Lang Resources & Evaluation (2020) 54:1019–1058
https://doi.org/10.1007/s10579-020-09486-5

ORIGINAL PAPER

# Fake opinion detection: how similar are crowdsourced datasets to real data?

**Tommaso Fornaciari[1] · Leticia Cagnina[2] · Paolo Rosso[3] · Massimo Poesio[4]**

Published online: 28 March 2020


**Abstract** Identifying deceptive online reviews is a challenging tasks for Natural Language Processing (NLP). Collecting corpora for the task is difficult, because normally it is not possible to know whether reviews are genuine. A common workaround involves collecting (supposedly) truthful reviews online and adding them to a set of deceptive reviews obtained through crowdsourcing services. Models trained this way are generally successful at discriminating between 'genuine' online reviews and the crowdsourced deceptive reviews. It has been argued that the deceptive reviews obtained via crowdsourcing are very different from real fake reviews, but the claim has never been properly tested. In this paper, we compare (false) crowdsourced reviews with a set of 'real' fake reviews published on line. We evaluate their degree of similarity and their usefulness in training models for the detection of untrustworthy reviews. We find that the deceptive reviews collected via crowdsourcing are significantly different from the fake reviews published online. In the case of the artificially produced deceptive texts, it turns out that their domain similarity with the targets affects the models' performance, much more than their

✉ Tommaso Fornaciari
fornaciari@unibocconi.it

Leticia Cagnina
lcagnina@unsl.edu.ar

Paolo Rosso
prosso@dsic.upv.es

Massimo Poesio
m.poesio@qmul.ac.uk

[1] Bocconi University, Milan, Italy

[2] Universidad Nacional de San Luis, San Luis, Argentina

[3] Universitat Politècnica de València, Valencia, Spain

[4] Queen Mary University of London, London, UK

untruthfulness. This suggests that the use of crowdsourced datasets for opinion spam detection may not result in models applicable to the real task of detecting deceptive reviews. As an alternative method to create large-size datasets for the fake reviews detection task, we propose methods based on the probabilistic annotation of unlabeled texts, relying on the use of meta-information generally available on the e-commerce sites. Such methods are independent from the content of the reviews and allow to train reliable models for the detection of fake reviews.

**Keywords** Deception detection · Crowdsourcing · Ground truth · Probabilistic labeling

## 1 Introduction

Many E-commerce sites, such as Amazon[1], Ebay[2], Tripadvisor[3] and similar, give customers the opportunity to leave comments concerning their products. Shoppers appreciate the possibility of sharing their opinions, and often take advantage of other consumers' experience. However, the lack of controls on those who are enabled to publish reviews, exposes customers to the risk of finding texts which do not express honest opinions, but are concealed forms of commercial promotion. To identify such disguised advertisements is not trivial and the dimension of the phenomenon is difficult to estimate. Even so, there is a growing public awareness of the problem. At the same time, the awareness of the problem in academy and industry grew as well.

For example, it is now possible to use free services online, such as FAKESPOT[4], specifically dedicated to the detection of fake reviews. A number of attempts to solve the problem using Natural Language Processing (NLP) methods have also been made. But such methods require datasets to train models for the fake review detection task. Datasets of this kind have been released by Amazon and Yelp; but these companies did not provide much information concerning how the datasets were created (Sect. 2.3). An alternative approach was pursued by Ott et al. (2011), who adopted the crowdsourcing approach now widespread in NLP for the creation of corpora (Negri et al. 2011; Salloum et al. 2017; Skeppstedt et al. 2018). Ott et al. collected truthful reviews online and completed the data set with false crowdsourced texts. Their corpus has been extensively used by researchers including Feng et al. (2012), Banerjee and Chua (2014), Hernández Fusilier et al. (2015), and Lin et al. (2017).

In this paper, we assess this popular approach, using a corpus of online reviews called DeRev: DEception in REViews (Fornaciari and Poesio 2014). This corpus includes a substantial gold standard created by exploiting the appearance in the press of articles unveiling the falsity of some book reviews. This allowed us to

[1] www.amazon.com.

[2] www.ebay.com.

[3] www.tripadvisor.com.

[4] www.fakespot.com.

select a set of texts whose falsity or truthfulness was known with high confidence. In this work, we created using crowdsourcing services a second set of false reviews which mirrored our gold standard exactly. These matched data sets gave us the opportunity of evaluating:

1. the degree of similarity between the reviews published on line and those artificially produced via crowdsourcing;
2. the performance of models trained with crowdsourced reviews against the gold standard, which is the target in a real life scenario, but is unavailable in most studies.

Finally, we explore the use of methodologies for the heuristic annotation of unlabeled texts, as an alternative to crowdsourcing for the realization of training data sets.

## 2 Related work

To our knowledge, the paper by Jindal and Liu (2008) is the first study where linguistic and other 'reviewer/product centric' features were exploited to identify deceptive reviews. The authors carried out an interesting analysis of reviews and reviewers' behavior on Amazon: for example, they pointed out that 'a large number of reviewers write only a few reviews, and a few reviewers write a large number of reviews' and, similarly, 'a large number of products get very few reviews and a small number of products get a large number of reviews'. This is also called 'activity bias', or 'wisdom of a few' (Baeza-Yates 2018). They also were first to address the difficulty for human annotators in distinguishing truthful from fake reviews: indeed, they developed a method aimed to detect only reviews duplicated and/or clearly identifiable as spam.

Given such difficulty, effective algorithms for the identification of fake reviews would be particularly useful, both for researchers and for shoppers; nonetheless, the lack of reliably annotated corpora makes it problematic to apply supervised methods to this classification task.

In order to obtain reliable gold standards, that is data sets of reviews whose truthfulness or deceptiveness is known, three main approaches have been pursued.

### 2.1 Semi-supervised methods

The first approach involves using semi-supervised methods, that is to find a way to annotate a set of reviews, and then to exploit them with the aim of going ahead with the automatic annotation of unlabeled data. In this context, two main options are available, known as co-training methods and Positive Unlabeled (PU) learning.

### 2.1.1 Co-training methods

An example of use of a co-training method is the work of Li et al. (2011), who made use of the algorithm described by Blum and Mitchell (1998). The method is employed for the automatic annotation of unlabeled data and involves two steps: first, the training of a number $n$ (usually two) classifiers, relying on independent feature sets, on a given set of annotated data; then, the predictions of a classifier on unlabeled data are exploited as labels for the other(s).

Li et al. (2011) created a big corpus of around 60,000 reviews, of which 6000 where manually labeled as spam or not spam. Starting from the labeled data, they annotated the whole corpus through an iterative process which exploited two different classifiers, one based on review-centric features, and the other relying on reviewer-centric features, independently trained.

The features concerning the reviewer, however, were motivated by the consideration that "the spammers consistently write review spam" (Li et al. 2011, p. 2490), and the notion of spam, in turn, was derived from the helpfulness of the reviews, which is rated online by the readers.

Another case where the method of Blum and Mitchell (1998) was employed is the study of Zhang et al. (2016). They called their approach CoSpa: co-training for Spam review identification. The authors used the corpus created by Ott et al. (2011) (discussed below) to build two different views of the data set, one relying on lexical terms and the other on the Probabilistic Context-Free Grammar (PCFG) rules employed also by Feng et al. (2012). Then they developed two versions of their co-training algorithm, namely CoSpa-C and CoSpa-U, showing that both of them outperform single classifiers trained with the same features.

In both studies, however, the annotation of the data is problematic: in the first case the focus is on reviews which may not be necessarily defined as fake, in the second one the fake reviews are artificially produced through crowdsourcing services.

### 2.1.2 Positive unlabeled (PU) learning

Presented by Liu et al. (2002, 2003), the method is aimed to "investigate the following problem: Given a set of documents of a particular topic or class P, and a large set M of mixed documents that contains documents from class P and other types of documents, identify the documents from class P in M" (Liu et al. 2002, p. 387). In particular, the theoretical discussion of the authors led them to the conclusion that "by using positive and mixed document sets, one can build accurate classifiers with high probablity when sufficient documents in P and M are available" (Liu et al. 2002, p. 390).

