Patent Text Novelty and Firm Market Value
---- SD4611 Interim Report

Yukyee CHAN

1. Introduction

Novelty detection is classifying validation data, which differs from the data available during training. This can be seen as a "one-class classification", where a model is established to describe "normal" training data (Pimentel Marco et al., 2014). It is an important ability of a signal recognition scheme. The principal reason is that classification work is an open set in the actual world (Bendale & Boult, 2016), in other words, the classification system will reject the unknown classes at the testing time since a traditional classification system may not provide enough classification categories in the training set. A classifier does not work or perform poorly when recognizing a class that has never been seen before. The challenge of assessment, or the problem of information overflow (Gerken & Moehrle, 2012), is the second important issue. Manually detecting the novelty will result in an information overflow issue. The personnel will spend too much time and work inefficiently, particularly if the data format is text. Therefore, novelty detection methods in machine learning are very valuable in improving performance and efficiency.

The past has seen a wide variety of textual novelty detection methods. Pimentel Marco et al. (2014) integrate research papers about novelty detection that have appeared in the machine learning literature, for instance, probabilistic-based novelty detection, distance-based novelty detection, distribution-based novelty detection, and such. The development and advancement of the economy are driven mainly by technological innovation. The outcome of technological innovation is a patent. Measuring the novelty of patents through novelty detection methods has been the focus of attention in recent years. Investors, companies, and researchers are interested in patents as a form of intellectual property. An enormous amount of prior research has developed a stock price forecast model based on the patent indicator, for instance, citation links, keywords, abstract documents, and so on (Shibayama et al., 2021). However, nobody is measuring the novelty in patents without considering the evaluations of relevant scientists and subject matter experts.

Considering the impact of syntactic and semantic novelty, we adopted the patent's abstract documents as the predictor. In this paper, we want to study the influences of the textual novelty of the patent on the stock value of the firm without using relevant scientific or expert evaluations as quality indicators. We will measure and evaluate existing novelty detection methods, then apply the best performing method to our patent data to find the relationship between patent novelty and firm value.

When our work is finished, we expect that it will enhance the field of novelty detection research and help with the investigation of the relationship between patent novelty and firm value, enabling others to build on our discoveries.

2. Textual Novelty Detection

Textual novelty detection involves categorizing text/document validation data, which is different from the data accessible during training, i.e., given a new document *p* and a set of existing documents D=$\{d_i\}$, the textual novelty detection is to define a function *TN(p, D)* that tells how novel *p* is given the existence of *D*. This section introduces the existing available approaches and evaluation framework.

2.1 Related Work

Distance-based methods, which assume that previously seen or known data is clustered together and new data is farther away from the cluster, are perhaps the most common novelty detection method (Hautamaki et al., 2004). The drawback of Hautamaki (2004) is that the supervised classification method, KNN Graph, adopted in the paper, relies on all classes present in the data at the training set. A traditional multi-class classification scheme is inappropriate for this case study, as some anomalies may not be known a priori (Pimentel et al, 2014). Gerken and Moehrle(2012) employed semantic patent analysis to measure novelty in semantic patent analysis to identify highly novel inventions.

Distribution-based method, which assumes that known or observed data has its own probability distribution and that those distributions may be thresholded to define the boundaries of different classes in the dataset, is another common method for detecting novelty (Pimentel et al, 2014). The traditional way to achieve novelty detection is to threshold the entropy of the class probability distribution (Hendrycks & Kevin Gimpel, 2016). These methods do not actually measure novelty, but rather closeness to the decision boundary.

Neural network-based method, which classifies whether test data belongs to the training data class by defining a description of normal data and building a model for it (Markou & Singh, 2003). One of the deep learning for novelty detection -- the autoencoder model extract a low-dimensional representation from the high-dimensional input space. It tries to copy its input to output, and anomalies/novelties are harder to reconstruct than normal samples. Data points that differ from the normal distribution cannot be restored well, resulting in large reconstruction errors. Mei (2018), Adarsh (2021), and Goudarzvand (2022) employed autoencoders for textual novelty analysis.

Table 1. Existing Studies on Textual Novelty Detection

Studies	Technique	Evaluation Metrix	Shortcomings
(Bhattarai et al., 2020)	Tsetlin Machine Text Classifier	Accuracy scores	It relies on all classes present in training data
(Shibayama et al., 2021)	Distance-based method: Q-percentile Similarity Method	Pearson Correlation and Logistic Regression	Suffers from insufficient validation.
(Gerken & Moehrle, 2012)	Distance-based method: Maximum Similarity Method	Spearman's rank correlation coefficients, Recall and Precision	Need to be concern the scope of the case study.
(Luo et al., 2022)	Distance-based method: Maximum Similarity Method with BERT	Correlation test and p-values	Disciplinary limitations
(Hendrycks & Kevin Gimpel, 2016)	Distribution-based method: Threshold decision in PDF	The Area Under the Receiver Operating Characteristic curve (AU-ROC) ,and Area Under the Precision-Recall curve (AUPR)	It does not actually measure novelty, but rather closeness to the decision boundary.
(Hautamaki et al., 2004)	Distance-based method: KNN Graph	Receiver Operating Characteristics (ROC)	It relies on all classes present in training data
(Adarsh et al., 2021)	Neural network-based method: Using Latent Dirichlet Allocation with Auto-Encoders	Precision and recall	

(Mei et al., 2018)	Neural network-based method: Using Semantic Clustering and Autoencoders	Pearson correlation coefficients	It is not well correlated with novelty assignments made by a human rater
(Goudarzvand et al., 2022)	Neural network-based method: Similarity-based second chance autoencoders	Precision, recall, and F1 score	

2.2 Existing Methods

To measure textual novelty in this study, we employ a few representative methodologies from the above literature and apply them. We provide further detail about them below.

