Some Practical Analyses of the Judgment Documents of Labor Litigations for Social Conflicts and Similar Cases

Chao-Lin Liu and Yi-Fan Liu

National Chengchi University, Taipei, Taiwan

Abstract

We report two applications of the analysis of the judgment documents of labor litigations. The disputes listed in the judgment documents provide a very good foundation for recommending similar cases with explanations. For this narrowly focused "similarity", we could achieve a 70% accuracy in our recommendations. Analyzing and learning about the disputes that the litigants argued in their cases help the lawyers, the social workers, and ordinary people to know more about and improve their society.

Keywords

Machine learning, legal informatics, civil cases, user interaction, social studies, social services

1. Introduction

The analysis of contents of judgment documents can have a wide variety of applications. Certainly, the results of analysis can be used for legal informatics. The lawsuits, both criminal and civil ones, are about people's lives. Analyzing the judgment documents and deep understanding the causes of the lawsuits can also be useful for understanding our society and thus the needs and expectations of our clients.

The analysis for the judgment documents of the civil cases that are about the support of the elderly may shed light on the family issues of both the plaintiffs and the defendants [14]. If we can do a large-scale analysis of this category of documents, we probably can figure out how to improve the social services of our governments such that people do not have to resort to litigations to solve problems.

Analogously, the analysis of judgment documents of labor litigations leads us to learn about the conflicts between the employees and their employers. This knowledge can be used for legal informatics and can be used for social studies and social services.

In: *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workspace (LegalAIIA 2023)*, held in conjunction with ICAIL 2023, June 19, 2023, Braga, Portugal.

chaolin@g.nccu.edu.tw, 108753213@nccu.edu.tw

0000-0002-4093-1497 (C.-L. Liu)

Identifying similar cases is a more important goal in our current work. If we can identify similar cases for a new case, then we are more empowered to predict the future judgements of the new case. Similar prior cases can also be used to support or to challenge the predicted judgments of a prediction system.

"Similarity" is a subjective idea, however. People may consider two objects similar, even when the others do not [10].

One may approach the task of identifying similar cases from a wholistic or theoretical perspective, e.g., [7] and [9]. We adopt a relatively narrow and more specific focus. We would consider two labor litigations similar if the involved **disputes** between the plaintiffs and the defendants are similar.

By analyzing the disputes recorded in the judgment documents, we gain insights into the labor-related conflicts. By building a system for recommending similar cases based on the clusters of disputes, we would improve the explainability of the recommender system. Experimental results indicate that achieving a 70% accuracy is possible for our recommender. Although this is not a highly impressive result, it is encouraging.

We offer information about the sources of the judgment documents that we used in our research in Section 2, and define our research problems more officially in Section 3. We explain the necessary preprocessing of the online files to make them useful for our studies in Section 4, and present the design concepts of our recommender system in Section 5. In Section 6, we outline the methods of how we use additional information to

assist the readers to acquire information from the disputes recorded in the judgment documents. In Section 7, we report the results of the evaluations of the recommender system, and we offer more technical discussions in Section 8.

2. Source and Selection of Judgment Documents

2.1. Data Source

The Judicial Yuan is the highest government unit that governs the judicial matters in Taiwan. The Judicial Yuan publishes the judgment documents of several types of courts, including the local courts, the high courts, the supreme courts, and some special courts three months after the judgment date whenever possible. Namely, the judgment documents for February judgments will not be published until May. Some of the judgment documents may not be published by legal reasons, e.g., for protecting the minors or for protecting the litigants. In addition, the contents of the documents were also anonymized for privacy reasons.

After these cautionary steps, the publishable documents were placed on the Internet, and we will refer to this website as TWJY for representing the Taiwan Judicial Yuan.[1]

As of May 2023, there are about 18 million documents available on the website. The website offers documents for judicial decisions of as early as January 1996. In the first few years, only documents of a limited number of special courts were available. The coverage started to broaden since 2000.

2.2. Selection

After downloading all of the published documents from the TWJY, we need to identify the documents that belong to the category of labor litigations. The TWJY includes documents for a wide variety of lawsuits, certainly including the criminal and the civil cases, and there are myriad subcategories of lawsuits of them.

Each document in the TWJY is a JSON file, and adopts a top-level structure. The JSON file has several fields for document: JID is the long identification number; JYEAR is the year when the case occurred in terms of Taiwan calendar; JCASE is the abbreviated code for the type of the lawsuit; JNO is the short identification number for the lawsuit; JDATE is the date for the current judgment in terms of the Western calendar; JTITLE is the category of the lawsuit, and JFULL is the full text for the judgment document.

Therefore, using the contents of JCASE and JTITLE fields to find relevant documents is a basic step. We focus on the judgments of the local courts, where the judges would consider the material parts of the lawsuit. Using the codes in the JCASE and the JTITLE fields could help us exclude cases for appeal.

