Massimo Poesio / Tommaso Fornaciari
Queen Mary University / Bocconi University, Italy

Detecting deception in text using NLP methods

SIGNAL
May 30thth, 2018

Outline

Outline

Detecting Lies

- Lies are much more common in communication than we would expect (Vrij, 2008)
- There are different types of lies. Many of these are harmless or even beneficial, but some cases of lying are harmful and even criminal, so the ability to detect them could be useful
  - Lying in court
  - Deceptive online reviews
- People however are not very good at detecting lies
  - Their performance is not better than chance (Bond and De Paulo, 2006)
  - ...and does not improve after specific training, either (Levine *et al.*, 2005).
- Well-known tools like the polygraph ('lie detector' also are far less successful than we would expect

Approaches to Deception Detection

Several approaches to detection deception are possible, relying on the analysis of

- non-verbal
  - E.g., ‘averting gaze’
- physiological responses.
  - On the basis that liars are more agitated, they should sweat more, etc.
- verbal

Detecting Deception using verbal clues

- The surest way to tell that somebody is telling a lie is when you know for sure that what the person is saying is not true
  - E.g., Jeffrey Archer telling journalists he was on the phone with the Prime Minister when one of the journalists knew that wasn't possible because the Prime Minister was delivering a speech at the time
  - In most successful trials for lying in court, the police knew for sure that a certain statement is false
- However, in most cases we have no such certain knowledge. It may still be possible however to tell whether somebody is lying purely on the basis of the style they are using Vrij (2008)
  - On the assumption that liars feel guilty, and such guilt may 'leak through' their speech
  - Or that telling a lie requires effort, so the liar may use a simplified form of language, more generic terms, etc.
- For most applications, methods relying on verbal clues only are easiest to apply:
  - no need to ask the potential liar to wear a lie detector
  - can carry out the analysis offline

NLP and Deception Detection

Among the methods relying on the analysis of verbal clues, Natural Language Processing techniques have been reasonably successful in a variety of experimental conditions, such as dealing with:

- Samples of spoken and written language collected in laboratory conditions Newman *et al.* (2003); Strapparava and Mihalcea (2009);
- Computer-Mediated-Communication Hancock *et al.* (2008); Zhou *et al.* (2004); Zhou (2005);
- Samples of spoken and written language collected on the field in judicial context Bachenko *et al.* (2008); Fornaciari and Poesio (2011a,b).

NLP

- Modern NLP is based on the use of Machine Learning Techniques to create CLASSIFIERS capable of assigning labels to (parts of text) or documents. Examples include
  - Spam Detectors that classify email messages into SPAM / NON SPAM
  - Sentiment analyzers that classify (parts of) text into positive / negative
- In the case of deception detection, Stylometric methods have been used to classify text in DECEPTIVE / NON DECEPTIVE

Using machine learning in NLP

- In traditional Artificial Intelligence, systems for, e.g., analyzing natural language or images were developed by writing algorithms by hand
- Around the mid 1980s the realization came that this approach, apart from being very different from the way humans learn how to do things (which need not be a problem as AI chess-playing systems are much better than humans), was unlikely to achieve good results as no human or team of humans can ever hope to think of all the possibilities
- So the focus of AI switched to developing algorithms that could learn how to carry out such tasks from datasets of examples
- Such systems typically extract features from the object they have to classify (e.g., a review) and use them to decide on a category
- A particularly successful approach to choosing such features has been the stylometric approach in which only surface features are used

Stylometry

In NLP, stylometry studies texts on the basis of its stylistic features only. As Koppel *et al.* (2006) point out, the features used in stylometric analyses belong to two main families:

Surface features. This type of features includes the frequency and use of function words or of certain *n*-grams of words or part-of-speech (POS tag), without taking into consideration their meaning.