2.2.1 TFIDF-based Maximum Similarity Method

TF-IDF is a statistical method used to evaluate the importance of a term to a collection of archives or to one of the archives in a corpus. The importance of a word increases proportionally to the number of times it occurs in the archive but decreases inversely proportional to the frequency it appears in the corpus. The TF-IDF representation calculates and multiplies two measures for each word in a document, term frequency, and inversed document frequency. Term frequency is the frequency with which a given the word appears in the document:

$$tf(t,d) \coloneqq \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} = \frac{(Number\ of\ occurrences\ of\ term\ t\ in\ document\ d)}{(Total\ number\ of\ terms\ in\ the\ document\ d)} \qquad (1)$$

Inverse document frequency represents how common a word is in the entire document set D:

$$idf(t,D) \coloneqq \ln\left(\frac{N}{1 + |\{d \in D | t \in d\}|}\right) = \ln\left(\frac{Total\ number\ of\ documents\ in\ the\ corpus}{Number\ of\ documents\ with\ term\ t\ in\ them}\right) \qquad (2)$$

Cosine similarity is a commonly used similarity calculation method in information retrieval. It can be used to calculate the similarity between documents, it can also calculate similarity between words, and it can also calculate the similarity between query strings and documents. Here, we define similarity as the cosine similarity build upon the TF-IDF vector representation of documents. After transforming each document into a vector of TF-IDF values, the cosine similarity of any pair of vectors is obtained by taking their dot product and dividing it by the product of their norm, which also indicates the cosine of the angle between the vectors:

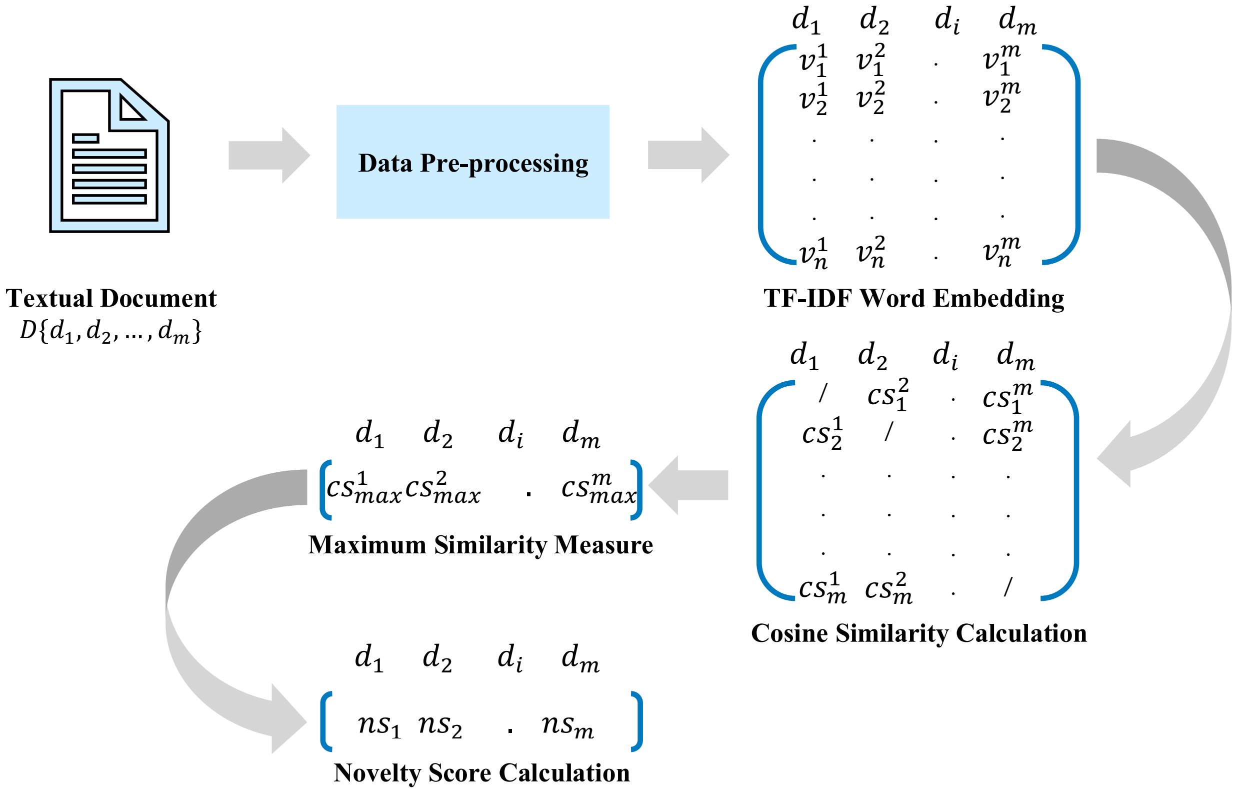
$$cosine(\boldsymbol{v}, \boldsymbol{w}) = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{|\boldsymbol{v}||\boldsymbol{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}} \qquad (3)$$

Finally, we employ a simple that calculates the largest similarity between current document p and every document d_i in D, and novelty score defines one minus maximum similarity, which also represents distance:

$$MaximumSimilarity(p, D) = \ max_i(cosine\ similarity\ (p, d_i)) \qquad (4)$$

$$TFIDF\_NoveltyScore(p,D)=1-MaximumSimilarity(p,D) \qquad (5)$$

Figure 1. TFIDF-based novelty calculation process


Textual Document
$D\{d_1, d_2, \dots, d_m\}$
Data Pre-processing
TF-IDF Word Embedding
Cosine Similarity Calculation
Maximum Similarity Measure
Novelty Score Calculation

d_i = document for the i^{th} data instance
$v_n^i = n^{th}$ *word embedding vector of* i^{th}*document*
$cs_k^i =$ *cosine similarity of* i^{th}*document and* k^{th}*document*
cs_{max}^i = *maximum cosine similarity of* i^{th}*document*
ns_i = novelty score for the i^{th} data instance

2.2.2 Bert-based Maximum Similarity Method

BERT is a pre-trained model released by Google, trained using Wikipedia and book corpus data (Devlin et al., 2018). BERT offers an advantage over other contextual embedding models because it dynamically generates word representations based on surrounding words. BERT can also be used for extraction from text data. Based on the existing literature on novelty detection, (Luo et al., 2022), we refer to their method, which defines similarity as the cosine similarity of document-based BERT vector representations. Reimers and Gurevych (2019) present Sentence-BERT (SBERT), a modification of a pre-trained BERT network that uses Siamese and Triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine similarity. Here we use SBERT, and we choose "sentence-transformers/all-MiniLM-L6-v2". It maps sentences and paragraphs to a fixed 384-dimensional dense vector space, which can be used for tasks such as semantic search.

