Words Worth: Verbal Content and Hirability Impressions in YouTube Video Resumes

Skanda Muralidhar
Idiap and EPFL
Switzerland
smuralidhar@idiap.ch

Laurent Son Nguyen
Idiap
Switzerland
lnguyen@idiap.ch

Daniel Gatica-Perez
Idiap and EPFL
Switzerland
gatica@idiap.ch

Abstract

Automatic hirability prediction from video resumes is gaining increasing attention in both psychology and computing.Most existing works have investigated hirability from the perspective of nonverbal behavior, with verbal content receiving little interest.In this study, we leverage the advances in deep-learning based text representation techniques (like word embedding) in natural language processing to investigate the relationship between verbal content and perceived hirability ratings.To this end, we use 292 conversational video resumes from YouTube, develop a computational framework to automatically extract various representations of verbal content, and evaluate them in a regression task.We obtain a best performance of $R^2 = 0.23$ using GloVe, and $R^2 = 0.22$ using Word2Vec representations for manual and automatically transcribed texts respectively.Our inference results indicate the feasibility of using deep learning based verbal content representation in inferring hirability scores from online conversational video resumes.

1 Introduction

First impressions play an important role in many social interactions, be it in personal life (like a first date) or in the professional contexts (like job interviews) (Ambady and Skowronski, 2008).Psychologists define first impressions as the “mental image formed about something or someone after a first meeting”.People form impressions about others’ attractiveness, personality, hirability or trustworthiness within a very short amount of time; nonverbal cues have been shown to play an important role in the formation of first impressions (Ambady and Rosenthal, 1992; Willis and Todorov, 2006).Despite the importance of verbal content and its relationship with various social constructs, it has been studied relatively rarely in comparison with nonverbal behavior.This work explores the relationship between verbal content and hirability impressions using a previously collected dataset consisting of noisy, real-world video resumes from YouTube (Nguyen and Gatica-Perez, 2016).

Literature in NLP and social computing have investigated the relation between verbal content and various social contrasts.In particular, Sinha et al. (Sinha et al., 2015) infered personality traits (HEXACO) of employees from Enterprise Social Media posts.Plank et al. (Plank and Hovy, 2015) collected a novel corpus of 1.2M English tweets annotated with Myers-Briggs personality type and reported the feasibility of using linguistic content from social media data to reliably predict some personality dimensions.Biel et al. (Biel et al., 2013), using 442 YouTube video blogs, investigated the relation between verbal content and personality impressions.The authors reported a performance of $R^2 = 0.31$ in inferring *Agreeableness* using manual transcriptions.

In the context of job interviews, literature has examined face-to-face interviews (Muralidhar and Gatica-Perez, 2017; Chen et al., 2016) and video interviews (Chen et al., 2017) to understand the relationship between verbal content and hirability impression.In this study, we investigate this relationship in the context of “in-the-wild”, real-world conversational video resumes.To the best of our knowledge, we are the first to utilize advances in natural language processing (Doc2Vec, Word2Vec, GloVe) to understand verbal behavior in this context.In particular, using a dataset of 292 YouTube video resumes, we address three research questions; (1)How can verbal content be represented to infer hirability impressions in video resumes?(2)What is the effect of automatic speech recognition (ASR) on inference performance compared to manual transcription? (3)What is the impact of video duration on inferring hirability impressions using verbal content?

Towards this goal, we develop a computational model to automatically extract various verbal representations from text corpus and evaluate their performance in a regression task.The contribution of this work are: (1)We transcribe 292 videos both manually and automatically; (2) We extract various representations of verbal content (Doc2Vec, Word2Vec and GloVe); (3) For manual transcription, we evaluate the various representations in an inference task and observe best inference performance for *Overall Hirability* ($R^2 = 0.23$) using GloVe; (4) We then assess the performance of automatic transcription versus manual and observe comparable inference performances, with $R^2 = 0.21$ for *Overall Hirability*; (5) We assess the difference in performance between automatic transcription of 2 minutes versus full video duration and observe that inference performance improve slightly with $R^2 = 0.22$ for *Overall Hirability*.

