On the Use of Homogenous Sets of Subjects in Deceptive Language Analysis

Tommaso Fornaciari
Center for Mind/Brain Sciences
University of Trento
tommaso.fornaciari@unitn.it

Massimo Poesio
Language and Computation Group
University of Essex
Center for Mind/Brain Sciences
University of Trento
massimo.poesio@unitn.it

Abstract

Recent studies on deceptive language suggest that machine learning algorithms can be employed with good results for classification of texts as truthful or untruthful. However, the models presented so far do not attempt to take advantage of the differences between subjects. In this paper, models have been trained in order to classify statements issued in Court as false or not-false, not only taking into consideration the whole corpus, but also by identifying more homogenous subsets of producers of deceptive language. The results suggest that the models are effective in recognizing false statements, and their performance can be improved if subsets of homogeneous data are provided.

1 Introduction

Detecting deceptive communication is a challenging task, but one that could have a number of useful applications. A wide variety of approaches to the discovery of deceptive statements have been attempted, ranging from using physiological sensors such as lie detectors to using neuroscience methods (Davatzikos et al., 2005; Ganis et al., 2003). More recently, a number of techniques have been developed for recognizing deception on the basis of the communicative behavior of subjects. Given the difficulty of the task, many such methods rely on both verbal and non-verbal behavior, to increase accuracy. So for instance De Paulo et al. (2003) considered more than 150 cues, verbal and non-verbal, directly observed through experimental subjects. But finding clues indicating deception through manual inspection is not easy. De Paulo et al. asserted that "behaviors that are indicative of deception can be indicative of other states and processes as well".

The same point is made in more recent literature: thus Frank et al. (2008) write "We find that there is no clue or clue pattern that is specific to deception, although there are clues specific to emotion and cognition", and they wish for "real-world databases, identifying base rates for malfeasant behavior in security settings, optimizing training, and identifying preexisting excellence within security organizations". Jensen et al. (2010) exploited cues coming from audio, video and textual data.

One solution is to let statistical and machine learning methods discover the clues. Work such as Fornaciari and Poesio (2011a,b); Newman et al. (2003); Strapparava and Mihalcea (2009) suggests that these techniques can perform reasonably well at the task of discovering deception even just from linguistic data, provided that corpora containing examples of deceptive and truthful texts are available. The availability of such corpora is not a trivial problem, and indeed, the creation of a realistic such corpus is one of the problems in which we invested substantial effort in our own previous work, as discussed in Section 3.

In the work discussed in this paper, we tackle an issue which to our knowledge has not been addressed before, due to the limitations of the datasets previously available: this is whether the individual difference between experimental subjects affect deception detection. In previous work, lexical (Fornaciari and Poesio, 2011a) and surface (Fornaciari and Poesio, 2011b) features were employed to classify deceptive statements issued in Italian Courts. In this study, we report the results

Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection, pages 39–47,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

of experiments in which our methods were trained either over the whole corpus or over smaller subsets consisting of the utterances produced by more homogenous subsets of subjects. These subsets were identified either automatically, by clustering subjects according to their language profile, or by using meta-information about the subjects included in the corpus, such as their gender.

The structure of the paper is as follows. In Section 2 some background knowledge is introduced. In Section 3 the data set is described. In Section 4 we discuss our machine learning and experimental methods. Finally, the results are presented in Section 5 and discussed in Section 6.

2 Background

2.1 Deceptive language analysis

From a methodological point of view, to investigate deceptive language gives rise to some tricky issues: first of all, the strategy chosen to collect data. The literature can be divided in two main families of studies:

- Field studies;
- Laboratory studies.

The first ones are usually interesting in forensic applications but in such studies verifying the sincerity of the statements is often complicated (Vrij, 2005). Laboratory studies, instead, are characterized by the artificiality of participants' psychological conditions: therefore their findings may not be generalized to deception encountered in real life.

Due to practical difficulties in collection and annotation of suitable data, in literature finding papers in which real life linguistic data are employed, where truthfulness is surely known, is less common and Zhou et al. (2008) complain about the lack of "data set for evaluating deception detection models". Just recently some studies tried to fill this gap, concerning both the English (Bachenko et al., 2008; Fitzpatrick and Bachenko, 2009) and Italian language (Fornaciari and Poesio, 2011a,b). Just the studies on Italian language come from data which have constituted the first nucleus of the corpus analysed here.

