SDSC4116 Final Report

Innovation Novelty and Firm Performance ---- Textual Measures

Yukyee CHAN (56230549)

Abstract:

The purpose of this study is to investigate the correlation between the originality of patent text and firm value. Technological innovation is a key factor for economic development, and this study evaluates and compares three different methods for quantifying the originality of patent texts, specifically TFIDF-based Maximum Similarity, BERT-based Maximum Similarity, and Variational Autoencoder. The results show that the BERT-based model outperforms the other two methods in detecting the uniqueness of the patent text. Furthermore, this study establishes a positive link between BERT-based text novelty and firm value using ordinary least squares regression analysis of biotech start-ups. This study highlights the importance of detecting novel concepts to improve the efficiency and performance of identifying the impact of novelty in patent texts on firm value.

Keywords:

Patent innovation, Textual novelty, Maximum Similarity, Variational Autoencoder

1. Introduction

Economic development and progress are primarily driven by technological innovation. Patents are frequently regarded as signals of an organization's innovation capacity, intellectual property, and research and development (R&D). 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).

In this project, we are interested in the impact of patent novelty on firms. Patent novelty can be measured by the technical fields and citations. In this study, I study textual novelty of patents which is measured on patent's abstract using machine learning. This would support us study the influences of the textual novelty of the stock value of the firm.

In text mining, novelty detection can be seen as a "one-class classification" problem, where a model is established to describe "normal" training data (Pimentel Marco et al., 2014) and testing data is compared with training data for novelty. The past has seen a wide variety of textual novelty detection methods. Pimentel Marco et al. (2014) reviewed novelty detection research papers that have been published in the machine learning field. These papers cover topics like distribution-based novelty detection, distance-based novelty detection, probabilistic-based novelty detection, and others.

In the project, I evaluate existing novelty detection methods and develop three textual novelty measures, which are based on TFIDF, BERT (Bidirectional Encoder Representations from Transformers), and auto-encoder. I then choose the best performing method for further analysis on the relationship between patent novelty and firm value. In the project, I find that BERT outperforms the other two models in

deciphering patent textual novelty. We also find that the textual novelty affect firm's abnormal return on the financial market.

2. Related Work

Pimentel Marco et al. (2014) reviewed novelty detection research papers that have been published in the machine learning field. There are three approaches, distance-based, distribution-based, and classification-based methods.

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). For example, the entropy of the class probability distribution is used for novelty detection (Hendrycks & Kevin Gimpel, 2016).

The classification-based method build classifiers to classify whether test data belongs to the training data of normal data (Markou & Singh, 2003). One of the deep learning for novelty detection -- the autoencoder model extracts 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 summarizes the major studies on textual novelty detection. It describes the categories of textual methods, the methods used and evaluation matrices.

Table 1. Existing Studies on Textual Novelty Detection

Studies	Category	Method	Evaluation Metrix
(Shibayama et al., 2021)	Distance-based method	Q-percentile Similarity Method	Pearson Correlation
(Gerken & Moehrle, 2012)	Distance-based method	Maximum Similarity Method	Spearman's rank correlation coefficients, Recall and Precision
(Luo et al., 2022)	Distance-based method	Maximum Similarity Method with BERT	Correlation
(Hautamaki et al., 2004)	Distance-based method	KNN Graph	Receiver Operating Characteristics (ROC)
(Adarsh et al., 2021)	Classification-based method	Using Latent Dirichlet Allocation with Auto-Encoders	Precision and recall
(Mei et al., 2018)	Classification-based method	Using Semantic Clustering and Autoencoders	Pearson correlation
(Bhattarai et al., 2020)	Classification-based method	Tsetlin Machine Text Classifier	Accuracy
(Hendrycks & Kevin Gimpel, 2016)	Distribution-based method	Threshold decision in PDF	Area Under the Receiver Operating Characteristic curve (AU- ROC), and Area Under the Precision-Recall curve (AUPR)

3. Textual Novelty Measure Development

Textual novelty detection problem can framed as follows: 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. In this project, we develop three methods for textual novelty.

