Event Isolation Using Deep Bidirectional Transformers

Data Science Institute COLUMBIA UNIVERSITY

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)

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.

procedures for BERT. (1)

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

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

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:

	Primary Article Secondary Article	Example o
	Random Cluster 1	Calculating Int
Grat	nam: Whistleblower	Graham: Whistleblov
CBS	Poll: Most Approve of achment Inquiry	Graham: Whistleblo Complaint 'A Politica
Dona	ald Trump Demands	
Jeffr	ey Wright: 'Impeachment Corrupt Lowlife Clown'	Graham: Whistleblov Complaint 'A Politica
11113		
Inve: 'Frau	stigate Adam Schiff for id & Treason'	Graham: Whistleblo Complaint 'A Politica
Inve 'Frau Clus Mac Hous anno	stigate Adam Schiff for id & Treason' ter X (Most Similar to 1 Thornberry: Sixth GOP se member from Texas bunces plans to leave	Graham: Whistleblor Complaint 'A Politica) Calculating In Graham: Whistleblo Complaint 'A Politic
Investigation of the second se	stigate Adam Schiff for Id & Treason' ter X (Most Similar to 1 Thornberry: Sixth GOP se member from Texas bunces plans to leave Thornberry, top GOP naker on Armed Services, bunces retirement	Graham: Whistleblor Complaint 'A Politica Calculating In Graham: Whistleblo Complaint 'A Politic Graham: Whistleblo Complaint 'A Politic

rigure 5: waikthrou similarity metric.

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

- of

Data Science Capstone Project with Microsoft Research

f Metric Calculation data from Sep 4, 2019)				
a-group Similarities (1 = no similarity, 5 = exact same topic)				
er Setup' VS	CBS Poll: Most Approve of Impeachment Inquiry	= 4		
er Setup' VS	Donald Trump Demands Meeting with 'Whistleblower'	= 5		
er Setup' VS	Jeffrey Wright: 'Impeachment This Corrupt Lowlife Clown'	= 4		
er Setup' VS	Investigate Adam Schiff for 'Fraud & Treason'	= 3		
er-group \$	Similarities (1 = no sim	ilarity, 5 = exact same topic)		
er Setup' VS	Mac Thornberry: Sixth GOP House member from Texas announces plans to leave	= 3		
er Setup' VS	Mac Thornberry, top GOP lawmaker on Armed Services, announces retirement	= 3		$\sum (S_i) + c \sum W_i = \int 1$
er Setup' VS	Rep. Mac Thornberry of Texas is latest House Republican to retire	= 3	score	$e = \sum_{i \in \mathcal{C}t} \frac{\langle \mathcal{C}t \rangle}{ \mathcal{C}t } \mathcal{S}_i = \sum_{j \in \mathcal{S}_i} \frac{\mathcal{H}_i}{ \mathcal{I} } \text{ where } \mathcal{W}_i = \begin{cases} 0 \\ 0 \end{cases}$
1 + 1 + 1	+ 0 = 0.75		Where 2	$\mathcal{I}a_{i,j}$ and $\mathcal{I}b_{i,j}$ are the intra-cluster and inter-cluster score for test <i>i</i> and article <i>j</i>
ugh of calculating the				Figure 6: Score calculation form
tv metric.				final results.

(1) Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training

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.





