Language that Captivates the Audience: Predicting Affective Ratings of TED Talks in a Multi-Label Classification Task

Elma Kerz
RWTH Aachen University

Yu Qiao
RWTH Aachen University

Daniel Wiechmann
University of Amsterdam

elma.kerz@ifaar.rwth-aachen.de
yu.qiao@rwth-aachen.de
d.wiechmann@uva.nl

Abstract

The aim of the paper is twofold: (1) to automatically predict the ratings assigned by viewers to 14 categories available for TED talks in a multi-label classification task and (2) to determine what types of features drive classification accuracy for each of the categories. The focus is on features of language usage from five groups pertaining to syntactic complexity, lexical richness, register-based n-gram measures, information-theoretic measures and LIWC-style measures. We show that a Recurrent Neural Network classifier trained exclusively on within-text distributions of such features can reach relatively high levels of overall accuracy (69%) across the 14 categories. We find that features from two groups are strong predictors of the affective ratings across all categories and that there are distinct patterns of language usage for each rating category.

1 Introduction

The ability to communicate competently and effectively is key to personal contentment, academic achievement and professional career success. The ability to communicate competently is even found to enhance social relationships (Burleson, 2007; Morreale and Pearson, 2008). In educational and vocational contexts, competent speakers experience more success in sharing their knowledge, ideas and views (De Vries et al., 2010). One of the essential communication skills is that of giving an informative and impactful public speech. The development and mastery of public speaking is recognized as a core communicative competence (Schreiber et al., 2012) and has been incorporated into the educational curricula and standards for both first and second/foreign language (Common Core State Standards Initiative, 2010). Across various assessment grids and evaluation forms, a speech is considered competent and effective when the communicative intention is fulfilled – such as that of informing, persuading or entertaining an audience (and most speeches aim at achieving one or more of these) is achieved – when it is appropriate to the specific communicative context and when it matches the expectations of the audience. Given its central role, it is hardly surprising that research on public speaking has given rise to a vast field, scattered across disciplines using a range of methodological approaches (Backlund and Morreale, 2015). This research has been directed towards understanding the role of both verbal and non-verbal components in the assessment of the multidimensional construct of public speaking competence (Morreale, 2007). In more recent years, there has been a growing interest in automatic assessment of speaking skills. Most studies in this area have primarily focused on the role of auditory and acoustic measures of prosody (such as loudness, voice quality and pitch) and non-verbal cues (such as the use of gestures, eye-contact and posture) in predicting human ratings (see Section 2 on related work for more details).

The present paper contributes to and expands this emerging line of research by modeling public speaking skills in a multi-label classification task on the basis of five groups of features of language usage. More specifically, the aim of the paper is twofold: (1) to automatically predict the affective ratings of public speeches assigned by online viewers across fourteen rating

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 13–24
April 19, 2021. ©2021 Association for Computational Linguistics

categories and (2) to determine what types of features drive classification accuracy in predicting each of the categories. In pursuit of these aims, we use a large open repository of public speaking, TED Talks[1]. TED (Technology, Entertainment and Design) Talks are designed to not exceed the length of 18 minutes and provide succinct and enlightening insights on various topics or ideas that are "worth spreading." Topics presented in these talks range from global warming to what keeps us happy and healthy as we go through life. Most popular TED Talks, such as Brené Brown's "The Power of Vulnerability" has garnered almost 48.000.000 million views. TED presenters are often selected not only on the basis of their expertise on a given topic but also for their ability to effectively and succinctly communicate.

The dataset used in the paper included 2392 public speaking videos aligned with 5.89 million human ratings. Each TED speech is rated in terms of 14 categories: (1) beautiful, (2) confusing, (3) courageous, (4) fascinating, (5) funny, (6) informative, (7) ingenious, (8) inspiring, (9) jaw-dropping, (10) long-winded, (11) obnoxious, (12) OK, (13) persuasive and (14) unconvincing. The speech transcripts were automatically analyzed using CoCoGen, a computational tool that implements a sliding-window technique to compute a series of measurements for a given language feature (contours) that captures the distribution of that feature within a text. A Recurrent Neural Network (RNN) classifier – that exploits the sequential information in those contours – was trained on these distributions to predict 14 rating categories investigated in the paper. The remainder of the paper is organized as follows: Section 2 briefly presents a concise review of related work. Section 3 describes the TED talk corpus we used alongside with viewer ratings. Sections 4 presents the tool used for automated text analysis and the five groups of features used in the paper. Section 5 gives a description of the classification model architecture. Section 6 presents the main results and discusses them. Finally, Section 7 summarizes the main findings reported in the paper and proposes future research directions.