We use SBERT deep learning model to train the corpus and obtain the word embedding representation of documents. Then, we define similarity as the cosine similarity, as seen in Equation (3), built upon the SBERT 384-dimensional dense vector representation of documents. As same Section 2.2.1TFIDF-based Method as seen in (4), we also employ the maximum similarity measure that calculates the largest similarity between current document p and every document di in D, and novelty score defines one minus maximum similarity, which also represents distance:

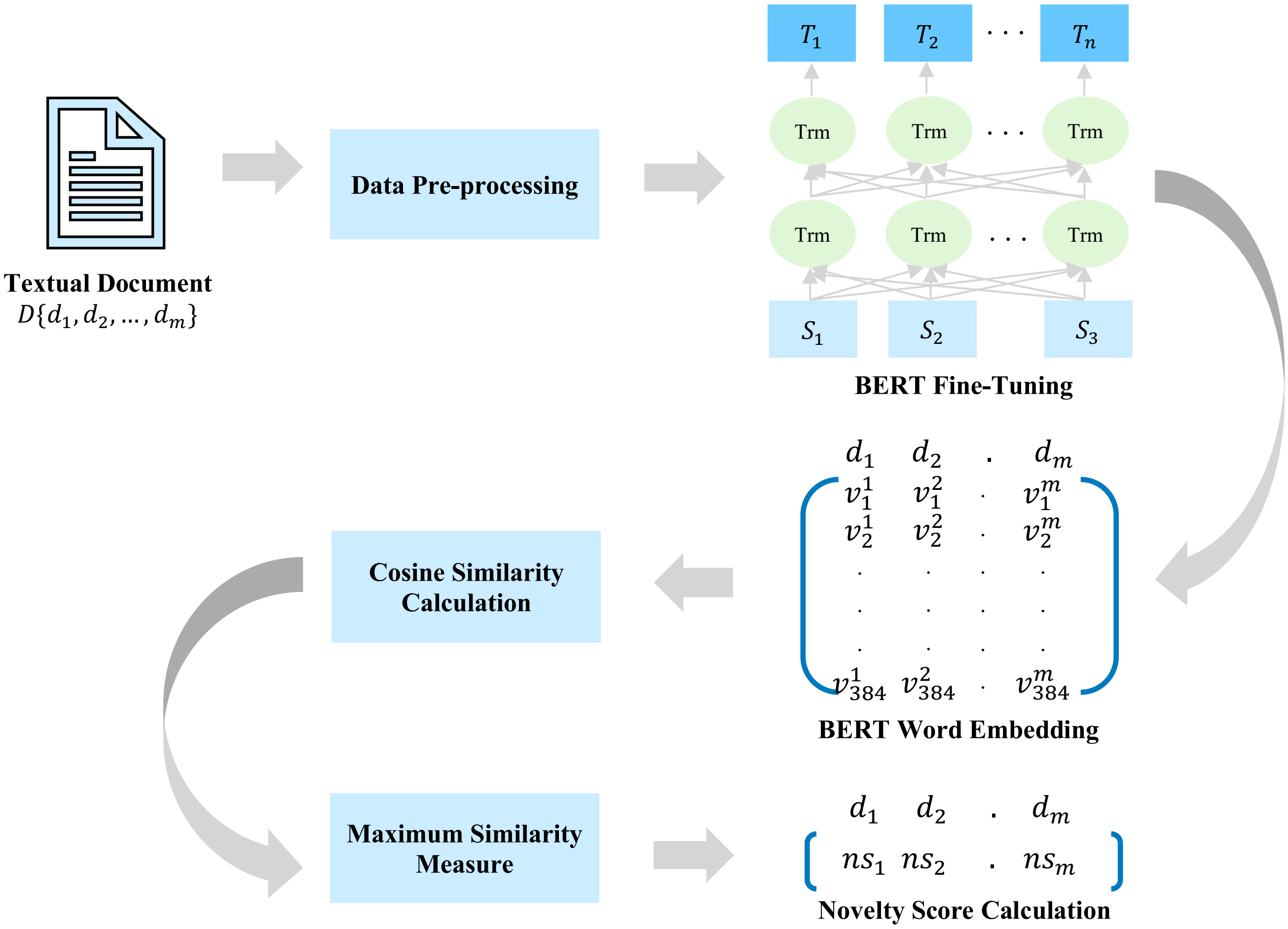
$$BERT\_NoveltyScore(p,D)=1\text{-}MaximumSimilarity(p,D) \tag{6}$$

Additionally, we normalized (without changing the magnitude of the score) the novelty scores to better represent the data:

$$Normalized\_NoveltyScore_i = \frac{NoveltyScore_i - NoveltyScore_{min}}{NoveltyScore_{max} - NoveltyScore_{min}} \tag{7}$$

For BERT-based word vector model fine-turning, we split patent data into sentence levels as a training data, set the number of SBERT model warm-up steps to 500, the training batch sample size to 32, the epochs step size to 10, and the rest of the parameters are default settings. After training the model, a word vector representation model is obtained.

