Implicit readability ranking using the latent variable of a Bayesian Probit model

Johan Falkenjack
SICS East Swedish ICT AB
Linköping, Sweden
johan.falkenjack@liu.se

Arne Jönsson
SICS East Swedish ICT AB
Linköping, Sweden
arne.jonsson@liu.se

Abstract

Data driven approaches to readability analysis for languages other than English has been plagued by a scarcity of suitable corpora. Often, relevant corpora consist only of easy-to-read texts with no rank information or empirical readability scores, making only binary approaches, such as classification, applicable. We propose a Bayesian, latent variable, approach to get the most out of these kinds of corpora. In this paper we present results on using such a model for readability ranking. The model is evaluated on a preliminary corpus of ranked student texts with encouraging results. We also assess the model by showing that it performs readability classification on par with a state of the art classifier while at the same being transparent enough to allow more sophisticated interpretations.

1 Introduction

Modern models of readability analysis for classification often use classification algorithms such as SVM (Petersen, 2007; Feng, 2010; Falkenjack et al., 2013) which give us an assessment whether a text is easy-to-read or not. Such models have a very high accuracy, for instance, a model using 117 parameters from shallow, lexical, morphological and syntactic analyses achieves 98,9 % accuracy (Falkenjack et al., 2013). However, these models do not tell us much about whether a given text is easier to read than any other text, other than the binary classification. In order to perform a more fine grained prediction we normally need to train the models using a corpus of graded texts, for an overview of such methods see Collins-Thompson (2014).

There are attempts to grade texts without an extensive corpus of graded texts (Pitler and Nenkova, 2008; Tanaka-Ishii et al., 2010). Tanaka-Ishii et al. (2010) present an approach which predicts relative difficulty based on pairwise comparison between texts and thus models degree of readability on an ordinal scale. Probabilistic models have also been used, by for instance Martinez-Gómez and Aizawa (2013). Their model is based on a Bayesian network that comprises 22 linguistic features. The model is trained on a corpus with eye fixations as a measure of reading difficulty. The paper focuses on readability diagnosis and presents a variety of linguistic features important for readability.

In this paper we present a novel way of assessing the relative readability of texts based on the latent variable of a Bayesian Probit model trained for classification. The Probit model is a straightforward way to model the probability that a text is considered easy-to-read or not. In its traditional interpretation it belongs to the wider family of linear classification models. However, in this paper we illustrate that when using a latent variable interpretation, the Probit model lends itself particularly well to readability assessment.

The probabilistic nature of the Probit model allows us to interpret and assess the relative readability between texts. Roughly, if $P(Y_{Text_1} = 1|X_{Text_1}) > P(Y_{Text_2} = 1|X_{Text_2})$ then $Text_1$ is easier to read than $Text_2$. Meanwhile, the latent variable underlying the Probit model is even easier to interpret and might even be viewed as a semi-linear and affine measure of degree of readability in itself.

2 Method

We construct both a Probit model and a Support Vector Machine for classifying easy-to-read texts. However, here we focus on presenting the Probit model as the SVM is already well established in the field of readability assessment. We also present the various evaluation methods used to assess the Probit method. We begin, however, by describing our corpora.

2.1 Corpora

We use three different Swedish corpora. Two are used to train the model and evaluate the classification performance of the model. The third corpus is used to evaluate the ranking performance of the model.

2.1.1 Training and classification evaluation

The source of easy-to-read texts is LäSBarT (Mühlenbock, 2008). We also use the general text corpus SUC (Ejerhed et al., 2006). From each of these two corpora we have selected 700 texts with a similar distribution of lengths. We make the assumption that texts from SUC are not easy-to-read but in reality we expect a small portion of the SUC-texts to actually be easy-to-read. This means that the sets are not well separated making perfect classification infeasible.

This set of 1400 texts from LäSBarT and SUC are labelled according to their source corpus and split into a training set of 350 texts from each corpus and a test set of the same size.