2 Related work

A combined use of automated text analysis of authentic language use in large corpora and machine learning techniques has received increasing interest in recent years. Such an approach has been successfully applied in various classification tasks, including detection of personality, gender, and age in the language of social media (Schwartz et al., 2013), Alzheimer's dementia detection in spontaneous speech (Luz et al., 2020), author identification and/or verification (Khanh and Vorobeva, 2020), fake news detection (Pérez-Rosas et al., 2017), etc. Most closely related to this paper is research focused on predicting human behavioral responses and human subjective/affective ratings/judgements through automated analysis of speaking samples. Several studies have used verbal and non-verbal cues in predicting audience laughter in humorous speeches (Chen and Lee, 2017), human ratings of politicians' speaking performance along multiple dimensions, such as expressiveness, monotonicity and persuasiveness (Scherer et al., 2012) or predicting performance assessments of students' oral presentation (Luzardo et al., 2014). For example, Luzardo et al. (2014) trained a binary classifier to predict the quality of the presentation. The accuracy of the trained binary classifier is 65% and 69% respectively for the features extracted from slides and audio track. In (Pfister and Robinson, 2011), real-time recognition of affective states and its application to the assessment of public speaking skills are proposed by using acoustic features. The skills are predicted with 61-81% accuracy. In another interesting application, (Weninger et al., 2012) analysed 143 online speeches hosted on YouTube to classify an individual as achievers, charismatic speakers, and team players with 72.5% accuracy on unseen data. The most related work to ours is Weninger et al. (2013), which predicted the affective ratings for online TED talks using lexical features, where online viewers assigned 3 out of 14 predefined rating categories that resulted in the affective state invoked in them listening to the talks. Their models reached average recall rates of 74.9 for positive categories (jaw-dropping, funny, courageous, fas-

[1] https://www.ted.com/

cinating, inspiring, ingenious, beautiful, informative, persuasive) and 60.3 for neutral or negative ones (OK, confusing, unconvincing, long-winded, obnoxious). In summary, as reviewed above, the available literature on automatic assessment and evaluation of speaking competencies based on human judgments or affective ratings have primarily drawn on audio features (prosody of the speech) and/or visual cues.

3 Data

We analysed the TedTalk data gathered on the ted.com website[2]. We crawled the site and obtained every TED Talk transcript and its metadata from 2006 through 2017, which yielded a total of 2668 talks. Viewers on the Internet can vote for three impression-related labels out of the 14 types of labels: beautiful, confusing, courageous, fascinating, funny, informative, ingenious, inspiring, jaw-dropping, longwinded, obnoxious, OK, persuasive, and unconvincing. The labels are not mutually exclusive and users can select up to three labels for each talk. If only a single label is chosen, it is counted three times. All talks that featured more than one speaker as well as talks that centered around music performances were removed. This resulted in a dataset of 2392 TED talks with a total number of views of 4139 million and a total number of 5.89 million ratings (see Table 1). All ratings were normalized per million views to account for differences in the amount of time that talks have been online. All ratings were binarized by their medians, such that each category has a value 1 when the rating of a text in this category was above or equal to the median and 0 if not.

4 Automated Text Analysis

The speech transcripts were automatically analyzed using CoCoGen, a computational tool that implements a sliding window technique to calculate within-text distributions of scores for a given language feature (for current applications of the tool in the context of text classification, see (Kerz et al., 2020; Qiao et al., 2020; Ströbel et al., 2020)). Here, in this paper, we employ a total of 119 features derived from multi-disciplinary integrated approaches to language (Christiansen and Chater, 2017) that fall into five categories: (1) measures of syntactic complexity, (2) measures of lexical richness, (3) register-based n-gram frequency measures, (4) information-theoretic measures, and (5) LIWC-style (Linguistic Inquiry and Word Count) measures. The first four categories of features are derived from the literature on language development showing that, in the course of their lifespan, humans learn to produce and understand complex syntactic structures, more sophisticated and diverse vocabulary and informationally denser language (Berman, 2007; Hartshorne and Germine, 2015; Ehret and Szmrecsanyi, 2019; Lu, 2010, 2012; Brysbaert et al., 2019). The inclusion of features in the fourth category is motivated by recent research on language adaptation (Chang et al., 2012) and research that looks at language from the perspective of complex adaptive systems (Beckner et al., 2009; Christiansen and Chater, 2016a) indicating that, based on accumulated language knowledge emerging from lifelong exposure to various types of language inputs, humans learn to adapt their language to meet the functional requirements of different communicative contexts. The fifth group of features is based on insights from many years of research conducted by Pennebaker and colleagues (Pennebaker et al., 2003; Tausczik and Pennebaker, 2010), showing that the words people use in their everyday life provide important psychological cues to their thought processes, emotional states, intentions, and motivations.

In contrast to the standard approach implemented in other software for automated text analysis that relies on aggregate scores representing the average value of a feature in a text, the sliding-window approach employed in CoCoGen tracks the distribution of the feature scores within a text. A sliding window can be conceived of as a window of size ws, which is defined by the number of sentences it contains. The window is moved across a text sentence-by-sentence, computing one value per window for a given indicator. The series of measurements generated by CoCoGen captures the progression of language performance within a text for a given indicator and is

[2] https://www.ted.com/

Table 1: Descriptive statistics for TED talk dataset (N=2392 talks); total of 5.89 million ratings from a total of 4162 million views. Descriptive statistics of rating scores are normalized per million views.

