ESGBERT: Language Model to Help with Classification Tasks Related to Companies' Environmental, Social, and Governance Practices

Srishti Mehra, Robert Louka, Yixun Zhang

UCB MIDS W266: Natural Language Processing, July 2021

{srishtimehra, robertlouka, yixunz}@berkeley.edu

Abstract

Environmental, Social, and Governance (ESG) are non-financial factors that are garnering attention from investors as they increasingly look to apply these as part of their analysis to identify material risks and growth opportunities. Some of this attention is also driven by clients who, now more aware than ever, are demanding for their money to be managed and invested responsibly. As the interest in ESG grows, so does the need for investors to have access to consumable ESG information. Since most of it is in text form in reports, disclosures, press releases, and 10-Q filings, we see a need for sophisticated NLP techniques for classification tasks for ESG text. We hypothesize that an ESG domain-specific pre-trained model will help with such and study building of the same in this paper. We explored doing this by fine-tuning BERT's pre-trained weights using ESG specific text and then further fine-tuning the model for a classification task. We were able to achieve accuracy better than the original BERT and baseline models in environment-specific classification tasks.

1. Introduction

The importance of Environmental, Social, and Governance (ESG) issues has risen in prominence over the last decade. In the early 1990's fewer than 20 publicly listed companies issued reports that included ESG data; that number grew to almost six thousand by 2014 (Serafeim and Yoon, 2021). Regulations for SEC filings, Acts to follow certain standards for responsibility about Climate Change and Human Governance, and Investor and Shareholder support has driven the motivation for these disclosures.

There has been little research analyzing the non-financial information content (Kölbel et al., 2020, p. 8) in financial disclosures. The most common methods for analyzing the non-financial, narrative information content still remain a manual- or dictionary-based approach (Berkman et al., 2019; Matsumura et al., 2018; Reverte, 2016; Verbeeten et al., 2016; Clarkson et al., 2008; Cormier and Magnan, 2007, i.a.). Also, only a few studies concerning non-financial information focus on the actual narrative information content, and rather address the quantity of non-financial information published (Armbrust, Schäfer and

Klinger, 2020 p. 2; Hummel and Schlick, 2016).

Domain-specific BERT variations like FinBERT (Araci, 2019) and BioBERT (Lee and Yoon et al., 2019), that have been either fine-tuned or pre-trained on domain corpus instead of or in addition to the generic English language, have achieved great results in domain-specific classification tasks. The primary interest of this research is to harness that benefit for ESG specific text classification tasks. For the same, we study building an environment-specific variation of BERT by fine-tuning the pre-trained BERT weights using a Masked Language Model (MLM) task on ESG corpus and then further fine-tuning our model for the classification tasks to predict:

1. A change or no change, and

2. An upward or downward change (if any)

in environmental scores of companies using ESG related text in their 10-Q filings. We use a Sequence Classification that we fine-tune with our classification task.

2. Background

Accounting for Sustainability[1] is a project that aims to inspire action by finance leaders to drive a fundamental shift towards resilient business models and a sustainable economy. To do so the project publishes guides, case studies, blogs, reports and surveys, and hosts webinars. All of this material is available on their knowledge hub and is reflective of the opportunities and risks posed by environmental and social issues. These are what we chose as our ESG corpus to pre-train our BERT model on top of the English Wikipedia and BooksCorpus it has been trained on (Devlin et al., 2018).

In 2010, the SEC published an interpretive release on climate change-related disclosures “to remind companies of their obligations under existing federal securities laws and regulations to consider climate change and its consequences as they prepare disclosure documents" (SEC, 2010, p. 6297). Therefore, the company’s disclosures should not only consist of financial narratives but also contain information about the environmental aspects concerning the firm (Armbrust, Schäfer and Klinger, 2020). This is why we chose to use 10-Q filings as our input for the classification task.

Sustainalytics’ ESG Risk Ratings measure a company’s exposure to industry-specific material ESG risks and how well a company is managing those risks. This multi-dimensional way of measuring ESG risk combines the concepts of management and exposure to arrive at an absolute assessment of ESG risk. These risk scores are also broken down into environmental, social, and governance risks. Of these, for our research, we use the change in total environmental risk score for each

[1] https://www.accountingforsustainability.org/en/knowledge-hub.html

company quarter over quarter to indicate whether there was:

1. A change or no change, and

2. An upward or downward change.

3. Related Work

This section describes previous research conducted on domain-specific variations of BERT (3.1) and ESG related NLP research (3.2).

