Cecilia Magnusson Sjöberg & Rebecka Weegar

Means for Memo Matching (MMM): A Study of Legal Informatics and Language Technology

1 Project approach

1.1 Project team

This chapter is about the Means for Memo Matching (MMM) Project and how it has enabled studies of legal informatics[1] and natural language processing[2] in higher education[3]. Artificial intelligence (AI) tools have been one attribute for promising results. The research has increasingly been carried out over the last couple of years on an ad hoc basis at two Stockholm university departments, namely the Law Department and the Department of Computer and System Sciences.[4] It should be empha-

[1] *Legal informatics* is commonly understood as a technologically oriented intersection of the research field Law & ICT (information and communication technology). For more on that kind of approach, see *Legal Management of Information Systems – Incorporating Law in e-Solutions*, Cecilia Magnusson Sjöberg (ed) (Studentlitteratur 2005). The other field within this context is usually labelled (substantive) *ICT Law* and takes an interest in how to interpret and apply law in digital environments, such as the internet. See further, e.g., *Rättsinformatik i det digitala informationssamhället*, Cecilia Magnusson Sjöberg (ed) (Studentlitteratur 2021). See also Cecilia Magnusson Sjöberg, 'Legal Automation: AI in Law revisited' in Marcelo Corrales, Mark Fenwick and Helena Haapio (eds), *Legal Tech, Smart Contracts and Blockchain* (Springer 2019) pp. 173–187.

[2] *Natural language processing* is an area of research and a set of methods and technologies for processing human language with computers.

[3] This refers primarily to university education.

[4] List of participants: *Cecilia Magnusson Sjöberg*, *Stockholm University*, LL.D., Professor of Law & Informatics, Subject director, *Rebecka Weegar, Stockholm University* PhD, Lec-

sised that the text here presented is merely a beginning of forthcoming research in this environment. More precisely the notion of MMM works as a trigger of investigations into the interplay of various kinds of matching of legal texts such as machine grading versus manual grading etc. So, in this project legal and computer science researchers collaborate on the question if the grading of a short-written assignment in higher education can be fully or partly automated with the use of AI (artificial intelligence) tools. The legal researchers in the project have an approach based in legal informatics. The computer scientists mainly draw on expertise from the field of language technology.

A good project team is essential in many aspects. The current MMM Project is an example thereof. In this context, the research requires an understanding of the interplay between law, language, and technology. In the MMM Project, emphasis is mainly placed on methodological issues, but knowledge of facts and other substantive matters is also taken into consideration. Examples of substantive issues include basic information about the normative hierarchy of legal sources such as constitutional laws and (decided) court cases as well as linguistic classification systems. In other words, *a mix of skills* is needed in a project of this kind, and these skills must in its turn be inserted and integrated into the analysis. For instance, it can be noted that the MMM Project team includes both junior and senior researchers.

1.2 Starting points

One initial and major assumption in the MMM Project is that grading at universities can under certain circumstances be performed wholly or partially *automatically*. This implies that full automation is not a goal. A second assumption is that the generic and multifaceted notion of *grading* needs to be specified. Thirdly, we assume that *AI-based* solutions are promising in learning analytics.

The *setting* of this study has as mentioned above been the Law programme at the Department of Law, Stockholm University, in collaboration with the Department of Computer and Systems Sciences (DSV). The test material is a compulsory short written assignment (one page memo, see Annex 1 for more details on the writing instructions) in which

turer at Department of Computer and System Sciences, *Johan Rosell, Stockholm University*, Research assistant.

students have to discuss the General Data Protection Regulation[5] from a methodological point of view. To be a bit more precise major grading features are (a) *facts*, i.e. how well a student is able to relate to adequate data in the current situation. Next step is (b) *focus*, i.e. the ability to apply an analytical approach. Finally (c) *form* is relevant, i.e. professional document management.

The students primarily concerned are those taking today's mandatory course 'Rättsinformatik' (Legal informatics) during the fourth year of the Law programme at Stockholm University. More information on relevant parts of the syllabus, etc., will follow below. It is important however, already here to note that the course in question reflects also in general terms how Law & information and Communication Technology (ICT), since the beginning of the 1980s, has played an important role at Stockholm University, not only within legal education but also in teaching for instance tech students.[6] Other components currently include digitalisation and internationalisation in the light of privacy and data protection, automatic and autonomous decision-making and legal aspects of information security. Letting law play a *proactive role*, instead of merely functioning as a reactive conflict-solving mechanism when things have already gone wrong, is a critical success factor.