2 Dataset

2.1 YouTube Video Resume Dataset

In this work, we use a dataset previously collected by our group (Nguyen and Gatica-Perez, 2016).Nguyen et al. collected 939 videos using various keywords (like video resume, video cv etc), collected these videos from YouTube.Of these, we randomly selected a subset of 313 videos (i.e. $1/3$ of the data) as manual transcriptions is an expensive and time consuming process.Furthermore, of the 313 videos, 21 were discarded due to difficulty in transcription (due to music, accent of speakers) and missing annotations.Hence in this work, we use a corpus of 292 YouTube video resumes.

2.2 Annotations

The 292 videos were manually annotated for demographics and hirability impressions (on a $1-5$ Likert scale) by Amazon Mechanical Turkers (Nguyen and Gatica-Perez, 2016) with each video rated by at least 5 workers.We use Intraclass Correlation Coefficient (ICC) to measure inter rater agreement, a commonly used metric in psychology and social computing.ICC values were greater than 0.5 and is considered acceptable (Nguyen and Gatica-Perez, 2016).

2.3 Transcriptions

Manual Transcription: It was carried out by a native English speaker, who transcribed the videos as is (with no changes or corrections).As manual transcription is a tedious and expensive process, only the first 2 minutes were transcribed.These transcriptions constitute the "gold-standard" as they can be considered the output of an ideal, errorless ASR system.

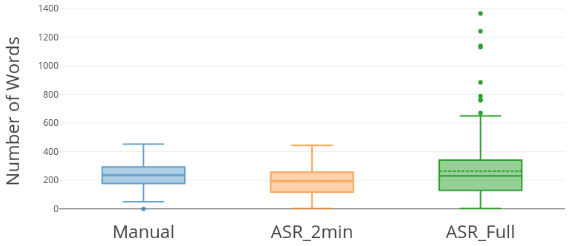

Figure 1: Box plot illustrating the distribution of number of words obtained by (a) manual transcription [Man] (b) ASR for first 2 minutes [ASR-2min] (c) ASR for full video [ASR-Full] for a random subset of 292 videos. The dotted line indicates the mean value.

Automatic Transcription: To address our research questions, we used an off-the-shelf ASR, Google Speech API (Cloud Services) for speech-to-text transcription.This API was selected as it is the best performing ASR system (Këpuska and Bohouta, 2017) and is readily available.Using Google Speech API, we generate two sets of transcriptions (a) first two minutes (to compare with manual transcription) (b) full video.Performance of the ASR was measured using word error rate (WER), and for this dataset was 41.5%.To put these results in perspective, Biel et al. reported an WER of 62.4% in their work (Biel et al., 2013) where the videos were comparable in terms of audio quality.Figure 1 shows the descriptive statistics of transcribed word count.

3 Method

Our methodology is illustrated in Figure 2. To obtain a feature representation of verbal content, we evaluated two distinct approaches: (a) representation at the document level; and (b) representation at the word level, followed by an aggregation step.For Doc2Vec and word-based representations, the text is pre-processed by converting them into lower case, removing the stop words, then stemmin and tokenizing.This was done using the Natural Language Toolkit (NLTK) python package (Bird et al., 2009).

3.1 Document-Based Representation

Linguistic Inquiry and Word Count (LIWC) is a software (Pennebaker and King, 1999) we use