Lexical features. These features attempt to capture the meaning of texts. Such information may come from:

- **Lexicons.** Lexicons associate each word to a variety of categories of different kinds: grammatical, lexical, psychological and so on. This results in a profile of texts with respect to those categories.
- **Linguistic analyses.** More complex analyses such as syntactic analyses, extraction of argument structure or coreference are also possible. Some can be carried out automatically, others are usually done by hand (Bachenko *et al.*, 2008).

Two applications of deception detection

Today I will discuss two examples from our own work of the use of stylometric techniques for deception detection:

- Identifying deceptive statements in court (Fornaciari and Poesio, 2013)
- Classifying online reviews (Fornaciari and Poesio, 2014; Fornaciari *et al.*, 2018)

The key problem: Lack of real data

In our view, the foremost problem with current research is the lack of real datasets. Most work relies on artificial datasets created in the lab, such as

- The fake points of view on various topics taken by the subjects in (Newman *et al.*, 2003)
- The fake reviews produced for the studies in (Strapparava and Mihalcea, 2009) and in the most widely used dataset for work on reviews, produced by (Ott *et al.*, 2011)

Outline

High stakes corpus

DeCour - DEception in COURt - is a corpus constituted by the transcripts of 35 hearings in front of the judge.

They come from criminal proceedings for calumny and false testimony, where the defendants were found guilty.

The proceedings end with a judgment which summarises the facts, pointing out the lies told by the speaker.

The hearings took place in 4 Italian Courts: Bologna, Bolzano, Prato and Trento.

Subjects

The testimonies of 31 subjects were collected, who played the role of:

- Witness in 19 hearings;
- Defendant in 14 hearings;
- Expert witness in 1 hearing;
- Victim in 1 hearing.

Their mean age was 36 and they were all fluent Italian speakers.

Sex	
Men	23
Women	7
Transgenders	1

Origin	
Italy, North	12
Italy, Center	2
Italy, South	9
Abroad	8

The education of 6 subjects was known, ranging from elementary to high school.

Annotation

Only the 3015 utterances of the heard subjects were taken into consideration.

They were annotated as:

- 1202 annotated as true, as coherent with the reconstruction of the facts contained in the judgment;
- 945 annotated as false, as pointed out in the judgment as false;
- 868 annotated as uncertain: their truthfulness was not known or not logically determinable (as in case of questions).

Mark up format

The hard-copies of the hearings were subjected to Optical Character Recognition - OCR and stored as text files.

After manual editing, aimed to emend the unavoidable errors of the OCR, the *corpus* was structured in XML format.


```xml
<hearing>
  <header birtharea="N" birthplace="BZ" birthyear="xxxxx" court="BZ" day="xxxxx"
      idsub="xxxxx" month="xxxxx" name="xxxxx" nrdoss="xxxxx" nrhear="xxxxx" sex="M"
      study="unk" typesub="defwit" typetest="false" year="2003" yeardoss="03"/>
  <intro> GIUDICE MONOCRATICO – DOTT. xxxxx xxxxx Viene introdotto il testimone ; questi
      viene avvertito dal Giudice dei suoi obblighi e rende la dichiarazione ex Art. 497
      C.P.P. . Fornisce le generalità : xxxxx xxxxx , nato a xxxxx il xx xxxxx xxxx , ivi
      residente . </intro>
  <turn nrgen="1" nrpros="1" speaker="pros">
    <utterance class="x" nrgen="1" nrpros="1">
      Lei nella primavera del 2001 ci può dire come ha conosciuto Mizar Roberto , in quali circostanze ?
    </utterance>
  </turn>
  <turn nrgen="2" nrsub="1" speaker="defwit">
    <utterance class="uncertain" nrgen="2" nrsub="1">
      Adesso non mi ricordo come l' ho conosciuto , comunque ci siamo conosciuti ...
    </utterance>
    <utterance class="uncertain" nrgen="3" nrsub="2">
      Non mi ricordo , in giro , al Cici anche , perché prendevo il Metadone tempo fa .
    </utterance>
    <utterance class="uncertain" nrgen="4" nrsub="3">
      Adesso sono due anni che sono a posto , quasi due anni .
    </utterance>
  </turn>
```


Sensitive data were anonymised, as agreed with the Courts.