The algorithm, exploited with success in the field of biology by Elkan and Noto (2008), was applied for the detection of fake reviews by Hernández Fusilier et al. (2013, 2015). Similarly to Zhang et al. (2016), Hernández Fusilier et al. worked on the data set of Ott et al. (2011), carrying out an iterative process with two steps, which they described as follows: "In the first step the whole unlabeled set is

considered as the negative class. Then, we train a classifier using this set in conjunction with the set of positive examples. In the second step, this classifier is used to classify (that is, to label) the unlabeled set. The instances from the unlabeled set classified as positive are eliminated; the rest of them are considered as the reliable negative instances for the next iteration. This iterative process is repeated until a stop criterion is reached. Finally, the latest built classifier is returned as the final classifier" (Hernández Fusilier et al. 2013, p. 40). The results of their experiments showed that effective classifiers can be built, even with a small amount of positive cases.

The same approach was applied by Li et al. (2014a), for the first time on a Chinese corpus. The authors claimed that the Collective PU learning framework outperforms the state-of-the-art baseline algorithms (Li et al. 2014a, p. 904). Nonetheless, as observed by (Rout et al. 2017, p. 1320), the "assumption regarding continual refining of negative instances over iterations will not always hold in practice": for example, the "continual but gradual reduction of the negative instances over iterations […] unfortunately is not always true" (Li et al. 2014b, p. 468). However, a careful selection of the negative, that is non-deceptive reviews, can improve the performance of the algorithm (Hernández Fusilier et al. 2015).

### 2.2 Creating deceptive reviews via crowdsourcing

The issues with semi-supervised methods discussed above led researchers to use crowdsourcing to build dedicated data sets, where the ground truth is precisely known. Such researchers typically created their own fake products' reviews and then they merged those reviews with reviews collected online, whose truthfulness is considered sure. This is the approach followed by Ott et al. (2011) with respect to online reviews, but also by Strapparava and Mihalcea (2009) regarding essays written about a number of different topics, such as abortion and suicide. The data set created by Ott et al. (2011) has been widely used in literature, for example by Feng et al. (2012) and Banerjee and Chua (2014). Another corpus of deceptive opinions called Paraphrased OPinion Spam (POPS) was created by Kim et al. (2017), just with the aim of producing false reviews paraphrasing the truthful ones. The clear advantage of this method is the possibility of working with surely false reviews. However the underlying assumption is that the reviews artificially produced are assimilable to the texts published on line. In Martens and Maalej (2019) a fake review dataset was collected using information available in the web about 'how to write a fake review' in app stores. Thus, the authors obtained 8607 fake reviews to be used with a dataset of true reviews crawled from the Apple App Store. Seven different supervised machine learning approaches were training with both datasets.

### 2.3 The Amazon and Yelp datasets

A third way to train NLP models is to rely on the datasets released by the platforms which collect reviews on line. Companies such as Amazon and Yelp are making a great deal of effort to develop methods for in identifying spamming and deceptive reviews, including releasing datasets of such reviews.

The Amazon dataset was published in early 2018, and can be found online.[5] According to Amazon, the dataset includes reviews "non-compliant with respect to Amazon policies." Unfortunately, it is not entirely clear whether these reviews are merely suspicious or have been demonstrated to be false, so we could not use this dataset in the research discussed here as we wanted to use reviews whose status as deceptive or otherwise was completely transparent.[6]

Mukherjee et al. (2013b) used the corpus of reviews released on Yelp[7]. The texts in such corpus are annotated as filtered or not filtered by Yelp's algorithm for the detection of deceptive opinion spam, whose details however are a trade secret. This suggests that the class of the reviews was probabilistically determined, rather than known *a priori*. Mukherjee et al. carried out experiments using the corpus of Ott et al. (2011) to train models employed for the detection of fake reviews belonging to the Yelp data set. In this setting, it turned out that "models trained using AMT[8] generated (crowdsourced) fake reviews are not effective in detecting real-life fake reviews in a commercial website with detection accuracies near chance" (Mukherjee et al. 2013b, p. 7). Mukherjee et al. (2013b) are not alone in raising this issue. Li et al. (2014c) discuss the work of Mukherjee et al. (2013b), and while they claim "that it is possible to detect fake reviews with above-chance accuracy" (Li et al. 2014c, p. 1574), they also admit that "it is still very difficult to estimate the practical impact of such methods, as it is very challenging to obtain gold-standard data in the real world" (Li et al. 2014c, p. 1574).

In this paper, we addressed the same issue as Mukherjee et al. that is, the effectiveness of models trained with artificially false texts, when tested against false texts produced in natural conditions. Our work addresses one of the limitations of Mukherjee et al. (2013b), who don't know the ground truth regarding the deceptiveness of the reviews, as the classes are identified by the Yelp algorithm.

The Amazon and Yelp datasets were also used by Shehnepoor et al. (2017), who developed an algorithm called NetSpam, which relies on the kind of features discussed in the next Section.

### 2.4 Deception indicators

A great deal of effort has been also spent trying to identify effective indicators of deception. Two types of features have been considered in the literature: (i) behavioral and (ii) linguistic features.

[5] https://s3.amazonaws.com/amazon-reviews-pds/readme.html.

[6] We know from discussions with Amazon researchers (p.c.) that Amazon has created a substantial task-force dedicated to identifying fake reviews, removing them from the site, and pursuing their authors, and that the fake reviews in the published dataset were identified by this task-force as being almost certainly false, but this status is unofficial.

[7] www.yelp.com.

[8] Amazon Mechanical Turk.

### 2.4.1 Behavioral features

So-called behavioral features aim to capture the behavior of the writers of deceptive reviews.

Xie et al. (2012), for example, observed that most reviewers write just one review, and then analyze through time series the behavior of those who write many reviews. Fei et al. (2013) exploit the same idea, referring in particular to the "burstiness" of the reviews, as "bursts of reviews can be either due to sudden popularity of products or spam attacks" (Fei et al. 2013, p. 175). In this paper we make use of this concept as well. Li et al. (2017) rely on the same observation in order to study the reviewers behavior and to classify their reviews as spam or genuine opinions. Saini and Sharan (2017) try to consider personality traits of the authors, according the Big Five Factor Model. This model defines five bipolar traits and has become a standard over the years (Costa and MacCrae 1992). Mukherjee et al. (2013a) carried out a study where they pointed out a variety of reviewers' practices, considered as possible indicator of spam. For example they stigmatized the fact of "posting many reviews in a single day" (p. 633), the presence of "duplicated/near duplicated versions of previous reviews" (p. 634) and the early time frame of the reviews, as "early reviews can greatly impact people's sentiment on a product" (Mukherjee et al. 2013a, p. 634).

While in general the intuitions of the researchers look reasonable and seem worth of being exploited, the lack of reliable ground truth makes difficult to evaluate their real effectiveness. In this study, we make use of the behavioral cues of deception identified by Fornaciari and Poesio (2014) and we estimate their value in Sect. 5.1.1.

### 2.4.2 Linguistic features

The linguistic features used in the literature range from very simple to highly complex. Hernández Fusilier et al. (2015) employ simple character *n*-grams and word uni-grams and bi-grams; Li et al. (2014a) refer to have tried to use Chinese character *n*-grams, even though they found that the performance was poorer than that with word uni-grams and bi-grams. In contrast Ott et al. (2011), besides uni- and bi-grams, employ relatively more complex features, that is Part-Of-Speech (POS) and the set of lexical, psychological and semantical dimensions detected by LIWC (Pennebaker et al. 2001), a lexicon commonly used in the field of deception detection. On the other hand, in Cardoso et al. (2018), word *n*-grams with *n* selected using a grid search over {1,2,3} constitute the feature set of a comprehensive analysis. Artificial and real-world datasets were used in different settings considering chronological order and posting time of the reviews in context-based classification algorithms.

A similar study carried out by Cagnina and Rosso (2017), who studied the performance of Naïve Bayes and SVM classifiers using character *n*-grams in tokens (with n 3 and 4), the sentiment score and specific LIWC linguist features such as pronouns, articles and verbs (present, past and future tenses), for the detection of deception in intra and cross domain cases. Feng et al. (2012), and later Zhang et al.

(2016), made use of deep syntactical features, extracted through the rules of the Probabilistic Context-Free Grammar (Jelinek et al. 1992).