Common denominators within the MMM Project are *grading* and *graders*. In this context it is important to note that the grading of the mandatory task of completing a written assignment – a methodologically oriented memo – on the topic *General Data Protection Regulation* is not equivalent to the more differential grading scale used in the students' final course grades. Instead, students receive feedback in the form 'fail', 'good' or 'very good'. In order to receive a final course grade, the requirement is a passing grade ('good' or 'very good'). The more fine-grained categorisation is made so as to reflect the structure of the final *exam*. When it comes to graders, there are a variety of set-ups. Graders can be more or less qualified, interested in the topic area, pedagogically skilled, etc. In the MMM Project, we included one senior grader and one junior grader. To conclude, there is a major distinction to be made between

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation), commonly referred to as the GDPR.

[6] For such an approach see, e.g., *IT Law for IT Professionals – an Introduction*, Cecilia Magnusson Sjöberg (ed) (Studentlitteratur 2005).

grading of the one-page written assignment on the one hand and the final general course grade on the other. As long as a student pass the written assignment the result will not have any particular impact on the final course grade as such which is given on the scale "AB", "BA", "B", or "underkänd"/"fail".

A project of this kind can easily become comprehensive when it comes to data collection, processing, and management. For an overview, see parts of the Memo corpus in Annex (B).[7] For reasons that we will describe below, we chose to delimit the primary scope to only 'fail' in certain parts of the study. A consequence of this delimitation was a need for 'fail' features that became a task in itself within the project.

1.3 Structure of contents

The contents of this chapter are structured in the following way. The text begins with the *project approach* described in terms of how the project team was composed as regards scientific skills and seniority. Points of departure are then conventionally expressed as hypotheses to be verified or falsified. The part that reports on *project activities* is vital for the study. From a legal point of view, an overview of the *legal framework* is also crucial. The concluding remarks will probably be of greatest interest to the reader. In addition, there is some documentation to be found in the *annexes*. *References* are naturally listed in footnotes.

2 Project activities

2.1 Research set-up and the data used in the study

Given the fact that this book chapter is quite atypical contributing to the legal domain in a digital setting rather than in the traditional theoretical dogmatic format. This is the overall explanation why conditions and outcome of the studies are presented in a seemingly abstract format way of notes rather than full sentences in the traditional way. General aspects of the included studies are thus listed as a next step.

[7] On this kind of research, see for instance Cecilia Magnusson Sjöberg, *Critical Factors in Legal Document Management: A Study of Standardised Markup Languages. The Corpus Legis Project* (Jure 1998).

Included MMM studies – general aspects

Memos were generated each semester (approx. 200–250)
Project Manager: study outline (design)

Manual grading Grader 1 and Grader 2
Pilot study 1A Memos 1–24
Pilot study 1B Memos 25–49
These memos were graded separately and then jointly

Manual grading Grader 1 and Grader 2
Major study 2 Memos 50–499
Grader 1 Memo 50–274
Grader 2 Memo 275–499
Graded separately
Grades were not negotiated in the major study.

In summary, when using the machine with training and validation data, the researchers used:
– 432 (450 - 18) memos in the range 50–499.
– Five extra memos graded 'fail' were added (to increase the number of memos in that particular dataset).
– the 49 memos from the pilot study.

For the purpose of use as test data, the grading of 18 memos (the ones divisible by 25) were extracted, see above.

Pilot studies (1)

1A Memos 1–24 (= 24)
All 24 memos were first graded separately by two graders (without any prior discussion).
Grades were negotiated for the purpose of future consistency in grading.

1B Memos 25–49 (= 25)
All 25 memos were graded separately (there was some synchronization between the two graders from pilot study 1A).
Grades were negotiated for the purpose of future consistency in grading.

Major study (2)

2

Memos 50–499 (= 450) plus 5 additional memos graded 'fail' and the 49 memos from the pilot study, i.e., in total 450 + 5 + 49 = 505 memos Multiple applications were run.
274 were graded by Grader 1 and memos 275–499 were graded by Grader 2. Note that there was some synchronization between the two graders from pilot studies 1A and 1B.

– 49 (randomly selected) memos from within the range 275–499 were also graded by Grader 1, to be used for consistency between the two graders in the major study. These 49 memos were used as the validation dataset.

– The results of the grading of 18 memos (every 25th in the range 50–499) were retained by the graders for use as test data in the comparison between 'man' (grader) and 'machine' (algorithm) as a final test dataset. The other results of the grading (432 memos) were used as training and validation data. As mentioned, there was some synchronization between graders from pilot studies 1A and 1B.

The following observations appear to be of particular interest within the MMM Project. To begin with, *the end result* including the classifier is in itself interesting. The *classifier* can simplified be described as the machine generated classification "model", based on training data, for sorting written assignments into the categories 'fail' or 'pass' respectively. We have also noted a co-existing consistency as well as discrepancy among *junior and senior graders*. This applies also internally with regard to each *individual grader's* consistency with himself/herself. Furthermore, there are potentials associated with a *combinational approach* (human beings and AI: training data, validation data and test data). Mention should also be made of the impact of *negotiations* among graders), e.g. in terms of unwanted vagueness.

