Protecting Privacy in the Archives: Supervised Machine Learning and Born-Digital Records

Tim Hutchinson
University Archives & Special Collections
University of Saskatchewan Library
Saskatoon, Canada
tim.hutchinson@usask.ca

***Abstract*— This paper documents the iterations attempted in developing training sets for supervised machine learning relating to identification of documents relating to human resources and containing personal information. Overall, these results show promise, although we have so far been unable to propose a more systematic approach to developing training sets. This suggests that supervised machine learning could be a viable approach for a "triage" method of reviewing collection for restrictions.**

Keywords—natural language processing, NLP, personal information, PII, digital archives, supervised machine learning, probabilistic classification, Naïve Bayes classifier

I. Introduction

This paper follows up on the author's earlier experimentation with topic modelling applied to the identification of records with personal information, reported at the 2017 Computational Archival Science workshop [1]. In that paper, we concluded in part that:

"Based on the results so far, topic modeling seems to be more successful for high-level identification of topics than drilling down to the document level. Topic modeling can help identify documents that need to be reviewed, but there is a potential for them to be buried among other documents. Training topic models focused on management, planning, human resources, etc. might at least narrow down the documents needed to be individually reviewed."

One of the proposed directions for further research was described as: "Train topic models based on documents known to be relevant, e.g. a group of HR documents." Colleagues at the 2017 Computational Archival Science workshop kindly pointed the author in the right direction: this idea really should have been framed as supervised machine learning.

More generally, Baron and Borden have articulated a research agenda as part of a call to action about the use of computational methods to enable access to digital archives [2].

This paper documents the iterations attempted in developing training sets for supervised machine learning relating to identification of HR-related documents, with preliminary findings. The earlier rounds in particular also unavoidably demonstrate the author's learning curve relating to supervised machine learning and the Weka tool.

II. Methodology

A. Overview

We used Weka, Java-based open source software for data mining, focusing on it classification module. Weka ("Waikato Environment for Knowledge Analysis") is developed by the University of Waikato, New Zealand [3].

The text corpus used for this investigation is drawn from the records of a University of Saskatchewan Associate Vice-President for Information and Communications Technology, accessioned in 2000. The full collection has an estimated 2500 documents, for the purpose of this research we were able to identify and convert to text a total of 1784 documents, primarily Word and related office formats.

For most of the classification "rounds", we defined three categories:

Personal human resources records (HR-personal): HR records relating to one or more identified individuals.

General human resources records (HR-general): Records relating to human resources but not identifying individuals, such as position postings, policy statements, correspondence or minutes relating to policy development, complement planning, etc. No attempt was made to distinguish between records that would be considered confidential or open, which will depend on the institutional context.

Records not related to human resources (non-HR): all other records

With a couple early exceptions, we used a Naïve Bayes (multinomial) classifier, a well known probabilistic classifier. Unless noted, the following Weka configuration was used:

Classifier: NaiveBayesMultinomial

Filters:

StringToWordVector (attributeIndices: 1);

RemoveType

This was implemented through the FilteredClassifier.

B. Training set development

Round 1a:

As an initial proof of concept (and experimentation with the tool), the first round used all the identified documents, at this stage identified as HR vs. non-HR.

The filters StringToWordVector and NumericToBinary were applied, with no additional cleanup, and then the Naïve Bayes classifier was applied.

Round 1b:

As an additional cleanup step, the AttributeSelection filter was applied (prior to using the Naïve Bayes classifier).

Round 2:

HR: 50 documents (first 50 in order) from HR documents

Non-HR: 50 documents from full non-HR set. Did not include any documents with relevance to HR, i.e. documents eliminated from HR set above. Went through in order, and took the first 50 good examples.

Due to the pre-processing required to extract text, the document filenames had UUID prefixes, so the order was reasonably random, not based on the original filenames.

HR: subset of identified HR documents (removed more routine items, e.g. employee thanks; need smaller set so that remaining documents can be used for actual classification); non-HR: set of similar size (select good examples, e.g. avoiding very short text)

In both cases, documents with little text were avoided (e.g. charts and tables, brief cover notes).

Round 2a:

Still using the full set of documents, this time FilteredClassifier was used.

Round 2b:

Full set of documents, with Naïve Bayes Multinomial.