Recently Hernández-Castañeda and Calvo (2017) tested their methods on the version of our data set released in 2014 (Fornaciari and Poesio 2014). They employed semantic features relying on a continuous semantic space model based on Latent Dirichlet Allocation (LDA) topics (Blei et al. 2003).

The use of dense word representations (Mikolov et al. 2013) in conjunction with Deep Learning techniques, which is becoming ubiquitous in NLP, has been also tested in the last years. In these studies the Bag-Of-Words approach is overtaken, as each word is projected into an abstract feature space, where the semantic similarities between words—more precisely, the similarity between the contexts where the words are found—can be measured. Pre-trained versions of these word representations—word embeddings—are available and widely employed in many studies (Mikolov et al. 2013; Pennington et al. 2014).

Word embeddings are typically employed to feed deep neural network architectures. This is the case of Ren and Ji (2017). Kim et al. (2017) applied a Bidirectional Long Short-Term Memory network (BiLSTM) (Graves et al. 2013) to the corpus of Ott et al. (2013) and to their own POPS. Zhang et al. (2018) used a Recurrent Convolutional Neural Network (RCNN), and they tested their classifier on the data set of Ott et al. (2013) as well. Similar models were used in Bhargava et al. (2018) in order to analyze and compare convolutional neural network, long short-term memory and recurrent neural networks.

Lastly, an original study has been carried out by Hovy (2016), where he reverses the problem, wondering: "So far, NLP has been used mostly for detection, and works well on human-generated reviews. But what happens if NLP techniques are used to generate fake reviews as well?" (Hovy 2016, p. 351). Having posed the problem in these terms, he employed generative models to produce fake reviews, and then he tested the effectiveness of models for deception detection against reviews automatically created: to our best knowledge, this is the first experiment of this kind in the field of deceptive opinion spam, and it shows a very plausible scenario of the forthcoming challenges for the researchers. Even in this case, however, the author employed simple *n*-grams and some word-centric meta-information as features (Hovy 2016, p. 354).

## 3 Contributions

The contributions of this paper are threefold:

– We provide a new resource available on github, consisting of two matched corpora for opinion spam detection. The two corpora—DeRev and Crowd-DeRev—consist of reviews published on Amazon and produced by crowd-sourcing respectively. DeRev is composed of labeled and unlabeled reviews published online. The labeled ones are divided in truthful and deceptive reviews and constitute a gold standard, as the deceptive reviews were identified

exploiting external sources of information (Sect. 4). The crowdsourced reviews of Crowd-DeRev are perfectly aligned with the gold standard of DeRev.
– We show that crowdsourced and online fake reviews substantially differ from each others, even when they are about the same products (Setion 6.2.3). This finding should make researchers wary of training classifiers for the detection of opinion spam with crowdsourced corpora, as such classifiers might not generalize to real use cases.
– We address the problem of how to get datasets of an appropriate size by proposing alternative methods for the probabilistic annotation of online unlabeled reviews. Our proposed methods methods rely on information commonly provided by e-commerce websites and lead to the creation of training data having stronger validity than the crowdsourced ones.

## 4 The datasets

### 4.1 The opportunity

On September 4th, 2012, an article by Alison Flood was published on the Guardian,[9] reporting about the crime writer Jeremy Duns's activities to unmask writers of deceptive reviews. He had carried out real investigation activities and had been able to discover a number of 'sock puppeteers' colleagues, that is, authors writing and/or paying for the composition of glowing reviews of their own books, or disparaging reviews of their competitors' work. After reading that article, we contacted Duns and he was extremely helpful, giving us several hints to recognize possible cues of deception in the reviews. We exploited his suggestions for the creation of our data set, as discussed in Sect. 4.2 and thanks to him we collected further information from some other articles on the topic, in particular some that appeared on *The New York Times*.

On July 25th, 2011, an article was published on Moneytalksnews, entitled '3 Tips for Spotting Fake Product Reviews—From Someone Who Wrote Them',[10] in which the author Sandra Parker discussed her experience as professional review writer. She stated that advertising agencies were used to pay her $10–20 for writing reviews on e-commerce sites like Amazon.com, and that, even though she was not formally asked to lie, 'if the review wasn't five star, they didn't pay'. Later, David Streitfeld, in his article published on the New York Times,[11] August 19th, 2011, interviewed again Sandra Parker, who partially modified her previous statement as follows: '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'. However, it seems that in both cases only five

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

[10] http://www.moneytalksnews.com/2011/07/25/3-tips-for-spotting-fake-product-reviews.

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

stars reviews would have been published: this is useful information which we exploited (Sect. 4.4.2).

Sandra Parker also made a number of common-sense suggestions that could be used to detect suspicious reviews. And the name Sandra Parker itself was a valuable hint, as it enabled us to find on Amazon her reviews, which can be considered definitely false. Furthermore, knowing for which products reviews had been purchased enabled us to collect other users' comments about products false opinions of which has been posted online.

Finally, yet another article on the New Yor Times,[12] written by David Streitfeld on August 25th gave us the opportunity. to know the titles of four books, whose authors confessed to have paid for receiving reviews of their texts.

### 4.2 DeRev 2014: corpus creation

Thanks to the information provided by the articles discussed above we were able to create a corpus we called DeRev (DEception in REViews), initially constituted by 6819 book reviews posted on Amazon.

The products reviewed in DeRev were 68 different books, chosen with the purpose of identifying products whose reviews could have been genuine or fake with high probability. Concretely, we distinguished two categories of books: Suspect Books (SB) and Innocent Books (IB).

We characterized Suspect Books as follows:

1. We considered as Suspect Books the four books discussed in the mentioned article written by David Streitfeld (August 25th, 2012);
2. We added four further books, written by three of the authors of the previous group;
3. We added the 22 books for which Sandra Parker wrote a review;
4. Lastly, we realized that some reviewers of the books pointed out by David Streitfeld tended to write reviews relatively to a small and defined set of books: we identified 16 such books, and considered them as suspect as well.

On November 17th, 2012[13], we scraped the reviews of the 46 books considered as suspect according to the definition above, which received 2707 reviews in total.

We also collected the reviews of 22 so called Innocent Books. These were chosen among either classics written by authors such as Arthur Conan Doyle or Rudyard Kipling, or among books written by living writers of such renown that buying reviews for them would have been pointless: examples include Ken Follett and Stephen King. 4112 reviews of Innocent Books were scraped in total. This version of the corpus was employed for the experiments previously carried out by Fornaciari and Poesio (2014).

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

[13] We specify the date of the scraping because, obviously, the amount of reviews changes as time passes.

Two further observations can be made. First, the ratio between the number of reviews and books reveals that the Innocent Books (4112/22) received an amount of reviews much greater than the Suspect Books (2707/68): an average of about 187 reviews per Innocent Book as opposed to 40 for each book. Second, 4811 different reviewers can be counted, but this figure is obtained by considering all anonymous reviewers as one (874 anonymous reviews appear in DeRev—that is, 12.82% of the corpus). Thus in reality the number of reviewers is much greater.

### 4.3 DeRev 2018: corpus filtering

A more careful examination of the corpus, suggested by the study of Jindal and Liu (2008), made us realize that a number of reviews in the original DeRev were perfectly identical to each other, having been repeatedly published. This occurred 45 times. In 41 cases, the same review was published twice; but in the remaining 4 cases the reviews were posted more than two times, and in one case the same text was posted 8 (!) times.

We were aware that duplicated reviews are probably spam, but in order to avoid carrying out the training and the test of our models on identical reviews, instead of keeping the multiple copies in the corpus and using that information for spam detection, we only kept one copy of the repeated reviews, removing the duplicates. As a result, 60 reviews were discarded: 6 from the 2707 of Suspect Books, 54 from the 4112 of the Innocent Books, and the final size of DeRev was reduced to 6759 reviews.

The fact that we found a higher number of duplicated—namely, spam—reviews of Innocent Books than of the Suspect Books would appear counterintuitive. But in fact, this suggests that the fake reviews we are interested in are a specific form of spam reviews: those which convey false opinions. This is not necessarily the case for all duplicated reviews, which may simply be the result of users exploiting others' opinions. By contrast, those who write fake reviews are typically asked to produce those false opinions themselves, and are likely subject to some quality check of their activity.