3.1 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:

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

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

$$idf(t,D) \coloneqq \ln\left(\frac{N}{1+|\{d \in D | t \in d\}|}\right) \quad (2)$$

$$= \ln\left(\frac{Total\ number\ of\ documents\ in\ the\ corpus}{Number\ of\ documents\ with\ term\ t\ in\ them}\right)$$

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:

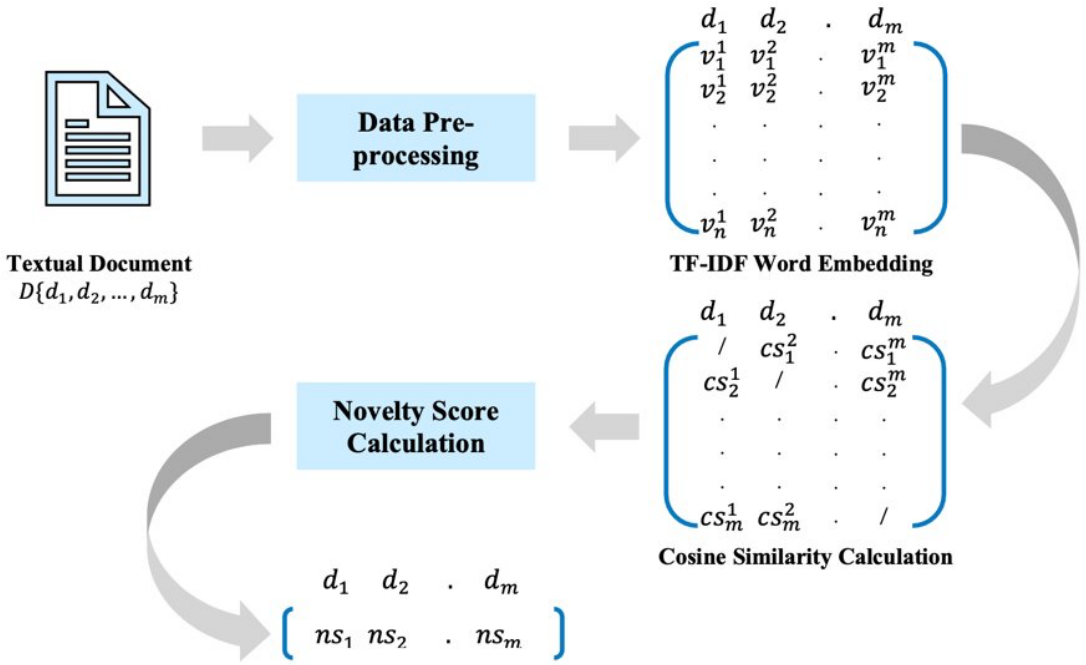
$$cosine(\boldsymbol{v}, \boldsymbol{w}) = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{|\boldsymbol{v}||\boldsymbol{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}} \quad (3)$$

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:

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

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

Figure 1. TFIDF-based novelty calculation process



$d_i =$ 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

3.2 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:

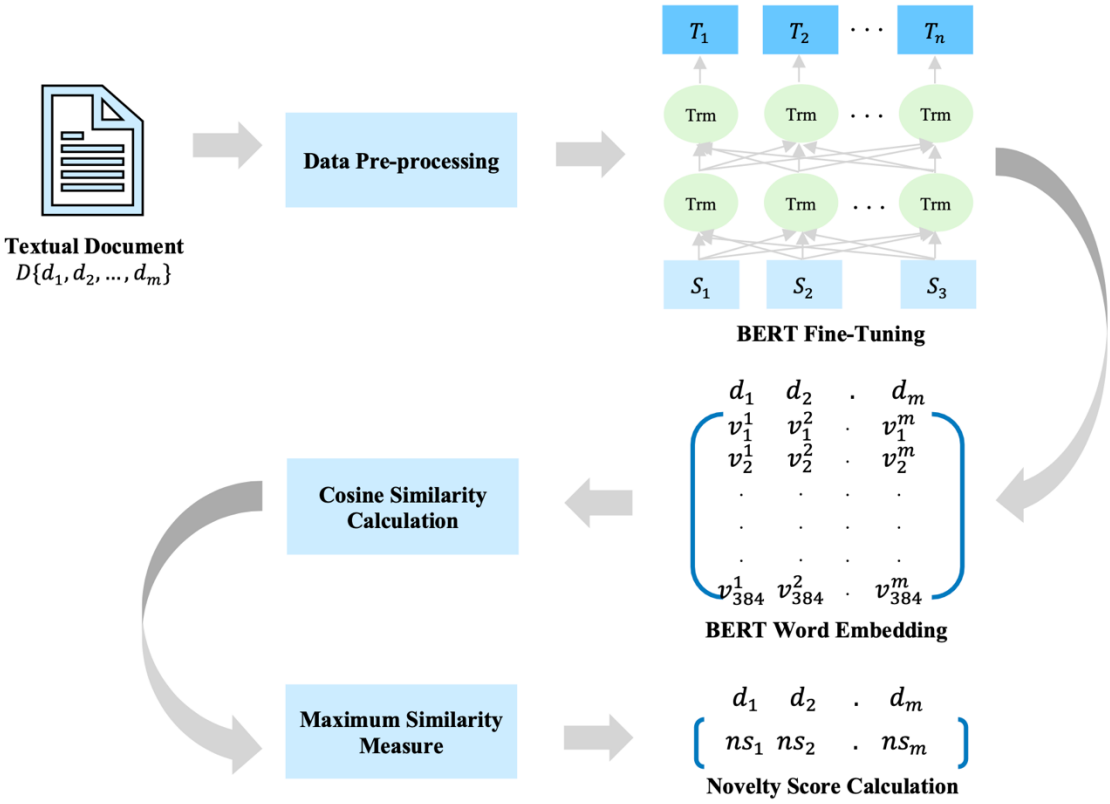
$$BERT\_NoveltyScore(p,D)=1-MaximumSimilarity(p,D) \quad (6)$$

