Identifying book reviews in Swedish newspapers

Niklas Zechner
Språkbanken
University of Gothenburg
niklas.zechner@gu.se

Abstract

Several classification algorithms are applied to the problem of identifying book reviews in Swedish newspapers from 1906. Word frequency methods perform well compared to a large language model, even on short texts, especially with large numbers of frequencies.

1 Background

With the rise of rapid digitisation methods, historical newpapers have become a treasure trove of information. Centuries of news writing gives us a new window into the lives and thoughts of the people of bygone ages. With such a vast amount of data, it can be very difficult to identify specific information. Some topics can be found with a keyword search, but others are less simple.

In this study, we focus our attention on spotting book reviews, in particular reviews of fiction. In previous work, we have used the resulting data to look at those reviews and their content from a more humanist perspective. (Ingvarsson et al., 2022)

We use data from 1906, one of the latest years for which newspapers are available without copyright restrictions, but with an eye to future applications on later texts as well. When the data is still in copyright, we may be faced with the perplexing problem of needing to analyse text without having access to the text. In some cases, we may have unlimited access to word frequencies, but limited access to the complete text, which could mean that we need a method which uses only word frequencies.

In recent years, deep learning models such as BERT have become the standard for many language processing tasks, including classification. They have overall been found to be effective, but come with a few downsides. One is that constructing the model itself takes a lot of processing time and requires very large amounts of data, making it unfeasible for smaller languages. This is perhaps less relevant in this case; smaller languages tend not to have centuries of newspaper data, and for Swedish there are well-performing BERT models (Malmsten et al., 2020). A second issue is, as mentioned, copyright might limit access to the full text, and BERT models are not applicable to word frequency data. A third issue is that the method is not transparent; we cannot see what the decision is based on, as is possible with many earlier models. Transparency can help us further develop the methods, or find flaws in them, but is also of particular interest here, as we pass on the results to humanities researchers: What is it about a text that makes it look like a review, and what are the differences between different categories of reviews?

2 Experiments

We will test a few approaches to classification on reviews in newspaper text, and some parameters that might affect it. First, we test different classification algorithms – several varieties of a method based on word frequencies, along with a BERT model. Our primary goal is identifying reviews of fiction literature, compared to general newspaper text. As a secondary goal, we also try distinguishing between reviews based on what is being reviewed – fiction literature, non-fiction literature, or non-literature. Our non-review data comes from several different newspapers, so for comparison, we do another test, classifying each article based on which publication it comes from.

Second, we want to see how the accuracy of the frequency-based method depends on how many different features – here, word frequencies – we analyse. Fewer features means a faster method, but more features may give higher accuracy.

Third, we are interested in how the length of the analysed text affects the results. Previous studies (Zechner, 2017) have shown that this is often the biggest factor for successful classification.

3 Data

The National Library of Sweden (Kungliga Biblioteket) has a growing collection of digitised newspapers going back to 1645 (Börjeson et al., 2023). From these, annotators extracted the texts of all reviews written in nine newspapers written in 1906. Two of the papers were removed from the data due to digitalisation errors, leaving seven: *Arbetet, Dagens Nyheter, Göteborgs Handels- och Sjöfartstidning, Göteborgsposten, Sydsvenskan, Socialdemokraten, Upsala Nya Tidning*. The data had been separated into arbitrary chunks by the digitisation process, so we combined them, tracking each separate review as a single text. The annotators also marked each review as one of fiction literature, non-fiction literature, or other (such as reviews of theatrical performances).

For comparison, we needed texts which are not reviews. Annotator time being a precious resource, we opted to simply extract the longest available text chunks. This naturally introduces a bias, as some types of text are likely to have larger chunks. It also means that the chunk might not include a whole article, but it should be a large part of one, and only one. In the future, we can expect article segmentation to improve, so for now, we are to some extent simulating the problem of distinguishing between reviews and non-review articles.

The data used in this study consists of all the reviews found by the annotators, namely 248 reviews of fiction literature (FIC), 52 reviews of non-fiction literature (NONF), and 505 other reviews (OTHER), along with 100 non-review article texts from each of the seven publications (ART).

4 Method

4.1 Frequency method

We calculate the average and standard deviation for the 1000 most common tokens (i.e. words and punctuation marks) in the full set of data. We then create a frequency profile for each individual text, containing the frequencies of those 1000 tokens.

For each test, a subset of the relevant profiles are chosen as training data. Each profile in the test set is compared to the training set to find the closest matching class. A comparison was made between nearest-neighbour and centroid methods; the results of the former were unremarkable and are left out here for brevity. For the centroid method, the profiles in the training set are combined into one, so we can then find, for each profile in the test set, the most similar class in the training set. Combining the frequencies can be done either unweighted (the per-profile average) or weighted (the overall frequency), letting the larger texts have a greater impact. Since the texts are relatively close in size, we would not expect any large difference. Preliminary tests confirm this, and the results presented here are based on unweighted frequencies.