After this filtering process, the reviews written by an anonymous author went down to 858–12.69% of the reviews–and the number of reviewers to 4810: compared to the previous version of the corpus, one reviewer was lost as the same review was signed with two different nickname the same day. The size of DeRev was 1,160,015 tokens, considering punctuation blocks (such as the ellipsis) as a single token. The mean size of the reviews was 171.63 tokens. Column 'Whole DeRev' in Table 1 summarizes these statistics. The titles of the reviews were neither included in these statistics nor in the following analyses.

### 4.4 The gold standard

#### *4.4.1 The gold standard in Fornaciari and Poesio (2014)*

In the first study where we made use of this corpus (Fornaciari and Poesio 2014), the ground truth was determined by relying not only on the information that a book

could be considered as Suspect or Innocent (Sect. 4.2), but also on the non-linguistic cues of deception discussed in the following Sect. 5.1.1. This led to the selection of 236 reviews from the 6819 of DeRev, whose deceptiveness or truthfulness was known with a very high degree of confidence.

Identifying deceptive reviews this way, however, had several shortcomings. First, since the cues of deception were used for the determination of the gold standard, they could not be employed as features in the models for the classification task. This was an opportunity lost, as the meta-data provided by Amazon could convey good information regarding the truthfulness of the reviews. Second, the small amount of reviews in the gold standard prevented us from carrying out experiments involving only the gold standard itself, both for the training and for the test of the models. In fact, in order to carry out the training of the models, in Fornaciari and Poesio (2014) algorithms for the probabilistic annotation of the whole corpus were employed, which prevented from evaluating the performance of the models when trained with a set of instances whose class was fully reliable.

#### *4.4.2 The gold standard used in this study: the DeRev 2018 gold standard*

In order to overcome these shortcomings, in this study we used a new gold standard, which does not depend on the non-linguistic cues of deception, and whose size is greater than that of the previous version. We achieved this by considering as false the reviews:

1. for the books of the four writers who admitted to have bought reviews (Mark Husson, Peter Biadasz, Roland Hughes and John Locke), which were ranked with 5 stars;
2. written by Sandra Parker, regardless of the number of stars they received.

By contrast, the truthful reviews were randomly selected from the reviews of Innocent Books, which received a ranking of 5 stars as well. Basically, we only used the cue of deception coming from our *a priori* knowledge—the Suspect Books—while the other cues of deception, which are purely heuristic, were not considered. In this way, we ended up with 776 false and 776 true reviews, for a total of 1552. Column 'DeRev gold standard' in Table 1 summarizes the statistics of this new gold standard.

Interestingly, the overlap between old and new gold standard is minimal: only 62 reviews belong to both. This is due both to the different selection criteria of the false reviews and to the different randomization process in the collection of the truthful reviews. However, as far as the 62 common reviews are concerned, the correspondence between the two gold standard is complete: there are 38 reviews labeled as false and 24 labeled as true in both the gold standards, and no cases where a review in the old gold standard is labeled as false and in the other one as true, or the other way round.

### 4.5 Crowd-DeRev: a crowdsourced replica of the gold DEREV

The ideal conditions created by the existence of the gold standard just described—that is, the possibility of carrying out an experiment using genuinely truthful and deceptive reviews posted online—rarely occur in deception detection. For this reason, many researchers build artificially their own corpora resorting to crowdsourcing services for the creation of false texts, as discussed in Sect. 2.2. In order to assess this approach to creating a dataset for the detection of fake reviews, and to compare it with the approach followed by (Fornaciari and Poesio 2014) to use a gold standard, we applied the same strategy to create a second corpus of artificially created fake reviews. Using CrowdFlower[14], we created a replica of our DeRev corpus, that we called crowdsourced DeRev-Crowd-DeRev. This new corpus was designed to be a mirror as accurate as possible of the gold standard.

Specifically, we commissioned 1552 reviews of the exact same books for which we have in our gold standard reviews whose falsity or truthfulness is known, as described in Sect. 4.4.2. The crowdsourced corpus consists of 776 reviews of the books that received genuine reviews in the gold standard, and 776 reviews of the books that received false reviews in the gold standard. The statistics of this new set of reviews are shown in Table 1, column 'Crowd-DeRev'.

#### *4.5.1 Task description and postprocessing*

The workers were given the following instructions:

> *Imagine you work for a company which sells reviews for a number of products.*
> *You are asked to write a fake review of a book (as if you were a reader) to be posted on amazon.com.*
> *The review needs to sound realistic and portray the book in a positive light.*
> *Feel free to look at their page if you are not familiar with the book, but do not copy or plagiarize the content of other reviews already posted.*
> *The size of the review should be at least 750 characters and the task should be carried out in about 5 min.*
>
> *Writer: ...*
> *Title: ...*
> *Link: ...*

After the creation, the reviews were cleaned as follows:

- All reviews produced by workers were checked for plagiarism using multiple online tools such as *grammarly*[15], *PaperRater*[16] and *SmallSEOTools*[17].
- Duplicate and non English reviews were discarded.
- Reviews with 30% or more of the text plagiarized were discarded.

[14] www.crowdflower.com.

[15] https://www.grammarly.com/plagiarism-checker.

[16] https://www.paperrater.com/plagiarism_checker.

[17] http://smallseotools.com/plagiarism-checker/.

**Table 1** Statistics for DeRev 2018

| | Whole DeRev 2018 | DeRev 2018 gold standard | DeRev |
|---|---|---|---|
| Total reviews | 6759 | 1552 | 1552 |
| Truthful | 776 | 776 | 0 |
| False | 776 | 776 | 1552 |
| Unlabeled | 5207 | 0 | 0 |
| Reviewers | 4810 | 1276 | 1552 |
| "Anonymous" reviewer | 858 (12.69%) | 163 (10.50%) | (Crowdsourcing) |
| Books reviewed | 68 | 52 | 52 |
| Tokens number | 1,160,015 | 248,496 | 289,463 |
| Reviews' mean tokens | 171.63 | 160.11 | 186.51 |

## 5 Deception detection methods

In the following Section we report the results of a series of experiments aiming at comparing DeRev with Crowd-DeRev. A key method used for this comparison is training deception detection models using one corpus or the other. In this Section, we discuss how these models were trained.

### 5.1 Linguistic features

In our experiments, we made use of a set of common linguistic features, similarly to the studies discussed in Sect. 2.4. A light form of preprocessing was carried out on the text of the reviews. First, the corpus was tokenized, considering as single tokens single words and punctuation marks, which were selected in blocks. This means that terminal punctuation, such as a dot or a question mark, is collected as a token, but the same is done for groups of punctuation marks such as the three dots of the ellipsis '…' or the sequence of the smile ':-)'. The whole corpus was also put in lower-case, and then lemmas were retrieved and Part Of Speech were tagged using TreeTagger[18] (Schmid 1994).

In our experiments, we considered the following kinds of features:

- from uni- to tetra-grams of lemmas;
- from uni- to tetra-grams of POS;
- tri- and tetra-grams of characters, not considering the spaces: that is, collected from words of at least three/four characters;

The choice of considering relatively long *n*-grams was motivated by the purpose of not discarding formulaic sequences of words, which may occur more frequently in fake or in genuine reviews.

The most informative *n*-grams were then selected according to their Information Gain (IG) Forman (2003). Information Gain is a widely used metric %citeforman:ig,

[18] www.ims.uni-stuttgart.de.

which measures the decrease in entropy when the feature is given vs. absent, according to the formula:

$$IG = e(pos, neg) - [P_{n-gram}e(tp,fp) + P_{\neg n-gram}e(fn, tn)]$$

in which $e$ is the entropy:

$$e(x, y) = -\frac{x}{x+y}\log_2\frac{x}{x+y} - \frac{y}{x+y}\log_2\frac{y}{x+y}$$

and $P_{n-\text{gram}}$, $P_{\neg n-\text{gram}}$ are defined as follows:

$$P_{n-\text{gram}} = \frac{tp+fp}{all}$$
$$P_{\neg n-\text{gram}} = 1 - P_{n-\text{gram}}$$

where:

- $tp$ = true positives: presence of the cue of deception with respect to a false review;
- $fp$ = false positives: presence of the cue with respect to a genuine review;
- $tn$ = true negatives: absence of the cue with respect to a genuine review;
- $fn$ = false negatives: absence of the cue with respect to a false review;
- $pos$ = positives: number of false reviews, $tp + fn$;
- $neg$ = negatives: number of true reviews, $fp + tn$.

