Words that fascinate the listener: predicting affective ratings of on-line lectures

Felix Weninger, Pascal Staudt, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Weninger, Felix, Pascal Staudt, and Björn Schuller. 2013. "Words that fascinate the listener: predicting affective ratings of on-line lectures." *International Journal of Distance Education Technologies* 11 (2): 110–23. https://doi.org/10.4018/jdet.2013040106.

Words that Fascinate the Listener:
Predicting Affective Ratings of On-Line Lectures

Felix Weninger, Institute for Human-Machine Communication, Technische Universität München, Munich, Germany

Pascal Staudt, Institute for Human-Machine Communication, Technische Universität München, Munich, Germany

Björn Schuller, Institute for Human-Machine Communication, Technische Universität München, Munich, Germany

ABSTRACT

In a large scale study on 843 transcripts of Technology, Entertainment and Design (TED) talks, the authors address the relation between word usage and categorical affective ratings of lectures by a large group of internet users. Users rated the lectures by assigning one or more predefined tags which relate to the affective state evoked in the audience (e. g., 'fascinating', 'funny', 'courageous', 'unconvincing' or 'long-winded'). By automatic classification experiments, they demonstrate the usefulness of linguistic features for predicting these subjective ratings. Extensive test runs are conducted to assess the influence of the classifier and feature selection, and individual linguistic features are evaluated with respect to their discriminative power. In the result, classification whether the frequency of a given tag is higher than on average can be performed most robustly for tags associated with positive valence, reaching up to 80.7% accuracy on unseen test data.

Keywords: *Emotion Recognition, Linguistic Features, Online Lectures, Technology-Entertainment and Design Talks (TEDTalks), Text Classification*

INTRODUCTION

Sensing affect related states, including interest, confusion, or frustration, and adapting behavior accordingly, is one of the key capabilities of humans; consequently, simulating such abilities in technical systems through signal processing and machine learning techniques is believed to improve human-computer interaction in general (Schuller & Weninger, 2012) and computer based learning in particular (Aist, Kort, Reilly, Mostow, & Picard, 2002; Forbes-Riley & Litman, 2010). Important abilities of affective tutors or lecturers, besides emotional expressivity (Huang, Kuo, Chang, & Heh, 2004), include the choice of appropriate wording, which has been found to be highly important in computer

DOI: 10.4018/jdet.2013040106

based tutoring to support the learning outcome (Narciss & Huth, 2004). Furthermore, there is increased evidence for the influence of affect related states on the learning process (Craig, Graesser, Sullins, & Gholson, 2004; Bhatt, Evens, & Argamon, 2004; Forbes-Riley & Litman, 2007). In particular, previous studies highlighted the relation between system responses in a tutoring dialogue and student affect (Pour, Hussein, Al Zoubi, D'Mello, & Calvo, 2011); it turned out, for example, that dialogue acts of an automated tutor influence student uncertainty (Forbes-Riley & Litman, 2011). However, these studies do not take into account the linguistic content of lectures as a whole; hence, we aim to bridge this gap by addressing the automatic assignment of categorical affective ratings by a large audience to on-line lectures from the TED talks website (www.ted.com/talks). This prediction is based on learning the relation between linguistic features of the speech transcripts and the ratings given by the audience, which comprises many thousands of internet users in our case. Such automatic predictions can be immediately useful to evaluate the quality of lectures given by a distant education system, and to gain insight into which lecture topics or lecturing strategies are related to certain affective states. The aspect of predicting the induced affect from the lecturers' speech has—to our knowledge—not been addressed in a systematic fashion so far: Rather, in (Forbes-Riley & Litman, 2011), features from student responses to the system and abstract goals of the dialogue manager are used to analyze student affect. In this respect, that study is somewhat related to sentiment analysis (Schuller & Knaup, 2010) or opinion mining (Turney, 2002), where the goal is to deduce the affect of the users from written reviews. However, in our study we aim at predicting the users' affective ratings based on the lectures themselves. This also distinguishes our contribution from the large body of literature on prediction of (ordinal-scale) movie ratings—for a recent study on the public Internet Movie Database (IMDB), we refer to (Marovic, Mihokovic, Miksa, Pribil, & Tus, 2011). In that field, in contrast to our study, the vast majority of approaches seem to be exploiting similarities in user profiles rather than features of the rated objects (*instances* in terms of machine learning), such as in (Marlin, 2003).