For these purposes, the text files were manipulated using Perl (Wall *et al.*, 2004).

Lexical features: LIWC

The Linguistic Inquiry and Word Count - LIWC is perhaps the best-known lexical resource for deception detection, developed by Pennebaker *et al.* (2001).

In particular, it is a validated lexicon, whose English dictionary is constituted of around 4500 words (or roots of words), whereby each term is associated with an appropriate set of syntactical, semantical and/or psychological dimension, such as emotional words, cognitive words, self references, different kind of pronouns, and so on.

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 dimensions grows. The output is a profile of the text which relies on the rate of incidence of the different dimensions in the text itself.

LIWC also includes different dictionaries for several languages, amongst which Italian (Alparone *et al.*, 2004).

LIWC dimensions

The most representative LIWC dimensions, employed in the experiments of Newman *et al.* (2003):

Standard linguistic dimensions	Psychological processes	Relativity
Word Count	Affective or emotional processes	Space
% words captured by the dictionary	Positive emotions	Inclusive
% words longer than six letters	Negative emotions	Exclusive
Total pronouns	Cognitive processes	Motion verbs
First-person singular	Causation	Time
Total first person	Insight	Past tense verb
Total third person	Discrepancy	Present tense verb
Negations	Tentative	Future tense verb
Articles	Certainty	
Prepositions	Sensory and perceptual processes	
	Social processes	

Surface Features

As surface features, in our experiments we considered:

- Utterances' length with punctuation;
- Utterances' length without punctuation;
- 7 kind of *n*-grams considered, from unigrams to eptagrams, of:
  - Lemmas;
  - Part-Of-Speech - POS.

Lexical features

In the experiments where LIWC features are employed, there were included:

- The rate of words found in the text which are also present in the LIWC dictionary;
- The number of words longer than six letters.
- 82 out of the 85 lexical categories of the LIWC Italian dictionary (the three remaining - 'They', 'Passive' and 'Formal' - were empty in our *corpus*.)

The mean number of words per sentence is omitted as meaningless for our analysis units.

Feature selection

- The most informative lexical / n-gram features were chosen using a method called Information Gain - IG.
- Only chosen from utterances classified as True or False

Training models

We trained models in order to classify the utterances of DeCour, according to the classes they belong to.

We tested a variety of classification methods, finding that the best performance was obtained with Support Vector Machines (SVMs) Cortes and Vapnik (1995).

Our SVM models were trained and then tested via *n*-fold cross-validations.

- In all the experimental conditions, each hearing of DeCour constitutes a fold for the cross-validations, so that the experiments run on the whole corpus have been carried out with a 35-fold cross-validation.
- In other experiments, some hearings were discarded and thence the *n*-fold cross-validation corresponded to the number of the employed hearings.

Experimental designs

Thirteen experiments were carried out, divided in three groups.

- The first group of 5 experiments were concerned with replicating the methodology of Newman *et al.* [2003] in a high-stakes deception scenario and comparing the performance of the lexical features used in that work with that of surface features;
- The goal of the second group of 5 experiments was to compare the performance of the classifier on the entire corpus with the performance on the subset of utterances classified as true or false only, that is discarding the uncertain utterances, which in the previous group of experiments were grouped together with the true ones into the generic class of not-false utterances;
- In the last group of 3 experiments we focused on more cohesive sets of subjects:
    - only male speakers: 25 hearings;
    - only Italian native speakers: 26 hearings;
    - only over 30 years old speakers: 21 hearings.

Whole DeCour

Classes: False vs. True and Uncertain utterances.

