Explanation Versus Prediction: Statistical Differences in Detecting Fraudulent Events Do Not Necessarily Have Predictive Power

Angelica M. Tinga (A.M.Tinga@uvt.nl)

Welmoed Kuperus (Welmoed.Kuperus@gmail.com)

Maira B. Carvalho (M.BrandaoCarvalho@uvt.nl)

Max M. Louwerse (M.M.Louwerse@uvt.nl)

Tilburg University, Department of Cognitive Science and Artificial Intelligence, Tilburg, The Netherlands

Abstract

A large body of research in the cognitive sciences relies on examining statistical differences. While the approach of examining differences can aid in explaining behavior, it does not necessarily mean that these differences have predictive power. Yet, understanding behavior both involves explaining and predicting behavior. As a point in case, the current study used a naturalistic email dataset to examine statistical differences and predictive power in fraudulent activities. Differences between 1st and 3rd person pronoun use in liars and people telling the truth are widely reported in the literature. The current study aimed to test for the effect of fraudulent events on pronoun use in emails using the Enron corpus and additionally applied a machine learning approach to estimate whether pronoun use predicts fraud. While the ratio between 1st and 3rd person pronoun use was related to fraud, this construct did not have predictive power. The current study highlights an important conclusion for the cognitive sciences: The importance of not only testing for differences, but of also applying predictive models. In this way it can be determined whether effects of a construct on an outcome can also predict the outcome.

Keywords: corpus linguistics; machine learning; deception; pronouns

Introduction

Many studies in the cognitive sciences rely on examining statistical differences. This approach provides us with important knowledge about differences in for example behavior between extroverts and introverts (Lu & Hsiao, 2010), clinical populations and non-clinical ones (Garnefski et al., 2002) and males and females (Bleidorn et al., 2016). While examining differences can aid in explaining behavior, it does not necessarily mean these differences have predictive power. Understanding behavior both involves explaining and predicting behavior (Rosenberg et al., 2018). A model focused on explanation could be appealing theoretically, but could be very limited in predicting actual human behavior (Yarkoni & Westfall, 2017).

One field of study in which differences have been widely examined is that of deception, in which comparisons are made between when people are lying and when they are telling the truth (DePaulo et al., 2003). Lying is cognitively more complex than telling the truth. To make a lie convincing we have to exert a lot of cognitive control, which might paradoxically be reflected in cues that betray our deception (Zuckerman et al., 1981), both verbally and non-verbally (DePaulo et al., 2003).

Several studies have examined these cues to deception using experimental manipulations, for example by asking participants to lie or to tell the truth, testing for a statistical difference between the manipulations. These studies demonstrated that there is a difference between liars and people that tell the truth: Liars provide fewer details and tell fewer compelling stories, as they are uncertain and less engaged (DePaulo et al., 2003). Liars apparently try to distance themselves from the content of the communication, with content increasing in abstractness (Louwerse et al., 2010). Abstractness in communication may be reflected in pronoun use (Hancock et al., 2008; Humpherys et al., 2011; Louwerse et al., 2010; Newman et al., 2003), with a decrease in self-references and an increase in other-references reflecting increasing abstractness. Even though experienced liars may be avoiding tainted words that reveal their intentions, pronoun use is outside of conscious control of speakers and writers and therefore a useful measure to determine whether statements are truthful or not (Pennebaker, 2011).

Newman et al. (2003) examined 1st person pronouns (self-references) and 3rd person pronouns (other-references) when participants were instructed to produce a story on abortion which matched their opinion or not. They demonstrated that participants who wrote a story they did not agree with used fewer 1st person singular and fewer 3rd person pronouns than participants that agreed with their story. Similarly, Hancock et al. (2008) asked participants to either write a truthful or untruthful story on several different topics. The participants that were untruthful used fewer 1st person pronouns and more 3rd person pronouns than truthful participants. A meta-analysis on 116 studies on lying and deceptive cues by DePaulo et al. (2003) also demonstrated that there is an effect of being truthful on pronoun use, with fewer self-references and more other-references showing up in liars. However,

Louwerse et al. (2010) found that fraudulent events were associated mostly with an increase in 1st person pronouns.

