Editorial News Classification

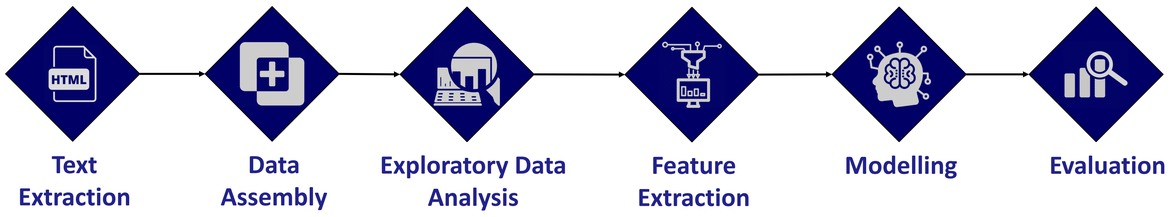
Data Science Institute
COLUMBIA UNIVERSITY

Aastha Joshi, Ameya Karnad, Nirali Shah, Sarang Gupta, Ujjwal Peshin
Dr. Daniel Preotiuc-Pietro, Kai-Zhan Lee
Dr. Smaranda Muresan

Bloomberg

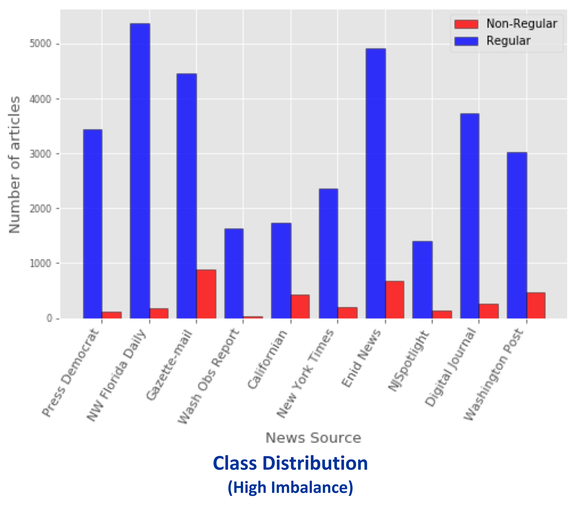
Problem Statement & Motivation

The main objective of the project is to build a classifier which would be able to predict whether a news article is an editorial or not. It was motivated by the need to improve on the editorial tab of the Bloomberg terminal. This would remove the reliance on news sources to provide such information.

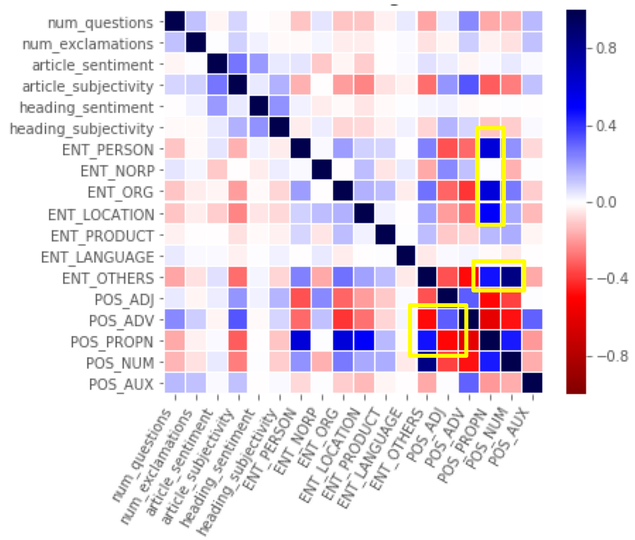


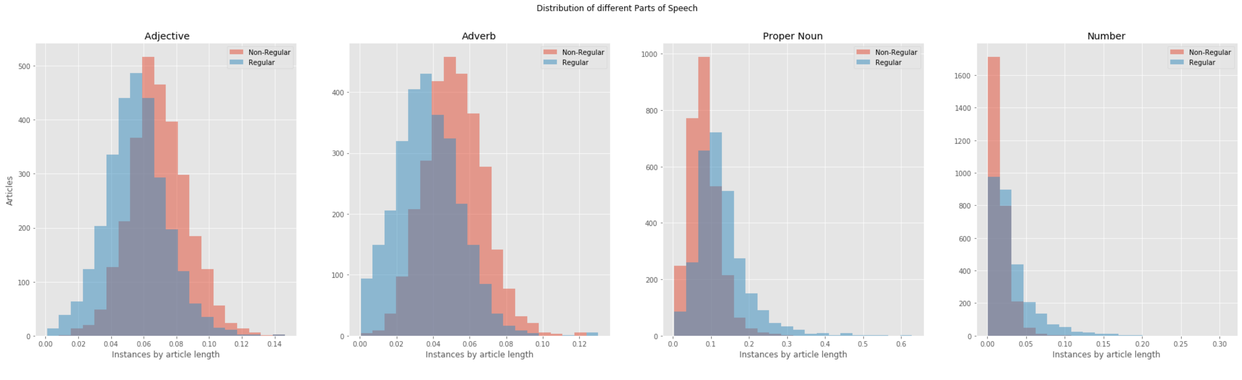
Exploratory Data Analysis

News articles are heavily imbalanced towards regular news articles, with an average of every 10th article being non-regular. An initial analysis on NER and POS was performed and there was a difference in distributions for each class. After extracting features based on sentiment, NER, and POS, it was interesting to see that several features were highly correlated, such as person entity, and POS of proper noun.