2.1.2 Ranking evaluation

We also use 9 sets of ordered texts from the MASTER project (Kanebrant et al., 2015). This corpus consists of 30 texts split into partially overlapping sets. There are 14 samples of general fiction split in 3 partially overlapping sets of 6 texts, 8 samples from social science textbooks split in 3 partially overlapping sets of 4 texts, and 8 samples from natural science textbooks split in 3 partially overlapping sets of 4 texts.

Each set is ordered through the texts being labelled with the average performance of weak readers on reading comprehension tests based on the text, where weak readers are defined as those readers scoring below the mean on all texts they were assigned (3 for general fiction or 2 for social and natural sciences). The readers were students in the Swedish 4th grade, 6th and 8th grade corresponding to an age of 11, 13 and 15 years respectively, giving rise to the 3 sets in each genre. The number of weak students we had access to varied, but for grades 4 and 6 we had roughly 200 weak students per text each year, whereas the number of weak students for grade 8 were rarely above 30 per text.

We refer to this corpus as preliminary as the sample of students used in this paper does not correspond to all data collected by Kanebrant et al. (2015) and has not gone through rigorous post-processing. At this time however, this is the only empirically based readability ranked data in Swedish.

2.2 Feature set

We use the small set of features for readability classification presented in Falkenjack and Jönsson (2014). This feature set is the result of a genetic feature selection scheme in an attempt to eliminate parsing based features[1].

The set consist of 8 features, of which 6 are part-of-speech (in which delimiters are included) tag unigrams for adverbs (AB), interrogative/relative possessive pronouns (HS), cardinal numbers (RG), and major (MAD), minor (MID) and pairwise (PAD) delimiters. The last two features consist of OVIX, or Word variation index (Ordvariationsindex), a type-token ratio based measure normalized for use on texts with different lengths, and SweVocH, the ratio of words in the text belonging to a lexicon of ”highly frequent words” in a reference corpus. The feature set was optimized specifically for use with an SVM classifier (Falkenjack and Jönsson, 2014) but we assume it will work well enough with other models as well.

[1] Accuracy of the small feature set on the task of classifying the SUC and LäsBart corpora is 98.5% compared to 98.9% for a 119 parameter model, that also includes parsing based features, although these high accuracies might be a result of some over-fitting due to the lack of a separate test set not used for cross-validation.

2.3 For comparison: Support Vector Machine

The Support Vector Machine (SVM) is a linear classification model that can be viewed as state-of-the-art in easy-to-read classification with high accuracies achieved even with small feature sets (Falkenjack and Jönsson, 2014). The SVM works by finding a hyperplane in feature space which optimally separates two classes. This can be extended to non-linear models using different kernel functions but in this paper we use a linear SVM. Fitting an SVM to data entails solving a Quadratic Programming problem, in this paper we use Platt's Sequential Minimal Optimization algorithm (Platt, 1998) accessed through the `kernlab` library (Karatzoglou et al., 2004) for the `R` statistical programming language.

2.4 Probit model

The Probit model is a well established model in statistical learning, introduced in the 1930s (Gaddum, 1933; Bliss, 1934) and used primarily for classification. It is closely related to the younger but somewhat more well known Logit model, or Logistic regression, but has some properties which makes it especially suitable to Bayesian modelling (McCulloch et al., 2000). The Probit model takes the form given in Equation 1

$$\Pr(Y = 1 \mid X) = \Phi(X^T\beta) \tag{1}$$

where Φ is the Cumulative Distribution Function for the Standard Normal Distribution (if Φ is replaced by the logistic function we get the Logit model), Y is the dependent variable, or label, and X are the covariates, or features, on which Y depend. In the readability classification case, Y is an indicator value indicating whether the text is easy-to-read or not, while X is a vector of feature values of the features covered in Section 2.2.