Finally, in contrast to many previous studies focusing on singular affects or ratings on a single 'good or bad' scale, we investigate the multi-label categorical ratings from the TED talks website which are given by internet users through assignment of tags to each talk. The tag set is determined by the creators of the TED talks website. These tags are on the one hand directly associated with the emotion evoked in the audience (e. g., 'obnoxious', 'funny'); on the other hand, they can refer to perceived attributes of the speaker resulting in a certain affect of the audience (e. g., 'courageous', which may result in 'feeling moved'). Third, the tags may describe the argumentative structure of the talk (e. g., 'long-winded', 'unconvincing'), which is arguably reflected in affective states such as 'boredom' or 'confusion' which are often investigated in the context of tutoring (Pour et al., 2011; Forbes-Riley & Litman, 2011). For the most part, these tags refer to emotion-related states (e.g., confusion, feeling inspired) rather than full blown 'Big 6' emotions, while 'obnoxious' (arguably, the 'strongest' of the 14 tags) could roughly correspond to disgust. Furthermore, in most cases, tags can be classified into those inducing positive or negative valence (e.g., 'confusing' and 'long-winded' for negative valence, and 'inspiring' or 'fascinating' for positive valence).

On lectures from the TED talks website, a slightly tongue-in-cheek analysis has been performed as to the relation of linguistic content and the user ratings (www.get-tedpad.com). This analysis is based on n-gram language modeling, and simplifies the categories to a 'good/bad' classification; still, we are not aware of a scientifically rigorous study on this topic.

In the remainder of this article, we first detail our evaluation database, which contains 843 TED talk transcripts of roughly two million words. There, we also explain how we turn the analysis of categories into a dimensional problem by defining binary classification tasks,

which we consider relevant for practical applications in lecture evaluation. Then, the experimental setup and results, including feature extraction and classifiers, are laid out. Finally, feature relevance analysis is performed and conclusions are drawn.

EVALUATION DATABASE

Overview

The TED talks web site offers over thousand lectures from a wide range of topics. The speakers have a maximum of 18 minutes to talk. The language is English. The lectures are released under the Creative Commons BY-NC-ND license and are thus freely available for research. Transcripts are available for a majority of the lectures. For this study, we collected all 843 transcribed lectures which were available at the time of data collection (April 2011); we expect this number to be further increasing in the future. While audio and video are available in addition to the transcript, we use text data exclusively in this study. In a real-life system for affect classification, one cannot always rely on ground truth transcripts, but one would have to use automatic speech recognition (ASR) instead. Yet, as our study focuses on linguistic features for a novel affect classification paradigm, we thereby eliminate ASR inaccuracy as a confounding factor.

The TED website allows the users to rate each talk by selecting up to three of 14 predefined tags; the number of times that a certain tag has been assigned to a single talk will be called *tag frequency* in the following. In the collected data set, on average, the total amount of tags given is 1,695. This number indicates that each talk is rated by a few hundred users at least, supposing non-malicious system use. Since the ratings are anonymous, no information is available on the background of the raters; however, since all talks are given in English, a certain familiarity of the raters with the Western culture can be assumed safely. The available tags are shown in Table 1 along with statistics on the tag frequency.

One can see that the average tag frequencies vary strongly for different tags. For instance, the tag 'inspiring' is assigned 290 times per talk on average while 'confusing' is only assigned 18 times. The maximum tag frequency of 13,989 occurs for 'jaw-dropping' while 'long-winded' was assigned only 238 times at most. Overall, it is evident that positive tags seem to be assigned much more frequently than negative or neutral tags.

Subdivision

We split the corpus of 843 TED talks into a training, validation and test set. While the instances in the training set are used to build a model for classification, the disjoint test set is used to evaluate its ability to label unseen test data (cf. the section 'Automatic Classification Experiments' below). The validation set is used to optimize design decisions or 'hyperparameters' in the process of building classification models on a set that is disjoint from both the training and the test set. While the precision of the estimated model parameters is generally expected to increase with larger training sets, the statistical significance of the evaluation decreases with smaller validation and test set sizes. Taking into account these requirements, we split the corpus of 843 TED talks into a training, validation and test set of roughly equal sizes, following a straightforward protocol to foster reproducibility. Defining s_j as the unique ID given by the website to talk j modulo 3, we assigned all lectures j with $s_j = 0$ to the training, all with $s_j = 1$ to the validation and those with $s_j = 2$ to the test set. Since the talk ID depends on the order in which the lectures were published, this splits the corpus in a way that the distribution of newer and older lectures is nearly the same in all three sets—this ensures a near equal amount of user ratings in practice, since influence factors such as popularity can be assumed as random when following this partitioning strategy. In the result, 277 lectures are contained in the training, 285 in the validation, and 281 in the test set. The slightly differing numbers of lectures per set are due to the fact

Table 1. Minimum, maximum and mean tag frequency per tag in the TED talks database

	Tag Frequency		
Tag	**Min**	**Max**	**Mean**
Positive			
Jaw-dropping	0	13,989	176
Funny	0	7,185	113
Courageous	0	5,497	90
Fascinating	2	8,729	231
Inspiring	3	10,601	290
Ingenious	0	2,334	121
Beautiful	0	6,000	105
Informative	0	3,680	210
Persuasive	0	4,965	171
Neutral/Negative			
OK	2	348	58
Confusing	0	526	18
Unconvincing	0	2,020	51
Long-winded	0	238	34
Obnoxious	0	1,026	24

that the talk IDs are not continuous, probably since some of the lectures have been removed from the website.