2.2 Stylometry

Our own work and that of other authors that recently employed machine learning techniques to detect deception in text employs techniques very similar to that of stylometry. Stylometry is a discipline which studies texts on the basis of their stylistic features, usually in order to attribute them to an author - giving rise to the branch of author attribution - or to get information about the author himself - this is the field of author profiling.

Stylometric analyses, which relies mainly on machine learning algorithms, turned out to be effective in several forensic tasks: not only the classical field of author profiling (Coulthard, 2004; Koppel et al., 2006; Peersman et al., 2011; Solan and Tiersma, 2004) and author attribution (Luyckx and Daelemans, 2008; Mosteller and Wallace, 1964), but also emotion detection (Vaassen and Daelemans, 2011) and plagiarism analysis (Stein et al., 2007). Therefore, from a methodological point of view, Deceptive Language Analysis is a particular application of stylometry, exactly like other branches of Forensic Linguistics.

3 Data set

3.1 False testimonies in Court

In order to study deceptive language, we created the DECOUR - DEception in COURt - corpus, better described in Fornaciari and Poesio (2012). DECOUR is a corpus constituted by the transcripts of 35 hearings held in four Italian Courts: Bologna, Bolzano, Prato and Trento. These transcripts report verbatim the statements issued by a total of 31 different subjects - four of which have been heard twice. All the hearings come from criminal proceedings for **calumny** and **false testimony** (artt. 368 and 372 of the Italian Criminal Code).

In particular, the hearings of DECOUR come mainly from two situations:

- the defendant for any criminal proceeding tries to use calumny against someone;
- a witness in any criminal proceeding lies for some reason.

In both cases, a new criminal proceeding arises, in which the subjects can issue new statements or not, and having as a body of evidence the transcript of the hearing held in the previous proceeding.

The crucial point is that DECOUR only includes text from individuals who in the end have been found guilty. Hence the proceeding ends with a judgment of the Court which summarize the facts, pointing out precisely the lies told by the speaker in order to establish his punishment. Thanks to the transcripts of the hearing and to the final judgment of the Court, it is possible to annotate the statements of the speakers on the basis of their truthfulness or untruthfulness, as follows.

3.2 Annotation and agreement

The hearings are dialogs, in which the judge, the public prosecutor and the lawyer pose questions to the witness/defendant who in turn has to give them answers. These answers are the object of investigation of this study. Each answer is considered a **turn**, delimited by the end of the previous and the beginning of the following intervention of another individual. Each turn is constituted by one or more **utterances**, delimited by punctuation marks: period, triple-dots, question and exclamation marks. Utterances are the analysis unit of DECOUR and have been annotated as **false**, **true** or **uncertain**. In order to verify the agreement in the judgments about truthfulness or untruthfulness of the utterances, three annotators separately annotated about 600 utterances. The agreement study concerning the three classes of utterances, described in detail in (Fornaciari and Poesio, 2012), showed that the agreement value was k=.57. Instead, if the problem is reduced to a binary task - that is, if true and uncertain utterances are collapsed into a single category of **not-false** utterances, opposed to the category of false ones - the agreement value is k=.64.

3.3 Corpus statistics

The whole corpus has been tokenized and sensitive data have been made anonymous, according to the previous agreement with the Courts. Then DECOUR has been lemmatized and POS-tagged using a version of TreeTagger[1] (Schmid, 1994) trained for Italian.

DECOUR is made up of 3015 utterances, which come from 2094 turns. 945 utterances have been annotated as false, 1202 as true and 868 as uncertain. The size of DECOUR is 41819 tokens, including punctuation blocks.

4 Methods

In this Section we first summarize our classification methods from previous work, then discuss the three experiments we carried out.

4.1 Classification methods

Each utterance is described by a feature vector. As in our previous studies (Fornaciari and Poesio, 2011a,b) three kinds of features were used.

First of all, the feature vectors include very basic linguistic information such as the length of utterances (with and without punctuation) and the number of words longer than six letters.

The second type of information are lexical features. These features have been collected making use of LIWC - Linguistic Inquiry and Word Count, a linguistic tool realized by Pennebaker et al. (2001) and widely employed in deception detection (Newman et al., 2003; Strapparava and Mihalcea, 2009). LIWC is based on a dictionary in which each term is associated with an appropriate set of syntactical, semantical and/or psychological categories. When a text is analysed with LIWC, the tokens of the text are compared with the LIWC dictionary. Every time a word present in the dictionary is found, the count of the corresponding categories grows. The output is a profile of the text which relies on the rate of incidence of the different categories in the text itself. LIWC also includes different dictionaries for several languages, amongst which Italian (Agosti and Rellini, 2007). Therefore it has been possible to apply LIWC to Italian deceptive texts, and the approximate 80 linguistic dimensions which constitute the Italian LIWC dictionary have been included as features of the vectors.

