Discriminative Models for Predicting Deception Strategies

Scott Appling, Erica Briscoe, C.J. Hutto
Georgia Tech Research Institute
Georgia Institute of Technology
Atlanta, GA 30332
{scott.appling, erica.briscoe}@gtri.gatech.edu, cjhutto@gatech.edu

ABSTRACT

Although a large body of work has previously investigated various cues predicting deceptive communications, especially as demonstrated through written and spoken language (e.g., [30]), little has been done to explore predicting *kinds* of deception. We present novel work to evaluate the use of textual cues to discriminate between deception strategies (such as exaggeration or falsification), concentrating on intentionally untruthful statements meant to persuade in a social media context. We conduct human subjects experimentation wherein subjects were engaged in a conversational task and then asked to label the kind(s) of deception they employed for each deceptive statement made. We then develop discriminative models to understand the difficulty between choosing between one and several strategies. We evaluate the models using precision and recall for strategy prediction among 4 deception strategies based on the most relevant psycholinguistic, structural, and data-driven cues. Our single strategy model results demonstrate as much as a 58% increase over baseline (random chance) accuracy and we also find that it is more difficult to predict certain kinds of deception than others.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: [Statistical Computing]; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic Processing*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*; I.5.1 [**Pattern Recognition**]: Models—*Statistical*

General Terms

Measurement

Keywords

Deception; Deception Strategies; Deception Strategy Prediction; Social Computing; Natural Language Processing

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742575.

1. INTRODUCTION

The recent commonality of social media communication has drastically changed the way many individuals receive information [20]. With increasingly available technology, almost anyone can readily and effectively address single individuals or large numbers of people instantly. Along with this increased access is the potential to influence, either benignly or maliciously, where deception is a common utilized method [5].

Research into 'tells' that indicate when a person is lying have been a subject of interest for many years (e.g., [15]) and has resulted in a substantial list of potential cues ([12]). These cues to deception are usually described as those that occur significantly more or less frequently when a communicator expresses a lie as compared to when they tell the truth and may differ according to the point of view of the receiver and sender [1].

The use of linguistic cues to indicate potentially deceptive communications has also been studied in a variety of modalities and contexts (e.g., [24, 13, 7]). While text-based deception in computer mediated communication is a specific area of study [29, 30, 3], there has been less of a concentration on the various *types* of deception strategies that individuals utilize in their duplicitous communications, especially as people more often use informal, succinct messages as a result of the widespread adoption of social media. In this work we study: 1) the difficultly in automatically discriminating between single strategies from one another as well as recognizing statements including both; 2) the robustness of single strategy prediction models choosing between four different strategies; 3) the difficulties in predicting certain strategies using generalizable features alone.

2. DECEPTION STRATEGIES

Individuals may employ deception through a variety of strategic means, such as in an attempt to mislead or misrepresent information [26]. These strategies are often used in order to change the beliefs of the message receiver [8]. [28, 8] identified the following strategies:

- **Falsification** - e.g. lies, contradictions, or "distortions" [28]
- **Exaggeration** - e.g. more or modified information via superlatives [28]
- **Omission** - e.g. secrets (missing information), half-truths (less or modified information), and what [28] refers to as concealment

- **Misleading** - e.g. topic changes, irrelevant information, or equivocation; [28] refers to these statements as diversionary responses

Following the strategies taxonomy above, we conducted human subject experimentation to simulate an online social network environment where participants were tasked with engaging in a conversation where they employed various deception strategies. Then, using the obtained data, we develop and evaluate discriminative models of deception strategies.

The rest of the paper is organized as follows: first we describe the experimental setup and the collected data; next, we discuss the development of discriminative models to automatically identify different deception strategies using several kinds of linguistic features along with the results; afterwards, we discuss the implications of the results and the difficulty in detecting certain strategies; we close with conclusions and future work.

3. METHODS AND DATA

Eighty-three subjects (recruited from the student body at Georgia Tech) participated in the experiment. Each subject was seated at a computer station and asked to use a mock online social media site (FaceFriend). The mock site was intended to resemble a popular social networking site (see Figure 1). Participants completed three pre-experiment surveys, including a personality test and a survey regarding their social media and email usage, so as to ensure consistency among familiarity with social media. Another survey not relevant to these results was also conducted. The procedure and results reported below are from one of a three part experiment, the other two parts of which are not reported here.

