Experiments in Open Domain Deception Detection

Verónica Pérez-Rosas and Rada Mihalcea
Computer Science and Engineering
University of Michigan
vrncapr@umich.edu, mihalcea@umich.edu

Abstract

The widespread use of deception in online sources has motivated the need for methods to automatically profile and identify deceivers. This work explores deception, gender and age detection in short texts using a machine learning approach. First, we collect a new open domain deception dataset also containing demographic data such as gender and age. Second, we extract feature sets including n-grams, shallow and deep syntactic features, semantic features, and syntactic complexity and readability metrics. Third, we build classifiers that aim to predict deception, gender, and age. Our findings show that while deception detection can be performed in short texts even in the absence of a predetermined domain, gender and age prediction in deceptive texts is a challenging task. We further explore the linguistic differences in deceptive content that relate to deceivers gender and age and find evidence that both age and gender play an important role in people's word choices when fabricating lies.

1 Introduction

Given the potential ethical and security risks associated with deceitful interactions, it is important to build computational tools able not only to detect deceivers but also to provide insights into the nature of deceptive behaviors. In particular, information related to the demographics of the deceivers could be potentially useful, as recent studies have shown that online users lie frequently about their appearance, gender, age or even education level.

There are multiple scenarios where it would be desirable to identify deceivers' demographics; for instance, identifying the age and gender of SMS senders or Twitter users might help improve parental controls, spam filtering, and user's security and privacy.

In this paper, we present a study on deception detection in an open domain, and also present an analysis of deceptive behavior in association with gender and age. Unlike previous studies, where domain-specific conversational transcripts and reviews have been used, this research targets the identification of deceit in short texts where domain and context are not available. We aim to build deception, age, and gender classifiers using short texts, and also explore the prediction of gender and age in deceptive content. Moreover, we present an analysis of the topics discussed by deceivers given their age and gender based on the assumption that, when lying in an open domain setting, deceivers will show natural bias towards specific topics related to gender and age.

2 Related work

To date, several studies have explored the identification of deceptive content in a variety of domains, including online dating, forums, social networks, and consumer reviews. (Toma and Hancock, 2010) conducted linguistic analyses in online dating profiles and identified correlations between deceptive profiles and self references, negations, and lower levels of words usage. A study for deception detection on essays and product reviews is presented in (Feng et al., 2012). (Ott et al., 2011) addressed the identification of spam in consumer reviews and also studied the human capability of detecting deceptive reviews, which was found not better than chance. In a following study, (Ott et al., 2013) presented an analysis of the sentiment associated to deceitful reviews focusing particularly in those containing negative

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1120–1125,
Lisbon, Portugal, 17-21 September 2015.

sentiment as it largely affects consumer purchase decisions. More recently (Yu et al., 2015) presented a study where authors analyze the role of deception in online networks by detecting deceptive groups in a social elimination-game.

This previous work has shown the effectiveness of features derived from text analysis, which frequently includes basic linguistic representations such as n-grams and sentence counts statistics (Mihalcea and Strapparava, 2009; Ott et al., 2011) and also more complex linguistic features derived from syntactic context free grammar trees and part of speech tags (Feng et al., 2012; Xu and Zhao, 2012). Other studies have focused on deception clues inspired from psychological studies. For instance, following the hypothesis that deceivers might create less complex sentences (DePaulo et al., 2003), researchers have incorporated syntactic complexity measures into the analysis. (Yancheva and Rudzicz, 2013) presented a study based on the analysis of syntactic units and found that syntactic complexity correlates with deceiver's age. Psycholinguistics lexicons, such as Linguistic Inquiry and Word Count (LIWC) (Pennebaker and Francis, 1999), have also been used to build deception models using machine learning approaches (Mihalcea and Strapparava, 2009; Almela et al., 2012) and showed that the use of semantic information is helpful for the automatic identification of deceit.

While there is a significant body of work on computational deception detection, except for (Yancheva and Rudzicz, 2013) who considered the relation between syntactic constructs and deceivers' age, to our knowledge there are no computational analyses of demographics in deceptive content. However, there have been a number of psychological studies on the role of gender and age in deceptive behavior. These studies have found interesting associations between deception and gender. For instance, (Toma et al., 2008) identified differences in self-presentation among genders. In this study men were found to lie more about their height and women lied more about their weight. (Kaina et al., 2011) found that females are more easily detectable when lying than their male counterparts. (Tilley et al., 2005) reported that females are more successful in deception detection than male receivers.

