Automatic Detection of Deceit in Verbal Communication

Rada Mihalcea
Computer Science and Engineering
University of Michigan
mihalcea@umich.edu

Verónica Pérez-Rosas
Computer Science and Engineering
University of North Texas
veronica.perezrosas@gmail.com

Mihai Burzo
Computer Science, Engineering, and Physics
University of Michigan - Flint
mburzo@umich.edu

ABSTRACT

This paper presents experiments in building a classifier for the automatic detection of deceit. Using a dataset of deceptive videos, we run several comparative evaluations focusing on the verbal component of these videos, with the goal of understanding the difference in deceit detection when using manual versus automatic transcriptions, as well as the difference between spoken and written lies. We show that using only the linguistic component of the deceptive videos, we can detect deception with accuracies in the range of 52-73%.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Natural Language Processing

Keywords

deception detection, speech transcription, crowdsourcing

1. INTRODUCTION

Work on deception has received significant attention from several fields of study, ranging from psychology to sociology, linguistics, and computer vision. A number of studies have been carried out in the past for the automatic detection of deception, focusing primarily on either physiological [20, 15] or psycholinguistics [11, 2] traits. More recent work has also considered the automatic detection of deceit from linguistic data, by using machine learning applied on either written [9, 12] or spoken [6, 14] deceptive statements. To our knowledge, however, there is no computational work that has specifically evaluated the effect on classifier accuracy when using written versus spoken deceptive statements, or, in the case of the latter statements, the role played by the quality of the transcriptions.

In this paper, we address the task of deception detection in verbal communication. Using a crowdsourced dataset consisting of 140 deceptive and truthful videos collected using the Amazon Mechanical Turk website, we analyse and compare the quality of a deception detection tool by using the verbal (linguistic) component of these statements, obtained through either manual or automatic transcriptions.

Specifically, our goal is to answer the following research questions. First, can we build an automatic deception detection tool by relying on the linguistic component of these deceptive videos, obtained through manual transcriptions? Second, is there a loss in accuracy when the manual transcriptions are replaced with automatic transcriptions? Finally, is there a significant difference between the accuracy of a deception detection system built from spoken statements as compared to written statements?

2. RELATED WORK

In psychology, it is worthwhile mentioning the study reported in [2], where more than 100 cues to deception are mentioned. Many of these cues involve speaker's behavior, including facial expressions and eye shifts, but linguistic cues such as word and phrase repetitions are also included. Pennebaker and colleagues [11] report on a psycholinguistic study, where they conduct a qualitative analysis of true and false stories by using word counting tools. Studies have also been made in neuroscience, such as the one reported in [10], where EEG signals are used to predict deceit.

Computational work applied to language includes the studies described in [19, 13], which included linguistic cues for deception detection in text-based and face-to-face communication. Deception in speech was also addressed in [6, 14], where the main focus was on acoustic and prosodic features. In more recent work, larger scale experiments were performed with the help of data collected through crowdsourcing. Mihalcea and Strapparava [9] proposed a method for the collection of linguistic deceptive datasets using crowdsourcing, and they showed through classification experiments that the data was indeed useful for this task. Similar to this, a review dataset was collected and used in experiments reported in [12], with the goal of identifying spam reviews.

There is also a significant amount of work carried out on the recognition of deception through visual clues. Some of the earliest work is due to Ekman [3], who studied how deception is expressed on the face and body. Following his work, a corpus of deceptive videos was compiled [5]. More recently, a number of studies have been made on the use of thermal imaging for deception detection, focusing primarily on the identification of thermal signatures for specific areas of the face [20, 15]. Deceptive expressions have also been studied through computer vision methods, which can separate between expressions of genuine and posed pain [7], facial expressions [18], brow movements [17], or smile [1, 16].

ICMI'13, December 9–13, 2013, Sydney, Australia
Copyright 2013 ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522888 .

3. DATASET

The dataset of deceptive videos was created using the Amazon Mechanical Turk service, which is a crowdsourcing platform provided by Amazon.com. A HIT (Human Intelligence Task) was set up on Mechanical Turk, in which workers were provided specific instructions about how to record and upload both deceptive and truthful videos. They were asked to look into the camera when recording the video, and speak clearly. They were also asked to avoid background noise and music, and had a clearly lit setting so that their face could be seen clearly. Finally, the length of the recording had to be between 1.5 - 3 minutes.

The workers had to produce and upload two videos. For the first video, the guidelines asked the contributor to think about their best friend, and record a video of them talking about the best friend. The video could include a description of their friendship, mentioning the reasons for which they were such good friends, including anecdotes or anything that seemed relevant to their relationship and what kept them together. Thus, the first video consisted of a truthful recording about a best friend. Next, for the second video, the workers were asked to think about a person they could not stand, and record a video of them talking about this person and describing him/her as though he/she were their best friend. Therefore, in this video they recorded a deceptive description of a (fake) best friend.

