Identifying fake Amazon reviews as learning from crowds

Tommaso Fornaciari
Ministero dell'Interno
Dipartimento della Pubblica Sicurezza
Segreteria del Dipartimento
ComISSIT
tommaso.fornaciari@interno.it

Massimo Poesio
University of Essex
CSEE
University of Trento
CIMeC
poesio@essex.ac.uk

Abstract

Customers who buy products such as books online often rely on other customers reviews more than on reviews found on specialist magazines. Unfortunately the confidence in such reviews is often misplaced due to the explosion of so-called **sock puppetry**–authors writing glowing reviews of their own books. Identifying such deceptive reviews is not easy. The first contribution of our work is the creation of a collection including a number of genuinely deceptive Amazon book reviews in collaboration with crime writer Jeremy Duns, who has devoted a great deal of effort in unmasking sock puppeting among his colleagues. But there can be no certainty concerning the other reviews in the collection: all we have is a number of cues, also developed in collaboration with Duns, suggesting that a review may be genuine or deceptive. Thus this corpus is an example of a collection where it is not possible to acquire the actual label for all instances, and where clues of deception were treated as annotators who assign them heuristic labels. A number of approaches have been proposed for such cases; we adopt here the 'learning from crowds' approach proposed by Raykar et al. (2010). Thanks to Duns' certainly fake reviews, the second contribution of this work consists in the evaluation of the effectiveness of different methods of annotation, according to the performance of models trained to detect deceptive reviews.

1 Introduction

Customer reviews of books, hotels and other products are widely perceived as an important reason for the success of e-commerce sites such as amazon.com or tripadvisor.com. However, customer confidence in such reviews is often misplaced, due to the growth of the so-called **sock puppetry** phenomenon: authors / hoteliers writing glowing reviews of their own works / hotels (and occasionally also negative reviews of the competitors).[1] The prevalence of this phenomenon has been revealed by campaigners such as crime writer Jeremy Duns, who exposed a number of fellow authors involved in such practices.[2] A number of sites have also emerged offering Amazon reviews to authors for a fee.[3]

Several automatic techniques for exposing such deceptive reviews have been proposed in recent years (Feng et al., 2012; Ott et al., 2001). But like all work on deceptive language (computational or otherwise) (Newman et al., 2003; Strapparava and Mihalcea, 2009), such works suffer from a serious problem: the lack of a gold standard containing 'real life' examples of deceptive uses of language. This is because it is very difficult to find definite proof that an Amazon review is either deceptive or genuine. Thus most researchers recreate deceptive behavior in the lab, as done by Newman et al. (2003). For instance, Ott et al. (2001), Feng et al. (2012) and Strapparava and Mihalcea (2009) used crowdsourcing, asking turkers to produce instances of deceptive behavior. Finally, Li et al. (2011) classify reviews as deceptive or truthful by hand on the basis of a series of heuristics: they start by excluding anonymous reviews, then use their helpfulness and other criteria to decide

[1] The phenomenon predates Internet - see e.g., Amy Harmon, 'Amazon Glitch Unmasks War Of Reviewers', New York Times, February 14, 2004.

[2] See Andrew Hough, 'RJ Ellory: fake book reviews are rife on internet, authors warn', telegraph.co.uk, September 3, 2012

[3] See Alison Flood, 'Sock puppetry and fake reviews: publish and be damned', guardian.co.uk, September 4, 2012 and David Streitfeld, 'Buy Reviews on Yelp, Get Black Mark', nytimes.com, October 18, 2012.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 279–287,
Gothenburg, Sweden, April 26-30 2014.

whether they are deceptive or not. Clearly a more rigorous approach to establishing the truth or otherwise of reviews on the basis of such heuristic criteria would be useful.

In this work we develop a system for identifying deceptive reviews in Amazon. Our proposal makes two main contributions:

1. we identified in collaboration with Jeremy Duns a series of criteria used by Duns and other 'sock puppet hunters' to find suspicious reviews / reviewers, and collected a dataset of reviews some of which are certainly false as the authors admitted so, whereas others may be genuine or deceptive.