4.2 Distance measures

To measure which training profile is the most similar, we try several algorithms, which we can think of as distance measures in the vector space defined by the frequencies. Using example vectors (a, b, c) and (A, B, C), these are:

Manhattan distance (MAN),

$$|A-a|+|B-b|+|C-c|$$

Cartesian distance (CART),

$$\sqrt{(A-a)^2+(B-b)^2+(C-c)^2}$$

Negated scalar product (DOT),

$$-(Aa+Bb+Cc)$$

Cosine distance (COS),

$$1-\frac{Aa+Bb+Cc}{\sqrt{A^2+B^2+C^2}\sqrt{a^2+b^2+c^2}}$$

4.3 Normalisation

There are many ways to “normalise” a vector set. We try two of them here. First, we can choose to *relativise* (R) the vectors, that is, subtract for each frequency the average, so that the vectors are relative to the origin point formed by the overall average text. Second, we can choose to *equalise* (E) the frequency values, by dividing each one by the overall standard deviation, so that each frequency has equal impact on the result. Since the first two distance measures are independent of the origin point, we do not apply any normalisation to them. This leaves ten different distance measures, with normalisations included.

For the BERT method, we apply KB-BERT (Malmsten et al., 2020) with 768 dimensions on each of the texts. The output is already normalised in such a way that each vector has length 1, so we use cosine distance in all cases.

	base	MAN	CART	DOT	DOTE	DOTR	DOTRE	COS	COSE	COSR	COSRE	BERT
FIC/ART (balanced)	50	92	85	58	88	85	90	88	95	85	92	91
FIC/NONF/OTHER	33	83	67	60	83	63	83	68	87	69	87	83
publication	14	49	35	17	70	29	74	37	74	39	76	38

Table 1: Results for the three accuracy tests, in percent.

4.4 Evaluation

For each of these experiments, we do ten-fold cross validation; that is, use a tenth of the data as test set, repeat for each tenth, and combine the results.

For the first experiment (FIC/ART), we present the average per-class accuracy, since the ART class is much bigger than the FIC class and we want both to count. For the second experiment (FIC/NONF/OTHER), the reasoning is the opposite: the NONF class is much smaller and could skew the results, so we use the overall accuracy instead of the per-class accuracy. For the third experiment, the classes are all the same size.

5 Results

Table 1 shows the results of our comparison of classifiers. First, we compare all available fiction reviews (FIC) with all available articles, that is, general newspaper text (ART). Second, we compare the three different types of reviews. We compare all available reviews, and attempt to classify them as one of the three types. Third, we classify the articles based on which publication it comes from.

We see in Table 1 that cosine distance with equalised values is overall the most successful. Additional experiments not listed here confirm this. The BERT model is slightly behind on the first two experiments, and more so on the third. Since the two most popular methods – COSRE and BERT – are nearly the same for the main experiment, we forgo testing for statistical significance.

The table also lists the baseline accuracy for each experiment – what we would get with random guessing. In the first two cases, the classes are of different size, so we could pick the most common class as a baseline, but since the methods do not use that information, and the disparity is not inherent in the problem, this should not be relevant.

Figure 1 shows how the accuracy of the three different experiments varies with the number of features. This is using the frequency method, with relative and equalised cosine distance (COSRE). We see that the accuracy increases significantly up to a relatively high number of features, although about 100 seems to be enough for the first experiment.

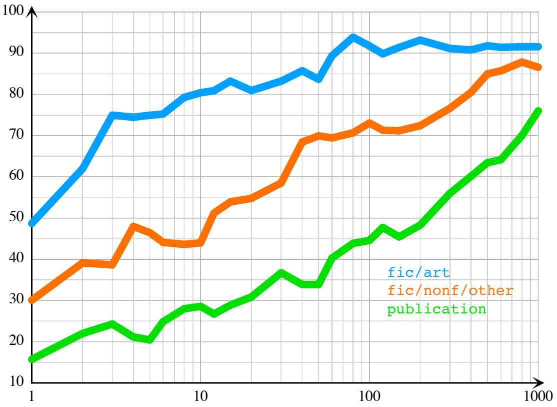

Figure 1: Accuracy, in percent, for each of the three experiments, as a function of the number of features.

To measure the effect of text length on classification, we would normally use only part of each text, and vary the included length. But since these texts are quite short, that might not be feasible. Instead, we look at which texts have been correctly classified, and see if there is a difference. Are longer texts more likely to be correctly classified? Figure 2 shows the distribution of text lengths and the correct/incorrect classifications for each of the three experiments. We see that the effect of length is not overwhelming. For another perspective, we can measure the ratio of average lengths (geometric mean) for incorrectly vs. correctly classified texts. For each experiment, they are: 1.10, 1.42, 0.99.

