Operational Risk Text Classification Using Weak Supervision

Xinyue Ma

1. Introduction

Operational risk events occur in almost all processes of banks' operations. According to Basel Committee on Banking Supervision (2003), Operational risk is defined as the risk of losses resulting from inadequate or failed internal processes, people, systems, or external events. There are different types of operational risks like fiduciary breaches, aggressive sales, breaches of privacy, account churning, failure of IT systems, health and safety, litigation, and misuse of confidential information.

According to the Basel II regulations for banks, there are seven classifications of operational risk events, including internal fraud, external fraud, employment practices and workplace safety, clients, products and business practices, damage to physical assets, business disruption and system failures, execution, and delivery and process management. Operational risk events can relate to small errors in the daily business that occur at high frequency and low losses, as well as large-scale frauds and natural disasters that occur at low frequency but result in high losses. Different industries are exposed to different types of operational risks. For example, the agriculture industry is very sensitive to extreme weather and pollution, which means the probability of occurrence of damage to physical assets and system failure events will be higher than, for example, the consumer service industry. Managing operational risk is very important for banks. Some activities for risk management include risk identification, risk control assessments, and governance frameworks.

Because of broad causes and complex types, it is difficult to identify operational risk events. Managers can have a picture of how many types of factors leading to bank operational risk totally and using such information to further manage and control operational risk. However, outsiders, like investors, don't have access to daily operational data. Researchers have found that we can identify operational risk events from textual disclosures of financial statements, which contain all related risk events. For example, most US public companies are required by the US Securities and Exchange Commission (SEC) to disclose significant developments, that occur between filings of the Form 10-K or Form 10-Q, in the format of the SEC 8-K form. Major organizational and company events that would necessitate the filing of a Form 8-K include bankruptcies or receiverships, material impairments, completion of acquisition or disposition of

assets, and departures or appointments of executives. 8-K fillings are important for investors to make investment decisions and for managers to manage risks. Some trading researchers predict stock returns by developing trading strategies using text-based features from 8-K documents and sentiment analysis.

This research uses weak supervision to build a training set to identify operational risk events from 8-K fillings. Text classification is a standard task in machine learning. The standard process relies on supervised and semi-supervised approaches. A successful machine learning model usually requires lots of human efforts to label a large set of training data. Since we had no sample labels available, we utilize the weak supervision strategy to train machine learning models on the weakly labeled training data created by automated methods.

2. Literature Review

Weak supervision is a simple and adaptable approach leveraging programmatically created weakly labeled training sets. Weak supervision has been widely applied in other common NLP tasks including knowledge-base completion, sentiment analysis, and information retrieval [3]. Some researchers have applied the weak supervision approach to classifying news documents into operational risk events categories. The approach to enrich the category labels is a combination of human expertise and language models[4].

Snorkel is a system that enables users to train models without hand labeling any training data, developed by Stanford. It allows users to write labeling functions that capture domain knowledge and resources. Snorkel significantly increase the speed of model building and predictive performance, compared to hand labeling[5].

Because of the great performance and convenience of Snorkel, we utilize Snorkel for our weak supervision task to generate the weak labeled train dataset. Then we train classification models based on the new dataset.

3. Empirical analysis

We use Snorkel to build a training set for identifying sections of 8-K filling including operational risk events or not. The key advantage of this approach is its simplicity and adaptability. We can just add new labeling functions to the model.

3.1 Dataset

An 8-K filing is a report of unscheduled material events or corporate changes at a company that could be of importance to the shareholders or the SEC. SEC required companies to file an 8-K to announce significant events and they can use 8-K as needed, instead of filed annually or quarterly. The SEC outlines nine sections that require form 8-K, where each section may have subsections. Our focus is on identifying operational risk events from these sections, which are reported as items in 8-K fillings. As thousands of events had to be identified, weak supervision seemed to be a promising approach to assist in the classification tasks. We use these as training datasets and generate labels for these data points with weak supervision.

To evaluate the performance of our model, we construct a new dataset with six operational risk events, three market risk events, and one credit risk event. The label for operational risk is 1 and or other risk evens is 0. The data comes from the Reserve Bank of New Zealand. This is a small set and is only used for the final evaluation of our classifier.

3.2 Data Preparation

To get the training dataset, we randomly draw ten 8-K fillings documents from the SEC website per quarter from 1995-2020, and get 1040 documents in total. After downloading the index list, we extract 8-K text from the website. There are two formats of 8-K fillings as the old and the new versions, which vary in structures. The new fillings include structured HTML code. We process these documents differently. We clean and preprocessing textual data with multiple steps, to remove headers and extract useful text from each 8-K filling document. This gives us 2575 items of text. First, we remove negation words from stop words because they might be useful to identify operational risk. Stopwords are words that have no significance when constructing meaningful features from the text, such as a, an, the. HTML tags don't add much value for analyzing text, so we remove unnecessary HTML tags to retain useful textual information. Accented characters are converted to standard ASCII characters since we only analyze the English language. We use regular expressions to remove special characters and symbols which will add extra noise in unstructured text and standard words to their base stem. Finally, we tokenize text using lemmatize to remove a word affixed to get the base form of a word.

