Leveraging Google BERT to Detect and Measure Innovation Discussed in News Articles

Keyu Chen
School of Data Science
University of Virginia
Charlottesville, VA, USA
km5ar@virginia.edu

Benjamin Cosgro
School of Data Science
University of Virginia
Charlottesville, VA, USA
bcc5d@virginia.edu

Oretha Domfeh
School of Data Science
University of Virginia
Charlottesville, VA, USA
od9cw@virginia.edu

Alex Stern
School of Data Science
University of Virginia
Charlottesville, VA, USA
acs4wq@virginia.edu

Gizem Korkmaz
Biocomplexity Institute
University of Virginia
Arlington, VA, USA
gkorkmaz@virginia.edu

Neil Alexander Kattampallil
Biocomplexity Institute
University of Virginia
Arlington, VA, USA
nak3t@virginia.edu

2021 Systems and Information Engineering Design Symposium (SIEDS) | 978-1-6654-1250-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/SIEDS52267.2021.9483744

***Abstract*—In this paper, we leverage non-survey data (i.e., news articles), natural language processing (NLP), and deep learning methods to detect and measure innovation, ultimately enriching innovation surveys. Our dataset is composed of 1.9M news articles published between 2013 and 2018 acquired from Dow Jones Data, News, and Analytics. We use Bidirectional Encoder Representation from Transformers (BERT), a neural network-based technique for NLP pre-training developed by Google. Our methods involve: (i) utilizing Google's BERT as a binary classifier to identify articles that mention innovation, (ii) developing BERT's named-entity recognition algorithm to extract company names from these articles, (iii) leveraging BERT's question and answering capabilities to extract company and product names. As a result, we obtain innovation indicators, i.e., company innovations in the pharmaceutical sector.**

***Index Terms*—BERT, NLP, Dow Jones, innovation**

I. Introduction

Innovation is traditionally measured through surveys of selected companies such as NSF's Business R&D and Innovation Survey [1]. In this paper, we focus on product innovation, defined in OECD's Oslo Manual [2] as "new or improved good that differs significantly from the firm's previous goods and that has been available to potential users." Our goal is to use news articles and natural language processing (NLP) methods while leveraging Google's BERT [3] to measure business innovation, particularly within the pharmaceutical industry due to its high rate of innovation. The three main tasks we need to accomplish in order to achieve this goal are: (i) text classification to identify news articles that mention innovation, (ii) named-entity recognition (NER), and (iii) question answering (QA) to extract company names from these articles to identify the innovators. The main contributions and findings of this paper are listed below:

- We develop a classification model to detect potential innovation articles and perform extensive fine tuning for optimization. We fine-tune the cut-off value, learning rate and number of epochs. We calculate and present the performance metrics, i.e., precision, recall, and F1-score.
- We implement multiple iterations of the NER model in order to identify companies that are mentioned within the potential innovation articles detected by the classification model. We obtained a test accuracy of 90%.
- We develop a QA model with various degrees of accuracy (see Table II) and use it to extract company names from 20K potential innovation articles detected by the classification model. When we compare the results with the NER model, we find a 67% overlap between the NER and QA models for company extraction.

II. Related Work

Natural Language Processing (NLP) has seen great advances through the use of Google BERT. BERT, which stands for Bidirectional Encoder Representations from Transformers, is novel in that it trains and encodes on the context of words both to the left and right of a subject [3]. Since releasing BERT, many experts have started to see how it could apply to their field of expertise. Researchers at Korea University in Seoul developed what they called BioBERT and trained it on a combination of four corpora in Named-Entity Recognition (NER) and Question Answering (QA). These included the entirety of English Wikipedia, PubMed abstracts, PubMed full articles, and BooksCorpus. In testing, results from Wikipedia and BooksCorpus alone varied between 71% and 91% accuracy. However, including PubMed corpora improved results in across all tests. By including articles similar in nature to those which would be used in testing, the results understandably improved. In QA, results were less appealing, topping out at 57%, but adding PubMed did improve accuracy [8]. Other researchers found that BERT could be fine-tuned to cover different disciplines [9]. As a result, they created a model called SciBERT that is pre-trained on scientific text. Some of their tasks included NER and text classification that relied on a corpus comprised of computer science literature, as well as a corpus of biomedical literature. They found BERT

978-1-6654-1250-6/21/$31.00 ©2021 IEEE

useful in both corpora. In the case of the biomedical corpus, researchers were able to obtain as high as 90% accuracy in NER. In comparison to BioBERT, SciBERT was able to further improve on accuracy in testing on the same corpora. In text classification, their model was able to obtain an accuracy as high as 85% in a separate corpus [9].