2.2 Narrowing down the scope of the analysis

One task that emerged during the study was a need to limit the scope of the analysis to only two output categories[8] (fail/pass) instead of three (fail/good/very good). From the start, there was a relatively clear distinction between 'fail' on the one hand and 'good'/'very good' on the other for both the human graders and the machine learning (ML) classifier. However, the distinctions within the category 'good/very good' were much vaguer. Therefore, the decision was made to focus on *the 'fail' assessments and underlying 'fail' features*. This has surely had some impact on the end results, but the authors believe the delimitation was justified.

To illustrate the concept, a few 'fail' features identified by the graders and later used by the ML classifier are listed below. The left-hand side shows what might be referred to as features that should be included in a memo, and on the right-hand side, a few features that should be excluded from a memo are mentioned.

'Fail' features

Should be included in text	Should be excluded from text
• Sufficient number of words	# Checklists
◦ Comprehensiveness	# Grammatical mistakes
• Paragraphs	# Plagiarism
◦ Readability	
• Editing language	
◦ English	
• References	
◦ Articles	
◦ Governing frameworks	
• Important concepts	
◦ Controller	
◦ Data subject	
◦ GDPR	
◦ Privacy	

[8] In section 4 these are called "labels". The terms output categories and "labels" can be used interchangeably.

3 Legal framework

The legal framework surrounding the project primarily in terms of applicable rules and regulations consists of a multitude of smaller parts. Here it is important to emphasize that the MMM Project is more or less completely methodologically oriented. The next stage in the work will however broaden the analysis (scope) towards substantive (material) law. The overall ambition has therefore been to review and ensure legal compliance, rather than to perform in-depth analyses of the law in force. Such exercises can already be found in the legal doctrine addressing the legal implications of information and communication technologies. Instead, at this stage of the MMM Project primary concern has been that personal data was processed in accordance with the General Data Protection Regulation, as there is no doubt that the MMM Project involves *personal data* processing that falls within the scope of the GDPR.

The kind of provisions that need to be taken into consideration can be exemplified by governing legal definitions (Article 4) and the important distinction between anonymisation (where the General Data Protection Regulation does not apply) and pseudonymisation (where the General Data Protection Regulation does apply). Further, there are general data protection principles (Article 5) that must be adhered to, such as lawfulness, fairness, and transparency. Of utmost importance is the requirement that the so-called controller has a legal ground for the processing, e.g., the data subject's consent, making processing lawful (Article 6). The information duties (Articles 12–15) can in practice be quite burdensome, as they comprise both information to be provided upon the initiative of the controller and also upon the request of the data subject (Articles 12–15). Applied automated decisions, including profiling, are another aspect of the algorithms and associated models in the project (Article 22). Attention should also be paid to legal system development (Article 25), i.e., data protection by design and by default. For further reflections concerning for instance fulfilment of information duties, see the annexes.

Another regulation of great importance is the European Commission's proposal COM(2021) 206 final, for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending certain Union legislative acts. As an example, mention could be made of the 44 legal definitions laid down in Article 3 of the proposal, comprising for instance the following concepts relevant to artificial intelligence (AI):

> (29) 'training data' means data used for training an AI system through fitting its learnable parameters, including the weights of a neural network;
>
> (30) 'validation data' means data used for providing an evaluation of the trained AI system and for tuning its non-learnable parameters and its learning process, among other things, in order to prevent overfitting; whereas the validation dataset can be a separate dataset or part of the training dataset, either as a fixed or variable split;
>
> (31) 'testing data' means data used for providing an independent evaluation of the trained and validated AI system in order to confirm the expected performance of that system before its placing on the market or putting into service;

In addition to data protection regulation and a compliance check with the proposed AI regulation, there is quite a lot of legislation that one must be aware of and comply with when the setting is the *public sector*. This is the case in the MMM Project: in Sweden publicly funded universities are government agencies (*myndigheter*). This means that the automated grading in the MMM project should be compliant with both general and special *administrative law* governing teaching activities, including examination measures of different kinds (such as those included in the MMM Project). Fundamental principles of *openness* capturing transparency[9] vs. *secrecy* (confidentiality) are also to be considered during system design, development, implementation, and management. From a legal point of view, legal *digital archives* that are synchronised with daily information flows are also on the regulatory wish list.

Last, but not the least, *ethical considerations* must be made. It is important to let ethical vetting and similar measures play a separate role, so as not to be directly incorporated as law (generally speaking).