Participants were asked to read a scenario involving a group decision making task called the 'subarctic survival problem' [14]. Participants were instructed to log in to a mock social media platform, *FaceFriend* and to communicate with another individual to decide the best ranking of items to take from a crashed plane site in an Arctic environment. Participants were further instructed to advocate for a specific ranked list of items, also provided, and to do so using deception, when possible. To ameliorate differences in creative ability, subjects were provided with expert explanations for the advantages provided by each item; these could be used as the basis for creating deceptive statements. For example, the list of items included an Axe, which the expert noted could be used to chop wood. Lastly, participants were ostensibly informed that the other conversation participant would be unaware of any potential deception. The participant's conversational partner was a confederate member of the research team, sitting in another room. During the conversation, to minimize variation in the confederate's response language, the confederates communicated using a list of statements conceived beforehand as suitable decision making conversation responses (e.g., "What about the canvas?"). After discussing the items for 12 minutes, the interaction was stopped, and subjects were asked to log out of FaceFriend.

Afterwards, participants were shown the list of statements they made during the task in a browser and asked to identify which ones were deceptive using a check box. Upon indicating a particular statement was deceptive, participants

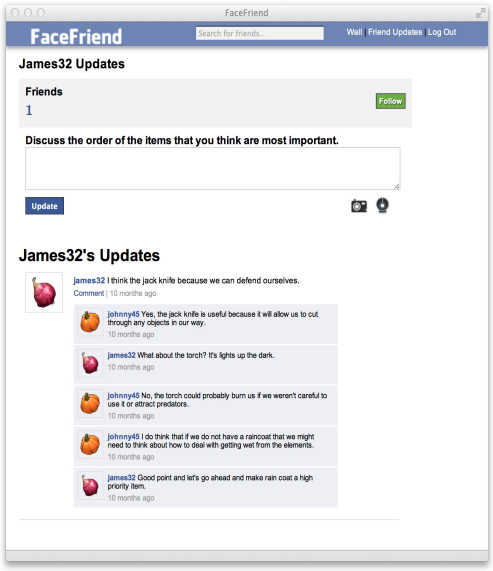

Figure 1: Screenshot of the FaceFriend Platform.

were prompted to categorize their strategy for that particular message. The possible strategies to select were the same bolded terms from Section 2, and more than one designation was allowed. An additional category, called "Other", was also given which allowed free response space for any strategies that the subject felt did not fit into one of the four provided categories.

Table 1 lists the frequencies of statements obtained by strategies, after filtering out statements which were uninterpretable, where more than 2 strategies were chosen or, composed of less than 4 words.[1] Most participants reported utilizing only one strategy in their deception attempts.

Table 1: Frequency of Deception Strategies

	Exaggeration	Misleading	Omission	Falsification
E	67	7	3	8
M	-	69	8	10
O	-	-	41	4
F	-	-	-	45

Elements on the diagonal represent statements where a single strategy was chosen while other elements represent two strategies that were chosen to categorize a single statement (e.g. Three statements were labeled as exaggeration and omission.)

4. MODELING AND RESULTS

Using the collected data with labeled deception strategies, we create discriminative models based on structural [3, 29], psycholinguistic (using LIWC) [23, 27], and data-driven (e.g., trigrams[2])[9] features. We create two different kinds

[1]Elements on the diagonal represent only a single strategy being selected.

[2]We do not use unigram or bigrams features as they were observed to be covered by LIWC features.

Table 2: Sample Means and Standard Deviations of Structural & Psycholinguistic Cues by Deception Strategy

	E		M		O		F	
Structural	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Word Count	12.91	7.17	13.44	7.72	12.87	8.49	14.76	6.46
Word Length	3.22	0.59	3.14	0.70	3.04	0.70	3.35	0.69
Pausality	0.87	0.58	0.69	0.80	0.87	0.76	1.10	1.17
Verb Count	2.03	1.11	2.63	1.80	2.26	1.92	2.31	1.48
FK Grade Level	0.30	4.77	0.54	4.96	-0.50	5.82	1.13	4.01
Modifiers	1.67	1.19	1.53	1.27	1.51	1.23	1.99	1.44
Psycholinguistic								
Personal Pronouns	0.06	0.08	0.04	0.04	0.03	0.03	0.04	0.04
Insight	0.01	0.02	0.02	0.03	0.01	0.02	0.01	0.01
Adverbs	0.03	0.04	0.03	0.05	0.04	0.05	0.03	0.04
Negations	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.03
Numbers	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.04

of discriminative models, multi-label models, e.g. statement X is both an exaggeration and misleading, and single label models, e.g. Statement X is primarily an omission. For the single label instance, we use only statements where one deception strategy has been chosen. For both multi- and single strategy discriminative models, we compared performance using both a Random Forest [2] and a series of Linear SVMs [16] and found the Random Forest classifier to give the best precision.

