

Detecting Pharmaceutical Innovations in News Articles using Machine Learning

DSPG Team: Quinton Neville (Columbia University), Raghav Sawhney (Virginia Tech)

SDAD Team: Devika Nair, Gizem Korkmaz, Neil Alexander Kattampallil

Sponsor: Gary Anderson, The National Center for Science & Engineering Statistics (NCSES) at the National Science Foundation (NSF)

BACKGROUND

- Innovation is traditionally measured through surveys of selected companies (e.g., Business R&D and Innovation Survey (BRDIS)).
- While BRDIS measures innovation incidence, i.e., the number of innovating firms, NCSES is interested in exploring the possibility of leveraging non-traditional data to richer and complementary innovation measures.

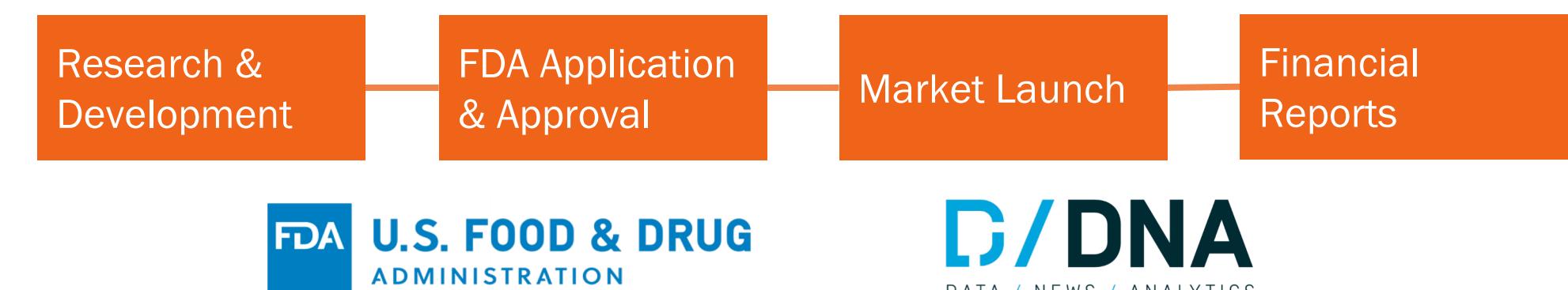
8 During the three years 2013 to 2015, did your company introduce:

- a. New or significantly improved goods (Exclude the simple resale of new goods purchased from other companies and changes of a solely aesthetic nature)? Yes No
- b. New or significantly improved services? Yes No

Figure 1. BRDIS survey question on innovation following OSLO manual definition of innovation [1].

- This project aims to develop machine learning methods to detect mentions of product innovations in news articles.

- Our current focus is on **Pharmaceutical Drugs and Medical Devices** that are heavily regulated by the **Food and Drug Administration (FDA)**.



FDA U.S. FOOD & DRUG ADMINISTRATION

D/DNA
DATA / NEWS / ANALYTICS

Figure 2. Abstract pipeline representing development of new drugs and medical devices.

- We use the **OSLO Manual definition of product innovation**, which requires that:
 - Product is new or significantly improved
 - Product is available on the market

DATA

- DNA Dataset:** We obtain about 2M news articles from Dow Jones about the pharma industry for years 2013-2018.

- The dataset contains over 30 variables on these articles, including publisher, subject, company codes, and publication date.