In sum, statistical differences between pronoun use in liars and people that are truthful have been established in the existing literature, although the direction of these effects varied across studies. Perhaps this is not entirely surprising, as the context and the ecological validity of these studies also differ. Most studies on pronoun use and deception induced lying with an experimental manipulation, by for instance asking participants to write about an opinion opposite to what they truly belief. These cases are considered to be deception (Newman et al., 2003), except that there is no consequence to participants' 'lying'. Such laboratory studies provide excellent insights in linguistic deceptive cues but lack ecological validity.

To use a case where the stakes of deception were higher than a manipulated laboratory setting, Louwerse et al. (2010), used an email dataset (Klimt & Yang, 2004), which contained 517,431 emails from about 150 Enron executives and employees from 1999 to 2001. The Enron Cooperation was one of the world's leading gas, electricity, and communication companies and is most famous for the elaborate and systematic way in which accounting fraud spread throughout the organization, which led to declaration of bankruptcy in 2001. The advantage of using this corpus is that, besides its ecological validity, it covers a relatively large time span and it has detailed information available on the company and its fraudulent activities (Diesner et al., 2005). The disadvantage of using a naturalistic corpus, however, is that it is very difficult to determine which emails actually contain deception and which ones do not. Louwerse et al. (2010) operationalized deception by identifying the periods during which fraudulent events took place, capitalizing on the sheer number of emails in these different time frames.

Although statistical differences show up between liars and non-liars in pronoun use and although this difference is theoretically making sense, it is not clear whether pronoun use allows for predicting deception, and if so, to what extent. The current study uses Louwerse et al. (2010) as an illustration. We used the Enron email dataset, but rather than only investigating whether there is an effect of fraudulent emails on linguistic variables as in Louwerse et al. (2010), we additionally applied a machine learning approach to estimate whether linguistic variables also predicted fraud. Moreover, rather than taking a large number of linguistic variables, we applied the principle of parsimony and only focused on pronoun use.

Table 1: Overview of events within the Enron Cooperation from 2000-2001. Marked events are considered fraudulent. Adapted from Louwerse et al. (2010) p. 964.

Event	Description of event	Date (month-year)
- Layoffs	Employees within Enron Corporation were laid off.	12-01
- CEO	Indicating involvement of the CEO within any coded event.	3-00, 08-00, 11-00, 01-01, 04-01, 08-01, 10-01, 11-01
- Fraudulent paperwork filled signed	Filing and/or signing of fraudulent paperwork (by the CEO or COO).	03-00, 08-00
- Fraudulent comments	Enron made fraudulent comments, to the employees and/or investors.	01-01, 09-01
- Discussion of ethics	A discussion of ethics occurred between Enron executives or between the CEO and employees.	07-00, 03-01, 05-01, 08-01, 09-01, 10-01
- Selling Enron shares	Selling of Enron stock by high-level executives occurs.	11-00, 05-01, 06-01, 07-01, 08-01, 09-01
- Rolling blackouts initiated	Intentional initiation of rolling blackouts in California.	01-01
- Meetings with national political figures	High-level Enron executives met with national political figures including the Secr. of the Treasury and the Secr. of Commerce.	02-01, 03-01, 04-01, 08-01, 10-01, 11-01
- Financial support of political candidate	High-level Enron executives (CEO & President) provided financial support for a newly elected national political figure.	01-01
- Profit announced	Profits were announced for the quarter.	04-01
- Loss announced	Losses were announced for the quarter.	10-01
- SEC inquiry developments	Beginning of the SEC inquiry and the point at which the SEC inquiry became a formal investigation.	10-01
- Shredding occurs	Shredding of Enron documents in Enron and/or Arthur Andersen accounting firm.	10-01
Shredding stopped	Shredding of Enron documents stopped in Enron and/or Arthur Andersen.	10-01, 11-01
- Fraud announced	Enron admitted to having overstated the company's profits.	11-01
- Bankruptcy filed	Bankruptcy was filed.	12-01

Methods

Selection and Classification of Data

Only emails sent by Enron employees were selected to filter the data from noise, such as advertisements, promotions, and other junk mail. Accordingly, we excluded duplicate emails and emails from other organizations (number of emails excluded in this step: 486,272). Next, we used the Interquartile Range rule for outlier removal: emails that had length above 1.5 times the Interquartile Range were excluded (number of emails excluded in this step: 2,157). This was necessary as some emails included the quoted replies from previous emails, thus providing redundancy. Finally, since our objective was to explore pronoun use, specifically the relationship between the use of different types of pronouns, we excluded emails that did not have at least one 1st person pronoun and one 3rd person pronoun (number of emails excluded in this step: 19,523). In summary, out of the 517,431 emails in the Enron dataset, 9,479 emails (1.83%) were included for further analyses.