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

$$Normalized\_NoveltyScore_i = \frac{NoveltyScore_i - NoveltyScore_{min}}{NoveltyScore_{max} - NoveltyScore_{min}} \quad (7)$$

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*

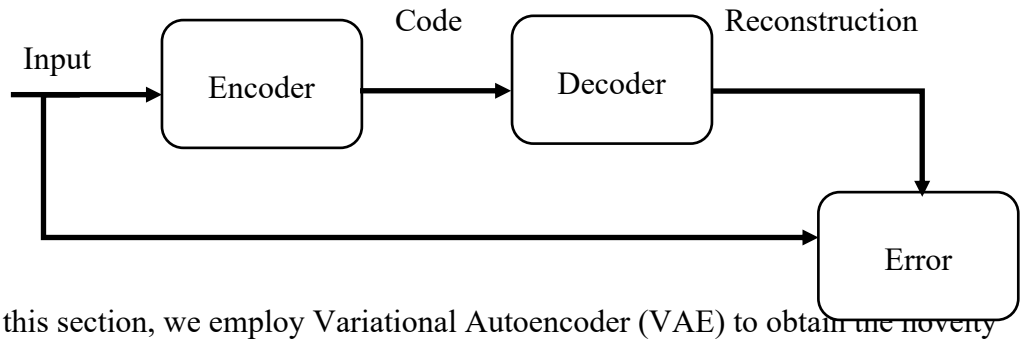
$v_n^m = n^{th}$ *word embedding vector of* m^{th} *document*

$ns_i =$ novelty score for the i^{th} data instance

3.3 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.)

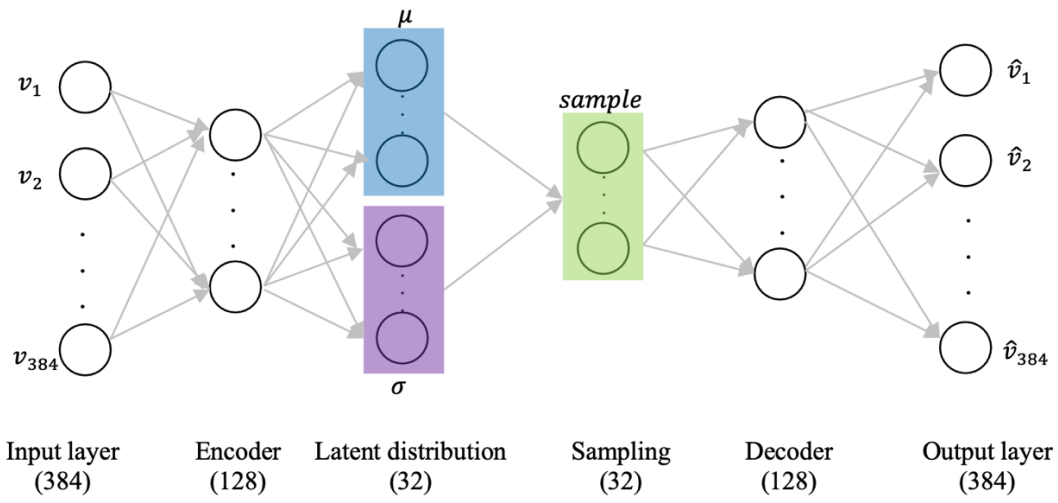
Figure 3. The concept of reconstruction error in autoencoder