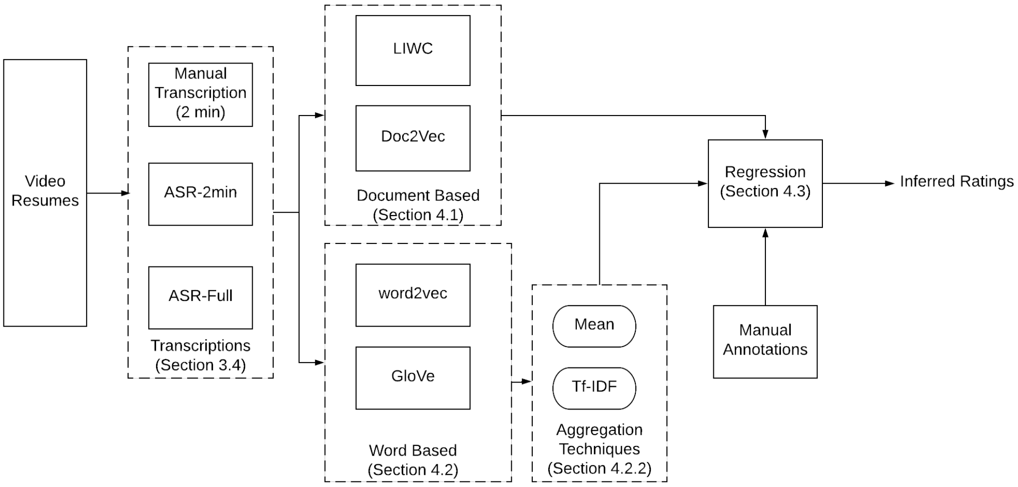

Figure 2: Overview of the work flow used in this study. The two classes of verbal content representation methods (a) document-based (b) word-based investigated is illustrated. For the document-based method, performance of LIWC and Doc2Vec in inferring hirability impressions is investigated. For the word-based method, all combinations of algorithm and aggregation techniques are investigated.

to extract lexical features.It computes these features by looking up each word in the transcript to the in-built English dictionary and is maps it to one of 70 categories.LIWC does not need text to be pre-processed and is a common text representation technique in computing literature (Muralidhar and Gatica-Perez, 2017; Biel et al., 2013).

Doc2Vec or paragraph vector was proposed by Le et al.(Le and Mikolov, 2014) to represent documents. After the text is pre-processed, we generate document vectors by training a model for word embedding using the Gensim package (Řehůřek and Sojka) in python. For the model generation, we use a constant learning rate for 10 epochs with 100 iterations and a vector of length 100. These numbers were empirically determined.

3.2 Word-Based Representations

For word-based representations, we use a two-step approach.First, word embedding from the transcripts are computed using pre-trained models (Word2Vec and GloVe).Next, these embedding are aggregated for a document level representation.

3.2.1 Word Representation

Word2Vec: developed by Mikolov et al., is an unsupervised learning algorithm that learns word embedding from a text corpus (Mikolov et al., 2013) with two models (a) continuous bag of words (CBOW) (b) continuous skip-gram (skip-gram).In both, the algorithm starts with a randomly initialized vectors and then learns the embedding by prediction.In this work, we use pre-trained CBOW model (300-dimensional) provided by Google which is trained on the Google News Dataset consisting of 100 billion words and a vocabulary of 3 million words (Mikolov et al., 2013).

GloVe: is a statistical method to learn word embedding developed by Pennington et al. (Pennington et al., 2014).This algorithm uses the global co-occurrence statistics, i.e count of word co-occurrences in a text corpus.In this work, we use GloVe with two different pre-trained models (both 300-dimensional vector) provided by the authors; (a) GloVe(S) trained on 6 billion words of Wikipedia (2014) with a vocabulary size of 400K words, and (b) GloVe(B) trained on a larger corpus of 840 billion words with a vocabulary of 2.2 million words.

3.2.2 Aggregation Techniques

In order to use Word2Vec and GloVe for representing documents (document embedding), various aggregation techniques were applied as not all words represent a sentence equally.The most common aggregation techniques are averaging and term frequency-inverse document frequency (TF-IDF). They have been shown in literature to work better than Doc2Vec for short sentences and small documents (Kenter et al., 2016; De Boom et al., 2016; Yih et al., 2011).

