Effects of Division of Data in Author Identification

Niklas Zechner

Department of Computing Science,
Umeå University
zechner@cs.umu.se

Abstract

When developing or evaluating a system for automatic author identification, it is usually necessary to divide the available text for each author in two parts, to play the roles of ”known” and ”unknown” data. This division step is often not given much attention. But as we will show, there are several ways to divide a text, which may lead to very different results. We work on a corpus of forum posts, and divide the posts of each user either chronologically in the middle, or by taking alternating posts. The alternating method gives a significantly higher accuracy. We argue that this is misleading, and the middle division is preferred.

1. Introduction

Author identification is one of the most common text classification tasks. It is a diverse area, in which there is little consensus on the best methods (Rudman, 1998). There are a few different variants, all of which amount to somehow finding a likely author of an unknown text or set of texts. We may have a supervised classification problem, where the classifier is given a set of candidate authors, and picks the most likely one. We may have a clustering problem, where a set of texts are grouped by which ones are likely to be written by the same author. And we may have a similarity problem, where texts are pairwise compared and given a similarity score, which can be interpreted as the likelihood that they are written by the same author. (Stamatatos, 2009)

We will use a technique based on similarity, but which we easily adapt to also work with supervised classification: Once we have a similarity measure, it is easy to take an unknown text, compare it with all known authors, and see which one is the most similar, and thus the most likely candidate.

Any text classification, in the algorithmic sense, usually starts with extracting some numerical features from the text - how long it is, how often certain words are used, how long the average word is, and so on. Many studies use an extensive list of features (Narayanan et al., 2012). With these, we can make a statistical model, and compare different texts. There are many possible algorithms that can be used to classify or compare texts based on those features, but we will only look at one simple algorithm in this study.

2. Method

The data used in this study comes from the ICWSM boards.ie corpus, a collection of posts to a webboard, from which we have chosen two years of data. We use a very simple set of features: We count the frequencies of the n most common words, where we define ”common” as the most frequent in the corpus as a whole, and ”word” as any token, including punctuation. For example, the five most common words are period, ”the”, comma, ”to”, ”a”. For each feature, the distribution is standardised, that is, the values are translated and scaled so that the mean is 0 and the standard deviation is 1. Then we are able to compare two feature value arrays by cosine similarity.

For each author, we list the posts chronologically, and divide them into two sets, a and b, by two different methods. The first method is by taking alternating posts - the first post goes in a, the second in b, the third in a, and so on. The second method is by splitting the list of posts in the middle. We call these methods alt and mid.

To evaluate this similarity measure, we use it for what is effectively supervised classification. For each a-part, we go through all of the b-parts. We calculate the similarity measure, and keep track of which b-part has the highest similarity. If that is the one which actually comes from the same author as the current a-part, it is considered a success, otherwise a failure. After going through all the a-parts, we can see what fraction of them have been successful; we consider that the accuracy of the method.

We could look at the difference between divisions by just running a single test on each and seeing the difference in accuracy, but we opt to go into more detail. First, we vary the amount of data given to the algorithm. We expect that the accuracy will increase with more data, and so this way we can compare the effects of different divisions to those of differences in amount of data. Second, we make the same comparison while varying the number of features, that is, in this case, the number of word frequencies counted.

3. Theory

One problem in author identification is avoiding influence from other differences between texts. When we try to identify an unknown text, we want the classifier to see only similarities and differences that are typical of authors, not of topics or contexts. If we evaluate a system on a corpus where many authors have written about the their own preferred topics, and those topics show up in both sides of the classification (training set and test set, or in this case, the a and b parts) we will get a high accuracy, since the system is effectively classifying topic along with author. If we then use the system on texts where an author does not stay on the same topic, we may get different results - for supervised classification, that could mean lower accuracy, and for applications where we only want to know if two texts are by the same author, it could mean more false negatives.

It seems likely that forum posts which are close in time are often on the same topic, so that choosing the alt method will lead to more posts on the same topic ending up in the same half, inflating the apparent accuracy. Our two methods can be seen as two extremes on a scale, with the alt division giving the highest possible similarity, and the mid division the lowest. Some studies divide the data randomly, whether by choice or because the corpus is scrambled for copyright or privacy reasons. We suspect that a random division should lead to a similarity somewhere between the alt and mid methods, but we have not tested that here. Where a real application would end up is an open question; if the unknown text is really indistinguishable from known texts by the same author, we might get a better accuracy than the mid division suggests, but in most cases, we can reasonably expect another text by the same author to have some systematic differences, so that even the mid division gives an overestimation.