Obtaining Dimensional Ratings

Unlike categorical annotation schemes often followed in emotion recognition (Schuller, Batliner, Steidl, & Seppi, 2011), the tags given by the users are not mutually exclusive. Hence, we treat the tags as dimensions in this study; an optimal way to assess and handle the possible interdependencies between those tags remains to be investigated in future research.

Thus, for each talk, a '14-dimensional' annotation is obtained from the tag frequencies. An essential part of this transformation is normalization: The total number of tags assigned to a talk strongly depends on the time that the talk has been already available at the website, and on the general popularity of the talk, which in turn may depend on the topic. In fact, the total number of tags assigned to a talk varies from 96 to 46 k.

Therefore, to obtain an annotation independent of both the total number of ratings and the overall frequency of the tags, we first calculate the 'relative tag frequency' r_{ij} for each talk:

$$r_{ij} = \frac{n_{ij}}{N_j}$$

where n_{ij} is the frequency of tag i for talk j and N_j is the total number of tags assigned to talk j. With this relative measure we set a threshold for each tag i, which allows to define a discrete class label $c_i(j) \in \{ \text{yes, no} \}$ indicating whether the tag i is assigned to talk j more frequently than 'it could be expected' or not. From an application point of view, this class label determines whether an automatic system should assign the tag i to the talk j. More precisely, the labels $c_i(j)$ are computed by discretizing r_{ij} at the median m_i of the r_{ij} among

all lectures in the data set, i.e., $c_i(j)$ = yes if and only if $r_{ij} > m_i$.

It is notable that in our study, ratings are obtained from thousands of anonymous users while in fields such as personality analysis or emotion recognition often a carefully chosen set of annotators is employed to get a stable ground truth (Schuller et al., 2011). We are aware of the fact that quality control is a non-trivial issue; however, we argue that in contrast to ratings by large groups of paid subjects as done in crowdsourcing (Parent & Eskenazi, 2011), there is no incentive to produce 'random' ratings in our case. The fact that a user may—intentionally or not—assign the same rating multiple times can even be seen as valuable information (emphasis) as long as such behavior does not occur excessively.

AUTOMATIC CLASSIFICATION EXPERIMENTS

The discrete class labels which are assigned to each talk allow viewing the prediction of user ratings as text classification problems. In general text classification, the goal is to automatically assign category labels to textual documents by means of linguistic features, such as occurrence of certain keywords. In our study, we consider one binary (yes/no) text classification problem per tag. We take a purely data-based approach where the relation of linguistic features and certain tags, including their relevance for classification, is learned fully automatically from a large set of training documents (cf. above).

Linguistic Features

This study focuses on the set of words and bag of words (BoW) models: In the set of words model, only binary features exist, which indicate the presence (1) or absence (0) of a word. In contrast, in a BoW model, features correspond to word counts. It has been observed that more sophisticated contextual models do not achieve fundamental improvements in comparison to the resulting blow-up of the feature space (Sebastiani, 2002). In addition, we considered the TF x IDF approach (term frequency times inverse document frequency). In this model, the first factor (TF) measures the frequency of a 'term' (in our case, a word) in a lecture. The second factor (IDF) is used to enhance precision, assuming that using terms which occur in a high percentage of the documents lead to many 'false positives' (Salton & Buckley, 1988). In our experiments, we compute the IDF factor on the training set. The size of the feature space, corresponding to the vocabulary size of the training set, is 36 k. Hence, there is a need for classifier that can handle large feature spaces efficiently (cf. the subsequent section); furthermore, we investigate feature selection methods (cf. below).

Classifiers

The classification algorithms used for the experiments are implemented in the Weka toolkit (Hall et al., 2009; Witten & Frank, 2011) for straightforward reproducibility. Naïve Bayes (NB) is designed for binary (set of word) features while Multinomial Naïve Bayes (MNB) handles multinomial features, i.e., word counts as in the bag of words model (McCallum & Nigam, 1998). Naïve Bayes classifiers are a popular technique for text classification, because they are fast and easy to implement (Rennie, Shih, Teevan, & Karger, 2003). They are probabilistic classifiers which consider the probability $P\left(\hat{c}|\mathbf{d}\right)$ that a document, represented as a vector $\mathbf{d}$, is assigned to the class $\hat{c} \in \{\text{ yes, no }\}$. By Bayes' theorem, this probability can be expressed as:

$$P\left(\hat{c}|\mathbf{d}\right) = \frac{P\left(\hat{c}\right)P(\mathbf{d} \mid \hat{c})}{P(\mathbf{d})}$$

where $P(\hat{c})$ is the prior probability of class $\hat{c}$, and $\mathrm{P}(\mathbf{d} \mid \hat{c})$ is approximated from the corresponding relative frequencies measured in training data, following the 'naïve' assumption of conditional independence of the features.