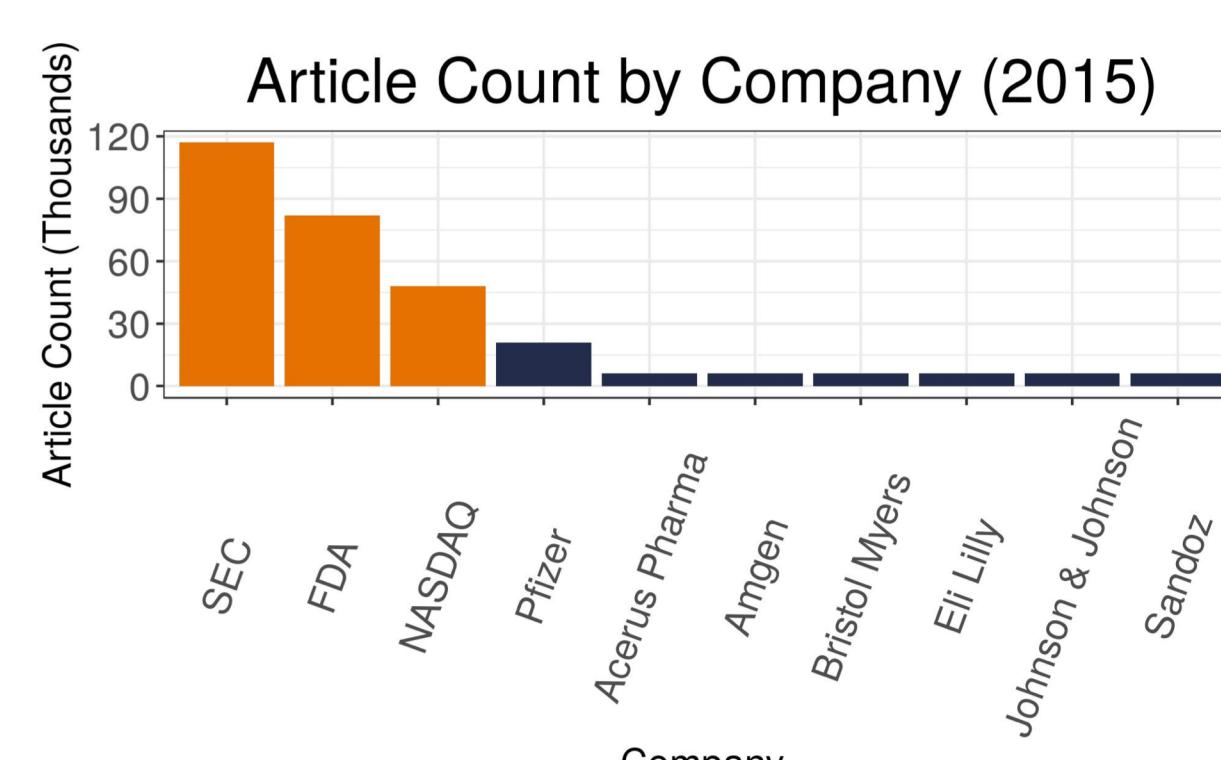
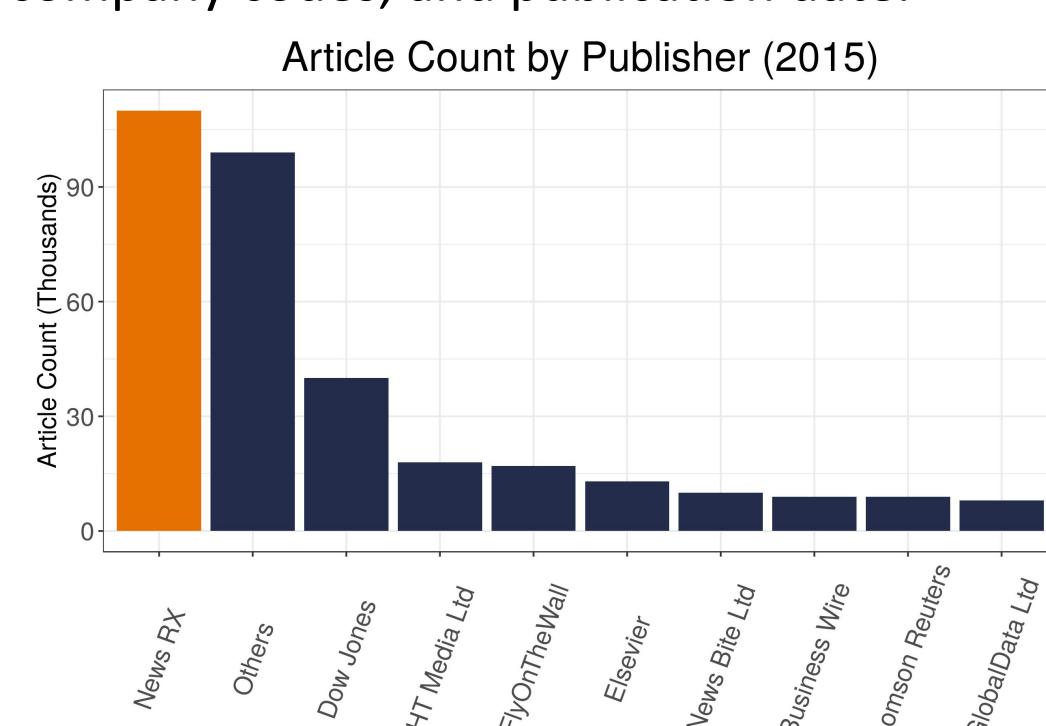


