Using Text Classification to Detect Alcohol Intoxication in Speech

Andreas Jauch[1,2], Paul Jaehne[1,2], and David Suendermann[2]

[1] IBM, Böblingen, Germany
[2] DHBW, Stuttgart, Germany

andreas.jauch@de.ibm.com paul.jaehne@de.ibm.com david@suendermann.com

Abstract. This paper focuses on text-based classification of the Munich Alcohol Language Corpus (ALC) which contains speech from persons in intoxicated as well as in sober state. In order to classify whether a person is intoxicated or not, several combinations of classifiers and feature extraction approaches have been examined. One major finding was that the expressiveness of a test was tightly coupled to its type of speech and topic. The best result was achieved by classifying picture description tests using logistic regression which resulted in an unweighted average recall of 89.4%.

1 Introduction

One of the most severe problems in traffic is the abuse of alcohol. In the United States, every day about 30 people die in crashes involving alcohol-impaired drivers totaling more than 50 billion US$ annual cost [1]. Therefore, measuring intoxication by interviewing drivers and analyzing their answers is an encouraging topic of current research activity. The Ludwig Maximilians University of Munich put together a foundation for work related to the detection of alcohol intoxication by producing a publicly available speech corpus. The *Alcohol Language Corpus* (ALC) [7, 8] contains speech recordings and their transcriptions in intoxicated as well as in sober state. To inspire research teams to work on classifying intoxication state on this corpus, the *Interspeech Speaker State Challenge 2011* has been brought up to serve as a stage for competitions [9]. Hence, there already exist a number of publications covering this topic, though to the best of our knowledge all of them—including the intoxication subchallenge winners from Interspeech [3]—used audio-based features either on its own or in combination with others to perform classification. Solely the team from the University of Erlangen [2] measured accuracy[3] of a text-only based system, but later on, they combined it with different kinds of features to improve their results. Furthermore, they excluded textual features from their major final result

[3] text-only based results are provided on their development set only with an unweighted average recall of 59,1% [2]

system because they decreased accuracy. Encouraged by the fact that no text-only focused publications on this challenge exist, we hereby present our results on classification using textual features only.

2 System Description

2.1 Basic Setup

Although the *ALC* was originally created as a speech corpus, it is exhaustively transcribed allowing for easy feature extraction from these transcripts[4]. Using the WEKA toolkit [6], bag-of-word feature vectors in form of word presence or word count vectors were generated on the transcribed speech. Additionally, information on speech irregularities like stutters, repetitions or noticeable pauses was added to the feature set. Apart from this, no other information provided by the corpus—like audio data or phonetic information—was used for feature generation.

In the beginning, the tests were executed using multilayer perception neural networks, decision tables, J48, JRip, naïve Bayes, logistic regression and SMO as classifiers. However, since the latter three produced considerably better results than the others, the final experiments—and such all of those presented in this paper—were only performed on those three. All experiments have been executed using 10-fold cross-validation. To be able to compare results with other publications on the same matter, we used unweighted average recall (UAR) as performance metric as calculated by

$$\mathrm{UAR} = \frac{\mathrm{recall(alc)} + \mathrm{recall(nonalc)}}{2} = \frac{\frac{tp}{tp+fp} + \frac{tn}{fn+tn}}{2} \tag{1}$$

We are making the code written to conduct these experiments available to the public in the form of an open-source GIT repository on

http://suendermann.com/corpus/alc.html

and researchers are encouraged to live up to these results.

2.2 Enhancements

Tests were done in several iterations, where the goal of each iteration was to top the resulting accuracies of the previous tests.

In the first iteration, feature selection using information gain was applied to discard features whose contribution of information is lower than a certain threshold. Information gain is computed by evaluating the differences in entropy [5] with or without the knowledge of that feature. In order to optimize the final

[4] The transcripts, which build the basis for this paper, where generated by manual annotation.

set of features, the ideal threshold needed to be found, which was achieved by running a series of test runs using different thresholds.

To profit from the different approaches of the used classifiers, SMO, logistic regression and naïve Bayes were combined into a majority voting system that predicts the resulting class by way of majority vote. As the corpus contains speech from different topics[5] ranging from simple tasks as reading a telephone number, over tongue twisters to picture description, the corpus was further divided into its 11 document classes to asses differences in their expressiveness. Since tests were executed on both the entire corpus and all individual document class sub-corpora in the same manner, the tests can be easily compared.