Meta		Mean	SD	Median	Min	Max
Duration (min)		13.62	5.29	14.14	2.25	44.63
Word Count		2044.56	885.54	2040	15	5800
Views		1730676.52	2540236.17	1139779.5	155895	47227110
Ratings	N ratings	Mean	SD	Median	Min	Max
Inspiring	1277876	534.23	1307.69	235	5	24924
Informative	862945	360.76	550.77	221.5	0	9787
Fascinating	758344	317.03	636.79	161.5	5	14447
Persuasive	544215	227.51	476.24	101	0	10704
Beautiful	444193	185.7	481.46	62	0	9437
Courageous	401316	167.77	437.11	53	0	8668
Funny	368139	153.9	603.88	20	0	19645
Ingenious	360870	150.87	285.87	68	0	6073
Jaw-dropping	344204	143.9	560.84	40	0	14728
OK	193184	80.76	90.16	55	0	1341
Unconvincing	127299	53.22	93.11	27	0	2194
Longwinded	78158	32.67	42.26	19	0	447
Obnoxious	60440	25.27	53.55	12	0	1361
Confusing	49815	20.83	31.88	12	0	531

referred here to as a 'complexity contour'. In general, for a text comprising n sentences, there are $w = n - ws + 1$ windows.[3] CoCoGen uses the Stanford CoreNLP suite (Manning et al., 2014) for performing tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic parsing (Probabilistic Context Free Grammar Parser (Klein and Manning, 2003)).

Table 2 provides a concise overview of the features used in this study. The first group consists of 18 features pertaining to syntactic complexity. These features are implemented based on descriptions in Lu (2010) and using the Tregex tree pattern matching tool (Levy and Andrew, 2006) with syntactic parse trees for extracting specific patterns. The second group subsumes 12 features pertaining to lexical richness: five measures of lexical variation, one measure of lexical density, seven measures of lexical sophistication. The operationalizations of these measures follow those described in Lu (2012) and Ströbel (2014). The third group includes 25 n-gram frequency features that are derived from the five register sub-components of the Contemporary Corpus of American English (COCA, (Davies, 2008)): spoken, magazine, fiction, news and academic language[4]. Our frequency-ngram measures differ from those used in the earlier studies reviewed in Section 2. Instead of using only bigrams and trigrams, we extend them to include longer word combinations (four- and five-grams) and use a more nuanced definition to operationalize the usage of such combinations given in equation (1):

$$\mathrm{Norm}_{n,s,r} = \frac{|C_{n,s,r}| \cdot \log \left[\prod_{c \in |C_{n,s,r}|} freq_{n,r}(c) \right]}{|U_{n,s}|} \tag{1}$$

Let $A_{n,s}$ be the list of n-grams ($n \in [1,5]$) appearing within a sentence s, $B_{n,r}$ the list of n-gram appearing in the n-gram frequency list of register r ($r \in \{\text{acad}, \text{fic}, \text{mag}, \text{news}, \text{spok}\}$) and $C_{n,s,r} = A_{n,s} \cap B_{n,r}$ the list of n-grams appearing

[3] Given the constraint that there has to be at least one window, a text has to comprise at least as many sentences at the ws is wide $n \geq w$.

[4] The Contemporary Corpus of American English is the largest genre-balanced corpus of American English, which at the time the measures were derived comprised of 560 million words.

Table 2: Overview of the 119 features investigated in the paper

Feature groups	Number of features	Sub-groups	Example/Description
Syntactic complexity	18	Length of production unit Subordination Coordination Particular structures	e.g. Mean length of clause e.g. Clauses per sentences e.g. Coordinate phrases per clause e.g. Complex nominals per clause
Lexical richness	12	Lexical density Lexical diversity Lexical sophistication	The number of lexical words divided by total number of words e.g. Type-token ratio e.g. Words from the New General Service List, see Browne et al. (2013)
Register-based	25	Spoken ($n \in [1,5]$) Fiction ($n \in [1,5]$) Magazine ($n \in [1,5]$) News ($n \in [1,5]$) Academic ($n \in [1,5]$)	Frequencies of uni-, bi- tri-, four-, five-grams from the five sub-components (genres) of the COCA, see Davies (2008)
Information theory	3	$\text{Kolmogorov}_{\text{Deflate}}$ $\text{Kolmogorov}_{\text{Deflate Syntactic}}$ $\text{Kolmogorov}_{\text{Deflate Morphological}}$	Measures use Deflate algorithm and relate size of compressed file to size of original file
LIWC-style	61	Linguistic dimensions Psychological processes Relativity Personal concerns	For a comprehensive description of LIWC features, see Pennebaker et al. (2015a)

both in s and the n-gram frequency list of register r. $U_{n,s}$ is defined as the list of unique n-gram in s, and $freq_{n,r}(a)$ the frequency of n-gram a according to the n-gram frequency list of register r.

A total of 25 measures results from the combination of (a) a 'reference list' containing the top 100,000 most frequent n-grams and their frequencies from one of five register subcomponents of the COCA corpus and (b) the size of the n-gram ($n \in [1,5]$). The fourth group includes three information-theoretic measures that are based on Kolmogorov complexity. These measures use the Deflate algorithm (Deutsch, 1996) to compress a text and obtain complexity scores by relating the size of the compressed file to the size of the original file (see (Ströbel, 2014) for the operationalization and implementation of these measures).

