Complex Intell. Syst. (2018) 4:1–10
https://doi.org/10.1007/s40747-017-0042-z

ORIGINAL ARTICLE

An innovative multi-segment strategy for the classification of legal judgments using the k-nearest neighbour classifier

S. Pudaruth[1] · **K. M. S. Soyjaudah**[2] · **R. P. Gunputh**[3]

Received: 13 November 2016 / Accepted: 11 April 2017 / Published online: 24 April 2017

Abstract The classification of legal documents has been receiving considerate attention over the last few years. This is mainly because of the over-increasing amount of legal information that is being produced on a daily basis in the courts of law. In the Republic of Mauritius alone, a total of 141,164 cases were lodged in the different courts in the year 2015. The Judiciary of Mauritius is becoming more efficient due to a number of measures which were implemented and the number of cases disposed of in each year has also risen significantly; however, this is still not enough to catch up with the increase in the number of new cases that are lodged. In this paper, we used the k-nearest neighbour machine learning classifier in a novel way. Unlike news article, judgments are complex documents which usually span several pages and contains a variety of information about a case. Our approach consists of splitting the documents into equal-sized segments. Each segment is then classified independently of the others. The selection of the predicted category is then done through a plurality voting procedure. Using this novel approach, we have been able to classify law cases with an accuracy of over 83.5%, which is 10.5% higher than when using the whole documents dataset. To the best of our knowledge, this type of process has never been used earlier to categorise legal judgments or other types of documents. In this work, we also propose a new measure called confusability to measure the degree of scatteredness in a confusion matrix.

Keywords Supreme court · Judgments · Classification · Nearest neighbour

✉ S. Pudaruth
s.pudaruth@uom.ac.mu

[1] Department of Ocean Engineering and ICT, Faculty of Ocean Studies, University of Mauritius, Moka, Mauritius

[2] Department of Electrical and Electronics Engineering, Faculty of Engineering, University of Mauritius, Moka, Mauritius

[3] Department of Law, Faculty of Law and Management, University of Mauritius, Moka, Mauritius

Introduction

According to the latest census that was carried out in the year 2011 [1], the Republic of Mauritius has a resident population of about 1,233,000. Despite this small population size, an astonishing number of cases are lodged and processed at the different courts of the Republic of Mauritius every year. In 2015 alone, 141,164 new cases were lodged and 137,920 were disposed of. A record number of 50,270 cases remained outstanding at the end of 2015 [2]. The introduction of several measures by the Judiciary of Mauritius to increase the efficiency of the courts have been quite successful but has not been able to match up to the rapid increase in the number of cases that are being lodged [3].

The k-nearest neighbour (kNN) algorithm was introduced by Fix and Hodges [4] and elaborated by Cover and Hart [5]. It basically works by finding the instance (in this case the instances are documents) from the dataset which is closest ($k = 1$) to the one that is being considered. It is also possible to find the three closest ($k = 3$), five closest ($k = 5$), seven closest ($k = 7$) or k-closest instances, where k is usually an odd integer. When $k = 1$, the category of the nearest neighbour is taken to be the category of the document for which the prediction is being done. When k is larger than 1, a majority voting system is used to select the best category [6]. Weights can also be assigned to the nearest neighbours. There are different ways in which the distance (degree of similarity) can be computed and this often depends on the nature of the

data. The Euclidean distance measure is the most popular but others like Chi-square distance, Minskowsky and cosine similarity measure also exist [7].

The kNN algorithm is often said to be a lazy machine learning classifier as there is no training per se [7], i.e. there is no learning or no model is actually built as it is an example-based classifier. Instead, the new instance is simply compared with all the existing instances to retrieve the closest matches [8]. In practice, an existing dataset is often separated into a training set and a testing set to evaluate its performance through cross-validation. The kNN classifier is a relatively simple algorithm compared to more complex approaches like artificial neural networks or support vector machines [9]. This simplicity, robustness, flexibility, and reasonably high accuracies have been exploited in diverse fields such as patent research [10], medical research [11], astrophysics [12], bioinformatics [13], and text categorisation [14,15]. The drawback of kNN lies in the expensive testing of each instance as every new instance must be compared with the whole dataset. It is also sensitive to peripheric attributes.

In this paper, we have applied the kNN classifier in a novel way. It is usually accepted that a legal judgment is a special kind of document, unlike news articles or medical reports. The size of a legal judgment can vary from one page to dozens of pages. Judgment usually contains abstract concepts written in complicated sentence structures [16,17]. An added difficulty in our study is multilingualism. Mauritian judgments, especially for civil cases, usually contain a significant proportion of French words. This is so because many Mauritius civil law documents are written entirely in French. Two major examples are the Code Civil Mauricien and the Code de Commerce. In previous studies, researchers have always considered a document as a single closed entity. Because legal judgments are rich in information, we split them into different pages and analyse them separately. A kNN classifier is run on each page. A voting process is then carried out to choose the predicted category.

The rest of the paper is organised as follows. In Sect. 2, we illustrate the problem in more detail. In Sect. 3, we describe several related works that have been done in the legal domain using the nearest neighbour approach. The methodology used and the data set for this study is described in Sect. 4. The different experiments that were performed and their results are discussed in Sect. 5. Finally, in Sect. 6, we conclude the paper with a brief summary of the work that has been implemented. Some ideas for future works are also presented.