Figure 3. Top companies mentioned in DNA articles (2015). The top 3 are federal and financial entities, then followed by actual pharmaceutical companies.

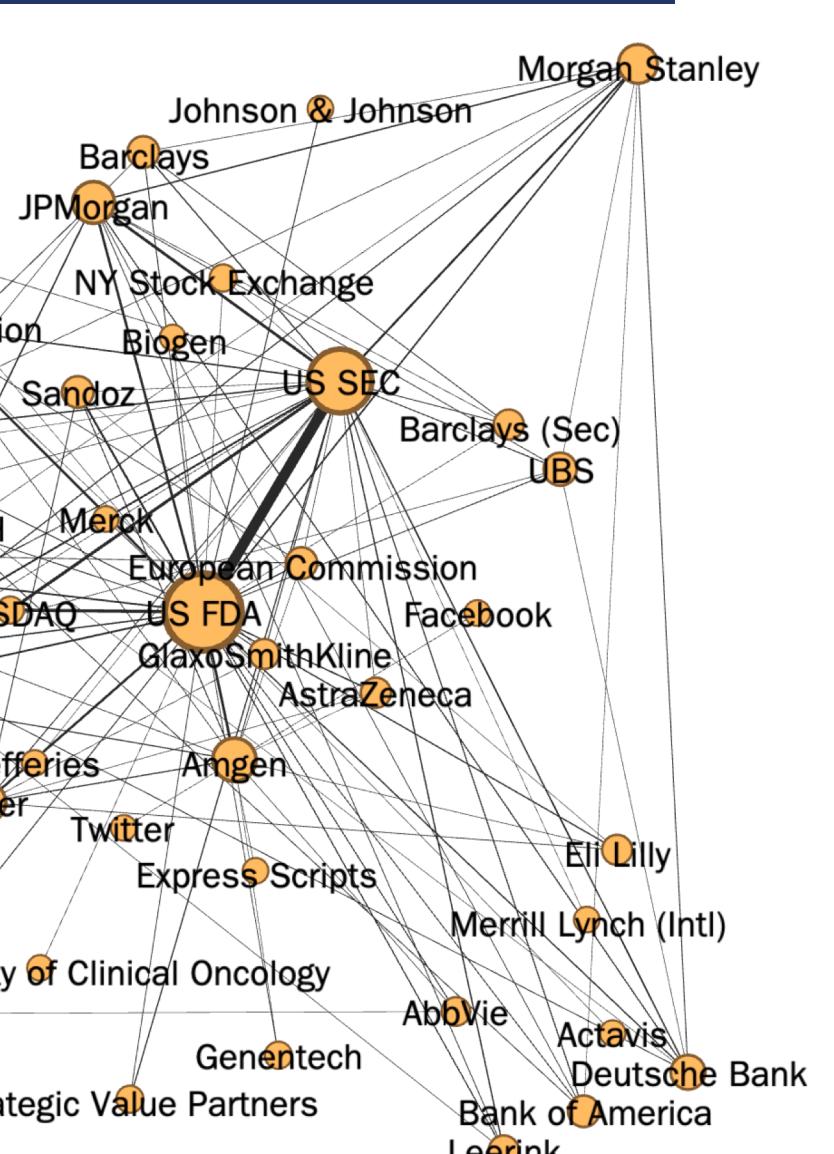
Figure 4. The network of DNA articles' company co-mentions in 2015.