Based on Louwerse et al. (2010), 16 types of events within the Enron Corporation from 2000-2001 were identified based on the timeline of the Enron case (Table 1). The event types 'fraudulent paperwork filed signed', 'fraudulent comments', 'rolling blackouts initiated' and 'shredding of documents' were identified as clearly fraudulent. Additionally, as Enron admitted to having overstated the company's profit, the events of 'profit announced' and 'loss announced' were also considered fraudulent. All events considered fraudulent are marked in gray in Table 1.

The dataset primarily consisted of emails from higher executives, increasing the probability that the content of the emails involved decision-making processes related to the fraudulent events. Emails sent during those activities that were sent in periods of fraudulent events were labeled as fraudulent. All other emails were labeled as non-fraudulent. Obviously, this is an overgeneralization, but a useful one given the illustrative purposes of the current study examining significant differences and predictive power.

A total of 28.1% (N = 2,664) of the 9,479 included emails was classified as being related to fraudulent events (compared to 71.9% [N = 6,815] being not related to fraudulent events). An overview of the distribution of normalized 1st and 3rd person pronouns in fraudulent and non-fraudulent emails is depicted in Figure 1.

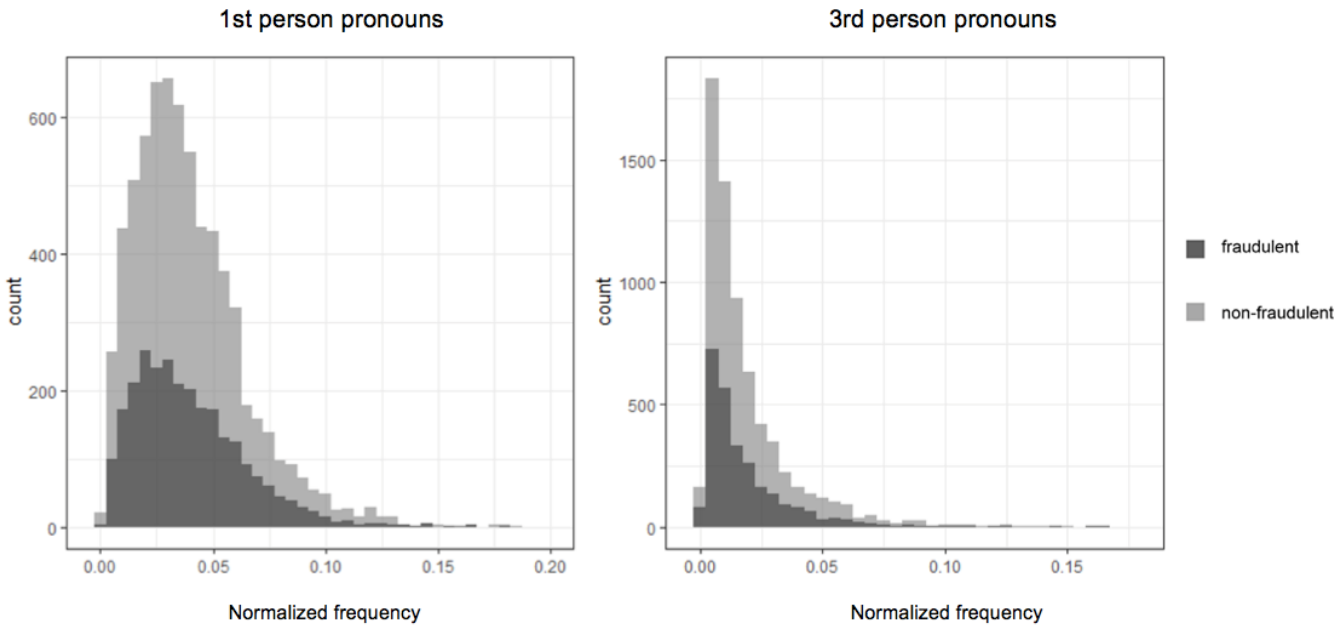

Figure 1: Distribution of normalized 1st and 3rd person pronouns in fraudulent (dark gray) and non-fraudulent (light gray) emails.

Feature Extraction

The number of occurrences of 1st person and 3rd person pronouns were computed for each of the 9,479 emails (see Table 2 for an overview of included pronouns). These occurrences were then normalized by the number of word tokens in each email.