3.3 Regression

We outline our proposed computational framework for evaluating the research questions posed as a regression task. We define this task as inferring the impressions of hirability and soft skills using various verbal content representations. Towards this, we evaluate two regression techniques (Support Vector Machines regression (SVM-R) and Random Forest regression (RF)) implemented in the "scikit-learn" package for Python (Pedregosa et al., 2011). The hyper-parameters of the machine learning algorithms were optimized for best performance using 10-fold inner cross-validation (CV) and grid search, while the performance was assessed using the 100 independent runs of Leave-one-video-out CV. The performance of machine learning algorithms was evaluated using the coefficient of determination (R^2). We report the best performing algorithm only (RF).

4 Results and Discussion

4.1 RQ-1: Manual Transcriptions

Regression results using manual transcriptions are presented in Table 1.We observe that in an ideal case (i.e. using manual transcriptions) best inference performance for *Overall Hirability* is obtained using GloVe(S) with $R^2 = 0.23$.This implies that raters, at least partially, formed their hirability impressions based on verbal content.

In terms of inference performance, Doc2Vec consistently performs worse for all the hirability variables with $R^2 = 0.08$ (*Overall Hirability*) being highest.We hypothesize that this poor results could be the relatively short length of the documents (mean number of words = 232.67; min=50; max=453).As the performance is much lower than the other representation methods, thereon we will not discuss the results of Doc2Vec.Competitive results were obtained for LIWC features, with highest inference performance for *Professional* ($R^2 = 0.24$), followed by *Overall Hirability*, indicating that simple features like LIWC captures some of the variances in data.

Using GloVe(B) (Tf-Idf), best performance was obtained for *Overall Hirability* ($R^2 = 0.19$), while GloVe(B) (Avg) performed little lower, (highest for *Overall Hirability* with $R^2 = 0.14$).The GloVe(S) (Avg), performed best amongst all the representations for all hirability variables except *Professional*.The best performance was achieved for *Overall Hirability* ($R^2 = 0.23$) and lowest for *Professional* and *Social* ($R^2 = 0.17$).It is interesting to note that GloVe(S) performed better than GloVe(B) trained on a much larger data.

The Word2Vec representation performed better than LIWC features for *Overall Impression, Social* and *Communication*, but slightly lower for *Overall Hirability* and *Professional*.Word2Vec (TF-IDF) performed better then Word2Vec (Avg) for *Overall Impression* ($R^2 = 0.2$ and $R^2 = 0.18$) and *Professional* ($R^2 = 0.20$ and $R^2 = 0.16$).In the context of existing works, these results are better than those reported in the literature.Muralidhar et al., (Muralidhar and Gatica-Perez, 2017) using LIWC features extracted from 169 videos, reported an inference performance of $R^2 = 0.11$.

Using 1891 video interviews, Chen et al. (Chen et al., 2017) obtained *Precision* and *Recall* of 0.67 and 0.66 respectively in a classification task.The authors obtained the text corpus using ASR provided by IBM Bluemix platform and representation was achieved using Bag-of-Words (BoW).Nguyen et al., (Nguyen and Gatica-Perez, 2016) investigated the impact of nonverbal behavior in inferring first impressionn and reported a inference performance of $R^2 = 0.15$ for *Overall Hirability* ($N = 939$).

In summary, using manually transcribed text, GloVe(S) (Avg) achieves the best inference performance for *Overall Hirability*.Our results indicate the improved performance of word-based representations of verbal content in inferring hirability impressions, thus answering RQ-1.

4.2 RQ-2: Effect of Automatic Transcriptions

We observe that for ASR-2min corpus, the best inference performance (*Overall Hirability* with $R^2 = 0.21$) is obtained using Word2Vec (Table 1).We also observe that LIWC features extracted from Manual perform slightly better than those from ASR-2min for *Overall Hirability* ($R^2 = 0.20$ compared to $R^2 = 0.17$). Interestingly, GloVe(S) model, which performed best for Manual, does not perform as well for the ASR-2min corpus with best performance for *Overall Impression* ($R^2 = 0.14$). Similarly, GloVe(B) model performs worse than other models individually and in comparison with results from Manual for *Overall Impression* ($R^2 = 0.12$).