III. Data

In this paper, we focus on the pharmaceutical industry (drugs and medical devices) which is heavily regulated by the Food & Drug Administration (FDA). The regulation aspect allows us to use publicly available FDA databases to identify companies that are truly coming out with new products. Our data consists of 1.9M news articles published between 2013 and 2018 acquired from Dow Jones Data, News, and Analytics (DNA). Here we focus on 360K observations from 2013. The dataset includes variables such as the publisher, publication date, and companies mentioned (codes and names) in each article.

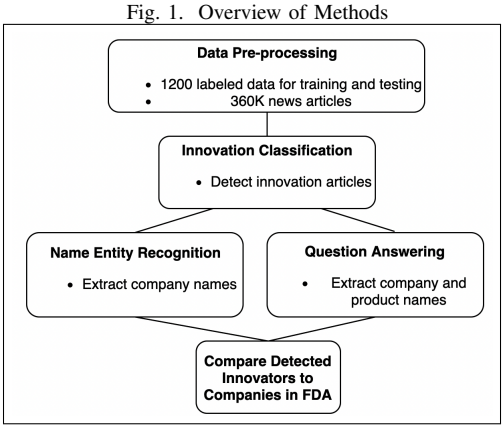
Labeled News Articles

To train our models, we use a randomized labeled sample set of 1,200 articles with response variables that indicate: (i) whether the article mentions an innovation or not, (ii) the name of the company/innovator, (iii) the launched product. Our innovation criteria includes product launches but excludes FDA approvals and patents as these do not imply that the products are available in the market (per our definition). After numerous reviews, we identified 32 articles, less than 2.7% of all sampled articles, that mentioned an innovation. This is a challenge for our classification model since these algorithms often perform better when positive and negative observations are evenly distributed in the training and test sets.

IV. Methods

We use multiple methods summarized in Figure 1 to develop innovation indicators. We use the labeled dataset to train our models, starting with our classification task. After the optimization of the classification model's hyper-parameters and detection of potential innovation articles, we run QA and NER algorithms to extract company names from these articles. Our goal is to obtain innovation counts by company for 2013 and to compare the list of potential innovators identified using our methods to the companies found in FDA databases.

Fig. 1. Overview of Methods



A. Innovation Classification

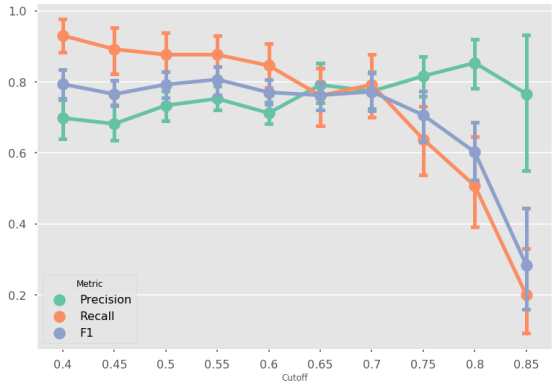
The first step in this process is to identify articles which contain information about an innovation. In its simplest form, this is a binary classification problem. Those articles flagged by our innovation classifier will be passed on to the subsequent models which will extract relevant details such as company name, product name.

1) Pre-processing: Utilizing BERT requires a number of pre-processing steps to be taken before a single byte can be fed into the model. One of the incredible advances BERT has made in the field of NLP is being able to derive context across varying sizes of text input. The limit to this is that all batches of text must not exceed a token limit of 512. Article titles were extracted and subsequently the body of each article was appended to the title in a single text string. BERT then requires special tokens to be added to the beginning and end of each string of text it takes in. From here, each string of text is converted into symbolic tokens that BERT has an extensive working knowledge of. Short and commonly used words have their own embeddings. The rest are split grammatically such that BERT can recognize prefixes and suffixes common in the English language. These strings were then pre-processed by the BERT model, tokenized, and clipped to 512 tokens. The rationale for doing this follows standard journalistic practice which is to include the most relevant information earlier on in the article because there is no guarantee the reader makes it to the end. Additionally, earlier rounds of testing indicated that the title of an article alone contained a wealth of information related to our task. Therefore, we took each article's title plus as many characters from the beginning of the article on before reaching BERT's limit.