However, the filtering of relevant and usable documents needs more steps. Sometimes, even when the JCASE and JTITLE fields seem to qualify a document, we may find clues in the JFULL field that indicates the case does not meet our needs.

In our current study, we look for cases in which the courts explicitly recorded the disputes between the plaintiffs and the defendants in the judgment document. According to the Code Civil Procedure, which governs how the judges, the plaintiffs, and the defendants should proceed with the civil lawsuits, the litigants should prepare a list of their disputes. These disputes provide important information about the lawsuits, and help the lawsuits to proceed more effectively. Nevertheless, not all of the judgment documents would record the disputes.

We will explain that we build our current work on the assumption that the published documents would contain the list of disputes in Section 3. Hence, at the stage of filtering for useful documents, we look into the JFULL field to make sure that the documents meet this requirement.

At this moment we selected **3835** cases from 21 local courts in Taiwan. Figure 1 shows the distribution over the years when the cases took place. The horizontal axis shows the years, and the vertical axis shows the number of cases. The long-term trend is that the number of cases has been increasing over the years.

Figure 2 shows the sources of the selected cases. The horizontal axis lists the code in English letters for the 21 local courts, and the vertical axis shows the contribution of the individual local court in the selection. Most of the selected came from the top five courts, which happened to locate in the metropolitan areas.

[1] Open data of Taiwan Judicial Yuan (TWJY): https://opendata.judicial.gov.tw/, last accessed 2023/05/05

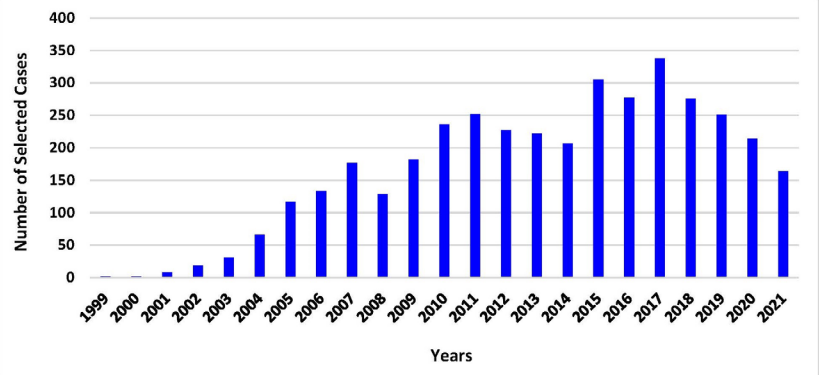

Figure 1: The temporal distribution of the selected cases

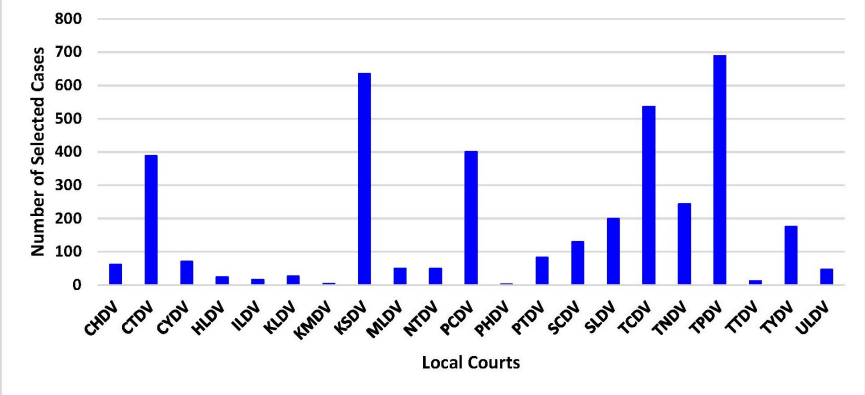

Figure 2: Most of the selected cases came from the top five local courts.

3. Problem definitions

We aim at identifying similar cases of labor litigations. For this task, when given three cases, C_X, C_Y, and C_Z, we want to determine whether two C_Y or C_Z are more similar to C_X, denoted by $C_Y \succcurlyeq_{C_X} C_Z$. To this end, we hope to find a function f of two cases, such that the relationship (1) holds.

$$C_Y \succcurlyeq_{C_X} C_Z \Rightarrow f(C_X, C_Y) \geq f(C_X, C_Z) \qquad (1)$$

This formulation of relatively similarity should be intuitive and convincing, but it is also incomplete, if not controversial. People of different interests or needs may have different definitions for "similarity". A person might be interested in finding cases in which the judges' decision styles are similar or in finding cases in which the causes of the lawsuits are similar. "Similarity" is a dynamic concept, and it is not easy to define with a function.