Subsequent to this round, the full set was refined. Some documents had been omitted from the earlier round due to processing errors. Sixty-four documents were restored, resulting in a new total of 1747.

Round 3:

Expanded the Round 2 training set to distinguish between HR-personal and HR-general, with just under 50 documents in each.

Round 4:

Took the largest 50 documents from each folder.

Rounds 3 and 4 represent the key methodologies for developing training sets, with the sets for the remaining rounds either refining those sets or using similar methods. Results are quite different, see Tables I and II.

TABLE I. ROUND 3 RESULTS

Classified as	HR-general	HR-personal	Non-HR	Not classified	*Recall*
HR-general	92	4	15	0	***82.9%***
HR-personal	25	82	8	0	***71.3%***
Non-HR	298	9	1208	6	***79.4%***
Precision	***22.2%***	***86.3%***	***98.1%***		
Correctly identified instances		1382 (79.4%)			

TABLE II. ROUND 4 RESULTS

Classified as	HR-general	HR-personal	Non-HR	Not classified	*Recall*
HR-general	108	3	0	0	***97.3%***
HR-personal	36	79	0	0	***68.7%***
Non-HR	996	117	445	0	***28.6%***
Precision	***9.5%***	***39.7%***	***100.0%***		
Correctly identified instances		632 (35.4%)			

Round 5:

Combined the Round 4 training set for HR-personal and HR-general; kept 25 largest documents from the Round 4 HR-personal and HR-general training sets.

Round 6:

Reality check – re-run with Round 3 training set, to ensure the configuration was valid.

Round 7:

Minor adjustments to Round 4 training set (e.g. removed a few documents with bad encoding).

Round 8:

For HR-personal and HR-general, the middle set (based on size) was used, choosing 50 documents each. For non-HR, we started at the 100th largest file, and took the next 50 by size.

Round 9:

Starting with the Round 3 training set, we combined HR-general and HR-personal, resulting in two categories, HR and non-HR.

The combined set was not weeded, so the new HR set was double the size of non-HR.

Round 10:

New training set: Created new training set from the full set – 12 documents in each. However, for non-HR, not a lot of good examples were found, and a few may have been misclassified.

The results from cross-validation were quite poor, so this training set was not tested further.

Round 11:

Combined HR-general and HR-personal, with some minor adjustments, resulting in 16 HR and 22 non-HR documents.

Round 12:

Returned to the training set from Round 3, but some refinements to the full data set had been made by this point.

Round 13:

Combined HR-general and non-HR using the Round 5 training set. Added 50 more documents to HR-personal (next 50 largest)

Round 14:

With the Round 4 training set, added the AttributeSelection filter.

Round 15:

With the Round 3 training set, added the AttributeSelection filter

Round 16:

Refined the Round 4 training set using Round 3 non-HR data. There is no overlap between Round 3 and Round 4 data, in the non-HR folder. Swapped in the non-HR folder from Round 3, leaving the others as is.

III. General analysis

A. *Two categories*

For rounds with two categories – generally HR and non-HR – the results are as follows. See Table III for details.

Recall:

- Recall for HR is excellent across the board (lowest is 89%); highest score is in round 13 (99%)
- Recall for non-HR is mixed; rounds 2a, 2b and 9 score between 76% and 91%; the others range between 29% and 45%; the highest score is from round 2b.

Precision:

- Precision for HR is uniformly low; highest score is 40% (round 2b)
- Precision for non-HR is uniformly high, ranging from 98% to 100%; highest score is in round 13.

Overall:

- There are mixed results for the total rate of correctly identified instances; ranging from 78% to 92% (three rounds) and 38% to 49%. The highest score is in round 2b.
- Rounds 2a, 2b, and 9 have good to excellent relevance scores, but each has a low precision score for HR.

B. *Three categories*

For the rounds with three categories – HR-personal, HR-general, and non-HR – the results are as follows. See Table V for details.

Recall:

- Recall for HR-general is decent to excellent (70% to 97%) in all cases except round 14 (50%). The highest scores are in rounds 4 and 7 (97.3%)
- Recall for HR-personal is also reasonable, ranging from 64% to 88%. The highest score is in round 8 (87.8%)
- Recall for non-HR is mixed: between 72% and 81% in five rounds; 28% or 29% in two rounds; and 57% or 59% in two rounds.