5 Classification Models

We used a Recurrent Neural Network (RNN) classifier, specifically a bidirectional dynamic RNN model with Long Short-term Memory (LSTM) cells. A dynamic RNN was chosen as it can handle sequences of variable length[5]. As shown in Figure 1, the input of the contour-based model is a sequence $X = (x_1, x_2, \ldots, x_l, x_{l+1}, \ldots, x_n)$, where x_i, the output of CoCoGen for the ith window of a document, is a 119 dimensional vector, l is the length of the sequence, $n \in \mathbb{Z}$ is a number, which is greater or equal to the length of the longest sequence in the dataset and $x_{l+1}, \cdots, x_n$ are padded $\mathbf{0}$-vectors. The input of the contour-based model is fed into a RNN which consists of 5 bidirectional LSTM layers with 400 hidden units in each cell. To predict the class of a sequence, we concatenate the hidden variable of the last LSTM

[5] The lengths of the feature vector sequences depends on the number of sentences of the texts in our corpus.

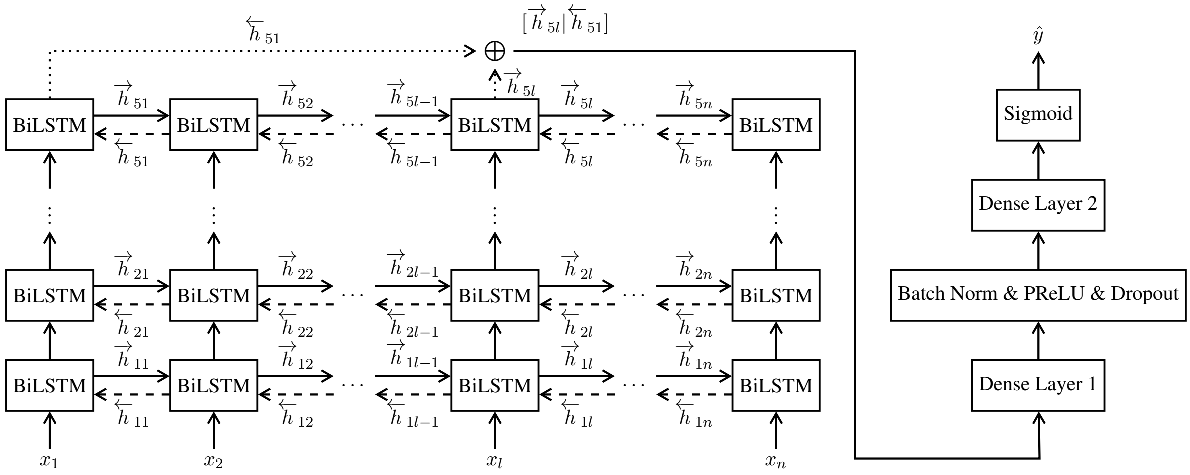
$\overleftarrow{h}_{51}$
$[\overrightarrow{h}_{5l}|\overleftarrow{h}_{51}]$
$\hat{y}$
Sigmoid
Dense Layer 2
Batch Norm & PReLU & Dropout
Dense Layer 1
BiLSTM
$\overrightarrow{h}_{51}$
$\overleftarrow{h}_{51}$
$\overrightarrow{h}_{52}$
$\overleftarrow{h}_{52}$
$\overrightarrow{h}_{5l-1}$
$\overleftarrow{h}_{5l-1}$
$\overrightarrow{h}_{5l}$
$\overleftarrow{h}_{5l}$
$\overrightarrow{h}_{5n}$
$\overleftarrow{h}_{5n}$
$\overrightarrow{h}_{21}$
$\overleftarrow{h}_{21}$
$\overrightarrow{h}_{22}$
$\overleftarrow{h}_{22}$
$\overrightarrow{h}_{2l-1}$
$\overleftarrow{h}_{2l-1}$
$\overrightarrow{h}_{2l}$
$\overleftarrow{h}_{2l}$
$\overrightarrow{h}_{2n}$
$\overleftarrow{h}_{2n}$
$\overrightarrow{h}_{11}$
$\overleftarrow{h}_{11}$
$\overrightarrow{h}_{12}$
$\overleftarrow{h}_{12}$
$\overrightarrow{h}_{1l-1}$
$\overleftarrow{h}_{1l-1}$
$\overrightarrow{h}_{1l}$
$\overleftarrow{h}_{1l}$
$\overrightarrow{h}_{1n}$
$\overleftarrow{h}_{1n}$
x_1
x_2
x_l
x_n