3 Open Domain Deception Dataset

We started our study by collecting a new open domain deception dataset consisting of freely contributed truths and lies. We used Amazon Mechanical Turk and asked each worker to contribute seven lies and seven truths, on topics of their own choice, each of them consisting of one single sentence. In an attempt to obtain truths and lies that represent everyday lying behavior, we asked our contributors to provide plausible lies and avoid non-commonsensical statements such as "I can fly." Since we did not enforce a particular topic, resulting truths and lies are open domain. Sample truths and lies are presented in Table 1. Note that the collected lies might include statements that are somehow unrealistic, even if plausible, e.g., "I own two Ferraris, one red and one black". We decided to also include these statements in order to aid the identification of differences in deceivers and true-tellers language, as we hypothesize that they might help reveal topics that naturally occur in truths and lies.

Additionally, we collect demographic data from the contributors, including their gender, age, country of origin, and education level. To avoid spam, contributions were manually verified by one of the paper authors. The final dataset consists of 7168 sentences from 512 unique contributors. Since each contributor provided seven lies and seven truths the dataset contains a total of 3584 truths and 3584 lies respectively. Participant's ages range from 18 to 72 years, with an average age of 34.14 and a standard deviation of 12.67.

4 Features

In this section, we describe the sets of features extracted, which will then be used to build our classifiers.

Unigrams We extract unigrams derived from the bag of words representation of truths and lies present in our dataset.

Shallow and deep syntax features These features consist of part of speech (POS) tags and lexicalized production rules derived from Probabilistic Context Free Grammar (PCFG) trees, obtained with the Berkeley Parser (Petrov et al., 2006).

Female	
Lie	Truth
I won 1 billion dollars in the Illinois state lottery last year and gave it all away to my mother.	My daughter is my best friend in the whole wide world, and i would give my life for hers.
On my last birthday i turned 119 years old and went sky diving as a gift to myself.	I graduated with a degree in information systems 10 years ago and still can't find a good job.
I'm allergic to alcohol	Giraffes are taller than zebras.
Male	
Lie	Truth
Barak obama was my guest last night; he offered me the administrative assistant job at white house in Washington.	Internet is one of the greatest invention of history of humankind with its ability to speed up the communication.
I own two Ferraris, one red and one black	I love to play soccer with my friends
I wake up at 11 o clock every day	I wake up at 6 am because I have to work at 7 am

Table 1: Sample open-domain lies and truths provided by male and a female participants

Semantic features These features include the 80 semantic classes present in the LIWC lexicon. Each feature represents the number of words in a sentence belonging to a specific semantic class.

Readability and Syntactic Complexity features This set includes the Flesch-Kincaid and Gunning Fog readability scores and 14 indexes of syntactic complexity derived from the syntactic analysis of each sentence; performed with the tool provided by Lu (Lu, 2010).

5 Classification of Deception, Gender, and Age in Short Texts

Our first experiment seeks to evaluate whether deception detection can be conducted using the open domain deception dataset described above. We performed the evaluations at user level, by collapsing all the lies from one user into one instance, and all the truths into another instance.

We build deception classifiers using the SVM algorithm[1] and the different sets of features. We performed a five-fold cross-validation, by training each time on 80% of the users and testing on the remaining 20%. During our evaluations truths and lies pertaining to a particular user were either on the training or testing set. Classification results on individual and combined sets of features are presented in Table 3. The best performing set of features are the POS tags, followed by features derived from production rules. The remaining sets of features achieved accuracy values ranging from 54% to 65%, which still represent a noticeable improvement over the random baseline. Note that we experimented with a few more feature sets combinations, including the use of all the features together, however we did not observe significant improvements.

