Topic and Emotion Classification of Customer Surveys

D. Suendermann[1,2], J. Liscombe[2], J. Bloom[2], and R. Pieraccini[2]
{david,jackson,jonathanb,roberto}@speechcycle.com

[1] Baden-Wuerttemberg Cooperative State University, Stuttgart, Germany
[2] SpeechCycle Labs, New York, USA

Abstract. A method to assess the quality of customer service phone interactions is to point callers to an online survey where they can express their opinions, wishes, complaints, commendations, etc. by way of free-form text input. This paper investigates to which extend semantic classification can be applied to large amounts of surveys (thousands) in order to answer questions such as those in the following examples:

- Which callers are calling about their bill, technical issues, product pricing, etc.?
- Has the percentage of callers complaining about long hold time on the phone increased from month to month?
- Who is asking for a call-back or is threatening to cancel service with the company being called?
- Is the caller conveying positive, negative, or neutral emotion referring to a certain topic?

Three statistical classifiers (Ripper, SVM, naïve Bayes) were evaluated on a manually annotated set of 5589 surveys using ten-fold cross-validation. In doing so, 15 different topics (classes) were investigated. In an additional set of experiments, each class was associated with an emotion flag (positive/negative/neutral) to add valence to the picture. In order to cope with the occurence of multiple classes and emotion flags in a single survey, we introduced a novel annotation language encoding semantics, emotion flags, and temporal sequence of topics. A demo system can be accessed at http://suendermann.com/verbatim.php5.

1 Introduction

In a world where product and service features barely differ among competitors of certain businesses, the quality of customer service is an important differentiator. E.g., in the telecommunication industry, bundle services nowadays include cable TV, high-speed Internet, landline and wireless service whose features are largely identical among different providers. In addition to lower pricing, providers try to differentiate their services by means of superior customer service and support. Consequently, one of the main focuses of the customer service departments of large companies is to constantly monitor the quality of services rendered [8].

A frequently used method to assess customer support is to survey customers [16, 7] . This can be done in a number of ways including

1) out-bound calling customers and asking a number of questions,
2) asking customers who are calling into a service hotline a number of questions right after their service interaction,
3) sending customers a personal e-mail after a completed service interaction with a link to a survey web portal.

Survey questions are generally of these types:

A) yes/no (e.g., *Were you satisfied with this customer service interaction?*),
B) multiple choice (e.g., *Which was the reason for your call: billing, payment, technical support, general inquiry, or something else?*), or
C) free-form (e.g., *What was the reason for your call?*).

Responses to questions of Type A or B can be evaluated in a rather straightforward fashion by calculating frequency distributions over the number of possible choices (e.g., 85% of the callers were satisfied [Type A], or 21% of the people called about billing, 18% wanted to make a payment, etc. [Type B]). Type C allows customers to express their opinions and desires in an unconstrained way, which has the potential of conveying lots of useful and detailed information. E.g., by matching customers to the call center representative serving them, it can provide very specific feedback. An example of a Type-3 survey response collected via a web interface of a large cable service provider is

> *Cynthia's assistance went above and beyond. However, even though Cynthia offered me a new contractual option with your company, (Which I will give it a 1 year trial) I feel that my rates for cable & internet are extremely high and if they continue to rise, I will discontinue my service with your company.*

It is certainly worthwhile for customer service managers to read such survey responses every now and then to hear the direct voice of the customers. However, in companies processing millions of customer interactions every week [13], the manual processing of free-form customer feedback becomes unfeasible. Instead, in this paper, we propose the application of semantic classifiers to textual features in order to identify surveys belonging to predefined topics (classes). This method can be useful to answer a variety of questions of primary interest to stakeholders in customer service departments. Examples include the ones listed in the abstract:

- Which callers are calling about their bill, technical issues, product pricing, etc.?
- Has the percentage of callers complaining about long hold time on the phone increased from month to month?
- Who is asking for a call-back or is threatening to cancel service with the company being called?
- Is the caller conveying positive, negativ, or neutral emotion referring to a certain topic?