(Since we are interested in creating a fake reviews detector, we consider as positive cases the deceptive reviews, and negative the truthful ones.)

Basically, we use IG to collect the features most imbalanced in our classes, i.e., which have the highest predictive value. In our case, we calculated it considering the texts belonging to the training set of each experiment, and adopting their training class (which, in some experimental conditions, is not the same class used for the test—Sects. 6.2.2, 6.2.3 and 8). In order to prevent the collection of features concerning only a small subset of the data, the IG was calculated for the $n$-grams recurring at least 100 times in the whole corpus. Then the $n$-grams of lemmas and POS were ranked in an unique list according to their IG value, and the 150 $n$-grams with the highest scores were employed as features for the experiments.

#### *5.1.1 Behavioral features*

We also employed few non-linguistic cues, identified according to the information we obtained thanks to Jeremy Duns and to the experience of other researchers mentioned in Sect. 2. In our settings, we computed the presence vs. the absence of these cues as sign of deceptiveness/truthfulness of the reviews. These clues are:

Cluster (Cl) The first clue comes from the suggestion made by Sandra Parker in her article. As she pointed out, the

| | |
|---|---|
| | agencies which provide review services usually gave her 48 h to write a text. 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, in the study of Fornaciari and Poesio (2014) we considered as positive this clue of deceptiveness if the review belonged to a group of at least two reviews posted within 3 days.<br>Since, literally, the advice of Sandra Parker was to pay attention to the reviews produced within 48 h, in this study we made the Cl clue more restrictive, and considered it as present if two reviews appeared within only two days. In this way, 522 reviews were found, where the Cl clue had been flagged as present by Fornaciari and Poesio (2014) and it was considered absent in this study. Remarkably, considering the gold standard employed in the study of Fornaciari and Poesio (2014)—described in Sect. 4.4.1—there is only one review where the clue was considered present according to the old criterion, and absent according to the new one: this means that the most suspect reviews were almost always posted simultaneously with others, within the little lapse of time of 48 h. Therefore the new, more restrictive threshold should have made the Cl clue more precise in identifying fake reviews. |
| 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 hypothesize it is less probable that the writers of fake reviews post their texts using their true name. This line of reasoning is not dissimilar to that of Li et al. (2011), who pointed out that the reviewers can be divided in habitual spammers or not spammers. |
| 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. |

While the Cluster cue was computed following the suggestions discussed in Sect. 2 and those given by Jeremy Duns and Sandra Parker in Sect. 4.1, NickName and Unknown Purchase were directly provided by the Amazon website.

**Table 2** Information gain for the cues of deception against the gold standard

| Cue | IG |
|---|---|
| Cluster (Cl) | 0.3079 |
| NickName (NN) | 0.0001 |
| Unknown purchase (UP) | 0.1559 |

After the identification of the cues, we evaluated their effectiveness calculating their Information Gain against the gold standard. The results are summarized in Table 2. The IG value of CL and UP is exceedingly high, compared to the values of the linguistic features, selected as described in Sect. 5.1: this is an advantage we exploit in our experiments. The surprise is represented by the cue NickName, whose IG value breaks down to a value which is close to zero and widely below the threshold we adopted for the selection of linguistic features. In this case, clearly a cue which we assumed to be a good predictor of deception turns out to be weak. However, in order to avoid the possible bias introduced by this information provided by the gold standard, in the following experimental conditions where these cues were used as features, we used all of them, not only those which turned out to be actually useful.

### 5.2 Training models

For this study we tested a number algorithms performing supervised classification, and we obtained the best results with Support Vector Machines (SVM) (Cortes and Vapnik 1995): a well-known method which has proven successful in a number of applications, including deception detection (Fornaciari and Poesio 2013; Yang and Liu 1999). Its success mostly depends on the ability of dealing with entities which would not be linearly separable in the feature space, applying them kernel functions that shift the entities in a higher dimensions space, where the linear separation is possible (Zhou et al. 2008). Hence, the choice of the kernel function is crucial for the effectiveness of the models. While in literature linear kernels are considered useful for text categorization, as texts in vector space are often represented by sparse vectors (Karatzoglou et al. 2006), in our experiments radial kernels gave the best performance.

Our use of SVMs rather than the Deep Learning methods widely used in recent literature on text classification is primarily due to the relatively small size of our corpus, which would have prevented training of reliable word embeddings specific for the task. A second reason is that in most literature on deception detection more traditional classifiers are used, and therefore our results can be compared more directly to those obtained by others.

The validation method for the training of the model was tenfold cross-validation.

### 5.3 Baselines

For the evaluation of the models, usually majority or random baselines are employed as the least challenging thresholds. Majority baseline simply corresponds

to the rate of the most frequent class in the data set. Since, in our case, in the gold standard used for the tests the two classes are divided in equal parts, the possible threshold of 50% cannot be considered, as it would be obtained simply flipping a coin.

Just to evaluate the possible performance of a coin is the basic idea of the random baselines. For example, the so called Monte Carlo simulation consists in flipping a coin having the same probability distribution of the classes in the data set, for a high number of times. In our case, we reiterated for 100 000 times 1552 random predictions with $p[y = 0] = 0.5$, as many entities we have in the gold standard. Comparing the random predictions with the gold standard, we found that in less than 0.01% of the cases the accuracy was higher than 54.70%, the precision higher than 54.69% and the recall higher than 56.70%.

## 6 Comparing DeRev and Crowd-DeRev

In order to compare the two matched corpora we carried out a series of experiments which can be distinguished according to the kind of training set that was employed, as follows:

1. In Sect. 6.1, we discuss an experiment in which we exploited the availability of a substantial gold standard to test fake review detection in a traditional supervised learning setting, where each instance is associated to a certain and unambiguous label. In this setting, we did not employ the 5207 DeRev reviews not belonging to the gold standard—whose class is unknown—, and we performed a tenfold cross-validation on the 1552 reviews of the DeRev 2018 gold standard.
   The performance of the model trained in this experiment could be considered as an upper bound: what performance fake review detection models could achieve when training on real fake reviews and real genuine reviews.
2. The experiments in Sect. 6.2 explore the most common approach to deception detection in the literature, where researchers, not having proper gold standards, replace them with data sets consisting of a combination of reviews collected online and reviews artificially produced using crowdsourcing.
   - As discussed above, in these studies typically the classifier has to distinguish (supposedly) genuine online and crowdsourced fake reviews: this experimental design is replicated in Experiment 2 (Sect. 6.2.1).
   - DeRev, however, includes fake reviews published on line. Crowd-DeRev contains crowdsourced deceptive reviews: they concern the same books which belong to the DeRev 2018 gold standard (truthful and deceptive reviews published on line). This makes it possible to try some novel experimental designs. In Experiment 3 (Sect. 6.2.2), we use again truthful online reviews and crowdsourced fake reviews, but in this case the crowsourced reviews are about books for which we have truthful reviews on line. This means that, unlike in experiment 2, truthful (online) and deceptive

**Table 3** Experiment 1

| Experimental design | | | | |
|---|---|---|---|---|
| Training set | DeRev 2018 gold standard | | | |
| Test set | DeRev 2018 gold standard | | | |
| Features | 147 linguistic + 3 behavioral | | | |
| | 150 linguistic | | | |
| Cross-validation | Tenfold | | | |
| **Performance** | | | | |
| | Accuracy (%) | Precision (%) | Recall (%) | F-measure (%) |
| Model (linguistic + behavioral) | **93.04** | **94.53** | **91.37** | **92.92** |
| Model (linguistic only) | 89.56 | 92.88 | 85.70 | 89.14 |
| Monte Carlo baseline | 54.7 | 54.69 | 56.7 | 55.68 |
| **Confusion matrix** | | | | |
| | False reviews | True reviews | | |
| Predicted false | 709 | 41 | | |
| Predicted true | 67 | 735 | | |

Bold indicates the highest value of the column

(crowdsourced) reviews are about exactly the same books. In this way we can evaluate the effect of domain similarity on the classifier. Even though the conditions are possibly harder than in Experiment 6.2.1, we still would expect the classifier could distinguish well between genuine and artificial reviews.