2. we developed an approach to the truthfulness of reviews based on the notion that the truthfulness of a review is a latent variable whose value cannot be known, but can be estimated using some criteria as potential indicators of such value–as *annotators*–and then we used the **learning from crowds** algorithm proposed by Raykar et al. (2010) to assign a class to each review in the dataset.

The structure of the paper is as follows. In Section 2 we describe how we collected our dataset; in Section 3 we show the experiments we carried out and in Section 4 we discuss the results.

2 Deception clues and dataset

2.1 Examples of Unmasked Sock Puppetry

After reading an article by Alison Flood on *The Guardian* of September 4th, 2012 [4], discussing how crime writer Jeremy Duns had unmasked a number of 'sock puppeteers,' we contacted him. Duns was extremely helpful; he pointed us to the other articles on the topic, mostly on *The New York Times*, and helped us create a set of **deception clues** and the dataset used in this work.

On July 25th, 2011, an article appeared on www.moneytalksnews.com, entitled '3 Tips for Spotting Fake Product Reviews - From Someone Who Wrote Them'.[5] Sandra Parker, author of the text, in that page described her experience as 'professional review writer'. According to her statements, advertising agencies were used to pay her $10-20 for writing reviews on sites like Amazon.com. She was not asked to lie, but 'if the review wasn't five star, they didn't pay'. In an article of August 19th, written by David Streitfeld on www.nytimes.com,[6] she actually denied that point: 'We were not asked to provide a five-star review, but would be asked to turn down an assignment if we could not give one'.

In any case, in her article Sandra Parker gave the readers some common sense-based advices, in order to help them to recognize possible fake reviews. One of these suggestions were also useful for this study, as discussed in Section 2.3. From our point of view, however, the most interesting aspect of the article relied in the fact that, letting know the name of an author of fake reviews, it made possible to identify them in Amazon.com, with an high degree of confidence.

A further article written on August 25th by David Streitfeld gave us another similar opportunity.[7] In fact, thanks to his survey, it was possible to come to know the titles of four books, whose the authors paid an agency in order to receive reviews.

2.2 The corpus

Using the suggestions of Jeremy Duns and the information in these articles we built a corpus we called DEREV (DEception in REViews), consisting of clearly fake, possibly fake, and possibly genuine book reviews posted on www.amazon.com. The corpus, which will be freely available on demand, consists of 6819 reviews downloaded from www.amazon.com, concerning 68 books and written by 4811 different reviewers. The 68 books were chosen trying to balance the number of reviews (our units of analysis) related to suspect books which probably or surely received fake reviews, with the number of reviews hypothesized to be genuine in that we expected the authors of the books not to have bought reviews. In particular, we put into the group of the suspect books - henceforth SB - the reviews of the four books indicated by David Streitfeld. To this first nucleus, we also added other four books, written by three of the authors of the previous group. We also in-

[4] *Sock puppetry and fake reviews: publish and be damned*, http://www.guardian.co.uk/books/2012/sep/04/sock-puppetry-publish-be-damned

[5] http://www.moneytalksnews.com/2011/07/25/3-tips-for-spotting-fake-product-reviews---from-someone-who-wrote-them/

[6] http://www.nytimes.com/2011/08/20/technology/finding-fake-reviews-online.html?_r=1&

[7] http://www.nytimes.com/2012/08/26/business/book-reviewers-for-hire-meet-a-demand-for-online-raves.html?pagewanted=all

cluded in the SB group the 22 books for which Sandra Parker wrote a review. Lastly, we noticed that some reviewers of the books pointed out by David Streitfeld tended to write reviews of the same books: we identified 16 of them, and considered suspect as well. In total, on November 17th, 2011 we downloaded the reviews of 46 books considered as suspect, which received 2707 reviews.[8] We also collected the reviews of 22 so called 'innocent books', for a total of 4112 reviews. These books were mainly chosen among classic authors, such as Conan Doyle or Kipling, or among living writers who are so renowned that any reviews' purchase would be pointless: this is the case, for example, of Ken Follett and Stephen King. As shown by the number of the reviews, the books of these authors are so famous that they receive a great amount of readers' opinions.