Admitting this intrinsic diversity in "similarity", we must define our perspective of "similarity." Since preparing the list of the disputes for a civil litigation is required (or highly recommended) by law, we would base our definition for the relative similarity between two civil cases on the similarity between the disputes of these two cases. We understand that this definition might not be very general, but should be useful in legal practice.

Let D_X, D_Y, and D_Z denote the list of disputes of C_X, C_Y, and C_Z, respectively. Assume that we can create a mechanism, g, for computing the similarity between two lists of disputes, then we can rewrite relationship (1) as relationship (2).

$$C_Y \succcurlyeq_{C_X} C_Z \Rightarrow g(D_X, D_Y) \geq g(D_X, D_Z) \qquad (2)$$

4. Data Preprocessing

For each of the judgment documents that we selected in Section 2.2, there is a section in each of the document that itemizes the disputes between the plaintiffs and the defendants. The format looks like the following, although the formats may vary and they are actually in Chinese. (See Appendix A for real examples.)

The disputes are following.
1. **dispute-1 statement**
2. **dispute-2 statement**
3. **…**

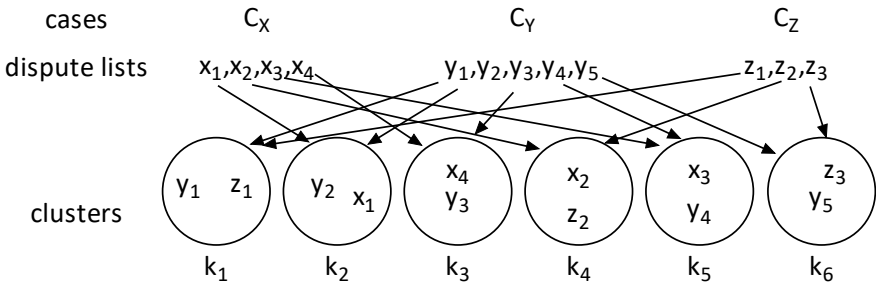

Figure 3: An illustration for using clustering for defining the function g

One direct challenge is that the number of disputes vary from case to case. Hence, it is not easy to define the function g in (2).

The second challenge is that, although the dispute statements are normally not very long, they can still contain a few sentences, and the number of sentences can vary from dispute to dispute.

The third challenge is that the statements contain specific information about their belonging cases, e.g., person names, place names, and time expressions. Comparing these named entities between two litigations may not make very much sense.

For the third challenge, our programs would recognize using the NER techniques (named entity recognition) to identify the named entities, and would replace those specific nouns (or noun phrases) with more general terms, e.g., someone, somewhere, and sometime. This would make the dispute statements more comparable. We refer to this as a "**blurring**" step. (See Appendix A for real examples.)

For the second challenge, we try to do sentence splitting in Chinese in some of our experiments. It is well known in the field of natural language processing that the Chinese texts do not use delimiters between words. Hence, sometimes, we need to do word tokenization. It is less well known that there is no specific sentence boundary in Chinese texts either, even though there is a punctuation mark " ◦ ", whose function is supposed to be similar to the period (.) in English.

Like many other researchers, we chose to split the statements at specific punctuation marks, e.g., " ? ". This punctuation mark appears often in the lists of disputes in the TWJY documents, partially due to that this is how the courts recorded some undermined questions in the litigations. Hence, our program might split multiple questions in one dispute statement into individual disputes in some experiments. (See Appendix B for a real example.)

5. Clustering: Motivation and Issues

5.1. Main Ideas

We outline the basic idea for how we may solve the first challenge with the simplified situation shown in Figure 3. Assume that C_X has four disputes, $\{x_1, x_2, x_3, x_4\}$, in its dispute list, that C_Y has five disputes, $\{y_1, y_2, y_3, y_4, y_5\}$, and that C_Z has three disputes, $\{z_1, z_2, z_3\}$. Now, assume that we put these 12 disputes in a pool, and that we run a clustering procedure to cluster them into six clusters [5], as shown at the bottom of Figure 3. Let's name these six clusters, k_1, k_2, k_3, k_4, k_5, and k_6.

Hence, as we may inspect, in Figure 3, and find that the disputes of C_X and C_Y appear together in $\{k_2, k_3, k_5\}$, and that the disputes of C_X and C_Z appear together only in $\{k_4\}$.

If the results of clustering the pooled disputes is like topic modeling, which we normally hope so, then, qualitatively, the three dispute pairs of C_X and C_Y are similar, and one dispute pair of C_X and C_Z is similar. Namely, if we define the assignments in (3), then we have $C_Y \geqslant_{C_X} C_Z$ based on (2)

$$\begin{cases} f(C_X, C_Y) = g(D_X, D_Y) = 3 \\ f(C_X, C_Z) = g(D_X, D_Z) = 1 \end{cases} \quad (3)$$

More specifically, we can define the g function for two cases as the number of clusters in which their disputes appear together.