Section 2 will focus on the derivation of topics and emotion flags; the annotation scheme will be discussed in Section 3. Then, in Section 4 we will provide details on the experimental setup around this work and present results.

2 Topics and Emotion Flags

Topics of particular interest to customer service departments, e.g. in the cable provider market vertical, include surveys about

- an **A**utomated system,
- the **B**illing department,
- the **C**osts of services,
- a billing **D**ispute,
- an **E**mergency situation (e.g., callers threatening to cancel service),
- a request to **F**ollow up with the caller (call-back request),
- a **H**uman representative,
- the automated **I**nternet troubleshooting system [1],
- **O**ther topics,
- a **P**roduct,
- the general-purpose call **R**outer [5],
- a vague mentioning of an automated trouble-**S**hooting system [1],
- a **T**ruck roll or a **T**echnician on site,
- the automated cable **TV** troubleshooting system [1],
- **W**ait time in line.

The bolded letters are unique to each topic and will be used to refer to topics in the scope of the annotation scheme introduced in Section 3.

A fixed number of unique classes to distiguish in written documents generally suggests the application of a semantic classifier similar to what is being used for the task of call routing [4]. There, callers are asked to briefly describe the reason for their call in response to a system prompt such as

> *Briefly tell me what you are calling about today.*

After applying large-vocabulary speech recognition to the caller response, a semantic classifier is applied to the recognition hypthesis returning one of a number of possible call reasons (classes). High-resolution call routers sometimes distinguish hundreds of classes [14].

However, it turns out that responses to call routing system prompts and survey responses of unlimited input length differ considerably in their nature. The example given above is prototypical for free-form responses in that they are not limited to a unique topic but contain a time sequence of topics. The topic sequence of this particular example is decoded in Table 1.

Reviewing this example, we observe that the mentioning of a topic can be associated with a certain emotion. The emotional flavor of a customer comment is clearly of special interest to the customer service department. It is crucial to know whether people like or hate their services, whether they had a positive or negative experience with the call center agent or spoken dialog system, or whether product costs are considered cheap or expensive. For this purpose, we introduce a three-point emotion scale (positive/neutral/negative).

In Table 1, each topic is also associated with an emotion flag, so, we see for instance that a human agent is mentioned twice, once in a positive way (*went above and beyond*) and once neutral (*Cynthia offered me*).

Table 1. Example for a time sequence of topics

text	annotation	emotion flag
Cynthia's assistance went above and beyond.	H	+
However, even though Cynthia offered me...	H	
a new contractual option with your company, (Which I will give it a 1 year trial)...	O	
I feel that my rates for cable & internet are extremely high...	C	-
and if they continue to rise, I will discontinue my service with your company.	E	-

3 Annotation

As motivated in Section 1, we want to apply statistical classifiers in order to automatically analyze the membership of a given survey to the classes and emotion flags introduced in Section 2. In order to train the classification models, we need to establish respective training data. In our case, we need to map the survey text to the canonical classes and emotion flags it represents. This process is often done in a supervised manner (i.e., manually) and is referred to as *annotation*.