4.1 Pre-Processing and Descriptions of Cues

For each statement we employ POS tagging [21] to extract structural features. Next, we lowercase and tokenize the statement before extracting psycholinguistic and data-driven features.

Table 2 describes the means and standard deviations of various sample cues[3] used as model features for all single statement deceptions. We include average modifier count as an additional measure of complexity where traditional measures of statement complexity [19] may fail to accurately assess statement complexity for social media style texts (e.g. Tweets, Facebook-style Posts, interaction chat).

4.2 Single & Multi-Strategy Prediction

Models were 10-Fold cross validated on the set of statements labeled as either employing one or both strategies, i.e. each model is able to predict a statement as being one or both strategies, taking into account the non-mutually exclusive nature of deception strategies. In total, we train 6 models and Table 3 describes the precision scores for each different pair. Discriminating between exaggerations and falsifications results in the highest score of 71% precision while the most difficult decision is between omission and falsification strategies. In single strategy prediction we filter out statements where multiple strategies were selected and build a single 10-Fold cross validated discriminative model to determine the most likely deception strategy. The majority of training examples collected are single strategy deception statements and we expect that in practice models built from these collections of statements, removing statements with a plurality of labels would effectively serve as noise reduction given the frequency counts (see Table 1).

[3]Definitions of cues: FK Grade Level-Flesh Kincaid Grade level; Pausality - average count of punctuation marks used within a statement; Modifiers - a count of adjectives and adverbs; Psycholinguistic cues are measured as a percentage of the statements that is composed of the category; Insight words e.g. think, know, consider; Numbers - second, thousand

Table 3: Multi-label Precision and Semantic Relatedness By Strategy Pairs

	Exagg.		Mislead.		Falsif.	
	P	**SR**	**P**	**SR**	**P**	**SR**
Omiss.	65.4%	0.30	57.4%	0.23	48.0%	0.27
Exagg.	-	-	57.4%	0.68	71.3%	0.29
Mislead.	-	-	-	-	65.7%	0.43

Table 4: Single Strategy Confusion Matrix

		Actual					
		O	**E**	**M**	**F**	**Precision**	**Recall**
Predicted	**O**	8	20	12	5	28.6%	19.5%
	E	11	38	17	4	41.9%	54.2%
	M	11	24	30	7	38.8%	48.0%
	F	6	14	19	7	25.8%	15.5%

In the single strategy analysis we look at predicting 1 of 4 deception strategies where the baseline model is random chance i.e. 1 in 4 chance or 25% accuracy. Here we saw a 58% improvement in accuracy over the baseline model with 42% precision and 54% recall for correctly predicting exaggerations over other strategies.

Motivated by the literature on strategy utilization [11], we see the multi-label classification problem additionally interesting and perhaps more natural as humans engage in deception. To fully evaluate the multi-label prediction problem where any statement could exhibit 1 of 15 labels (i.e. choose

Table 5: Ranked Features By Deception Strategy

	Exaggeration	Misleading	Omission	Falsification
1	liwc_personal_pron	liwc_3rd_singular	"more important than"	liwc_negations
2	"and keep us"	liwc_anxiety	sf_avg_pau_cnt	sf_avg_mod_cnt
3	"because you can"	liwc_friends	liwc_3rd_singular	liwc_numbers
4	"from the cold "	liwc_insight	liwc_adverbs	"be last because"
5	"is better than"	"and can be"	"also be used"	sf_avg_pau_cnt

1 of 2 labels {uses strategy, does not use} for each of the 4 strategies where at least one strategy is employed) would require less sparse data and instead we reduced this set to 1 in 3 labels (first strategy, second, or both) to focus on identifying discriminative features between strategies.