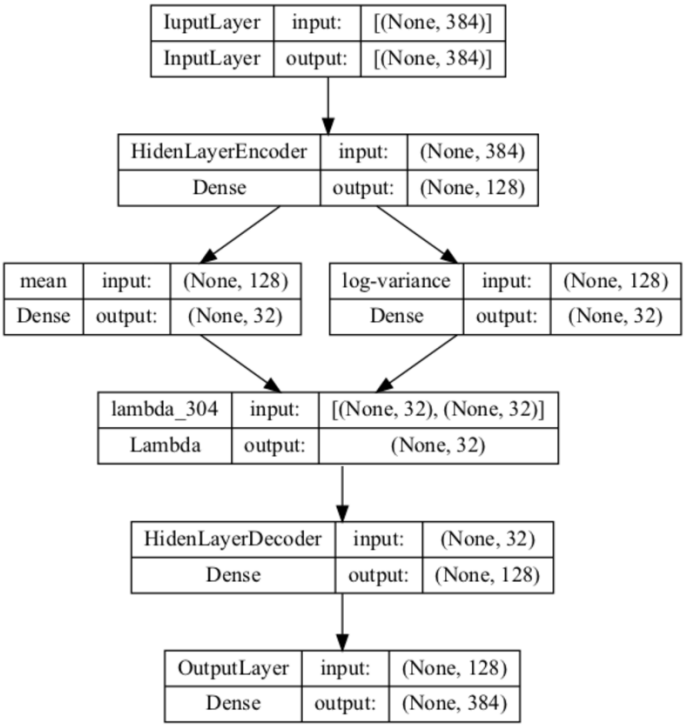


In this section, we employ Variational Autoencoder (VAE) to obtain the novelty scores. VAE is an autoencoder whose latent distribution is regularized during training, sampling through a normal distribution to ensure its latent space is well-characterized for better results.

Here, we use SBERT to generate 384-dimensional dense vectors for each document. The VAE used in this study is a seven-layer deep neural network, with the input layer (Layer 1) having 384 features, the encoding layer (Layer 2) having 128 features, the bottleneck layer (Layers 3, 4, and Layers 3 and 2) having 32 features, the decoding layer (Layer 6) having 128 features, and the output layer (Layer 7) having 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





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):

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

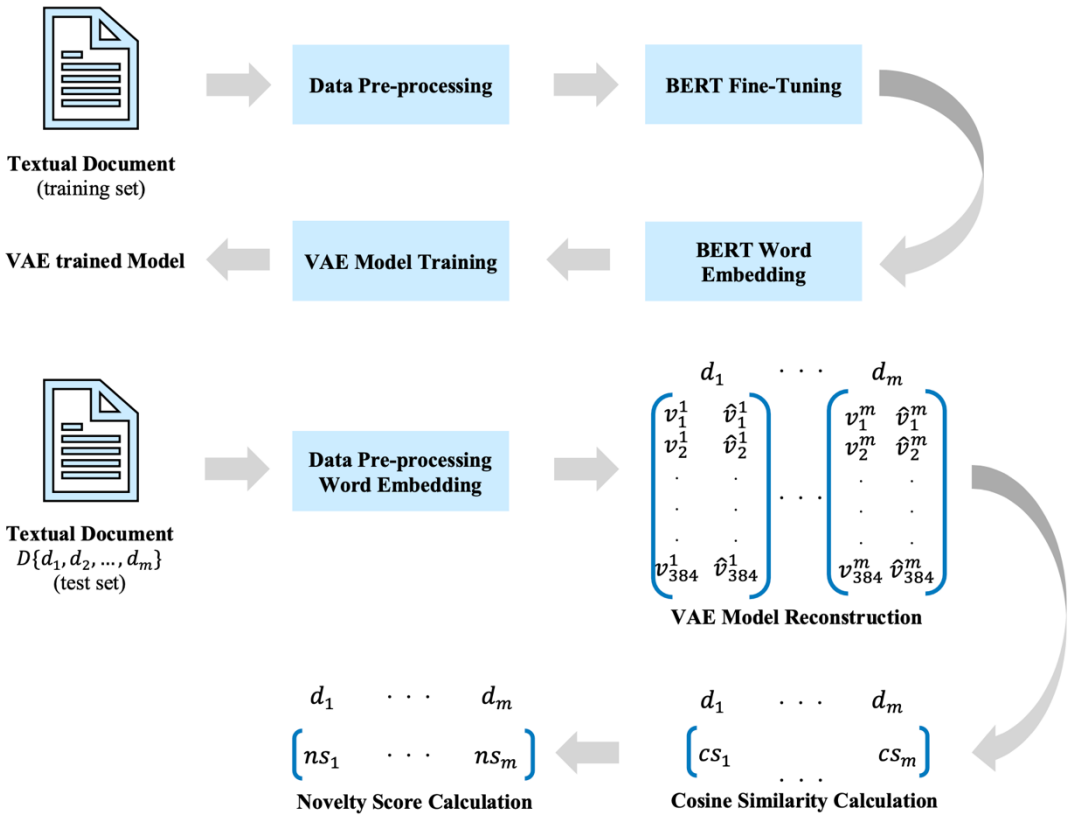
$NoveltyScore_i =$ novelty score for the i^{th} data instance

$p_i =$ 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



4. Evaluation

Before employing these measures, we evaluate their performances whether they are definitely capable of capturing the novelty of the text.