3 Experiments

3.1 Corpus Description

The ALC provides German speech of 77 female and 85 male speakers, recorded both sober and intoxicated. Its vocabulary size is 15776 words, where all the different dialectical forms of one word are counted separately. The ALC focuses not only on read speech, but it also contains a variety of different spontaneous speech samples. In addition to that, the corpus also contains tests on command and control speech for its applicability in an automotive environment. Altogether, there are 11 different document classes that can be divided into spontaneous and non-spontaneous speech as shown in Table 1.

When running experiments using feature selection on the corpus, some words resulted in a surprisingly high information gain in contrast to the rest of the corpus. This turned out to be due to the fact that there are some tests not available in both states—e.g., one tongue twister only appears in intoxicated state. Of course, this would give a text classifier considerable advantage. Thus, we removed all those tests not available in both states from the corpus corresponding to a decrease in size by about one third. Still containing 4698 intoxicated samples and 4978 sober ones, this now modified corpus is almost equally balanced with a distribution of 48.6% to 51.4%. Hence, WEKA's default classifier ZeroR always picking the most frequent class produced a baseline UAR of 51.4%. The modified ALC corpus has a total of 11386 words vocabulary.

[5] in the following referred to as *document classes*

Table 1. description of sub-corpora

Doc Class	Description	Speech Type	#Types	#Samples	%Samples ALC	%Samples NonALC
LN	list numbers	read	264	1660	48.80%	51.20%
LT	list tongue twister	read	174	344	47.09%	52.91%
LS	list spelling	read	107	344	47.09%	52.91%
RT	read tongue twister	read	527	1316	49.24%	50.76%
RR	read command	read	175	1356	47.79%	42.21%
RA	read address	read	491	1356	47.79%	42.21%
DQ	dialogue question	spontaneous	4651	324	50.00%	50.00%
DP	dialogue picture description	spontaneous	2974	344	47.09%	52.91%
MQ	monologue question	spontaneous	2961	344	47.09%	52.91%
MP	monologue picture	spontaneous	4177	648	50.00%	50.00%
EC	elicited command	spontaneous	979	1640	49.39%	50.61%
all	complete corpus	various	11386	9676	48.55%	51.45%

3.2 Experiment I - Entire Corpus

Our first approach was to feed the entire corpus into all three classifiers, which produced results hardly outperforming the ZeroR baseline. Furthermore, this approach comes along with immense computational requirements due to more than 11000 features to be processed. While naïve Bayes and SMO were able to produce results in a reasonable time frame, logistic regression took more than two weeks to complete. Concluding from this test, it can be said that this set of features is too large to be applicable for a text-only-based classification.

After application of feature selection, logistic regression achieved the best accuracy with 58.80% UAR[6] which relates to a relative improvement of more than 14% over the ZeroR baseline. Nevertheless, it was still lower than the Interspeech Speaker State Challenge 2011 baseline[7] of 65.9% UAR [9].

3.3 Experiment II - Individual Document Classes

Since the combination of all features did not lead to the desired results, the next experiment was targeted on checking whether it is useful to further divide the corpus into its different document classes. This approach was also motivated by the question how the performance of a simple task like reading a telephone number compares to that of a rather difficult test such as describing a picture. The

[6] This result was achieved with an information gain threshold of 0.0002 which included 971 featues.

[7] That number was provided in the call for participation which was, however, not restricted to text-only-based features. As some of the samples have been removed to avoid discrepancies in the corpus—as described in section 3.1—these results are not directly comparable.

expectation was that a classifier specifically trained on one document class could be more effective than a classifier trained on the whole corpus. Consequently, the corpus was divided into 11 sub-corpora each of which containing only transcriptions from one document class. Table 1 shows details about each sub-corpus. A similar idea was published in [11], where a comparison between the different prompt types *spontaneous speech*, *tongue twister* and *command-and-control* was performed.