5.2. Some Technical Discussions

The example in Figure 3 shows the main ideas, and the procedure shows us a way to compute the similarity between any two cases at the same time, even though the number of disputes in their dispute lists are different.

Figure 4 shows the distribution of the original number of disputes in the dispute lists of the

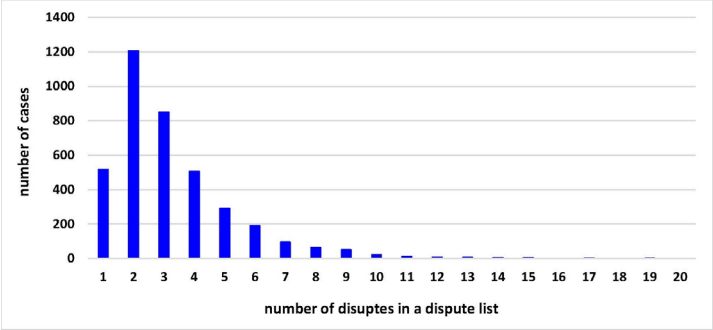

Figure 4: The distribution of the original number of disputes in 3835 cases

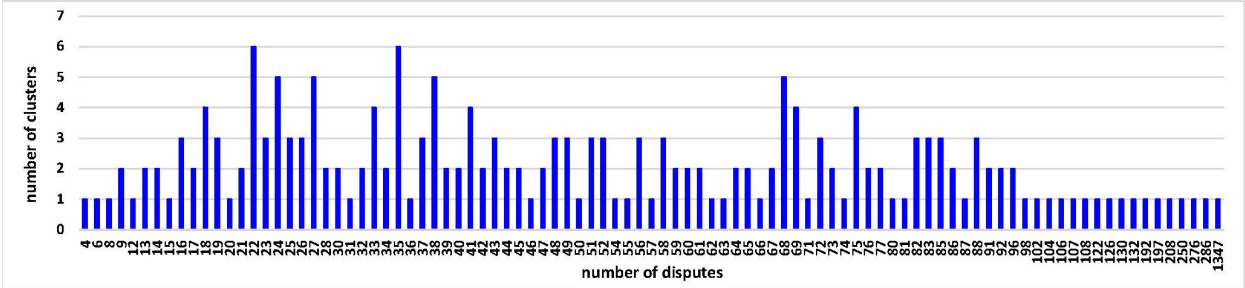

Figure 5: The distribution of the original number of disputes in 200 clusters

selected 3835 cases. In total, we have **12523** disputes in them.

Although we showed only six clusters in Figure 3, a clustering procedure for a realistic analysis task may have to consider hundreds of clusters, hoping that some of the clusters offer useful insights into our data. Figure 5 shows the distribution of the number of clusters that include a certain number of disputes, when we ran a *k*-means clustering procedure to produce **200** clusters for the 12523 disputes.

Given that we will generate hundred of clusters from a large number of disputes, how could we make sure that the disputes that were placed in the same clusters had related and even close legal meanings? We will discuss this problem in Section 6.

There are some more technical questions for this procedure. The current definition favors the cases that have more disputes in the dispute lists. In practice, although the numbers of disputes of cases concentrate within a small range, this factor is influential, so more sophisticated methods may be needed.

In addition, it appears that the current definition disregards the number of disputes that belong to the same cluster. If C_X and C_Y both have one dispute that belongs to a certain cluster k_i, their g function will receive a credit of 1. If C_X and C_Y both have more disputes that belong to a certain cluster k_i, their g function will still receive a credit of 1. Although this may be concerning, how should we define a new score for such cases? Should we consider the Dice coefficient?

On the other hand, if there are many disputes of a case that can be placed into the same cluster, the meanings or functions of these disputes may be similar (and redundant), which is not reasonable for a well-written judgment document. Hence, we expect that having multiple and many disputes in one cluster may not be a frequent problem.

6. Clusters for Social Conflicts

6.1. Steps for Clustering

We vectorized each of the pooled disputes, and clustered the resulting vectors. We may apply the techniques of singular value decomposition (SVD) to reduce the dimension of the resulting vectors in the middle.

For this preliminary exploration, we applied the TFIDFvectorizer function of *scikit learn* to vectorize the disputes.[2] We certainly should try the vectorization with the BERT family, but we would prefer to check whether using the traditional models would give ourselves better ability of explainability first. We chose to use the

[2] The website of scikit learn is https://scikit-learn.org/stable/, last visited May 5, 2023.