4 Machine learning approach

The topic of the MMM study is the automatic scoring of written assignments or essays[10], which falls within the research area of natural language processing. Natural language processing is a subfield of artificial intelli-

[9] Regarding this interplay, see Cecilia Magnusson Sjöberg, 'Legal AI from a Privacy Point of View: Data Protection and Transparency in Focus' in Sonya Petersson (ed), *Digital Human Sciences* (Stockholm University Press 2021) DOI: https://doi.org/10.16993/bbk.h.

[10] AES is an acronym for automated essay scoring.

gence that concerns automatic processing of spoken and written human language.

Often, some kind of machine learning is used for automatic essay scoring[11]. Machine learning (ML), another subfield of AI, can be used as a tool to accomplish a wide array of tasks involving data, and the goal is often to train a classifier to mimic the behaviour of an expert in some field. This can be accomplished by applying a learning algorithm to examples labelled by an expert. In the case described here, the expert is the grader, the task is assessment of student texts, and the data are texts written by students, with associated feedback labels: 'pass' or 'fail'. The result of the training is a classifier adapted to assessing student texts that are similar, but not identical (this would be plagiarism, for which there are other tools for discovery), to those in the training data. In this case, the classifier is adapted for the particular student assignment described in this study.

4.1 Text features

When applying machine learning to assessment of students' assignments, it is important to consider how to represent the students' texts. One main component is of course the contents of each text: the words that the student has used. However, there are also other characteristics of a text that can influence the assessment or that can indicate the overall quality of the text. Examples include the length of the text, the vocabulary, and any errors in spelling or grammar. The characteristics selected to represent a text are called *features* and the selection of features has a large impact on how well a machine learning classifier can be trained to perform classification.

In this study, a number of different features have been included. They include the 'features of fail' discussed in section 2.2, and also features that describe the structure of a text, such as the number of paragraphs and headings therein[12] (see further Annex C).

[11] For a survey on notable AES systems, see: Semire Dikli, 'An Overview of Automated Scoring of Essays' (2006) The Journal of Technology, Learning and Assessment 5.1. For an overview of recent AES research, see Ke, Zixuan, and Vincent Ng, 'Automated Essay Scoring: A Survey of the State of the Art' (2019) IJCAI vol. 19.

[12] Three libraries were used to generate features, pyspellchecker by Peter Norvig (https://norvig.com/spell-correct.html), language-tool-python (https://pypi.org/project/language-tool-python/), and Natural Language Toolkit, see Steven Bird, Edward Loper and Ewan Klein, *Natural Language Processing with Python* (O'Reilly Media Inc. 2009).

4.2 Outlier features

A majority of the students' texts were similar in form and content, but some texts had more unusual characteristics. We therefore included features to identify texts that were significantly different from the 'average' text. For example, a majority of the texts consisted of around five paragraphs, but a smaller number of texts consisted of either one long paragraph or many, very short paragraphs, resembling a list. If the number of paragraphs deviated significantly from the norm, information about this was included as a binary feature[13]. Other examples of features in this category were unusually few references to articles in the General Data Protection Regulation and a text being unusually short. The total number of outlier features for each text was also included as a feature.

These outliers were used to automatically generate feedback comments which could be displayed either to a grader or directly to the student who had handed in the text.

A feature that is useful in machine learning should correlate with the target label, in this case the 'pass'/'fail' assessment. The figure in Annex C shows the correlation between a selection of the included features and the 'fail' label. In that figure, it can be noted that the 'outlier features' have a stronger correlation with the feedback label ('pass'/'fail') than the original features based on the number of occurrences of some characteristic. For example, knowing the number of paragraphs is less informative than knowing that the number of paragraphs is either very low or very high compared with the average for all texts.

4.3 Training the classifier

To develop a classifier based on machine learning, a suitable dataset must be selected. As mentioned in the examples of AI relevant concepts in section 3, the dataset is usually divided into distinct parts. In this study, a majority of the data were used for training and the remaining data were used for validation and evaluation of the trained classifier.

A common strategy is to use a validation dataset during development, for instance to select which features to include. An alternative (or complementary) approach is to apply cross-validation. During cross-validation, the training data are further divided into smaller segments. A

[13] A binary feature represents a characteristic that is either present in the text or not.

model is trained on all but one of the segments and then evaluated on the remaining data. This is repeated until all segments have been used for evaluation. This method is suitable when limited training data are available, since it eliminates the need for a separate validation dataset. Here, cross-validation was used to select and finetune the learning algorithms, and to select which features to include. The best choices found during cross-validation were evaluated against the validation data and the final test dataset.

In all, 18 memo texts were set aside as a test dataset, 49 memo texts were used for validation and the remaining 437 texts were used to train the classifier. For the training data stemming from the pilot study, the negotiated grades ('feedback') of the two human graders were used as labels. The remaining training data was only graded by one of the two graders meaning that no negotiation was employed for labelling. The labels, 'pass' and 'fail', assigned by the graders were thus used as the 'ground truth', where the goal of training is for the classifier to be able to assign these labels in a way that mimics the human labelling.