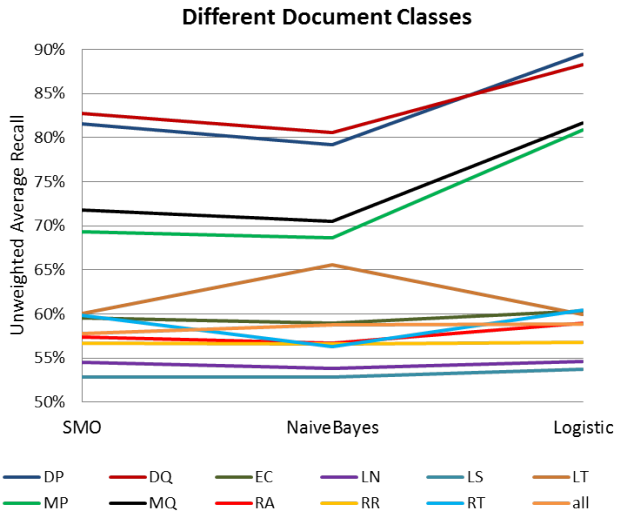

Fig. 1. accuracy of document classes

As shown in Figure 1, there are considerable differences on achieved accuracies supporting our conjecture that document classes vary in terms of expressiveness. The diagram compares the unweighted average recall achieved on each document class using SMO, naïve Bayes and logistic regression classifiers. Just as before, feature selection was applied on the basis of an information gain threshold optimized for each document class. Figure 2 shows the influence of the information gain threshold for the class DP (picture description) in detail. As expected, classes containing spontaneous speech performed best. Among them, dialogue-speech-based document classes (DP, DQ) achieved an unweighted average recall between 79.17% and 89.43% performing considerably better than monologue classes. The latter form the next group in the result set, performing between 68.67% UAR and 80.86% UAR. Table 1 shows that these classes have a considerably higher number of word types than the non-spontaneous classes, being a likely reason for the superior performance.

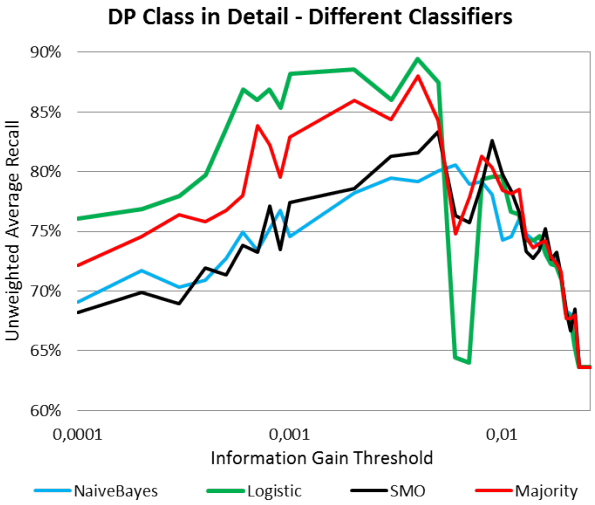

Fig. 2. DP (picture description) optimizing IG threshold

This diagram shows the positive impact of feature selection using information gain on the results. Furthermore, a strange finding from this experiment is that SMO and logistic regression show a steep decline right after their peaks, whereas naïve Bayes does not suffer from such a drop. We were able to recreate this effect in multiple test iterations, but were not able to positively identify its cause as of yet.

3.4 Experiment III - Word Counts

Since the previous tests were all performed using word presence bag-of-word features, the next test examined the potential of adding word count information. Figure 3 reveals that, contrary to our hope, the change from word presence to word count deteriorates accuracy. When using logistic regression the difference in accuracy between word count and word presence is rather small, whereas much higher differences can be observed when using naïve Bayes.

3.5 Experiment IV - Combining Document Classes

The preceding results made us wonder whether a combination of the best performing document classes, e.g. DP, DQ, and MQ, could further improve results due to synergetic effects. At first, we merged sub-corpora containing these three classes into one corpus. Although this combination performed considerably better than the run on the undivided corpus, synergetic effects were not as strong as anticipated and did not result in an improvement compared to individual document classes.