We use a subset of DNA articles filtered using *innovation keywords* to generate the network.

For illustrative purposes we remove edges occurring less than 150 times.

The graph shows 43 nodes and 113 edges.

Node sizes and edge weights represent degree centrality and number of occurrences of co-mentions, respectively. Layout is random.

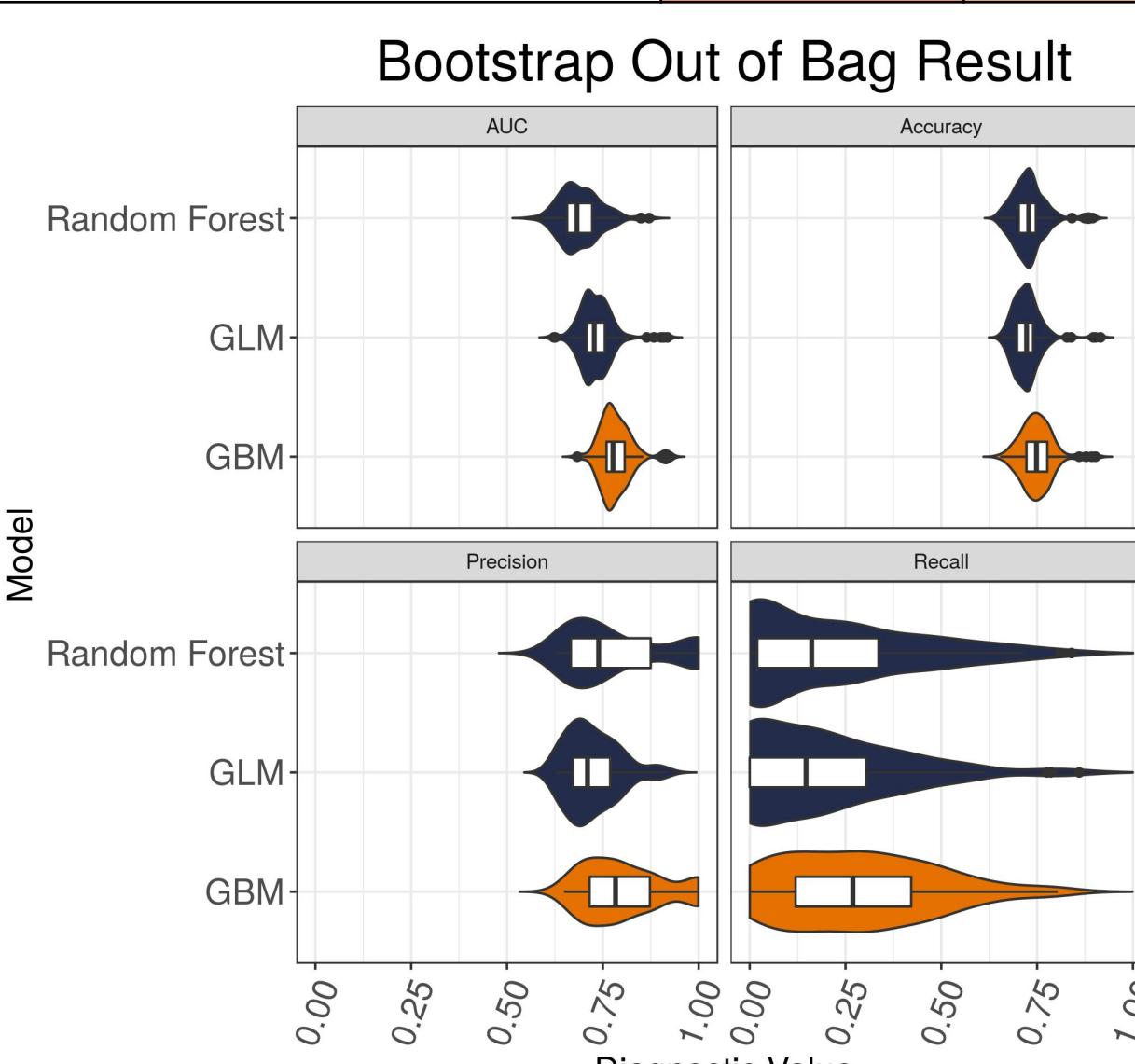
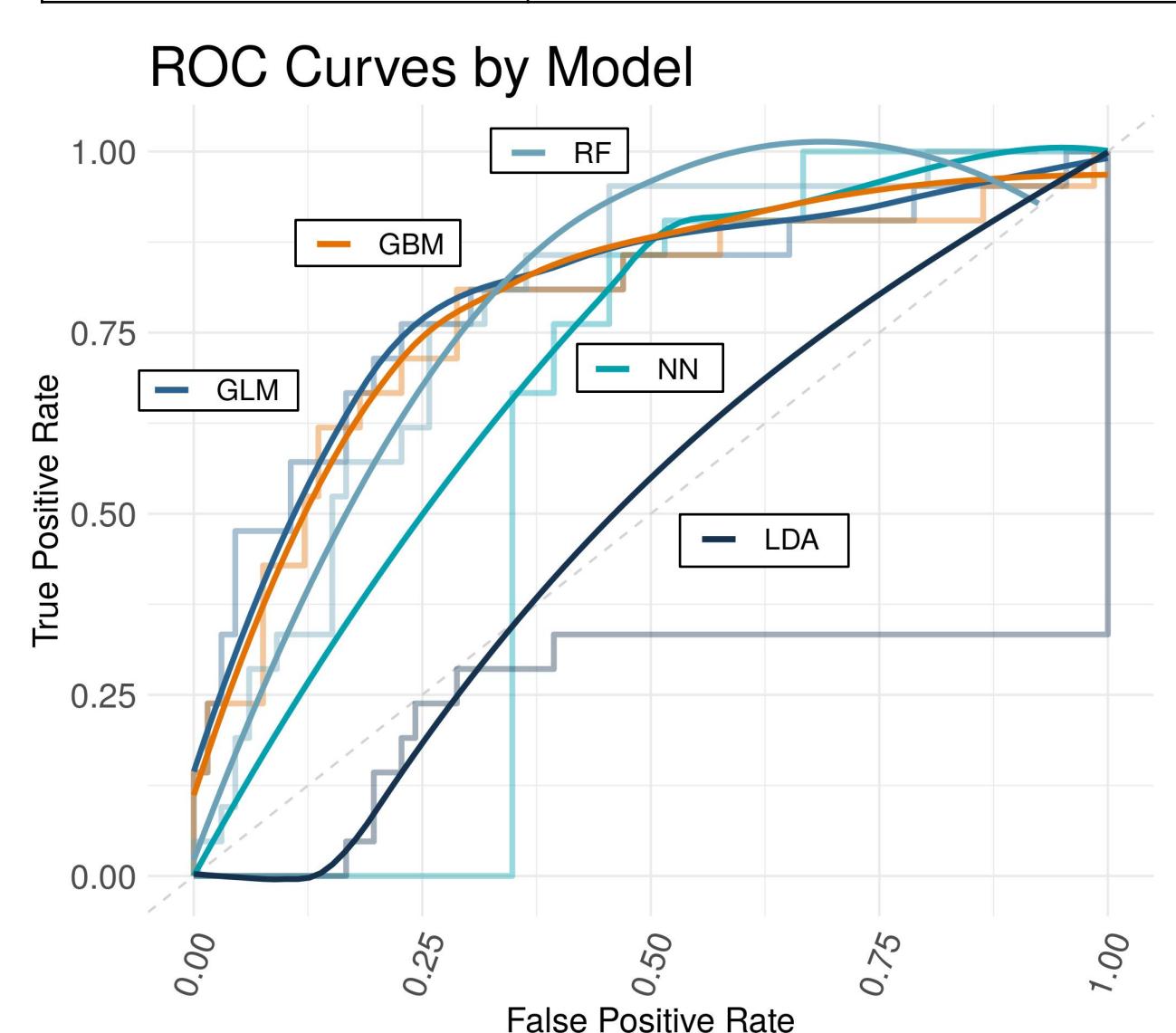


