The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language

Rada Mihalcea
University of North Texas
rada@cs.unt.edu

Carlo Strapparava
FBK-IRST
strappa@fbk.eu

Abstract

In this paper, we present initial experiments in the recognition of deceptive language. We introduce three data sets of true and lying texts collected for this purpose, and we show that automatic classification is a viable technique to distinguish between truth and falsehood as expressed in language. We also introduce a method for class-based feature analysis, which sheds some light on the features that are characteristic for deceptive text.

You should not trust the devil, even if he tells the truth.
– Thomas of Aquin (medieval philosopher)

1 Introduction and Motivation

The discrimination between truth and falsehood has received significant attention from fields as diverse as philosophy, psychology and sociology. Recent advances in computational linguistics motivate us to approach the recognition of deceptive language from a data-driven perspective, and attempt to identify the salient features of lying texts using natural language processing techniques.

In this paper, we explore the applicability of computational approaches to the recognition of deceptive language. In particular, we investigate whether automatic classification techniques represent a viable approach to distinguish between truth and lies as expressed in written text. Although acoustic and other non-linguistic features were also found to be useful for this task (Hirschberg et al., 2005), we deliberately focus on written language, since it represents the type of data most frequently encountered on the Web (e.g., chats, forums) or in other collections of documents.

Specifically, we try to answer the following two questions. First, are truthful and lying texts separable, and does this property hold for different datasets? To answer this question, we use three different data sets that we construct for this purpose – consisting of true and false short statements on three different topics – and attempt to automatically separate them using standard natural language processing techniques.

Second, if truth and lies are separable, what are the distinctive features of deceptive texts? In answer to this second question, we attempt to identify some of the most salient features of lying texts, and analyse their occurrence in the three data sets.

The paper is organized as follows. We first briefly review the related work, followed by a description of the three data sets that we constructed. Next, we present our experiments and results using automatic classification, and introduce a method for the analysis of salient features in deceptive texts. Lastly, we conclude with a discussion and directions for future work.

2 Related Work

Very little work, if any, has been carried out on the automatic detection of deceptive language in written text. Most of the previous work has focused on the psychological or social aspects of lying, and there are only a few previous studies that have considered the linguistic aspects of falsehood.

In psychology, it is worthwhile mentioning the study reported in (DePaulo et al., 2003), where more than 100 cues to deception are mentioned. However, only a few of them are linguistic in nature, as e.g., word and phrase repetitions, while most of the cues involve speaker's behavior, including facial expressions, eye shifts, etc. (Newman et al., 2003) also report on a psycholinguistic study, where they conduct a qualitative analysis of true and false stories by using word counting tools.

Computational work includes the study of (Zhou et al., 2004), which studied linguistic cues for deception detection in the context of text-based asynchronous computer mediated communication, and (Hirschberg et al., 2005) who focused on deception in speech using primarily acoustic and prosodic features.

Our work is also related to the automatic classification of text genre, including work on author profiling (Koppel et al., 2002), humor recognition

TRUTH	LIE
ABORTION	
I believe abortion is not an option. Once a life has been conceived, it is precious. No one has the right to decide to end it. Life begins at conception,because without conception, there is no life.	A woman has free will and free choice over what goes on in her body. If the child has not been born, it is under her control. Often the circumstances an unwanted child is born into are worse than death. The mother has the responsibility to choose the best course for her child.
DEATH PENALTY	
I stand against death penalty. It is pompous of anyone to think that they have the right to take life. No court of law can eliminate all possibilities of doubt. Also, some circumstances may have pushed a person to commit a crime that would otherwise merit severe punishment.	Death penalty is very important as a deterrent against crime. We live in a society, not as individuals. This imposes some restrictions on our actions. If a person doesn't adhere to these restrictions, he or she forfeits her life. Why should taxpayers' money be spent on feeding murderers?
BEST FRIEND	
I have been best friends with Jessica for about seven years now. She has always been there to help me out. She was even in the delivery room with me when I had my daughter. She was also one of the Bridesmaids in my wedding. She lives six hours away, but if we need each other we'll make the drive without even thinking.	I have been friends with Pam for almost four years now. She's the sweetest person I know. Whenever we need help she's always there to lend a hand. She always has a kind word to say and has a warm heart. She is my inspiration.

Table 1: Sample true and deceptive statements

(Mihalcea and Strapparava, 2006), and others.

3 Data Sets

To study the distinction between true and deceptive statements, we required a corpus with explicit labeling of the truth value associated with each statement. Since we were not aware of any such data set, we had to create one ourselves. We focused on three different topics: opinions on abortion, opinions on death penalty, and feelings about the best friend. For each of these three topics an annotation task was defined using the Amazon Mechanical Turk service.