Lastly, frequencies of lemmas and part-of-speech n-grams were used. Five kinds of n-grams of lemmas and part-of-speech were taken into consideration: from unigrams to pentagrams. These frequency lists come from the part of DECOUR employed as training set. More precisely, they come from the utterances held as true or false of the training set, while the uncertain utterances have not been considered. In order to emphasize the collection of features effective in classifying true and false statements, frequency lists of n-grams have been built considering true and false utterances separately. This means that, in the training set, homologous frequency lists of n-

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

Table 1: The most frequent n-grams collected

N-grams	Lemmas	POS	Total
Unigrams	50	15	
Bigrams	40	12	
Trigrams	30	9	
Tetragrams	20	6	
Pentagrams	10	3	
Total	150	45	195

grams - unigrams, bigrams and so on - have been collected from the subset of true utterances *and* form the subset of false ones. From these lists, the most frequent n-grams have been collected, in a decreasing amount according to the length of the n-grams. Table 1 shows in detail the number of the most frequent lemmas and part-of-speech collected for the different n-grams. Then the couples of frequency lists were merged into one.

This procedure implies that the number of surface features is not determined *a priori*. In fact the 195 features indicated in Table 1, which are collected from true and false utterances, are unified in a list where each feature has to appear only once. Therefore, theoretically in the case of perfect identity of features in true and false utterances, a final list with the same 195 features would be obtained. In the opposite case, if the n-grams from true and false utterances would be completely different, a list of 195 + 195, then 390 n-grams would result. The aim of this procedure is to get a list of n-grams which could be as much as possible representative of the features of true and false utterances. Obviously, the smaller the overlap of the features of the two subsets, the greater the difference in the appearance of true and false utterances, and greater the hope to reach a good performance in the classification task.

We used the Support Vector Machine implementation in R (Dimitriadou et al., 2011). As specified above, the classes of the utterances are false vs. not-false, where the category of not-false utterances results from the union of the true and uncertain ones.

4.2 Corpus division

With the aim of training models able to classify the utterances of DECOUR as false or not-false, the corpus has been divided as follows:

Training set The 20 hearings coming from the Courts of Bologna and Bolzano have been employed as training set. In terms of analysis units, this means 2279 utterances, that is 75.59% of DECOUR. The features of the vectors come from this set of data.

Test set The 9 hearings of the Court of Trento have been employed as test set, in order to evaluate the effectiveness of the trained models. This test set was made up by 426 utterances, which are 14.13% of DECOUR.

Development set The 6 hearings of the Court of Prato have been employed as development set during the phase of choice and calibration of vector features, therefore this set of utterances is not directly involved in the results of the following experiments. The develpment set was constituted by 310 utterances, that is 10.28% of DECOUR.

In the various experimental conditions, some subsets of DECOUR have been taken into consideration. Hence, different hearings have been removed from the test and/or training set in order to carry out different experiments. Since the test sets vary in the different experiments, in relation to each of them different chance levels have been determined, in order to evaluate the effectiveness of the models' performance.

4.3 Experiments

Three experiments were carried out. In the first experiment, the entire corpus was used to train and test our algorithms. In the second and third experiment, sub-corpora were identified.

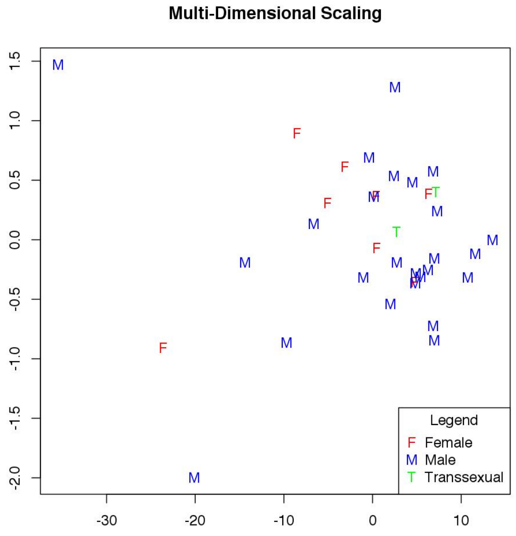
4.3.1 Experiment 1: whole test set

In the first experiment, the classification task has been carried out simply employing the training set and the test set as described above, in order to have a control as reference point in relation to the following experiments.