Problem statement

The classification of legal documents is often considered as a difficult problem. In our previous work on the categorisation of Supreme Court cases [18], we identified several reasons for this. These are the high variability in document length, presence of abstract concepts and multilingualism (for Mauritian cases). Also, cases often do not fit neatly in one category only. For example, under Mauritian laws, it is possible for a person to bring a civil law case (under the law of tort) for the criminal offence of assault, which is defined under Section 230 of the Criminal Code [19]. Another major problem with judgments which has not been mentioned before is that judgments do not contain only legal terms describing a single issue. In reality, the legal issues are usually embedded into mountains of factual information which are usually irrelevant to the legal issues which are being considered. This is the main cause of mis-classification for legal documents. All machine learning classifiers directly or indirectly rely on words and words' frequencies to classify a document. Thus, the presence of these extra information usually has a negative impact on performance measures such as accuracy, precision, and recall. In this work, we shall study the impact of using different parts of the document on classification rate.

Literature review

The kNN classifier is a simple machine learning technique which looks for the instance which is most similar to the current one. To calculate similarity between two documents, a number of functions can been used. Mutual information and cosine similarity are two commonly used metrics [20]. Furthermore, in the basic algorithm, all the words are assumed to have the same weight, i.e. a weight of one. Over the years, many adjustments have been made to the basic classifier to improve its accuracy.

Han et al. [21] used a weight-adjusted cosine similarity measure to classify several real data sets and showed that this is more effective than traditional approaches. Their weight-adjusted kNN outperformed traditional kNN approaches, decision tree techniques and rule-based systems. The only drawback is that their weight-finding algorithm has a complexity of $O(x^2)$, where x is the number of instances. Kwon and Lee [22] used the k-nearest neighbour approach to classify web pages. They also used a weighted mechanism for the different features.

Baoli et al. [23] conducted extensive experiments on choosing the right value for k. They understood that if the number of documents in each category in not equal (which is usually the case for real datasets), choosing a fixed value for k will favour the larger classes. Thus, they designed a new decision function to choose an appropriate value for k. The decision function is based on the number of instances for each category in the dataset. They showed that their approach was better and more stable than the traditional kNN for large values of k. More recently, Bhattacharya et al. [24] conducted a similar study but using a probabilistic framework to estimate a value for k dynamically.

Jo proposed to use string vectors rather than numerical vectors to encode documents [25]. A string vector is simply a list of words arranged in descending order of their frequencies as found in a document. The kNN algorithm was then used to calculate the semantic similarity between two string vectors. The proposed algorithm was tested on the Reuters 21578 dataset. It was found that the new way of encoding documents has a small positive impact on the classification rates. An interesting study was performed by Kumar et al. [26] on judgments from the Supreme Court of India. They used kNN with cosine similarity and showed that it is better to use legal terms only rather than all terms for classification.

A novel classification technique for legal documents based on data compression was designed and implemented by Mastropaolo et al. [27]. Their idea is that if two documents are very similar and they are combined and compressed, the overall compressed size should be only slightly smaller than one of the individually compressed files. In the same way, if the two files are very different, their combined compressed size will be significantly greater than one of the individually compressed files. They proposed two algorithms to compute similarity using compression. Both of them used the concept of nearest neighbour to determine the similarity. The authors used a dataset of 70 documents with 7 categories. Their novel algorithm outperformed classical approaches like J48, Naïve Bayes and SMO.

Eunomos is a legal information system which has been developed as a result of the ICT4LAW project [28]. It is intended to become a full-fledge legal document management system for legal firms, legal practitioners and legal scholars [29,30]. Boella et al. used the SVM (Support Vector Machines) classifier to classify 233 legal documents into six different areas of tax law [31]. They achieved an accuracy of 76.23%. In 2005, Biagoli et al. [32] classified 582 paragraphs from Italian laws into 10 categories using SVM, with an accuracy of 92%. However, the classification was not in terms of traditional areas of law but instead they used classes such as obligations, substitutions and prohibitions.

In his 2013 PhD thesis [33], Medina re-assessed completely the kNN classifier from all angles. In particular, he was interested in measuring the uncertainly of predictions made through kNN. Although his work was based on prediction of the chemical and physical properties of materials, the principles are broad enough to be applicable to many other areas. In the next section, we describe how our dataset has been created and how it will be used in the experiments.

Methodology

All the Supreme Court judgments for the year 2013 were downloaded from the website of the Supreme Court of the Republic of Mauritius. However, only 401 judgments were considered for this study as judgments falling into categories having less than six instances were discarded. The corpus contains a total of 855,150 words, after stripping off the header and footer information in each case. These 401 documents had been manually classified into 18 distinct categories by an LLB graduate from the University of Mauritius. Each document was then segmented into small pieces of 50, 100, 200, 300, 400, 500 and 600 words. If the last segment was less than half the segment size, it was added to the previous segment otherwise it was considered as a segment on its own. The predictions into the appropriate categories were made using the kNN classifier using Weka 3, which is an open source data mining software developed in the Java programming language [33]. The name of each category, the number of documents and the number of 100-word segments in each category are provided below in Table 1.