Precision:

- Precision is uniformly poor for HR-general, with a range between 9% and 33%; highest score in round 8.
- There is a significant range for HR-personal – 17% to 86%, with the best result from round 3, and the other scores all lower (mostly significantly) than 70%
- But precision for non-HR is excellent – in the mid- to high-nineties (and as high as 100%) across the board; the highest scores are in rounds 4 and 7.

Overall:

- Over 70% of instances are correctly identified in five of the nine rounds, otherwise ranging from 35% to 62%. The best result is in round 10 (81%).
- In rounds 3, 6, and 12, all recall scores are over 70%. The corresponding precision scores are highest in round 3, but with a low precision for HR-general.
- Contrary to the cross-validation scores, none of the rounds have high precision scores across the board.

C. *Precision and recall*

A trade-off between recall and precision is not unexpected. Using F1 scores [4], we can do a further assessment of the "best" results.

- For two categories, the best overall results are in round 2b, followed by round 2a then 8.
- For three categories, the best overall results are in round 4, followed by round 10 (recall that these use the same training set)

In our scenario – attempting to classify documents with personal information – it seems reasonable to argue that a high recall score for HR-personal is the most important goal. That is, if a document is HR-personal, it has a high probability of being classified as such. Along the same lines, good precision for HR-personal is also desirable, to minimize the number of documents incorrectly classified as HR-personal. The low precision scores for HR-general are not necessarily a problem, unless they correspond to low recall scores for HR-personal.

The manually generated training set from round 3 (used again in round 12) appears to have been the most effective. Unfortunately, attempts to make the process of developing a training set more systematic – for example, choosing the largest documents – were largely unsuccessful.

However, in most cases there were good results from cross-validation (see Tables IV and VI), so more analysis to determine why the full classification of those sets had poor results should be helpful.

IV. Conclusions and Further Research

Overall, these results show promise, although we have so far been unable to propose a more systematic approach to developing training sets. This suggests that supervised machine learning could be a viable approach for a "triage" method of reviewing collection for restrictions. At the very least, it could help measure the risk of whether documents requiring restrictions are to be found in a large document set. It might be part of an institution's due diligence before providing access based on other arrangements such as non-disclosure agreements and exceptions under privacy legislation that allow research use of personal information.

There are a number of potential further research directions, both to refine and better understand the current results and to build on these experiments.

- Refine the models, with more data cleanup;
- Rerun each round of testing with the final data set, as there were a few changes in the course of the current research;
- More granular training sets within the HR-personal classification, e.g., performance evaluations, benefits information, etc.[1];
- Further analysis of these results in the context of theoretical and other research about the Naïve Bayes classifier;
- Test other classification methods;
- More analysis of false positives and false negatives: what characteristics led to documents being incorrectly classified?
- Run the models against new data sets. Will the AVP Information Technology training data work with other collections?
- Different training sets, e.g. attempt to develop training sets relating to restricted vs. non-restricted records. With HR records, we were able to depend on keywords and phrases relevant to that topic. More general records will likely pose different challenges.

TABLE III. Comparative results for two categories – test set

Round #	2a	2b	5	10	11
Recall					
HR	92.9%	93.9%	98.7%	88.9%	96.9%
non-HR	81.5%	90.9%	28.8%	75.8%	40.1%
Precision					
HR	35.7%	39.6%	16.8%	36.0%	19.0%
non-HR	99.5%	99.6%	99.3%	98.1%	98.9%
Classification accuracy	89.3%	91.5%	37.7%	78.4%	47.3%
F1 score					
HR	51.5%	55.7%	28.6%	51.2%	31.8%
non-HR	89.6%	95.1%	44.6%	85.5%	57.0%

TABLE IV. Comparative results for two categories – cross-validation

Round #	2a	2b	5	10	11
Recall					
HR	90.0%	90.0%	96.0%	**96.8%**	95.5%
non-HR	92.0%	92.0%	**100.0%**	85.4%	37.5%
Precision					
HR	91.8%	91.8%	**100.0%**	92.9%	67.7%
non-HR	90.2%	90.2%	**96.2%**	93.2%	85.7%
Classification accuracy	91.0%	91.0%	**98.0%**	93.0%	71.1%
F1 score					
HR	90.9%	90.9%	**98.0%**	94.8%	79.2%
non-HR	91.1%	91.1%	**98.0%**	89.1%	52.2%

[1] Thanks to one of the anonymous reviewers for this suggestion.

