Cross-cultural Deception Detection

Verónica Pérez-Rosas
Computer Science and Engineering
University of North Texas
veronicaperezrosas@my.unt.edu

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

Abstract

In this paper, we address the task of cross-cultural deception detection. Using crowdsourcing, we collect three deception datasets, two in English (one originating from United States and one from India), and one in Spanish obtained from speakers from Mexico. We run comparative experiments to evaluate the accuracies of deception classifiers built for each culture, and also to analyze classification differences within and across cultures. Our results show that we can leverage cross-cultural information, either through translation or equivalent semantic categories, and build deception classifiers with a performance ranging between 60-70%.

1 Introduction

The identification of deceptive behavior is a task that has gained increasing interest from researchers in computational linguistics. This is mainly motivated by the rapid growth of deception in written resources, and in particular in Web content, including product reviews, online dating profiles, and social networks posts (Ott et al., 2011).

To date, most of the work presented on deception detection has focused on the identification of deceit clues within a specific language, where English is the most commonly studied language. However, a large portion of the written communication (e.g., e-mail, chats, forums, blogs, social networks) occurs not only between speakers of English, but also between speakers from other cultural backgrounds, which poses important questions regarding the applicability of existing deception tools. Issues such as language, beliefs, and moral values may influence the way people deceive, and therefore may have implications on the construction of tools for deception detection.

In this paper, we explore within- and across-culture deception detection for three different cultures, namely United States, India, and Mexico. Through several experiments, we compare the performance of classifiers that are built separately for each culture, and classifiers that are applied across cultures, by using unigrams and word categories that can act as a cross-lingual bridge. Our results show that we can achieve accuracies in the range of 60-70%, and that we can leverage resources available in one language to build deception tools for another language.

2 Related Work

Research to date on automatic deceit detection has explored a wide range of applications such as the identification of spam in e-mail communication, the detection of deceitful opinions in review websites, and the identification of deceptive behavior in computer-mediated communication including chats, blogs, forums and online dating sites (Peng et al., 2011; Toma et al., 2008; Ott et al., 2011; Toma and Hancock, 2010; Zhou and Shi, 2008).

Techniques used for deception detection frequently include word-based stylometric analysis. Linguistic clues such as n-grams, count of used words and sentences, word diversity, and self-references are also commonly used to identify deception markers. An important resource that has been used to represent semantic information for the deception task is the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker and Francis, 1999). LIWC provides words grouped into semantic categories relevant to psychological processes, which have been used successfully to perform linguistic profiling of true tellers and liars (Zhou et al., 2003; Newman et al., 2003; Rubin, 2010). In addition to this, features derived from syntactic Context Free Grammar parse trees, and part of speech have also been found to aid the deceit detection (Feng et al., 2012; Xu and Zhao, 2012).

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 440–445,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

While most of the studies have focused on English, there is a growing interest in studying deception for other languages. For instance, (Fornaciari and Poesio, 2013) identified deception in Italian by analyzing court cases. The authors explored several strategies for identifying deceptive clues, such as utterance length, LIWC features, lemmas and part of speech patterns. (Almela et al., 2012) studied the deception detection in Spanish text by using SVM classifiers and linguistic categories, obtained from the Spanish version of the LIWC dictionary. A study on Chinese deception is presented in (Zhang et al., 2009), where the authors built a deceptive dataset using Internet news and performed machine learning experiments using a bag-of-words representation to train a classifier able to discriminate between deceptive and truthful cases.

It is also worth mentioning the work conducted to analyze cross-cultural differences. (Lewis and George, 2008) presented a study of deception in social networks sites and face-to-face communication, where authors compare deceptive behavior of Korean and American participants, with a subsequent study also considering the differences between Spanish and American participants (Lewis and George, 2009). In general, research findings suggest a strong relation between deception and cultural aspects, which are worth exploring with automatic methods.