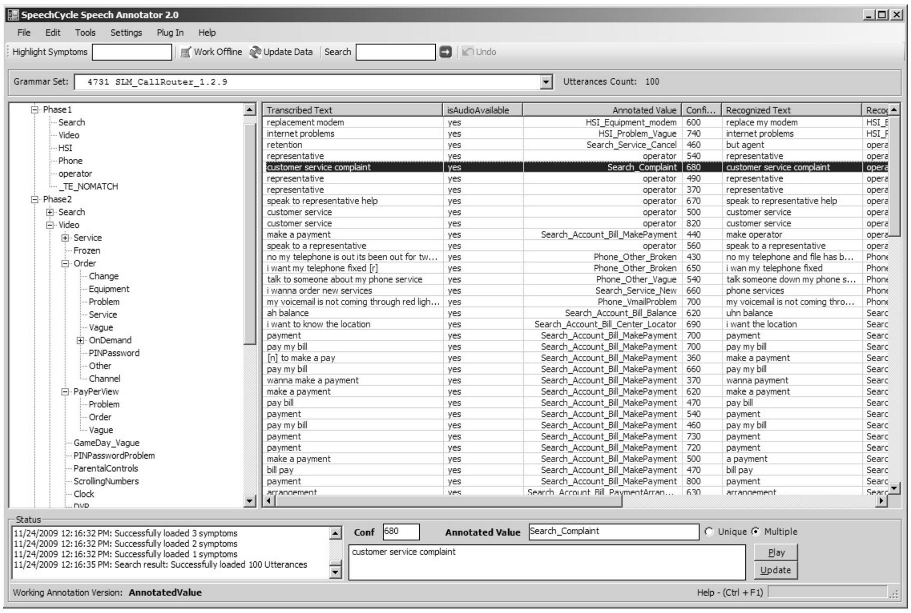
Semantic annotation as required for a standard call routing task (see Section 2) maps exactly one class to a given utterance/text [5]. Figure 1 shows annotation software which lists a number of caller responses to the aforementioned example prompt *Briefly tell me what you are calling about today.* On the left, possible classes are shown in a hierarchical fashion (similar to a folder structure). The annotation task consists now of dragging and dropping utterances into one of the classes on the left. For example, the utterance *I am having a problem ordering a movie* refers to cable TV service (aka *Video*), it is about an *Order* and describes a *Problem*. The correct class would hence be

```
Video_Order_Problem
```

Sometimes, callers refer to multiple reasons at once (e.g., *I'd like to order a show and pay last month's bill*). Since the above described annotation method is not designed to accomodate multiple classes for a single utterance, this utterance would be mapped to a generic multiple-symptom class. Since, usually, these cases are negligible (0.4% for our example call router), no special handling for mappings to multiple classes is required.

As shown in Section 2, the situation is completely different for the case of unrestricted surveys. In fact, the corpus used in our experiments (see Section 4), contained 64% surveys with multiple classes.

To cover all possible scenarios of classes and emotion flags which can be associated with a given survey, we came up with a simple language describing



Grammar Set: 4731 SLM_CallRouter_1.2.9

Utterances Count: 100

Transcribed Text	isAudioAvailable	Annotated Value	Confi...	Recognized Text	Reco
replacement modem	yes	HSI_Equipment_modem	600	replace my modem	HSI_E
internet problems	yes	HSI_Problem_Vague	740	internet problems	HSI_F
retention	yes	Search_Service_Cancel	460	but agent	opera
representative	yes	operator	540	representative	opera
customer service complaint	yes	Search_Complaint	680	customer service complaint	opera
representative	yes	operator	490	representative	opera
representative	yes	operator	370	representative	opera
speak to representative help	yes	operator	670	speak to representative help	opera
customer service	yes	operator	500	customer service	opera
customer service	yes	operator	820	customer service	opera
make a payment	yes	Search_Account_Bill_MakePayment	440	make operator	opera
speak to a representative	yes	operator	560	speak to a representative	opera
no my telephone is out its been out for tw...	yes	Phone_Other_Broken	430	no my telephone and file has b...	Phone
i want my telephone fixed [r]	yes	Phone_Other_Broken	650	i wan my telephone fixed	Phone
talk to someone about my phone service	yes	Phone_Other_Vague	540	talk someone down my phone s...	Phone
i wanna order new services	yes	Search_Service_New	660	phone services	Phone
my voicemail is not coming through red ligh...	yes	Phone_VmailProblem	700	my voicemail is not coming thro...	Phone
ah balance	yes	Search_Account_Bill_Balance	620	uhn balance	Searc
i want to know the location	yes	Search_Account_Bill_Center_Locator	690	i want the location	Searc
payment	yes	Search_Account_Bill_MakePayment	700	payment	Searc
pay my bill	yes	Search_Account_Bill_MakePayment	700	pay my bill	Searc
[n] to make a pay	yes	Search_Account_Bill_MakePayment	360	make a payment	Searc
pay my bill	yes	Search_Account_Bill_MakePayment	660	pay my bill	Searc
wanna make a payment	yes	Search_Account_Bill_MakePayment	370	wanna payment	Searc
make a payment	yes	Search_Account_Bill_MakePayment	620	make a payment	Searc
pay bill	yes	Search_Account_Bill_MakePayment	470	pay bill	Searc
payment	yes	Search_Account_Bill_MakePayment	540	payment	Searc
pay my bill	yes	Search_Account_Bill_MakePayment	460	pay my bill	Searc
payment	yes	Search_Account_Bill_MakePayment	730	payment	Searc
payment	yes	Search_Account_Bill_MakePayment	720	payment	Searc
make a payment	yes	Search_Account_Bill_MakePayment	500	a payment	Searc
bill pay	yes	Search_Account_Bill_MakePayment	470	bill pay	Searc
payment	yes	Search_Account_Bill_MakePayment	800	payment	Searc
arrangement	yes	Search_Account_Bill_PaymentArran...	630	arrangement	Searc