3.3 Labeling Function

Labeling functions are noisy, programmatic rules, and heuristics that assign labels to unlabeled training data. It is a powerful way to encode domain knowledge and other supervision sources programmatically. When building labeling functions, we try to brainstorm and also pay attention to the correlation between labeling functions.

3.3.1 Keyword Search

We apply the TF-IDF method to a text document, which contains the definition of seven Basel II operational risk events and typical examples, to get keywords for each classification. Seven labeling functions are developed based on matching texts with these keywords. After constructing seven keyword searching labeling functions, we apply them to our train dataset to get a label matrix, where each column is for one labeling function and each row is for the data point. We evaluate the performance by calculating the coverage and overlap of these labeling functions. From Table 1, we notice that the employment practice and workplace safety labeling function has the highest coverage and the business disruption and system failures labeling function have the lowest coverage rate. This can be explained by the low frequency of occurrence of system failures. Or because we have fewer keywords for the category. The overlap between the employment practice and workplace safety labeling function and damage to physical assets labeling function has high overlaps. This can be explained by that both categories are affected by environments.

Table 1: Evaluation of Keyword Searching Labeling Functions

	j	Polarity	Coverage	Overlaps	Conflicts
internalfraud	0	[1]	0.135534	0.081553	0.0
externalfraud	1	[1]	0.142913	0.048544	0.0
employmentworkplace	2	[1]	0.249320	0.235728	0.0
clientsproducts	3	[1]	0.093981	0.088544	0.0
physicalassets	4	[1]	0.168544	0.166990	0.0
disruptionsystem	5	[1]	0.009709	0.007379	0.0
executiondelivery	6	[1]	0.160000	0.121165	0.0

3.3.2 Pattern Matching

We use the similarity between strings to construct a labeling function. Based on the properties of operations, there are lots of string similarity algorithms. We use the token-based method to find similar tokens in both sets. The more the number of common tokens, the more is the similarity between the sets. Tokens of different lengths have equal importance. We use the previous keyword dataset and compute the similarity with it and each data point in the training dataset. We get a low coverage rate. The similarity labeling function doesn't perform as well as others. According to Tabel 2, we get a low coverage rate. The similarity labeling function doesn't perform as well as others.

Table 2: Evaluation of Similarity Labeling Functions

	j	Polarity	Coverage	Overlaps	Conflicts
similarity	0	[1]	0.092039	0.0	0.0

3.3.3 Third-party Models

We use the pre-trained sentiment analyzer provided by TextBlob to build labeling functions. We believe risk events and non-risk events have a different distribution of sentiment scores. Risk events are often associated with negative. We extract the polarity and subjectivity scores and pick the threshold of polarity equal to -0.05 and subjectivity equal to 0.4. Then apply the labeling functions to the text and evaluate the performance. We notice that the subjectivity labeling function has a higher coverage rate than the polarity labeling function. They have the same data points overlaps and conflicts.

Table 3: Evaluation of Third-party Model Labeling Functions

	j	Polarity	Coverage	Overlaps	Conflicts
textblob_polarity	0	[1]	0.149903	0.062136	0.062136
textblob_subjectivity	1	[0]	0.299806	0.062136	0.062136

3.3.4 SpaCy Name Entity Recognition

SpaCy is a natural language processing tool. Considering operational risk events are associated with the person, system, and process. We use it to identify name entities like the person, events, and product in the text. From Table 4, events are not recognized from our training dataset and there are only a few products are recognized.

Table 4: Evaluation of SpaCy Name Entity Recognition Labeling Functions

	j	Polarity	Coverage	Overlaps	Conflicts
has_people	0	[1]	0.063216	0.0	0.0
has_product	1	[1]	0.000395	0.0	0.0
has_event	2	[]	0.000000	0.0	0.0