A particular strength of the Probit model for readability analysis is that it can be viewed as a **latent variable model**. A latent variable model is a model which assumes one or more unobserved, or latent, variables which connect observed variables. If we assume a latent variable Y^* and an error term $\epsilon \sim N(0, 1)$ then a latent variable equivalent to Equation 1 can be written as Equation 2.

$$\Pr(Y^* > 0 \mid X) \quad \text{where} \quad Y^* = X^T\beta + \varepsilon \quad \text{and} \quad Y = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

This formulation allows us to view the Probit model as a linear regression over an unobserved, or latent, real valued variable which underlies the assigned labels in the classification problem. This is particularly useful when different classes are defined by the degree of some linear property as in the case with easy-to-read classification where the underlying property is degree of readability which now is being indirectly modelled on an interval scale.

2.5 Model estimation: Gibbs Sampling

Fitting a Probit model to data can be done in a number of ways. For the traditional Frequentist Probit model, often referred to as Probit regression, Maximum Likelihood Estimation is the preferred method. This approach is well established but yields only a point estimate of β.

In the Bayesian framework we want to estimate the full posterior distribution for the model. For the Probit model, as for many other statistical models, this problem turns out to be infeasible to solve analytically so a numerical approach is generally applied instead. A Markov Chain Monte Carlo sampler can be used to draw a sample from the posterior distribution of a Probit model. If the prior distributions over β are Normal we have conjugacy with the Normal distribution of the errors and a Gibbs sampler can be constructed using a data augmentation scheme based on the latent variable interpretation presented above (Albert and Chib, 1993).

We use a `C++` implementation of such a Gibbs sampler, accessed through the `R` statistical programming language using the `MCMCpack` library (Martin et al., 2011).

2.5.1 Prior

As we use a Bayesian fitting method we need to supply a prior. However, we opt to use only a weak regularization prior rather than a more informative prior based on any belief about specific coefficient values.

We use 0 as prior mean for all coefficients and $100 \times I$ as prior covariance ($= 1/100 \times I$ prior precision), effectively putting an independent $N(0, 100)$ prior on each coefficient. This prior is only weakly informative but keeps the coefficients from growing towards infinity and the covariances from growing towards 1, thus avoiding over-fitting. We also standardize the covariates, which avoids introducing a bias to any specific coefficient when using equal priors.

2.5.2 Sampling scheme and MCMC diagnostics

For a Gibbs sampler of a Bayesian Probit model, the output is draws from the posterior distribution over the coefficients β. As successive draws from an MCMC sampler are not independent, we thin our sample by keeping only every 200th generated draw. We also discard the first 500 thinned draws in a burn-in phase. We first show that the sample converges to the true posterior by running 5 parallel chains, taking a sample of 5 000 draws from each chain. To illustrate that the chains converge on the same distribution we calculate the potential scale reduction factor (Gelman and Rubin, 1992) for each feature and plot the value of these as the chain gets longer (Brooks and Gelman, 1998). The so called Gelman plots in Figure 1 show values very close to 1 for all features and thus indicate that the chains have converged to the same distribution which we can assume is the true posterior. The MCMC diagnostics are performed using the `CODA` library (Plummer et al., 2006).

