News Absorption Project

Data Science Institute
COLUMBIA UNIVERSITY

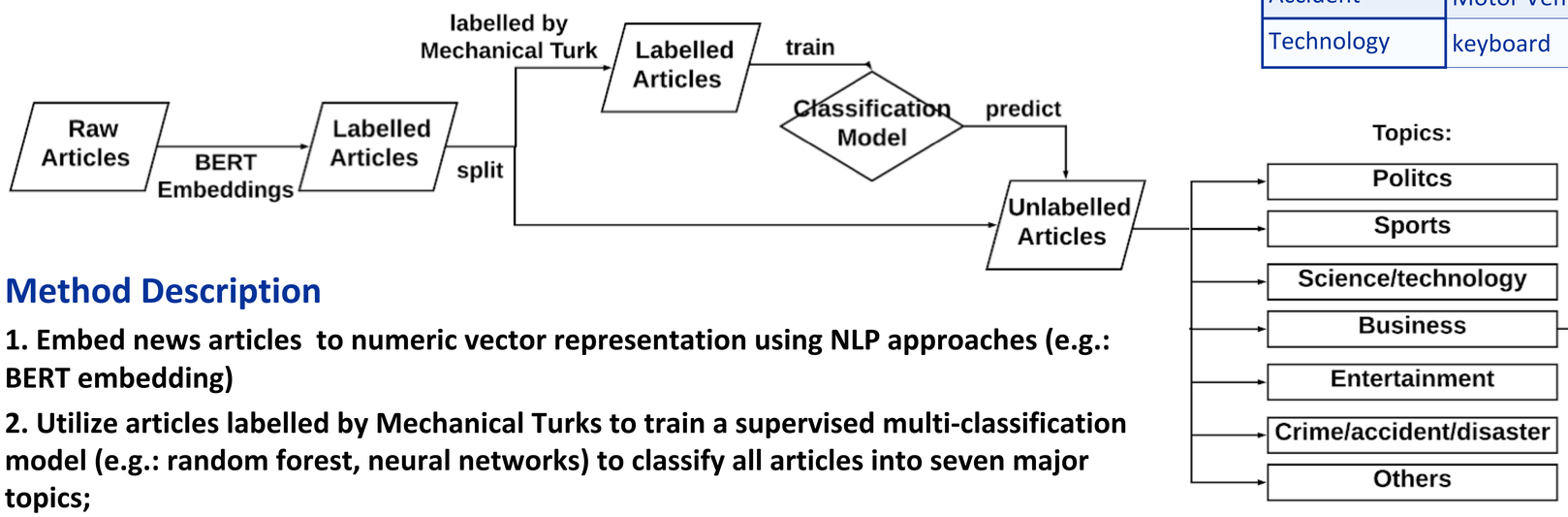
Authors: Harin Sanghirun, Thanaspakorn Niyomkarn, Jingwen Guo, Xuehan Zhao, Ze Chen
Industry Mentors: David Rothschild
Faculty Mentors: Andrew Gelman

Data Science Capstone Project with Microsoft Research

Project Background

This project aims to provide insights about the comprehensive and dynamic news ecosystem by clustering news articles. Over 200,000 articles will be first transformed to vector representation using NLP approaches and then more complex clustering algorithms which combine both supervised learning and unsupervised learning will be applied on them. The result will be clusters with keywords describing them in each predefined news topics. The result of news clustering will be used to support academic research and informs journalists and publishers.

Figure 1. project workflow



Method Description

1. Embed news articles to numeric vector representation using NLP approaches (e.g.: BERT embedding)

2. Utilize articles labelled by Mechanical Turks to train a supervised multi-classification model (e.g.: random forest, neural networks) to classify all articles into seven major topics;

3. Run clustering algorithms (community detection algorithm) on each of topics to obtain subclusters of each news topic. Small clusters will be merged to reduce noise.

4. Use Name Entity Recognition (NER) name each subclusters and explore patterns

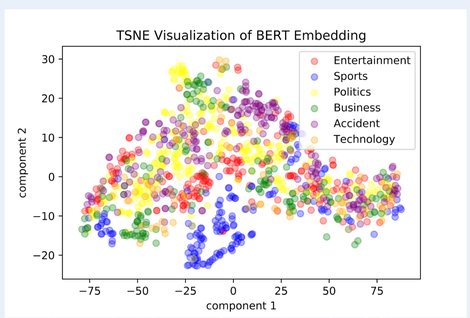


Figure 2. TSNE of BERT embeddings

Model	Precision	Recall	F1 Score
BOW	0.398	0.399	0.356
Entity Count	0.394	0.223	0.266
BERT	0.489	0.469	0.445
Doc2Vec	0.650	0.107	0.158

Table 1. performance metrics of embeddings

	Politics	Business	Entertainment	Sports	Accident	Technology	Others	Avg(micro)
F1 Score	0.78	0.75	0.68	0.93	0.78	0.12	0.06	0.72

Table 2. F1 score of Supervised Multi-Classification Model (Random Forest)

Results

Table 3. Keywords of Top Clusters in Each Type of News Topics using Tf-idf on Name Entities

Topics	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Politics	Internal politics: bipartisanship	Middle East, China, North Korea	Britain, Europe, Canada	Russia
Business	Amazon, Boeing	Saudi Arabia, India, Africa	Military, Army	China
Entertainment	Fox, instagram	Jeff Goldblum	Charlotte Murphy	Marijuana
Sports	World Cup	NBA	NFL(Football)	TBT(Basketball))
Accident	Motor Vehicle Crash	car accidents	Duck Boat Accident	Cumberland County
Technology	keyboard	Apple	NA	NA

Conclusions

In this project, we have explored the usage of various recent embedding models, supervised learning, and community detection to help extract patterns of news events from daily aggregation of online articles. We have found that, based on the F1 Score, BERT embedding performs the best. However, it is still challenging to interpret the resulting cluster. We will further tune and improve our clustering algorithms to reduce noises and make clusters more meaningful.

Acknowledgments

We thank Dr. Rothschild for constantly providing us with advice and help. We thank Markus Mobius for setting up the Virtual Machine. We thank Ling Dong for helping us with Mechanical Turk. We thank Baird Howland for clarifying our confusion about dataset.

References

Fortunato, Santo, and Darko Hric. "Community Detection in Networks: A User Guide." Physics Reports 659 (2016): 1–44. doi: 10.1016/j.physrep.2016.09.002

Microsoft Research, "Project Ratio", https://www.microsoft.com/en-us/research/project/project-ratio/, (accessed on Oct 16, 2019)