The size of DEREV is 1175410 tokens, considering punctuation blocks as single token. The mean size of the reviews is 172.37 tokens. The titles of the reviews were neither included in these statistics nor in the following analyses.

2.3 Deception clues

Once created the corpus, we identified a set of clues, whose presence suggested the deceptiveness of the reviews. These clues are:

Suspect Book - SB The first clue of deceptiveness was the reference of the reviews to a suspect book, identified as described above. This is the only clue which is constant for all the reviews of the same book.

Cluster - Cl The second clue comes from the suggestions given by Sandra Parker in her mentioned article. As she pointed out, the agencies she worked for were used to give her 48 hours to write a review. Being likely that the same deadline was given to other reviewers, Sandra Parker warns to pay attention if the books receive many reviews in a short period of time. Following her advice, we considered as positive this clue of deceptiveness if the review belonged to a group of at least two reviews posted within 3 days.

Nickname - NN A service provided by Amazon is the possibility for the reviewers to register in the website and to post comments using their real name. Since the real identity of the reviewers involves issues related to their reputation, we supposed it is less probable that the writers of fake reviews post their texts using their true name. Moreover, a similar assumption was probably accepted by Li et al. (2011), who considered the profile features of the reviewers, and among them the use or not of their real name.

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

2.4 Gold and silver standard

The clues of deception discussed above give us a heuristic estimate of the truthfulness of the reviews. Such estimation represents a silver standard of our classes, as these are not determined through certain knowledge of the ground truth, but simply thanks to hints of deceptiveness. The methods we used in order to assign the heuristic classes to the reviews are described in the next Section; however for our purposes we needed a gold standard, that is at least a subset of reviews whose ground truth was known with a high degree of confidence. This subset was identified as follows.

First, we considered as false the 22 reviews published by Sandra Parker, even though not all her reviews are characterized by the presence of all the deception clues. Even though we cannot really say whether her reviews reflect her opinion of the books in question or not, she explicitly claimed to have been paid for writing them; and she only bought on Amazon three of these 22 books. This is the most accurate knowledge about fake reviews not artificially produced we have found in literature. Then we focused on the four books whose authors admitted to have bought the reviews.[9] Three of them received many reviews, which made it difficult to understand if they were truthful or not. However, one of these

[8] We specify the date of the download because, obviously, if the data collection would be repeated today, the overall number of reviews would be greater.

[9] http://www.nytimes.com/2012/08/26/business/book-reviewers-for-hire-meet-a-demand-for-online-raves.html?pagewanted=all

Table 1: The distribution of deception clues in the reviews

	Nr. clues	Reviews	Tot.	%
False	4	903		
rev.	3	1913	2816	41.30%
True	2	2528		
rev.	1	1210		
	0	265	4003	58.70%

books ('Write your first book', by Peter Biadasz) received only 20 reviews, which therefore could be considered as fake with high degree of probability. Even though we have no clear evidence that a small number of reviews correlates with a greater likelihood of deception, since we know this book received fake reviews, and there are only few reviews for it, we felt it is pretty likely that those are fake. Therefore we examined the reviews written by these twenty authors, and considered as false only those showing the presence of all the deception clues described above. In this way, we found 96 reviews published by 14 reviewers, and we added them to the 22 of Sandra Parker, for a total of 118 reviews written by 15 authors.

Once identified this subset of fake reviews, we selected other 118 reviews which did not show the presence of any deception clue, that is chosen from books above any suspicion, written by authors who published the review having made use of their real name and having bought the book through Amazon and so on.

In the end, we identified a subset of DEREV constituted by 236 reviews, whose class was known with high degree of confidence and considered them as our gold standard.

3 Experiments

We carried out two experiments, in which the classes assigned to the reviews of DEREV were found adopting two different strategies. In the first experiment the classes of the reviews were determined using majority voting of our deception clues. This experiment is thus conceptually similar to those of Li et al. (2011), who trained models using supervised methods with the aim of identifying fake reviews. We discuss this experiment in the next Section. In the second experiment, learning from crowds was used (Raykar et al., 2010). This approach is discussed in Section 3.2.1.