Figure 2. BERT-based novelty calculation process

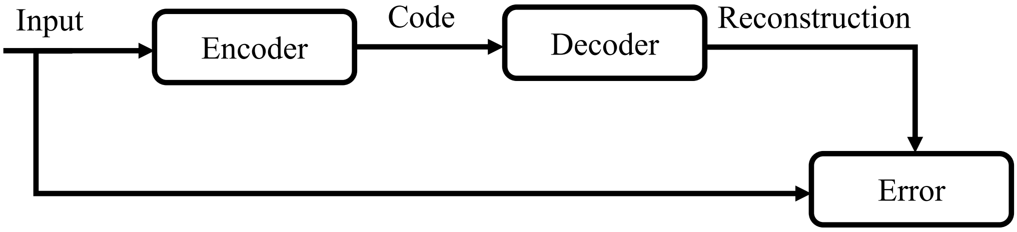


$S_i = i^{th}\ sentence\ data\ in\ traing\ set$
$T_i = i^{th}\ output\ from\ the\ SBERT\ model$
$Trm = transformer\ encoders\ and\ decoders$
$v_n^m = n^{th}\ word\ embedding\ vector\ of\ m^{th} document$
$ns_i =$ novelty score for the i^{th} data instance

2.2.3 Variational Autoencoding

Autoencoders are trained using the same data at the input layers and output layers, where the neural network learns the representation of the dataset, mostly for dimensionality reduction, eliminating unwanted signals during training. Known/normal data inputs can be passed directly from layer to layer with minimal loss, but unknown/novelty data will have more data loss as it deviates from the hidden data pattern. Based on the existing works of literature on textual novelty detection, (Adarsh et al., 2021), (Mei et al., 2018), and (Goudarzvand et al., 2022), we refer to their method, adopting a deep learning autoencoder model to extract a low-dimensional representation from the high-dimensional input space. Figure 4 showed that the architecture of the autoencoder model consists of two symmetric deep neural networks - an encoder and a decoder that apply backpropagation to set the target value equal to the input. Anomalies/novelties are harder to reconstruct than normal samples. Data points that differ from the normal distribution cannot be restored well, resulting in large reconstruction errors. (See Figure 3.)

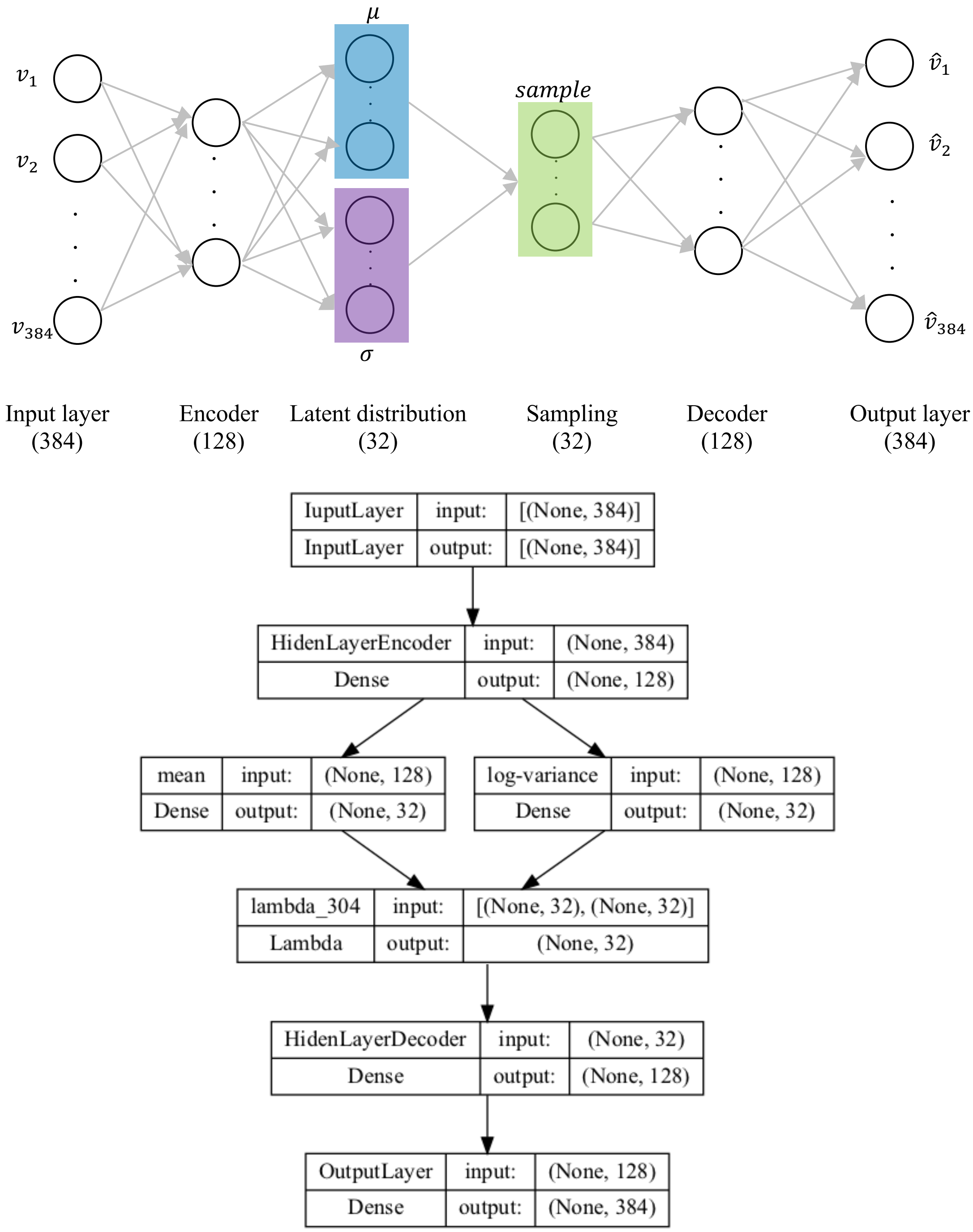
Figure 3. The concept of reconstruction error in autoencoder



In this section, we employ Variational Autoencoder (VAE) to obtain the novelty scores. VAE is an autoencoder whose latent distribution is regularized during training, sampling through a normal distribution to ensure its latent space is well-characterized for better results.

Here, we use SBERT to generate 384-dimensional dense vectors for each document. The VAE used here is a seven-layer deep neural network with a network structure of 384 features in the input layer (Layer 1), 128 features in the encoding layer (Layer 2), 32 features in the bottleneck layer (Layers 3,4,5), decoding layer (Layer 6) with 128 features, and the output layer (layer 7) have 384 features. Layers 2, 3, 4, 5, and 6 form the hidden layers of the autoencoder. In the latent space, we build a standard normal distribution to represent the data. The input is mapped to a normal distribution. Mean and variance are learned during model training (Layers 3,4). Then, the latent vector is sampled from the normal distribution with mean and variance (Layer 5) and passed to the decoder to obtain the predicted output. The Figure 4 shows the architecture of our variational autoencoder model.