Figure 1: Roll-out of the RNN model based on complexity contours

cell in layer 5 $\overrightarrow{h}_{5l}$, i.e. the hidden variable of $5th$ RNN layer right after the feeding of x_l, with hidden variable of the last LSTM cell in the backward direction $\overrightarrow{h}_{51}$. The result vector of concatenation $[\overrightarrow{h}_{5l}|\overrightarrow{h}_{51}]$ is then transformed through a feed-forward neural network. The feed-forward neural-network consists of two fully connected layers (dense layer), whose output dimensions are 400, 14. Between the first and sencond fully connected layer, a Batch Normalization layer, a Parametric Rectifier Linear Unit (PReLU) layer and a dropout layer were added. Before the final output, a sigmoid layer was applied. The dataset was splitted into training and testing sets with a ration of 80:20 and 5-fold cross-validation is applied. As the loss function for training, binary cross entropy was used:

$$\mathcal{L}(\hat{Y},c) = -\frac{1}{C}\sum_{i=1}^{C}(y_i \log(\hat{y})+(1-y_i)\log(1-\hat{y}))$$

in which $c = (y_1, y_2, \ldots, y_N), C = 14$ is number of responses and $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$ is the output vector of the sigmoid layer rounded to closest integer. For optimization, we used Adamax with a learning rate $\eta = 0.0001$. The dropout rate of the RNN layers and the dropout layer is set to 0.5. The minibatch size is 32, which was shown as a reasonable value for modern GPUs (Masters and Luschi, 2018). All models were implemented using PyTorch (Pytorch, 2019).

Our baseline model has the same network architecture as the model described above, but instead of being trained on complexity contours, it was trained on sentence embeddings extracted from the Universal Sentence Encoder (USE) (Cer et al., 2018). USE takes a sentence as input and a 512-dimensional sentence representation as its output. The pretrained USE model we used was obtained from the Tensorflow (TF) Hub website and, according to TF Hub, the model was trained with a deep averaging network (DAN) encoder (Iyyer et al., 2015).

6 Results and Discussion

We report the results of multi-label classification with five-fold cross validation. The performance metrics of the RNN classification model (global accuracy, precision, recall and macro F1 scores) are presented in Table 3. The model achieved an average total accuracy of 69% averaged across the 14 rating categories. The highest accuracy was reached for the persuasive category (77%) and lowest for the long-winded category (62%). The results of RNN models trained on each feature set introduced in Section 5 are presented in Table 4. These results revealed that classification accuracy was mainly driven by LIWC-style and ngram-based features: The LIWC set was most predictive for 8/14 rating categories (persuasive, courageous, fascinating, inspiring, funny, ingenious, jaw-dropping and confusing), whereas the n-gram set ranked first in 3/10 rating categories (beautiful, unconvincing, obnoxious). In two rating categories (informative, OK) these two feature sets were equally predictive. Averaged across rating categories, the LIWC-based model achieved slightly better accuracy (+1%) at the cost of using 61 features (36 more features) compared to that of the n-gram based model trained on 25 features. RNN models trained on the lexical and syntactic features both reached 61% classification accuracy. The RNN trained on syntactic features reached the highest accuracy for the long-winded rating category (61%). The classifier based on the three information theoretic features achieved 56% accuracy. The finding that n-gram measures figure prominently in the classification is consistent with a growing body of studies indicating that n-gram measures are good predictors of human judgments/ratings of writing and speaking skills both in first and second language (Christiansen and Arnon, 2017; Garner et al., 2020; Saito, 2020). More generally, the findings are also in support of recent theoretical proposals emphasizing the importance of knowledge of such 'chunks' for human language processing to ameliorate the effects of the 'real-time' constraints on language processing imposed by the limitations of human sensory system and human memory in combination with the continual deluge of language input (cf., Christiansen and Chater, 2016, 2017, for the 'Now-or-Never bottleneck') . The result that LIWC-style features emerged as strong predictors is not surprising since previous research employing these measures has provided valuable insight into psychological processes and behavioral outcomes (Tausczik and Pennebaker, 2010; Pennebaker et al., 2015b). A closer examination of how individual features within each feature-group distinguished between higher-rated and lower-rated TED talks in a given rating category revealed some interesting patterns. For reasons of space, we focus on the results for the seven most frequently selected categories (*persuasive, courageous, fascinating, beautiful, inspiring, informative and funny*) for which classification accuracy was consistently greater than 70%. For each feature, we determined the differences between the group means for each indicator of language use ($M_{\text{higher rated}}$ - $M_{\text{lower rated}}$) for the *persuasive* category. For the *persuasive* rating category, the top LIWC-style features of highly rated talks concern words associated with words related to core drives and needs (*power*) and personal concerns (*risk, money and work*), while words related with perceptual processes (*see, hear, feel*) are associated with talks with lower ratings on the *persuasive* dimension. Highly persuasive talks are further characterized by a stronger reliance on higher frequency n-grams from more formal language registers: the top-five n-gram features associated with highly persuasive talks concern frequent 3-, 4- and 5-grams from the academic register and 3- and 4-grams from the news register. At the same time, frequent n-grams from the fiction register are indicative of lower persuasive talks. Regarding lexical richness, we found that highly-rated talks in the persuasive rating category exhibit high lexical diversity (CTTR, RTTR, NDW) in combination with high word prevalence and low lexical sophistication (NAWL, NGSL, ANC, BNC), indicating that persuasive talks use words that are widely known rather than those that are advanced and infrequent. At the syntactic level, persuasive talks show tendencies towards higher degrees of subordination (e.g. dependent clauses per T-unit) and phrasal complexity (complex nominals per T-unit) and lower degrees of coordination (coordinate phrases per clause). Figure 2 in the appendix shows the frequency with which features of particular types appeared in the top-five most associated (left) or top-five most dissociated features (right).[6] The figure discloses distinct patterns of language usage for particular rating categories: TED talks with high ratings in the categories *beautiful*, *inspiring*, *courageous* and *funny* are characterised by a strong reliance on frequent n-grams from the fiction register, words from the LIWC types *pronoun* and *social*, and relatively higher scores on indicators of lexical sophistication. At the same time, these talks exhibit relatively low proportions of frequent academic n-grams, low lexical diversity and shorter length of production units. Highly rated talks on the

[6] In case a top-five list involved a change in sign only features with positive values (for the associated feature list) or negative values (for the dissociated feature list) were included.