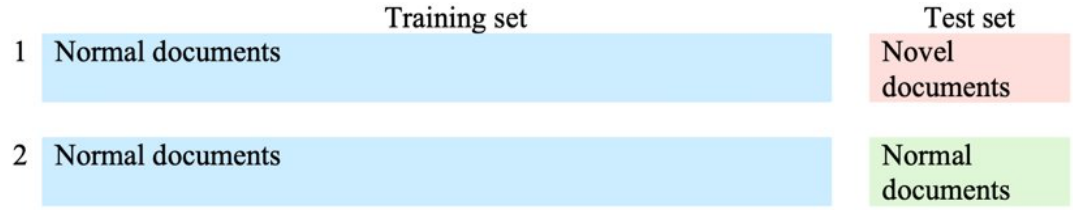
4.1 Evaluation Dataset

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.

4.2 Evaluation Framework

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 with only normal documents and one with 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



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.

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, as seen in Equations (9). To determine the difference between normal and novel documents, we calculate the correlation between the novelty score and the novelty/normal classes.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}} \tag{9}$$

$$where\ r = \text{Pearson correlation coefficient}$$

In addition, we calculate the Kolmogorov-Smirnov (KS) test, which is a non-parametric statistical test that can be used to compare the distributions of two comparison groups to determine if they are significantly different from each other, as follows:

$$KS = Maxmum|F(X) - F(Y)| \tag{10}$$

$$where\ F(X) = \ cumulative\ frequency\ distribution\ of\ X$$

$$F(Y) = \ \ cumulative\ frequency\ distribution\ of\ X$$

Besides, we calculate the Jaccard coefficient and Dice coefficient to measure the similarity/overlap level between two comparison groups, as seen in Equations (11) and (12) 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 (13).

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{11}$$

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{12}$$

$$Jaccard_{continuous(A,B)} = \frac{a^2 r(1+r) - 2rax}{2a^2(1+r^2) - a^2 r(1+r) + 2rax} \tag{13}$$

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

4.4 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 seem to have novelty scores between 0.8 and 1, whereas normal classes seem to have an irregular distribution, as shown 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