Figure 4. The Architecture of Variational Autoencoder model


v_1
v_2
v_{384}
μ
σ
$sample$
$\hat{v}_1$
$\hat{v}_2$
$\hat{v}_{384}$
Input layer (384)
Encoder (128)
Latent distribution (32)
Sampling (32)
Decoder (128)
Output layer (384)
IuputLayer | input: | [(None, 384)]
InputLayer | output: | [(None, 384)]
HidenLayerEncoder | input: | (None, 384)
Dense | output: | (None, 128)
mean | input: | (None, 128)
Dense | output: | (None, 32)
log-variance | input: | (None, 128)
Dense | output: | (None, 32)
lambda_304 | input: | [(None, 32), (None, 32)]
Lambda | output: | (None, 32)
HidenLayerDecoder | input: | (None, 32)
Dense | output: | (None, 128)
OutputLayer | input: | (None, 128)
Dense | output: | (None, 384)

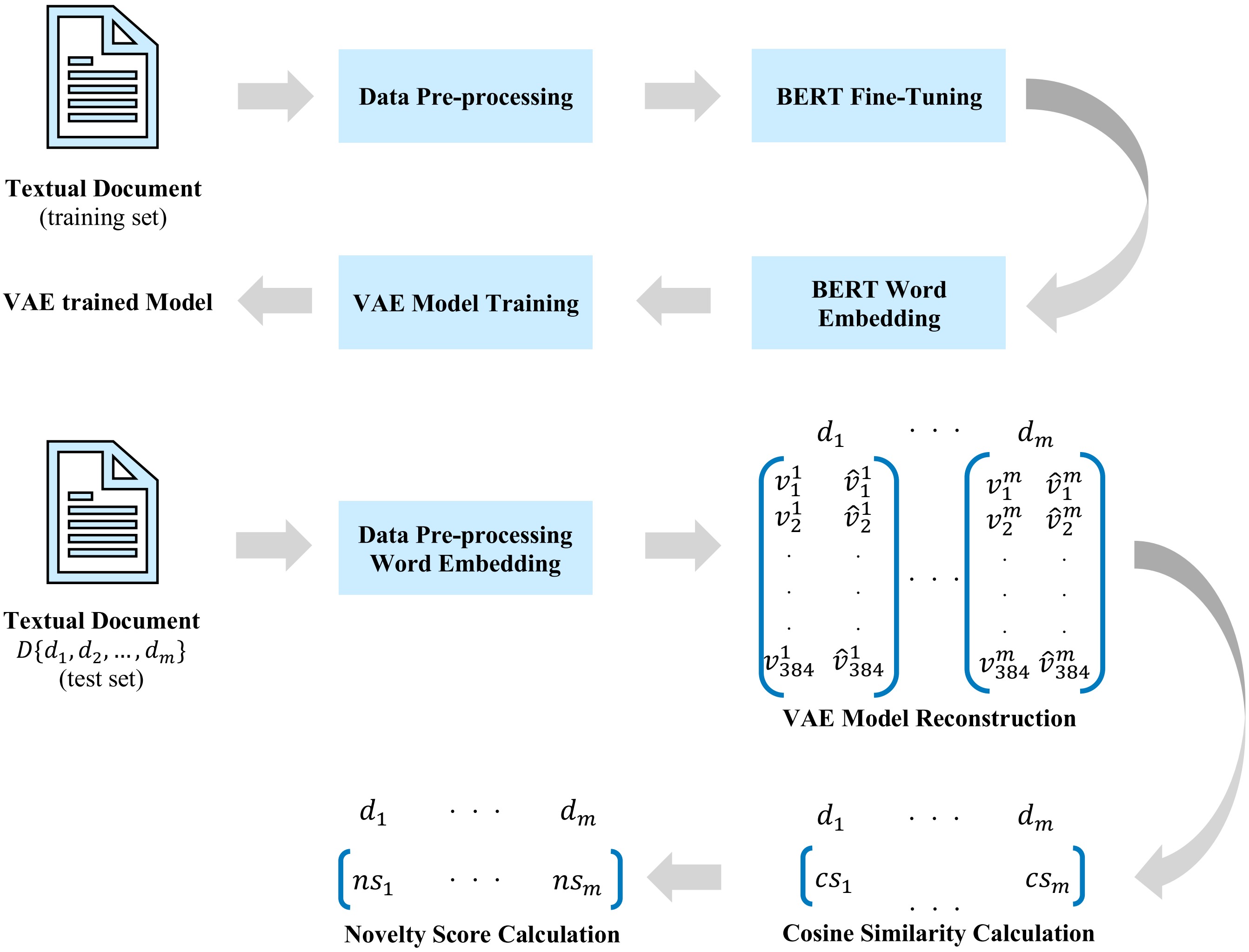
In the training phase, we train the data through the Adam optimizer, and set reLU as the activation function in the input layer and hidden layer and set Sigmoid as the activation function in the output layer. We set a batch size of 32, 100 training epochs, and use 25 percent of the data for validation. A deep-learning VAE model is produced once the model has been trained. Then, we define similarity as the reconstruction errors, that is calculate the cosine similarity of the predicted data and actual data to measure the reconstruction error/loss to determine the novelty score (Adarsh et al., 2021):

$$VAE\_NoveltyScore_i = 1 - cosine\ similarity(p_i, a_i) \tag{8}$$

$NoveltyScore_i =$ novelty score for the i^{th} data instance
$p_i =$ predicted data for the i^{th} data instance
$a_{i=}$ actual data for the i^{th} data instance

Additionally, we normalized (without changing the magnitude of the score) the novelty scores computed by reconstruction errors to better represent the data, as seen in Equation (7).