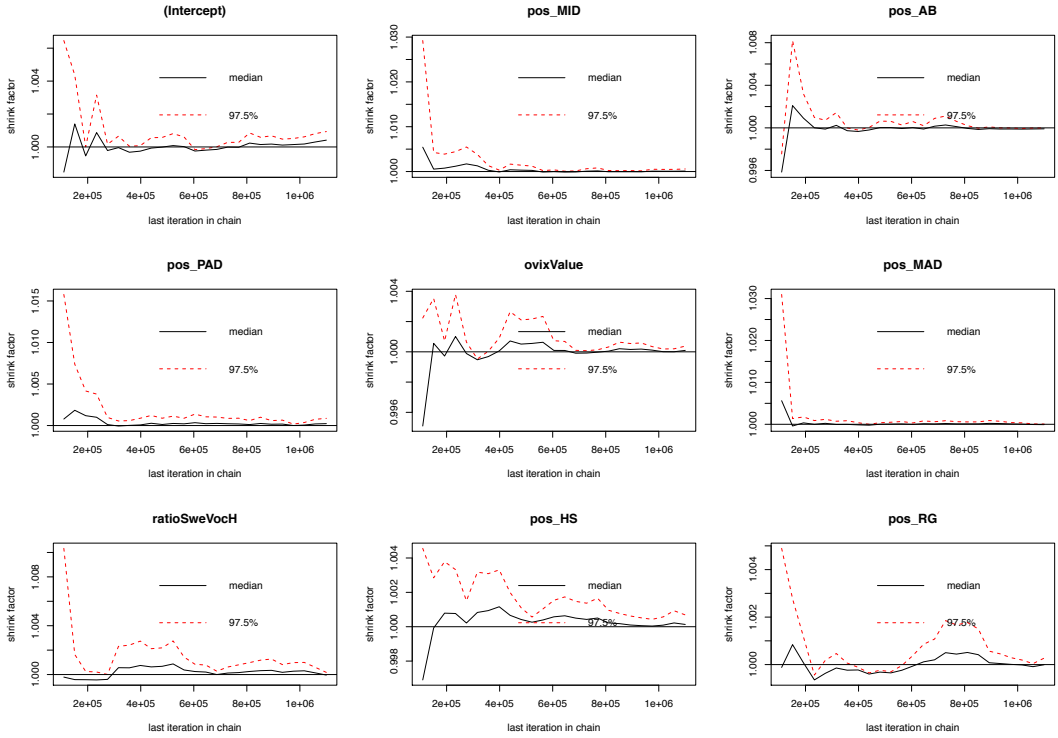

Figure 1: Gelman plots for the 8 feature coefficients and the intercept.

After illustrating that all the chains converge to the true posterior, we simulate a single longer chain, keeping a total of 100 000 draws. We compute the maximum inefficiency factor (IF) for the sample. The IF is a measure of the loss of efficiency from using a dependent rather than independent sample and is calculated for each feature (Heckman and Leamer, 2001). In this case, the maximum IF is $\sim$1.02 which is acceptable.

2.6 Prediction

The fitting approach we use model the class label only as a function of the latent variable and fits the latent variable to the data. This means that besides traditional measures of classifier performance we can also inspect this latent readability value for each text.

For instance, we can plot the distributions for all true positives and all false positives and see whether true positives are generally more readable, according to the latent variable, than false positives (see Figure 3). This latent variable can also be viewed as a measure of classifier confidence and can, by definition, be converted into a probability for use in ensemble methods.

2.6.1 Posterior predictive distribution

Going one step further we should realize that we do not have a single instance of the Probit model. We do not, as is usual in Frequentist statistics, have only a Maximum Likelihood Estimate (MLE) of the "best fitting" instance.

Rather, we have 100 000 draws from the posterior distribution of the model. That is, effectively we have 100 000 instances of the Probit model. Each of these instances can be used to predict a value for a new observation, giving us 100 000 values for this observation. These values constitute a sample from the posterior predictive distribution, or PPD, for that observation.

This PPD can be used to calculate, for instance, the probability that text T_1 is easier to read than text T_2 according to the model, by simply calculating what proportion of the draws that result in $Y^*_{T_1} < Y^*_{T_2}$.

We can also inspect the posterior predictive distribution directly. The variance of which can be viewed as another measure of confidence, but rather our confidence in the model itself than the model's confidence when making a specific prediction.

2.7 Evaluation

We present results from two evaluations, one classification task using the SUC/LäSBarT dataset and one ranking evaluation on the smaller MASTER data set.