4.4 Evaluation and performance measures

Two important measures for evaluating the performance of a classifier are recall and precision. When the task is to classify texts into assessment categories, a high precision value for a category like 'fail' means that the texts assigned to that specific category truly belong in that category. A high recall means that the classifier correctly, and among all available texts, identifies the texts that should belong in a specific category. Precision and recall values can range from 0 to 1, where 1 is a perfect score. A recall of 1 for a category means that the classifier has correctly identified all the texts in that category, while a recall of 0.5 means that the classifier has missed half of the texts.

A good classifier should have high scores for both precision and recall, but there is often a trade-off: when you increase precision, you might decrease recall and vice-versa. Imagine that we have a classifier that classifies all assignments as 'fail'. This corresponds to a recall of 1 for the label 'fail': every text that deserves the feedback 'fail' will be labelled as 'fail', but the precision would be low, as we would fail many students that deserve to pass. On the other hand, if the classifier would only assign the feedback 'fail' to a single text deserving of that feedback, the classifier would have perfect precision for the label fail, since no students that should pass

would be failed. Yet, this would also result in many false negatives (students that should have failed get a pass), meaning that recall would be very low. While the ideal is that a classifier has both high precision and recall, one of the measures can be prioritized over the other.

In this case, we were particularly interested in a high recall for the category 'fail', as one goal was to correctly identify all the texts that lacked some quality necessary to get a passing grade on subsequent assignments in the course.

4.5 Agreement and Cohen's kappa

Ideally, a classifier that has been trained for a specific task should agree with human experts performing the same task. However, how well a classifier can be expected to perform varies depending on the type and complexity of the task. One way of estimating an upper bound for the expected performance is to measure the agreement between two or more experts on the same task. A well-defined task should yield a high level of agreement, while a more complex or less well-defined task can result in lower levels of agreement. A common way of measuring agreement between experts in text classification and automatic grading and assessment is to use Cohen's kappa, κ[14]. The maximum possible value for this measurement is 1.0. Here, the texts in the validation data were graded by two graders, independently, and the κ for them was calculated to be 0.43[15], which indicates a certain level of agreement, but not complete agreement. One of the graders assigned 'fail' as feedback to five texts in the validation dataset, and the other grader assigned 'fail' as feedback to seven texts. In all, three texts were given the feedback 'fail' by both graders.

[14] Cohen's kappa, κ, measures agreement between two assessments while adjusting for chance agreement, see Jacob Cohen, 'A coefficient of agreement for nominal scales' (1960) Educational and Psychological Measurement 20.1, 37–46.

[15] For an interpretation of kappa scores, see J. Richard Landis and Gary G. Koch, 'The measurement of observer agreement for categorical data' (1977) Biometrics, 159–174. They denote a score between 0.41 and 0.6 as moderate agreement, and require a score of at least 0.61 for substantial agreement.

4.6 Partially automated grading

There were two main challenges in constructing a useful classifier for the task described here. First, 'fail' was a *minority class*, meaning that there were only a handful of texts (13%) that received a 'fail' assessment in the training data. Therefore, the learning algorithm had only a few instances to learn from, making it difficult for it to determine which feature patterns corresponded to a likely assessment of 'fail'. Second, the manual grading of the validation dataset showed a fairly low agreement between the two graders. This means that some texts with similar characteristics were possibly assigned contradictory labels in the training data.

For these two reasons, it was not expected that a classifier trained on these data could fully replace an expert grader, which caused us to set a second goal: to reduce the number of texts needing manual grading by half. Such 'partially automated grading' could be achieved by letting the classifier divide the texts into two groups: one group of texts with a high probability of a passing grade and one group of texts with some probability of a failing grade. Manual grading would only be needed for the group with some probability of 'fail', while all other texts could automatically be given a passing grade.

The classifiers were therefore used to rank all the texts in the test and validation datasets, from highest to lowest probability of 'fail', the goal being that all texts in the category 'fail' should end up in the top half of the list, while the texts in the bottom half could be considered as passing. This would be possible for a classifier that assigns a probability for each text to belong to the class *Fail*. Usually, if this probability exceeds 0.5, the class *Fail* would be assigned by the classifier. Here, we lowered the probability threshold until half of the texts were placed in the 'possible fail' group. This can also be understood as increasing the threshold for considering a text as belonging to the class *Pass*. Only texts with a very high probability of belonging to the class *Pass* would be assigned to that class by the classifier.