2) Class Imbalance Problem: This refers to the challenge regarding training a model to not simply expect every observation to come from the dominating class. In our case, we find that a baseline model which classifies each and every article as not containing an innovation would reach an accuracy of over 97%. This can be incredibly misleading when it comes to tuning the model and even more so when it comes to drawing real-world conclusions based on said model. To address this issue, we used a training set that contained a $50/50$ split of innovation and non-innovation articles. This is a common technique used to bypass the significant class imbalance problem. This does, however, dramatically reduce the size of our training set. This is acceptable in theory because BERT does not have to be trained from scratch. Upon its creation, BERT was trained on a representative corpus of the English language. When training BERT for a specific task, such as identifying articles which contain information about an innovation, a technique called transfer learning is utilized. Transfer learning involves simply re-training the final layers of the network on the specific problem. A smaller, more targeted training set can be sufficient for such a task. We prove this method is acceptable in practice as well by repeating this process multiple times with different random samples of non-innovation containing articles. The fact that the spread of each

metric remains stable in Figure 2 across random subsets of negative training examples confirms this.

Fig. 2. Classification Model Metrics: Precision, Recall, and F1-score



3) Training BERT: Since the majority of the BERT network was trained to contextually interpret the English language in a task-agnostic manner, we were able to simply replace the final layer with a single sigmoid node in order to accomplish our task. The sigmoid node represents the logistic function. Often used to generate a probability estimate, the sigmoid node in this case will provide us with an estimated probability that the article contains information about an innovation.

4) Hyperparameter Tuning: There are a number of different hyper-parameters to identify and tune for this model. The Adam optimizer was selected for its versatility, widespread industry use, and adaptive learning rate capabilities. The model's learning rate and the number of epochs it was trained for were decided upon using a manual grid search process in which we identified optimal values. Due to the nature of this problem, our goals in tuning were to maximize the values of the precision, recall, and F1 metrics.

Additionally, we tuned our method for the cut-off value used in our binary classification step. The default value is traditionally 0.5 but we would rather gather as many innovation containing articles as we can while also picking up a minimal number of mis-labels along the way. In this vain, we tuned our cut-off value between 0.4 and 0.85. Based on Figure 2, we concluded 0.7 would be the optimal cutoff value for this task.

B. Named-Entity Recognition (NER)

The next step was to process the potential innovation articles detected by the classifier using BERT's Named-Entity Recognition (NER) to extract company names.

1) Training: To train the NER model, we use Conll-2003 dataset [11] – an existing pre-labeled news article dataset. In addition to utilizing this set for training, we decided to utilize BERT's uncased pre-trained model to start with before training further. BERT's pre-trained models come with a dictionary of words that are already assigned to the four categories which BERT uses to identify words. These are location [LOC], organization [ORG], person [PER], and miscellaneous [MISC]. The model is built to output a list of words and a corresponding list of tags identifying what the model has identified it as. For example, a sentence reading "John Hancock signed the Declaration of Independence in Philadelphia." would output:

['John', 'Hancock', 'signed', 'the', 'Declaration', 'of', 'Independence', 'in', 'Philadelphia']

with the corresponding list:

[B-PER, I-PER, O, O, B-MISC, I-MISC, I-MISC, O, B-LOC].

As seen in the example above, the model marks the beginning of an entity with a B and every subsequent part with and I. The following segments identify what type of entity this has been identified as. In this case, PER is person, MISC is miscellaneous, and LOC is location. Any O denotes a word that did not require recognition likely due to either not being a noun or not being significant.

By pulling these words into the model, BERT starts with a considerable advance in textual analytical abilities. BERT also offered the option to use cased or uncased models. Due to the fact that branding for many companies involves use of lower cased words, we decided to use the uncased version to decrease bias against uncased company names. Much of this model was adapted from an article written by Tobias Sterbak [11]. Upon training the BERT pre-trained uncased model with the Conll-2003 dataset, we obtained a validation accuracy of 97.69% at 3 epochs.