2.7.1 Classification

While this paper focuses on ranking, it should be noted that what we are actually modelling is a classifier. Thus, we should say something about how the PPD is used for classification. To be clear, within the Bayesian paradigm, the full PPD is the answer to the question "What is the degree of readability of text T?", however, we can do a lot with the PPD. Firstly, we can transform it into a probability of belonging to either class by simply computing the ratio of draws generating each class label.

A more refined way of assigning actual class labels utilizes Bayesian decision theory, which, in short, combines the PPD with a loss function. This loss function can take many forms but can be described as a way to quantify the cost of making different erroneous decisions. The task is then to find and make those decisions which minimize the expected loss.

In our case, the decision to make is what label or probability to apply to a new observation. However, as classification is not the main focus of this paper we will simply use the sign of the PPD mean value as decision rule. This also makes comparisons to the SVM reference classifier easy.

2.7.2 Latent variable ranking

To evaluate the model for ranking we use the small sets of ordered texts from the MASTER project (Kanebrant et al., 2015) covered in Section 2.1.

We compute PPD of the latent variable readability for each text and the corresponding ranking of the texts. Using these distributions and the reference readability, as defined in Section 2.1, we can compute the posterior distribution of regular Pearson correlation as well as Kendall rank correlation, also known as Kendall's τ. Kendall's τ is a correlation metric used for comparing different total orders (or with some tweaking, partial orders as well) on the same set (Kendall, 1948).

Support Vector Machine		
	Easy	Non-easy
Easy	341	17
Non-easy	9	333

Probit model, posterior mean		
	Easy	Non-easy
Easy	340	14
Non-easy	10	336

Table 1: Confusion matrices for the two classifiers.

3 Results

We present three results. First we present the estimated posterior of the Probit model, then we compare its ability to classify texts to a more conventional SVM model, and finally we give preliminary results on the Probit model's ability to also rank texts.

3.1 Fitted model

For completeness we plot the posterior densities for all coefficients in Figure 2.