4.7 Results of the automatic text assessment

Two different learning algorithms performed well during the cross-validation: Random Forest (RF) and Gaussian Naive Bayes (GNB)[16]. These final classifiers were evaluated against both the validation dataset and the test dataset. The validation dataset consisted of 49 texts that had been graded by two graders independently and the performance of the classification model was evaluated against both expert graders for both the RF and the GNB classifiers. Cohen's kappa values for the validation dataset with both classifiers and both graders are shown in the table below:

Evaluation set:	κ, Random Forest	κ, Gaussian Naive Bayes
Validation set assessed by Grader 1 (49 instances)	0.56	0.76
Validation set assessed by Grader 2 (49 instances)	0.46	0.56
The 43 instances in the validation set given the same feedback by both graders	0.79	0.84

These numbers can be compared to the agreement value between Grader 1 and Grader 2 for the validation dataset, which was 0.43. In terms of κ, both classifiers agreed with each of the graders individually to a greater degree than the graders agreed with each other.

When evaluating the classifiers on the subset of the validation dataset where Grader 1 and Grader 2 had assigned the same grade (the last line in the table), GNB achieved the highest agreement, 0.84. This dataset could be considered as containing student texts which are 'easier' to grade, since the borderline cases where the two graders disagreed were removed. Still, this subset could also be considered a more reliable evaluation dataset, as there were no disagreements regarding these texts.

Comparing the two classifiers in terms of precision and recall (here, both are compared with Grader 2), the RF classifier had higher precision overall, while the GNB classifier had higher recall overall. This means, that while all texts that were classified as 'fail' by the Random Forest

[16] Implementation from the Scikit-learn library: Pedregosa *et al.*, 'Machine Learning in Python' (2011) Journal of Machine Learning Research 12, 2825–2830.

classifier, were also labelled as 'fail' by Grader 2, less than half of the texts that should have been classified as 'fail' were identified. The higher recall of the GNB classifier, on the other hand, meant that the GNB classifier was more suitable for identifying the texts that should be classified as 'fail' which corresponds with our goal: to identify all the texts that should possibly be assessed as 'fail':

	Precision	Recall
Random Forest, Fail	1.00	0.43
Random Forest, Pass	0.91	1.00
Gaussian Naive Bayes, Fail	0.75	0.86
Gaussian Naive Bayes, Pass	0.98	0.95

For the final test dataset, consisting of 18 student texts graded by Grader 1, both classifiers were applied again. This dataset had not been available during the development of the classifiers and contained only one text assessed as 'fail'. When applying the classifiers to this dataset, the RF classifier did not manage to correctly identify that text. However, it was identified by the GNB classifier, which achieved a precision of 0.33 and a recall of 1.0 for the '*fail*' class and a precision of 1.0 and a recall of 0.88 for the '*pass*' class.

The GNB classifier assigning the label 'fail' to three of the texts in the test set, with the text manually labelled as 'fail' being among those three texts. If we could trust the classifier to produce the same quality of results in the future, it would be possible to manually review only the texts classified as 'fail' by the GNB classifier, while automatically passing the remaining texts. This would lead to a large reduction of manual work in grading, as about 80% of the texts would be automatically labelled.

However, for the partially automated grading, discussed in section 4.6, a threshold of 50% was set, where half of the texts would be automatically labelled as 'pass', while the other half would be given manual feedback. This approach requires more manual work, but reduces the risk of incorrectly labeling a text with 'pass'. For this partially automated grading, both classifiers managed to correctly divide the texts in the validation

dataset and in the test dataset so that only 50% would require manual grading, with no 'fail' texts being missed. This corresponds to a 'fail' recall of 1.0 for both classifiers and both datasets.

5 Concluding remarks

The MMM Project could briefly be described as a legally oriented text analysis of grading assessments in higher education (HE). The text analysis is based on language technology used for the purpose of classification by means of machine learning. More precisely, the project is about means for memo matching and enhanced equal treatment of students by automation, combined with individual manual feedback. The combination of *man (humans)* and *AI (here machine* learning) proved to be important.[17]

The overall goal was to achieve accurate and efficient grading in a specific exam situation that could involve one or several grader(s). *Consistency* was considered essential. Mention could here be made of a very limited test in the pilot study that resulted in four consistent, manually self-graded memos. Ultimately, there are two major categories of consistent grading (Grader 1 to Grader 2, Grader X to machine) and one category of consistent self-grading (Grader X to Grader X). (Here "Grader X" stands for a grader that has not been specified as either Grader 1 or Grader 2, the overall purpose being remaining anonymity.)

Along with the interplay of man and machine and issues of consistency comes the role of *negotiations* within a framework that allows *consultation* among graders. *Diversified manual grading* is another result of the study. This implies that there is a considerable variety among human graders, which in turn opens for unwanted vagueness and limitations in foreseeability. A rather complicated outcome is when two seemingly qualified graders make strikingly diverging assessments. For instance, in this study, there was one memo assessed as 'fail' by one grader and as 'very good' by another.