Status

```
11/24/2009 12:16:32 PM: Successfully loaded 3 symptoms
11/24/2009 12:16:32 PM: Successfully loaded 2 symptoms
11/24/2009 12:16:32 PM: Successfully loaded 1 symptoms
11/24/2009 12:16:35 PM: Search result: Successfully loaded 100 Utterances
```

Conf 680 **Annotated Value** Search_Complaint ○ Unique ● Multiple

customer service complaint

Working Annotation Version: **AnnotatedValue**

Help - (Ctrl + F1)

Fig. 1. Annotation software processing data of a call routing task

the time sequence of topics and emotion flags encountered in the survey. Here, the coding scheme of Table 1 is used, so, for the table's example, the semantic annotation string is

```
H+HOC-E-
```

Generally, our annotation language l can be expressed as

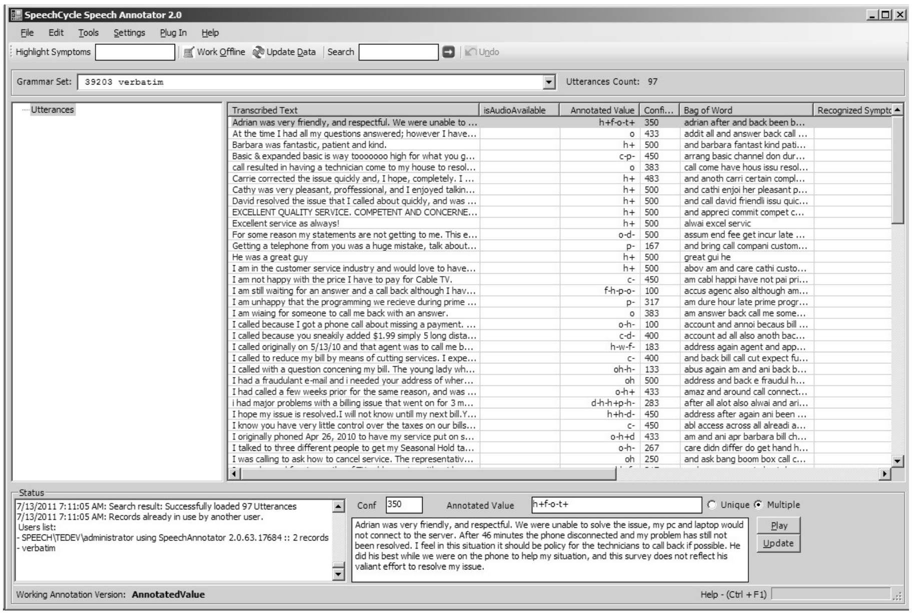
$$
\begin{aligned}
l &:= c[l] \\
c &:= t[e] \\
t &\in \{\mathtt{P}, \mathtt{H}, \mathtt{W}, \mathtt{T}, \mathtt{B}, \mathtt{D}, \mathtt{A}, \mathtt{R}, \mathtt{I}, \mathtt{V}, \mathtt{C}, \mathtt{F}, \mathtt{E}, \mathtt{O}\} \\
e &\in \{+, -\}
\end{aligned}
$$

Figure 2 shows how the same annotation software we have applied to the call routing scenario can be used to produce the annotation string. While reading the survey, the annotating person writes the string into the *Annotated Value* field.