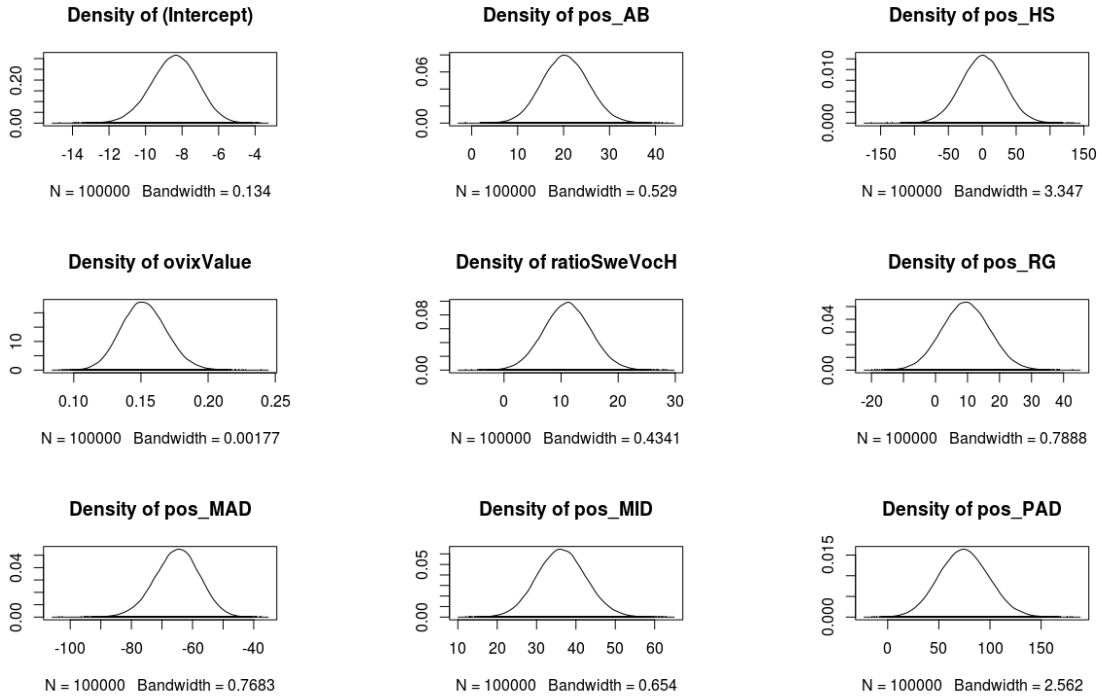

Figure 2: The posterior densities for all coefficients.

3.2 Classifier performance

Table 1 depicts the confusion matrices for the two classifiers. As can be seen the two models are comparatively good as classifiers with roughly 96.3% accuracy for the SVM and 96.6% accuracy for Probit.

Figure 3 shows kernel density estimates of the posterior predictive distributions for all texts, separated into the four false/true positive/negative categories. As we can see, the distributions for incorrectly classified texts lie closer to 0 for both false positives and false negatives than for correct classifications. Generally this can be interpreted to mean that the confidence of the classifier is lower for these examples or that their estimated degree of readability is less extreme.

However, it is also illuminating to look at a the PPDs of the first correctly and the first incorrectly classified texts.

In the first plot in Figure 4 we can see the posterior distribution of readability for a correctly classified easy-to-read text. This distribution is centred far from 0 implying a high degree of class confidence by

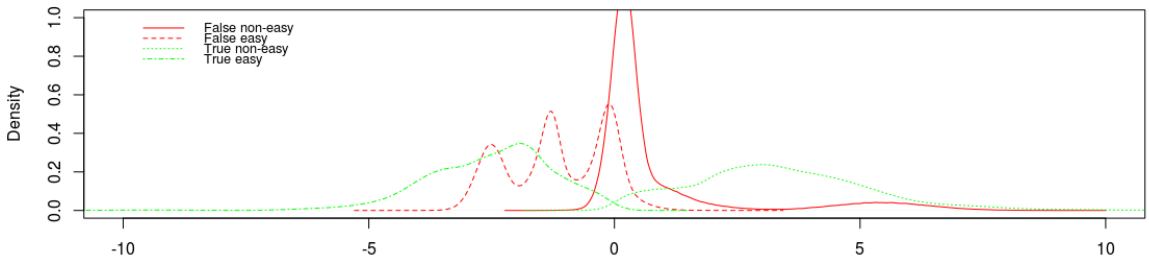

Figure 3: Total densities for false labels and true labels. Negative values correspond to easy to read texts and positive are non-easy.

Figure 4: The posterior predictive distribution of readability values for one correctly and one erroneously classified easy-to-read text.

the model as well as a low degree of difficulty, i.e. a high degree of readability.

If we instead look at the second plot in Figure 4 we see the posterior distribution of readability for an incorrectly classified easy-to-read text. This text belongs to the easy-to-read corpus but has a posterior mean on the non-easy-to-read side of 0, i.e. it is classified as non-easy-to-read. We can see that the posterior mean lies relatively close to zero and a not insignificant part of the distribution actually lies on the correct side of 0, note also that the x-axis scales differ between the two plots, the variance of the first, ”correct”, PPD is smaller than that of the second, ”erroneous”, PPD.

3.3 Latent variable ranking

Figure 5 presents the posterior distributions of the Pearson ρ and Kendall τ correlations. The bars indicating posterior mass of Kendall τ are scaled to the height of the graphs. A large τ indicate strong rank order correlation and a large ρ indicate a strong linear correlation.

For both types of correlation, the distributions vary among different test sets but the majority of the sets show a definite correlation; general fiction for grades 4 and 6, social science for grade 4 and natural science for grades 6 and 8. Some sets show a strong correlation, social science grade 8 and natural science grade 4. However, social science grade 6 and general fiction grade 8 show basically no correlation at all. It should be restated that the test sets are small and preliminary, and not investigated for any specific problems with the data, but even so, we view the fact that 5 out of the 9 test sets show strong correlations with our predicted rankings, and another 2 shows some correlation, as promising.