4.3.2 Experiment 2: no outliers

In the second experiment, a more homogeneous subset of DECOUR was obtained by automatically identifying and removing outliers. This was done in an unsupervised way by building vector descriptions of the hearings and clustering them. The features of these vectors were the same n-grams described above, collected from the whole

Figure 1: Multi-Dimensional Scaling of DECOURĖach entity corresponds to a hearing; the letters represent the sex of the speakers.



corpus (not from the only test set); their values were the mean values of the frequencies of the utterances belonging to the hearing.

This data set has been transformed into a matrix of between-hearing distances and a Multi-Dimensional Scaling - MDS function has been applied to this matrix (Baayen, 2008). Figure 1 shows the plot of MDS function. Each entity corresponds to a hearing, and is represented by a letter indicating the sex of the speaker. Getting a glimpse at Figure 1, it is possible to notice that, in general, almost all the hearings are quite close - that is, similar - to each other. Only three hearings seem to be clearly more peripheral than all the others, particularly the three most to the left in Figure 1. These hearings have been considered as outliers and shut out from the experiment. They are two hearings from Trento and one from Prato. In practice, it means that the training set, coming from the hearings of Bologna and Bolzano, remained the same as the previous experiment, while two hearings have been removed from the test set, which was constituted only by the hearings of Trento.

4.3.3 Experiment 3: only male speakers

Different from the previous one, the third experiment does not rely on a subset of data automatically identified. Instead, the subset comes from personal information concerning the subjects involved in the hearings. In fact, their sex, place of birth and age at the moment of the hearing are known. In this paper, places of birth and age have not been taken into consideration, since grouping them together in reliable categories raises issues that do not have a straightforward solution, and the size of the subsets of corpus which would be obtained must be taken into account.

Therefore this experiment has been carried out taking into consideration only the sex of the subjects, and in particular it concerned only the hearings involving men. This meant reducing the training set consistently, where seven hearings of women were present and thence removed. Instead from the test set just three hearings have been taken off, one involving a woman and two involving a transsexual.

4.4 Baselines

The chance levels for the various test sets have been calculated through Monte Carlo simulations, each one specific to every experiment. In each simulation, 100000 times a number of random predictions has been produced, in the same amount and with the same rate of false utterances of the test set employed in the single experiment. Then this random output was compared to the real sequence of false and not-false utterances of the test set, in order to count the amount of correct predictions. The rate of correct answers reached by less than 0.01% of the random predictions has been accepted as chance threshold for every experiment.

As a baseline, a simple majority baseline was computed: to classify each utterance as belonging to the most numerous class in the test set (not-false).

5 Results

The test set of the first experiemnt, carried out on the whole test set, was made up of 426 utterances, of which 190 were false, that is 44.60%. While the majority baseline is 55.40% of accuracy, a Monte Carlo simulation applied to the test set showed that the chance level was 59.60% of correct predictions. The results are shown in Table 2. The overall accuracy - almost 66% - is clearly above the chance level, being more than six points greater than the baseline.

Table 2: Whole training and test set

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-measure
False utterances	59	131	80.82%	31.05%	44.86%
True utterances	222	14	62.89%	94.07%	75.38%
Total	281	145			
Total percent	65.96%	34.04%			
Monte Carlo simulation	59.60%				
Majority baseline	55.40%				

Table 3: Test set without outliers

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-measure
False utterances	51	90	80.95%	36.17%	50.00%
True utterances	180	12	66.67%	93.75%	77.92%
Total	231	102			
Total percent	69.37%	30.63%			
Monte Carlo simulation	61.26%				
Majority baseline	57.66%				

Table 4: Training and test set with only male speakers

	Correctly classified entities	Incorrectly classified entities	Precision	Recall	F-measure
False utterances	32	85	74.42%	27.35%	40.00%
True utterances	179	11	67.80%	94.21%	78.85%
Total	211	96			
Total percent	68.73%	31.27%			
Monte Carlo simulation	63.19%				
Majority baseline	61.89%				

In the second experiment, the test set without outliers was made up of 333 utterances; 141 were false, which means 42.34% of the test set. The majority baseline was then at 57.66%, while the chance threshold determined with a Monte Carlo simulation had an accuracy rate of 61.26%. Table 3 shows the results of the analyses. Taking the outliers out of the test set allows tthe best performance of the three experiments to be reached. In fact the accuracy is more than 69%, which is more than eight points above the highest chance level of 61.26%.