4 Experiments

4.1 The Classification Framework

A practical way to answer the questions raised in the introduction of this paper is to train separate classifiers for each topic (class). These classifiers would be



Grammar Set: 39203 verbatim

Utterances Count: 97

Transcribed Text	isAudioAvailable	Annotated Value	Confi...	Bag of Word	Recognized Sympto
Adrian was very friendly, and respectful. We were unable to ...		h+f-o-t+	350	adrian after and back been b...	
At the time I had all my questions answered; however I have...		o	433	addit all and answer back call ...	
Barbara was fantastic, patient and kind.		h+	500	and barbara fantast kind pati...	
Basic & expanded basic is way tooooooo high for what you g...		c-p-	450	arrang basic channel don dur...	
call resulted in having a technician come to my house to resol...		o	383	call come have hous issu resol...	
Carrie corrected the issue quickly and, I hope, completely. I ...		h+	483	and anoth carri certain compl...	
Cathy was very pleasant, proffessional, and I enjoyed talkin...		h+	500	and cathi enjoi her pleasant p...	
David resolved the issue that I called about quickly, and was ...		h+	500	and call david friendli issu quic...	
EXCELLENT QUALITY SERVICE. COMPETENT AND CONCERNE...		h+	500	and appreci commit compet c...	
Excellent service as always!		h+	500	alwai excel servic	
For some reason my statements are not getting to me. This e...		o-d-	500	assum end fee get incur late ...	
Getting a telephone from you was a huge mistake, talk about...		p-	167	and bring call compani custom...	
He was a great guy		h+	500	great gui he	
I am in the customer service industry and would love to have...		h+	500	abov am and care cathi custo...	
I am not happy with the price I have to pay for Cable TV.		c-	450	am cabl happi have not pai pri...	
I am still waiting for an answer and a call back although I hav...		f-h-p-o-	100	accus agenc also although am...	
I am unhappy that the programming we recieve during prime ...		p-	317	am dure hour late prime progr...	
I am wiaing for someone to call me back with an answer.		o	383	am answer back call me some...	
I called because I got a phone call about missing a payment. ...		o-h-	100	account and annoi becaus bill ...	
I called because you sneakily added $1.99 simply 5 long dista...		c-d-	400	account ad all also anoth bac...	
I called originally on 5/13/10 and that agent was to call me b...		h-w-f-	183	address again agent and app...	
I called to reduce my bill by means of cutting services. I expe...		c-	400	and back bill call cut expect fu...	
I called with a question concening my bill. The young lady wh...		oh-h-	133	abus again am and ani back b...	
I had a fraudulant e-mail and i needed your address of wher...		oh	500	address and back e fraudul h...	
I had called a few weeks prior for the same reason, and was ...		o-h+	433	amaz and around call connect...	
i had major problems with a billing issue that went on for 3 m...		d-h-h+p-h-	283	after all alot also alwai and ari...	
I hope my issue is resolved.I will not know untill my next bill.Y...		h+h-d-	450	address after again ani been ...	
I know you have very little control over the taxes on our bills...		c-	450	abl access across all alreadi a...	
I originally phoned Apr 26, 2010 to have my service put on s...		o-h+d	433	am and ani apr barbara bill ch...	
I talked to three different people to get my Seasonal Hold ta...		o-h-	267	care didn differ do get hand h...	
I was calling to ask how to cancel service. The representativ...		oh	250	and ask bang boom box call c...	

Status

7/13/2011 7:11:05 AM: Search result: Successfully loaded 97 Utterances
7/13/2011 7:11:05 AM: Records already in use by another user.
Users list:
- SPEECH\TEDEV\administrator using SpeechAnnotator 2.0.63.17684 :: 2 records
- verbatim

Conf 350

Annotated Value h+f-o-t+

○ Unique ● Multiple

Adrian was very friendly, and respectful. We were unable to solve the issue, my pc and laptop would not connect to the server. After 46 minutes the phone disconnected and my problem has still not been resolved. I feel in this situation it should be policy for the technicians to call back if possible. He did his best while we were on the phone to help my situation, and this survey does not reflect his valiant effort to resolve my issue.