3.1 Domain-specific BERT variants

FinBERT (Araci, 2019) author explored pre-training BERT on Financial corpus based on their learning from a previous study by Howard and Ruder (2018) which shows that further pre-training a language model on a target domain corpus improves the eventual classification performance. They studied pre-training BERT on financial corpora and using that model on financial sentiment classification and saw improved results when compared to original BERT (pre-trained on generic English language corpora).

BioBERT (Lee and Yoon et al., 2019) authors similarly (similar architecture as followed by FinBERT) pre-trained BERT with Biomedical corpora in addition to the English language corpora it was already trained on. They went on to find that BioBERT largely outperformed BERT in a variety of biomedical text mining tasks.

3.2 ESG related NLP research

Armbhurst, Schäfer, and Klinger, 2020 studied the effect of the environmental performance of a company (as learned from MD&A sections in 10-K and 10-Q filings) on the relationship between the company's disclosures and financial performance. They found that the disclosures (textual information) did not predict financial performance, however, found evidence that environmental performance could be extracted from that textual information using NLP techniques.

Serafeim and Yoon, 2021 showed that ESG ratings predict future ESG news and market reactions to the news, particularly when there is disagreement amongst raters. This study is similar to our study in that it uses ESG scores (from TruValue, not Sustainalytics) to predict the news, whereas we are using information from public documents ("news") to predict scores. The news in their study was aggregated by TruValue using machine learning from a wide variety of sources.

4. Method

This section will be divided into BERT (4.1), pre-trained BERT weights on ESG specific corpus (4.2), and fine-tuning for the classification task mentioned in earlier sections (4.3).

4.1 Bidirectional Encoder Representations from Transformers

BERT (Devlin et al., 2018) is a pre-trained model that builds word representations learned through bi-directional tasks. It has a Masked Language Model (MLM) that predicts some portion of words that are masked and Next Sentence Prediction (NSP) that captures the relationship between two sentences which is not directly captured by language modeling. For finetuning, the BERT model is first initialized with the pre-trained parameters learned in the bi-directional approach, and all of the parameters are fine-tuned using labeled data from the downstream tasks. Each downstream task has separate fine-tuned models, even though they are initialized with the same pre-trained parameters (Devlin et al., 2018).

4.2 Pre-training on ESG specific corpus

BERT's pre-training procedure largely follows the existing literature on language model pre-training. For the pre-training corpus, they use the BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words).

Since the text in our research is also in the English language, we did not want to miss out on the benefit of pre-trained weights on such large English language corpora. Thus, we decided to further train these pre-trained model weights by building a Masked Language Modeling Task on top of BERT.

We use text from the Knowledge Hub of Accounting For Sustainability for our Masked Language Modeling task. These occur in the form of guides, case studies, blogs, reports, and surveys. We tokenized using BERT's WordPiece Tokenizer, masked 15% of the words, and learned to predict those masked words. In doing so, we updated the pre-trained weights of BERT per the language modeling task being run on the ESG corpus.

4.3 Fine-Tuning for Classification Task

In order to test our hypothesis of a domain-specific variation of BERT working better than that trained on generic language, we chose two classification tasks that we fine-tuned on. The classification tasks were to predict whether there was:

1. A change or not

2. Positive or negative change

in the environmental risk scores quarter-over-quarter of the companies we were experimenting with.

Input for Fine-Tuning with Classification

BERT takes up to 512 tokens as input. Since our input per company per quarter was an entire 10-Q document, we needed a way to choose 512 tokens for each 10-Q document. Since 10-Q reports contain small portions that address environmental factors, our approach was to extract the most relevant (to environmental factors) sentences and choose 512 tokens from those. In order to do so, we needed a method to order all sentences in the report (or pick the top 3) by relevance (to environmental factors). To do so, we encoded sentences in the reports and compared them using cosine similarity with some compare sentences that we thought would help us extract the most relevant sentences.