In both experiments we carried out a 10-fold cross-validation where in each iteration feature selection and training were carried out using 90% of the part of the corpus with only silver standard annotation and 90% of the subset with gold. The test set used in each iteration consisted of the remaining tenth of reviews with gold standard classes, which were employed in order to evaluate the predictions of the models. This allowed to estimate the efficiency of the strategies we used to determine our silver standard classes.

3.1 Majority Voting

3.1.1 Determining the class of reviews by majority voting

The deception clues discussed in Section 2.3 were used in our first experiment to identify the class of each review using majority voting. In other words, those clues were considered as independent predictors of the class; the class predicted by the majority of the annotators/clues was assigned to the review. Specifically, if 0, 1 or 2 deception clues were found, the review was classified as true; if there were 3 or 4, the review was considered false. Table 1 shows the distribution of the number of deception clues in the reviews in DEREV.

3.1.2 Feature selection

In both experiments each review was represented as feature vector. The features were just of unigrams, bigrams and trigrams of lemmas and part-of-speech (POS), as collected from the reviews through TreeTagger[10] (Schmid, 1994).

Since in each experiment we applied a 10-fold cross-validation, in every fold the features were extracted from the nine-tenths of DEREV employed as training set. Once identified the training set, we computed the frequency lists of the n-grams of lemmas and POS. The lists were collected separately from the reviews belonging to the class 'true' and to the class 'false'. Such separation was aimed to take into consideration the most highly frequent n-grams of both genuine and fake reviews. However, for the following steps of the feature selection, only the n-grams which appeared more than 300 times in every frequency list were considered: a threshold empirically chosen for ease of calculations. In fact, among the most

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

Table 2: The most frequent n-grams collected

N-grams	Lemmas	POS	Total
Unigrams	34	21	
Bigrams	21	13	
Trigrams	13	8	
Total	68	42	110

frequents, in order to identify the features most effective in discriminating the two classes of reviews, the Information Gain (IG) of the selected n-grams was computed (Kullback and Leibler, 1951; Yang and Pedersen, 1997).

Then, after having found the Information Gain of the n-grams of lemmas and part-of-speech, a further reduction of the features was realized. In fact, we selected a relatively small amount of features, in order to facilitate the computation of the Raykar et al.'s algorithm (discussed in Sub-section 3.2.1), and only the n-grams with the highest IG values were selected to be taken as features of the vectors which represented the reviews. In particular, the n-grams were collected according to the scheme shown in Table 2. By the way, 8, 13, 21 and 34 are numbers belonging to the Fibonacci series (Sigler, 2003). They were chosen because they grow exponentially and are used, in our case, to give wider representation to the shortest n-grams.

Lastly, two more features were added to the feature set, that is the length of the review, considered with and without punctuation. Therefore, in each fold of the experiment, the vectors of the reviews were constituted by 112 values: 2 corresponding to the length of the review, and 110 representing the (not normalized) frequency, into the review itself, of the selected n-grams of lemmas and POS.

3.1.3 Baselines

The best way to assess the improvement coming from the algorithm would have been with respect to a supervised baseline. However this was not possible as we could only be certain regarding the classification of a fraction of the reviews (our gold standard: 236 reviews, for a total of about 23,000 tokens). We felt such a small dataset could not be used for training, but only for evaluation; therefore we used instead two simple heuristic baselines.

Majority baseline. The simplest metric for performance evaluation is the majority baseline: always assign to a review the class most represented in the dataset. Since in the subset of DEREV with gold standard we had 50% of true and false reviews, simply 50% is our majority baseline.

Random baseline. Furthermore, we estimated a random baseline through a Monte Carlo simulation. This kind of simulation allows to estimate the performance of a classifier which performs several times a task over random outputs whose distribution reflects that of real data.

In particular, for this experiment, since we had 236 reviews whose 50% were labeled as false, 100000 times we produced 236 random binomial predictions, having $p = .5$. In each simulation, the random prediction was compared with our real data. It turned out that in less than .01% of trials the level of 62.29% of correct predictions was exceeded. The thresholds for precision and recall in detecting deceptive reviews were 62.26% and 66.95% respectively.