Word2Vec (TF-IDF) representation performs best using ASR-2min text corpus with *Professional* ($R^2 = 0.26$) and worse for *Social* ($R^2 = 0.13$).We observe that except for *Communication* and *Overall Impression*, use of ASR-2min performs slightly better than manual transcriptions.We hypothesize that this improvement could be due to Word2Vec, being a predictive model is less sensitive to ASR errors (WER) than

Table 1: Results of the inference task using the random forest algorithm (N=292) using manually transcribed (Manual), automatically transcribed (ASR-2min and ASR-Full) text corpus.The best performance is highligted in bold.

	Overall Impression			Overall Hirability			Professional Skills			Social Skills			Communication Skills		
	Manual	ASR-2min	ASR-Full	Manual	ASR-2min	ASR-Full	Manual	ASR-2min	ASR-Full	Manual	ASR-2min	ASR-Full	Manual	ASR-2min	ASR-Full
LIWC	0.13	0.13	0.14	**0.20**	0.17	0.19	**0.24**	0.18	0.20	0.07	0.11	0.09	0.13	0.17	0.20
Doc2Vec	0.03	0.01	0.01	0.08	0.04	0.03	0.03	0.01	0.06	0.03	0.0	0.0	0.05	0.02	0.06
Word2Vec															
- Avg	0.18	0.09	0.16	0.18	0.08	**0.26**	0.16	0.17	0.13	0.14	0.09	**0.22**	0.22	0.14	0.21
- Tf-Idf	0.20	0.18	0.16	0.17	**0.21**	0.16	0.20	**0.26**	0.10	0.10	0.13	0.19	0.22	0.19	0.14
Glove(S)															
- Avg	0.21	0.14	0.19	**0.23**	0.12	0.14	0.17	0.12	0.09	0.17	0.13	0.11	0.20	0.07	0.12
- Tf-Idf	0.15	0.14	0.20	**0.23**	0.14	0.18	0.15	0.11	0.14	0.12	0.15	0.10	0.15	0.09	0.16
Glove(B)															
- Avg	0.12	0.13	0.16	0.14	0.09	0.12	0.11	0.06	0.12	0.14	0.11	0.14	0.16	0.08	0.16
- Tf-Idf	0.16	0.12	0.11	**0.19**	0.11	0.08	0.13	0.07	0.09	0.13	0.10	0.12	0.15	0.09	0.13

GloVe.Biel et al. (Biel et al., 2013) investigated the use of manual and automatic transcription to infer personality impressions in YouTube video blogs.The authors reported a much lower performance using ASR ($R^2 = 0.18$) as compared to manual transcriptions $R^2 = 0.31$ for *Agreeableness*.This can be attributed to the high WER (62.4%) of the ASR system used (Hain et al., 2012) rather than the text representation methods.

In summary, the results indicate that the performance of ASR-2min is slightly lower compared to Manual (albeit with a different representation (Word2Vec)), and suggest the potential of using this approach (RQ-2).

4.3 RQ-3: Effect of Duration

The best performance using LIWC features extracted from ASR-Full text corpus (Table 1) was obtained for *Communication* and *Professional* ($R^2 = 0.20$), and lowest for *Social* ($R^2 = 0.09$).This seems to suggest that transcription of the extra duration of the videos improves inference performance.Word2Vec (Avg) performed better than Word2Vec (TF-IDF) method for all social variables with best performances for *Overal Hirability* ($R^2 = 0.26$) and worse for *Professional* ($R^2 = 0.13$).Using this representation method, ASR-Full out-performed the ASR-2min corpus for all variables except *Professional* ($R^2 = 0.13$)

Inference performance of GloVe(S)(TF-IDF) performed slightly better than GloVe(S)(Avg) for all variables (best performance for *Overall Impression* ($R^2 = 0.20$), worse for *Social* ($R^2 = 0.10$)) and is better compared to ASR-2min (except *Social*).The performance of GloVe(B) was lower than that of all other representations with best results for *Overall Impression* and *Communication*($R^2 = 0.16$).Although the performance of GloVe(B) method was lower than other word-based representations, these results are better than those obtained using ASR-2min.