TABLE V. COMPARATIVE RESULTS FOR THREE CATEGORIES – TEST SET

Round #	3	4	6	7	8	12	14	15	16
Recall									
HR-general	82.9%	97.3%	82.9%	97.3%	74.8%	77.5%	49.5%	70.3%	87.4%
HR-personal	71.3%	68.7%	71.3%	67.8%	87.8%	75.7%	79.1%	66.1%	63.5%
non-HR	79.4%	28.6%	77.5%	27.7%	58.8%	81.1%	57.4%	73.0%	71.8%
Precision									
HR-general	22.2%	9.5%	21.9%	9.4%	33.2%	25.0%	9.9%	20.1%	18.2%
HR-personal	86.3%	39.7%	65.1%	39.6%	16.9%	60.8%	31.8%	34.9%	68.9%
non-HR	98.1%	100.0%	98.1%	100.0%	98.0%	97.8%	94.8%	97.5%	98.6%
Classification accuracy	79.4%	35.4%	77.5%	34.6%	61.8%	80.8%	58.4%	72.8%	72.7%
F1 score									
HR-general	35.0%	17.3%	34.6%	17.1%	46.0%	37.8%	16.5%	31.2%	30.1%
HR-personal	78.1%	50.3%	68.0%	50.0%	28.4%	67.4%	45.4%	45.6%	66.1%
non-HR	87.8%	44.4%	86.6%	43.4%	73.5%	88.7%	71.5%	83.5%	83.1%

TABLE VI. COMPARATIVE RESULTS FOR THREE CATEGORIES – CROSS-VALIDATION

Round #	3	4	6[a]	7	8	12[a]	14	15[a]	16
Recall									
HR-general	80.4%	82.0%	80.4%	80.0%	42.0%	80.4%	60.0%	80.4%	86.0%
HR-personal	68.8%	86.0%	68.8%	90.0%	96.0%	68.8%	86.0%	68.8%	82.0%
non-HR	95.8%	100.0%	95.8%	100.0%	100.0%	95.8%	100.0%	95.8%	85.4%
Precision									
HR-general	72.5%	85.4%	72.5%	88.9%	100.0%	72.5%	85.7%	72.5%	78.2%
HR-personal	82.5%	87.8%	82.5%	84.9%	68.6%	82.5%	74.1%	82.5%	91.1%
non-HR	90.2%	94.3%	90.2%	96.6%	84.7%	90.2%	87.7%	90.2%	85.4%
Classification accuracy	81.7%	89.3%	81.7%	90.4%	79.3%	81.7%	82.0%	81.7%	84.5%
F1 score									
HR-general	76.3%	83.7%	76.3%	84.2%	59.2%	76.3%	70.6%	76.3%	81.9%
HR-personal	75.0%	86.9%	75.0%	87.4%	80.0%	75.0%	79.6%	75.0%	86.3%
non-HR	92.9%	97.1%	92.9%	98.3%	91.7%	92.9%	93.5%	92.9%	85.4%

[a] Rounds 6, 12, and 15 use the Round 3 training set; data repeated to facilitate comparison to test set results

REFERENCES

[1] Tim Hutchinson, "Protecting privacy in the archives: preliminary explorations of topic modeling for born-digital collections," Proceedings of the 2017 IEEE International Conference on Big Data. Boston, MA: 11-14 December 2017, pp. 2251-2255.

[2] Jason R. Baron and Bennett B. Borden, "Opening up dark digital archives through the use of analytics to identify sensitive content,"Proceedings of the 2016 IEEE International Conference on Big Data, Washington, DC, 5-8 December 2016, pp. 3324-3229.

[3] Weka 3: Data Mining Software in Java, available from https://www.cs.waikato.ac.nz/ml/weka/. Version 3.8.2 was used for this research.

[4] William Koehrsen, "Beyond Accuracy: Precision and Recall," Towards Data Science website. Available from https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c, accessed 15 October 2018.