The final collection consists of 140 videos, out of which 70 include deceptive recordings, and 70 contain truthful recordings. The recordings were made by: 33 women and 37 men; 6 teenagers, 62 adults between 18-60 years of age, 2 adults over 60 years old; 51 Whites, 5 African Americans, 4 Hispanics, 10 Asians.

Transcriptions of the videos in these two collections were obtained using two approaches. First, we collected manual transcriptions by using again crowdsourcing via the Amazon Mechanical Turk. Second, we used automatic speech recognition to generate transcriptions for the videos.

3.1 Manual Transcriptions

A HIT (Human Intelligence Task) was set up on Mechanical Turk, in which workers were provided specific instructions about how to transcribe a video. The guidelines asked for complete, correctly spelled sentences, with punctuation included as needed. The workers were also asked to use filler words, such as "um," "like," "you know." For this task, we did not receive a significant amount of spam, perhaps due to the fact that this is a widespread task type, and there appears to be a skilled workforce on Mechanical Turk.

Nonetheless, the transcriptions were manually verified for correctness. We first used simple criteria to accept/reject the transcriptions, such as length (e.g., a transcription that has only one or two lines of text is clearly spam when the corresponding video has a length of 2 minutes). One of the authors has then further verified the quality of the transcriptions by checking for the presence of randomly selected utterances from the spoken review inside the transcription. The videos corresponding to those transcriptions that were rejected were returned to the site for another transcription.

3.2 Automatic Transcriptions

There are several speech recognition systems that are commercially or freely available, such as the Dragon Naturally Speaking tool,[1] or the CMU Sphinx toolkit.[2] However, most of these tools require a training step, and we did not have a training set for our data. We thus opted to use the Google automatic speech recognition engine, which is a ready to use resource available through the YouTube API.[3] We requested automatic transcriptions for our entire dataset, and we obtained captions in the SubRip text format. The API was unable to generate transcriptions for a few of our spoken reviews due to poor quality issues. Thus, after the transcription process, we ended up with a total of 114 transcribed files, consisting of 58 truthful and 56 deceptive statements, for an average of 1,394 characters per statement.

Table 1 shows sample segments of manual and automatic transcriptions. Table 2 shows the statistics over this dataset.

Source	Deceptive	Truthful
Video	112.08 sec.	115.68 sec.
Transcription		
Manual	252 words	292 words
Automatic	209 words	237 words

Table 2: Deception dataset statistics. Averages for video duration in seconds and transcription word counts

Metric	
Aligned words	33323
% WRR	42.4
% Substitutions	32.8
% Deletions	24.8
%Insertions	3.3

Table 3: Word recognition performance measures for automatic transcriptions

3.3 Performance Measures for Automatic Transcriptions

To evaluate the quality of the automatic transcriptions, we use the Sclite tool, which is a freeware resource distributed with the NIST SCTK Scoring Toolkit.[4] Sclite implements an alignment algorithm that evaluates the relation between an hypothesized text (HYP) and a reference (REF) text, and provides statistics such as word recognition rate (WRR), and the number of substitutions, deletions and insertions found while comparing the two sources. Table 3 shows the speech quality statistics for the automatic transcriptions. Since each statement is considered as a single sentence, we are not presenting the sentence recognition performance. As it can be observed in the table, the average word recognition rate of the speech recognition system is 42.4%, which can be partly explained by the recording settings (i.e., home recordings, surrounding environment noise).

4. LINGUISTIC FEATURES FOR DECEPTION DETECTION

Our goal in this paper is to perform comparative analyses of deception detection tools that can be derived from the linguistic component of deceptive video. We decided to focus on those features that were successfully used in the past for deception detection [9, 12, 11].

Specifically, we use a bag-of-words representation of the transcripts to derive unigram counts, which are then used as input features. First, we build a vocabulary consisting of all the words, including stopwords, occurring in the transcripts