Table 1: Dimensions of vectors for disputes before SVD

ngram_range	(1,3)	(1,4)	(1,5)
dimension	149381	322280	555441

'char' model, which allowed us to adopt the concept of FastText in tokenizing the Chinese text [2]. When using the 'char' model, we could manipulate the longest length of the *n*-grams for tokenization at will. The computational costs would increase as we increased *n*. Table 1 offers information about the resulting size of the vector for each of the disputes when we set *n* to 3, 4, and 5. Most Chinese words contain no more than four characters, but some Chinese legal terms are longer.

A common strategy to avoid huge computational costs is to do the SVD step for dimensionality reduction, after we use a large *n* for capturing longer legal terms in the tokenization step. We replied on the TruncatedSVD function of scikit learn to do SVD, and the only important parameter to choose was the number of target dimension.

We can then apply clustering methods to cluster the vectors of the disputes. Top-down clustering, like the *k*-means algorithm, can be a choice. Bottom-up clustering, like the agglomerative clustering, is another common choice [1].

We need to choose the number of clusters for inspection. The problem of selecting the number of clusters is a common issue for people who apply clustering in their work, and we can find some standard method in the literature. We will not pursue that venue of topic in this manuscript. Instead, we shall show that choosing a large number of clusters would give us promising results for now.

6.2. Inspecting the Clusters

We implement a mechanism for assisting readers to inspect the contents of the clusters. This is necessary because that, although the results of clustering (also topic modeling) are potentially useful, not all of the clusters (topics) may make sense in practice. A researcher can read the clusters to find out the useful ones diligently and patiently, as many have done so [13]. This inspection process does require some hard work because a cluster can include many dozens of disputes, as we have shown in Figure 5. It is better to offer assistive information to facilitate this inspection process.

After we had produced the clusters of disputes, we could analyze the disputes in the clusters. We can find, list, and offer the most frequent words in each of the clusters to the readers. The readers may grasp a rough idea about the possible main subject of the disputes in a cluster.

Since we used the TFIDFvectorizer function that was provided by the *scikit learn* to vectorize the disputes, and used the vectors in the clustering step, we could use the words that have highest TFIDF values as the summary for a given cluster.

We provide a partial screenshot for the inspection of a cluster that contains 49 disputes in Appendix C.

Given the summaries and the list of disputes, we found that it was easy to browse the clusters of the disputes, and figured out the main subjects for the clusters. We could change the settings for the clustering steps, and confirmed that clustering the disputes provide a practically viable way to find the topics of the disputes. In fact, before we could offer the summaries, we have read our clusters directly, found the viability of the clustering methods, and discussed some preliminary results in a digital humanities conference [8].

By studying the clusters of the disputes, we found some common types of labor-employer disputes. They are for the benefits after retirement, for unlawful or debatable layoff, for body injuries or fatality during worktime, for the salaries and the late-night meals as a result of overtime work, etc. These findings are consistent with the annual reports of the Ministry of Labor [11]. Hence, analyzing the judgment documents offer a viable window to view the conflicts of our society.

7. Similar Cases Recommendations

7.1. Labeling Similarity

To verify whether we recommend similar cases correctly, we need to prepare for the "correct" answers, i.e., the ground truth, for each pair of judgments. That we selected 3835 cases means we need to determine any two cases in this selection are similar or not. That would be 3835×3835 decisions. We certainly cannot afford this workload. Based on our previous work, we had a collection of 3030 labeled pairs, and we have 1453 "similar" pairs, 599 "barely similar" pairs, and 977 "dissimilar" pairs.

We chose to label the pairs by three labels because we would like to allow the annotator the

freedom to express her negative judgments in a soft way. Allowing the annotator to say "barely similar" is an alternative method to say "dissimilar." Hence, in our analysis, when the label was "dissimilar" for a recommendation, we would just consider that as an incorrect recommendation.

An assistant with a bachelor degree in law conducted the labeling. We have hired a second assistant with expertise in law, but have not completed the analysis of inter-annotator agreement yet. During the labeling period of the first assistant, we gave the same the pair of cases to her, and we found that she might assign the same pair into different categories, but she remained consistent most of the time.

7.2. Parameters for Clustering

We mentioned several parameters that influence the clustering procedure in the previous sections. We use **SP** to denote whether or not we split the original dispute statement into separate statements (cf. the end of Section 4). We denote whether we replace some named entities with more general words by **B** (cf. middle of Section 4). We denote the longest n-gram that we used when we called TFIDFvectorizer by ***n*** (cf. Section 6.1). We denote the number of clusters we set when we clustered the disputes by ***noc***. To simplify the procedure for the experiments, we selected a threshold, $\boldsymbol{\tau}$. If $g(D_X, D_Y) \geq \boldsymbol{\tau}$, then our programs would recommend that D_X and D_Y were similar. If D_X and D_Y are a pair that was previously labeled, we could verify this recommendation. If they are not previously labeled, then we will ignore this pair in the evaluation process. In Section 6.1, we mentioned that one might need to do SVD to reduce computational costs.

