Classifying News Article into Negative or Non-Negative

Coord.: Kristjan Roosild, TransferWise

P10 Team members: Wanting Huang, Shakshi Sharma, Carel Kuusk, Sebastien Boire

How to identify an article with negative bout an entity?

Binary classification

Negative

- Precise subject
- Justified accusation
- Illegal activity

Not negative

- Wide subject
- Uncertain illegal activity

About Dataset

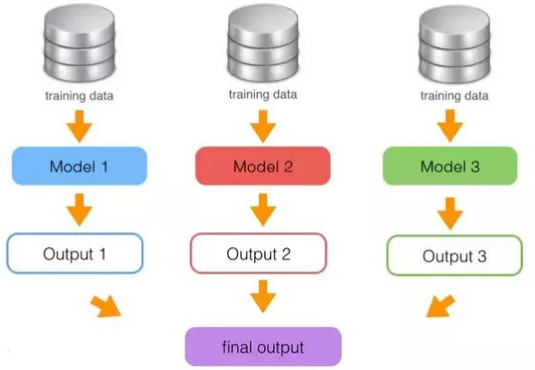
- Gathered both type of data manually using pointers provided by Kristjan
- Did Quality Assurance of the data
- Considered Random articles also part of the Non-Negative News article
- 622 articles classified as adverse media, 465 as non-adverse media

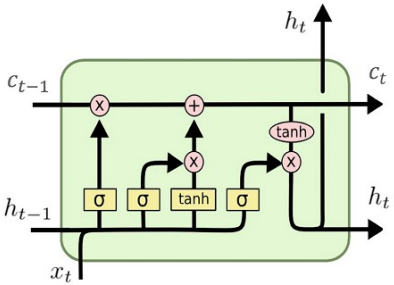
Pre-processing of the dataset

- We used two columns from the dataset - ‘article’ and ‘label’
- Check for null values if any
- Shuffled the data
- Generates embedding vectors of size 100 for each word using Word2Vec technique after performing some basic NLP stuff such as converting to lowercase, removing stop words and punctuations, tokenize the sentence into words
- Also removed noisy words such as ‘zvemushonga’, ‘zvavamw’
- About 1000 articles with 2000 attributes

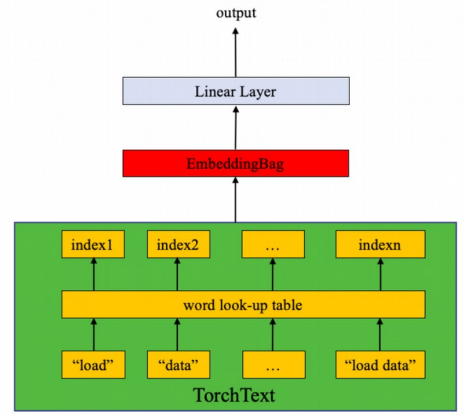
Models

- Ensemble method
  (RF+XGBoost+SVC+GBM)
- LSTM (with embedding matrix obtained from our data and from Wikipedia)
- TextSentiment (EmbeddingBag + LinearLayer)
- Term frequency-inverse doc frequency (tf-idf) + dense r
- BERT





LSTM
(Long-Short Term Memory)



Results obtained: F1-score

Overfitting:

- LSTM with embeddings from our data: 0.65
- LSTM with embeddings from Wikipedia: 0.67
- TextSentiment: 0.86
- BERT (accuracy, not f1): 0.53
- Tf-idf: 0.88-0.9

Most promising:

- Ensemble (RF+XGBoost+SVC+GBM): 0.91

Conclusion

- Choice of model: depends on the amount of data available
- Success in isolating negative news articles with classification algorithms
- The dataset is somehow small, a problem of overfitting

Thank you for listening !