– Finally, in Experiment 4 (Sect. 6.2.3), we measure the ability of a review classifier to discriminate between deceptive online reviews and crowdsourced deceptive reviews about the same books that received online fake reviews (the Suspect Books). In this case, the data set contains only deceptive reviews, both online and crowdsourced, and also the texts regard the same topics. Therefore, assuming that the deceptive reviews are similar to each others, regardless to the way they were created, we would expect a poor classifier performance.

3. Lastly, as an alternative to crowdsourcing, in Sect. 7 we discuss probabilistic methods for labeling on line reviews, in order to create data sets for deception detection. Thanks to the DeRev 2018 gold standard, we can also measure the reliability of the probabilistic labels. The performance of the probabilistic labels, then, becomes the baseline for the models which are trained with the probabilistic labels. The last experiment, in fact, is trained using the probabilistic labels applied for the whole DeRev, that is both to the 5207 unlabeled reviews, not belonging to the gold standard, and to the 1552 reviews of the gold standard. In both cases, however, we trained the model with the

probabilistic labels, and we measured the performance against the labels of the gold standard.

## 6.1 Experiment 1: training and testing models with the DeRev 2018 gold standard

Table 3 summarizes the results of the experiment where both training and test set are genuine reviews and 'genuinely fake' reviews in the DeRev 2018 gold standard. In such setting, truthful and deceptive reviews can be classified with very high performance: a classifier using both linguistic features and the behavioral cues described in Sect. 5.1.1 achieved a F-measure of almost 93%. This is actually not surprising, as it is consistent with the previous literature in the field: authors such as Ott et al. (2011) and Feng et al. (2012), for example, found similar levels of accuracy in their experiments. Even without using behavioral cues (which are not available for the crowdsourced texts used in the next experiments), we still achieved a F-measure of 89.14%; again, in line with the previous literature. This outcome underlines the usefulness of non-linguistic cues of deception, which determine an improvement of more than 3 percentage points in the overall performance, but also that a very good performance can be achieved using just the linguistic features used in the following experiments.

## 6.2 Experiments 2, 3 and 4: training models with mixed training sets

### *6.2.1 Experiment 2: training truthful online and 'false' crowdsourced reviews*

In this experiment we tested the extent to which deception detection models trained according to the practice most commonly found in the literature are able to identify genuine fake reviews. To do this, we use to train the models the most common experimental design found in the literature: use as training set a mix of truthful reviews published online and deceptive ones artificially created using crowdsourcing. (The 776 truthful reviews of the gold standard, and the 776 crowdsourced reviews of the same books reviewed by the genuinely fake reviews in the gold standard.) This training set is thus conceptually similar to those employed by authors such as Ott et al. (2011), Feng et al. (2012), Banerjee and Chua (2014), Zhang et al. (2016) and so on.

The key difference from this previous work is that in this experiment the models trained on genuine online reviews and crowdsourced fake reviews are not then tested on other crowdsourced 'fake' reviews, but on genuine and 'genuinely fake' reviews collected online from our gold standard. (Tenfolds cross-validation is used to ensure that there is no overlap between the genuine reviews used for training and those used for testing.)

The results, summarized in Table 4, show that the model trained using crowdsourced fake reviews still performs well above the baseline—75.77% of accuracy and 72.67% of F-measure—but its performance is much lower than that of the equivalent (i.e., linguistic features-only) model trained on 'genuinely fake'

**Table 4** Experiment 2

| Experimental design | |
|---|---|
| Training set | **True** reviews from DeRev 2018 gold standard |
| | False reviews from the Crowd-DeRev reviews mirroring the **false** ones of the DeRev 2018 gold standard |
| Test set | DeRev 2018 gold standard |
| Features | 150 linguistic |
| Cross-validation | Tenfold |

Performance

| | Accuracy (%) | Precision (%) | Recall (%) | F-measure (%) |
|---|---|---|---|---|
| Model | **75.77** | **83.33** | **64.43** | **72.67** |
| Monte Carlo baseline | 54.7 | 54.69 | 56.7 | 55.68 |

Confusion matrix

| | False reviews | True reviews |
|---|---|---|
| Predicted false | 500 | 100 |
| Predicted true | 276 | 676 |

Bold indicates the highest value of the column

reviews (89.56% accuracy and 89.14 F value) even if both models are trained on reviews *of the same books*. This suggests that there may be some difference between the *language* used in 'crowdsourced fake' and the language used in the 'genuinely fake' reviews. Experiment 4 was specifically designed to test this hypothesis.

But there is something else that we need to check. This is the fact that even this lower performance for the models trained using 'crowdsourced fake' reviews could still be the result of overly favourable experimental conditions. This is the fact that genuine and 'crowdsourced fake' reviews are about different books. So even though feature selection is independently carried out for each fold during cross-validation process, some domain-dependent information may have still positively biased the classification. This could not happen if these methods were applied in a real-life scenario, in which the reviews to be classified could be about unseen books, or books might receive both genuine and fake reviews. The next experiment was designed to explore this concern.

### *6.2.2 Experiment 3: training truthful online and 'truthful' crowdsourced reviews*

Since the configuration of the training set used in Experiment 2 is actually not realistic, as it is normally impossible to know in advance the content of the test set, in this experiment we replicated Experiment 2 but modified one detail, exploiting the fact that Crowd-DeRev contains fake reviews not only of the Suspect Books in

DeRev, but of the Innocent Books. In Experiment 3, the set of 'crowdsourced fake' reviews employed in training consists of fake reviews that match the truthful reviews in the gold standard, instead of reviews matching the reviews of Suspect Books. i.e., truthful and 'crowdsourced fake' reviews are about *the same books*, the Innocent Books that received truthful reviews online.

But now, when we evaluate our model against the gold standard, as we did in the previous experiment, the performance of the model lowers to chance level: accuracy falls to 50.84% and the F-measure to 43.19% (Table 5).

This outcome shows that, even though in both cases the training set contains truthful and false reviews, the effectiveness of the models is heavily affected by the degree of domain correspondence between training and test set. Therefore, one can suppose that, in Experiment 2 (and perhaps in other similar studies), what the models really classify is not the texts' deceptiveness, but a mix of stylistic and content-related differences.

#### *6.2.3 Experiment 4: 'false' crowdsourced and false online reviews*

The previous experiment suggested that detecting deception could be affected by the domain similarity between texts. In this Experiment we return to the other question raised by Experiment 2: the extent to which crowdsourced fake reviews are similar to the fake reviews published online. If this were the case, it would be possible to claim that, in the absence of 'genuinely fake' reviews, 'crowdsourced fake' reviews could be used as a replacement. This is the assumption behind studies where artificial reviews created in laboratory are employed to classify reviews published online.

We already discussed how this experiment employs 776 genuinely fake reviews collected online and belonging to the gold standard, and the 776 artificial reviews about the very same books. These data can be used to remove any effect due to a domain difference between artificial and genuine reviews.

In order to check the similarity between the two types of 'fake' reviews of the same books, a classifier was trained to distinguish 'artificial fake' from 'genuinely fake' reviews. Note also that any difference between them is unlikely to depend on their domain, but it would have to do with differences in the form of language used by professional writers of fake reviews and crowdsourced writers. If there were no difference between these two kinds of fake reviews, the models should perform poorly in classifying them.

Table 6 shows that the model can discriminate between 'genuinely fake' and 'crowdsourced fake' reviews of the same books very easily: overall accuracy is almost 86%, and F-measure 87%. Such high levels of accuracy suggest that crowdsourced and genuinely fake reviews are not interchangeable. Moreover, considering this outcome, it would become difficult to interpret the results of experiments conducted in that way. In particular, one might wonder if the classifiers really detect deception, or other characteristics of the texts, ranging from the differences between different groups of writers to those possibly related to the texts' topics.