	Accuracy		False utterances		
	Mean	Total	Precision	Recall	F-measure
LIWC	68.28%	69.35%	51.57%	36.40%	42.68%
BF	68.29%	69.95%	53.42%	32.28%	40.24%
IG	69.89%	70.18%	53.11%	41.59%	46.65%
LIWC+BF	68.96%	70.55%	54.77%	34.60%	42.41%
LIWC+IG	68.59%	69.88%	52.54%	40.42%	45.69%

Baseline	Accuracy	Precision	Recall
Random	60.03%	37.03%	35.97%
Majority	68.66%	NaN	0%
Algorithmic	62.39%	40.06%	41.80%

True vs. False utterances

Classes: False vs. True utterances. The Uncertain ones are removed.

	Accuracy		False Utterances		
	Mean	Total	Precision	Recall	F-measure
LIWC	66.48%	68.23%	65.56%	58.62%	61.90%
BF	68.62%	69.86%	69.05%	57.14%	62.53%
IG	68.25%	69.54%	68.77%	56.40%	61.97%
LIWC+BF	69.84%	70.61%	70.60%	56.93%	63.03%
LIWC+IG	68.90%	70.24%	71.31%	54.18%	61.58%

Baseline	Accuracy	Precision	Recall
Random	54.54%	49.95%	48.36%
Majority	55.98%	NaN	0%
Algorithmic	59.57%	54.38%	52.80%

Discussion

In every experimental condition:

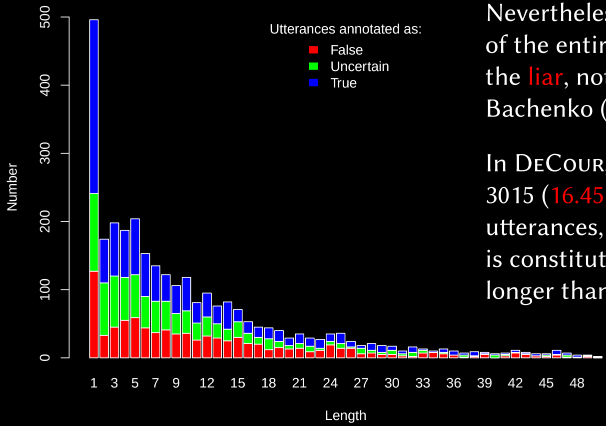
- the models can identify deceptive statements with an accuracy around 70%, which is well above chance and much better than the simple heuristic algorithm;
- The precision is considerably higher than the baselines;
- In "whole DeCour", "male speakers" and "over 30 speakers" conditions the recall is lower instead.

Therefore:

- This suggests that the type of methods proposed by Pennebaker *et al.* (2001) can be applied with a certain degree of success to identify deception even with real-life data collected in high-stakes situations.
- The results of the experiments relying on more homogeneous subsets of subjects do not show remarkable improvement in the effectiveness of the models, also because if in one hand the accuracy rises slightly, the baselines too are shifted upwards.

Deception at utterance level

The task of classifying single utterances is much more challenging than the one attempted by, e.g., Pennebaker *et al.* (2001), who classified full texts.

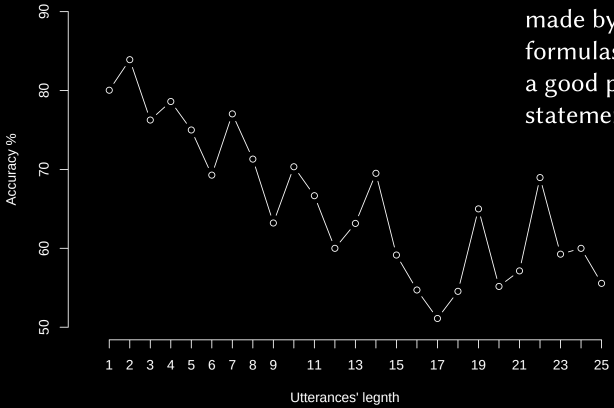


Nevertheless, working at the level of the entire narrative identifies the liar, not the lie Fitzpatrick and Bachenko (2012).

In DeCour, 496 utterances out of 3015 (16.45%) are single-word utterances, and 70.44% of DeCour is constituted by utterances no longer than 15 words.