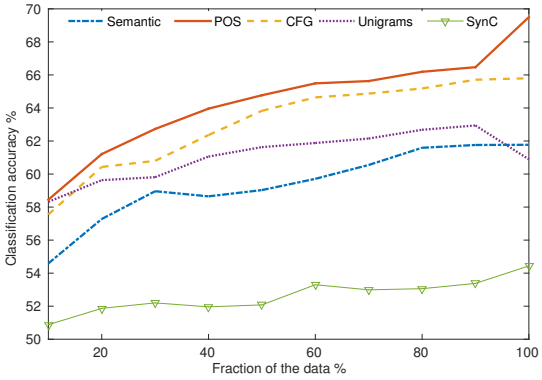

Figure 1: Learning curves for deception detection using five feature sets

To analyze the impact of the amount of data on the classifier learning process, we plot the learning curves on the different sets of features using incremental amounts of data as shown in Figure 1. Evaluations were conducted using five-fold cross validations on each incremental fraction of data. The learning trend suggests that most classifiers benefit from increasing amounts of training data.

[1] As implemented in the Weka toolkit, with default parameter settings.

Gender	Female	298
	Male	214
Age	Young ($\leq$ 35 years)	319
	Middle-aged/Elder (>35 years)	193

Table 2: Class distribution for gender and age

Feature set	Deception	Gender	Age
Baseline	50.00%	58.00%	62.00%
Unigrams	60.89%	54.25%	51.12%
Semantic	60.21%	57.28%	61.83%
POS	69.50%	49.95%	52.39%
CFG	65.39%	52.19%	54.74%
Readability	54.44%	58.16%	62.26%
Uni+Semantic	62.17%	63.04%	51.51%

Table 3: SVM classifiers trained for three prediction tasks: deception, gender, and age.

However, except for the POS features, the overall performance seems to stabilize when using 90% of the training data.

As a second experiment, we evaluate the ability of the classifier to predict gender and age in short open domain deceptive texts. Given the contributors' age distribution, which lies mainly in the range of 30-45 years, we opted to cluster the participants age into into two groups: young ($\leq$ 35 years) and middle-aged/elder (>35 years). Class distributions for age and gender are shown in Table 2. We performed the age prediction task on the two groups using the different sets of features and SVM classifiers. Classification accuracies are shown in Table 3. Reported baselines consist of a majority class baseline. Results show low to moderate improvement over the baseline for gender classification, with the combination of semantic features and unigrams being the best performing feature set. However, our classifiers performed poorly in the age prediction task, with accuracies below the majority class baseline.

Overall, the results suggest that age and gender prediction are challenging tasks when conducted in open domain deception data. One possible explanation for this is that the lack of context introduces noise into the analysis. For instance, the following sentence: "I'm 50 years old" can belong to either a male or a female, and it might be a lie for younger people or a truth for older people.

Lies			
Male		Female	
Other	2.22	Certain	1.87
Negate	2.08	Negate	1.63
Certain	2.06	You	1.59
Death	2.04	Motion	1.47
Anger	2.03	Down	1.45
You	1.77	Money	1.35
Friends	1.71	Anger	1.28
Othref	1.67	Future	1.20
Truths			
Male		Female	
Religion	1.67	Sleep	1.64
Family	1.65	Religion	1.61
Groom	1.60	See	1.50
Music	1.49	Discrepancy	1.39
Sports	1.45	Anxiety	1.36
School	1.42	Posfeel	1.33
Posfeel	1.35	Metaphor	1.33
Feel	1.32	TV	1.31

Table 4: Results from LIWC word class analysis for short open domain truths and lies.

6 Analyzing Language Used by Deceivers Given Age and Gender

In order to explore language differences among deceivers and true-tellers, we use the linguistic ethnography method (Mihalcea and Pulman, 2009) and obtain the most dominant semantic word classes in the LIWC lexicon associated to truth and lies provided by males and females. Results are shown in Table 4. From this table, we observe interesting patterns in word usage that are shared among genders. For instance, spontaneous lies often include *negation*, *certain*, and *you* words, which is in line with previous work on domain-specific deception (Mihalcea and Strapparava, 2009) that suggested that liars try to reinforce their lies through the use of stronger wording and detachment from the self. On the other hand, people appear to be less likely to lie when talking about their *family*, *religion*, and describing *positive* experiences. There are also LIWC classes associated to a specific gender. Male lies contain references to friends and others, while female lies contain references to money and future. Similarly, female true-tellers use metaphor words while male true-tellers use words related to sports and music.