For the encoding of the sentences, we experimented with SentenceBERT (Reimers and Gurevych, 2019) and Universal Sentence Encoder (Cer et al., 2018). We found that the Deep Averaging Network (DAN) version of the Universal Sentence Encoder works best for us. Since we were using the DAN version of the Universal Sentence Encoder, we created a compare sentence that was a scramble of words that would be important in the relevant sentences. We hypothesize that the scramble of words gave us a deeper, more diverse set of relevant sentences to extract from the reports.

After encoding and comparing each sentence in the report with the compare sentence(s), we chose the top 3 sentences for each document and fed that as input to our model. We let there be a truncation for those that exceeded 512 tokens and padding for those that had less than 512 tokens in the 3 sentences chosen.

Fine-Tuning

The approach for fine-tuning for both the classification tasks was to use BERT embeddings and attention masks of the chosen 512 tokens and feed them into the BERT that was fine-tuned on the ESG corpus. There is then a classification layer that takes the BERT outputs as inputs. The models were then hyperparameter tuned to achieve the results discussed in section 6.

5. Data

ESG Corpus for Fine-Tuning

The ESG corpus that we fine-tuned BERTs pre-trained weights on was obtained from the Knowledge Hub of the Accounting For Sustainability project.

Input and Output for Classification

For our input, we got 10-Q reports for S&P 500 companies from Notre Dame's Software Repository for Accounting[2] and Finance for the time frame 2014-2018. For the output, we used Wharton's research platform WRDS[3] to obtain quarterly Sustainalytics scores for the same companies for the same time frame.

EDA

Charts that display descriptive statistics about the data set are shown in the appendix. The overall distribution (Figure 1) of scores is not huge. Roughly 60% of the quarterly changes in environmental scores are zero. While the tails of the distribution do contain score changes on the larger side, most of the changes are quite small. This

[2] https://sraf.nd.edu/data/stage-one-10-x-parse-data

[3] https://wrds-www.wharton.upenn.edu/data-dictionary/sustainalytics_all/

does not affect our architecture much since we are doing binary classifications (change or no change; positive change or negative change). Figure 2 shows the relative frequency of the sentence lengths. Figures 3 and 4 show the portion of text in the reports that belong to the parts of speech and label explanations respectively.

6. Results and Discussion

Naive Bayes barely beat the common class in training and could not beat it in the validation or test data sets. ESGBERT is able to outperform BERT and the other classification techniques we compared against in both tasks:

1. Predicting whether the company will have an environmental score change or not.

2. Predicting whether a company's environmental score change will be positive or negative.

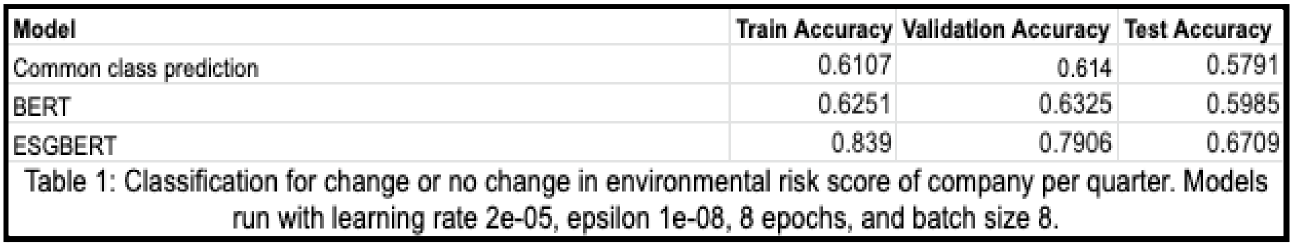
The results are illustrated in Table 1 and Table 2 respectively, and details about the relevant hyperparameters are mentioned in the description below each table.

A key driver of achieving the higher results is the use of the scrambled word sentence in the Universal Sentence Encoder. We were unable to achieve such high accuracy without this comparison sentence.

For future developments, we would like to try our model on Social and Governance risk scores as well as on ESG disclosures that companies have now started to release. We believe that since those disclosures will fully focus on the Environmental, Social, and Governance standings and practices, they will have more to inform about the scores than 10-Q filings did.