Table 1 Judgments dataset split into fixed-size segments

Categories	Number of judgments	Number of 100-word segments
Homicide (murder and manslaughter)	14	266
Involuntary homicide by imprudence	8	116
Road traffic offences	23	294
Drugs	44	1481
Other criminal offences	47	621
Company law	22	522
Labour law	17	490
Land law	32	819
Habere facias possessionem (tenancy)	23	431
Contract law	47	1293
Attachment proceedings	6	127
Road accident and insurance	9	199
Customs act	6	148
Election matters	14	310
Family issues	10	196
Injunctions	50	934
Judicial review	16	330
Application for Leave to the Judicial Committee of the Privy Council (JCPC)	13	171

Source: Authors

Table 1 shows the distribution of cases for each area of law. We have a total of 136 criminal cases (first five categories) and 265 civil cases (remaining categories). In our previous work [17], we used only eight (8) categories but this time we have increased it to eighteen (18). For example, the homicide category has been split into two different categories: murder and manslaughter and involuntary homicide by imprudence.

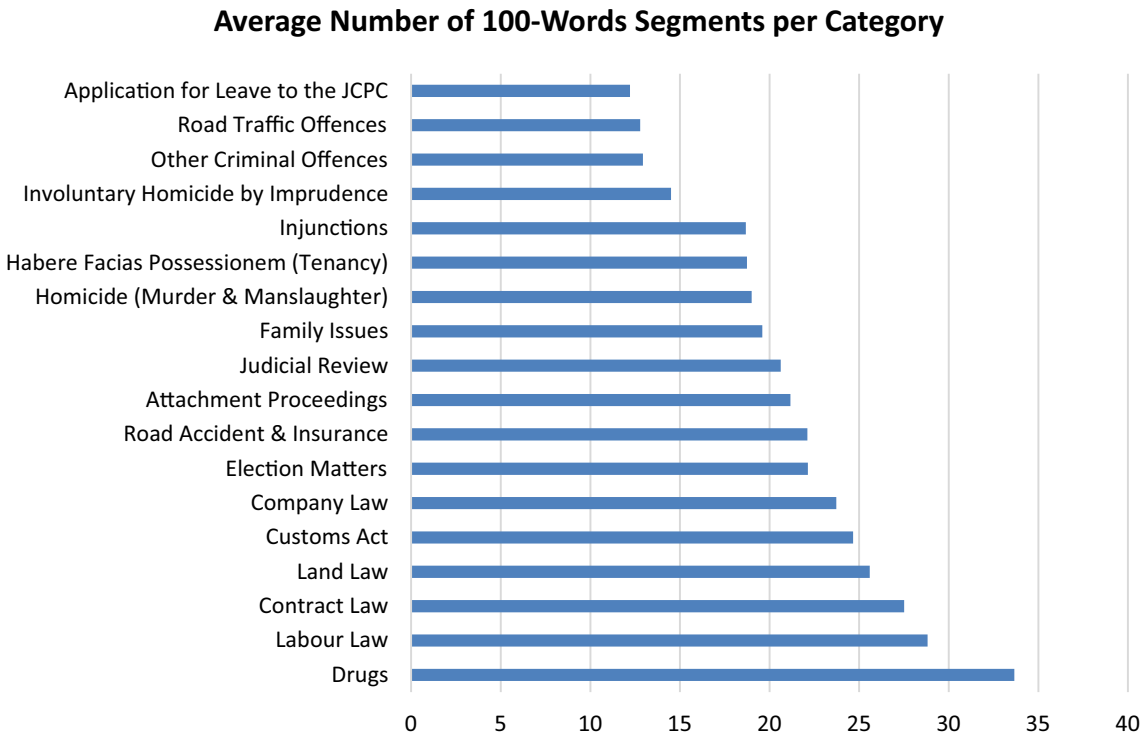

Fig. 1 Average number of 100-word segments per category

Land law has also been split into general land issues and tenancy issues. New categories like road accident and insurance, election matters, family issues, injunctions and judicial review have also been added. The number of cases has also rose from 294 to 401. The biggest drawback of the horizontal thesaurus approach is the need to create a lexicon (dictionary of terms) for each area of law that is added to the system. This is a time-consuming procedure and we would expect that for large number of classes, the accuracy would decrease significantly. The purpose of classifying legal materials is to address the needs of legal practitioners who need to find relevant materials quickly to prepare their cases. The information has to be provided at a sufficient granular level to be useful to them.

Figure 1 shows the average number of 100-word segment for each of the eighteen categories in which the judgments have been classified. A total of 8748 document segments were extracted from the 401 judgments. The Drugs category has the highest average number of segments per case, followed by Labour Law, Contract Law, Land Law, etc. The average number of segments for criminal law cases is 20 while that of civil and administrative law cases is slightly higher at 22. The smallest document in our dataset contains 262 words while the largest document contains 19,789 words.

Besides fixed-length segments, we also investigated the effect of variable-length segments on the accuracy of document classification. In our case, the variable-length segments are the paragraphs in the document. In our work, a paragraph is a body of text which is separated on both sides (top and bottom) of the text by at least one blank line. Three scenarios have been considered. In the first one, all the paragraphs are processed. Due to the inherent noisiness in textual data, a paragraph can be as short as one word and as long as whole document. Second, only paragraphs which contained more than 50 words were used for classification. And our third dataset consisted of paragraphs with more than 50 words but less than 200 words. We also considered the possibility of using only paragraphs which are more than 100 words; however, upon an in-depth analysis, we found that many documents did not have at least one paragraph of this size.