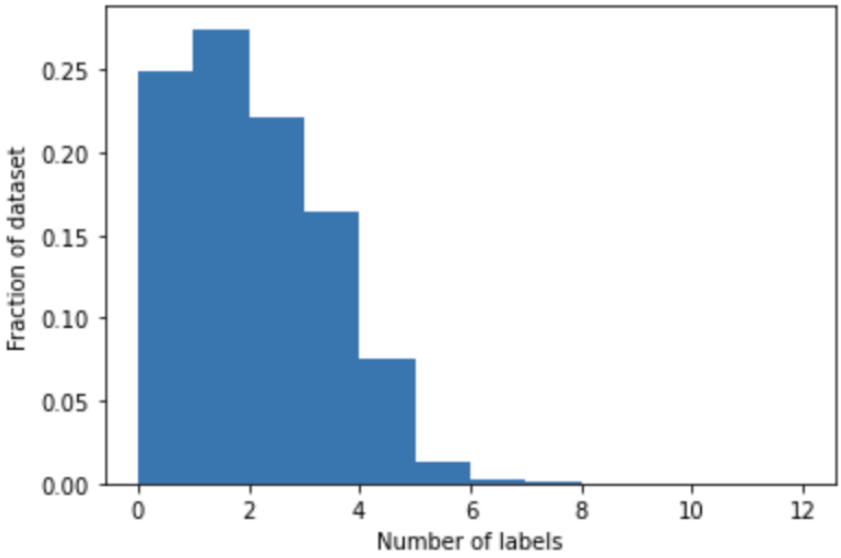
3.4 Aggregation

We combine all labeling functions and generate the label matrices. The labeling functions have various coverage. Using the histogram we can see that over half of the dataset has 2 or fewer labels from labeling functions.

Table 5: Aggregation of All Labeling Functions

	j	Polarity	Coverage	Overlaps	Conflicts
internalfraud	0	[1]	0.137890	0.112604	0.026867
externalfraud	1	[1]	0.145397	0.088898	0.037535
employmentworkplace	2	[1]	0.253655	0.246938	0.101936
clientsproducts	3	[1]	0.095614	0.093244	0.022126
physicalassets	4	[1]	0.171474	0.170684	0.067562
disruptionsystem	5	[1]	0.009878	0.009087	0.003556
executiondelivery	6	[1]	0.162782	0.137495	0.062426
similarity	7	[1]	0.093639	0.086922	0.045832
textblob_polarity	8	[1]	0.152509	0.118530	0.063216
textblob_subjectivity	9	[0]	0.305018	0.208613	0.208613
has_people	10	[1]	0.063216	0.043461	0.005531
has_product	11	[1]	0.000395	0.000395	0.000395
has_event	12	[]	0.000000	0.000000	0.000000

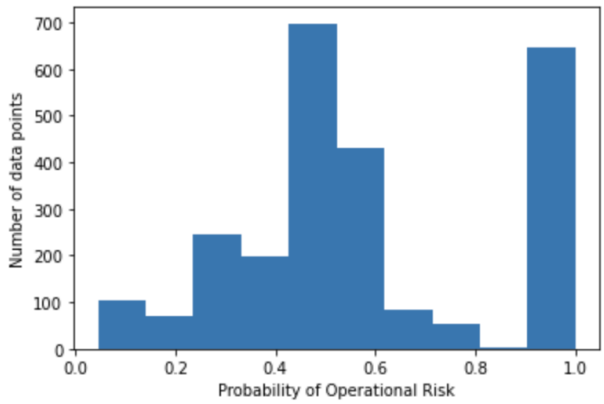
Image 1: Label Frequency



Considering that some labeling functions may be correlated, some data points may be overvoted in the majority vote-based model. Snorkel provides a useful wool LabelModel to combine the outputs of the labeling functions and also produce a single set of noise-aware training labels.

We convert the labels from labeling functions into a single noise-aware probabilistic label per data point. We use the majority vote method on a per data point basis to vote Yes if more LFs voted YES and label it YES and vice versa. Image 3 shows the confidence we have that each data point including the operational risk event. We can conclude that more than one-third of the data points are very likely to include operational risk events.

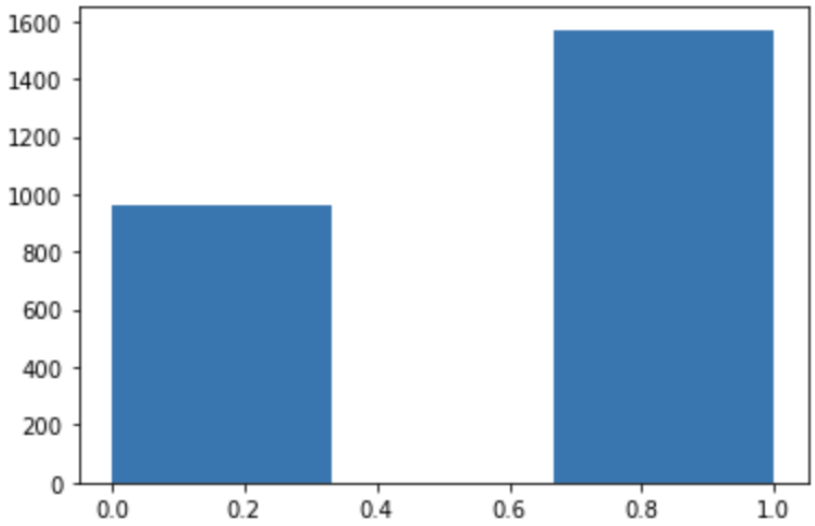
Image 2: Confidence of Operational Risk



3.5 Training Models

Now we get the probabilistic training labels, we can get labels for the dataset. From image 3, we can see that about 60% of datasets are exposed to operational risk events.

