Event Isolation Using Deep Bidirectional Transformers

Data Science Institute
COLUMBIA UNIVERSITY

Antonio L. Ruiz, Yu Han Huang, Jacob Klein
Emily Cassell, Deepak Ravishankar
Industry Mentor: Dr. David Rothschild
Faculty Supervisor: Dr. Andrew Gelman

Data Science Capstone Project
with Microsoft Research

Project Purpose

Microsoft implemented a pipeline to systematically scrape and process news stories on a daily basis. Their current system can identify the top fifty news stories daily, while also collecting metadata such as the articles' publisher and prominence (where it is located on the page). Our analysis enables users to discover a myriad of insights related to news networks and the flow of information. The news article clusters map networks of shared articles and can be used in tracking which organizations choose to publish articles on specific topics. By contributing to the study of these topics, this team is contributing to the general understanding of how digital media is generated and shared.

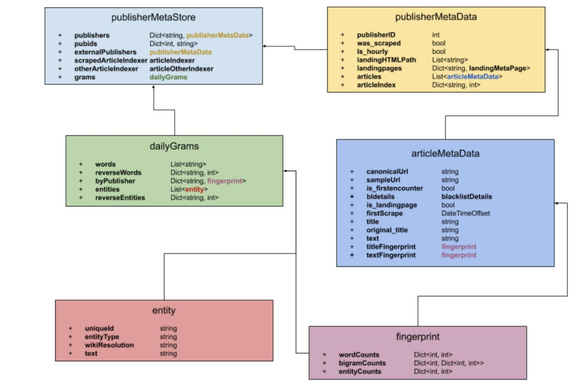


Figure 1: Original data pipeline (by Microsoft Research team)

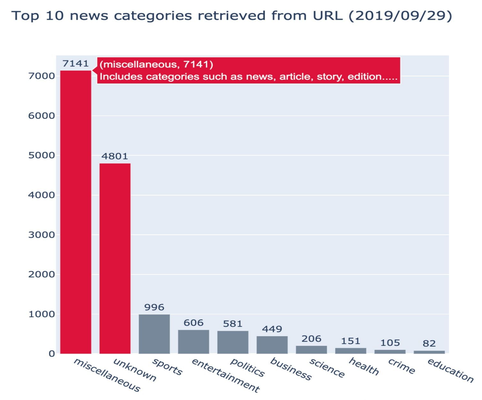


Figure 2: Categories obtained from URLs show us an empirical proof of news misclassification.

Methodology

The group's process is as follow: First, categories are obtained from the URLs of the articles through our self-designed architecture. Second, embeddings are computed using BERT (Bidirectional Encoder Representations from Transformers) on said articles' summary text. Next, embeddings are used to create cluster centroids and map the remaining articles to the high level clusters (categories). Finally, the best clustering separation (given by the Calinski Harabasz Score) is calculated for every individual high level cluster using grid search to retrieve optimal low level clusters.

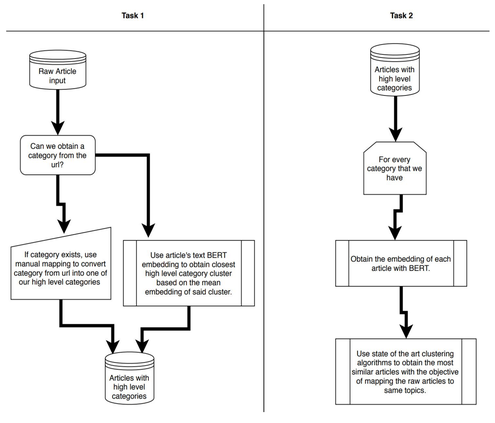


Figure 3: Methodology flowchart.

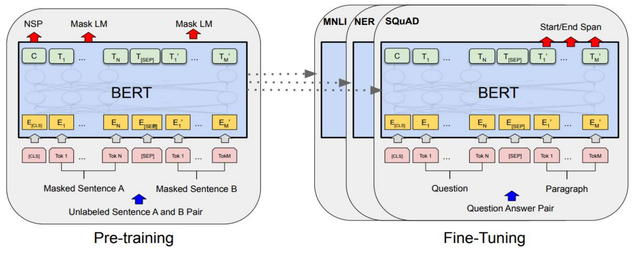


Figure 4: Pre-training and fine-tuning procedures for BERT. (1)

Results

The group utilized MTurk to determine article similarities and calculated a score based on the proportionality of articles that had higher within-cluster similarity scores than between-cluster similarity scores. The methodology and results are described below:

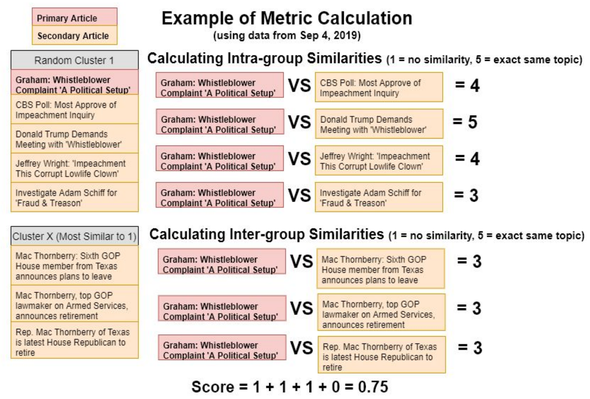
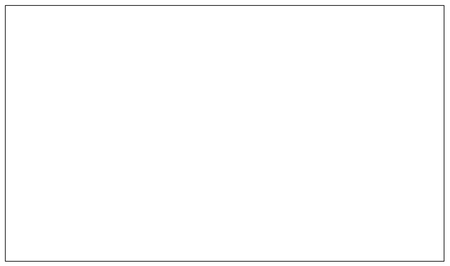


Figure 5: Walkthrough of calculating the similarity metric.



$$score = \sum_{i \in \mathcal{C}t} \frac{(\mathcal{S}_i)}{|\mathcal{C}t|} \mid \mathcal{S}_i = \sum_{j \in \mathcal{S}_i} \frac{W_i}{|\mathcal{I}|} \quad \text{where} \quad \mathcal{W}_i = \begin{cases} 1 & \text{if } \mathcal{I}a_i \geq \mathcal{I}b_i \\ 0 & \text{if } \mathcal{I}a_i < \mathcal{I}b_i \end{cases}$$

Where $\mathcal{I}a_{i,j}$ and $\mathcal{I}b_{i,j}$ are the intra-cluster and inter-cluster score for test i and article j; and $\mathcal{C}t$ the cluster tests.

Figure 6: Score calculation formula and final results.

Conclusion & Future Recommendations

In collaboration with Microsoft Research, the team presented a deep bidirectional transformer system for cluster based news classification. Considering the overwhelming volume of online news available every day, the group successfully deployed the BERT-large-cased model with BERT architecture to generate improved embeddings for the larger latent manifold data representation. Furthermore, they constructed a hierarchical clustering model architecture to improve cluster accuracy and create more meaningful low-level event-wise news clusters given the high level category-wise clusters. The unsupervised clusters were then examined through a mechanical turk process to evaluate the performance of the clustering algorithm. This is one of the first implementations to apply such a technique to news article clustering.

Acknowledgments

This project would not have been possible without the help and guidance from Ling Dong, Baird Howland, Markus Mobius and our project mentor Dr. David Rothschild.

References

(1) Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL

(2) Koppula, Hema Swetha, et al. "Learning URL Patterns for Webpage De-Duplication." Proceedings of the Third ACM International Conference on Web Search and Data Mining - WSDM 10, 2010, doi:10.1145/1718487.1718535.