Table 2 shows the number of paragraphs in each of the three scenarios. There is a huge drop in the number of paragraphs from scenario 1 to scenario 2. On average, the drop is more than 300% and this value is quite consistent across all the categories. However, there is only a 7% drop in the number of paragraphs from scenario 2 to scenario 3, although the drop would have been more significant if it was measured in terms of the number of words, as in scenario 3, we are discarding all paragraphs having 200 words or more.

Experiments, results and evaluation

In this section, we describe the experiments that were performed on the dataset and then evaluate the results. The results are presented in the form of tables and figures. These

Table 2 Judgments dataset split into paragraphs

Categories	No. of paragraphs (any size)	No. of paragraphs (>50 words)	No. of paragraphs (>50 and <200 words)
Homicide (murder and manslaughter)	512	190	174
Involuntary homicide by imprudence	297	87	82
Road traffic offences	944	225	220
Drugs	2722	1049	954
Other criminal offences	1821	454	433
Company law	1127	368	344
Labour law	1017	361	337
Land law	1747	618	587
Habere facias possessionem (tenancy)	1017	330	307
Contract law	2653	879	795
Attachment proceedings	330	87	83
Road accident and insurance	459	149	142
Customs act	368	110	104
Election matters	854	217	210
Family issues	472	146	135
Injunctions	2533	703	669
Judicial review	817	235	222
Judicial Committee of the Privy Council (JCPC)	449	125	117
Total	20139	6333	5915

Source: Authors

are then compared with our own previous results and with existing works in the literature.

An overall accuracy of 53.4% was obtained using the default parameters in the kNN classifier available in Weka when applied on the 401 cases (whole documents). Accuracy is the proportion of instances that are correctly classified. A tenfold stratified cross-validation approach was used in all experiments. After an extensive set of experiments were performed using various parameters, we found that by converting the dataset to lowercase, using unigrams, bigrams and trigrams together, limiting the number of words per class to 200 and using a value of 3 for k with distance weighing (1/d) resulted in an overall accuracy of 73.0%, which is about 20% higher than when using the default values. Term frequency (tf), inverse document frequency (idf) and stemming did not have significant impact on the accuracy.

In the next series of experiments, the inputs to the classifier were equal-sized segments of 50-, 100-, 200-, 300-, 400-, 500- and 600-words. Each segment was independently classified into an appropriate area of law using the kNN classifier. For example, the first case in our dataset belongs to the Attachment Proceedings category and it contains 2160 words. Thus, this case is divided into 22 equal-sized segments of 100 words each after the splitting operation. Since the last segment (22nd) contains more than half the number of words in a normal segment, it is considered as a full segment even if the size is less than 100 words. After the kNN operation is over, 10 out of the 22 segments are classified into the Attachment Proceedings category, 7 of them are classified into the Contract category, 4 into the Injunction category and 1 into the Labour category. The category with the maximum number of votes is taken as the predicted category. Using this innovative procedure, we have been able to achieve an overall accuracy of 83.5%, which is 10.5% higher than when using the whole document for classification.

The recall and precision for each category are shown in Table 3. Recall is the fraction of relevant documents that are correctly identified compared to the total number of documents in that category. The average recall has increased from 0.70 to 0.78 when the documents are analysed in parts rather than in whole. Precision is the fraction of correctly identified instances out of all documents that have been predicted as belonging to that category. There has also been an increase in the average precision value which has jumped from 0.77 to 0.98. Except for Contract Law, all the precision values are above 90% (Fig. 2).

Although there has been a general improvement in the average recall value, the individual recall values for 6 out of the 18 categories are less for segmented documents than for whole documents. The sharpest drop is from the Family Law

Table 3 Performance measures using kNN

	Categories	Whole document		Segmented documents	
		Recall	Precision	Recall	Precision
1	Homicide (murder and manslaughter)	0.79	0.92	0.57	**1.00**
2	Involuntary homicide by imprudence	0.38	0.75	**0.50**	**1.00**
3	Road traffic offences	0.74	0.94	**0.96**	**1.00**
4	Drugs	0.82	0.88	**0.89**	**0.91**
5	Other criminal offences	0.96	0.65	0.74	**0.97**
6	Company law	0.55	0.43	**1.00**	**1.00**
7	Labour law	0.47	0.67	**0.88**	**1.00**
8	Land law	0.44	0.78	**0.97**	**1.00**
9	Habere facias possessionem (tenancy)	0.78	0.82	0.74	**1.00**
10	Contract law	0.64	0.54	**1.00**	0.44
11	Attachment proceedings	0.67	0.80	0.50	**1.00**
12	Road accident and insurance	0.67	0.86	**0.89**	**1.00**
13	Customs act	0.50	0.60	**1.00**	**1.00**
14	Election matters	0.93	1.00	**1.00**	**1.00**
15	Family issues	0.60	0.75	0.10	**1.00**
16	Injunctions	0.82	0.77	**0.84**	**1.00**
17	Judicial review	0.94	0.88	0.56	**1.00**
18	Application for leave to the JCPC	0.85	0.85	**0.92**	**1.00**
Average		**0.70**	**0.77**	**0.78**	**0.96**

Source: Authors
The bold values indicate the categories in which the results are better for segmented documents compared with whole documents