The settings of these parameters need some explanations. **SP** and **B** are both either TRUE or FALSE. We set ***n*** to 3, 4, and 5, as we mentioned that most Chinese words contain four characters. When **SP** is FALSE, we had 12523 dispute statements. If we set ***noc*** to 200, 600, 800, and 1000, then these clusters will have an average of 61, 21, 16, and 13 disputes. When SP is TRUE, we may further split the original dispute statements, and we obtained 18770 dispute statements. If we set ***noc*** to 200, 600, 800, and 1000, then these clusters will have an average of 94, 31, 23, and 19 disputes. By setting ***noc*** to four different values, we may observe its influences.

Table 2: A total of 120 combinations of parameters.

B	***n***	***noc***	**SP**	$\boldsymbol{\tau}$	combinations
2	3	4	TRUE	3	72
			FALSE	2	48

Table 3: Better results among the 120 experiments.

B	***n***	***noc***	**SP**	$\boldsymbol{\tau}$	S	BS	NS
T	4	200	F	3	67.1%	18.0%	15.0%
T	3	600	F	2	68.8%	15.8%	15.4%
T	3	800	F	2	69.3%	16.3%	14.4%
T	3	1000	F	2	71.7%	15.7%	12.6%
F	3	200	F	3	71.5%	17.2%	11.3%
F	3	600	F	2	71.5%	17.0%	11.5%
F	5	600	F	2	69.9%	16.8%	13.3%
F	3	800	F	2	69.6%	16.5%	13.9%
F	5	800	F	2	69.4%	17.8%	12.8%
F	3	1000	F	2	70.2%	16.1%	13.7%
F	4	1000	F	2	73.0%	17.5%	9.5%
F	5	1000	F	2	72.5%	16.3%	11.1%
T	3	800	T	3	69.2%	15.1%	15.7%
F	3	600	T	3	68.0%	17.1%	14.9%
F	5	800	T	3	68.2%	19.5%	12.3%

Depending on whether **SP** is FALSE or TRUE, we set $\boldsymbol{\tau}$ to 2, 3, 4, or 5. When **SP** is FALSE, the selected cases have 12523/3835=3.27 disputes in their dispute lists, on average. Therefore, when SP is FALSE, we set $\boldsymbol{\tau}$ to 2 and 3. When SP is TRUE, the selected cases will have 18770/3835=4.89 disputes on average. Hence, when SP is TRUE, we set $\boldsymbol{\tau}$ to 3, 4, or 5.

We did not activate the SVD step because our computers could handle the computations when we set ***n*** to 5.

In conclusion, we conducted experiments for 120 different combinations of these parameters. This is from 2(for **B**)×3(for ***n***)×4(for ***noc***)×2(for $\boldsymbol{\tau}$)= **48**, when **SP** is FALSE, and from 2(for **B**)×3(for ***n***)×4(for ***noc***)×3(for $\boldsymbol{\tau}$)=72, when **SP** is TRUE. We summarize these calculations in Table 2. These recommenders would recommend different number of pairs of similar cases from the 3835 cases. Some recommended hundreds, but others very few. Most important of all, the 3030 labeled pairs cannot cover all of the 3835×3835 pairs. Hence, we report only the experiments in which there were at least 250 recommended and labeled pairs.

Table 3 lists data about the experiments that had relatively better performances. In this table, "S", "BS", and "NS" represent "similar", "barely similar", and "not similar", respectively. "T" and "F" are for "TRUE" and "FALSE", respectively.

Assume that a recommender recommended n pairs and that there were x, y, and z pairs of "S", "BS" and "NS". Normally, we n will be larger than $x + y + z$, as we explained in the previous paragraph. Let $m = x + y + z$, we report $\frac{x}{m}$, $\frac{y}{m}$, and $\frac{z}{m}$ for "S", "BS", and "NS" in Table 3. The average of the "S" column is just about 70.0%. Note that the summation of a row in Table 3 may become 100.1% due to accumulated rounding errors.

7.3. Empirical Observations

A recommendation for a pair of cases was considered correct only if that pair was labeled as "similar" by the annotator.

When we chose a large $\boldsymbol{\tau}$, the number of recommended pairs would reduce because of the high standard. If the number of recommended pairs was less than 250, we would not consider the results reliable in the current work, even though results of these experiments would have a larger proportion of "similar". When the outcomes of the experiments were reliable, increasing $\boldsymbol{\tau}$ would lead to better experimental results, which was expected.