**Table 5** Experiment 3

| Experimental design | | | | |
|---|---|---|---|---|
| Training set | **True** reviews from DeRev 2018 gold standard | | | |
| | False reviews from the Crowd-DeRev reviews mirroring the **true** ones of the DeRev 2018 gold standard | | | |
| Test set | DeRev 2018 gold standard | | | |
| Features | 150 linguistic | | | |
| Cross-validation | Tenfold | | | |
| Performance | | | | |
| | Accuracy (%) | Precision (%) | Recall (%) | F-measure (%) |
| Model | 50.84 | 51.15 | 37.37 | 43.19 |
| Monte Carlo baseline | **54.7** | **54.69** | **56.7** | **55.68** |
| Confusion matrix | | | | |
| | False reviews | True reviews | | |
| Predicted false | 290 | 277 | | |
| Predicted true | 486 | 499 | | |

Bold indicates the highest value of the column

## 7 An alternative to crowdsourcing: creating a dataset through probabilistic labeling

The experiments in the previous Section suggest that 'crowdsourced fake' reviews are clearly different from 'genuinely fake' reviews, even when those reviews are of the same products. (Experiment 3 further suggests that some of the positive results obtained with crowdsourced reviews may be due to a domain effect.) The implication is that datasets for deception detection created via crowdsourcing may not be genuinely representative. But unfortunately, we cannot expect to have for other types of deceptive reviews the same opportunity to collect true gold data that we had for Amazon book reviews thanks to the efforts of Jeremy Duns and the confession of Sarah Parker and others. In this Section we explore an alternative approach to create such datasets for other domains: a method for the heuristic annotation of reviews which is completely agnostic with respect to their topics and does not involve the use of crowdsourcing services. The approach involves creating a silver standard through probabilistic annotation.

In the experiments discussed in this Section we probabilistically annotated the whole DeRev (that is, both unlabeled and labeled reviews—or, in other words, both those belonging and those not belonging to the gold standard) and then we trained models using such probabilistic labels. These models were tested against the gold standard.

**Table 6** Experiment 4

| Experimental design | |
|---|---|
| Training set | **False** reviews from DeRev 2018 gold standard |
| | False reviews from the DeRev reviews mirroring the **False** ones of the DeRev 2018 gold standard |
| Test set | **False** reviews from DeRev 2018 gold standard |
| | False reviews from the Crowd-DeRev reviews mirroring the **False** ones of the DeRev 2018 gold standard |
| Features | 150 linguistic |
| Cross-validation | Tenfold |

Performance

| | Accuracy (%) | Precision (%) | Recall (%) | F-measure (%) |
|---|---|---|---|---|
| Model | **85.89** | **80.57** | **94.59** | **87.02** |
| Monte Carlo baseline | 54.7 | 54.69 | 56.7 | 55.68 |

Confusion matrix

| | False reviews | True reviews |
|---|---|---|
| Predicted false | 734 | 177 |
| Predicted true | 42 | 599 |

### 7.1 The silver standard

The typical situation in detecting opinion spam is having (lots of) real life data, but not knowing their ground truth. This is the situation in DeRev as well for the 5207 reviews not belonging to the 2018 gold standard. In such scenario, a possibility is to assign probabilistic labels to the texts thus creating a silver standard. This means to carry out the task of text classification in absence of training, just applying some form of unsupervised learning.

Once probabilistic labels are given to the whole corpus, including the reviews belonging to the gold standard, it is possible to evaluate the performance of the classes obtained thanks to the unsupervised method, against those of the gold standard itself. Such performance can be considered as a qualified threshold that should be beaten by the subsequent supervised models: otherwise, it would be pointless to train them and they could not be considered worth of being used.

The usual way to assign probabilistic labels is to rely on the prediction of (supposed) experts, even considering the possibility of weighting their degree of expertise. The alternative pursued here is to assign the classes exploiting the behavioral cues of deception described in Sect. 5.1.1, treating the heuristics as a set of labelers. For each instance, the prediction of all labelers/cues can be overall used

to assign the class, with reliability hopefully higher than that of each single labeler, individually considered. In the following subsections we describe various methods for deriving probabilistic labels from the predictions of our clues, with the aim of maximizing their accuracy identifying the deceptive reviews.

### 7.2 Majority voting

The simplest strategy to use the cues of deception to obtain a label is majority voting (MV). It simply consists in assigning to each entity the class predicted by the majority of the annotators:

$$y_i = \begin{cases} 1 & \text{if} \quad 1/R \sum_{j=1}^{R} y_i^j > 0.5 \\ 0 & \text{if} \quad 1/R \sum_{j=1}^{R} y_i^j < 0.5 \end{cases} \tag{1}$$

where the instance $y_i$ receives the class $j$ assigned by the majority of the annotators set $R$.

In our corpus, we estimated the majority voting class using the 3 behavioral cues of deception. In particular, the reviews were considered true in presence of 0 or 1 cue of deception, and false in front of 2 or 3.

Table 7 summarizes the distribution of the number of deception cues in DeRev. As one can see, 67.41% of the reviews are considered false. This unexpected outcome, which would suggest that more than two-thirds of the reviews online could be false, can be interpreted as a sign of weakness of the heuristic cues to detect deception: after all, it may be common that customers review products not using their real name, having written more or less simultaneously with other customers, or even without having bought the product in the same website. Indeed, even though we believe that the phenomenon of false reviews needs to be carefully investigated, we would be reluctant to admit that a great majority of reviews online are fake. However, the effectiveness of the cues is discussed in Sect. 5.1.1 and in Sect. 7.5, where we compare the rate of correspondence between silver standard and the gold standard.

### 7.3 Learning from crowds (LFC)

As pointed out by Carpenter (2008); Dawid and Skene (1979); Raykar et al. (2010); Whitehill et al. (2009), among others, the majority voting assumption is that the annotators are equally reliable: but this is never the case in real life. Therefore, the output of the majority voting may be affected by unevaluated biases. To address this problem, Raykar et al. (2010) developed the Learning From Crowds (LFC) algorithm, a maximum-likelihood estimator that *jointly* learns the classifier (or regressor), the annotators' accuracy, and the actual true label.

For ease of exposition, Raykar et al. (2010) use as classifier the logistic regression, even though the algorithm would work in the same way 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

**Table 7** The distribution of deception cues in DeRev

| | Clues | Reviews | Tot. | % |
|---|---|---|---|---|
| False | 3 | 1308 | | |
| | 2 | 3248 | 4556 | 67.41 |
| True | 1 | 1820 | | |
| | 0 | 383 | 2203 | 32.59 |
| Total | | | 6759 | |

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 $\gamma$, the class $y = 1$ if $\sigma(w^\top x) \geq \gamma$.

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 $\alpha^j$ is the rate of positive cases identified by the annotator—i.e., the recall of positive cases—while the specificity $\beta^j$ is the annotator's recall of negative cases. In formal terms:

$$\alpha^j = p[y^j = 1|y = 1]. \tag{2}$$

Instead the specificity $\beta^j$ (annotator $j$ recall of negative cases) is:

$$\beta^j = p[y^j = 0|y = 0]. \tag{3}$$

Given a data set $D$ 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:

$$p[D|\theta] = \prod_{i=1}^{N} p[y_i^1, \ldots y_i^R|x_i, \theta], \tag{4}$$

and the maximum-likelihood estimator is obtained by maximizing the log-likelihood, that is:

$$\widehat{\theta}_{ML} = \{\widehat{\alpha}, \widehat{\beta}, \widehat{w}\} = arg\max_{\theta}\{ln\, p[D|\theta]\}. \tag{5}$$

In particular, given:

$$p_i = \sigma(w^\top x), \tag{6}$$

$$a_i = \prod_{j=1}^{R} [\alpha^j]^{y_i^j} [1 - \alpha^j]^{1-y_i^j}, \tag{7}$$

$$b_i = \prod_{j=1}^{R} [\beta^j]^{1-y_i^j} [1-\beta^j]^{y_i^j}, \tag{8}$$

the likelihood can be rewritten as:

$$ln\, p[D, \mathbf{y}|\theta] = \sum_{i=1}^{N} y_i \, ln\, p_i a_i + (1-y_i)\, ln(1-p_i) b_i. \tag{9}$$