Working Annotation Version: **AnnotatedValue**

Fig. 2. Annotation software processing data of the survey task

binary when discarding emotion flags at the first place, i.e., the classifier would return `1` in the case it is confident that the survey is about a certain topic, otherwise `0`. This means that as many classifiers have to be trained as there are distinct classes, i.e., in our case 15.

When adding emotion flags to the picture, one has to be aware of the fact that a single survey can possibly contain multiple mentionings of the same topic with different emotion flags each. Principly, every single combination of positive, negative, and neutral are possible in a single survey for a single class (in our example in Section 2, we had positive and neutral for the class `H`. Consequently, when we would intend to use a single classifier per topic, it would have to be able to return every possible combination of emotion flags: `+`, `-`, `0`, `+-`, `+0`, `-0`, `+-0`, so, seven distinct return values. Here, `0` stands for *neutral*.

Another possibility to cope with emotion flags in this framework is to train separate binary classifiers for each topic/emotion flag combination. I.e., we would have an `H+` classifier, an `H-` classifier, and an `H0` classifier for the topic `H`.

4.2 Measuring Performance

In addition to the substantial difference between the annotation scheme of a call router and that of the free-form surveys we introduced in Section 3, there is also a major difference in the way classifier performance should be measured. In spoken-language understanding tasks as for instance in call routing, the classification hypothesis is simply compared with the canonical class (which a human annotator produced for the utterance in question). Here, the hypothesis is either correct of wrong. The metric True Total is the number of correct matches divided by the total number of samples in a test corpus, i.e., it is the percentage of correct responses of the classifier on a given test corpus [15].

Theoretically, one can calculate the True Total also for the binary classification scenario of the current work. However, as it turns out, the result can be misleading. This is because some of the topics have a very low likelihood of occurrence. For instance, only 0.2% of the surveys analyzed in this work mentioned I (see Table 2). That means, if we build a trivial classifier that exclusively returns the majority vote (in this case 0), it would be correct in 99.8% of the cases, a True Total that seems extraordinarily good. However, it missed all the cases that *did* mention I rendering it completely useless.

Table 2. Distribution of topics in the corpus.
Note: Percentages describe the fraction of surveys in which the topic/the topic with a certain emotion flag was found. Due to multiple occurrences of topics/emotion flags in some surveys, total does not add up to 100%, and + and - do not necessarily add up to total.

topic	description	total	+	-
A	automation	3.8%	0.2%	3.4%
B	billing	0.6%	0.0%	0.5%
C	cost	10.5%	0.3%	9.9%
D	dispute	4.4%	0.1%	3.6%
E	emergency	8.4%	0.0%	6.2%
F	follow-up	3.8%	0.5%	3.0%
H	human	66.6%	50.3%	17.6%
I	Internet	0.2%	0.0%	0.1%
O	other	34.0%	6.7%	20.2%
P	product	23.1%	2.5%	20.3%
R	call router	1.0%	0.0%	0.9%
S	troubleshooter	1.4%	0.2%	1.2%
T	truck	10.9%	6.3%	3.7%
V	TV	0.2%	0.0%	0.2%
W	wait	3.5%	0.3%	3.2%

In cases like these, the machine learning community usually considers the standard metrics Precision, Recall, and F-Measure [11]. Precision is the percentage of correctly accepted tokens in the set of accepted tokens. So, Precision describes the *quality* of accepted tokens. Recall, on the other hand, is the percentage of the correctly accepted tokens in the set of all tokens which *should have been* accepted. That is, Recall describes the *completeness* of accepted tokens. Finally, F-Measure is a harmonic mean of Precision and Recall.

Depending on the specifics of the classification task, Precision and Recall may not be of equal importance, a fact that is accounted for by different flavors of F-Measures. F_1, the most commonly used metric, treats Precision and Recall identically, whereas F_2 weights Recall twice as strong as Precision. In the current work, F_2 turned out to be a more appropriate metric than F_1 because missing tokens of some of the topics (such as emergency callers, requests for follow-up, or billing disputes) are considered critical, i.e., missing instances of such topics are more expensive than false alarms. At any rate, since the above mentioned trivial majority vote classifier would not accept any tokens, its Recall would consequently be zero, so would be *any* F-Measure, including F_2.