Table 2: Included 1st and 3rd person pronouns.

Type of pronouns	Included pronouns
1st person	*I, me, my, mine, myself, we, us, our, ours, ourselves*
3rd person	*he, she, him, her, his, hers, himself, herself, they, them, their, theirs, themselves*

Data Analysis

Relationship Pronoun Use and Fraudulent Events The relationship between 1st person and 3rd person pronoun use frequency and fraudulent events was examined by computing two logistic regression models.

In Model 1, we used the (normalized) frequency of 1st person pronouns and the (normalized) frequency of 3rd person pronouns as independent variables, and the class label (fraudulent/non-fraudulent) as dependent variable. In Model 2, we computed whether the ratio between 1st person and 3rd person pronoun use had any relationship to the class label. Both models were fitted using a maximum likelihood estimator, using the Python package StatsModels.

Predicting Fraudulent Events Through Pronoun Use In order to predict whether an email was related to fraudulent events, we used two logistic regression classifiers. The classifiers were fitted using the same features as the logistic regression models; specifically, Classifier 1 was trained on the (normalized) frequency of 1st person pronouns and (normalized) frequency of 3rd person pronouns, while Classifier 2 was trained on the ratio between 1st person and 3rd person pronouns. In order to deal with imbalanced data in the training phase, class weight was set to "balanced". In this way, the classifier penalizes mistakes in each class with a weight inversely proportional to the frequency of that class, in order to avoid favoring only the overrepresented class. Both classifiers were evaluated on accuracy, precision, recall, and F1. We also plotted the ROC curve to facilitate the visualization of the relationship between precision and recall. The performance scores were calculated using 10 x stratified 10-fold cross validation, and we report the mean value of all 100 individual scores. For the implementation, we used the LogisticRegression class from the Python library Scikit-learn, with all default parameters (except for class_weight, set to "balanced" to deal with the imbalance over the classes).

Results and Discussion

Relationship Pronoun Use and Fraudulent Events Model 1, which uses normalized 1st and 3rd person pronouns as separate independent variables, did not show a significant relationship between pronoun use and fraudulent events, p = .108 (Table 3).

Table 3: Logistic regression results for Model 1 (normalized 1st and 3rd person pronoun frequency).

			p
Model likelihood			.108
Variable	β	S.E.	p
Intercept	-0.95	0.04	<.001
1st person pronouns	1.38	0.97	.155
3rd person pronouns	-2.61	1.34	.051

Table 4: Logistic regression results for Model 2 (ratio between 1st and 3rd person pronoun frequency).

			p
Model likelihood			.023
Variable	β	S.E.	p
Intercept	-0.99	0.03	<.001
1st/3rd person pronouns	0.01	0.01	.022

Model 2, which uses the ratio between 1st and 3rd person pronouns, did show a significant relationship between pronoun use and fraudulent events, p = .023 (Table 4). The results demonstrated that 1st and 3rd person pronoun use were not individually related to fraudulent events, but that the ratio between the two was. The average ratio for emails that were and were not related to fraudulent events was 3.93 and 3.74 respectively, demonstrating that during times of fraudulent events the use of 1st person pronouns relative to the use of 3rd person pronouns increased. This relationship conflicts with the notion that people try to distance themselves from the information they are conveying when they are being untruthful by increasing abstractness by reducing self-references and increasing other-references in their communication. Yet, these findings are in line with the study of Louwerse et al. (2010) which also did not find support for pronouns reflecting increased abstractness during fraud, but did find support for abstractness in verbs.

As it is important not only to examine the relationship between a construct and an outcome, but also to examine whether the construct has predictive power, we also report the results of the logistic regression classifiers, to predict whether or not an email is related to fraudulent events based on 1st and 3rd person pronoun use.

Table 5: Average results from the 10 x 10-fold cross validation for Classifier 1 (normalized 1st and 3rd person pronoun frequency).