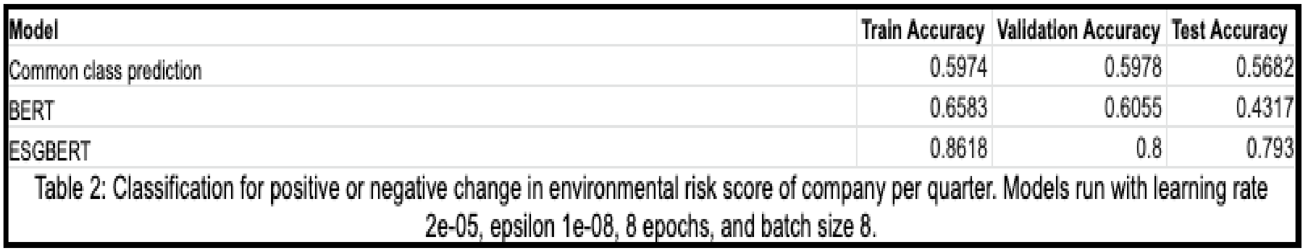
7. Conclusion

With our research, we strengthen the learning that pre-training BERT on domain-specific corpus yields better results in classification tasks related to that domain. As it has been done in other domains, we hope that ESGBERT can be used for multiple ESG specific text classification tasks and developers using this will be able to see improvements in their ESG related tasks.



Model	Train Accuracy	Validation Accuracy	Test Accuracy
Common class prediction	0.6107	0.614	0.5791
BERT	0.6251	0.6325	0.5985
ESGBERT	0.839	0.7906	0.6709

Table 1: Classification for change or no change in environmental risk score of company per quarter. Models run with learning rate 2e-05, epsilon 1e-08, 8 epochs, and batch size 8.



Model	Train Accuracy	Validation Accuracy	Test Accuracy
Common class prediction	0.5974	0.5978	0.5682
BERT	0.6583	0.6055	0.4317
ESGBERT	0.8618	0.8	0.793

Table 2: Classification for positive or negative change in environmental risk score of company per quarter. Models run with learning rate 2e-05, epsilon 1e-08, 8 epochs, and batch size 8.

References

1. Serafeim, George and Yoon, Aaron, Stock Price Reactions to ESG News: The Role of ESG Ratings and Disagreement (January 13, 2021). Harvard Business School Accounting & Management Unit Working Paper No. 21-079, Available at SSRN: https://ssrn.com/abstract=3765217 or http://dx.doi.org/10.2139/ssrn.3765217
2. Kölbel, J., Leippold, M., Rillaerts, J., & Wang, Q. (2020). Does the CDS market reflect regulatory climate risk disclosures?. SSRN, (3616324).
3. Berkman, H., Jona, J., & Soderstrom, N. S. (2019). Firm-specific climate risk and market valuation. Available at SSRN 2775552.
4. Matsumura, E. M., Prakash, R., & Vera-Muñoz, S. C. (2018). Capital market expectations of risk materiality and the credibility of managers' risk disclosure decisions. Available at SSRN, 2983977.
5. Reverte, C. (2016). Corporate social responsibility disclosure and market valuation: evidence from Spanish listed firms. Review of Managerial Science, 10(2), 411-435.
6. Verbeeten, F. H., Gamerschlag, R., & Möller, K. (2016). Are CSR disclosures relevant for investors? Empirical evidence from Germany. Management Decision.
7. Clarkson, P. M., Li, Y., Richardson, G. D., & Vasvari, F. P. (2008). Revisiting the relation between environmental performance and environmental disclosure: An empirical analysis. Accounting, organizations and society, 33(4-5), 303-327.
8. Cormier, D., & Magnan, M. (2007). The revisited contribution of environmental reporting to investors' valuation of a firm's earnings: An international perspective. Ecological economics, 62(3-4), 613-626.
9. Armbrust, F., Schäfer, H., & Klinger, R. (2020, December). A Computational Analysis of Financial and Environmental Narratives within Financial Reports and its Value for Investors. In Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (pp. 181-194).
10. Hummel, K., & Schlick, C. (2016). The relationship between sustainability performance and sustainability disclosure–Reconciling voluntary disclosure theory and legitimacy theory. Journal of accounting and public policy, 35(5), 455-476.
11. Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.
12. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.
13. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