Class Distribution
(High Imbalance)

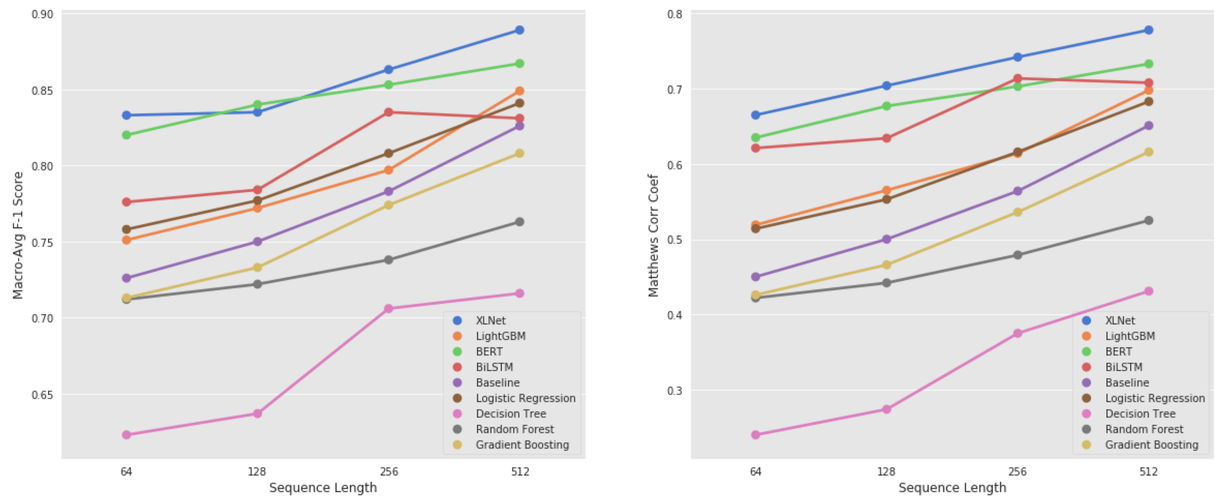




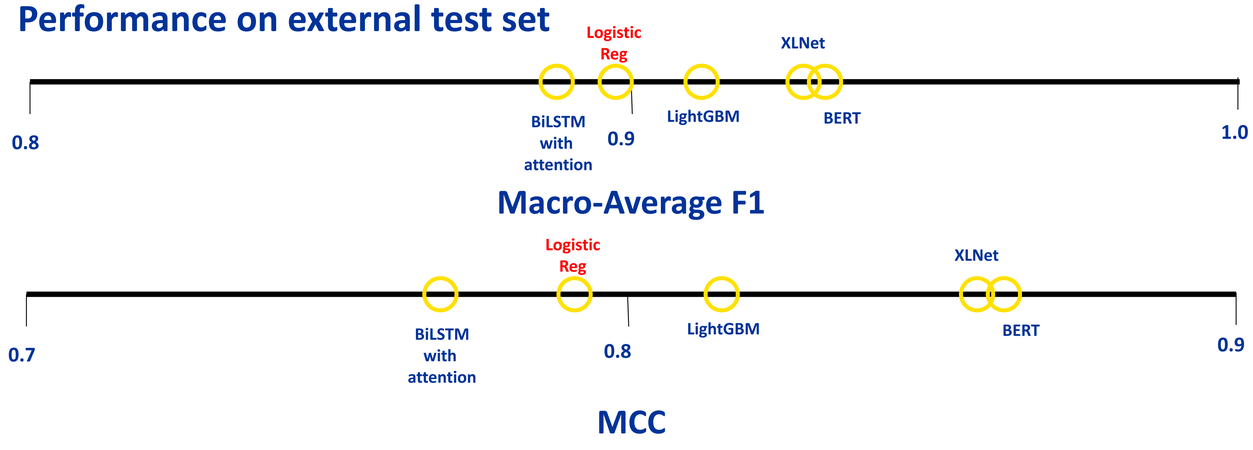
Modelling and Results

The training dataset was undersampled and contained 6386 articles. The testing was performed on two datasets, test (3436), and external test (1385), with the external test being a dataset from another news source. On the test set, XLNet performed the best, and on the external test set, BERT was the best model.

Performance on test set



Performance on external test set



Conclusions and Recommendations

BERT and XLNET are the best performing models. However, Logistic Regression has a comparable performance and does not lose out on explainability. Hence, we would recommend to use Logistic Regression in production.

Acknowledgments

We would like to thank Dr. Daniel Preotiuc-Pietro, Kai-Zhan Lee and Dr. Smaranda Muresan for providing us an opportunity to work on the project and guiding us throughout the project.

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237.