Accuracy and utterance length

There seems to be a correlation between length of the utterance and classification accuracy: the longer the utterances, the lower the accuracy.

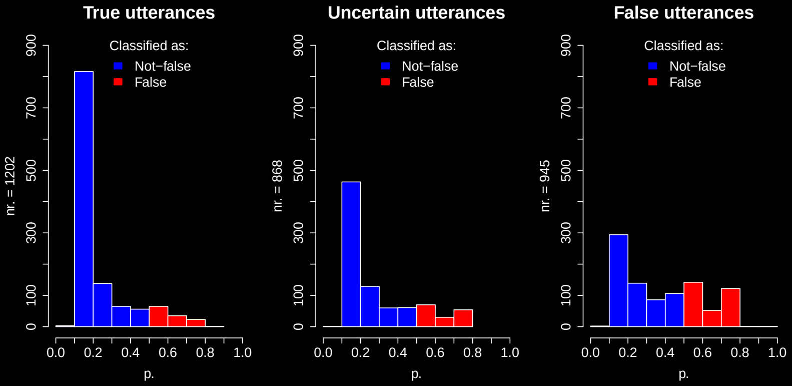


Since short statements are typically made by stereotypical linguistic formulas, formulaic language could be a good predictor in order to classify statements as true or false.

Uncertainty and noise

When uncertain utterances are removed, the gap between classification accuracy and heuristic baseline grows from about 6 to about 9 percent points.

The probabilities assigned by classifier to the utterances of belonging to the classes suggests that the uncertain ones are a mix of true and false statements.



English deceptive linguistic style

Newman *et al.* (2003), evaluating lab-produced samples of (spoken and written) deceptive English language through the LIWC categories, found that this is characterized by:

- Fewer first-person singular pronouns;
- Fewer third-person pronouns;
- Fewer exclusive words;
- More negative emotion words;
- More motion verbs.

These findings were confirmed by most subsequent research on English.

Italian deceptive linguistic style

Our assumptions about the prevalence of positive statements among true utterances and of negative statements among false ones are confirmed.

Confirming the results of Newman *et al.*, false utterances have higher values for the dimensions of:

- Negative Emotions;
- Exclusive words;
- Discrepancy.

False utterances have higher values for content expressing cognitive/perceptual processes, while true utterances have greater values for references to time, space, concrete topics and positive feelings.

LIWC categories most prevalent in True utterances

LIWC dimensions	False Utterances' mean values	True Utterances' mean values	Difference
Certainty	0.0973	0.2681	-0.1708
Prepositions	0.1472	0.1691	-0.0219
Space	0.0256	0.0348	-0.0093
Time	0.0603	0.0669	-0.0066
Home	0.0028	0.0086	-0.0058
Positive feelings	0.0160	0.0217	-0.0057
Leisure	0.0047	0.0094	-0.0047
Numbers	0.0067	0.0102	-0.0036
Nonfluencies	0.0015	0.0047	-0.0033
Optimism and energy	0.0066	0.0096	-0.0030
Occupation	0.0068	0.0093	-0.0024
We	0.0072	0.0096	-0.0024
Work	0.0026	0.0048	-0.0022
Past tense verb	0.0904	0.0920	-0.0017
They verb	0.0196	0.0209	-0.0014
Money	0.0034	0.0046	-0.0012
Eating, drinking, dieting	0.0021	0.0032	-0.0011
School	0.0002	0.0012	-0.0010
Friends	0.0029	0.0038	-0.0009
Inhibition	0.0040	0.0047	-0.0007

LIWC categories most prevalent in False utterances:

LIWC dimensions	False Utterances' mean values	True Utterances' mean values	Difference
Negations	0.2682	0.0742	0.1940
Cognitive processes	0.1794	0.0997	0.0797
Present	0.2146	0.1454	0.0692
I verb	0.1580	0.0957	0.0623
Total pronouns	0.1885	0.1473	0.0412
Transitive	0.0527	0.0192	0.0335
I	0.1099	0.0794	0.0305
Introspection	0.0584	0.0353	0.0231
To have	0.0561	0.0336	0.0225
Perceptual processes	0.0537	0.0316	0.0221
If	0.0642	0.0485	0.0157
Discrepancy	0.0309	0.0162	0.0147
Past participle	0.0764	0.0622	0.0142
Causation	0.0382	0.0270	0.0112
Communication	0.0452	0.0354	0.0098
Exclusive	0.1044	0.0946	0.0098
Negative emotion	0.0209	0.0112	0.0097
Articles	0.1735	0.1642	0.0093
Hearing	0.0304	0.0214	0.0091
Seeing	0.0148	0.0067	0.0082

Pronouns and verbs in false utterances

- Even though in Italian the pronouns can be omitted, the recurrent finding that liar use less pronouns and less self-references is not confirmed in DeCour.
  - This outcome is related to the large use of expressions concerning cognitive processes and speculations;

	False Utterances	True Utterances
First person pronouns/number of utterances	0.4158	0.2138
First person pronouns/number of tokens	0.0246	0.0166
Pronoun "Io"/First person verbs	0.2753	0.2526
First person pronouns/First person verbs	0.3718	0.3399
First person verbs/number of utterances	1.1185	0.6290
First person verbs/number of tokens	0.0664	0.0489

	True utterances	False utterances
non mi ricordo	20	49
non ricordo	6	68

The χ^2 test gives a $p = 0.0025$ for this contingency table.

Frequent *n*-grams in DeCour

Table: N-grams Frequency in DeCour

True utterances					
Tokens	Freq.	Bigrams	Freq.	Trigrams	Freq.
sì	431	xxxxx xxxxx	66	non mi ricordo	20
che	389	c’era	53	c’era un	13
xxxxx	327	mi hanno	40	che c’era	12
e	284	mi ricordo	36	mi ha detto	10
di	268	l’ho	32	mi ricordo che	9
False utterances					
Tokens	Freq.	Bigrams	Freq.	Trigrams	Freq.
non	644	l’ho	85	non mi ricordo	49
che	394	non mi	84	non lo so	38
ho	317	mi ricordo	69	non l’ho	28
e	302	non ricordo	68	non è che	17
mi	302	io non	61	io l’ho	16

Next steps

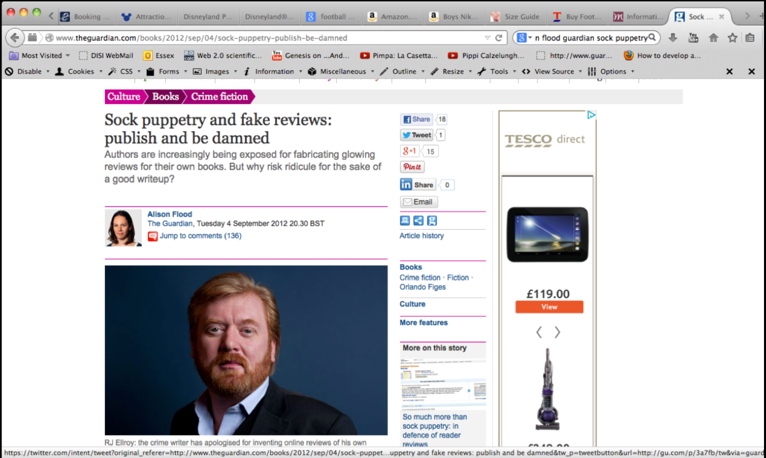
- To improve the feature selection, taking into consideration:
  - Expressions of doubt: edges;
  - Syntactical structure related features: parsing;
  - Dialogic elements: Linguistic Style Matching (Niederhoffer and Pennebaker, 2002).
- To open to multimodal analyses.

Outline

A different type of deception: ‘Sock puppetry’

As the Kubicki’s fifth addition of the Colton Banyon series, ‘A Dubious Plan’ is by far the most daring of all the series. It’s amazing how Kubicki incorporates history into a mix of mystery and sensuality. There are a significant amount of mysteries surrounding the finding of an old war plane in Death Valley, which is centered around the World War II era. Although the story begins with such a romantic spin, it transitions into action and suspense with the unraveling of a journey of survival. Do yourself a favor and make time to please yourself by reading this book.