	Accuracy	Precision	Recall	F1
Predicting fraud	48.24%	28.40%	55.39%	37.51%
Predicting non-fraud	48.24%	72.29%	45.44%	55.71%

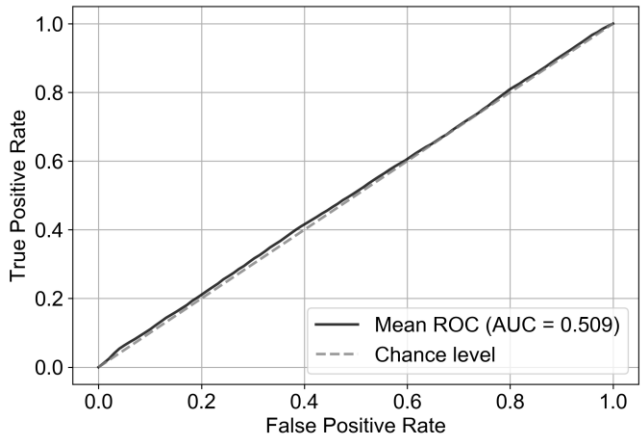

Figure 2: ROC curve for Classifier 1 (normalized 1st and 3rd person pronoun frequency).

Predicting Fraudulent Events Through Pronoun Use The evaluation scores from the 10 x 10-fold cross validation of Classifier 1 (trained on the normalized frequency of 1st person and 3rd person pronouns) are presented in Table 5 and the model's ROC curve is depicted in Figure 2.

As can be seen in Table 5 and in Figure 2, Classifier 1 did not perform above chance level (accuracy = 48.24%). F1 scores were also relatively low, reaching a maximum of 55.71% for predicting non-fraud. Precision scores were considerably higher for predicting non-fraud than for predicting fraud, indicating that the model favored the more common class. The evaluation scores and the ROC curve thus demonstrated the classifier based on the normalized individual frequencies has limited predictive power. This finding is in line with the absence of a significant relationship between these individual frequencies and fraudulent events.

The evaluation scores from the 10 x 10-fold cross validation for Classifier 2 (ratio between 1st and 3rd person pronoun use) are presented in Table 6 and the model's ROC curve is depicted in Figure 3. As can be seen in Table 6 and in Figure 3, the model using the ratio between 1st and 3rd person pronouns performed slightly above chance level (accuracy = 57.37%). F1 scores were again relatively low, reaching a maximum of 68.80%. As was the case for the classifier using the normalized individual frequencies, precision scores were a lot higher for the most common class. Considering all evaluation scores and the ROC curve, the model using the ratio between different pronouns also had limited predictive power.

Table 6: Average results from the 10 x 10-fold cross validation for Classifier 2 (ratio between 1st and 3rd person pronoun frequency).

	Accuracy	Precision	Recall	F1
Predicting fraud	57.37%	29.35%	36.79%	32.64%
Predicting non-fraud	57.37%	72.59%	65.41%	68.80%

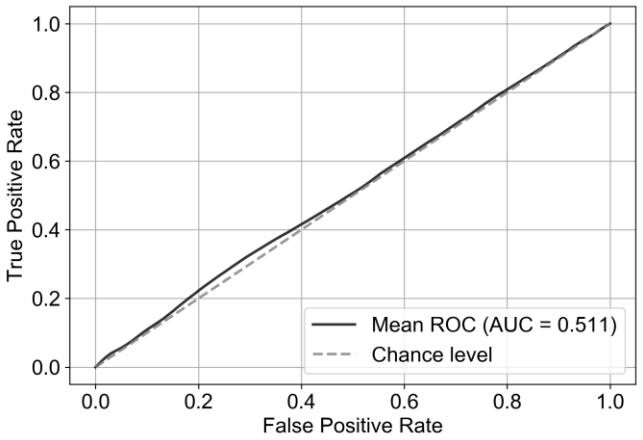

Figure 3: ROC curve for Classifier 2 (ratio between 1st and 3rd person pronoun frequency).

The mean ROC AUC of 0.511 indicates that this classifier is not able to classify deception any better than chance level. In conclusion, even though there was a significant relationship between pronoun use ratio and fraud, this construct had relatively little predictive power.

General Discussion

In the cognitive sciences many studies focus on examining statistical differences. This approach in which differences are examined provides us with valuable insights to explain behavior. Yet, it does not necessarily mean that these differences have predictive power. Understanding behavior both involves explaining and predicting behavior.

As a point in case, the current study examined the relationship between 1st and 3rd person pronoun use in emails sent by Enron employees and fraudulent activities. Additionally, we attempted to predict fraudulent events using 1st and 3rd person pronoun use in the emails.

Previous research demonstrated statistical differences between pronoun use in liars and people that are truthful, but the direction of the effects varied across studies. These studies generally examined the separate effects of 1st and 3rd person pronoun use. The current study demonstrated that the ratio between 1st and 3rd person pronouns was related to fraudulent events. This relationship indicated that the use of 1st person pronouns relative to 3rd person pronouns increased during times of fraudulent activities.