Figure 5. VAE novelty calculation process


Textual Document
(training set)
Data Pre-processing
BERT Fine-Tuning
VAE trained Model
VAE Model Training
BERT Word Embedding
Textual Document
$D\{d_1, d_2, \dots, d_m\}$
(test set)
Data Pre-processing Word Embedding
d_1 . . . d_m
v_1^1 $\hat{v}_1^1$ v_2^1 $\hat{v}_2^1$ v_{384}^1 $\hat{v}_{384}^1$
v_1^m $\hat{v}_1^m$ v_2^m $\hat{v}_2^m$ v_{384}^m $\hat{v}_{384}^m$
VAE Model Reconstruction
d_1 . . . d_m
ns_1 . . . ns_m
Novelty Score Calculation
d_1 . . . d_m
cs_1 . . . cs_m
Cosine Similarity Calculation

2.3 Proposed Deep Learning Novelty Measure

This study proposes a method to measure the novelty of patent data based on textual data. The proposed method considers article novels if the article contains a combination of semantically distant vectors. To this end, we first assign the word embedding vector representation of each vocabulary to each textual information. After that, we employ a simple maximum similarity measure that calculates the novelty scores for both the normal and novel documents. In the upcoming semester, we expect to employ deep-learning novelty measures to measure and assess novelty.

2.4 Evaluation Framework

Before employing these measures to study the impact of patents on company stock prices, we are necessary to evaluate whether they are definitely capable of capturing the novelty of the text. This section introduces the experimental setup including the preparation of benchmark datasets, the process of empirical evaluation, the selection of evaluation metrics, and the performance comparison. It aims to perform an experimental comparative evaluation of selected representative novelty detection methods.

2.4.1 Dataset

To ensure consistency and robustness of measures comparisons, we concentratedly use the 20 Newsgroups Dataset to evaluate the different measures. This dataset is a collection of newsgroup documents, which is one of the most common and popular datasets for experimentation with natural language processing of machine learning techniques. It contains 20 categories with a total of 18,828 text documents. In our experiments, we treat the three classes "alt.atheism", "comp.graphics" and "comp.os.ms-windows.misc" as normality classes and the "rec.motorcycles" class as a novelty class.

2.4.2 Empirical Evaluation

Following the existing novelty detection literature (Bhattarai et al.,2020), as shown in Figure 6 below, we set up a baseline document set with normal (non-novel) documents and two comparison groups, one containing only normal documents and one containing both normal and novel documents. The former is simulating the situation of the occurrence of normal documents, and the latter is simulating the situation of the occurrence of novel documents. In this way, we enable to discover the difference between normal documents and novel documents through the following processing.

Figure 6. The difference between comparison groups

	Training set	Test set
1	Normal documents	Novel documents
2	Normal documents	Normal documents

To discover the difference between normal documents and novel documents, we then divided the task into two experimental steps. After performing data pre-processing operations such as stop words removal, punctuation removal, lower casing, and tokenization, we employ different methods to distinguish two classes.

The first step is novelty score calculation. For the TFIDF-based and BERT-based methods, we employ a simple maximum similarity measure that calculates the novelty scores for both the normal and novel documents (Gerken & Moehrle, 2012), as seen in (5) and (6). For the VAE method, we define similarity as the reconstruction errors, then obtained the novelty score computed by cosine similarity of reconstruction errors (Adarsh et al., 2021), as shown in (8). The second step is the measurement and validation of the novelty score based on different evaluation metrics.

2.4.3 Evaluation Metrics

Following the existing novelty detection literature (Shibayama et al.,2021) and (Luo et al., 2022), we calculate the Pearson correlation coefficient between the maximum similarity and two comparison groups. To determine the difference between normal and novel documents, we calculate the correlation between the novelty score and the novelty/normal classes.

Besides, we calculate the Jaccard coefficient and Dice coefficient to measure the similarity/overlap level between two comparison groups, as seen in Equations (9) and (10) below. The larger the Jaccard/ Dice coefficient value, the higher the sample similarity /overlap. Kabir et al (2017) indicated that the commonly used similarity measures are the Jaccard and Dice measures, which are currently more popular when calculating the similarity/overlap of the interval. Additionally, we also employ the Jaccard coefficient in the continuous version, proposed by Costa (2021), as shown in (11).

$$Dice(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{9}$$

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{10}$$

$$Jaccard_{continuous(A,B)} = \frac{a^2r(1+r) - 2rax}{2a^2(1+r^2) - a^2r(1+r) + 2rax} \tag{11}$$

$$where\ a = size\ of\ A, b = size\ of\ b, r = \frac{b}{a}\ with\ 0 \le r \le 1$$

2.5 Performance Comparison

Figure 7, Figure 8, and Figure 9 showed that the results of an analysis that compared the distribution of novelty scores for normal documents and novel documents using three different methods: TFIDF-based Maximum Similarity method, BERT-based Maximum Similarity method, and Variational Autoencoder method. The results show that there is a difference in the novelty distribution between normal documents and novel documents. Specifically, novel documents have a higher distribution of novelty scores than normal documents. This shows that novel documents tend to have more uniqueness compared to normal documents.

When looking at the different methods used, the TFIDF-based method showed that most novel classes appear to have scores between 0.8 and 1, while normal classes appear to have an irregular distribution, as seen in Figure 7 below. In contrast, according to Figure 8 and Figure 9, the BERT-based and VAE methods showed that both normality and novelty classes followed a normal distribution, but the normality class was more centralized in the VAE method than in the BERT-based method and the novelty class was more centralized in the BERT-based method than in the VAE method. This means that the BERT-based method may be more sensitive to identifying unique feature in novel documents, and VAE method may better at distinguishing between normal and novel classes.