4 Conclusions

We have presented the Probit model, its latent variable interpretation, a Bayesian approach to fitting a Probit model to data based on this interpretation, as well as interpretations of other aspects of the model. The Bayesian approach gives us posterior distributions over all coefficients and predicted values. These express the uncertainty of the model and, though outside the scope of this paper, lend themselves to advanced approaches to decision making.

To further assess the model we also compared its ability to classify texts to a state of the art SVM classification model. We show that for classification the Probit model performs more or less on par with

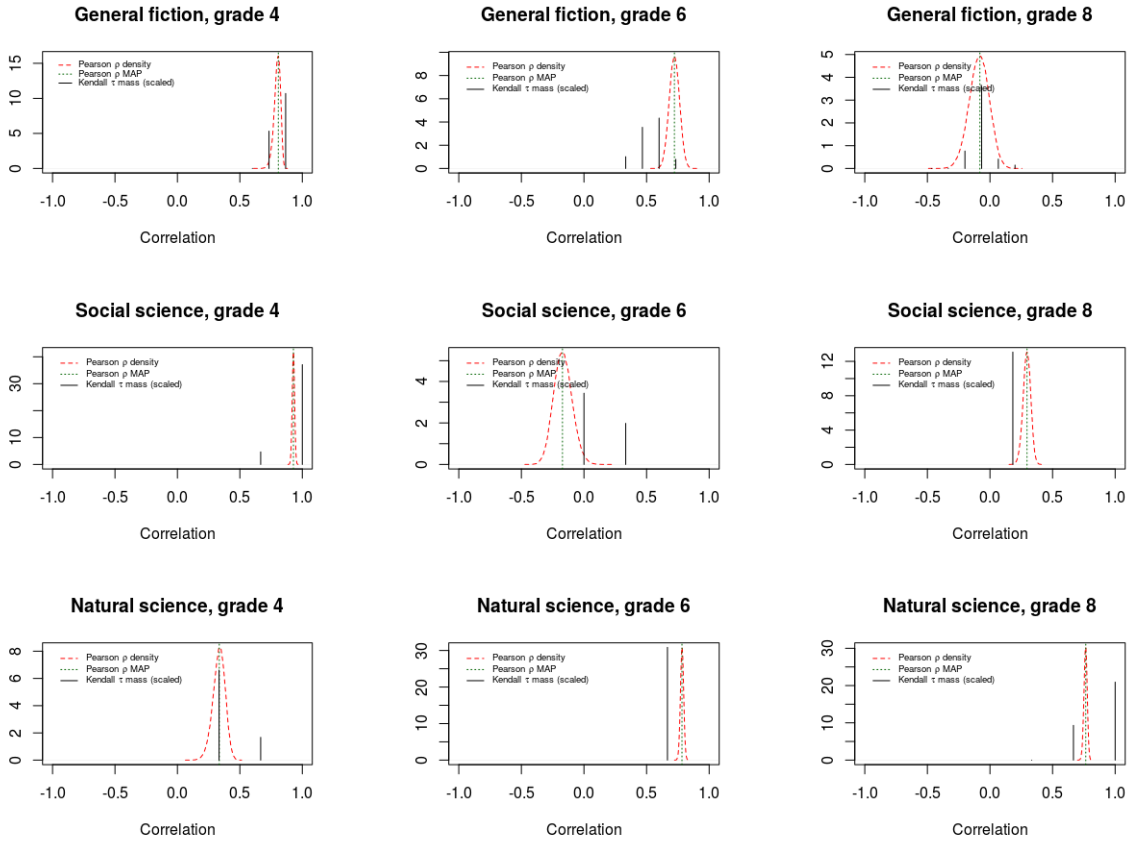

Figure 5: Posterior density of Pearson ρ and scaled mass of Kendall τ correlations.

a Support Vector Machine model even when using a feature set developed for SVM.