Table 3: Performance statistics of the RNN classifiers aggregated over five crossvalidation runs

	Baseline: RNN$_{\text{USE}}$		**RNN model based on complexity contours**							
	Acc	SD	Acc	SD	Precision	SD	Recall	SD	F1	SD
Overall	0.73	0.01	0.69	0.01	0.69	0.07	0.67	0.09	0.68	0.08
Persuasive	0.80	0.02	0.77	0.03	0.77	0.07	0.80	0.07	0.78	0.07
Courageous	0.81	0.03	0.76	0.03	0.80	0.05	0.69	0.09	0.74	0.07
Fascinating	0.82	0.03	0.75	0.03	0.75	0.06	0.78	0.06	0.76	0.06
Beautiful	0.79	0.02	0.74	0.02	0.77	0.06	0.66	0.05	0.71	0.05
Inspiring	0.79	0.02	0.73	0.02	0.76	0.02	0.65	0.18	0.70	0.03
Informative	0.78	0.02	0.73	0.04	0.71	0.13	0.82	0.11	0.76	0.12
Funny	0.76	0.02	0.71	0.02	0.71	0.04	0.70	0.08	0.70	0.05
Ingenious	0.77	0.02	0.70	0.02	0.68	0.04	0.75	0.05	0.71	0.05
Jaw-dropping	0.67	0.03	0.66	0.02	0.66	0.11	0.63	0.05	0.65	0.07
Unconvincing	0.67	0.01	0.65	0.03	0.66	0.04	0.58	0.18	0.62	0.07
OK	0.65	0.01	0.64	0.02	0.63	0.12	0.63	0.13	0.63	0.12
Confusing	0.67	0.03	0.63	0.03	0.63	0.04	0.56	0.23	0.59	0.06
Obnoxious	0.63	0.01	0.63	0.02	0.61	0.09	0.54	0.21	0.57	0.12
Longwinded	0.65	0.02	0.62	0.03	0.59	0.13	0.55	0.17	0.57	0.15

Table 4: Mean classification accuracy after five-fold cross validation with standard deviations across rating categories and feature-sets.

	All Features		LIWC		N-gram		Lexical		Syntactic		Inf. Theo.	
	(N = 119)		(N=61)		(N=25)		(N=13)		(N=18)		(N=3)	
Rating category	M_{Acc}	SD	M_{Acc}	SD	M_{Acc}	SD	M_{Acc}	SD	M_{Acc}	SD	M_{Acc}	SD
Overall	0.69	0.01	0.67	0.01	0.66	0.01	0.61	0.00	0.61	0.01	0.56	0.01
Persuasive	0.77	0.03	**0.75**	0.03	0.74	0.02	0.65	0.04	0.64	0.04	0.58	0.05
Courageous	0.76	0.03	**0.76**	0.02	0.74	0.02	0.65	0.01	0.62	0.01	0.53	0.04
Fascinating	0.75	0.03	**0.75**	0.03	0.72	0.03	0.67	0.02	0.63	0.03	0.55	0.03
Beautiful	0.74	0.02	0.71	0.01	**0.73**	0.01	0.63	0.02	0.63	0.03	0.56	0.04
Inspiring	0.73	0.02	**0.72**	0.02	0.69	0.02	0.65	0.01	0.65	0.01	0.54	0.02
Informative	0.73	0.04	**0.71**	0.03	**0.71**	0.03	0.64	0.03	0.67	0.04	0.61	0.05
Funny	0.71	0.02	**0.69**	0.03	0.68	0.02	0.60	0.03	0.64	0.02	0.60	0.02
Ingenious	0.70	0.02	**0.69**	0.02	0.66	0.01	0.64	0.02	0.60	0.02	0.53	0.02
Jaw-dropping	0.66	0.02	**0.62**	0.03	0.57	0.02	0.59	0.02	0.60	0.03	0.57	0.05
Unconvincing	0.65	0.03	0.60	0.01	**0.61**	0.03	0.57	0.02	0.56	0.04	0.53	0.05
OK	0.64	0.02	**0.59**	0.01	**0.59**	0.02	0.59	0.02	0.57	0.06	0.58	0.04
Confusing	0.63	0.03	**0.60**	0.02	0.58	0.02	0.56	0.02	0.57	0.02	0.51	0.02
Obnoxious	0.63	0.02	0.58	0.02	**0.61**	0.01	0.56	0.01	0.54	0.03	0.54	0.03
Longwinded	0.62	0.03	0.59	0.02	0.59	0.03	0.60	0.03	**0.61**	0.04	0.60	0.10

fascinating and *ingenious* dimensions group together and exhibit higher proportions of n-grams from the spoken register, higher proportions of words from the LIWC *function*-type – notably the personal pronoun *I* and are characterized by higher lexical sophistication. Finally, *persuasive*, *ingenious*, *fascinating* and *inspiring* talks display higher word prevalence score, which estimate the proportion of the population that knows a given word. This result indicates that the inclusion of this newly introduced crowd-sourced language metric (Johns et al., 2020) is a valuable addition to the existing features employed in research on automated assessment of spontaneous speech and calls for the development of additional crowd-sourced language features.