3 Datasets

We collect three datasets for three different cultures: United States (English-US), India (English-India), and Mexico (Spanish-Mexico). Following (Mihalcea and Strapparava, 2009), we collect short deceptive and truthful essays for three topics: opinions on Abortion, opinions on Death Penalty, and feelings about a Best Friend.

For English-US and English-India, we use Amazon Mechanical Turk with a location restriction, so that all the contributors are from the country of interest (US and India). We collect 100 deceptive and 100 truthful statements for each of the three topics. To avoid spam, each contribution is manually verified by one of the authors of this paper.For Spanish-Mexico, while we initially attempted to collect data also using Mechanical Turk, we were not able to receive enough contributions. We therefore created a separate web interface to collect data, and recruited participants through contacts of the paper's authors. The overall process was significantly more time consuming than for the other two cultures, and resulted in fewer contributions, namely 39+39 statements for Abortion, 42+42 statements for Death Penalty, and 94+94 statements for Best Friend. For all three cultures, the participants first provided their truthful responses, followed by the deceptive ones.

Interestingly, for all three cultures, the average number of words for the deceptive statements (62 words) is significantly smaller than for the truthful statements (81 words), which may be explained by the added difficulty of the deceptive process, and is in line with previous observations about the cues of deception (DePaulo et al., 2003).

4 Experiments

Through our experiments, we seek answers to the following questions. First, what is the performance for deception classifiers built for different cultures? Second, can we use information drawn from one culture to build a deception classifier for another culture? Finally, what are the psycholinguistic classes most strongly associated with deception/truth, and are there commonalities or differences among languages?

In all our experiments, we formulate the deception detection task in a machine learning framework, where we use an SVM classifier to discriminate between deceptive and truthful statements.[1]

4.1 What is the performance for deception classifiers built for different cultures?

We represent the deceptive and truthful statements using two different sets of features. First we use unigrams obtained from the statements corresponding to each topic and each culture. To select the unigrams, we use a threshold of 10, where all the unigrams with a frequency less than 10 are dropped. Since previous research suggested that stopwords can contain linguistic clues for deception, no stopword removal is performed.

Experiments are performed using a ten-fold cross validation evaluation on each dataset.Using the same unigram features, we also perform cross-topic classification, so that we can better understand the topic dependence. For this, we train the SVM classifier on training data consisting of a merge of two topics (e.g., Abortion + Best Friend) and test on the third topic (e.g., Death Penalty). The results for both within- and cross-topic are shown in the last two columns of Table 1.

[1] We use the SVM classifier implemented in the Weka toolkit, with its default settings.

	LIWC					Unigrams	
Topic	Linguistic	Psychological	Relativity	Personal	All	Within-topic	Cross-topic
		English-US					
Abortion	72.50%	68.75%	44.37%	67.50%	73.03%	63.75%	80.36%
Best Friend	75.98%	68.62%	58.33%	54.41%	73.03%	74.50%	60.78%
Death Penalty	60.36%	54.50%	49.54%	50.45%	58.10%	58.10%	77.23%
Average	69.61%	63.96%	50.75%	57.45%	69.05%	65.45%	72.79%
		English-India					
Abortion	56.00%	48.50%	46.50%	48.50%	56.00%	46.00%	50.00%
Best Friend	68.18%	68.62%	54.55%	53.18%	71.36%	60.45%	57.23%
Death Penalty	56.00%	52.84%	57.50%	53.50%	63.50%	57.50%	54.00%
Average	60.06%	59.19%	52.84%	51.72%	63.62%	54.65%	53.74%
		Spanish-Mexico					
Abortion	73.17%	67.07%	48.78%	51.22%	62.20%	52.46%	57.69%
Best Friend	72.04%	74.19%	67.20%	54.30%	75.27%	66.66%	50.53%
Death Penalty	73.17%	67.07%	48.78%	51.22%	62.20%	54.87%	63.41%
Average	72.79%	69.45%	54.92%	52.25%	67.89%	57.99%	57.21%

Table 1: Within-culture classification, using LIWC word classes and unigrams. For LIWC, results are shown for within-topic experiments, with ten-fold cross validation. For unigrams, both within-topic (ten-fold cross validation on the same topic) and cross-topic (training on two topics and testing on the third topic) results are reported.