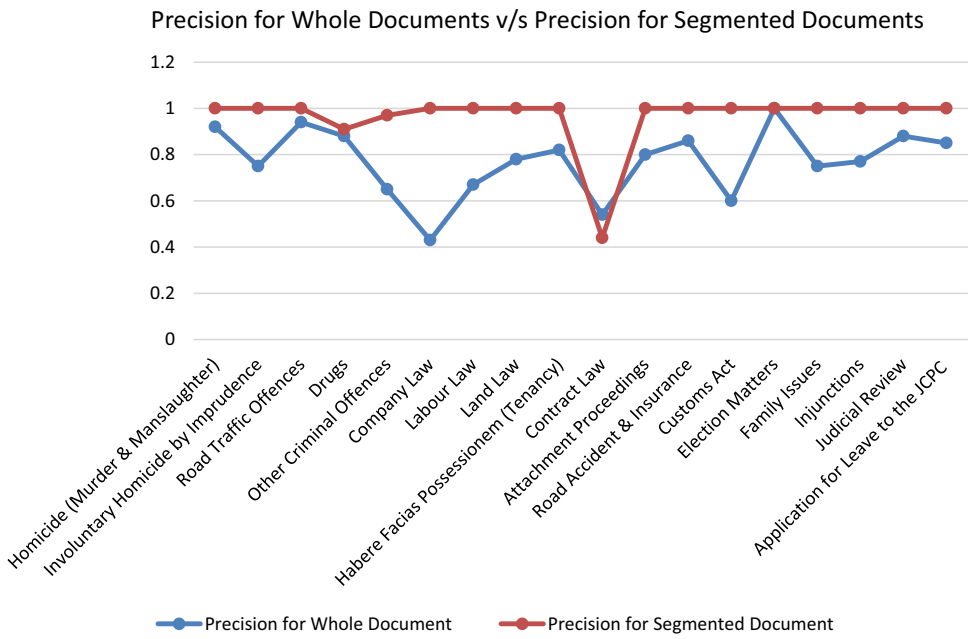

Fig. 2 Precision for whole documents v/s precision for segmented documents

cases, which has fallen from 0.6 to 0.1 (Fig. 3). The drop in Judicial Review is also very significant. On closer examination, we found that most of the misclassified cases are going to the Contract Law category and some of them to the Drugs category (Fig. 3). This is so because these are the two largest categories and together they make up about 32% of the dataset both in terms of the number of segments and the number of words. Thus, it is more likely that segments from other categories find their best matches in these two categories when a good match is not available in the correct category. Four categories (including Contract) now have a recall value of one (1) compared to none in the previous approach (Fig. 3).