Differences in pronoun use between fraudulent and non-fraudulent communication do not necessarily imply that this construct has any predictive power. In our attempt to predict fraudulent events through pronoun use in emails, classification models scored relatively low on all evaluation measures. These models are therefore limited in their predictive power, not being able to classify deception any better than chance level.

The finding of the current study that differences in 1st and 3rd person pronoun use had no predictive power warrants reported differences in pronoun use between truthful and deceptive communication to not be interpreted as providing a meaningful tool for predicting fraud.

Possibly, classification models were limited in their predictive power in the current study due to the way in which the data were labeled. Whether an email was considered fraudulent or not was based on a general timeline, which might add extra noise to the data. One cannot be sure about the number of emails containing deception and the amount of emails containing no deception that was correctly labeled. However, this issue is inherent in using a naturalistic dataset. The fact remains that it is of utmost importance when one wants to gain a comprehensive insight to not only examine constructs in the laboratory, but also in settings that are of higher ecological validity.

Although effects of deception on 1st and 3rd person pronoun use in communication are widely reported and have been demonstrated in the current study, this construct seems to lack in predictive power. The current study highlights an important conclusion for the cognitive sciences: The importance of not only testing for differences, but of also applying predictive models in order to determine whether effects of a construct on an outcome are also meaningful in predicting the outcome.

References

Bleidorn, W., Arslan, R. C., Denissen, J. J., Rentfrow, P. J., Gebauer, J. E., Potter, J., & Gosling, S. D. (2016). Age and gender differences in self-esteem—A cross-cultural window. *Journal of Personality and Social Psychology*, *111*(3), 396. DOI: 10.1037%2Fpspp0000078.

Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., & Torres-Carrasquillo, P. A. (2006). Support vector machines for speaker and language recognition. *Computer Speech & Language*, *20*(2), 210-229. DOI: 10.1016/j.csl.2005.06.003.

Cohen, W.W. (2015). *Enron Email Dataset*. Retrieved by https://www.cs.cmu.edu/~enron on 02/01/2017.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*, 74-118. DOI: 10.1037/0033-2909.129.1.74.

Diesner, J., Frantz, T., Carley, K.M. (2005). Communication networks from the Enron email corpus "It's always about the people. Enron is no different". *Computational and Mathematical Organization Theory, 11*, 201-228. DOI: 10.1007/s10588-005-5377-0.

Garnefski, N., Van Den Kommer, T., Kraaij, V., Teerds, J., Legerstee, J., & Onstein, E. (2002). The relationship between cognitive emotion regulation strategies and emotional problems: comparison between a clinical and a non-clinical sample. *European Journal of Personality*, *16*(5), 403-420. DOI: 10.1002/per.458.

Hancock, J.T., Curry, L., Goorha, S., & Woodworth, M.T. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 45*, 1-23. DOI: 10.1080/01638530701739181.

Humpherys, S. L., Mott, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems, 50*(3), 585-594. DOI: 10.1016/j.dss.2010.08.009.

Klimt, B. & Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. *Proceedings of the Fifteenth European Conference on Machine Learning*, pp. 217–225. DOI: 10.1007/978-3-540-30115-8_22.

Louwerse, M., Lin, K. I., Drescher, A., & Semin, G. (2010). Linguistic cues predict fraudulent events in a corporate social network. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 961-966.

Lu, H. P., & Hsiao, K. L. (2010). The influence of extro/introversion on the intention to pay for social networking sites. *Information & Management, 47*(3), 150-157. DOI: 10.1016/j.im.2010.01.003.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. N. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665–675. DOI: 10.1177/0146167203029005010.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 309-319.

Pennebaker, J. W. (2011). *The secret life of pronouns: How our words reflect who we are.* New York: Bloomsbury. DOI: 10.1016/S0262-4079(11)62167-2.

Rosenberg, M. D., Casey, B. J., & Holmes, A. J. (2018). Predicting complements explanation in understanding the developing brain. *Nature Communications, 9*(1), 589.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques*. Burlington, Mass.: Morgan Kaufmann. DOI: 10.1145/507338.507355.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122. DOI: 10.1177/1745691617693393.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology, 14*, 1-59. DOI: 10.1016/S0065-2601(08)60369-X.