For classification, the class maximizing (3) is selected; with respect to this maximization, the term $P(\mathbf{d})$ is constant and can hence be neglected. Due to the feature independence assumption, both types of NB models can be trained very efficiently even with a high number of features, and it has been shown that while the assumption of feature independence might be violated in real-life applications, there are many data sets in which strong dependencies exist among attributes, yet Naïve Bayes achieves high accuracy (Domingos & Pazzani, 1997); for a possible explanation of this behavior we refer to the study by Zhang (2004).

Besides, we use Support Vector Machines (SVM) (Vapnik, 1995) with linear kernel, trained with the Sequential Minimal Optimization (SMO) algorithm especially suited for the sparsity of our linguistic features, i.e., the situation where each feature has only few non-zero values among the instances (Platt, 1999). Linear support vector machines define a hyperplane which separates positive and negative instances in the vector space of features, while maximizing the margin, i.e., the distance between the hyperplane and the nearest positive example respectively negative example. Mathematically, if the 'yes' and 'no' classes are mapped to 1 and –1, SVM classify an instance $\mathbf{d}$ by means of:

$$\hat{c}\left(\mathbf{d}\right) = \operatorname{sgn}(\mathbf{w}^T\mathbf{d} + b)$$

where sgn is the sign function, and $\mathbf{w}$ and b can be interpreted as a weight vector for the individual features and the classifier bias, respectively. Both these parameters are learnt from training data. Linear SVM are popular for text classification since they are robust to overfitting to high dimensional input spaces as their complexity is not determined by the number of features, but on the separation of the training examples by the margin, and text classification problems are likely to be linearly separable (Joachims, 1998).

For each of the three classifiers, the hyperparameters of the linguistic feature extraction methods were optimized on the validation set. For all of the classifiers it turned out to be beneficial to convert all words to lower case, and remove stopwords (e.g., *the*, *for*, *but*) according to Weka's (Hall et al., 2009) built-in list of English stopwords. Since every classifier was able to deal with the resulting number of features, no periodic pruning or selection of the top words ranked by frequency is applied. For the NB classifier, the performance of simple binary features could not be further improved by modifications such as multiplication of with the IDF. As expected, MNB showed the highest performance for TFs instead of binary features. Furthermore, it appeared beneficial to transform the TFs to their logarithm, without measuring IDF; finally, normalization of the feature vectors to unity Euclidean length turned out to be advantageous. SVM surprisingly showed better effectiveness when only classifying by binary word presence. Moreover, in contrast to MNB, taking into account the IDF achieved better accuracy here. An overview over the combinations of classifier and linguistic features is given in Table 2.

Feature Selection

In text classification, besides removal of stopwords, often task-specific relevant features are selected. In this study, we apply two frequently used criteria to assess the relevance of features: the χ^2 statistic, measuring the statistical dependence between class labels and occurrence of terms, and information gain, quantifying the 'bits' of information obtained for the prediction of the class label by knowing whether a term occurs in a training instance (Gabrilovich & Markovitch, 2004; Yang & Liu, 1999; Yang & Pedersen, 1997). As these methods are supervised, i.e., they require class labels, feature selection criteria are evaluated on the training set only.

Results and Discussion

With the features parameterized as above and the default classifier hyperparameters defined in Weka, classifiers are trained on the union of

Table 2. Parameterization of linguistic feature extraction for each classifier, optimized on the validation set. Normalization refers to enforcing unity Euclidean length of feature vectors.

Classifier	Features	Normalization
NB	0/1	No
MNB	Log. TF	Yes
SVM	0/1 x IDF	Yes

training and validation set, and evaluated on the test set. For each of the tags the unweighted average recall (UAR) of the three different classifiers is measured. That is, the percentage of correctly classified instances (recall) is measured for both the 'yes' and 'no' classes and the unweighted average is taken. This measure is arguably better suited to imbalanced classification problems than conventional accuracy (Witten & Frank, 2011; Schuller et al., 2011): As the discretization threshold for the class labels is computed on the whole data set, the test set is not necessarily balanced for all tags; in our data set, the 'most imbalanced' tag in the test set is 'obnoxious', for which 153 instances exist in the 'no' and 128 in the 'yes' class. Table 3 shows the UAR of the binary classification tasks for the 14 tags.

Overall, the results are encouraging: We observe remarkable performances of up to 80.7% UAR for 'funny' (by MNB) and 80.2% UAR for 'fascinating' (by SVM). Furthermore, for all except three tag / classifier combinations, results are observed significantly above chance level UAR (50%, $p < .05$ according to a one-tailed z-test). As a rule of thumb, 57% UAR have to be surpassed to ensure significance, which is not given for 'confusing' (MNB and SVM) as well as 'obnoxious' (NB). Still, for each tag, there is at least one classifier that performs above chance.