2) Testing and Refinements: We used 32 articles that were labeled as mentioning an innovation in the labeled news article dataset, and extracted companies using the pre-trained BERT mentioned above. When outputting a list from the articles of our study, we only included items marked as an organization [ORG] as our goal is to extract names of companies. Once we extracted a list of companies for each article with NER, we used fuzzy matching to compare it to the manually identified company names in the labeled dataset as well as to the company names that are provided with the DNA dataset [14]. For the matching, we used the full company names rather than single word matching which resulted in higher performance.

In addition, created a dictionary of industry descriptors and removed these from the company names to improve the matching results. The words we identified as common are listed as: *"Industries, Inc, Incorporated, Group, Labs, Laboratories, Corporation, Corp, Companies, Medical, Pharmaceuticals, Technologies, Organization."* This list fit the pharmaceutical industry best but could easily be adapted to other industries. For example, the food and beverage industry might want to include words like "foods," "bakery," and "confectionery."

For each article, we used (i) the title, (ii) the leading paragraph, (iii) the title and the leading paragraph combined to extract organization names for comparison.

3) Performance: Table I presents the matching accuracy rates for the various inputs mentioned above. We observe that using title only results in lower match scores for both label comparisons, and we obtain the highest rates when we use both the title and the leading paragraph to extract company names. When we match the extracted company names to the companies identified manually, we obtain an accuracy of 90%, and when we use the company names provided by the DNA dataset, we obtain a 75% match rate.

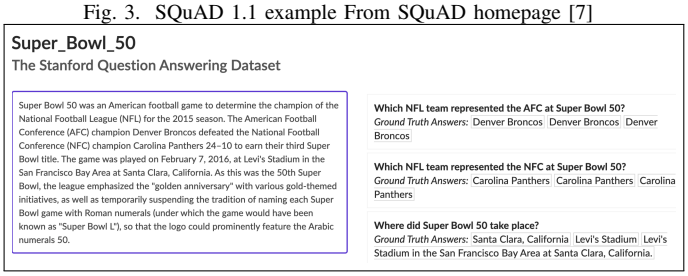
TABLE I
PERFORMANCE OF NER: COMPANY NAME MATCHING RATES

Subset of Article Processed	Pre-labeled Column for Comparison	
	Hand-Labeled Innovator	*Companies from DNA*
Title Only	62.5%	53.1%
Paragraph Only	87.5%	65.6%
Title and Paragraph	90.6%	75%

C. Question Answering (QA)

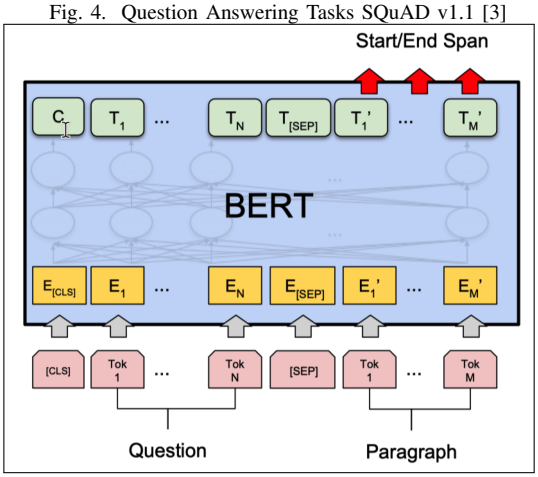
In this step, our goal is to retrieve key information (company and product names) from each of the potential innovation articles detected by the classifier using BERT's Question Answering (QA) [5]. To use BERT's QA, we need to feed a passage of text with a maximum of 512 tokens into the algorithm, then ask a question based on the information we would like to retrieve. See example in Figure 3. In our case we ask: "what's the name of the company?" or "what's the new product?" for each article. The result will be a piece of text from the passage it passes through.

Fig. 3. SQuAD 1.1 example From SQuAD homepage [7]



This algorithm is based on the Transformers Architecture and currently is the state-of-the-art algorithm for performing QA models. Figure 4 illustrates the process of passing through a paragraph, asking a question, and then getting an answer based on that text. More information can be found in [7].