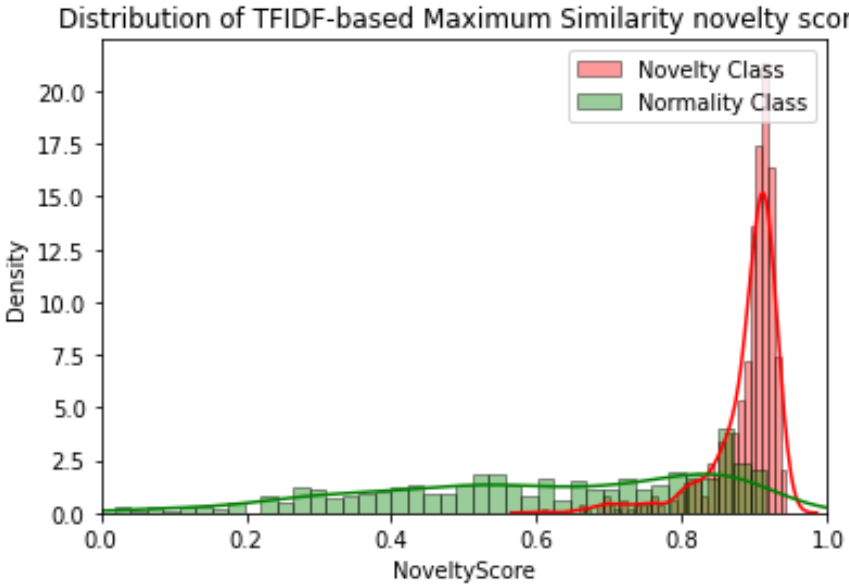


Figure 8. Novelty distribution of BERT-based method between two class

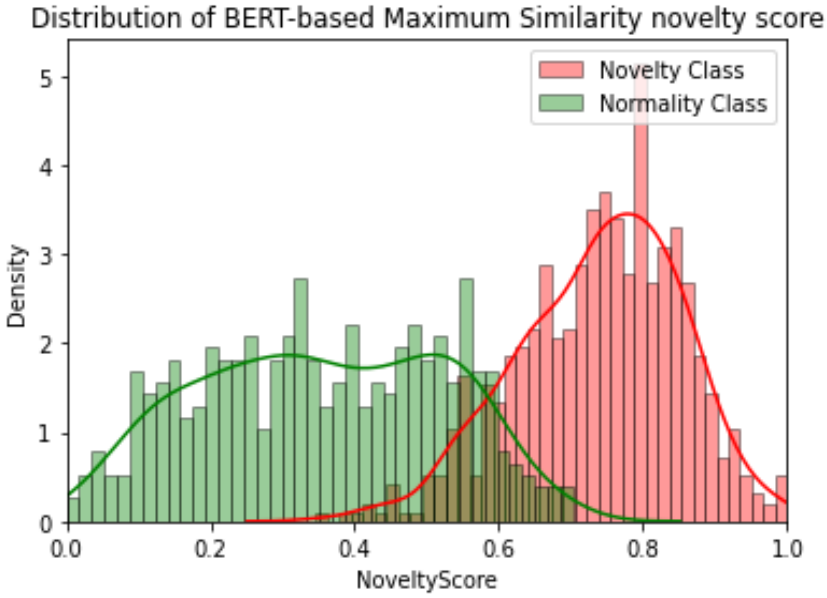


Figure 9. Novelty distribution of VAE method between two class

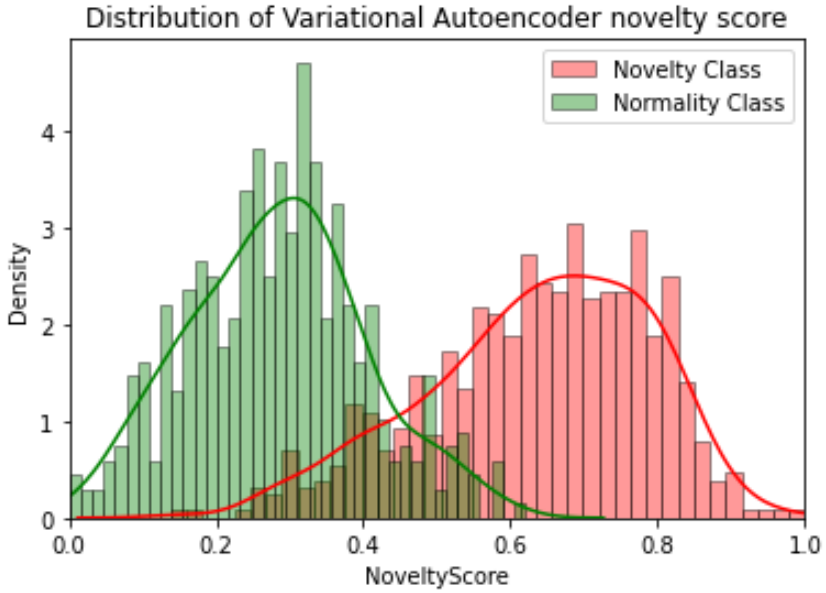


Table 2. Measure Performance Comparison

Measures	**Pearson correlation coefficient**	**Kolmogorov-Smirnov test**	**Jaccard coefficient**	**Jaccard coefficient (continuous version)**	**Dice coefficient**
TFIDF-based Maximum Similarity	0.679	0.698	0.313	0.063	0.477
Bert-based Maximum Similarity	0.808	0.865	0.301	0.213	0.462
Variational Autoencoder	0.782	0.802	0.367	0.321	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 has a range of -1 to 1, with -1 indicating a perfect negative correlation, 1 indicating a perfect positive correlation, and 0 indicating no correlation at all. 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.

When comparing two probability distributions, the KS test determines whether they come from the same distribution. It has a range of 0 to 1, with 0 indicating complete overlap and 1 indicating no overlap. As seen, the KS test of the Bert-based Maximum Similarity method is 0.865. It is the highest among the three methods, indicating the distribution of the novelty scores for novelty/normal classes produced by the Bert-based method is significantly, compared with Variational Autoencoder and TFIDF-based Maximum Similarity.

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. The Jaccard coefficient and Dice coefficient of the Variational Autoencoder is respectively 0.367 and 0.537. The Jaccard coefficient and Dice coefficient of the Bert-based method is respectively 0.313 and 0.477. However, they are both slightly lower than that of the Bert-based method.

5. Case Study: Firm Market Value Analysis

5.1 Data

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. 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.

I collect a name list of biotech companies from The U.S. Securities and Exchange Commission. We select SIC: 2836 —Biological Products, Except Diagnostic Substances in Standard Industrial Classification Codes as a sample of biotech firms. 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.