In a follow-up analysis, there were some indications of *seniority* among graders as a critical factor for giving a passing grade to 'odd', but accept-

[17] In total, 450 memos were included in the project. Initially, the assumption was that half of the included memos would be graded completely automatically. Furthermore, it was assumed that half of the included memos would require supplementary manual assessment (based on a digital selection). However, conditions changed throughout the pilot study (see above).

able written assignments. With an outlook focusing on *learning analytics* and achieving progress within the project, the decision was made to narrow down the scope of written assignments included. More precisely, the perspective was shifted from a general approach to *fail assessments and features*.

From an automatic grading perspective, the main result was that the hypothesis that it would be possible to reduce manual grading by at least 50% was confirmed with both the validation dataset and the test dataset using two different classifiers.

The outlier features were found to be particularly useful for classification. For example, while a student text in this case should contain content from the General Data Protection Regulation, too much overlap with the GDPR could indicate that the student might not have added much content of their own. With a larger dataset, the classifier could be expected to identify such a large-but-not-too-large overlap pattern. However, with only few examples of this particular pattern, the simple fact that a text diverges from the norm can be highly informative for classification. This approach could be further developed for automatic grading in general.

Annexes

Annex (A). General Data Protection Regulation compliance

Annexes

Annex (A). General Data Protection Regulation compliance

Stockholms universitet

Written guidelines concerning memo format

- "En A4-sida i Word (teckenstorlek 12, Callibri eller motsvarande, enkelt radavstånd, marginaljusterat och avstavat)."

"One A4 page in Word (font size 12, Calibri or similar, single line spacing, justified text with hyphenation)."

2021-06-24 11

Stockholms universitet

Information to data subjects

Aktuellt på kurswebben 2020

Information om ett forskningsprojekt

I syfte att frigöra tid för genomförande av den så kallade Stockholmsmodellen som lyfter fram vikten av kritiskt tänkande undersöker vi inom rättsinformatiken vilka lärandeprocesser som går att digitalisera på ett meningsfullt, säkert och effektivt sätt. Vi har så långt introducerat digitala föreläsningar och digitala hemtentor. Nu undersöker vi de rättsliga och tekniska förutsättningarna för helt eller delvis automatiserade bedömningar av de metod-PM avseende dataskyddsregleringen som utgör ett obligatoriskt moment under det första blocket i RINF -kursen.

12

Stockholms universitet

- Planen är att jämföra våra lärares manuella bedömningar av PM:en med vad som går att åstadkomma genom maskininlärning. Det aktuella forskningsprojektet MMM (Means for Memo Matching) sker i samverkan med Institutionen för data - och systemvetenskap också här vid Stockholms universitet. Målet är i slutänden att förbättra kursen för studenterna och att frigöra tid för andra moment. Om vi kommer fram till att det på ett rättssäkert och ändamålsenligt går att använda helt eller delvis automatiserade bedömningar av PM:en kan de undervisningstimmar kursen har till sitt förfogande användas till annan meningsfull undervisning för studenterna (istället för de timmar som nu går åt för PM -rättning).
- Skyddet av studenternas personliga integritet står i centrum varför alla texter kommer att pseudonymiseras (avidentifieras) för att undvika att enskilda individer ska kunna identifieras.
- Har du frågor kring forskningsprojektet är du välkommen att höra av dig

Course director/Course adm.

13

Currently on the course website 2020

Information on a research project

With the aim to free up time for implementation of the so-called Stockholm model, which focuses on the importance of critical thinking, we are within legal informatics studying which learning processes can be digitalised in a meaningful, secure and effective way. Thus far, we have introduced digital lectures and digital home exams. We are now studying the legal and technical conditions for fully or partially automated assessments of the method memos regarding the data protection regulation that are an obligatory part of the first block in the legal informatics course.

- The plan is to compare our teachers' manual assessments of the memos with what can be achieved through machine learning. The research project in question, MMM (Means for Memo Matching), is conducted in collaboration with the Department of Computer and Systems Sciences within Stockholm University. The aim is ultimately to improve the course for the students and free up time for other aspects. If we find that it is possible to use fully or partially automated assessments of the memos in a legal and suitable way, the teaching hours allotted to the course can be used to provide other meaningful education to the students (rather than being spent on assessing memos).
- Protection of the students' privacy is paramount, for which reason all texts will be pseudonymised (deidentified), to avoid the possibility of identifying any individual person.
- If you have any questions on the research project, you are welcome to contact [the course director].