Figure 7. Novelty distribution of TFIDF-based method between two class

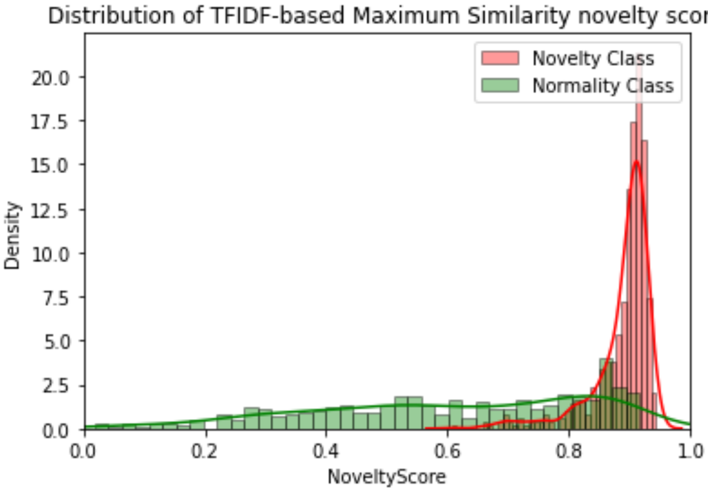


Figure 8. Novelty distribution of BERT-based method between two class

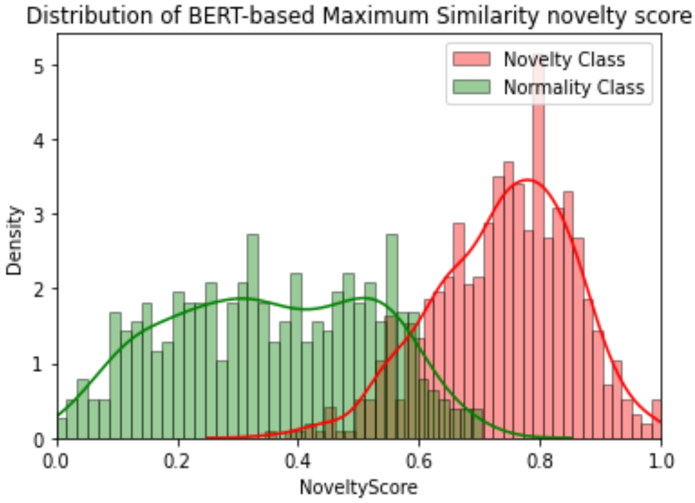


Figure 9. Novelty distribution of VAE method between two class

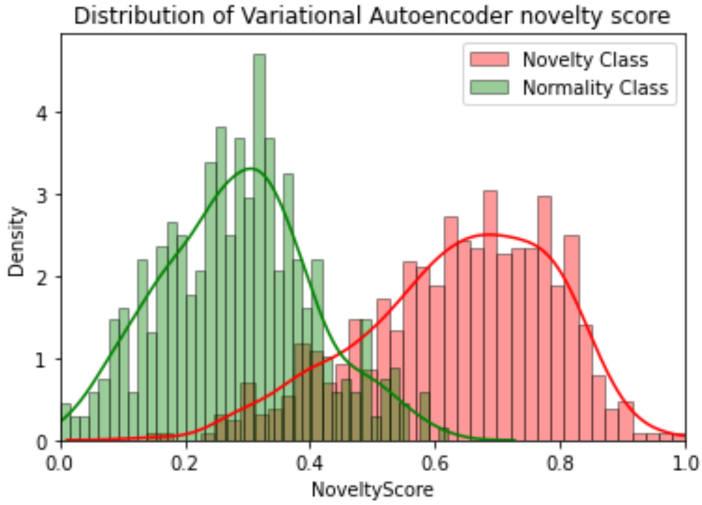


Table 2. Measure Performance Comparison

Measures	**Pearson correlation coefficient**	**Jaccard coefficient**	**Jaccard coefficient (continuous version)**	**Dice coefficient**
TFIDF-based Maximum Similarity	0.679	0.313	0.063	0.477
Bert-based Maximum Similarity	0.808	0.301	0.3	0.462
Variational Autoencoder	0.782	0.367	0.459	0.537

The experimental results for all measures are shown in Table 2. The Pearson correlation coefficient is used to measure the linear relationship between two variables - the predicted novelty score and the novelty/normal classes. It ranges from -1 to 1, where -1 means a perfect negative correlation, 1 means a perfect positive correlation, and 0 means no correlation. As seen, the correlation coefficient of the Bert-based Maximum Similarity method is 0.808. It is the highest among the three methods, indicating that exists a strong positive correlation between the similarities computed by the predicted novelty score and the novelty/normal classes. The correlation coefficient of the Variational Autoencoder and TFIDF-based Maximum Similarity method is also relatively high, respectively 0.782 and 0.679. However, they are both slightly lower than that of the Bert-based method.

The Jaccard coefficient and Dice coefficient are used to measure the similarity/overlap level between two comparison groups. They both also range from 0 to 1, where 0 indicates no similarity/overlap and 1 indicates perfect similarity/overlap. The Jaccard and Dice coefficient of the Bert-based method is the lowest among the three methods, respectively 0.301 and 0.462.

3. Firm Market Value Analysis

3.1 Data Collection

In this study, we concentrate on high-tech companies in their early stage to examine the effect of patent textual novelty on company value, since we believe patent has a greater impact on them. Specifically, Biotech startups are nascent industries. Moreover, biotech startups were chosen for the study as they obtained more abundant patent data than other startups, such as Fintech startups and IT(SAAS) startups. Due to privacy concerns or legal concerns, most of these two categories of firms will present fewer patent applications, which will result in less patent evidence of their innovation.

We choose The Nasdaq Stock Market to collect the firm's stock data since it is the second largest stock exchange in the world, after NYSE Euronext. According to internal statistics, 82% of the US biotech companies are listed on the NASDAQ. Hence, the stock market data of the US biotech companies listed on the NASDAQ was decided to collect.

We collect a name list of biotech companies from The U.S. Securities and Exchange Commission, which is an independent agency of the U.S. federal government, established after the Wall Street Crash in 1929. We also use its Standard Industrial Classification Codes as a standard selection. A company's business type is identified by the Standard Industrial Classification Codes that appear in its disseminated EDGAR filings. We select SIC: 2836 —

Biological Products, Except Diagnostic Substances as a sample. As shown in appendix Table A1, it contains a list of the chosen startup company listed on the NASDAQ from 2000 to 2019. In total, we have 196 companies.