For the first two topics (*abortion* and *death penalty*), we provided instructions that asked the contributors to imagine they were taking part in a debate, and had 10-15 minutes available to express their opinion about the topic. First, they were asked to prepare a brief speech expressing their true opinion on the topic. Next, they were asked to prepare a second brief speech expressing the opposite of their opinion, thus lying about their true beliefs about the topic. In both cases, the guidelines asked for at least 4-5 sentences and as many details as possible.

For the third topic (*best friend*), the contributors were first asked to think about their best friend and describe the reasons for their friendship (including facts and anecdotes considered relevant for their relationship). Thus, in this case, they were asked to tell the truth about how they felt about their best friend. Next, they were asked to think about a person they could not stand, and describe it as if s/he were their best friend. In this second case, they had to lie about their feelings toward this person. As before, in both cases the instructions asked for at least 4-5 detailed sentences.

We collected 100 true and 100 false statements for each topic, with an average of 85 words per statement. Previous work has shown that data collected through the Mechanical Turk service is reliable and comparable in quality with trusted sources (Snow et al., 2008). We also made a manual verification of the quality of the contributions, and checked by hand the quality of all the contributions. With two exceptions – two entries where the true and false statements were identical, which were removed from the data – all the other entries were found to be of good quality, and closely following our instructions.

Table 1 shows an example of true and deceptive language for each of the three topics.

4 Experimental Setup and Results

For the experiments, we used two classifiers: Naïve Bayes and SVM, selected based on their performance and diversity of learning methodologies. Only minimal preprocessing was applied to the three data sets, which included tokenization and stemming. No feature selection was performed, and stopwords were not removed.

Table 2 shows the ten-fold cross-validation results using the two classifiers. Since all three data sets have an equal distribution between true and false statements, the baseline for all the topics is 50%. The average classification performance of 70% – significantly higher than the 50% baseline – indicates that good separation can be obtained

between true and deceptive language by using automatic classifiers.

Topic	NB	SVM
ABORTION	70.0%	67.5%
DEATH PENALTY	67.4%	65.9%
BEST FRIEND	75.0%	77.0%
AVERAGE	70.8%	70.1%

Table 2: Ten-fold cross-validation classification results, using a Naïve Bayes (NB) or Support Vector Machines (SVM) classifier

To gain further insight into the variation of accuracy with the amount of data available, we also plotted the learning curves for each of the data sets, as shown in Figure 1. The overall growing trend indicates that more data is likely to improve the accuracy, thus suggesting the collection of additional data as a possible step for future work.

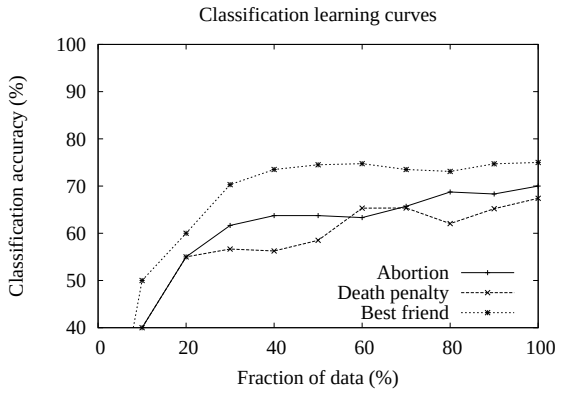

Figure 1: Classification learning curves.

We also tested the portability of the classifiers across topics, using two topics as training data and the third topic as test. The results are shown in Table 3. Although below the in-topic performance, the average accuracy is still significantly higher than the 50% baseline, indicating that the learning process relies on clues specific to truth/deception, and it is not bound to a particular topic.

5 Identifying Dominant Word Classes in Deceptive Text

In order to gain a better understanding of the characteristics of deceptive text, we devised a method to calculate a score associated with a given class of words, as a measure of saliency for the given word class inside the collection of deceptive (or truthful) texts.

Given a class of words $C = \{W_1, W_2, ..., W_N\}$, we define the class coverage in the deceptive corpus D as the percentage of words from D belonging to the class C:

$$Coverage_D(C) = \frac{\sum_{W_i \in C} Frequency_D(W_i)}{Size_D}$$

where $Frequency_D(W_i)$ represents the total number of occurrences of word W_i inside the corpus D, and $Size_D$ represents the total size (in words) of the corpus D.