3.1.4 Models

We tested a number of supervised learning methods to learn a classifier using the classes determined by majority voting, but the best results were obtained using Support Vector Machines (SVMs) (Cortes and Vapnik, 1995), already employed in many applications involving text classification (Yang and Liu, 1999).

3.1.5 Results

The results obtained by training a supervised classifier over the dataset with classes identified with majority voting are shown in the Table 3. The highest results are in bold. The methodological approach and performance achieved in this experiment seems to be comparable to that of Strapparava and Mihalcea (2009) and, more recently, of Li et al. (2011). However Li et al. (2011) evaluate the effectiveness of different kind of features with the aim of annotating unlabeled data, while we try to evaluate the reliability of heuristic classes in training.

3.2 Learning from Crowds

3.2.1 The Learning from Crowds algorithm

As pointed out by Raykar et al. (2010), majority voting is not necessarily the most effective way to determine the real classes in problems like

Table 3: The experiment with the majority voting classes

	Correctly classified reviews	Incorrectly classified reviews	Precision	Recall	F-measure
False reviews	75	43	**83.33%**	63.56%	72.12%
True reviews	103	15			
Total	178	58			
Total accuracy	**75.42%**				
Random baseline	62.29%		62.26%	**66.95%**	

those of reviews where there is no gold standard. This is because annotators are not equally reliable, and the reviews are not equally challenging. Hence the output of the majority voting may be affected by unevaluated biases. To address this problem, Raykar et al. (2010) presented a maximum-likelihood estimator that *jointly* learns the classifier/regressor, the annotator accuracy, and the actual true label.

For ease of exposition, Raykar et al. (2010) use as classifier the logistic regression, even though they specify their algorithm would work with any classifier. In case of logistic regression, the probability for an entity $x \in X$ of belonging to a class $y \in Y$ with $Y = \{1, 0\}$ is a sigmoid function of the weight vector w of the features of each instance x_i, that is $p[y = 1|x, w] = \sigma(w^\top x)$, where, given a threshold γ, the class $y = 1$ if $w^\top x \geq \gamma$.

Annotators' performance, then, is evaluated 'in terms of the sensitivity and specificity with respect to the unknown gold standard': in particular, in a binary classification problem, for the annotator j the sensitivity α^j is the rate of positive cases identified by the annotator –i.e., the recall of positive cases– while the specificity β^j is the annotator's recall of negative cases.

Given a dataset D constituted of independently sampled entities, a number of annotators R, and the relative parameters $\theta = \{w, \alpha, \beta\}$, the likelihood function which needs to be maximized, according to Raykar et al. (2010), would be $p[D|\theta] = \prod_{i=1}^{N} p[y_i^1, ...y_i^R|x_i, \theta]$, and the maximum-likelihood estimator is obtained by maximizing the log-likelihood, that is

$$\widehat{\theta}_{ML} = \{\widehat{\alpha}, \widehat{\beta}, \widehat{w}\} = \underset{\theta}{arg\,max}\{ln\, p[D|\theta]\}. \quad (1)$$

Raykar et al. (2010) propose to solve this maximization problem (Bickel and Doksum, 2000) through the technique of Expectation Maximization (EM) (Dempster et al., 1977). The EM algorithm can be used to recover the parameters of the hidden distributions accounting for the distribution of data. It consists of two steps, an Expectation step (E-step) followed by a Maximization step (M-step), which are iterated until convergence. During the E-step the expectation of the term y_i is computed starting from the current estimate of the parameters. In the M-step the parameters θ are updated by maximizing the conditional expectation. Regarding the third parameter, w, Raykar et al. (2010) admit there is not a closed form solution and suggest to use the Newton-Raphson method.

3.2.2 Determining the class of reviews using Learning from Crowds

In order to apply Raykar's algorithm, we proceeded as follows. First, we applied the procedure for feature selection described in Subsection 3.1.2 to create a single dataset: that is, the corpus was not divided in folds, but the feature selection involved all of DEREV. This dataset was built using the classes resulting from the majority voting approach and included these columns:

- The class assignments of the four clues discussed in Sub-section 2.3 – SB, Cl, NN, UP;
- The majority voting class;
- The 112 features identified according to the procedure presented in Sub-section 3.1.2.