Annex (B) Overview – 1A & 1B

Overview- 1A

Memo Numbe	Grader 1	Grader 2	Pre negotiat	Negotiat	Autc	Final Grade
Memo 1	Very Good	Very Good	Very Good	Very Good	0	0
Memo 2	Good	Good	Good	Good	0	0
Memo 3	Good	Good	Good	Good	0	0
Memo 4	Very Good	Very Good	Very Good	Very Good	0	0
Memo 5	Fail	Fail	Fail	Fail	0	0
Memo 6	Very Good	Good	TBD	Very Good	0	0
Memo 7	Fail	Good	TBD	Fail	0	0
Memo 8	Good	Fail	TBD	Fail	0	0
Memo 9	Very Good	Good	TBD	Very Good	0	0
Memo 10	Good	Good	Good	Good	0	0
Memo 11	Very Good	Good	TBD	Very Good	0	0
Memo 12	Good	Good	Good	Good	0	0
Memo 13	Fail	Fail	Fail	Fail	0	0
Memo 14	Very Good	Good	TBD	Good	0	0
Memo 15	Good	Good	Good	Good	0	0
Memo 16	Very Good	Very Good	Very Good	Very Good	0	0
Memo 17	Good	Fail	TBD	Fail	0	0
Memo 18	Good	Good	Good	Good	0	0
Memo 19	Good	Good	Good	Good	0	0
Memo 20	Very Good	Good	TBD	Good	0	0
Memo 21	Good	Good	Good	Good	0	0
Memo 22	Very Good	Good	TBD	Very Good	0	0
Memo 23	Very Good	Very Good	Very Good	Very Good	0	0
Memo 24	Very Good	Good	TBD	Very Good	0	0

2021-06-23

Statistics

Grade	Grader 1		Grader 2	
Fail	3	13%	4	17%
Good	10	42%	16	67%
Very Good	11	46%	4	17%
Sum	24	100%	24	100%

Grade	Pre negotiated		Negotiated	
Fail	2	8%	5	21%
Good	8	33%	10	42%
Very Good	4	17%	9	38%
TBD	10	42%	0	0%
Sum	24	100%	24	100%

Overview-1B

Memo Numbe	Grader 1	Grader 2	Pre negotiat	Negotiat	Automat	Final Grade
Memo 25	Good	Fail	TBD	Good	0	0
Memo 26	Good	Very Good	TBD	Good	0	0
Memo 27	Very Good	Very Good	Very Good	Very Good	0	0
Memo 28	Good	Good	Good	Good	0	0
Memo 29	Fail	Very Good	TBD	Good	0	0
Memo 30	Very Good	Very Good	Very Good	Very Good	0	0
Memo 31	Very Good	Very Good	Very Good	Very Good	0	0
Memo 32	Fail	Very Good	TBD	Good	0	0
Memo 33	Good	Good	Good	Good	0	0
Memo 34	Good	Good	Good	Good	0	0
Memo 35	Fail	Very Good	TBD	Good	0	0
Memo 36	Very Good	Good	TBD	Good	0	0
Memo 37	Good	Good	Good	Good	0	0
Memo 38	Good	Very Good	TBD	Very Good	0	0
Memo 39	Very Good	Very Good	Very Good	Very Good	0	0
Memo 40	Good	Very Good	TBD	Good	0	0
Memo 41	Very Good	Very Good	Very Good	Very Good	0	0
Memo 42	Good	Good	Good	Good	0	0
Memo 43	Fail	Good	TBD	Good	0	0
Memo 44	Good	Very Good	TBD	Good	0	0
Memo 45	Very Good	Very Good	Very Good	Very Good	0	0
Memo 46	Very Good	Good	TBD	Very Good	0	0
Memo 47	Good	Very Good	TBD	Good	0	0
Memo 48	Good	Good	Good	Good	0	0
Memo 49	Fail	Good	TBD	Fail	0	0

Statistics

Grade	Grader 1		Grader 2	
Fail	5	20%	1	4%
Good	12	48%	10	40%
Very Good	8	32%	14	56%
Sum	25	100%	25	100%

Grade	Pre negotiated		Negotiated	
Fail	0	0%	1	4%
Good	6	24%	16	64%
Very Good	6	24%	8	32%
TBD	13	52%	0	0%
Sum	25	100%	25	100%

2021-06-23

Annex (C) Correlations between features and the 'fail' category

The correlation between the feedback 'fail' and each feature can be seen in the last row. *Nouns, verbs, numbers, pronouns* and *determiners* represent how common each of these parts of speech is in each student text. *General Data Protection Regulation references* correspond to the number of references made to specific articles in the General Data Protection Regulation and *Vocabulary* corresponds to the number of different words used in each text. The remaining features are 'outlier' features, except *General Data Protection Regulation content*, which shows a (weak) correlation between getting a passing grade and a high degree of *General Data Protection Regulation content*.