[1] http://www.nuance.com/dragon

[2] http://cmusphinx.sourceforge.net/

[3] https://developers.google.com/youtube/

[4] http://www.itl.nist.gov/iad/mig//tools/

Truthful	Deceptive
Manual transcriptions	
OK. My best friend, um, name's Kaley. We have been friends since third grade. Um, we have a lot in common. We were both bullied as kids in school, so that's something that we connect on. Um, she is super hyper just a little bit hard to handle sometimes, but because of that makes me optimistic and makes me feel better during the day whenever she's very bubbly and excited about life and stuff.	My best friend, her name is Jill. Um, her and I have been friends since high school. Uh, we both were in choir together, we loved to sing together and do duets. Her and I had a lot in common in that aspect. Um, we both have a lot of health problems which brought us kind of closer together because we could relate to each other and understand um how to comfort one another in our hard times. Uh, both of us are in long term relationships. She's on the way to have a kid.
Automatic transcriptions	
my best friends have been friends since third grade million a lot in common we were both for joining us kids in school so net sentiment connect on timesheets progress just a little bit cardin jane doe sometimes bad because of that indecent options taking makes me feel better gained a whenever she's very violent kid i like his death	my best friend contains cell i have been friends since high school through the right-wing and i have loved to sing together indeed u␣s␣ headline common masback the boat have a lot of health problems guys come close to your isn't too early to each other roger stanton what's your comfort one another in our hard times this semester and most relationships sheets finally have a kid

Table 1: Manual and automatic transcriptions of sample truthful and deceptive statements.

of the training set. We then remove those words that have a frequency below 10 (value determined empirically on a small development set). The remaining words represent the unigram features, which are then associated with a value corresponding to the frequency of the unigram inside each video transcription.

5. EXPERIMENTS AND EVALUATIONS

Through our experiments, we address the three main research questions posed in the introduction.

5.1 Can we build an automatic deception detection tool by relying on the linguistic component of the deceptive videos?

To build the deception detection tool, we use linguistic features consisting of unigrams, as described in section 4. For the classification, we use the Support Vector Machines (SVM) and the Naive Bayes classifiers available in the Weka machine learning toolkit. For each experiment, a ten-fold cross validation is run on the entire dataset. Table 4 presents the accuracy results, also showing the effect of using a list of stopwords in the classifier or not.

	SVM	Naive Bayes
Unigrams	73.7%	64.0%
Unigrams, no stopwords	64.0%	63.2%

Table 4: Deception detection results for manual transcriptions.

The best classification accuracy is obtained by using an SVM classifier. The removal of stopwords appears to significantly impact the classification for the SVM classifier, although the loss in performance is less clear for the Naive Bayes. This may be explained by the fact that many of the deception cues observed in the past consist of function words (e.g., I, we), and thus their removal affects the accuracy of the classifier.

5.2 Is there a loss in accuracy when the manual transcriptions are replaced with automatic transcriptions?

Our next experiment consists of evaluating the performance of automatically transcribed statements in the deception detection task. We run experiments using the same set of features described above, and once again we use the SVM and the Naive Bayes classifiers from the Weka toolkit. The results obtained during these experiments are presented in Table 5.

	SVM	Naive Bayes
Unigrams	52.6%	58.8%
Unigrams, no stopwords	57.0%	60.5%

Table 5: Deception detection results for automatic transcriptions.

When using the automatic transcriptions, we observe a loss in accuracy between 5-20%, which is also explained by the high word error rate measured on these transcriptions. The SVM appears to be less robust to noise, with significantly higher losses in accuracy as compared to the Naive Bayes. Interestingly, the removal of stopwords increases the accuracy of both classifiers, which is an unexpected result. We hypothesize that the noise in the automatic transcriptions may have resulted in the over-generation of stopwords in statements where they did not belong, and therefore the removal of such stopwords had the effect of increasing the accuracy.

5.3 Is there a significant difference between sentiment analysis for spoken and written opinions?

Previous work has suggested that text extracted from spoken statements contains more spontaneous and richer emotional expressions than written statements and this may provide additional clues for paralinguistic tasks [8]. However, when working with transcriptions, additional challenges appear. For instance, differences in variable utterance lengths and disfluences such as hesitations (e.g. "uh", "um"), repetitions and corrections [4] introduce additional noise to the analysis, compared with "cleaner" text from written versions.

To explore the differences in deception detection when using written or spoken statements, we decided to empirically compare them using a machine learning approach. We used a subset of a dataset collected in our previous research [9]. The dataset consists of truthful and deceptive statements for the same scenario used in our work, but collected in a written format (i.e., participants were asked to type their statements). We used the same distribution as in our dataset, i.e., 58 truthful statements and 56 deceptive statements.

Table 6 presents the results obtained using the same linguistic features for the written statements. Once again, as it was the case with the manual transcriptions, the removal of stopwords results in a loss in performance, which supports our hypothesis that stopwords play an important role for deception detection from correct natural language statements. Comparing the results on written statements with those in Table 4, we can infer that spoken statements lead to equal or lower performance as compared to written statements, which implies that information verbally encoded in multimodal statements is less informative than the one available in written statements. One possible explanation for this phenomenon is the fact that when knowing that they are being observed, which is the case of video recordings, people tend to use additional resources such as gestures and intonations, which help them deliver their messages more accurately.