Image 3: Frequency of Labels



Then we build the Scikit-Learn classification model, neural network model, and LSTM(long short-term memory) model. Then we compare the performance of these models. As a result, all models have the same accuracy of 60%. This doesn't mean all three models have the same performance, considering the limited number of the test datasets.

Image 4: Neural Network Model

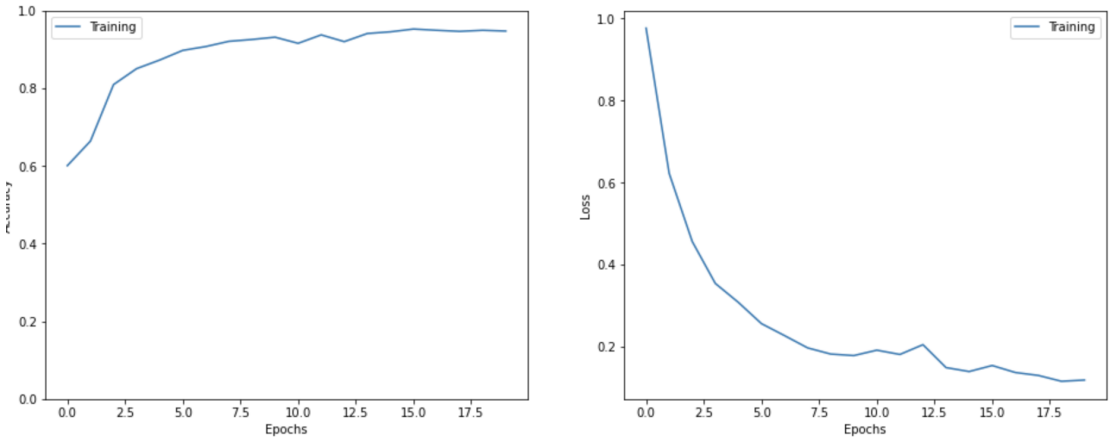
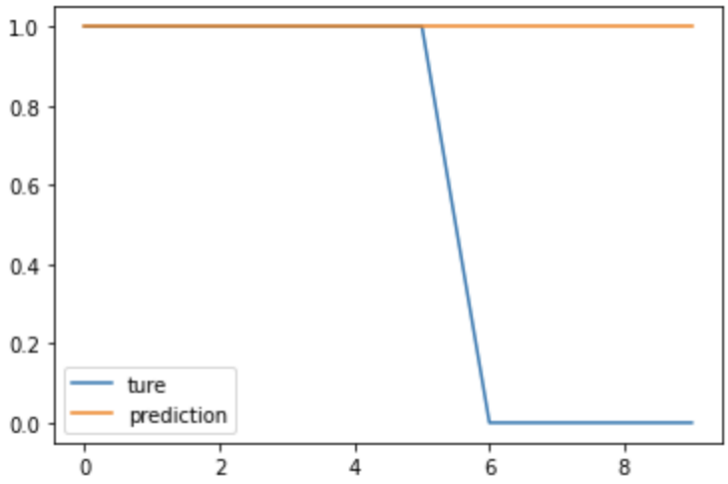


Image 5: Prediction Result of Neural Network Model



4. Conclusion

In this research, we present a method for weak supervision text classification based on the snorkel model. We constructed 13 labeling functions in total and aggregate all labeling functions. Finally, we use these labels to build classification models. In the next step, we can construct other labeling functions using other models. We also can conduct multiply classification tasks to classify each operational event into one of seven Basel II categories. Some events might be exposed to several categories. We can use labels such as high, medium, and low for each category. Comprehensive identification of operational risk events is of great importance for operational risk management. After identifying risks, managers can choose key risk indicators and perform more risk management activities,

Reference

1. https://www.sciencedirect.com/science/article/pii/S1877050918318982
2. https://en.wikipedia.org/wiki/Operational_risk
3. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0723-6
4. https://aclanthology.org/P19-1036.pdf
5. https://arxiv.org/pdf/1711.10160.pdf
6. https://www.cia-ica.ca/docs/default-source/2014/214118e.pdf
7. https://www.ior-institute.org/public/IORKRIGuidanceNov2010.pdf
8. https://itnext.io/string-similarity-the-basic-know-your-algorithms-guide-3de3d7346227
9. https://en.wikipedia.org/wiki/Form_8-K
10. https://www.rbnz.govt.nz/financial-stability/financial-stability-report/fsr-may-2019