5.2 Patent Data Processing and Patent 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.

To measure the degree of originality of each new patent, it is necessary to compare it with a set of earlier patents that have been published and a set of cited

patents in the new patent, from which to calculate the textual novelty score, which is simply to find the difference between the new patent and the historical patent. The novelty score is between 0 and 1, with 1 being the most novel and 0 being less novel. After the calculation of the textual novelty score of each patent was completed, the statistics of patent's novelty score for three measures was obtained, as seen in Table 3.

Table 3. The Statistics of Patent's Novelty Score between three Measures

	TFIDF-based Maximum Similarity	Bert-based Maximum Similarity	Variational Autoencoder
Mean	0.55	0.48	0.41
Standard Deviation	0.27	0.22	0.1

Figure 10. Novelty Score distribution of TFIDF-bsed method

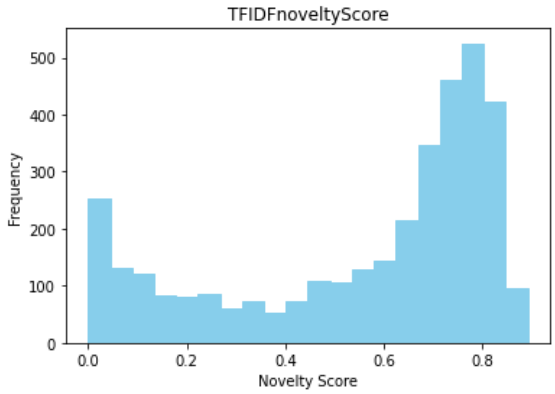


Figure 11. Novelty Score distribution of BERT-based method

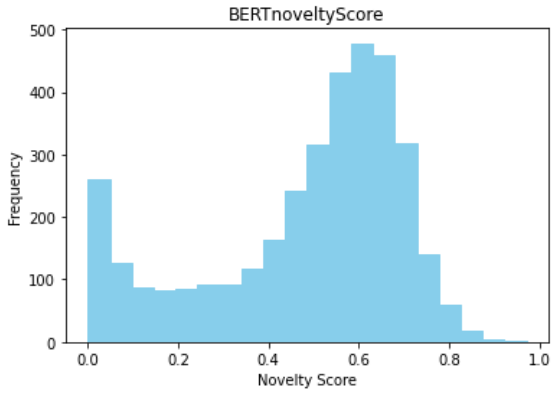


Figure 12. Novelty Score distribution of VAE method

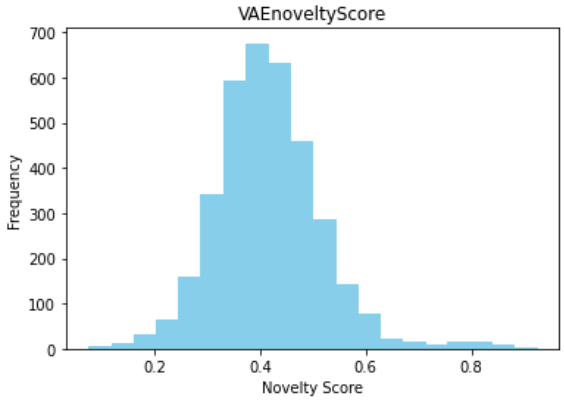


Table 4 the correlation between three different methods for measuring novelty scores, namely TFIDF-based, Bert-based, and Variational Autoencoder (VAE) methods. Table 4 showed that there is a strong positive correlation (0.9538) between the TFIDF-based and Bert-based Maximum Similarity methods. The correlation between VAE and the other two methods is moderate (0.5436 for TFIDF-based and 0.5997 for Bert-based).

Table 4. Patent's Novelty Score Correlation between three Measures

	TFIDF-based Maximum Similarity	Bert-based Maximum Similarity	Variational Autoencoder
TFIDF-based Maximum Similarity	1	0.9538	0.5436
Bert-based Maximum Similarity	0.9538	1	0.5997
Variational Autoencoder	0.5436	0.5997	1

5.3 Regression Analysis

This section describes a regression analysis that investigates the effect of textual patent novelty on firm value by combining using three methods. The analysis was conducted on a monthly dataset in Ordinary Least Squares regression to examine the

effects of textual innovation novelty on firm value. The results indicate that the BERT-based method performed the best in predicting the firm value. The other two measures do not identify significant correlation between patent novelty and firm abnormal return.

As shown in Table 5, there is a positive correlation between BERT-based textual novelty score and firm value, with the coefficient of the BERT-based method having positive (0.092). The findings are trustworthy because the p-value is 0.042, which is significant at the 0.05 level.