Fig. 4. Question Answering Tasks SQuAD v1.1 [3]



1) Pre-trained model – Hugging Face: We used the pre-trained model from Hugging Face rather than training our own based on the original BERT paper [3]. The authors state that "a distinctive feature of BERT is its unified architecture across different tasks. There is minimal difference between the pre-trained architecture and the final downstream architecture." [3]

Below are the traits that define the uncased BERT-Large model with whole word masking and fine tuned on SQuAD (details about the model can be found in [6]):

- Pre-trained on BookCorpus, which includes 11,038 unpublished books and Wikipedia articles in English.
- Whole word masking variant of BERT-Large.
- Uncased indicating that it does not take the capitalization into account. Thus, upper or lower-case versions of the same word will be treated the same in this model.
- Fine-tuned on SQuAD 1.1 dataset (Stanford Question Answering Dataset [7]).
- Model configuration [6]: 24 layers, 1024 hidden dimensions, 16 attention heads, and 336M parameters

We used SQuAD 1.1 due to its easy implementation relative to SQuAD 2.0. SQuAD 1.1 is "a reading comprehension dataset, consisting of questions posed by crowd-workers on a set of Wikipedia articles, where the answer to every question is a segment of text or span from the corresponding reading passage [7]." One limitation is that SQuAD 1.1 is not capable of determining whether or not to answer a question. In the case that there is not answer within the text provided, SQuAD 1.1 will always return its best guess for an answer [7].

2) QA Implementation and Testing: As done with NER, we used the 32 innovation articles in the labeled news article dataset, and extracted companies using the pre-trained QA model described above. BERT has a limitation of maximum length of 512 tokens, so for each article we used: (i) the title, and (ii) the title and the leading paragraph combined (first 512 tokens).

We used a total of 12 questions addressing company and product names including various versions of:

- (company) What is the company name?
- (company) Who is the innovator?
- (company) Which company announced the product?
- (product) What is the new product?
- (product) What is the new device?

Each question gave us one answer per article, which we then matched with our hand-labeled company and product names.

In our initial tests, when we compared the accuracy of the models with "title" and "title+leading paragraph" tokens, we obtained a higher performance for the latter as not all the innovations were mentioned in the article titles. Moreover, we tested and compared the time needed to process the data for the two inputs. Even though the title included far fewer tokens, both inputs took the same amount of time to process (6 seconds for each article). Thus we decided to focus on the "title+lead" tokens for each article in our 2013 dataset as it would take the same amount of time and improve our accuracy rate.

3) Performance: To evaluate the performance of the QA model, we compared the obtained answers from title and title+leading paragraph inputs to the hand-labeled company and product names. Here, we used three strategies once we obtained answers to each 12 questions (3 questions for companies, 9 questions for products):

I. we compared every single answer obtained from the title (12 individual answers) and title+lead (12 individual answers) inputs to the respective ground-truth in the labeled dataset;

II. we combined the answers from title and title+lead inputs

to the company-related questions (6 answers combined), and to the product-related questions (18 answers combined), and compared these two separate answer strings to the respective company and product names in the labeled dataset; and

III. we obtained answers only from title+lead (to reduce run time), and combined the company-related answers (3 answers combined), and product-related answers (9 answers combined), and compared these two strings to the respective company and product names in the labeled dataset.

TABLE II
PERFORMANCE OF QA: COMPANY AND PRODUCT NAME MATCHING

	Company Names	***Product Names***
I. Individual answers **(title and title+lead)**	78%	81%
II. Combined answers **(title and title+lead)**	97%	91%
III. Combined answers **(title+lead)**	94%	91%

Table II summarizes the fuzzy matching results using the answers with the highest matching probability for each evaluation strategy. When we matched individual answers to the company and product names, we obtained the highest accuracy values as 78% (25/32) for company and 81% (26/32) for product names. When we compared the combined answers from title and title+lead inputs, we obtained 97% and 91% accuracy rates, for company and product names, respectively. There was a slight decrease in the accuracy rate (3% for company names) when we used combined answers only from "title + lead" inputs, decreasing the computing time by half. Since QA is a computationally intensive task compared to NER, we suggest bearing such a minimal decrease in accuracy to save in computational resources.