4. Results

Figure 1 shows how the accuracy varies with the amount of data. We can clearly see that the alt division is considerably higher than the mid division, for both small and large amounts of data.

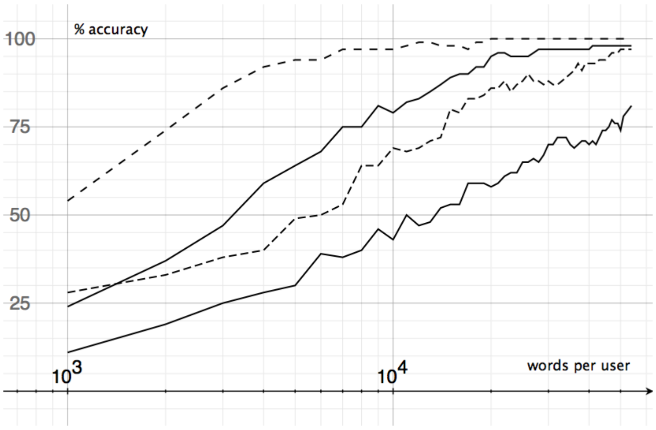

Figure 1: Accuracy as a function of amount of data (in each half). The dashed line is alt, the solid is mid. The upper two are for 100 features, the lower for 10. This is for 100 candidate authors.

Figure 1 compares two different feature sets, a smaller where we only count 10 word frequencies, and a larger where we count 100. The difference seems to apply to both, but to make sure, we make another test, varying the number of features while keeping the amount of data constant. In figure 2 we can see that they are indeed consistently different; the alt division gives considerably higher values regardless of not only amount of data but also number of features. Similar tests on other features, including syntactical measurements, are not shown here, but show similar results.

5. Conclusions

We have seen that the effect of changing how a corpus is divided can be quite remarkable, at times making the difference between a near-perfect method and a moderately reliable one, and yet this difference is often ignored. It is certainly worth taking into account when creating, evaluating, and demonstrating an author identification algorithm. The same may apply to other classification problems.

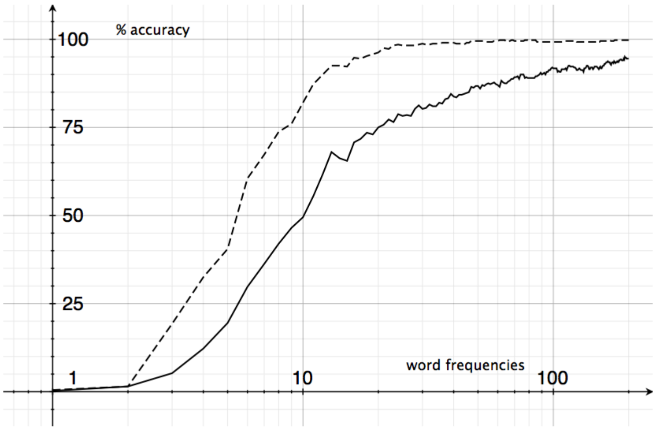

Figure 2: Accuracy as a function of number of features. The dashed line is alt, the solid is mid. This is for 400 candidate authors, with 22 000 words in each half.

There are still parameters that we have not considered. We have used only one classification algorithm, and only one corpus. Further research is needed to determine to what extent the results are universal, but there is no obvious reason why this corpus or this classifier should be a special case.

Our results also show that measurements of accuracy for this kind of algorithms might not be reliable. Since there is such a large variation just from changing the division, there is no way of knowing what the accuracy would be when applied to a real life problem. This stresses the need for more advanced corpuses, with large amounts of text by authors writing in different situations, on different topics. Whether our theory about the cause of these differences is correct is another matter. We can see a clear difference even when using only the ten most common words, none of which appear to be particularly topic-dependent, which is quite surprising, and there might be another explanation altogether. Either way, the difference is there, and needs to be taken into account.

The fact that the variation is so large even for those few common words could also have another far-reaching consequence. Unless there is some other reason for the difference in performance, we have to assume that these words are topic-dependent after all. This means that the choice of certain features on the basis that they are assumed to be topic independent has to to be called into question.

Based on our theoretical reasoning, we find that choosing the alt division, or a random division, could make the results misleading, so that too much trust is put into an applied system, or too much importance is placed on an academic study. Therefore we suggest that the mid division is the more appropriate choice for most situations, if one does not want to redo our tests with every new study. This is not the kind of world where methods often work better in practice than in theory.

References

Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Eui Chul Richard Shin, and Dawn Song. 2012. On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy (SP)*, pages 300–314. IEEE, May.

J Rudman. 1998. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31:351–365.

E. Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60.