	SVM	Naive Bayes
Unigrams	73.7%	70.2%
Unigrams, no stopwords	71.9%	64.9%

Table 6: Deception detection results for written statements.

6. CONCLUSION

In this paper, we addressed the task of deception detection for deceptive videos, with a focus on the linguistic component of these videos. Using a crowdsourced dataset consisting of truthful and deceptive statements, we performed evaluations to: (1) determine the accuracy of a deception detection tool that can be built using only the verbal component of the videos; (2) measure the role played by the quality of the transcription (manual versus automatic) on the accuracy of the deception classifier; and (3) compare the performance obtained with written versus spoken statements. Our findings show that using only the linguistic component of deceptive videos, we can build a deception detection classifier with accuracies in the range of 52-73%. We also found that the quality of the transcription can have a significant impact on the deception detection tool, with losses in accuracy of up to 20% for automatic transcriptions as compared to manual transcriptions. Moreover, we found that written and spoken deceptive statements are different in nature, and that the verbal channel of the spoken statements appears to be less informative than the one in written ones. Interestingly, the use of stopwords was found useful for the cases where the statements are linguistically correct (i.e., manual transcriptions or written statements), but appear to harm the classifier in the case of noisy automatic transcriptions.

ACKNOWLEDGMENTS

We are grateful to Prof. Bill Buckles from the University of North Texas for allowing us to use the thermal camera. This material is based in part upon work supported by National Science Foundation award #1355633. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] J. F. Cohn and K. L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2, 2004.

[2] B. DePaulo, J. Lindsay, B. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological Bulletin*, 129(1):74—118, 2003.

[3] P. Ekman. Detecting deception from the body or face. *Journal of Personality and Social Psychology*, 29(3), 1974.

[4] S. Ezzat, N. Gayar, and M. Ghanem. Investigating analysis of speech content through text classification. In *Soft Computing and Pattern Recognition (SoCPaR), 2010 International Conference of*, pages 105 –110, dec. 2010.

[5] M. Frank, J. Movellan, M. Bartlett, and G. Littleworth. RU-FACS-1 database, Machine Perception Laboratory, U.C. San Diego, 2012.

[6] J. Hirschberg, S. Benus, J. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, B. Pellom, E. Shriberg, and A. Stolcke. Distinguishing deceptive from non-deceptive speech. In *Proceedings of INTERSPEECH-2005*, Lisbon, Portugal, 2005.

[7] G. Littlewort, M. Bartlett, and K. Lee. Faces of pain: automated measurement of spontaneous all facial expressions of genuine and posed pain. In *Proceedings of the International Conference on Multimodal Interaction*, 2007.

[8] F. Metze, T. Polzehl, and M. Wagner. Fusion of acoustic and linguistic features for emotion detection. In *Semantic Computing, 2009. ICSC '09. IEEE International Conference on*, pages 153 –160, sept. 2009.

[9] R. Mihalcea and C. Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the Association for Computational Linguistics (ACL 2009)*, Singapore, 2009.

[10] A. Mohammadian, V. Abootalebi, M. H. Moradi, and M. A. Khalilzadeh. Multimodal detection of deception using fusion of reaction time and p300 component. In *Biomedical Engineering Conference*, 2008.

[11] M. Newman, J. Pennebaker, D. Berry, and J. Richards. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29:665–675, 2003.

[12] M. Ott, Y. Choi, C. Cardie, and J. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Association for Computational Linguistics (ACL)*, 2011.

[13] T. Qin, J. K. Burgoon, J. P. Blair, and J. F. Nunamaker. Modality effects in deception detection and applications in automatic deception-detection. In *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.

[14] J. F. Torres, E. Moore, and E. Bryant. A study of Glottal waveform features for deceptive speech classification. In *International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[15] P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. Pavlidis, M. Frank, and P. Ekman. Imaging facial physiology for the detection of deceit. *International Journal of Computer Vision*, 71(2), 2006.

[16] M. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, 2007.

[17] M. Valstar, M. Pantic, Z. Ambadar, and J. Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the ACM International Conference on Multimodal Interfaces*, 2006.

[18] Z. Zhang, V. Singh, T. E. Slowe, S. Tulyakov, and V. Govindaraju. Real-time automatic deceit detection from involuntary facial expressions. In *Conference on Computer Vision and Pattern Recognition*, 2007.

[19] L. Zhou, J. Burgoon, J. Nunamaker, and D. Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, 13:81–106, 2004.

[20] Z. Zhu, P. Tsiamyrtzis, and I. Pavlidis. Forehead thermal signature extraction in lie detection. In *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, France, 2007.