Similarly, we define the class C coverage for the truthful corpus T:

$$Coverage_T(C) = \frac{\sum_{W_i \in C} Frequency_T(W_i)}{Size_T}$$

The *dominance score* of the class C in the deceptive corpus D is then defined as the ratio between the coverage of the class in the corpus D with respect to the coverage of the same class in the corpus T:

$$Dominance_D(C) = \frac{Coverage_D(C)}{Coverage_T(C)} \quad (1)$$

A dominance score close to 1 indicates a similar distribution of the words in the class C in both the deceptive and the truthful corpus. Instead, a score significantly higher than 1 indicates a class that is dominant in the deceptive corpus, and thus likely to be a characteristic of the texts in this corpus. Finally, a score significantly lower than 1 indicates a class that is dominant in the truthful corpus, and unlikely to appear in the deceptive corpus.

We use the classes of words as defined in the Linguistic Inquiry and Word Count (LIWC), which was developed as a resource for psycholinguistic analysis (Pennebaker and Francis, 1999). The 2001 version of LIWC includes about 2,200 words and word stems grouped into about 70 broad categories relevant to psychological processes (e.g., EMOTION, COGNITION). The LIWC lexicon has been validated by showing significant correlation between human ratings of a large number of written texts and the rating obtained through LIWC-based analyses of the same texts.

All the word classes from LIWC are ranked according to the dominance score calculated with formula 1, using a mix of all three data sets to create the D and T corpora. Those classes that have a high score are the classes that are dominant in deceptive text. The classes that have a small score are the classes that are dominant in truthful text and lack from deceptive text. Table 4 shows the top ranked classes along with their dominance score and a few sample words that belong to the given class and also appeared in the deceptive (truthful) texts.

Interestingly, in both truthful and deceptive language, three of the top five dominant classes are related to humans. In deceptive texts however, the

Training	Test	NB	SVM
DEATH PENALTY + BEST FRIEND	ABORTION	62.0%	61.0%
ABORTION + BEST FRIEND	DEATH PENALTY	58.7%	58.7%
ABORTION + DEATH PENALTY	BEST FRIEND	58.7%	53.6%
AVERAGE		59.8%	57.8%

Table 3: Cross-topic classification results

Class	Score	Sample words
Deceptive Text		
METAPH	1.71	god, die, sacred, mercy, sin, dead, hell, soul, lord, sins
YOU	1.53	you, thou
OTHER	1.47	she, her, they, his, them, him, herself, himself, themselves
HUMANS	1.31	person, child, human, baby, man, girl, humans, individual, male, person, adult
CERTAIN	1.24	always, all, very, truly, completely, totally
Truthful Text		
OPTIM	0.57	best, ready, hope, accepts, accept, determined, accepted, won, super
I	0.59	I, myself, mine
FRIENDS	0.63	friend, companion, body
SELF	0.64	our, myself, mine, ours
INSIGHT	0.65	believe, think, know, see, understand, found, thought, feels, admit

Table 4: Dominant word classes in deceptive text, along with sample words.

human-related word classes (YOU, OTHER, HUMANS) represent detachment from the self, as if trying not to have the own self involved in the lies. Instead, the classes of words that are closely connected to the self (I, FRIENDS, SELF) are lacking from deceptive text, being dominant instead in truthful statements, where the speaker is comfortable with identifying herself with the statements she makes.

Also interesting is the fact that words related to certainty (CERTAIN) are more dominant in deceptive texts, which is probably explained by the need of the speaker to explicitly use truth-related words as a means to emphasize the (fake) "truth" and thus hide the lies. Instead, belief-oriented vocabulary (INSIGHT), such as *believe*, *feel*, *think*, is more frequently encountered in truthful statements, where the presence of the real truth does not require truth-related words for emphasis.

6 Conclusions

In this paper, we explored automatic techniques for the recognition of deceptive language in written texts. Through experiments carried out on three data sets, we showed that truthful and lying texts are separable, and this property holds for different data sets. An analysis of classes of salient features indicated some interesting patterns of word usage in deceptive texts, including detachment from the self and vocabulary that emphasizes certainty. In future work, we plan to explore the role played by affect and the possible integration of automatic emotion analysis into the recognition of deceptive language.

References

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

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. 2005. Distinguishing deceptive from non-deceptive speech. In *Proceedings of INTERSPEECH-2005*, Lisbon, Portugal.

M. Koppel, S. Argamon, and A. Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 4(17):401–412.

R. Mihalcea and C. Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142.

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

J. Pennebaker and M. Francis. 1999. Linguistic inquiry and word count: LIWC. Erlbaum Publishers.

R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii.

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