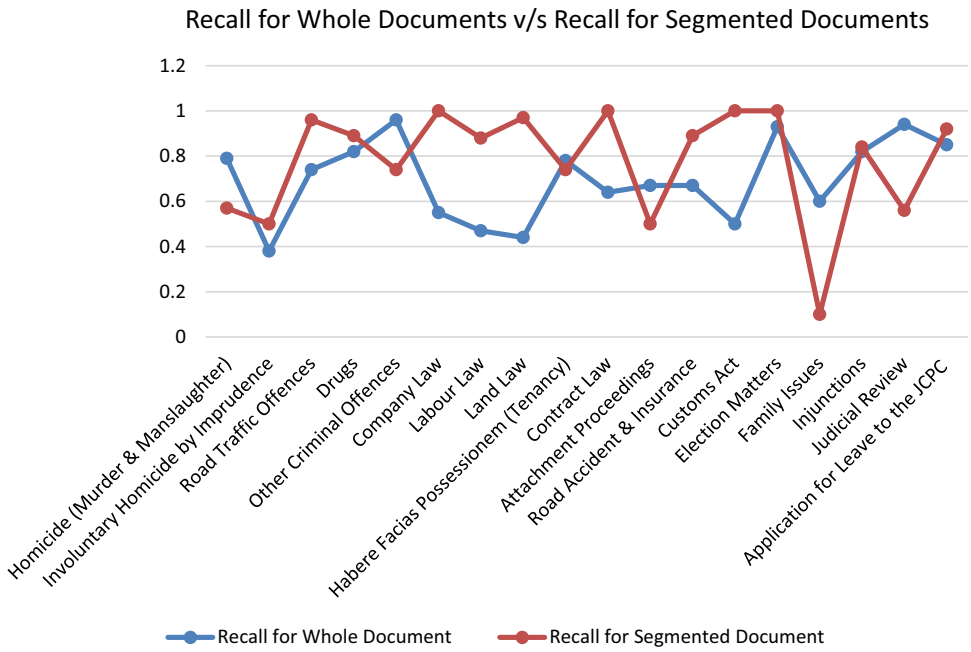

Fig. 3 Recall for whole documents v/s recall for segmented documents

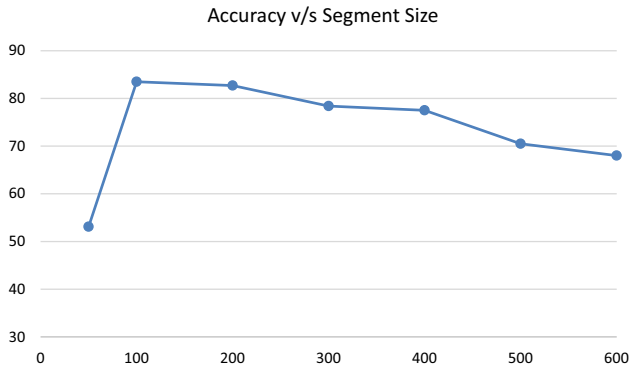

Fig. 4 Accuracy v/s segment size

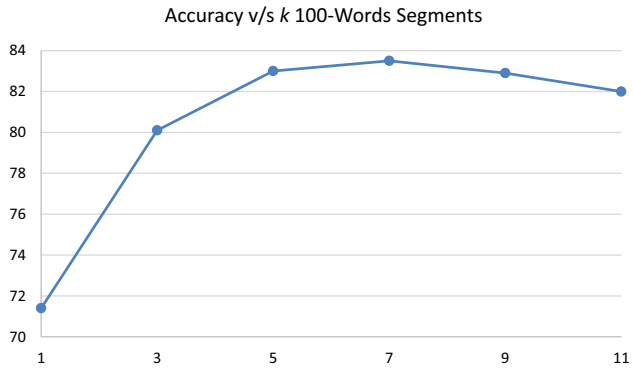

Fig. 5 Accuracy v/s k for 100-word segments

Splitting the documents into segments and then classifying them has a near-magical effect on precision. Except for the Contract Law category, there has been a significant increase in the precision values for all the other categories. More specifically, fifteen out of the eighteen categories now have a precision value of one (1) compared to only one category in the whole document approach. As explained earlier, the Contract category alone represents about 15% of the whole dataset and many segments from other categories are being drawn to it.

We also investigated the effect of segment size on the overall accuracy of the system. The highest accuracy of 83.5% is obtained when using segment size of 100 words. The accuracy is worst when using segments of 50 words. Furthermore, there is a steady decrease in accuracy when the segment size increases from 100 to 600 (Fig. 4).

Figure 5 shows the accuracy values for different values of k (kNN) for segments of 100 words. In all experiments with kNN, the Euclidean distance measure was used. Furthermore,

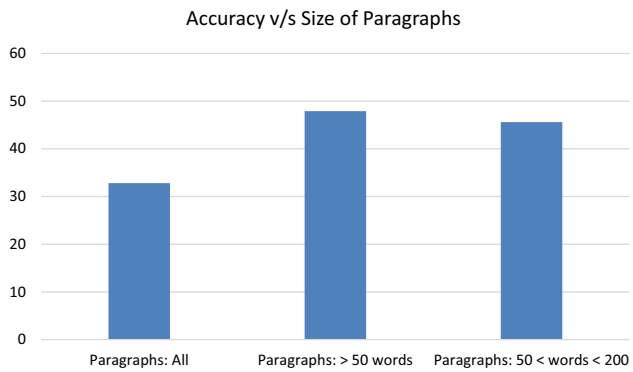

Fig. 6 Accuracy v/s size of paragraphs

a distance weighting method of 1/distance was used for $k \geq 3$. The best result of 83.5% was obtained when $k = 7$. The default value of $k = 1$ produced the worst result.

Figure 6 shows the accuracy values for paragraphs of different sizes. As was done earlier, all the data were converted

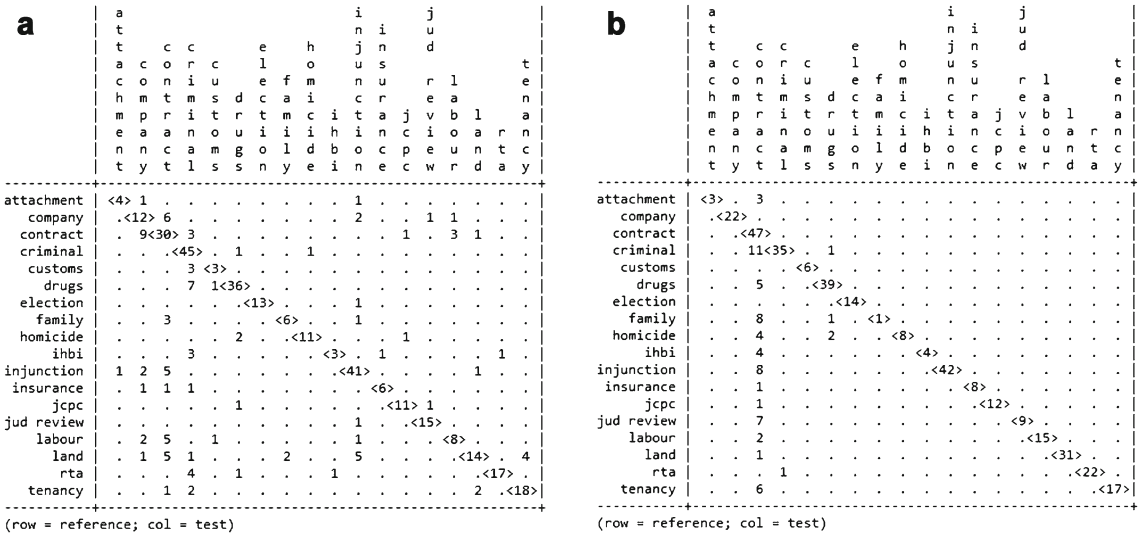

Fig. 7 **a** Confusion matrix for whole documents. **b** Confusion matrix for segmented documents

to lowercase. Bigrams and trigrams were used again and the number of terms was restricted to 200 for each class. The best result of 47.9% was obtained when using paragraphs with more than 50 words. These results were obtained with $k = 3$ and a weight of 1/distance. The worst result of 32.8% was obtained when using all paragraphs while using paragraphs which are between 50 and 200 words led to an accuracy of 45.6%. Compared with fixed-size segments or whole documents, we see that the results are much poorer for variable-sized paragraphs and, therefore, this is not a good approach for this dataset. This poor result can be explained by the way that the *k*NN algorithms works. It operates by comparing each paragraph with every other paragraph in the dataset looking for the most similar one and each term in a paragraph is given the same weight. When using variable-length segments, short paragraphs from one category will on average have more content in common with longer paragraphs, even if these are from other categories. In other words, most of the terms in a short paragraph are likely to be found in a large paragraph than in a small paragraph and this is what creates the poor outcome.