Sock puppetry in the media



Deceptive online reviews

- With the increasing reliance on online reviews, comes an increasing opportunity for unscrupolous book sellers / book writers / hotel managers to attract customers via fake reviews
- This has become an endemic problem
- In NLP, lots of work on detecting deceptive reviews
  - On e-commerce sites such as Amazon
  - On hotel recommendation sites such as Trip Advisor

Detecting deceptive Amazon reviews

- We applied the methods discussed in the previous experiment to detect fake Amazon reviews (Fornaciari and Poesio, 2014)
- Specifically
  1. We created a corpus of fake Amazon reviews called DEREV and consisting of
     - A number of reviews we knew to be fake because their authors confessed
     - A number of reviews we had good reason to believe were authentic because they were about classic books so famous that there was no need to write fake reviews
  2. We applied stylometric methods to classify those reviews
  3. We achieve around 72% accuracy

The DEREV corpus: The Idea

- On September 4th, 2012, Alison Flood published an article in *The Guardian* about the crime writer Jeremy Duns, who had unmarked a number of 'sock puppeteers' among his colleagues –authors writing and/or paying for glowing reviews of their own books. We contacted him and he was extremely helpful, giving us several hints to recognize possible cues of deception in the reviews.
- Afterwards we discovered a number of several other articles, in particular one from July 25th, 2011, on www.moneytalksnews.com. In the article, entitled 3 Tips for Spotting Fake Product Reviews - From Someone Who Wrote Them, Sandra Parker, shared her experience as professional review writer.

The DEREV corpus (2014, revised 2018)

The first release of the DEREV corpus consists of Amazon reviews of 68 books, of which

- 46 SUSPECT BOOKS
  - The 22 books for which Sandra Parker admitted writing a review
  - 4 books mentioned in another article by Streitfeld
  - 20 books reviewed by the same reviewers that had reviewed the 4 books mentioned by Streitfeld
- 22 INNOCENT BOOKS
  - Books written by classic authors, such as Arthur Conan Doyle or Rudyard Kipling
  - or by living writers who are so renowned that purchasing reviews would be pointless: e.g., Ken Follett and Stephen King.

We subsequently eliminated a number of duplicated reviews and ended up with 6759 reviews written by 4811 different reviewers, for a total of about 1 million tokens.

Gold and Silver Standard

- We have a reasonably plausible labelling for 1552 of the reviews in DEREV. We consider these our GOLD STANDARD
    - The 776 reviews written by the authors who admitted to producing fake reviews can be plausibly considered as fake
    - To these we added 776 randomly selected reviews out of the Innocent Books that can be plausibly considered as genuine
- But what about the other reviews?

Deception Cues

Jeremy Duns and Sandra Parker suggest a number of cues that can be used to recognize deceptive reviews and can be automatically extracted from Amazon:

Cluster - Cl
: Sandra Parker pointed out that agencies which provide review services gave her 48 hours to write a review. Being likely that the same deadline was given to other reviewers, Sandra Parker warned to pay attention if the books received many reviews in a short lapse of time. Following her advice, we considered as positive this clue if the review belonged to a group of at least two reviews posted within 3 days.

Nickname - NN
: Reviewers on Amazon can register and post comments using their real name. Since the real identity of the reviewers involves issues related to their reputation, it is less likely that the writers of fake reviews post their texts using their true name.

Unknown Purchase - UP
: One of the most interesting information provided by Amazon is whether the reviewer bought the reviewed book through Amazon itself. It is reasonable to think that, if this happened, the reviewer also read the book. Therefore, the absence of information about the certified purchase was considered a clue of deceptiveness.