Table 5. BERT-based method in OLS regression

OLS Regression Results

Dep. Variable:	abnretlog	R-squared:	0.138
Model:	OLS	Adj. R-squared:	0.132
Method:	Least Squares	F-statistic:	29.24
Date:	Wed, 22 Mar 2023	Prob (F-statistic):	6.91e-12
Time:	19:12:02	Log-Likelihood:	262.94
No. Observations:	548	AIC:	-515.9
Df Residuals:	543	BIC:	-494.3
Df Model:	4		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4415	0.062	-7.107	0.000	-0.563	-0.320
bertmsmlog	0.0917	0.045	2.030	0.042	0.003	0.180
USclass log	0.0296	0.010	2.968	0.003	0.010	0.049
volumelog	0.0222	0.004	6.237	0.000	0.015	0.029
turnoverlog	0.3951	0.234	1.691	0.091	-0.063	0.853

Omnibus:	127.697	Durbin-Watson:	1.566
Prob(Omnibus):	0.000	Jarque-Bera (JB):	439.077
Skew:	-1.056	Prob(JB):	4.53e-96
Kurtosis:	6.843	Cond. No.	223.

Table 6. TFIDF-based method in OLS regression

OLS Regression Results

Dep. Variable:	abnretlog	R-squared:	0.132
Model:	OLS	Adj. R-squared:	0.126
Method:	Least Squares	F-statistic:	29.74
Date:	Wed, 22 Mar 2023	Prob (F-statistic):	5.27e-12
Time:	19:12:22	Log-Likelihood:	261.16
No. Observations:	548	AIC:	-512.3
Df Residuals:	543	BIC:	-490.8
Df Model:	4		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4342	0.062	-6.975	0.000	-0.556	-0.312
tfidfmsmlog	0.0648	0.042	1.546	0.122	-0.017	0.147
USclass log	0.0294	0.010	2.983	0.003	0.010	0.049
volumelog	0.0222	0.004	6.113	0.000	0.015	0.029
turnoverlog	0.4043	0.243	1.661	0.097	-0.073	0.881

Omnibus:	129.178	Durbin-Watson:	1.566
Prob(Omnibus):	0.000	Jarque-Bera (JB):	456.511
Skew:	-1.059	Prob(JB):	7.41e-100
Kurtosis:	6.938	Cond. No.	224.

Table 7. VAE method in OLS regression

OLS Regression Results

Dep. Variable:	abnretlog	R-squared:	0.126
Model:	OLS	Adj. R-squared:	0.119
Method:	Least Squares	F-statistic:	20.65
Date:	Wed, 22 Mar 2023	Prob (F-statistic):	1.23e-09
Time:	19:12:37	Log-Likelihood:	259.06
No. Observations:	548	AIC:	-508.1
Df Residuals:	543	BIC:	-486.6
Df Model:	4		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3932	0.060	-6.528	0.000	-0.511	-0.275
vaemsmlog	-0.0198	0.047	-0.418	0.676	-0.112	0.073
USclass log	0.0289	0.010	2.974	0.003	0.010	0.048
volumelog	0.0211	0.004	5.696	0.000	0.014	0.028
turnoverlog	0.3971	0.236	1.685	0.092	-0.065	0.859

Omnibus:	127.710	Durbin-Watson:	1.558
Prob(Omnibus):	0.000	Jarque-Bera (JB):	443.614
Skew:	-1.052	Prob(JB):	4.68e-97
Kurtosis:	6.873	Cond. No.	223.

6. Conclusions

The study aims to investigate the relationship between the novelty of patent textual data and the value of a firm. In this project, I present a comparison of three different methods for measuring textual novelty - Bert-based Maximum Similarity, Variational Autoencoder, and TFIDF-based Maximum Similarity. Evaluation show the Bert-based method shows the highest correlation and KS test scores, while the

Variational Autoencoder and TFIDF-based Maximum Similarity show relatively high scores as well. Then, we conduct analysis on a monthly dataset using Ordinary Least Squares regression on nascent firms in biotechnology to study the impact of textual innovation novelty on firm value, and the results show a positive relationship between BERT textual innovation novelty and firm value. This study demonstrates the importance of novelty detection methods in improving performance and efficiency in identifying the impact of patent textual novelty on firm value.

There is still a room for improve this study. First, 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. More datasets may need to be included in the evaluation in subsequent studies. Second, we will explore further new novelty detection methods. 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.

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 and assist in Ordinary Least Squares regression analysis. 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

Implement Code:

https://drive.google.com/drive/folders/1miQdKyoY20MdkkNoOVVm5uCxNyuOl0tu?usp=share_link