Regarding the choice of the classifier, we observe that SVM outperform NB and MNB in six of the fourteen tags. MNB is observed most effective for five other tags while NB surpasses both MNB and SVM for three tags. Significant differences ($p < .05$) are, however, only encountered in two cases: For 'funny' and 'obnoxious', MNB outperforms NB by about 7% absolute UAR. This could indicate that TF instead of binary feature representation is particularly effective for these 'emotional' tags.

Conversely, examining the results among tags, it is evident that the effectiveness strongly varies. Particularly, the tag with the highest UAR and the one with the lowest UAR differ by more than 20% absolute UAR for any of the three classifiers. The highest average UAR across all classifiers is achieved for the tag 'beautiful', which delivered an average UAR of 77.6%. With 56.8% average UAR, the tag 'confusing' lagged behind all others. In general, it is obvious that categories with a positive meaning consistently lead to better classification effectiveness, as can be seen from the mean UAR across positively associated tags and all classifiers (74.9% UAR) as opposed to negatively associated or neutral ones (60.3% UAR). This surprising fact is however understandable when one considers that the classifiers for positive and negative tags are built on different amounts of data: Remember that either there is a clear bias of the users towards assigning positive tags—or, a lack of 'bad' lectures in general, cf. Table 1. Furthermore, tags such as 'confusing', 'unconvincing' or 'long-winded' arguably carry a high level (pragmatic) meaning that can hardly be captured by term frequency features in general. This phenomenon will be further discussed in the subsequent section.

Results using feature selection are shown exemplarily for the tag *beautiful* in Figure 1. In that case, feature selection can improve the results by up to 2% absolute, which is however

Table 3. Unweighted average recall (UAR) for each tag and for the three classifiers. The highest UAR per tag is typed in bold face. 'Mean' denotes average UAR across classifier, and average UAR across positive / negative tags.

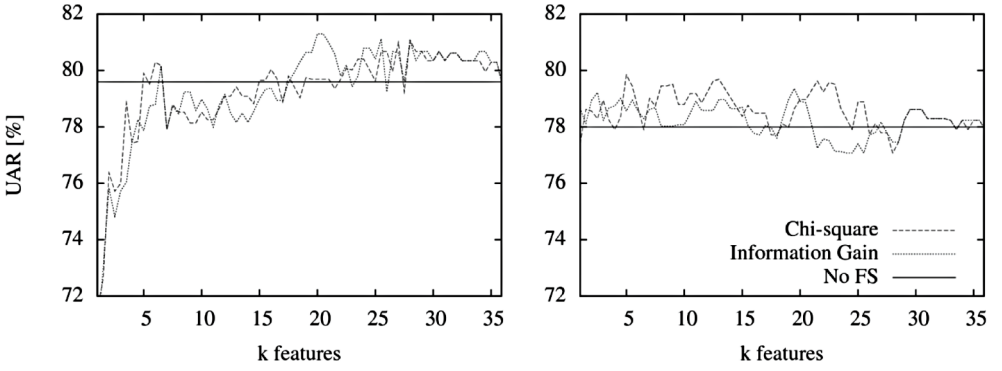
UAR [%]	Classifier			
Tag	**NB**	**MNB**	**SVM**	**Mean**
Positive				
Jaw-dropping	68.3	68.0	**71.6**	69.3
Funny	73.4	**80.7**	76.1	76.7
Courageous	73.3	77.5	**78.8**	76.5
Fascinating	76.1	76.3	**80.2**	77.5
Inspiring	69.4	**73.3**	71.7	71.5
Ingenious	74.1	**74.4**	73.7	74.0
Beautiful	75.2	78.0	**79.6**	77.6
Informative	72.4	75.7	**77.5**	75.2
Persuasive	73.3	76.2	**77.6**	75.7
			Mean	74.9
Neutral/Negative				
OK	**63.0**	60.9	61.1	61.7
Confusing	**57.8**	56.9	55.6	56.8
Unconvincing	60.2	**61.5**	61.0	60.9
Long-winded	**63.8**	61.1	63.2	62.7
Obnoxious	55.4	**63.0**	60.8	59.7
			Mean	60.3

not significant ($p > .05$) according to a one-tailed z-test. Furthermore, the optimal number of features for SVM classification is above 20 k. Similar results were obtained for the other tags; for none, significant UAR improvements could be obtained. This is in contrast to the behavior of feature selection reported in recent studies on sentiment analysis, e. g., (Schuller, 2011), where large performance gains could be obtained by feature selection, and the optimal number of features was observed at two orders of magnitude below; arguably, the task to predict user ratings is more multi-faceted than sentiment analysis, requiring a larger of number of features. Again, this motivates a closer look at relevance of individual features, as performed in the next section.

