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

Yukyee CHAN(Author)
Wanyue Zhou
Jianming Huang

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, full-text 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 full-text 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.

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.
(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

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

First, we employ a simple maximum similarity measure that calculates the largest similarity between p and every document d_i in D:

$$TN\_MS(p,D)=1-max_i(cosine\ similarity\ (p,\ d_i))$$

Here, we define similarity as the cosine similarity build upon the TF-IDF vector representation of documents. 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 word appears in the document:

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

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

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

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:

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

2.2.2 Bert-based Maximum Similarity Method

(To be continued)

2.2.3 Autoencoding

(To be continued)

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), 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.

To discover the difference between normal documents and novel documents, we then divided the task into two experimental steps. The first step is novelty score calculation. We assign the word embedding vector representation of each vocabulary to each textual information and then employ a simple maximum similarity measure that calculates the novelty scores for both the normal and novel documents (Gerken & Moehrle, 2012). 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), we calculate the Pearson correlation coefficient between the maximum similarity and two comparison groups. To figure out the difference between normal documents and novel documents, we calculate the correlation between the novelty score and the novelty/normal classes.

2.5 Performance Comparison

Figure 2. Distribution of maximum similarity score between two class

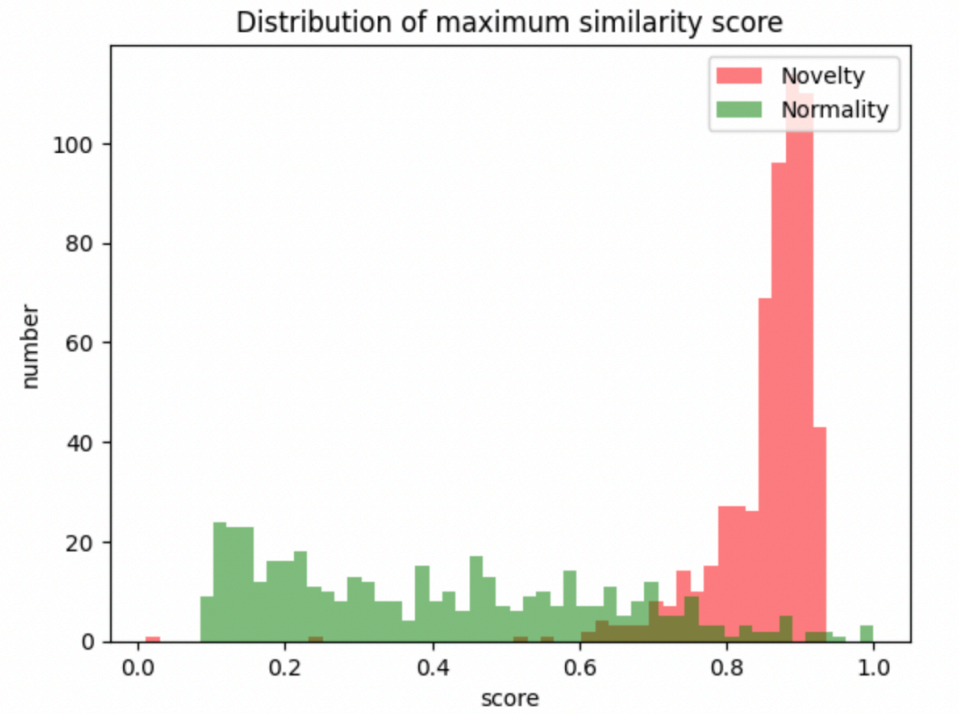

Figure 2 showed that the difference in the maximum similarity between normal documents and novel documents. It seems that most novel classes have scored around 0.8 and 1, and normal classes have irregular distribution.

Table 2. Measure Performance Comparison

Measures	Pearson correlation coefficient	Jaccard Coefficient	Jaccard Coefficient (continuous version)	Overlap Coefficient
Maximum Similarity	0.675	0.301	0.056	0.926
Bert Similarity	0.641	0.449	0.187	0.727
Autoencoding				

The experimental results for all measures are shown in Table 2. As seen, we found that Pearson's coefficient of the novelty score and the novelty/normal classes is 0.675. That means they are highly correlated.

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 full-text 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.

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