7 Conclusion

The ability to communicate competently and efficiently yields innumerable benefits across a range of social areas, including the enjoyment of congenial personal relationships, educational success, career advancement and, more generally, successful participation in the complex communicative environments of the 21st century. Public speaking is the epitome of spoken communication and, at the same time, the most feared form of communication. The paper contributes to the growing body of research that relies on automatic assessment of speaking skills and machine learning to better understand what makes a public speech effective. We performed a multi-label classification task to automatically predict human affective ratings associated with 14 categories on a large dataset of TED Talk speech transcripts. We demonstrated that a Recurrent Neural Network classifier trained exclusively on within-text distributions of 119 language features can reach relatively high levels of accuracy ($> 70\%$) on eight out of fourteen rating categories. We further showed that all rating categories are best predicted by (1) LIWC-style features, which counts words in psychologically meaningful categories, and (2) n-gram frequency measures, which reflect the use of register/genre-specific multiword sequences. Closer analysis of the distributions of particular language features disclosed distinct patterns of language usage for particular rating categories. More generally, the present paper responds to recent calls in the international scientific community of machine learning to use not only black box models but also explainable (white-box) models, since in any given application domain there is a need for both accurate but also understandable models (Rudin, 2019; Loyola-Gonzalez, 2019). In this paper, we have demonstrated in the domain of public speaking that models trained on human interpretable features in combination with deep learning classifiers can compete with black box models based on word embeddings.

References

Philip M Backlund and Sherwyn P Morreale. 2015. Communication competence: Historical synopsis, definitions, applications, and looking to the future. *Communication competence*, 22:11.

Clay Beckner, Richard Blythe, Joan Bybee, Morten H Christiansen, William Croft, Nick C Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, et al. 2009. Language is a complex adaptive system: Position paper. *Language learning*, 59:1–26.

Ruth A Berman. 2007. Developing linguistic knowledge and language use across adolescence. In M. Shatz E. Hoff, editor, *Blackwell handbooks of developmental psychology*, page 347–367. Blackwell Publishing.

Charles Browne et al. 2013. The new general service list: Celebrating 60 years of vocabulary learning. *The Language Teacher*, 37(4):13–16.

Marc Brysbaert, Paweł Mandera, Samantha F McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 english lemmas. *Behavior research methods*, 51(2):467–479.

Brant R Burleson. 2007. Constructivism: A general theory of communication skill. In W. Samter B. B. Whaley, editor, *Explaining communication: Contemporary theories and exemplars*, page 105–128. Lawrence Erlbaum Associates Publishers.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Franklin Chang, Marius Janciauskas, and Hartmut Fitz. 2012. Language adaptation and learning: Getting explicit about implicit learning. *Language and Linguistics Compass*, 6(5):259–278.

Lei Chen and Chong MIn Lee. 2017. Predicting audience's laughter using convolutional neural network. *arXiv preprint arXiv:1702.02584*.

Council of Chief State School Officers & National Governors Association. 2010. Common core state standards for english language arts and literacy in history/social studies, science, and technical subjects.

Morten H Christiansen and Inbal Arnon. 2017. More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, 9(3):542–551.

Morten H Christiansen and Nick Chater. 2016a. *Creating language: Integrating evolution, acquisition, and processing*. MIT Press.

Morten H Christiansen and Nick Chater. 2016b. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39.

Morten H. Christiansen and Nick Chater. 2017. Towards an integrated science of language. *Nature Human Behaviour*, 1(8).

Mark Davies. 2008. The Corpus of Contemporary American English (COCA): 560 million words, 1990-present.

Reinout E De Vries, Angelique Bakker-Pieper, and Wyneke Oostenveld. 2010. Leadership= communication? the relations of leaders' communication styles with leadership styles, knowledge sharing and leadership outcomes. *Journal of business and psychology*, 25(3):367–380.

Peter Deutsch. 1996. Deflate compressed data format specification version 1.3. *IETF RFC 1951*.

Katharina Ehret and Benedikt Szmrecsanyi. 2019. Compressing learner language: An information-theoretic measure of complexity in sla production data. *Second Language Research*, 35(1):23–45.

James Garner, Scott Crossley, and Kristopher Kyle. 2020. Beginning and intermediate l2 writer's use of n-grams: an association measures study. *International Review of Applied Linguistics in Language Teaching*, 58(1):51–74.

Joshua K Hartshorne and Laura T Germine. 2015. When does cognitive functioning peak? the asynchronous rise and fall of different cognitive abilities across the life span. *Psychological science*, 26(4):433–443.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 1681–1691.