Overall, these inference results tend to be comparable to those obtained using 2-min manual transcriptions (gold standard) and are higher than those reported using nonverbal cues (Nguyen and Gatica-Perez, 2016). We observe a moderate improvement in inference performance with full video duration transcribed for Word2Vec (Avg), thus answering RQ3.

5 Conclusion

This work investigated the relationship between verbal content and the formation of hirability impressions in conversational video resumes from YouTube. To this end, we use 292 video resumes previously collected by Nguyen et al. (Nguyen and Gatica-Perez, 2016).These videos were transcribed into text using manual and automatic (Google Speech API) transcriptions.Various text representations (word2vec, GloVe) were computed from both manual and automatic transcripts.We then investigated the effect of various document-based and word-based representations on inference performance in the two text corpora.

To conclude, we acknowledge that there are certain limitations to this work.Firstly, our experiments would benefit from having more data. In particular, this could help in experiments with doc2vec, which requires large amounts of data for accurate document representation.In future work, we will investigate the connect between verbal content and personality impressions as well as fuse other non-textual predictors.We will also analyze the impact of verbal content on the hirability impressions using the complete dataset (939 videos).

Acknowledgments

This work was partially funded by the Swiss National Science Foundation (SNSF) through the UBImpressed project. We thank Lesly Miculicich and Nikolaos Pappas (Idiap) for discussions.

References

Nalini Ambady and Robert Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2).

Nalini Ambady and John Joseph Skowronski. 2008. *First impressions*. Guilford Press.

Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. 2013. Hi youtube! personality impressions and verbal content in social video. In *Proc. 15th ACM ICMI*, pages 119–126. ACM.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Lei Chen, Gary Feng, Chee Wee Leong, Blair Lehman, Michelle Martin-Raugh, Harrison Kell, Chong Min Lee, and Su-Youn Yoon. 2016. Automated scoring of interview videos using doc2vec multimodal feature extraction paradigm. In *Proc. 18th ACM ICMI*, pages 161–168. ACM.

Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. 2017. Automated video interview judgment on a large-sized corpus collected online. In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*, pages 504–509. IEEE.

Google Cloud Services. Google Speech API.

Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. 2016. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156.

Thomas Hain, Lukáš Burget, John Dines, Philip N Garner, František Grézl, Asmaa El Hannani, Marijn Huijbregts, Martin Karafiat, Mike Lincoln, and Vincent Wan. 2012. Transcribing meetings with the amida systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):486–498.

Tom Kenter, Alexey Borisov, and Maarten de Rijke. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640*.

Veton Këpuska and Gamal Bohouta. 2017. Comparing speech recognition systems (microsoft api, google api and cmu sphinx). *Int. J. Eng. Res. Appl*, 7:20–24.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. 31st ICML*, pages 1188–1196.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Skanda Muralidhar and Daniel Gatica-Perez. 2017. Examining Linguistic Content and Skill Impression Structure for Job Interview Analytics in Hospitality. In *Proc. 16th ACM MUM*.

Laurent Son Nguyen and Daniel Gatica-Perez. 2016. Hirability in the wild: Analysis of online conversational video resumes. *IEEE Transactions on Multimedia*, 18(7):1422–1437.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *J. Personality and Social Psychology*, 77(6):1296.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitterorhow to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA.

Priyanka Sinha, Lipika Dey, Pabitra Mitra, and Anupam Basu. 2015. Mining hexaco personality traits from enterprise social media. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 140–147.

Janine Willis and Alexander Todorov. 2006. First impressions making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7):592–598.

Wen-tau Yih, Kristina Toutanova, John C Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 247–256. Association for Computational Linguistics.