Lies			
Age 18-34		Age 35-65	
Certain	2.04	Assent	2.13
Anger	1.98	Certain	1.83
Negate	1.82	Negate	1.68
Other	1.76	Anxiety	1.64
You	1.72	You	1.64
Down	1.64	Motion	1.54
Othref	1.53	Money	1.50
Death	1.49	Optim	1.38
Truths			
Age 18-34		Age 35-65	
Religion	1.83	Music	1.43
Tv	1.48	Sleep	1.42
Anxiety	1.45	Feel	1.34
Posfeel	1.37	Posfeel	1.33
See	1.30	See	1.31
Music	1.30	Sexual	1.28
School	1.29	Religion	1.27
Inhib	1.28	Family	1.25

Table 5: Results from LIWC word class analysis for short open domain truths and lies.

We also evaluate differences in word usage that might be attributed to deceiver's age. Resulting dominant classes and their scores are presented in Table 5. The analyses show interesting differences for deceiver's word usage across age. For instance, regardless of their gender, older deceivers use references to *anxiety*, *money*, and *motion*. On the other hand, younger deceivers language includes *anger*, *negate*, and *death* words. These differences suggest that indeed gender and age play a role on people words choices while fabricating lies.

7 Conclusions

In this paper, we presented our initial experiments in open domain deception detection. We targetted the deception detection on short text to address the cases where context is not available. In real settings, this can be useful when receiving a text message or when looking at anonymous posts in forums. We collected a new deception dataset consisting of one-sentence truths and lies, along with the demographics of the deceivers. Through several experiments, we showed that this data can be used to build deception classifiers for short open domain text. However, the classifiers do not perform very well while trying to predict gender and age. We further explored linguistic differences in deceptive content that relate to deceivers gender and age and found evidence that both age and gender play an important role on people's word choices when fabricating lies.

The dataset introduced in this paper is publicly available from http://lit.eecs.umich.edu.

Acknowledgments

This material is based in part upon work supported by National Science Foundation awards #1344257 and #1355633, by grant #48503 from the John Templeton Foundation, and by DARPA-BAA-12-47 DEFT grant #12475008. 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, the John Templeton Foundation, or the Defense Advanced Research Projects Agency.

References

Á. Almela, R. Valencia-García, and P. Cantos. 2012. Seeing through deception: A computational approach to deceit detection in written communication. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 15–22, Avignon, France, April. Association for Computational Linguistics.

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

S. Feng, R. Banerjee, and Y. Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 171–175, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Kaina, M.G. Ceruti, K. Liu, S.C. McGirr, and J.B. Law. 2011. Deception detection in multicultural coalitions: Foundations for a cognitive model. Technical report, DTIC Document.

X. Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

R. Mihalcea and S. Pulman. 2009. Linguistic ethnography: Identifying dominant word classes in text. In *Computational Linguistics and Intelligent Text Processing*, pages 594–602. Springer.

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

M. Ott, Y. Choi, C. Cardie, and J. Hancock. 2011. 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*, HLT '11, pages 309–319, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Ott, C. Cardie, and J. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Short Papers*, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

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

S. Petrov, L. Barrett, R. Thibaux, and D. Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 433–440, Stroudsburg, PA, USA. Association for Computational Linguistics.

P. Tilley, J. F. George, and K. Marett. 2005. Gender differences in deception and its detection under varying electronic media conditions. In *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 1 - Volume 01*, HICSS '05, pages 24.2–, Washington, DC, USA. IEEE Computer Society.

C. L. Toma and J. T. Hancock. 2010. Reading between the lines: Linguistic cues to deception in online dating profiles. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 5–8, New York, NY, USA. ACM.

C. L Toma, J. T Hancock, and N. B Ellison. 2008. Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin*, 34(8):1023–1036.

Q. Xu and H. Zhao. 2012. Using deep linguistic features for finding deceptive opinion spam. In *Proceedings of COLING 2012: Posters*, pages 1341–1350, Mumbai, India, December. The COLING 2012 Organizing Committee.

M. Yancheva and F. Rudzicz. 2013. Automatic detection of deception in child-produced speech using syntactic complexity features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 944–953, Sofia, Bulgaria, August. Association for Computational Linguistics.

D. Yu, Y. Tyshchuk, H. Ji, and W. A. Wallace. 2015. Detecting deceptive groups using conversations and network analysis. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 857–866.