6 Conclusion

6.1 What can we identify?

Is automatic identification of reviews feasible? With an accuracy in the vicinity of 90%, the method is performing well above random. The problem is that newspapers like these contain far more non-reviews than reviews, so if we were to classify all the text chunks in the original data, most of the ones marked as reviews would still be non-reviews.

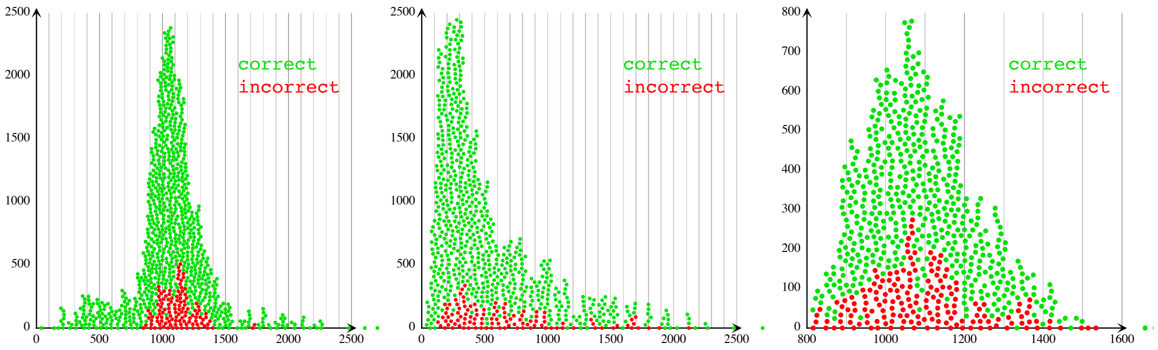

Figure 2: Distributions of lengths of texts for the three experiments, using the COSRE method. The x axis shows the number of words. Green points are texts correctly identified. Some outliers are not visible.

We notice in Table 1 that the second experiment gets surprisingly good results. It could be expected that distinguishing reviews of fiction literature from articles should be a much easier task than distinguishing them from other reviews – and furthermore, the second experiment has three classes rather than two.

The third experiment is also surprisingly effective. Even though the newspapers are written by many different people, we can identify them reasonably well. Without further analysing what the distinguishing features are, we might suspect that it has to do with place names, since many of these newspapers focus on different geographical areas. There also seems to be a difference in OCR errors, which could make the results overly optimistic.

If we had access to accurately segmented articles and reviews, would that significantly improve the classification? Initially, the obvious guess would have been yes – many previous studies have strongly suggested that the size of the data means everything. Here, we see surprisingly little evidence of that. The difference in accuracy between longer and shorter texts is measurable, but small. Still, these texts are relatively uniform in length. We have also chosen the longest text chunks for our non-review data, so if we were to apply the methods to all the chunks available, we should expect much lower accuracy – some chunks are only a headline or a few words. If we could apply them to fully segmented articles, the accuracy might improve a little, but we should not expect miracles.

6.2 Which method should we use?

As Table 1 shows, the cosine methods with equalised features performs well. Perhaps more surprising is that the non-relative version (COSE) performs on par with the relative version (COSRE), which is probably the most widely used. It seems odd that the absolute values should be effective, and this may well be due to chance.

In the first two experiments, the BERT method trails slightly behind the best frequency methods, which might also be no more than a random fluctuation, but it falls short more noticeably on the third experiment. Perhaps here the difference is more in choice of words than in content or sentence structure. It is overall surprising that we can reach such high accuracies on this seemingly much more difficult task. The average length of each text is slightly larger here, but not by much.

6.3 How many features do we need?

As we see in Figure 1, the accuracy for the first experiment seems to level out around 100 features – looking at the 100 most common words is enough to get the best possible accuracy. For the second experiment, we might be nearing the maximum with 1000, whereas with the third, the accuracy is still rising sharply. It is certainly expected that identifying publications is harder, but it looks like we can solve that with more features.

6.4 How long texts do we need?

Previous studies (Zechner, 2017) have strongly suggested that the length of a text is crucial for classification, but here we see a remarkable accuracy even for a few hundred words. The second experiment shows a noticeable difference in length between the correctly and incorrectly classified texts, but the other two show little difference.

References

Love Börjeson, Chris Haffenden, Martin Malmsten, Fredrik Klingwall, Emma Rende, Robin Kurtz, Faton Rekathati, Hillevi Hägglöf, and Justyna Sikora. 2023. Transfiguring the library as digital research infrastructure: Making KBLab at the national library of Sweden.

Jonas Ingvarsson, Daniel Brodén, Lina Samuelsson, Victor Wåhlstrand Skärström, and Niklas Zechner. 2022. The new order of criticism. Explorations of book reviews between the interpretative and algorithmic. In *The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022*. CEUR Workshop Proceedings.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of Sweden - making a Swedish BERT. *ArXiv*, abs/2007.01658.

Niklas Zechner. 2017. *A novel approach to text classification*. Ph.D. thesis, Umeå University.