A Silver Standard Using Aggregation Methods

- The deception cues just mentioned could be considered as VOTES for the review
- So that we could then use one of the AGGREGATION METHODS used in the literature on crowdsourcing to come up with a plausible labelling for the other reviews
- The aggregation methods we considered include:
  - MAJORITY VOTING as a baseline
  - The GLAD Bayesian aggregation method proposed by Whitehill *et al.* (2009)
  - The LEARNING FROM CROWDS Bayesian aggregation methods proposed by Raykar *et al.* (2010)

Comparing Aggregation Methods

Algorithm	First iteration	Rate of false reviews	Correspondance with the gold standard
MV	None	67.41%	52.58%
LFC	Majority Voting	76.15%	52.19%
LFC	Random classes	30.08%	**69.01%**
GLAD	Random classes	90.06%	45.10%

Using the DEREV silver standard to train a deceptive reviews detector

Experimental design

Training set	DEREV with LFC classes
Test set	gold standard
Features	147 linguistic, 3 behavioral

Confusion matrix

	False reviews	True reviews
Predicted false	446	102
Predicted true	330	674

Performance

	Accuracy	Precision	Recall	F-measure
Model	**72.16%**	**81.39%**	**57.47%**	**67.37%**
LFC baseline	69.01%	77.37%	53.74%	63.43%

Conclusions

- Deception detection a very interesting application for NLP - interesting uses both in forensics and in e-commerce
- Creating suitable datasets a big challenge
- Bayesian annotation methods potentially useful

Thanks!

Bibliography I

Alparone, F., Caso, S., Agosti, A., and Rellini, A. (2004). The Italian LIWC2001 Dictionary. Austin, TX: LIWC.net.

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.

Bond, C. F. and De Paulo, B. M. (2006). Accuracy of Deception Judgments. *Personality and Social Psychology Review*, **10**(3), 214–234.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**.

Fitzpatrick, E. and Bachenko, J. (2012). Building a data collection for deception research. In E. Fitzpatrick, J. Bachenko, and T. Fornaciari, editors, *Proc. of the EACL Workshop on Computational Approaches to Deception Detection*, pages 31–38.

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. (2013). Detecting deception in italian court testimonies. *Artificial Intelligence and Law*, **21**(3), 303–340.

Bibliography II

Fornaciari, T. and Poesio, M. (2014). Identifying fake amazon reviews as learning from crowds. In *Proc. of EACL*, Gothenburg.

Fornaciari, T., Cagnina, L., Rosso, P., and Poesio, M. (2018). Probabilistic annotation vs. crowd-sourced texts production for opinion spam detection. To be submitted.

Hancock, J. T., Curry, L. E., Goorha, S., and Woodworth, M. (2008). On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication. *Discourse Processes*, **45**(1), 1–23.

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*.

Levine, T. R., Feeley, T. H., McCornack, S. A., Hughes, M., and Harms, C. M. (2005). Testing the Effects of Nonverbal Behavior Training on Accuracy in Deception Detection with the Inclusion of a Bogus Training Control Group. *Western Journal of Communication*, **69**(3), 203–217.

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.

Niederhoffer, K. G. and Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, **21**(4), 337–360.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th ACL*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

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

Bibliography III

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, **11**, 1297–1322.

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*.

Vrij, A. (2008). *Detecting Lies and Deceit: Pitfalls and Opportunities*. Wiley Series in Psychology of Crime, Policing and Law. John Wiley & Sons, 2nd edition.

Wall, L., Christiansen, T., and Orwant, J. (2004). *Programming perl*. " O'Reilly Media, Inc.".

Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J. (2009). Whose vote should count more: optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, volume 22, pages 2035–2043.

Zhou, L. (2005). An Empirical Investigation of Deception Behavior in Instant Messaging. *IEEE Transactions on Professional Communication*, **48**(2), 147–160.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., and Twitchell, D. (2004). Automating Linguistics-Based Cues for Detecting Deception in Text-based Asynchronous Computer-Mediated Communication. *Group Decision and Negotiation*, **13**(1), 81–106.