5. DISCUSSION

A robust model will ideally employ topic-free features and we would expect good features thus to be highly ranked among thousands of data-driven features. The literature on linguistic properties of deception strategies is limited, however we note that three deception strategies (falsification, concealment, and equivocation) were employed by [4, 6] in a controlled experimental setting where they reported the misleading deception strategy as the, "most brief, vague, and hesitant" whereas in comparison, falsifications were the lowest on these characteristics. This finding would seem to indicate that in comparing misleading (equivocations) to falsifying statements, we would likely see longer statements with the later. We calculated the mean word counts for all statements including multi-label statements and found for falsifying statements more words on average than for misleading statements which supports the literature. In table 2 we also observe that the means for several structural cues (e.g. pausality, modifiers, Fleish-Kincaid grade level) are higher for falsification than other strategies.

Table 5 presents the Top 5 ranked features via the 1-Way ANOVA F-Test [18] for single strategy prediction. In analyzing these, the top feature for falsification was in the LIWC negation category (e.g., using 'no', 'not', 'never'), which is logical given that these statements are often disagreements in response to the statements made by the other participant. Overall, we find good evidence to support the generality of our models from feature inspection, as many of the top ranked features are psycholinguistic or structural. This would tend to indicate that with sufficient training data, discriminatory models can still be successful if these non-data-driven features are available.

The precision results of multi-label model performance between the different strategies indicates a greater difficulty in distinguishing certain strategies from others. The biggest difference in precision scores was approximately 23%, between models attempting to distinguish between omission and falsification and exaggeration and falsification.

Because there is a dearth of previous work on using discriminative models for deception strategy prediction, we evaluate the quality of our results in light of a comparison between the performance of each model and the semantic relatedness of strategy concepts to one another. Intuitively, we would expect that two deceptive message strategies that are highly dissimilar would be likely to exhibit different linguistic phenomena. For example, exaggerations are manifested by amplifiers like "too much" or "less important", whereas falsifications may manifest in either factually untrue statements or, as in conversations, as negations to questions. We expect to see better performance in discriminating between strategies whose semantic relatedness distance[4] is closer to 0.

We evaluated the semantic relation similarity between strategy concepts using the UMBC Phrase Similarity Service[5] with the LDC Gigawords Corpus[17]. Table 3 lists the similarity of concepts where higher scores are more similar. As expected, as relatedness scores decrease, model mean average precision increases.

We evaluated the omission deception strategy and while precision scores for multi-strategy prediction models were decent in some cases, there was little evidence in the literature for strong general features that could discriminate it from other strategies; it seems more likely that any good results found would be supported by data-driven features in practice. It is due to the contextual knowledge requirements (e.g. detecting omissions in machine translated texts[22] is done comparing missing information) entailed in reasoning about omissions that place it as a more challenging strategy to detect within language alone. For domain-free classifications, we suspect that verbal and visual cues are potentially more generally indicative of this kind deception strategy.

6. CONCLUSION

This study reports on the use of linguistic features in discriminative models determining the kinds of deception strategies a communicator is employing in a social media setting. Our best model achieved 71% precision distinguishing between exaggerations and falsifications. While our dataset is relatively small, less than 300 statements, we see the results as promising towards the development of deception detection techniques which can reason more robustly about different kinds of deception strategies.

7. FUTURE WORK

Given the relative nascence of social media and especially its effect on the human communication styles, a good deal of work is needed to understand how previously understood research into linguistically manifested deception may differ

[4]Semantic relatedness distance is evaluated on a scale of 0-1 where higher scores indicate more conceptual relatedness.
[5]http://swoogle.umbc.edu/SimService/index.html

with the use of new communication platforms. For example, new research supporting the differences between native/non-native message senders and communication accommodation theory in social media [10, 25] may provide better controls for understanding social media specific linguistic cues to deception. Future work will also focus specifically on how social media users utilize evidence (such as URLs or pictures) to substantiate their untruthful claims.

Additionally, we intend to use this experimental data set to evaluate how personality may be a predictor of the type of deception strategy chosen. We also intend to extend our features for classification to include discourse cues as [31], which will allow us to better account for the interaction and context in which deceptive statements are made.

8. ACKNOWLEDGMENTS

We thank the reviewers for their helpful and insightful comments. This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract No. W911NF-12-1-0043.