Then, we implemented the algorithm proposed by Raykar et al. (2010) in R.[11] We computed a Logistic Regression (Gelman and Hill, 2007) on the dataset to compute the weight vector w, used to estimate for each instance the probability p_i for the review of belonging to the class 'true'. For the logistic regression we used the 112 surface features

[11] http://www.r-project.org/

Table 4: The experiment with Raykar et al.'s algorithm classes

	Correctly classified reviews	Incorrectly classified reviews	Precision	Recall	F-measure
False reviews	85	33	**78.70%**	**72.03%**	75.22%
True reviews	95	23			
Total	180	56			
Total accuracy	**76.27%**				
Random baseline	62.29%		62.26%	66.95%	

mentioned above, adopting as class the majority voting, as suggested by Raykar et al. (2010).

The parameters α and β were estimated regarding the three clues Cl - Cluster, NN - Nickname and UP - Unknown Purchase. The attribute SB - Suspect Book was not used, in order to carry out the EM algorithm exclusively on heuristic data, removing the information obtained through sources external to the dataset. The parameters α and β of the three clues were obtained not from random classes, as the EM algorithm would allow, but again comparing the clues' labels with the majority voting class. In fact, aware of the local maximum problem of EM, in this way we tried to enhance the reliability of the results posing a configuration which could be, at least theoretically, better than a completely random one.

Knowing these values for each instance of the dataset, we computed the E-step and we updated our parameters in M-step.

The E-step and the M-step were iterated 100 times, in which the log-likelihood increases monotonically, indicating a convergence to a local maximum.

The final value of p_i determined the new class of each instance: if $p_i > .5$ the review was labeled as true, otherwise as false. In the end, the EM clusterization allowed to label 3267 reviews as false and 3552 as true, that is 47.91% and 52.09% of DEREV respectively.

3.2.3 Feature selection

The feature selection for this experiment was exactly the same presented for the previous one in Sub-section 3.1.2; the only, fundamental difference was that in the first experiment the classes derived from the majority voting rule, while in the second experiment the classes were identified through the Raykar et al.'s strategy.

3.2.4 Baselines

As in the first experiment, we compared the performance of the models with the same majority and random baselines discussed in Sub-section 3.1.3.

3.2.5 Models

We used the classes determined through the Learning by Crowds algorithm to train SVMs models, with the same settings employed in the first experiments.

3.2.6 Results

Table 4 shows the results of the classifier trained over the dataset whose the classes were identified through the Raykar et al.'s algorithm.

4 Discussion

4.1 Deceptive language in reviews

Of the 4811 reviewers who wrote reviews included in our corpus, about 900 were anonymous, and only 16 wrote 10 or more reviews. If, in one hand, this prevented us from verifying the performance of the models with respect to particular reviewers, on the other hand we had the opportunity of evaluating the style in writing reviews across many subjects.

In our experiments, we extracted simple surface features constituted by short n-grams of lemmas and part-of-speech. In literature there is evidence that also other kinds of features are effective in detecting deception in reviews: for example, information about the syntactic structures of the texts (Feng et al., 2012). In our pilot studies we did not obtain improvements using syntactic features. But even the frequency of n-grams can provide some insight regarding deceptive language in reviews; and with this aim we focused on the unigrams appearing more than 50 times in the 236 reviews constituting the gold standard of DEREV, whose un/truthfulness is known. The use of self-referred pronouns and adjectives is remarkably different in true and fake reviews: in the genuine ones, the pronouns 'I', 'my' and 'me' are found 371, 74 and 51 times respectively, while in the fake ones the pronoun 'I' is present only 149 and 'me' and 'my' less than 50 times. This reduced number of self-references is coherent with the findings of other well-known studies regarding deception detection (Newman et al., 2003); however, while in truthful reviews the pronoun 'you' appears only 84 times, in the fake ones the frequency of 'you' and 'your' is 151 and 75. It seems that while the truth-tellers simple state their opinions, the deceivers address directly the reader. Probably they tend to give advice: after all, this is what they are paid for. The frequency of the word 'read' - that is the activity simulated in fake reviews - is also quite imbalanced: 137 in true reviews and 97 in the fake ones. Lastly, it is maybe surprising that in the false reviews terms related to positive feelings/judgments do not have the highest frequency; instead in truthful reviews we found 52 times the term 'good' (and 56 times the ambiguous term 'like'): also this outcome is similar to that of the mentioned study of Newman et al. (2003).