14. SEC. 2010. Commission guidance regarding disclosure related to climate change. http://www.sec.gov/ rules/interp/2010/33-9106.pdf.
15. Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
16. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision (pp. 19-27).
17. Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.
18. Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Kurzweil, R. (2018). Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Appendix

Figure 1

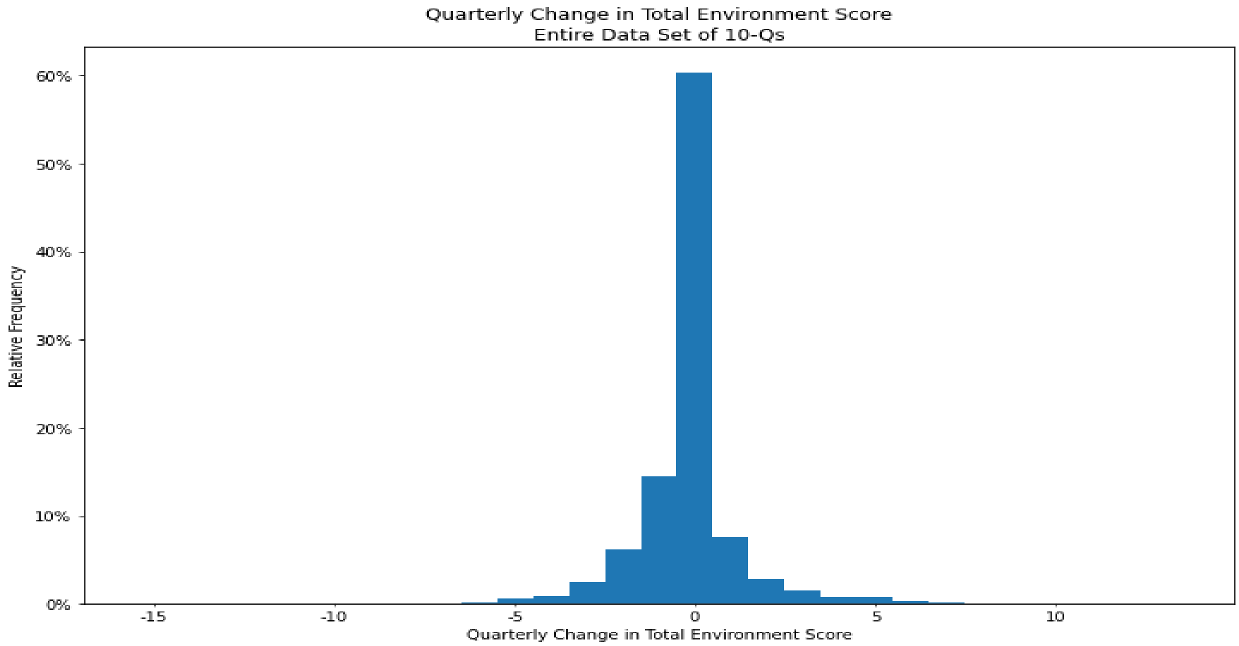


Figure 2

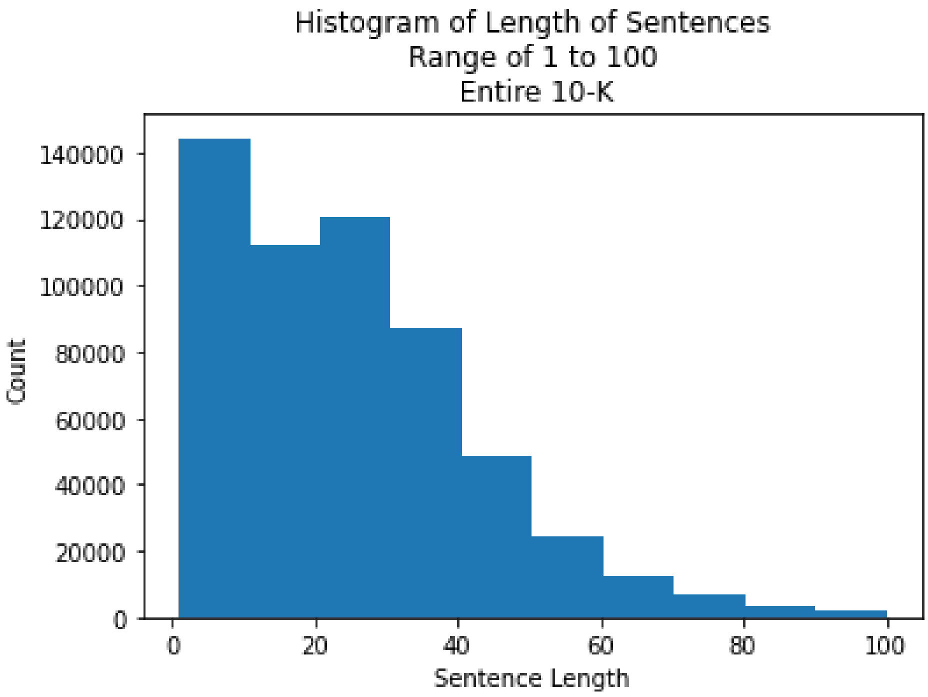


Figure 3

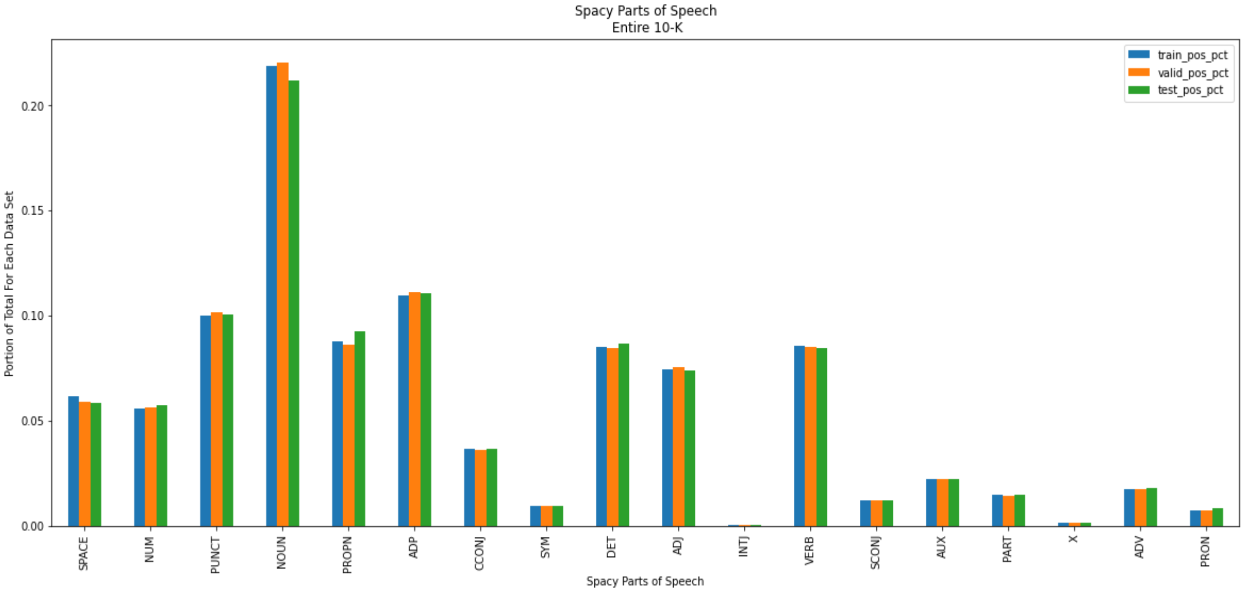
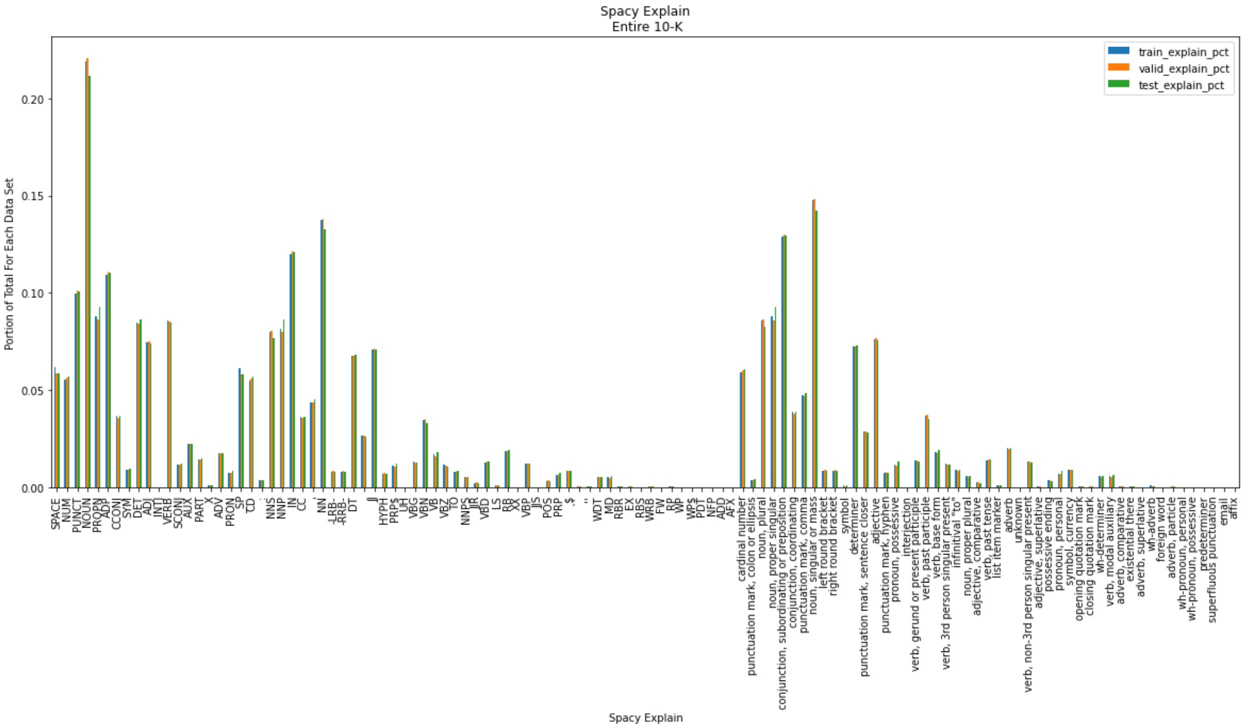