It was possible that the recommended pairs in an experiment were not labeled before. If the number of labeled pairs was low, results of these experiments were ignored because they were not statistically reliable as well.

On average, when we set **SP** to TRUE and further split the dispute statements, the average proportion of correct recommendations reduced than when we set **SP** to FALSE.

We had expected that the blurring step might help us achieve better recommendations. That was based on the intuition of using exact values of the named entities may make it harder for the recommenders to learn general concepts about similarity. However, the experimental results indicated that whether we did the blurring step or not did not affect the average results significantly, all else being equal.

Setting ***noc*** to 200 led to inferior results than setting ***noc*** to 600, 800, and 1000. We did not observe clear differences among setting ***noc*** to 600, 800, 1000. Using 200 clusters might not be sufficient to differentiate the underlying nature of the disputes of the selected cases, so may have confused the recommender. Choosing a larger ***noc*** may help us put more similar disputes in separate clusters, but we need to examine this manually. This may be worthwhile to do in the near future.

We did not observe clear relationships between the values of ***n*** and the performance of the recommenders. The dimensionality of the TFIDF vectors would increase exponentially with ***n***, so we might just set ***n*** to 3 in exploratory tests.

8. Discussions

We have conducted this preliminary research only with relatively more traditional machine approaches. We will extend our reach to deep learning and ChatGPT.[3] BERT [4] and legal BERT [3][6][12] may vectorize the dispute statements more precisely, if these models can be fine tuned for legal documents in Chinese. It should be clear that one should try to ask the ChatGPT to generate the summaries for the clusters of disputes, and verify their usability.

To enlarge and make our database public like CAIL [15], the quality and the quantity of the labeled pairs need additional work. We need more domain experts to help the annotation task. The analysis of inter-annotator agreement is an urgent mission. Since we do not expect that we can, nor do we think that we should even imagine that we should try to label all of the 3835×3835 pairs, we built recommenders to generate recommendations, asked the annotators to label, and used the labeled data to train our models. Effects of such an incremental procedure relies on the powerfulness of our recommenders. If the recommenders are good enough, we may produce better quality of labeled data at relatively low costs.

Acknowledgements

This research was supported in part by the project 110-2221-E-004-008-MY3 of the National Science and Technology Council of Taiwan. The authors are obliged to the reviewers who provided an abundant amount of information to study and improve our work. We will do more homework, though the revised version could not reflect our future progress. Yi-Fan Liu selected the judgment documents from TWJY and handled the preprocessing of the documents. Chao-Lin Liu

[3] ChatGPT: https://openai.com/blog/chatgpt, last visited May 5, 2023.

conducted the preliminary analysis and completed this manuscript.

References

[1] Ethem **Alpaydin**, *Introduction to Machine Learning*, fourth edition, The MIT Press, 2020.

[2] Piotr **Bojanowski**, Edouard **Grave**, Armand **Joulin**, and Tomas **Mikolov**, Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, vol. 5, 135–146, 2017.

[3] Ilias **Chalkidis**, Manos **Fergadiotis**, Prodromos **Malakasiotis**, Nikolaos **Aletras**, and Ion **Androutsopoulos**, LEGAL-BERT: The muppets straight out of law school, *Findings of the Association for Computational Linguistics*: EMNLP 2020, 2898–2904, 2020.

[4] Jacob **Devlin**, Ming-Wei **Chang**, Kenton **Lee**, and Kristina **Toutanova**, BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of NAACL-HLT 2019*, 4171–4186, 2019.

[5] Zhanxing **Hao**, Xiao **Wei**, and Hong **Hu**, A comparative method of legal documents based on LDA, *Proceedings of the 2018 International Conference on Applications and Techniques in Cyber Security and Intelligence*, 271–280, 2018.

[6] Peter **Henderson**, Mark S. **Krass**, Lucia **Zheng**, Neel **Guha**, Christopher D. **Manning**, Dan **Jurafsky**, and Daniel E. **Ho**, Pile of law: learning responsible data filtering from the law and a 256GB open-source legal dataset, *Advances in Neural Information Processing Systems* 2022, 29217–29234, 2022.

[7] Zhilong **Hong**, Qifei **Zhou**, Rong **Zhang**, Weiping **Li**, and Tong **Mo**, Legal feature enhanced semantic matching network for similar case matching, *Proceedings of the 2020 International Joint Conference on Neural Networks*, 1–8, 2020.

[8] Chao-Lin **Liu**, Yi-Fan **Liu**, Wei-Zhi **Liu**, and Hong-Ren **Lin**, Discovering civil disputes hidden in the court judgment documents for applications in social studies and legal informatics, *Proceedings of the 2022 International Conference on Digital Humanities*, 298–300, 2022.