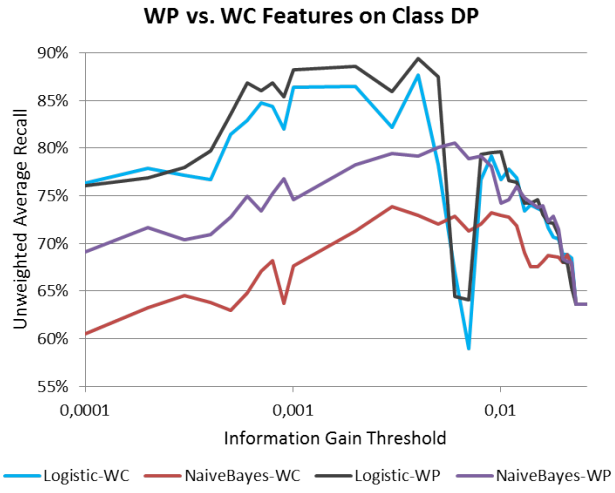

Fig. 3. word presence vs. word count features

3.6 Experiment V - Combining Classifiers

As a last experiment, the three classifiers under consideration were combined into a majority voting system, whose results are shown in Figure 2, too. It turned out that even classifier combination was not able to beat logistic regression, which seems to be the best on this specific domain.

4 Conclusion and Outlook

This study showed the power of pure text-based features to determine whether somebody is intoxicated or sober by analyzing speech transcriptions. There are two major findings. First, it is enough to limit analysis to a single spontaneous speech task as it is much more expressive than read speech. Second, the use of text-based features turned out to be very effective especially in conjunction with logistic regression.

Although the final result of 89.4% UAR on the most expressive document class (DP) is an excellent achievement, it needs to be said that the accuracy still needs to be improved before considering operational scenarios. Yet it is an important step forward to understand that it is not necessary to combine all document classes of the ALC, but that it is more worthwhile to concentrate on one or maybe two classes only. This also allows for further improvement of the test procedure itself since tests can now be developed with a special focus on spontaneous speech.

To improve classification accuracy even further, we will be considering n-gram features to model word order dependencies with a proven record of performance

gain in text classification [4, 10]. Furthermore, we plan to consider a weighting system for classifier combination, such that better classifiers are less likely to be outvoted by worse ones. Also, majority voting could be applied across multiple document classes as suggested in [11] as well.

Another area to look into is the applicability of this research to real-world scenarios. As manual transcription of speech is not available in real-time, software for analysis of intoxication based on text will have to rely on automatic speech recognition. It will therefore be necessary to analyze the influence of speech recognition performance on the accuracy of classification. This is particularly interesting considering that voice and speech characteristics of users may be subject to substantial change under the influence of alcohol.

References

1. L. Blincoe, A. Seay, E. Zaloshnja, T. Miller, E. Romano, S. Luchter, and R. Spicer. The Economic Impact of Motor Vehicle Crashes 2000. Technical report, U.S. Department of Transportation, 2002.
2. T. Bocklet, K. Riedhammer, and E. Nöth. Drink and Speak: On the Automatic Classification of Alcohol Intoxication by 46 Acoustic, Prosodic and Text-Based Features. In *Proc. of the Interspeech*, Florence, Italy, 2011.
3. D. Bone, M. Black, M. Li, A. Metallinou, S. Lee, and S. Narayanan. Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors. In *Proc. of the Interspeech*, Florence, Italy, 2011.
4. C. Boulis and M. Ostendorf. Text Classification by Augmenting the Bag-of-Words Representation with Redundancy-Compensated Bigrams. In *Proc. of the FSDM*, Newport Beach, USA, 2005.
5. I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer, New York, USA, 2006.
6. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.
7. F. Schiel and C. Heinrich. Laying the Foundation for In-Car Alcohol Detection by Speech. In *Proc. of the Interspeech*, Brighton, UK, 2009.
8. F. Schiel, C. Heinrich, S. Barfüsser, and T. Gilg. ALC - Alcohol Language Corpus. In *Proc. of the LREC*, Marrakesch, Marokko, 2008.
9. B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski. The INTERSPEECH 2011 Speaker State Challenge. In *Proc. of the Interspeech*, Florence, Italy, 2011.
10. C. Tan, Y. Wang, and C. Lee. The use of bigrams to enhance text categorization. *Information Processing and Management*, 38(4), 2002.
11. F. Weninger and B. Schuller. Fusing Utterance-Level Classifiers for Robust Intoxication Recognition from Speech. In *Proc. of the ICMI 2011*, Alicante, Spain, 2011.