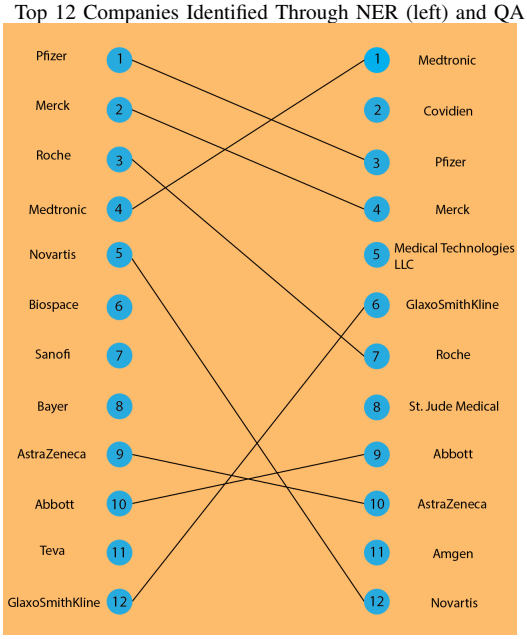
V. RESULTS

In this section, we present our final results. The first subsection presents the comparison between NER and QA methods using: (i) our labeled dataset, (ii) Dow Jones news articles (DNA) published in 2013. Next, we use our classification, NER, and QA methods to identify innovation articles from the 2013 DNA dataset, and extract companies that are mentioned in these articles. We compare these companies to companies in Food and Drug Administration (FDA) databases, and present our findings.

A. Comparison of NER and QA

1) Labeled Dataset: Here we compare the NER and QA models giving us insights about which method to choose when there is a lack of ground-truth (i.e., a labeled dataset). To do so, we took the list of organizations identified by the three company-related questions in the QA model and matched them to the organizations identified by NER. Using our 1,200 labeled article dataset, we were able to obtain 78% accuracy by choosing the highest match as our output. Additionally, we compared the NER organizations to the answers from all questions asked in the QA step including the questions relating to products. However, this only obtained 59% accuracy when using the highest match as output. We followed this by looking at the original list of answers identified by the three questions company-related questions in the QA model and ignored the NER answers. We then approached it from a method of boosting by using this method of choosing whichever company appeared most in the three questions. From this, an accuracy of 78% similarly appeared.

Fig. 5. Top 12 Companies Identified Through NER (left) and QA (right)



2) Dow Jones News Articles (2013): Here we use the 360,000 news articles obtained from DNA database. We use the classification method to identify 65,000 potential innovation articles. NER was used to extract the company names from these articles. Due to computational limitations, we were only able to run QA on the first 20,000 articles (out of 65,000) that were classified as innovation in the 2013 dataset. We compared the most frequent names extracted by NER to the most frequent company names identified as innovators in the first 20,000 articles identified by the QA. Figure 5 illustrates the top 12 companies identified by each method. Within the top 12 as well as in the top 50, we find an overlap of approximately 67% between two methods.

B. Comparison to the FDA Databases

1) National Drug Code (NDC) Database: We compared the potential innovators identified by our methods using the DNA 2013 articles to the companies listed in the National Drug Code (NDC) database provided by FDA [13]. We acquired the top 50 companies listed in the NDC database in 2013 and then compared them to our QA and NER results. Unfortunately, while the NER and QA seemed to correlate with each other well, the NDC data did not. Out of 50 companies listed by NDC as the top 50 innovators by new listings, only 3 made the NER list and only 1 (Johnson & Johnson) made the NER and QA lists. Currently, we do not suspect that our methods are not accurate, but that NDC is measuring

innovation in a different way. First of all, NDC database lists products that are already on the market, while our methods also capture newly approved and announced products even if they are not available to consumers yet. Second, while our top 50 list included big names in drug development such as GlaxoSmithKline, AstraZeneca, Pfizer, and Merck, NDC included names of manufacturers, labelers, and producers that were not commonly known or names that are not generally associated with drug development such as CVS and L'Oreal. We suspect that while our results may show the innovators in drug development, while names like CVS may indicate the production of generics and the smaller names may be the actual smaller producers that are subsidiaries of the larger companies, and these may not make be mentioned in the news articles.

2) FDA Approval Database: Finally, we compared the potential innovators in our list to the FDA database of newly approved drugs [12]. The comparison of the top 50 of their companies for number of approvals resulted in higher matches: 17 out of 50 companies appeared in the FDA database, and NER and QA lists. We obtained a match of 22 out of 50 companies from FDA database to wither the QA or the FDA lists. While 22 out of 50 matches were better than those with the NDC, it was not nearly the result we would hope for and would like to research more as to why we did not have as close of a match.

VI. Conclusion

Our methods with BERT showed that this tool can be properly applied to classification of innovation and determination of who an innovator is solely from news articles. However, while averaging about 80% accuracy in each step, better methods still are likely to appear as the field of Natural Language Processing (NLP) grows. As it stands, the NER phase of this study does little more than to confirm the answers from QA and could be dropped to reduce computing load, but this may prove more useful in future iterations of this model.