In the third experimental condition, where only male speakers were considered, the training set was made up of 13 hearings and the test set of 6 hearings. The utterances in the test set were 307, of which 117 were false, meaning 38.11% of the test set. In this last case, the majority baseline is at 61.89% of accuracy, while according to a Monte Carlo simulation the chance level was 63.19%. The overall accuracy reached in this experiment, shown in Table 4, was more than 68%: higher than the first experiment, but in this case the lower amount of false utterances in the test set led to higher chance thresholds. Therefore the difference between performance and the chance level of 63.19% is now the smallest of all the experiments: just five points and half.

From the point of view of detection of false utterances, although with internal differences, all the experiments are placed in the same reference frame. In particular, the weak point in performance is always the recall of false utterances, which remains more or less at 30%. Instead the good news comes from the precision in recognizing them, which is close to 80%. Regarding true utterances, the recall is always good, being never lower than 93%, while the precision is close to 65%.

6 Discussion

The goal of this paper was to verify if restricting the analysis to more homogeneous subsets could improve the accuracy of our models. The results are mixed. On the one end, taking the outliers out of the corpus results in a remarkable improvement of accuracy in the classification task, in relation to the performance of the models tested on the whole test set. On the other end, in other cases - most clearly, considering only speakers of the male gender - we find no difference; our hypothesis is that any potential advantage derived from the increased homogeneity is offset by the reduction in training material (seven hearings are removed in this case). So the conclusion may be that increasing homogeneity is effective provided that the remaining set is still sufficiently large.

Regarding the models' capacity to detect false rather than true utterances, the difference between the respective recalls is noteworthy. In fact, while the recall of not-false utterances is very high, that of false ones is poor. In other words, the results indicate that an amount of false utterances is effectively so similar to the not-false ones, that the models are not able to detect them. One challenge for future studies is surely to find a way to detect some aspect currently neglected of deceptive language, which could be employed to widen the size of false utterances which can be recognized.

On the other hand, in the two more reliable experiments the precision in detecting false utterances was about 80%. This could suggest that an amount of false utterances exists, whose features are in some way peculiar and different from not-false ones. The data seem to show that this subset could be more or less one third of all the false utterances.

However, this study was not aimed to estimate the possible performance of the models in an hypothetic practical application. The experimental conditions taken into consideration, in fact, are considerably different from those that would be present in a real life analysis.

The main reason of this difference is that in a real case to classify every utterance of a hearing would not be requested. A lot of statements are irrelevant or perfectly known as true. Furthermore it would not make sense to classify all the utterances which have not propositional value, such as questions or meta-communicative acts. In the perspective of deception detection in a real life scenario, to classify this last kind of utterances is useless. Only a subset of the propositional statements should be classified. In a previous study, carried out on a selection of utterances with propositional value of a part of DECOUR, machine learning models reached an accuracy of 75% in classification task (Fornaciari and Poesio, 2011b). In that study, precision and recall of false utterances are also quite similar to those of this study, the first being about 90% and the second about 50%.

From a theoretical point of view, the present study suggests that it is possible to be relatively confident in the effectiveness of the models in the analysis of any kind of utterance. This means that deceptive language is at least in part different from the truthful one and stylometric analyses can detect it. If this is true, the rate of precision with which false statements are correctly classified should clearly exceed the chance level.

Also in this case, Monte Carlo simulation is taken as reference point. Out of the 100000 random trials carried out to determine the baseline for the first experiment, less than 0.01% had a precision greater than 57.90% in classifying false utterances, in front of a precision of the models at 80.82%. Regarding the second experiment, the threshold for precision related to false utterances was 58.15% against a precision of the models at 80.95%. In the third experiment, the baseline for precision was 55.55% and the performance of models was 74.42%. In every experiment the gap is about twenty points per cent. The same cannot be said about the recall of false utterances: the baselines of Monte Carlo simulations in the three experiments were about 51-54%, while the best models' performance (of the second experiment) did not exceed 36%.