4.2 Estimating the gold standard

The estimation of the gold standard is a recurrent problem in many tasks of text classification and in particular with deceptive review identification, that is an application where the deceptiveness of the reviews cannot be properly determined but only heuristically assessed.

In this paper we introduced a new dataset for studying deceptive reviews, constituted by 6819 instances whose 236 (that is about 3.5% of the corpus) were labeled with the highest degree of confidence ever seen before. We used this subset to test the models that we trained on the other reviews of DEREV, whose the class was heuristically determined.

With this purpose, we adopted two techniques. First, we simply considered the value of our clues of deception as outputs of just as many annotators, and we assigned the classes to each review according to majority voting. Then we clustered our instances using the Learning from Crowd algorithm proposed by Raykar et al. (2010). Lastly we carried out the two experiments of text classification described above.

The results suggest that both methods achieve accuracy well above the baseline. However, the models trained using Learning from Crowd classes not only achieved the highest accuracy, but also outperformed the thresholds for precision and recall in detecting deceptive reviews (Table 4), while the models trained with the majority voting classes showed a very high precision, but at the expense of the recall, which was lower than the baseline (Table 3).

Since the results even with simple majority voting classes were positive, we carried out two more experiments, identical to those described above except that we included in the feature set the three deception clues Cluster - Cl, Nickname - NN and Unknown Purchase - UP. Both with majority voting and with learning from Crowds classes, the accuracy of the models exceeded 97%. This might seem to suggest that those clues are very effective; but given that the deception clues were used to derive the silver standard, their use as features could be considered to some extent circular (Subsection 2.4). Moreover, not all of our non-linguistic cues may be found in all review scenarios, and therefore the applicability of our methods to all review scenarios will have to be investigated. Specifically, Cluster is likely to be applicable to most review domains, Nickname and Unknown Purchase are Amazon features that may or may not be adopted by other services allowing users to provide reviews. However, our main concern was not to evaluate the effectiveness of these specific clues of deception, but to investigate whether better strategies for labeling instances than simple majority voting could be found.

In this perspective, the performance of our second experiment, in which the Learning from Crowds algorithm was employed, stands out. In fact in that case we tried to identify the classes of the instances abstaining from making use of any external information regarding the reviews: in particular, we ignored the Suspect Book - SB clue of deception which, by contrast, took part in the creation of the majority voting classes.

This outcome suggests that, even in scenarios where the gold standard is unknown, the Learning from Crowds algorithm is a reliable tool for labeling the reviews, so that effective models can be trained in order to classify them as truthful or not.

References

Bickel, P. and Doksum, K. (2000). *Mathematical statistics: basic ideas and selected topics*. Number v. 1 in Mathematical Statistics: Basic Ideas and Selected Topics. Prentice Hall.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Feng, S., Banerjee, R., and Choi, Y. (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 171–175, Jeju Island, Korea. Association for Computational Linguistics.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

Li, F., Huang, M., Yang, Y., and Zhu, X. (2011). Learning to identify review spam. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pages 2488–2493. AAAI Press.

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.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. (2001). 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*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

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.

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

Sigler, L., editor (2003). *Fibonacci's Liber Abaci: A Translation Into Modern English of Leonardo Pisano's Book of Calculation*. Sources and Studies in the History of Mathematics and Physical Sciences. Springer Verlag.

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*.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 42–49, New York, NY, USA. ACM.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *CiteSeerX - Scientific Literature Digital Library and Search Engine [http://citeseerx.ist.psu.edu/oai2] (United States)*.