As another contribution to literature emanating from this work, we are proposing a new measure call confusability, which measures the degree of confusion or scatteredness in a confusion matrix (error matrix or contingency table). A confusion matrix is a table which allows the visual assessment of a classifier's performance.

Confusability is defined as follows:

$$\text{Confusability} = \frac{\text{Number of Non-Zero Values in Confusion Matrix-Number of Non-Zero Diagonal Values}}{\text{Size of Confusion Matrix}}$$

Confusability for whole documents (Fig. 7a) = $\frac{(68-18)}{(18*18)} = \frac{50}{324} = 0.15$

Confusability for segmented documents (Fig. 7b) = $\frac{(35-18)}{(18*18)} = \frac{17}{324} = 0.05$

Thus, according to this measure, the values in the first matrix are about three times more dispersed than those in the second matrix. This means that it should be easier to improve the results for segmented documents compared with whole documents, even if the accuracy is already higher. A confusability score of zero (0) means that the prediction is perfect and all values have been correctly categorised. Confusability cannot have a value of one (1) because even if the matrix is full (has non-zero values in all cells), this would mean that all classes have at least one instance which has been correctly classified and non-zero diagonal values have to be subtracted from the total number of non-zero values in the matrix.

A lot of work has been done on the application of the k-nearest neighbour classifier in the field of document classification but to a much lesser extent on legal judgments. However, the accuracy of our system and that of previous systems are not directly comparable because of several reasons. First, we have used our own dataset which is unique and has some specificities. For example, about 4% of words in the dataset are French words and for some categories, this can be as high as 30% [15]. Many of the published works do not use English datasets. Second, most existing works deal

with news articles or medical reports which are generally much shorter than Supreme Courts judgments. Third, there is huge variation in the number of categories used in such studies. Some authors have used only two categories while others have used at least 40. There are also huge differences in the size of datasets (number of documents) used. Under these circumstances, it is very difficult to attempt to do a fair comparison. As regards our own work, we achieved an overall accuracy of 83.5% using 401 documents (of varied lengths) with 18 categories by applying the kNN in a novel way. This percentage is still higher than in comparable works [10,34,35].

Although we have proposed a new method for the classification of cases and the accuracy is satisfactory, there are still many avenues for further research. Although this aspect has not been considered in this work, cases are also often multi-topical which means that a case about one area of law may contain issues arising from other areas of law as well. In fact, this is one reason why document classification is challenging in this field. We intend to perform multi-label classification in some future work. Legal judgment often contains a lot of materials which is not centred around an area of law but rather contains general descriptions of the event and the people involved in the case. This adds to the complexity of classification. Furthermore, for injunctions, it could perhaps be more useful to know the cause and the type of injunction to which a judgment refers. The same applies for judicial review and for most of the other categories as well. Thus, in the future, we will attempt to perform multi-level (hierarchical) classification as we also strongly believe that if legal practitioners are able to classify cases into pre-defined categories with a finer level of detail, this will speed up their work to a considerable extent as much of their time is usually used up in legal research.

Conclusions

The objective of this paper was to classify judgments from the Supreme Court of Mauritius into eighteen pre-defined categories using a nearest neighbour classifier. The judgments were split into equal-sized segments and each segment was classified independently of the others. The category with the highest number of segments was chosen as the predicted category. To our knowledge, this is the first time that this type of procedure has been adopted in classification work. Using this novel approach, we obtained an overall accuracy of 83.5%, which is 10.5% higher than what the traditional approach of processing the document as a single entity could offer. This is also higher than in our previous work where we used a horizontal thesaurus for only eight categories but achieved an accuracy of only 79%. Thus, by applying the well-known k-nearest neighbour classifier in a novel way, we were able to improve the success rate by a significant amount even with a much higher number of categories. In the future, we intend to use several machine learning classifiers and then compare their performances.