The precision reached in recognizing false statements shows that the models were reliable in detection of deceptive language. On the other hand a remarkable amount of false utterances was not identified. The challenge for the future is to understand to which extent it will be possible to improve the recall in detecting false utterances, not losing and hopefully improving the relative precision. At that point, although in specific contexts, a computational linguistics' approach could be really employed to detect deception in real life scenarios.

7 Acknowledgements

To create DECOUR has been very complex, and it would not have been possible without the kind collaboration of a lot of people. Many thanks to Dr. Francesco Scutellari, President of the Court of Bologna, to Dr. Heinrich Zanon, President of the Court of Bolzano, to Dr. Francesco Antonio Genovese, President of the Court of Prato and to Dr. Sabino Giarrusso, President of the Court of Trento.

References

Agosti, A. and Rellini, A. (2007). The Italian LIWC Dictionary. Technical report, LIWC.net, Austin, TX.

Baayen, R. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge University Press.

Bachenko, J., Fitzpatrick, E., and Schonwetter, M. (2008). Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 41–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25(4):431–447.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D., Acharyya, M., Loughead, J., Gur, R., and Langleben, D. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, 28(3):663 – 668.

De Paulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 129(1):74–118.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A. (2011). r-cran-e1071. `http://mloss.org/software/view/94/`.

Fitzpatrick, E. and Bachenko, J. (2009). Building a forensic corpus to test language-based indicators of deception. *Language and Computers*, 71(1):183–196.

Fornaciari, T. and Poesio, M. (2011a). Lexical vs. surface features in deceptive language analysis. In *Proceedings of the ICAIL 2011 Workshop Applying Human Language Technology to the Law*, AHLTL 2011, pages 2–8, Pittsburgh, USA.

Fornaciari, T. and Poesio, M. (2011b). Sincere and deceptive statements in italian criminal proceedings. In *Proceedings of the International Association of Forensic Linguists Tenth Biennial Conference*, IAFL 2011, Cardiff, Wales, UK.

Fornaciari, T. and Poesio, M. (2012). Decour: a corpus of deceptive statements in italian courts. In *Proceedings of the eighth International Conference on Language Resources and Evaluation*, LREC 2012. In press.

Frank, M. G., Menasco, M. A., and O'Sullivan, M. (2008). Human behavior and deception detection. In Voeller, J. G., editor, *Wiley Handbook of Science and Technology for Homeland Security*. John Wiley & Sons, Inc.

Ganis, G., Kosslyn, S., Stose, S., Thompson, W., and Yurgelun-Todd, D. (2003). Neural correlates of different types of deception: An fmri investigation. *Cerebral Cortex*, 13(8):830–836.

Jensen, M. L., Meservy, T. O., Burgoon, J. K., and Nunamaker, J. F. (2010). Automatic, Multimodal Evaluation of Human Interaction. *Group Decision and Negotiation*, 19(4):367–389.

Koppel, M., Schler, J., Argamon, S., and Pennebaker, J. (2006). Effects of age and gender on blogging. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*.

Luyckx, K. and Daelemans, W. (2008). Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 513–520, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mosteller, F. and Wallace, D. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley.

Newman, M. L., Pennebaker, J. W., Berry, D. S., and Richards, J. M. (2003). Lying Words: Predicting Deception From Linguistic Styles. *Personality and Social Psychology Bulletin*, 29(5):665–675.

Peersman, C., Daelemans, W., and Van Vaerenbergh, L. (2011). Age and gender prediction on netlog data. *Presented at the 21st Meeting of Computational Linguistics in the Netherlands (CLIN21), Ghent, Belgium.*

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Lawrence Erlbaum Associates, Mahwah.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Solan, L. M. and Tiersma, P. M. (2004). Author identification in american courts. *Applied Linguistics*, 25(4):448–465.

Stein, B., Koppel, M., and Stamatatos, E. (2007). Plagiarism analysis, authorship identification, and near-duplicate detection pan'07. *SIGIR Forum*, 41:68–71.

Strapparava, C. and Mihalcea, R. (2009). The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In *Proceeding ACLShort '09 - Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*.

Vaassen, F. and Daelemans, W. (2011). Automatic emotion classification for interpersonal communication. In *2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*.

Vrij, A. (2005). Criteria-based content analysis - A Qualitative Review of the First 37 Studies. *Psychology, Public Policy, and Law*, 11(1):3–41.

Zhou, L., Shi, Y., and Zhang, D. (2008). A Statistical Language Modeling Approach to Online Deception Detection. *IEEE Transactions on Knowledge and Data Engineering*, 20(8):1077–1081.