4.3 Corpus and Experimental Results

For a large cable service provider [1] with a call volume of several million calls every month to its service hotline, we collected free-form online surveys as described in the introduction of this paper. The collected surveys amounted to about ten thousand every month. For a first proof of concept, we focused on a single month (May 2010) for which a number of 5589 randlomly selected surveys were annotated according to the scheme described in Section 3. We did not separate fixed training and test sets but instead used ten-fold cross-validation [3] in our experiments.

In a first round of experiments, we compared the performance of several state-of-the-art classifiers on this task. We selected the following classifiers from the WEKA toolbox [6] for this work:

- Ripper (a decision tree learner) [2],
- Sequential Minimal Optimization (SMO), a fast support vector machine implementation [9],
- naïve Bayes [4].

All these classifiers rely on sets of feature vectors and their associated class labels as training data, so, the survey text had to be converted into a feature representation. There are multiple techniques to represent utterances or texts in vector form, out of which we have been using the following ones:

- **wpres1.** Each vector element represents one word type in the vocabulary. For a specific text, all those elements representing words present in the respective text are `1`, all the others are `0`.
- **wpres5.** The same as **wpres1**, but only types whose total count in the training data is five or more are considered in the vector.
- **wcount1.** The same as **wpres1**, but instead of `1` to indicate the presence of a word in the text, the *count* of the word is used as element value.

- **wcount5.** The same as **wcount1** but discarding types with a total count of four or less.
- **bowpres1.** The same as **wpres1**, but before establishing vocabulary and vector elements, texts are converted into a bag-of-word representation, a compressed but semantically almost identical form of the text [10, 4].
- **bowpres5.** The same as **bowpres1** but discarding types with a total count of four or less.
- **tfidf1.** The same as **wpres1**, but the element values represent the text's words' TF-IDF scores [12].
- **tfidf5.** The same as **tfidf1**, but discarding types with a total count of four or less.
- **tfidfbow1.** The same as **tfidf1**, but after conversion into a bag-of-word representation.
- **tfidfbow5.** The same as **tfidfbow1**, but discarding types with a total count of four or less.

For the first experiment (to compare classifiers), we limited analysis to **tfidf1** features which are very common in information retrieval and data mining. We performed topic classification as well as joint classification of topics and emotion flags as discussed in Section 4.1.

At a first glance, the results seem to be slightly disappointing, with many results below 0.5 and even some 0. At this point, we have to remind the reader of the motivation behind using F_2 which was that a classifier can only be deemed useful when there is a Recall greater than 0 which means, at least one test sample has to be correctly identified. Given the extremely sparse and, at the same time, linguistically diverse set of examples for certain classes, it is almost impossible for a classifier to produce reasonable output. Nonetheless, this first experiment clearly indicates that the classification tree algorithm Ripper outperforms its competitors SMO and naïve Bayes and will therefore be used in the continuation of this project. Furthermore, we will use a consolidated score across classes (the weighted average as shown in the last row of Table 3) in order to help drawing conclusions more easily.

Looking at the joint classification of topics and emotion flags (in parenthesis in Table 3, classifier is Ripper), it is interesting how similar the results are to pure topic classification. For some of the topics, subdivision into more classes by adding emotion flags even results in a performance gain.

Results of our experiments to compare different feature vectors are shown in Table 4. Here, we used the Cost Sensitive Meta Classifier offered by WEKA which allowed us to optimize results towards our target metric F_2. This is why, this time, **tfidf1** achieved a higher score than in Table 3.