References

1. Statistics Mauritius (2016) Statistics Mauritius, ministry of finance and economic development. Available from http://statsmauritius.govmu.org/English/CensusandSurveys/Documents/ESI/toc1.htm. Accessed 04 Aug 2016
2. Supreme Court (2016). Annual Report of the Judiciary 2015. Republic of Mauritius
3. Ejudiciary Mauritius (2016) Available from: https://www.ejudiciary.mu/. Accessed 15 Aug 2016
4. Fix E, Hodges JL (1951) Discriminatory analysis—non-parametric discrimination: consistency properties. University of California, Berkeley, 21-49-004
5. Cover TM, Hart PE (1967) Nearest neighbour pattern classification. IEEE Trans Inf Theory 13(1):21–27
6. Mastropaolo A, Pallante F, Radiciono D (2013) Legal documents categorization by compression. In: Proceedings of the 2013 international conference on artificial intelligence and law, 10-14 June, Rome, Italy
7. Hu L, Huang M, Ke S, Tsai C (2016) The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus 5:1304
8. Kumar S, Reddy PK, Reddy VB, Singh A (2011) Similarity analysis of legal judgments. In: Proceedings of the 4thAnnual ACM Bangalore Conference (COMPUTE'11), No. 17, 25–26 March, Bangalore, Karnataka, India
9. Lorema AC, Jacintho LFO, Siqueira MF, de Giovanni R, Lohmann LG, de Carvalho Acplfandyamamoto M (2011) Comparing machine learning classifiers in potential distribution modelling. Expert Syst Appl 38:5268–5275
10. Kwon OW, Lee JH (2003) Text categorization based on k-nearest neighbor approach for web site classification. Inf Process Manag (Elsevier) 39:25–44
11. Han EHS, Karypis G, Kumar V (1999) Text categorization using weight adjusted k-nearest neighbor classification. Technical Report, Department of Computer Science and Engineering, University of Minnesota, USA
12. Medina JLV (2013) Reliability of classification and prediction in k-nearest neighbours. Thesis (PhD). University of Rovira, Virgili, Spain
13. Jo T (2010). Representation of texts into string vectors for text categorization. J Comput Sci Eng 4(2):110–127
14. Boella G, Caro LD, Humphreys L (2014) Requirements of legal knowledge management systems to aid normative reasoning in specialist domains. Lect Notes Comput Sci (Springer) 8417:167–182
15. Boella G, Caro LD, Humphreys L (2011) Using classification to support legal knowledge engineers in the Eunomos legal document management system. In: Proceedings of the 5thinternational workshop on Juris-informatics (JURISIN 2011), Springer
16. Baoli L, Shiwen Y, Qin L (2003) An improved k-nearest neighbor algorithm for text categorization. In: Proceedings of the 20thinter-

national conference on computer processing of oriental languages, Shenyang, China, 2003
17. Pudaruth KMS, Soyjaudah S, Gunputh RP (2016) Categorisation of supreme court cases using multiple horizontal Thesauri. University of Mauritius. Chapter in intelligent systems technologies and applications (Springer), pp 355–368
18. Boella G, Humphreys L, Martin M, Rossi P, van der Torre L (2012) Knowledge Management System to Build Legal Services. Lect Notes Comput Sci (Springer) 7639:131–146
19. ICT4LAW (2016) ICT4Law: ICT converging on law. Available from: http://ict4law.org/. Accessed 04 Aug 2016
20. Raju B, Vardhan V, Sowmya V (2014) Variant nearest neighbor classification algorithm for text document. Adv Intell Syst Comput 249:243–251
21. Hajlaoui K, Cuxac P, Lamirel JC, Francois C (2012) Enhancing patent expertise through automatic matching with scientific papers. Lect Notes Comput Sci (Springer) 7569:299–312
22. Streiter O, Voltmer L (2003) Document classification for corpus-based legal terminology. In: Proceedings of the 8thinternational conference of the international academy of linguistic law, May 2003, Iasi, Romania
23. Guo Q (2008) The similarity computing of documents based on VSM. Lect Notes Comput Sci (Springer) 5186:142–148
24. Zhuang Y (2012) An improved TFIDF algorithm in electronic information feature extraction based on document position. Lect Notes Electr Eng (Springer) 177:449–454
25. Liu M, Yang J (2012) An improvement of TFIDF weighting in text categorization. In: Proceedings of the international conference on computer technology and science (ICCTS 2012), Singapore, vol 47, pp 44–47
26. Nirmala devi M, Appavu S, Swathi UV (2013) An amalgam KNN to predict diabetes mellitus. In: Proceedings of the IEEE international conference on emerging trends in computing, communication and nanotechnology (ICECCN), 25-26 March, Tirunelveli, India, pp 691–695
27. Yigit H (2013) A weighting approach for kNN classifier. In: Proceedings of the IEEE international conference on electronics, computer and computation (ICECCO), 7–9 November, Ankara, Turkey, pp 228–231
28. Bhattacharya G, Ghosh K, Chowdhury AS (2015) A probabilistic framework for dynamic k estimation in kNN classifiers with certainty factor. In: Proceedings of the 8thIEEE international conference on advances in pattern recognition (ICAPR), 4-7 January, Kolkota, India, pp 1–5
29. Li L, Zhang Y, Zhao Y (2008) K-nearest neighbors for automated classification of celestial objects. Sci China Ser G: Phys Mech Astron 51(7):916–922
30. Xu H, Lu S, Zhou S (2012) A novel algorithm for text classification based on knn and chaotic binary particle swarm optimisation. Lect Notes Electr Eng (Springer) 211:619–627
31. Bichindaritz I (2011) Methods in case-based classification in bioinformatics: lessons learned. Lect Notes Comput Sci (Springer) 6870:300–313
32. Biagioli C, Francescono E, Passerini A, Montemagni S, Soria C (2005) Automatic semantics extraction in law documents. In: Proceedings of the 11thinternational conference on artificial intelligence and law (ICAIL). ACM, pp 43–48
33. Hall M, Eibe F, Holmes G, Pfahringer B, Reutemann P, Witten I (2009) The WEKA data mining software: an update. SIGKDD Explor 11(1)
34. Trstenjak B, Mikac S, Donko D (2013) KNN with TF-IDF based framework for text categorization. Procedia Eng (Sci Direct) 69:1356–1364
35. Basu T, Murthy A (2014) Towards enriching the quality of k-nearest neighbor rule for document classification. Int J Mach Learn Cybernet (Springer) 5(6):897–905