FEATURE RELEVANCE

In Table 4, we show for four selected tags the most discriminative words, corresponding to the binary word features with the highest positive or negative feature weight in an SVM classifier which was built on training and validation set TF features normalized to have the range [0, 1]. From the definition of linear SVM classification (cf. above), it follows that whenever a word corresponding to a feature with a positive weight occurs in a lecture, it will contribute to the classifier's decision to label the lecture as a 'yes' (1) instance, and vice versa, negative weights will foster a decision for the 'no' (-1) class. Hence, words corresponding to high absolute feature weights are most important for the model's decision between the two classes.

Figure 1. Feature selection: Unweighted average recall (UAR) for the Beautiful tag, with increasing numbers of features (total feature space: approx. 36 k features). No FS: No feature selection (cf. Table 3).



(a) Support Vector Machine (b) Multinomial Naïve Bayes

For the tag *fascinating* (80.2% UAR), it seems that that on the one hand music is likely tagged as 'fascinating', on the other hand it is striking that many of the words that are indicative for a positive ('yes') example can be connected with forms or appearance ('size', 'shape', 'dots', 'holes', 'patterns' and 'object'). One possible explanation could be that topics like nature, astronomy, architecture or art tend to be rated as fascinating by many people. Clearly, for the 'no' instances, words associated with society and economy are among the 'top ten', such as 'communities', 'campaign', 'management' or 'poverty'. Overall, this suggests a strong dependence of the rating on topic—it appears that the classifier learns an implicit model for topic classification. Next, when examining the most relevant words for the tag 'courageous' (78.8% UAR), results are more mixed: Among the positive feature weights, one notices a few topic-dependent ones such as 'political' and 'justice', but also words that are related directly or in a broader sense to 'courage', such as 'support' and 'fear' (the opposite of courage). Conversely, however, we observe that highly negative feature weights are given for words that can be considered 'non-political': music and wine.

Among the words indicative of 'persuasive' lectures (77.6% UAR), we find some that can be interpreted as being characteristic of the argumentation style—pointing out 'difference(s)', the 'average', or 'effective' (means)—but cannot be attributed to a single topic; neither can the words with negative feature weights which include 'photo', 'ultimate' or 'invisible'.

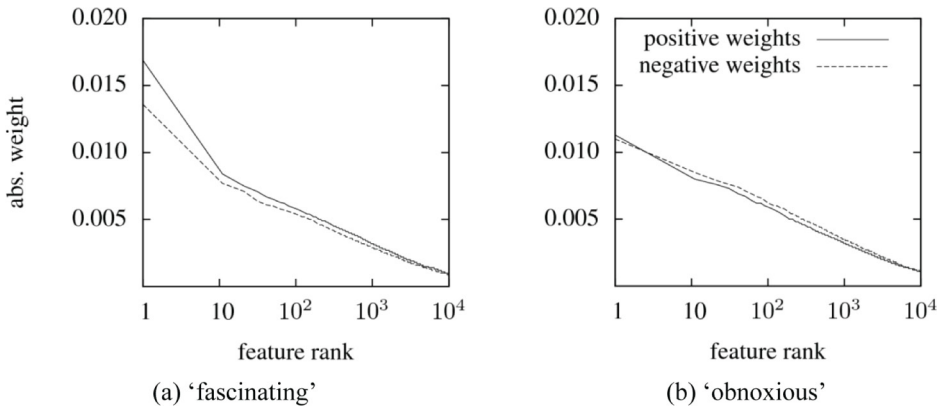
Finally, the results for the 'negative' tag 'unconvincing', which with 61.0% UAR lags behind the tags with positive meaning, suggest some overfitting to particularly 'convincing' or 'unconvincing' training examples: What strikes is the fact that numbers appear among the top ten words of the tag 'unconvincing' for both positive and negative feature weights. In general, as could be expected from the model performance, the results are much harder to interpret. For instance, the fact that the word 'technological' has a high weight for 'unconvincing' is surprising.

Generally, one should keep in mind that these ten words are just a fractional amount of all relevant features. In fact, the very high number of relevant features is an essential characteristic of text classification tasks (Joachims, 1998). To further shed light on the importance of the top-ranked features, the positive and negative feature weights in the order of their absolute values are shown in Figure 2, for the 'fascinating' and 'obnoxious' tags. We observe that weights decay rapidly for 'fascinating'

Table 4. The ten most discriminative words, in the order of their absolute feature weight in the SVM, for classifying into 'yes' (tag frequency above median) and 'no' instances

yes	**no**	**yes**	**no**
music	decline	fear	music
objects	communities	invited	wine
experiments	responsibility	political	web
surface	campaign	answers	fairly
size	father	women	blue
shape	solutions	support	pattern
patterns	conference	village	historical
evolution	initiative	woman	objects
dots	management	prepared	blocks
holes	poverty	justice	binary
(a) 'fascinating'		**(b) 'courageous'**	
yes	**no**	**yes**	**no**
average	built	music	16
groups	music	alternative	teachers
lose	beautiful	changing	school
difference	artificial	technological	dead
differences	ultimate	perception	training
country	june	marketing	forgot
staring	enter	decade	inspiring
effective	photo	truck	child
issue	invisible	35	jail
aids	motion	broader	impression
(c) 'persuasive'		**(d) 'unconvincing'**	

Figure 2. Absolute feature weights of the most relevant features in linear SVM classification



(a) 'fascinating' (b) 'obnoxious'

(which displays the highest UAR for SVM classification), indicating that the 'top' words are strongly indicative of the class, while weights are generally lower and decrease slower for 'obnoxious', which exhibits lowest performance in classification.