The stock market data of the US biotech companies listed on the NASDAQ from 2008 to 2019 was collected by using the Python library — yfinance, which provides users with current and historical stock market price data from Yahoo Finance. We collect this period because little has changed throughout this period, avoiding systematic risk in the market. We collect company symbols and daily close prices from 2009 to 2019. As shown in Appendix Table A2. In total, we have 112 companies.

3.2 Patent Data Processing and Novelty Detection

My group mate collected patent data for me. We have 4878 patents for the US biotech companies listed on the NASDAQ. Besides, we have 200 to 300 patents as baseline documents. Then, we apply the measures developed in Section 2 on to these patents.

3.3 Summary Statistics

In the upcoming semester, we expect that the summary statistics should be completed.

3.4 Regression Analysis

In the upcoming semester, we expect to finish the regression analysis.

4. Discussion

In this study, we discover there is a significant difference in maximum similarity between normal and new documents, shown in Figure 2. Additionally, we discovered that Pearson's coefficient of the maximum similarity score and the novelty/normal classes is 0.623 and they are closely related, shown in Table 2. That implies that using maximum similarity scores to represent novelty/common classes is possible.

The finding of this study is that novelty detection helps us determine what is original, innovative, and noteworthy. So that this study will further the study of novelty detection and help with the investigation of the relationship between patent novelty and company value. It will significantly advance the science of novelty detection.

4.1 Patent Textual Novelty's Impact on Firm Value

In the upcoming semester, we expect that the content of patent textual novelty's impact on firm value should be completed.

4.2 Comparison of Different Textual Novelty Measures

In the upcoming semester, we expect that the content of comparison of different textual novelty measures should be completed.

4.3 Comparison of different part of patent

In the upcoming semester, we expect that the content of comparison of different part of patent should be completed.

5. Conclusions

In conclusion, we completed nascent industry market research and biotechnology stock market data collection. We also conducted research on measures of novelty detection and implemented a portion of the measures for our subsequent study of patent innovation and company value. To differentiate between normal documents and novel documents, we deployed one of the existing novelty detection methods — Maximum Similarity Method. We discovered that the novelty/normal classes are closely related.

We hope that future researchers will explore further new novelty detection methods based on our findings. Furthermore, we believe that novelty detection methods will address the problem of information overflow, which will increase the effectiveness and efficiency of evaluating any document from a social perspective.

However, we still have a lot of room for improvement in this study. Since we only use one dataset — 20 Newsgroups Dataset for evaluation, we may suffer from insufficient validation since biological product patents often contain complex chemical components. Shibayama et al (2021) also suffer from this problem. Therefore, more datasets may need to be included in the evaluation in subsequent studies.

6. Acknowledgement

I would like to thank Prof. Xin LI and Dr. Wei HU for their guidance. I appreciate that they guide our research, check on our progress, and give useful advice in detail every week.

Besides, I would like to thank my group members Wanyue Zhou and Jianming Huang. Many thanks to Zhou for providing the related literature to enrich my study on novelty detection. Also, I appreciate gratefully to Huang for providing the patent abstract data to help me to complete the patent novelty detection in the future.

References

1. Hautamaki, V., Karkkainen, I., & Franti, P. (2004). Outlier detection using K-nearest neighbour graph. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*
2. Bhattarai, B., Granmo, O.-C., & Jiao, L. (2020). Measuring the novelty of natural language text using the conjunctive clauses of a Tsetlin machine text classifier. *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*.
3. Shibayama, S., Yin, D., & Matsumoto, K. (2021). Measuring novelty in science with word embedding. *PLOS ONE*, *16*(7).
4. Bendale, A., & Boult , T. E. (2016). Towards open set deep networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
5. Gerken, J. M., & Moehrle, M. G. (2012). A new instrument for Technology Monitoring: Novelty in patents measured by semantic patent analysis. *Scientometrics*, *91*(3), 645–670.
6. Hendrycks, D., & Gimpel, K. (2016). A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *International Conference on Learning Representations 2017*.
7. Pimentel, M.A., Clifton, D.A., Clifton, L.A.& Tarassenko, L. (2014). A review of novelty detection. *Signal Process., 99*, 215-249.
8. Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv.*

9. Luo, Z., Lu, W., He, J., & Wang, Y. (2022). Combination of research questions and methods: A new measurement of scientific novelty. *Journal of Informetrics*, *16*(2), 101282.
10. Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using Siamese Bert-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.
11. Markou, M., & Singh, S. (2003). Novelty detection: A review—part 2: neural network based approaches. *Signal Processing*, *83*(12), 2499–2521.
12. Adarsh, S., Asharaf, S., & Anoop, V. S. (2021). Sentence-level document novelty detection using latent Dirichlet allocation with auto-encoders. *Advances in Intelligent Systems and Computing*, 511–519.
13. Mei, M., Guo, X., Williams, B. C., Doboli, S., Kenworthy, J. B., Paulus, P. B., & Minai, A. A. (2018). Using semantic clustering and autoencoders for detecting novelty in corpora of short texts. *2018 International Joint Conference on Neural Networks (IJCNN)*.
14. Goudarzvand, S., Gharibi, G., & Lee, Y. (2022). Similarity-based second chance autoencoders for textual data. *Applied Intelligence*, *52*(11), 12330–12346.
15. Kabir, S., Wagner, C., Havens, T. C., Anderson, D. T., & Aickelin, U. (2017). Novel similarity measure for interval-valued data based on overlapping ratio. *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*.
16. Costa, L.D. (2021). Further Generalizations of the Jaccard Index. ArXiv.

Appendix

Table A1: Company list data
https://docs.google.com/spreadsheets/d/1DThPZ9S_Rr7c0GLgUJM5O7iLdeENMwZv/edit#gid=654600027
Table A2: Stock price data
https://docs.google.com/spreadsheets/d/1D1ADzSM4DY6JVBHKRLaSUDdTeuRq4n3O/edit?usp=sharing&ouid=114399266780834842979&rtpof=true&sd=true