Brendan T Johns, Melody Dye, and Michael N Jones. 2020. Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology*, 73(6):841–855.

Elma Kerz, Yu Qiao, Daniel Wiechmann, and Marcus Ströbel. 2020. Becoming linguistically mature: Modeling english and german children's writing development across school grades. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 65–74.

Bui Khanh and Alisa Vorobeva. 2020. A preliminary performance comparison of machine learning algorithms for web author identification of vietnamese online messages. In *2020 26th Conference of Open Innovations Association (FRUCT)*, pages 166–173. IEEE.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *LREC*, pages 2231–2234. Citeseer.

Octavio Loyola-Gonzalez. 2019. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *arXiv preprint arXiv:2004.06833*.

Gonzalo Luzardo, Bruno Guamán, Katherine Chiluiza, Jaime Castells, and Xavier Ochoa. 2014. Estimation of presentations skills based on slides and audio features. In *Proceedings of the 2014 acm workshop on multimodal learning analytics workshop and grand challenge*, pages 37–44.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Dominic Masters and Carlo Luschi. 2018. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*.

Sherwyn P Morreale. 2007. *The competent speaker speech evaluation form*. National Communication Association.

Sherwyn P Morreale and Judy C Pearson. 2008. Why communication education is important: The centrality of the discipline in the 21st century. *Communication Education*, 57(2):224–240.

James W Pennebaker, R.J. Booth, R.L. Boyd, and M.E Francis. 2015a. *Linguistic inquiry and word count: LIWC 2015. Operator's manual*. Pennebaker Conglomerates.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015b. The development and psychometric properties of liwc2015. Technical report.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Tomas Pfister and Peter Robinson. 2011. Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. *IEEE Transactions on Affective Computing*, 2(2):66–78.

Pytorch. 2019. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. https://github.com/pytorch/pytorch.

Yu Qiao, Daniel Wiechmann, and Elma Kerz. 2020. A language-based approach to fake news detection through interpretable features and brnn. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 14–31.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Kazuya Saito. 2020. Multi-or single-word units? the role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*, 70(2):548–588.

Stefan Scherer, Georg Layher, John Kane, Heiko Neumann, and Nick Campbell. 2012. An audio-visual political speech analysis incorporating eye-tracking and perception data. In *LREC*, pages 1114–1120.

Lisa M Schreiber, Gregory D Paul, and Lisa R Shibley. 2012. The development and test of the public speaking competence rubric. *Communication Education*, 61(3):205–233.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Marcus Ströbel. 2014. *Tracking complexity of l2 academic texts: A sliding-window approach*. Master thesis. RWTH Aachen University.

Marcus Ströbel, Elma Kerz, and Daniel Wiechmann. 2020. The relationship between first and second language writing: Investigating the effects of first language complexity on second language complexity in advanced stages of learning. *Language Learning*, 70(3):732–767.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Felix Weninger, Jarek Krajewski, Anton Batliner, and Björn Schuller. 2012. The voice of leadership: Models and performances of automatic analysis in online speeches. *IEEE Transactions on Affective Computing*, 3(4):496–508.

Felix Weninger, Pascal Staudt, and Björn Schuller. 2013. Words that fascinate the listener: Predicting affective ratings of on-line lectures. *International Journal of Distance Education Technologies (IJDET)*, 11(2):110–123.

A Appendix

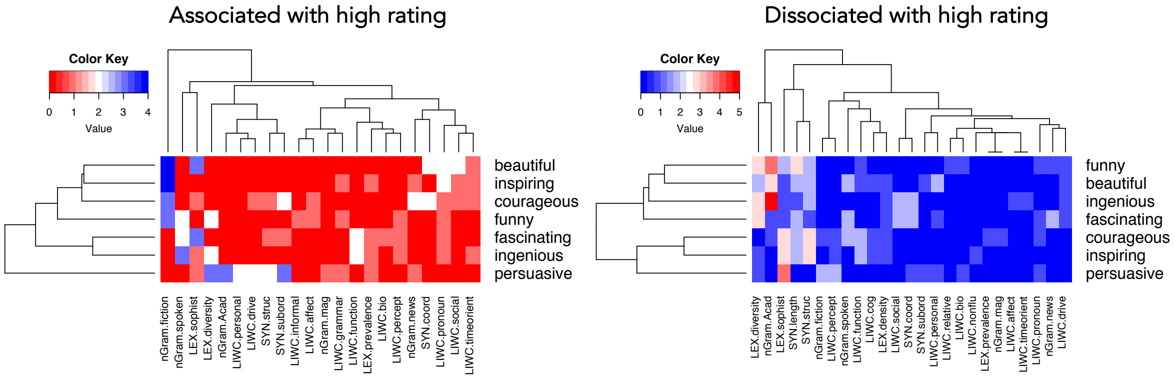

Figure 2: Heatmap of features associated (left; blue indicates higher counts) or dissociated (right, red indicates higher counts) with high ratings on each of the top-seven best predicted rating categories. Dendrograms represent Euclidean distances among rating categories and features respectively.