CONCLUSION AND FUTURE WORK

We have introduced a novel paradigm for textual affect classification: the automatic prediction of subjective user ratings in affective dimensions given to on-line lectures. These affective dimensions were deduced from the frequency of 14 prototypical adjectives (tags), assigned to a large database of transcripts of TED talks. Our results suggest that especially positive tags can be assigned robustly. Examinations of the classifier models reveal that for some tags, this might be due to implicit topic classification. Thus, future work could focus on multi-modal integration to combine linguistic features with acoustic and video features. Certainly, it could be believed that, e.g., incorporating prosodic anchors of charismatic speech (Rosenberg & Hirschberg, 2005) would benefit generalization of the models. Yet, first experiments with automatic classification based on the acoustic feature set of the 2011 Audio/Visual Emotion Challenge (Schuller et al., 2011) resulted in only 55.5% UAR across the 14 tags on the test set—interestingly, the maximum of 65.8% UAR was obtained for the tag 'long-winded'. Overall, this indicates that the methodologies for human affect recognition from speech cannot be transferred directly to the task at hand, in contrast to the linguistic features investigated in this article. Besides acoustics, useful video features that influence the audience's affect might include 'low level' global motion (optical flow) or histogram features, or 'higher level' features based on pre-classification, such as action units, gestures or body posture—the latter, however, might prove challenging to extract in real-life, web quality recordings such as the TED talks database.

Furthermore, having demonstrated the principal usefulness of linguistic features for the task at hand, we will investigate the effect of using ASR instead of ground truth transcripts in order to show how the proposed text classification methods could be applied in a real-life system. For affect recognition, it is well known that ASR inaccuracies can considerably impact performance (Wöllmer, Weninger, Steidl, Batliner, & Schuller, 2011). In that sense, the proposed evaluation database will serve as a challenging testbed for robust speech recognition algorithms.

REFERENCES

Aist, G., Kort, B., Reilly, R., Mostow, J., & Picard, R. (2002). Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, Pittsburgh, PA: IEEE.

Bhatt, K., Evens, M., & Argamon, S. (2004). Hedged responses and expressions of affect in human/human and human/computer tutorial interactions. In *Proceedings of Cognitive Science* (pp. 114–119). Chicago, IL: CogSci.

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, *29*(3), 241–250. doi:10.1080/1358165042000283101.

Domingos, P., & Pazzani, M. (1997). Beyond independence: Conditions for the optimality of the simple Bayesian classifier. *Machine Learning*, *29*, 103–130. doi:10.1023/A:1007413511361.

Forbes-Riley, K., & Litman, D. (2007). Investigating human tutor responses to student uncertainty for adaptive system development. In *Proceedings of Affective Computing and Intelligent Interaction* (pp. 678–689). Lisbon, Portugal: ACII. doi:10.1007/978-3-540-74889-2_59.

Forbes-Riley, K., & Litman, D. (2010). Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech & Language*, *25*(1), 105–126. doi:10.1016/j.csl.2009.12.002.

Forbes-Riley, K., & Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, *53*(9-10), 1115–1136. doi:10.1016/j.specom.2011.02.006.

Gabrilovich, E., & Markovitch, S. (2004). Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4.5. In *Proceedings of The Twenty-First International Conference on Machine Learning (ICML)* (pp. 321–328). Banff, Canada: AAAI.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, *11*(1), 10–18. doi:10.1145/1656274.1656278.

Huang, C.-C., Kuo, R., Chang, M., & Heh, J.-S. (2004). Fundamental analysis of emotion model for designing virtual learning companions. In *Proceedings of the 4th IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 326–330). Joensuu, Finland: IEEE.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Proceedings of the 10th European Conference on Machine Learning (ECML)* (pp. 137–142). Chemnitz, Germany: Springer.

Marlin, B. (2003). Modeling user rating profiles for collaborative filtering. In *Proceedings of Neural Information Processing Systems (NIPS)*. Vancouver, Canada: Neural Information Processing Systems Foundation.

Marovic, M., Mihokovic, M., Miksa, M., Pribil, S., & Tus, A. (2011). Automatic movie ratings prediction using machine learning. [Opatija, Croatia: IEEE.]. *Proceedings of MIPRO*, *2011*, 1640–1645.