Second, we use the LIWC lexicon to extract features corresponding to several word classes. LIWC 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 classes relevant to psychological processes (e.g., emotion, cognition), which in turn are grouped into four broad categories[2] namely: linguistic processes, psychological processes, relativity, and personal concerns. A feature is generated for each of the 70 word classes by counting the total frequency of the words belonging to that class. We perform separate evaluations using each of the four broad LIWC categories, as well as using all the categories together. The results obtained with the SVM classifier are shown in Table 1.

Overall, the results show that it is possible to discriminate between deceptive and truthful cases using machine learning classifiers, with a performance superior to a random baseline which for all datasets is 50% given an even class distribution. Considering the unigram results, among the three cultures considered, the deception discrimination works best for the English-US dataset, and this is also the dataset that benefits most from the larger amount of training data brought by the cross-topic experiments. In general, the cross-topic evaluations suggest that there is no high topic dependence in this task, and that using deception data from different topics can lead to results that are comparable to the within-topic data. Interestingly, among the three topics considered, the Best Friend topic has consistently the highest within-topic performance, which may be explained by the more personal nature of the topic, which can lead to clues that are useful for the detection of deception (e.g., references to the self or personal relationships).

Regarding the LIWC classifiers, the results show that the use of the LIWC classes can lead to performance that is generally better than the one obtained with the unigram classifiers. The explicit categorization of words into psycholinguistic classes seems to be particularly useful for the languages where the words by themselves did not lead to very good classification accuracies. Among the four broad LIWC categories, the linguistic category appears to lead to the best performance as compared to the other categories. It is notable that in Spanish, the linguistic category by itself provides results that are better than when all the LIWC classes are used, which may be due to the fact that Spanish has more explicit lexicalization for clues that may be relevant to deception (e.g., verb tenses, formality).

4.2 Can we use information drawn from one culture to build a deception classifier in another culture?

In the next set of experiments, we explore the detection of deception using training data originating from a different culture. As with the within-culture

[2] http://www.liwc.net/descriptiontable1.php

Topic	Linguistic	Psychological	Relativity	Personal	All LIWC	Unigrams
Training: English-US Test: English-India						
Abortion	58.00%	51.00%	48.50%	51.50%	52.25%	57.89%
Best Friend	66.36%	47.27%	48.64%	50.45%	59.54%	51.00%
Death Penalty	54.50%	50.50%	50.00%	48.50%	53.5%	59.00%
Average	59.62%	49.59%	49.05%	50.15%	55.10%	55.96%
Training: English-India Test: English-US						
Abortion	71.32%	47.49%	43.38%	45.82%	62.50%	55.51%
Best Friend	59.74%	49.35%	51.94%	49.36%	55.84%	53.20%
Death Penalty	51.47%	44.11%	54.88%	50.98%	39.21%	50.71%
Average	60.87%	46.65%	50.06%	48.72%	52.51%	54.14%
Training: English-US Test: Spanish-Mexico						
Abortion	70.51%	46.15%	50.00%	52.56%	53.85%	61.53%
Best Friend	69.35%	52.69%	51.08%	46.77%	67.74%	65.03%
Death Penalty	54.88%	54.88%	53.66%	50.00%	62.19%	59.75%
Average	64.92%	51.24%	51.58%	49.78%	61.26%	62.10%
Training: English-India Test: Spanish-Mexico						
Abortion	48.72%	50.00%	47.44%	42.31%	43.58%	55.12 %
Best Friend	68.28%	63.44%	56.45%	54.84%	60.75%	67.20%
Death Penalty	60.98%	53.66%	54.88%	60.98%	59.75%	51.21%
Average	59.32%	55.70%	52.92%	52.71%	54.69%	57.84%

Table 2: Cross-cultural experiments using LIWC categories and unigrams

experiments, we use unigrams and LIWC features. For consistency across the experiments, given that the size of the Spanish dataset is different compared to the other two datasets, we always train on one of the English datasets.