5 Conclusion

According to these results, the well-established TF-IDF metric performed lower than bag-of-word vectors. The absolute values of $F_2 = 0.71$ indicate that the

Table 3. Comparing classifiers for topics and emotion tags. In bold, results (F_2) greater than 0.5.

topic	Ripper ($\pm$)		SMO	naïve Bayes
A	**0.53**	(0.36)	0.04	0.02
B	0.23	(0)	0	0
C	**0.61**	**(0.58)**	0.08	0.32
D	0.20	(0.32)	0.01	0.05
E	0.25	(0.22)	0.05	0.19
F	0.19	(0.24)	0	0.02
H	**0.83**	**(0.85)**	**0.89**	**0.81**
I	0	(0)	0	0
O	0.31	(0.33)	0.24	0.49
P	0.26	(0.32)	0.11	0.50
R	0.15	(0.15)	0	0
S	0.08	(0.17)	0	0.02
T	**0.58**	**(0.61)**	0.07	0.22
V	0.10	(0)	0	0
W	**0.57**	**(0.59)**	0.07	0.47
avg	**0.63**	**(0.61)**	0.22	0.42

Table 4. Comparing features. Winners in bold.

feature	Precision	Recall	F_2
wpres1	0.56	0.75	0.70
wpres5	0.54	0.72	0.67
wcount1	0.50	0.74	0.67
wcount5	0.58	0.72	0.68
bowpres1	0.71	0.71	**0.71**
bowpres5	0.60	0.74	**0.71**
tfidf1	0.65	0.65	0.65
tfidf5	0.50	0.69	0.64
tfidfbow1	0.72	0.7	0.70
tfidfbow5	0.80	0.66	0.69

technique can indeed be useful when trying to detect infrequent surveys of specific topics in large amounts of data. Taking **bowpres5** as example: A Recall of 0.74 means that the classifier is missing only 26% of the topic's surveys. In contrast, a Precision of 0.6 means that 60% of the surveys returned by the classifier actually referred to the topic. Without classification, this percentage would be much much lower, e.g. < 10% for most of the topics shown in Table 2. Hence, topic classification as preprocessing step can significantly reduce the manual workload associated with the screening of tens of thousands of surveys every month specifically for rare topics and emotion flags.

References

1. Acomb, K., Bloom, J., Dayanidhi, K., Hunter, P., Krogh, P., Levin, E., Pieraccini, R.: Technical Support Dialog Systems: Issues, Problems, and Solutions. In: Proc. of the HLT-NAACL. Rochester, USA (2007)
2. Cohen, W.: Fast Effective Rule Induction. In: Proc. of the International Conference on Machine Learning. Lake Tahoe, USA (1995)
3. Devijver, P., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice Hall, Englewood Cliffs, USA (1982)
4. Evanini, K., Suendermann, D., Pieraccini, R.: Call Classification for Automated Troubleshooting on Large Corpora. In: Proc. of the ASRU. Kyoto, Japan (2007)
5. Gorin, A., Riccardi, G., Wright, J.: How May I Help You? Speech Communication 23(1/2) (1997)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
7. Hone, K., Graham, R.: Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI). Natural Language Engineering 6(3-4) (2000)
8. Neustein, A.: Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics. Springer, New York, USA (2010)
9. Platt, J.: Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Tech. rep., Microsoft Research, Seattle, USA (1998)
10. Porter, M.: An Algorithm for Suffix Stripping. Program 14(3) (1980)
11. van Rijsbergen, C.: Information Retrieval. Butterworths, London, UK (1979)
12. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw Hill, New York, USA (1983)
13. Suendermann, D.: Advances in Commercial Deployment of Spoken Dialog Systems. Springer, New York, USA (2011)
14. Suendermann, D., Hunter, P., Pieraccini, R.: Call Classification with Hundreds of Classes and Hundred Thousands of Training Utterances ... and No Target Domain Data. In: Proc. of the PIT. Kloster Irsee, Germany (2008)
15. Suendermann, D., Liscombe, J., Dayanidhi, K., Pieraccini, R.: A Handsome Set of Metrics to Measure Utterance Classification Performance in Spoken Dialog Systems. In: Proc. of the SIGdial Workshop on Discourse and Dialogue. London, UK (2009)
16. Walker, M., Litman, D., Kamm, C., Abella, A.: PARADISE: A General Framework For Evaluating Spoken Dialogue Agents. In: Proc. of the ACL. Madrid, Spain (1997)