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization* (pp. 41–48). AAAI Press.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multi-media learning. In D. L. H. Niegemann, & R. Brunken (Eds.), *Instructional design for multimedia learning* (pp. 181–195). Münster, Germany: Waxmann.

Parent, G., & Eskenazi, M. (2011). Speaking to the crowd: Looking at past achievements in using crowdsourcing for speech and predicting future challenges. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 3037–3040). Florence, Italy: ISCA.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods: Support vector learning* (pp. 185–208). Cambridge, MA: MIT Press.

Pour, P. A., Hussein, M. S., Al Zoubi, O., D'Mello, S., & Calvo, R. A. (2011). The impact of system feedback on learners' affective and physiological states. In *Intelligent Tutoring Systems* (Vol. 6094, pp. 264–273). Pittsburgh, PA: Springer. doi:10.1007/978-3-642-13388-6_31.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of International Conference on Machine Learning (ICML)* (pp. 616–623). Washington, DC: AAAI.

Rosenberg, A., & Hirschberg, J. (2005). Acoustic/prosodic and lexical correlates of charismatic speech. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 513–516). Lisbon, Portugal: ISCA.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. doi:10.1016/0306-4573(88)90021-0.

Schuller, B. (2011). Recognizing affect from linguistic information in 3D continuous space. *IEEE Transactions on Affective Computing*, *2*(4), 192–205. doi:10.1109/T-AFFC.2011.17.

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, *53*(9/10), 1062–1087. doi:10.1016/j.specom.2011.01.011.

Schuller, B., & Knaup, T. (2010). Learning and knowledge-based sentiment analysis in movie review key excerpts. In A. Esposito, A. M. Esposito, R. Martone, V. Müller, & G. Scarpetta (Eds.), *Toward autonomous, adaptive, and context-aware multimodal interfaces: Theoretical and practical issues* (Vol. 6456, pp. 448–472). Springer. doi:10.1007/978-3-642-18184-9_39.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., & Pantic, M. (2011). AVEC 2011 – The first international audio/visual emotion challenge. In *Proceedings First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011)* (pp. 415–424). Memphis, TN: Springer.

Schuller, B., & Weninger, F. (2012). Ten recent trends in computational paralinguistics. In A. Esposito, A. Vinciarelli, R. Hoffmann, & V. C. Müller (Eds.), *Proceedings of the 4th COST 2102 International Training School on Cognitive Behavioural Systems* (pp. 35-49). Springer.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. doi:10.1145/505282.505283.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)* (pp. 417–424).

Vapnik, C., & Cortes, V. (1995). Support vector networks. *Machine Learning*, *20*, 273–297. doi:10.1007/BF00994018.

Witten, I. H., & Frank, E. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Amsterdam, Netherlands: Elsevier.

Wöllmer, M., Weninger, F., Steidl, S., Batliner, A., & Schuller, B. (2011). Speech-based non-prototypical affect recognition for child-robot interaction in reverberated environments. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 3113–3116). Florence, Italy: ISCA.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR)* (p. 42-49). New York, NY: ACM.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 412–420). San Francisco, CA: AAAI.

Zhang, H. (2004). The optimality of naive Bayes. In *Proceedings of the Florida AI Research Society (FLAIRS)*. Miami, FL: AAAI.

Felix Weninger received his diploma in computer science (Dipl.-Inf. degree) from Technische Universität München (TUM), one of Germany's repeatedly highest ranked and among its first three Excellence Universities, in 2009. He is currently pursuing his PhD degree as a researcher in the Intelligent Audio Analysis Group at TUM's Institute for Human-Machine Communication, focusing his research on multi-source speech and audio recognition, including signal separation and robust back-ends for automatic speech recognition and paralinguistic information retrieval. Mr. Weninger is a member of the IEEE and (co-)authored more than 30 publications in peer reviewed books, journals and conference proceedings in the fields of speech and music signal processing, machine learning, and medical informatics.

Pascal Staudt received his BSc degree in Electrical Engineering and Information Technology from Technische Universität München for his study on text classification in affective dimensions. At the moment he is a graduate student in the Audio Communication and Technology program at Technische Universität Berlin. His research interests include audio synthesis, signal processing and multimedia content analysis.

Björn Schuller received his diploma in 1999 and his doctoral degree for his study on Automatic Speech and Emotion Recognition in 2006, both in electrical engineering and information technology from TUM. He is tenured as Senior Lecturer in Pattern Recognition and Speech Processing heading the Intelligent Audio Analysis Group at TUM's Institute for Human-Machine Communication since 2006. Best known are his works advancing Human-Computer-Interaction, Semantic Audio and Audiovisual Processing, Affective Computing, and Music Information Retrieval. Dr. Schuller is president-elect of the HUMAINE Association and member of the ACM, IEEE and ISCA, and (co-)authored 4 books and more than 270 publications in peer reviewed books (21), journals (39), and conference proceedings in the field of signal processing, and machine learning leading to more than 2,700 citations - his current H-index equals 28.