/2018 https://www.tensorflow.org/guide/keras

"SMOTE with Imbalanced Data." *Kaggle* https://www.kaggle.com/qianchao/smote-with-imbalance-data

"Keras LSTM tutorial – How to easily build a powerful deep learning language mode." Adventures in Machine Learning, 02/03/2018 http://adventuresinmachinelearning.com/keras-lstm-tutorial/

"GloVe: Global Vectors for Word Representation." 08/2014. https://nlp.stanford.edu/projects/glove/

"keras-team/keras." *GitHub*. 09/07. https://github.com/keras-team/keras/blob/master/examples/pretrained_word_embeddings.py

We don't know how others solve this problem because we didn't see any code to solving this problem online. We just saw one blog that indicates that he uses Natural language processing to do it. And then he finds that trying to detect the fake news is not as good as detecting the real news. So, he finally try to gathering real news information and detect if a news a real. We don't know the details about how he solves this problem. However, we also use Natural language processing approach this problem. We use LSTM algorithm to build the model because we think the contents in the news have order.

Detail of your approach

The basic idea is converting a news into vector and fitting this vector to TensorFlow model by adding LSTM layer. For the first part, we read data from json files and csv files and store them in panda dataframe. Then padding these data into integer. After that, we load GloVe dictionary. For the second part, we found the data is imbalance. There are many fake news and less real news. So we use oversampling method to make the data balance. For the third part, we split data after padding into train and validation. For the forth part, we start to train the model. We use packages from keras to build model. In the last part, we extract a new news content from BBC website and padding it. We predict this new news. The prediction result is close to 0.

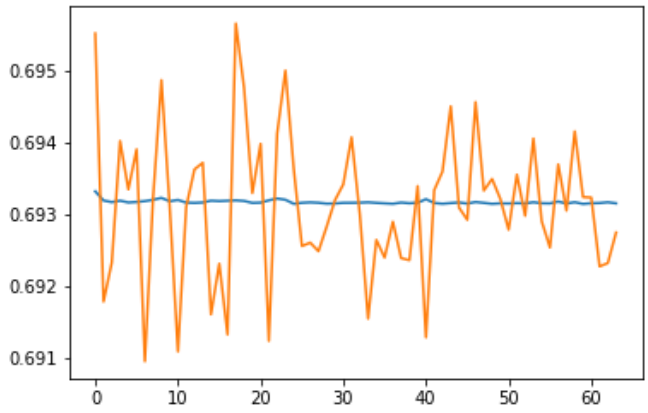
Dataset description

Our data include 443 bias news, 430 conspiracy news, 246 hate news, 146 satire news, 121 state news, 102 junksci news, 230 fake news and 211 real news. We consider bias, conspiracy, hate, satire, state, junksci and fake are both fake news and labeled these news as class 1 and real news as class 0.

Experiment detail

```
plt.plot(loss)
plt.plot(val_loss)
```

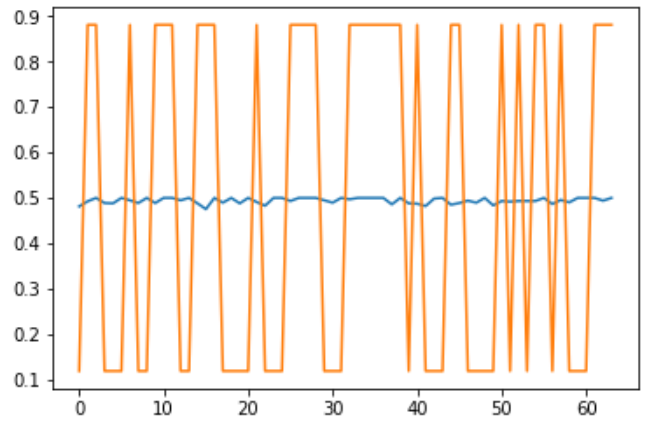
```
[<matplotlib.lines.Line2D at 0x1d64fe2e358>]
```



This graph shows the loss and validation loss, the blue line is loss and the orange line is validation loss. As we can see, the loss from training set doesn't change with training and it is close to 0.693. The loss from validations set goes up and down when training and finally close to 0.693. This means our model is not overfitting.

```
plt.plot(acc)
plt.plot(val_acc)
```

```
[<matplotlib.lines.Line2D at 0x1d64fe5d7b8>]
```



The blue line is training set and the orange line is validation set. This graph shows the accuracy of the model from training set is around 0.5. The accuracy goes up and down around 0.5 for validation set.

Error analysis

I think the reason loss and accuracy doesn't change in the training set is news data is not appropriate to oversampling. We assume there are half fake news and half fake news. So, we sample the data to half fake and half real news in the real word. The oversampling method will select each real news (less records) 8 times and this selection will make there are 50% fake news and 50% real news. We think we need search more real news content online to make the data balance rather than oversampling.

Conclusion

we extract a new news from BBC website and padding it. We found the prediction result is close to 0. The prediction result is real.