Figure 4


Spacy Explain
Entire 10-K
train_explain_pct
valid_explain_pct
test_explain_pct
Portion of Total For Each Data Set
0.00
0.05
0.10
0.15
0.20
Spacy Explain
SPACE
NUM
PUNCT
NOUN
PROPN
ADP
CCONJ
SYM
DET
ADJ
INTJ
VERB
SCONJ
AUX
PART
X
ADV
PRON
SP
CD
:
NNS
NNP
IN
CC
``
NN
-LRB-
-RRB-
DT
.
JJ
HYPH
PRP$
UH
VBG
VBN
VB
VBZ
TO
NNPS
JJR
VBD
LS
RB
XX
VBP
JJS
POS
PRP
$
,
''
WDT
MD
RBR
EX
RBS
WRB
FW
RP
WP
WP$
PDT
NFP
ADD
AFX
cardinal number
punctuation mark, colon or ellipsis
noun, plural
noun, proper singular
conjunction, subordinating or preposition
conjunction, coordinating
punctuation mark, comma
noun, singular or mass
left round bracket
right round bracket
symbol
determiner
punctuation mark, sentence closer
adjective
punctuation mark, hyphen
pronoun, possessive
interjection
verb, gerund or present participle
verb, past participle
verb, base form
verb, 3rd person singular present
infinitival "to"
noun, proper plural
adjective, comparative
verb, past tense
list item marker
adverb
unknown
verb, non-3rd person singular present
adjective, superlative
possessive ending
pronoun, personal
symbol, currency
opening quotation mark
closing quotation mark
wh-determiner
verb, modal auxiliary
adverb, comparative
existential there
adverb, superlative
wh-adverb
foreign word
adverb, particle
wh-pronoun, personal
wh-pronoun, possessive
predeterminer
superfluous punctuation
email
affix