We also have the predicted values of the latent variable, readability, which we use for readability ranking; one of the main reasons for using the Probit model. The results from the evaluation of readability are promising and encourage further research.

Future work will focus on developing a fitting algorithm allowing us to utilize mix of binary classified data and a small amount of ranked data to train a hybrid of the Probit model presented in this paper and the Ordered Probit model (Becker and Kennedy, 2010). Hopefully by then a less preliminary data set of ranked documents from the MASTER project might be available for evaluation.

Acknowledgements

This research was financed by VINNOVA, Sweden's innovation agency, and The Knowledge Foundation in Sweden.

References

James H. Albert and Siddhartha Chib. 1993. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679.

William E. Becker and Peter E. Kennedy. 2010. A graphical exposition of the ordered probit. *Econometric Theory*, 8(1):127–131, 10.

C. I. Bliss. 1934. The method of probits. *Science*, 79(2037):38–39.

Stephen P. Brooks and Andrew Gelman. 1998. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

Eva Ejerhed, Gunnel Källgren, and Benny Brodda. 2006. Stockholm Umeå Corpus version 2.0.

Johan Falkenjack and Arne Jönsson. 2014. Classifying easy-to-read texts without parsing. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), Gothenburg, Sweden*.

Johan Falkenjack, Katarina Heimann Mühlenbock, and Arne Jönsson. 2013. Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa-2013), Oslo, Norway*, NEALT Proceedings Series 16.

Lijun Feng. 2010. *Automatic Readability Assessment*. Ph.D. thesis, City University of New York.

John Henry Gaddum. 1933. Methods of biological assay depending on a quantal response. In *Reports on Biological Standard III.*, number 183 in Special Report Series of the Medical Research Council. Medical Research Council.

Andrew Gelman and Donald B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

James Heckman and Edward Leamer, editors. 2001. *Handbook of Econometrics*, volume 5. Elsevier, 1 edition.

Erik Kanebrant, Katarina Heimann Mühlenbock, Sofie Johansson Kokkinakis, Arne Jönsson, Caroline Liberg, Åsa af Geijerstam, Jenny Wiksten Folkeryd, and Johan Falkenjack. 2015. T-master – a tool for assessing students' reading abilitiesstudies on automatic assessment of students' reading ability. In *Proceedings of the 7th International Conference on Computer Supported Education (CSEDU 2015), Lisbon, Portugal*.

Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. 2004. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

Maurice G. Kendall. 1948. *Rank correlation methods*. Griffin, London.

Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park. 2011. MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, 42(9):22.

Pascual Martinez-Gómez and Akiko Aizawa. 2013. Diagnosing Causes of Reading Difficulty using Bayesian Networks. In *Proceedings of IJCNLP 2013*, pages 1383–1391, October.

Robert E. McCulloch, Nicholas G. Polson, and Peter E. Rossi. 2000. A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99(1):173–193.

Katarina Mühlenbock. 2008. Readable, Legible or Plain Words – Presentation of an easy-to-read Swedish corpus. In Anju Saxena and Åke Viberg, editors, *Multilingualism: Proceedings of the 23rd Scandinavian Conference of Linguistics*, volume 8 of *Acta Universitatis Upsaliensis*, pages 327–329, Uppsala, Sweden. Acta Universitatis Upsaliensis.

Sarah Petersen. 2007. *Natural language processing tools for reading level assessment and text simplification for bilingual education*. Ph.D. thesis, University of Washington, Seattle, WA.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: a unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08).*

John C. Platt. 1998. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical Report MSR-TR-98-14, Microsoft Research, April.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. 2006. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. 2010. Sorting Texts by Readability. *Computational Linguistics*, 36(2):203–227.