To enable the unigram based experiments, we translate the two English datasets into Spanish by using the Bing API for automatic translation.[3] As before, we extract and keep only the unigrams with frequency greater or equal to 10. The results obtained in these cross-cultural experiments are shown in the last column of Table 2.

In a second set of experiments, we use the LIWC word classes as a bridge between languages. First, each deceptive or truthful statement is represented using features based on the LIWC word classes. Next, since the same word classes are used in both the English and the Spanish LIWC lexicons, this LIWC-based representation is independent of language, and therefore can be used to perform cross-cultural experiments. Table 2 shows the results obtained with each of the four broad LIWC categories, as well as with all the LIWC word classes.

We also attempted to combine unigrams and LIWC features. However, in most cases, no improvements were noticed with respect to the use of unigrams or LIWC features alone. We are not reporting these results due to space limitation.

These cross-cultural evaluations lead to several findings. First, we can use data from a culture to build deception classifiers for another culture, with performance figures better than the random baseline, but weaker than the results obtained with within-culture data. An important finding is that LIWC can be effectively used as a bridge for cross-cultural classification, with results that are comparable to the use of unigrams, which suggests that such specialized lexicons can be used for cross-cultural or cross-lingual classification. Moreover, using only the linguistic category from LIWC brings additional improvements, with absolute improvements of 2-4% over the use of unigrams. This is an encouraging result, as it implies that a semantic bridge such as LIWC can be effectively used to classify deception data in other languages, instead of using the more costly and time consuming unigram method based on translations.

4.3 What are the psycholinguistic classes most strongly associated with deception/truth?

The final question we address is concerned with the LIWC classes that are dominant in deceptive and truthful text for different cultures. We use the method presented in (Mihalcea and Strapparava, 2009), which consists of a metric that measures the saliency of LIWC classes in deceptive versus truthful data. Following their strategy, we first create a corpus of deceptive and truthful text using a mix of all the topics in each culture. We then calculate

[3] http://http://http://www.bing.com/dev/en-us/dev-center

Class	Score	Sample words	Class	Score	Sample words
English-US					
Deceptive			Truthful		
Metaph	1.77	Die,died,hell,sin,lord	Insight	0.68	Accept,believe,understand
Other	1.46	He,her,herself,him	I	0.66	I,me,my,myself,
You	1.41	Thou,you	Optimism	0.65	accept, hope, top, best
Othref	1.18	He,her,herself,him	We	0.55	Our,ourselves,us,we,
Negemo	1.18	Afraid,agony,awful,bad	Friends	0.46	Buddies,friend
English-India					
Deceptive			Truthful		
Negate	1.49	Cannot,neither,no,none	Past	0.78	Happened,helped,liked,listened
Physical	1.46	Heart,ill,love,loved,	I	0.66	I,me,mine,my
Future	1.42	Be,may,might,will	Optimism	0.65	Accept,accepts,best,bold,
Other	1.17	He,she, himself,herself	We	0.55	Our,ourselves,us,we
Humans	1.08	Adult,baby,children,human	Friends	0.46	Buddies,companion,friend,pal
Spanish-Mexico					
Deceptive			Truthful		
Certain	1.47	Jamás(never),siempre(always)	Optimism	0.66	Aceptar(accept),animar(cheer)
Humans	1.28	Bebé(baby),persona(person)	Self	0.65	Conmigo(me),tengo(have),soy(am)
You	1.26	Eres(are),estas(be),su(his/her)	We	0.58	Estamos(are),somos(be),tenemos(have)
Negate	1.25	Jamás(never),tampoco(neither)	Friends	0.37	Amigo/amiga(friend),amistad(friendship)
Other	1.22	Es(is),esta(are),otro(other)	Past	0.32	Compartimos(share),vivimos(lived)

Table 3: Top ranked LIWC classes for each culture, along with sample words

the dominance for each LIWC class, and rank the classes in reversed order of their dominance score. Table 3 shows the most salient classes for each culture, along with sample words.