METHODS & RESULTS

Our research method is outlined below:

- Text processing
 - Bag of Words (BoW): Creating a document-term matrix with word frequencies
- Feature selection to reduce dimension: Least Absolute Shrinkage and Selection Operator (LASSO) applied to P-GLM, RF, GBM
- Model training using two labeling methods:
 - M1:** Manual labeling of 433 articles.
 - 137 (31.6%) articles mentioning new or improved drugs or devices
 - 49 (11.3%) of these articles' products were available on the market
 - M2:** Using DNA-generated subject-code label for "New Products or Services"
 - A random sample of 1,000 articles (50% with new product label)
- Determining cut-off probability thresholds using cost-optimization (using false-negative averse weights)
- Performance evaluation with metrics: ROC, accuracy, precision, recall and bootstrapped out-of-bag evaluation

Model	Description	Performance Metrics			
		M1: Manual Labeling Accuracy	M1: AUC	M2: DNA Labeling Accuracy	M2: AUC
Latent Dirichlet Allocation (LDA)	A bag-of-words model used for topic modeling [2]	0.76	0.78	0.54	0.52
Neural Networks (NN)	Represents data as computationally-weighted nodes organized into layers where learning occurs to build probabilistic edges. Pre-trained using Google AI's Bidirectional Encoder Representational Transformers (BERT) [3,4]	0.76	0.84	0.85	0.95
Penalized Generalized Linear Model (P-GLM)	Penalized logistic regression	0.77	0.78	0.66	0.78
Random Forest (RF)	Averages results of a bootstrapped 'forest' of independent decision trees	0.78	0.70	0.81	0.85
Gradient Boosting Machines (GBM)	A random forest where trees in succession learn from other trees	0.82	0.79	0.76	0.85



The models using manual labeling result in high accuracy but low recall values. This is due to the low proportion of innovation articles in the training set resulting in low predicted probabilities (of innovation).

FUTURE WORK

We have developed various machine learning methods to detect innovation in news articles. We observe the need for additional human-labelled training data. Our immediate next steps include:

- further investigating and testing the BERT-pretrained NN
- detecting company and product names in innovation articles
- comparing the findings (number of innovation per company) to FDA approvals
- expanding to other (less regulated, less data-rich) industries

REFERENCES

- [1] OECD/Eurostat (2018). OSLO Manual 2018: Guidelines for Collecting Reporting and Using Data on Innovation, 4th Edition, The Measurement of Scientific Technological and Innovation Activities, OECD Publishing, Paris/Eurostat, Luxembourg.
- [2] Lan, Man et al. "Text Representations for Text Categorization: A Case Study in Biomedical Domain." 2007 International Joint Conference on Neural Networks (2007): 2557-2562.
- [3] Siwei Lai, Lihegong Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15). 2015. AAAI Press 2267-2273.
- [4] Devlin, J; Chang, Ming-Wei; Lee, Kenton & Toutanova, Kristina. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." CoRR. October 2018.