More computing capability and better training sets will likely improve these methods. A future study could use more labeled data in order to improve the training and a future labeled set for NER that includes biomedical companies would help. More research into the fine-tuning of these models can also be done. For now, this method may be the best method available and would complement the traditional method of company surveys.

Other next steps can include searching for a dataset that might align further with our metrics for innovation in order to make a more clear comparison. Additionally, exploring why the FDA set did not match up with our dataset should be a priority. It is possible that our data is skewed by larger companies that have entire media departments to create publicity.

Further research should also be done in seeing how NER can be utilized. As seen in our results, NER and QA were close in their top 12 and top 50. If the hours of computing time can be avoided in attempting QA, this would make the model much more efficient.

VII. Acknowledgments

This material is based on work supported by U.S. Department of Agriculture (58-3AEU-7-0074) and the National Science Foundation (Contract #49100420C0015). Any opinions, finding, and conclusions or recommendations expressed in this material are those of the authors.

We extend our sincere gratitude to Dr. Gizem Korkmaz from the University of Virginia's Biocomplexity Institute for her consistent support and insights. We thank the University of Virginia School of Data Science for facilitating the success of this project and Prof. Gerard Learmonth for his invaluable guidance and for being our mentor. We would also like to thank Neil Kattampallil and Martha Czernuszenko for their support in bringing this project to fruition.

References

[1] NSF. (2020, Jul). Business R&D and Innovation Survey (BRDIS).National Science Foun-dation. Retrieved from https://www.nsf.gov/statistics/srvyberd/prior-descriptions/overview-brdis.cfm.

[2] OECD. (2018, Jan). Oslo Manual 2018. Organisation for Economic Co-operation and Development (OECD). Retrieved from https://www.oecd.org/science/oslo-manual-2018-9789264304604-en.htm.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv.org, 11-Oct-2018. [Online]. Available: https://arxiv.org/abs/1810.04805v1. [Accessed: 06-Apr-2021].

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," arXiv.org, 06-Dec-2017. [Online]. Available: https://arxiv.org/abs/1706.03762v5. [Accessed: 06-Apr-2021].

[5] P. Dwivedi, "Testing BERT based Question Answering on Coronavirus articles," Medium, 06-Apr-2020. [Online]. Available: https://towardsdatascience.com/testing-bert-based-question-answering-on-coronavirus-articles-13623637a4ff. [Accessed: 06-Apr-2021].

[6] Hugging Face. "Bert-Large-uncased-whole-word-masking-finetuned-squad." [Online]. Available: https://huggingface.co/bert-large-uncased-whole-word-masking-finetuned-squad. [Accessed: 07-Apr-2021].

[7] "SQuAD1.1," SQuAD - the Stanford Question Answering Dataset. [Online]. Available: https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/. [Accessed: 07-Apr-2021].

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," arXiv, 10-Sep-2019, [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1901/1901.08746.pdf . [Accessed: 07-Apr-2021].

[9] I. Beltagy, K. Lo, A. Cohan, "SCIBERT: A Pre-trained Language Model for Scientific Text," arXiv, 10-Sep-2019, [Online]. Available: https://arxiv.org/pdf/1903.10676.pdf . [Accessed: 07-Apr-2021].

[10] T. Sterbak, "Named-entity recognition with Bert," Depends on the definition, 24-Apr-2020, [Online]. Available: "https://www.depends-on-the-definition.com/named-entity-recognition-with-bert/" . [Accessed: 07-Apr-2021]

[11] Language-Independent Named-Entity Recognition (II). [Online]. Available: https://www.clips.uantwerpen.be/conll2003/ner/. [Accessed: 10-Apr-2021].

[12] "Drugs@FDA Data Files," U.S. Food and Drug Administration, [Online] Available: https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files . [Accessed: 17-Dec-2019]

[13] "National Drug Code Directory," U.S. Food and Drug Administration, [Online] Available: https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory . [Accessed: 31-May-2020]

[14] F. Arias, "Fuzzy String Matching in Python," DataCamp, [Online] Available: https://www.datacamp.com/community/tutorials/fuzzy-string-python . [Accessed 07-Dec-2020].