[9] Yinglong **Ma**, Peng **Zhang**, and Jiangang **Ma**, An efficient approach to learning Chinese judgment document similarity based on knowledge summarization, 2018. arXiv:1808.01843

[10] Arpan **Mandal**, Kripabandhu **Ghosh**, Saptarshi **Ghosh**, and Sekhar **Mandal**, Unsupervised approaches for measuring textual similarity between legal court case reports, *Artificial Intelligence and Law*, 29, 417–451, 2021.

[11] MOL, Ministry of Labor of Taiwan, Annual Statistics about the Labor in Taiwan, 2021. < https://www.mol.gov.tw/1607/2458/2464/2468/statisticalReportList>

[12] Shounak **Paul**, Arpan **Mandal**, Pawan **Goyal**, Saptarshi **Ghosh**, Pre-trained language models for the legal domain: a case study on Indian law, 2022. arXiv:2209.0604

[13] Daniel **Ramage**, Evan **Rosen**, Jason **Chuang**, Christopher D. **Manning**, and Daniel A. **McFarland**, Topic modeling for the social sciences, *Proceedings of the 2009 NIPS Workshop on Applications for Topic Models: Text and Beyond* 5: 1–4, 2009.

[14] Shahmin **Sharafat**, Zara **Nasar**, and Syed Waqar **Jaffry**. Legal data mining from civil judgments, *Proceedings of 2018 International Conference on Intelligent Technologies and Applications*, 426–436, 2018.

[15] Chaojun **Xiao**, Haoxi **Zhong**, Zhipeng **Guo**, Cunchao **Tu**, Zhiyuan **Liu**, Maosong **Sun**, Tianyang **Zhang**, Xianpei **Han**, Zhen **Hu**, Heng **Wang**, and Jianfeng **Xu**, CAIL2019-SCM: A dataset of similar case matching in legal domain, 2019. arXiv:1911.08962

Appendix A

We show two dispute lists in Chinese in this appendix.

Example 1

The source is CHDV,92,勞訴,32,20040102,1.json.

1. 系爭夜點費應否列入平均工資核發退休金？
2. 原告己○○、丙○○、乙○○、戊○○、辛○○、庚○○於具領退休金時，所簽立之收據之效力為何？

Example 2

The source is CHDV,98,勞訴,37,20100409,1.json.

1. (一)長森醫院於 97 年 7 月 31 日是否有歇業之事實？
2. (二)兩造間是否因歇業而終止勞動契約？

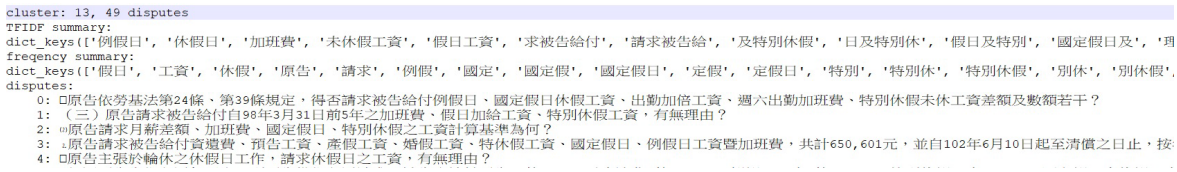

Figure A.1: The first few lines for reading a cluster with 49 disputes, along with its summaries.

After the blurring step (cf. Section 4), the anonymized person names in the second dispute in Example 1 were changed to "someone" in Chinese, which is listed below.

原告**某人、某人、某人、某人、某人、某人**於具領退休金時，所簽立之收據之效力為何？

The place name and the time expression in the first dispute in Example 2 were changed to "somewhere" and "sometime" in Chinese as well.

（一）**某地**於**某時**是否有歇業之事實？

Appendix B

Here is a dispute list that contains a dispute that can be split into two separate disputes. The source is CHDV,98,勞訴,3,20091224,1.json.

1. 被告終止勞動契約有無理由？
2. 原告請求被告給付資遣費是否有理由？被告是否尚積欠原告特休假未休工資及工傷假期間工資？

If we conduct sentence splitting, as we explained at the end of Section 4, we will split the second dispute into two separate disputes.

1. 被告終止勞動契約有無理由？
2. **原告請求被告給付資遣費是否有理由？**
3. **被告是否尚積欠原告特休假未休工資及工傷假期間工資？**

Appendix C

We adopted the concept of FastText [2] for creating Chinese tokens in our vectorization process, so the words were not realistic words, but the strings remained understandable for Chinese readers.

We showed the screenshot for inspecting a cluster of 49 disputes partially in Figure A.1. The "TFIDF summary" and the "frequent summary" were explained in Section 6 In this screenshot, we show only the first five disputes in this cluster.

Given the summaries, we hope that the readers can find that the "休假工資" (compensation salaries for working during holidays) was a subject in this cluster.