This analysis shows some interesting patterns. There are several classes that are shared among the cultures. For instance, the deceivers in all cultures make use of negation, negative emotions, and references to others. Second, true tellers use more optimism and friendship words, as well as references to themselves. These results are in line with previous research, which showed that LIWC word classes exhibit similar trends when distinguishing between deceptive and non-deceptive text (Newman et al., 2003). Moreover, there are also word classes that only appear in some of the cultures; for example, time classes (Past, Future) appear in English-India and Spanish-Mexico, but not in English-US, which in turn contains other classes such as Insight and Metaph.

5 Conclusions

In this paper, we addressed the task of deception detection within- and across-cultures. Using three datasets from three different cultures, each covering three different topics, we conducted several experiments to evaluate the accuracy of deception detection when learning from data from the same culture or from a different culture. In our evaluations, we compared the use of unigrams versus the use of psycholinguistic word classes.

The main findings from these experiments are: 1) We can build deception classifiers for different cultures with accuracies ranging between 60-70%, with better performance obtained when using psycholinguistic word classes as compared to simple unigrams; 2) The deception classifiers are not sensitive to different topics, with cross-topic classification experiments leading to results comparable to the within-topic experiments; 3) We can use data originating from one culture to train deception detection classifiers for another culture; the use of psycholinguistic classes as a bridge across languages can be as effective or even more effective than the use of translated unigrams, with the added benefit of making the classification process less costly and less time consuming.

The datasets introduced in this paper are publicly available from http://nlp.eecs.umich.edu.

Acknowledgments

This material is based in part upon work supported by National Science Foundation awards #1344257 and #1355633 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 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.

T. Fornaciari and M. Poesio. 2013. Automatic deception detection in italian court cases. *Artificial Intelligence and Law*, 21(3):303–340.

C. Lewis and J. George. 2008. Cross-cultural deception in social networking sites and face-to-face communication. *Comput. Hum. Behav.*, 24(6):2945–2964, September.

C. Lewis and Giordano G. George, J. 2009. A cross-cultural comparison of computer-mediated deceptive communication. In *Proceedings of Pacific Asia Conference on Information Systems*.

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

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.

H. Peng, C. Xiaoling, C. Na, R. Chandramouli, and P. Subbalakshmi. 2011. Adaptive context modeling for deception detection in emails. In *Proceedings of the 7th international conference on Machine learning and data mining in pattern recognition*, MLDM'11, pages 458–468, Berlin, Heidelberg. Springer-Verlag.

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

V. Rubin. 2010. 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.

C. Toma and J. 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. Toma, J. Hancock, and N. 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.

H. Zhang, S. Wei, H. Tan, and J. Zheng. 2009. Deception detection based on svm for chinese text in cmc. In *Information Technology: New Generations, 2009. ITNG '09. Sixth International Conference on*, pages 481–486, April.

L. Zhou and D. Shi, Y.and Zhang. 2008. A statistical language modeling approach to online deception detection. *IEEE Trans. on Knowl. and Data Eng.*, 20(8):1077–1081, August.

L Zhou, D. Twitchell, T Qin, J. Burgoon, and J. Nunamaker. 2003. An exploratory study into deception detection in text-based computer-mediated communication. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track1 - Volume 1*, HICSS '03, pages 44.2–, Washington, DC, USA. IEEE Computer Society.