9. REFERENCES

[1] D. E. Anderson, B. M. DePaulo, M. E. Ansfield, J. J. Tickle, and E. Green. Beliefs about cues to deception: Mindless stereotypes or untapped wisdom? *Journal of Nonverbal Behavior*, 23(1):67–89, 1999.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] E. J. Briscoe, D. S. Appling, and H. Hayes. Cues to deception in social media communications. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 1435–1443. IEEE, 2014.

[4] D. B. Buller, J. K. Burgoon, A. Buslig, and J. Roiger. Interpersonal deception theory: Examining deception from a communication perspective. Technical report, Army Research Institute for the Behavioral and Social Sciences, 1998.

[5] D. B. Buller, J. K. Burgoon, J. Daly, and J. Wiemann. Deception: Strategic and nonstrategic communication. *Strategic interpersonal communication*, pages 191–223, 1994.

[6] D. B. Buller, J. K. Burgoon, C. H. White, and A. S. Ebesu. Interpersonal deception vii behavioral profiles of falsification, equivocation, and concealment. *Journal of language and social psychology*, 13(4):366–395, 1994.

[7] J. K. Burgoon, J. Blair, T. Qin, and J. F. Nunamaker Jr. Detecting deception through linguistic analysis. In *Intelligence and Security Informatics*, pages 91–101. Springer, 2003.

[8] J. K. Burgoon, D. B. Buller, L. K. Guerrero, W. A. Afifi, and C. M. Feldman. Interpersonal deception: Xii. information management dimensions underlying deceptive and truthful messages. *Communications Monographs*, 63(1):50–69, 1996.

[9] W. B. Cavnar and J. M. Trenkle. N-grambased text categorization. In *In Proc. of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994.

[10] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754. ACM, 2011.

[11] B. M. DePaulo, D. A. Kashy, S. E. Kirkendol, M. M. Wyer, and J. A. Epstein. Lying in everyday life. *Journal of Personality and Social Psychology*, 70(5):979, 1996.

[12] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological bulletin*, 129(1):74, 2003.

[13] N. D. Duran, C. Hall, P. M. McCarthy, and D. S. McNamara. The linguistic correlates of conversational deception: Comparing natural language processing technologies. *Applied Psycholinguistics*, 31(03):439–462, 2010.

[14] P. M. Eady and J. C. Lafferty. *The subarctic survival situation*. Experimental Learning Methods, 1969.

[15] P. Ekman and W. V. Friesen. Nonverbal leakage and clues to deception. Technical report, DTIC Document, 1969.

[16] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[17] D. Graff and C. Cieri. English gigaword ldc2003t05. 2003.

[18] G. Heiman. *Basic statistics for the behavioral sciences*. Cengage Learning, 2013.

[19] J. Kincaid, R. Fishburn, R. Rogers, and B. Chissom. Derivation of new readability formulas for navy enlisted personnel (research branch report 8-75). *Memphis, TN: Naval Air Station, Millington, Tennessee*, 1975.

[20] A. Lenhart, K. Purcell, A. Smith, and K. Zickuhr. Social media & mobile internet use among teens and young adults. millennials. *Pew Internet & American Life Project*, 2010.

[21] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[22] I. D. Melamed. Automatic detection of omissions in translations. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 764–769. Association for Computational Linguistics, 1996.

[23] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319. Association for Computational Linguistics, 2011.

[24] J. Pennebaker. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54:547–577, 2003.

[25] C. Pérez-Sabater. The linguistics of social networking: A study of writing conventions on facebook. *Linguistik online*, 56(6/12):82, 2012.

[26] V. L. Rubin. On deception and deception detection: Content analysis of computer-mediated stated beliefs.

Proceedings of the American Society for Information Science and Technology, 47(1):1–10, 2010.

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

[28] R. E. Turner, C. Edgley, and G. Olmstead. Information control in conversations: Honesty is not always the best policy. *Kansas Journal of Sociology*, 1975.

[29] L. Zhou. An empirical investigation of deception behavior in instant messaging. *Professional Communication, IEEE Transactions on*, 48(2):147–160, 2005.

[30] L. Zhou and Y.-w. Sung. Cues to deception in online chinese groups. In *Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, page 146. IEEE Computer Society, 2008.

[31] L. Zhou and Y.-w. Sung. Discourse cues to online deception. In *Quality in Government and Business Symposium*, page 1. Citeseer, 2010.