Raykar et al. (2010) propose to solve this maximization problem (Bickel and Doksum 2015) through the technique of Expectation Maximization (EM) (Dempster et al. 1977). The EM algorithm can be used to recover the parameters of 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 defined above, according to the formula:

$$\mu_i = \frac{a_i p_i}{a_i p_i + b_i (1-p_i)}. \tag{10}$$

In the M-step , given the current estimate of $\mu_i$, the parameters $\theta$ are updated by maximizing the conditional expectation. In particular, equating the gradient of the said formula to zero, $\alpha$ and $\beta$ become:

$$\alpha^j = \frac{\sum_{i=1}^{N} \mu_i y_i^j}{\sum_{i=1}^{N} \mu_i}, \tag{11}$$

$$\beta^j = \frac{\sum_{i=1}^{N} (1-\mu_i)(1-y_i^j)}{\sum_{i=1}^{N} (1-\mu_i)}. \tag{12}$$

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, formalized as:

$$w_i^{t+1} = w_i^t - \eta \mathbf{H}^{-1} \mathbf{g}, \tag{13}$$

where $\eta$ is the step length, $\mathbf{g}$ is the gradient vector:

$$\mathbf{g} = \sum_{i=1}^{N} \left[\mu_i - \sigma(w^\top x)\right] x_i \tag{14}$$

and $\mathbf{H}$ is the Hessian matrix given by:

$$\mathbf{H} = -\sum_{i=1}^{N} \left[\sigma(w^\top x)\right] \left[1 - \sigma(w^\top x)\right] x_i x_i^\top. \tag{15}$$

In order to apply Raykar's algorithm, we proceeded as follows. First, we created a data set which was not divided in folds: the feature selection involved the whole DeRev. Moreover, no *a priori* knowledge of the classes was assumed during the feature selection process: we simply selected the most frequent uni-grams, bi-grams and tri-grams of lemmas and Part-Of-Speech (POS), according the scheme shown in Table 8, for a total amount of 150 surface features. With the same criterion of the highest frequency, we also collected 40 lexical features provided by the LIWC (Sect. 2.4.2), and finally we considered the token count of each review, with and without punctuation. In the end, we created a data set constituted by 192 features, plus the 3 cues of deception.

Then, we implemented the algorithm proposed by Raykar et al. (2010) in R.[19] As discussed above, the first step is the computation of the Logistic Regression (Gelman and Hill 2007) on the data set, followed by the estimation of the parameters $\alpha$ and $\beta$ for each annotator, which are iteratively updated during the Expectation–Maximization process, until convergence.

We performed these tasks in two different conditions, that is we estimated the initial values of $w$, the weights' vector for the features, determining the classes in the first execution of the logistic regression through:

1. Majority voting, according to the scheme of Table 7;
2. Random classes. In this case, we generated the classes assuming an amount of false reviews equal to 36.88%. This rate turned out from the majority voting, computed considering both the 3 behavioral cues of deception and the fact that the book addressed by the reviews was suspect or innocent: Suspect Book vs. Innocent Book. In that case, we concretely used four cues of deception, and the reviews were considered as false if characterized by the presence of three or four indicators.

   This choice reflects the purpose of exploiting our best inference about the possible amount of false reviews online. Regardless to the fact that it would be a worrying phenomenon, if more than one third of the reviews online was really deceptive, we are aware that this may not be the most friendly choice for the performance of our statistical models. In fact, since in our Gold Standard we set the distribution of truthful and false reviews as 50%, we could have computed the labels for the training of our models, starting from the same *a priori* probability $p = 0.5$. However, we preferred to use the most plausible distribution for the whole corpus, which is built as a valid sample of the reviews online, rather than for the gold standard, which is a set of instances where the classes' distribution is artificially determined.

We estimated the parameters $\alpha$ and $\beta$ considering the 3 behavioral cues of deception.

The likelihood maximization process, however, implied other not rigidly determined steps. In our case, we had to find a suitable number $n$ of iterations of the EM, and $\eta$, that is the step length for the update of the weight vector $w$. We

[19] www.r-project.org.

**Table 8** Amount of surface features for LFC

| | Uni-grams | Bi-grams | Tri-grams |
|---|---|---|---|
| Lemmas | 60 | 30 | 10 |
| POS | 30 | 15 | 5 |

found a good convergence adopting few iterations and small steps, that is $n = 10$ and $\eta = 0.001$.

### 7.4 Generative model of labels, abilities, and difficulties

The Generative model of Labels, Abilities, and Difficulties (GLAD), proposed by Whitehill et al. (2009), tackles the labeling problem considering, as suggested by the name itself, three factors:

- The real label $y$ of the entity $i$, that is $y_i \in Y$ where $Y = \{0, 1\}$;
- The expertise $\alpha$ of the annotator $i$, with $\alpha_j \in \{-\infty, +\infty\}$. In particular, $\alpha_j = +\infty$ when the labeler always detects the real class, $\alpha_j = -\infty$ when the labeler never assigns the correct label (in this case he is considered *adversarial*), and lastly $\alpha_j = 0$ if the annotator gives completely random answers, that is he has no discrimination power.
- The intrinsic difficulty $\beta$ of the entity $i$ to be labeled, so that $1/\beta_i \in \{0, \infty\}$. Imposing that $\beta$ is a positive number, $1/\beta_i = \infty$ represents the most highly ambiguous entity, which has 50% of probabilities to be well annotated, even by the most competent annotator, and $1/\beta_i = 0$ means the class of the entity is so clear, that even the most incompetent annotator identifies it.

Once defined $\alpha$ and $\beta$, according to the model of Whitehill et al. (2009), the labels $L$ given by the labeler $j$ to the entity $i$ are modeled as sigmoid function of the parameters described above:

$$p(L_{ji} = Y_i | \alpha_j, \beta_i) = \frac{1}{1 + e^{-\alpha_j \beta_i}} \tag{16}$$

Given a known prior distribution, Eq. 16 determines probabilistic labels, which are employed for the task of GLAD, which is to learn simultaneously the most likely values of $Y$, $\alpha$ and $\beta$. Similarly to LFC, the Maximum Likelihood algorithm employed is the Expectation–Maximization described above.

Summarizing, the algorithm of Whitehill et al. (2009) requires the previous determination of the distributions of the three parameters $Y$, $\alpha$ and $\beta$, that is the true labels, the expertise of the labelers and the difficulty of the entity, respectively. In our implementation, regarding the true labels we proceeded exactly the same way it was done with LFC, that is we assumed that 36.88% of the reviews in DeRev are false. As far as $\alpha$ and $\beta$ are concerned, we did not change their default values, that is mean and standard deviation equal to 1.

### 7.5 Silver standard's performance

Table 9 shows the rate of correspondence between our probabilistic and the gold standard labels, that is the accuracy of the firsts with respect to the seconds. We distinguish four cases:

1. Majority voting;
2. LFC starting the iterations from:
   (a) Majority voting;
   (b) Random classes.
3. GLAD.

Compared to the gold standard, except 2b), the accuracy is about 50%. By contrast in LFC, if in the first iteration we use random classes, the performance raises up to 69.01%.

The different performance of the case 2b) may be explained considering the rate of false cases that appear in each silver standard: in general, the use of the three cues of deception cause the creation of silver standards characterized by a high amount of positive cases—that is of false reviews—, around 70% in every algorithm. However, the iteration from random classes, where the rate of false reviews predicted was only 30.08%, leads in LFC to an annotation with a rate of positive cases around 30%. However, the distance from the distribution of the gold standard—50% of positive cases—is greater in 2b) rather than in 2a) and 1). Nonetheless, as we already supposed, to assume some too high amount of false reviews is probably wrong: therefore to perform the algorithms starting from a lower *a priori* probability of finding false reviews may have led to an annotation closer to the gold standard.

## 8 Experiment 5: training models with probabilistically annotated instances

In this last experiment, we trained models using the full DeRev (excluding Crowd-DeRev, but including the reviews not in the gold), using for training the probabilistic labels produced one of the methods described in the previous section, and testing these models on the reviews belonging to the DeRev 2018 gold standard (using tenfold cross validation to make sure the models are always tested against a fold not used for training).

This allows to compare the effectiveness of unsupervised algorithms with that of crowdsourcing for the creation of a training set. We use as baseline the performance of the heuristic labels with respect to the same gold standard. Table 10 shows the results for each annotation algorithm, and the following section discusses this outcome in comparison with the previous ones.

**Table 9** Amount of surface features for LFC

| Algorithm | First iteration | Rate of false reviews (%) | Accuracy against the gold standard (%) |
|---|---|---|---|
| MV | None | 67.41 | 52.58 |
| LFC-MV | Majority voting | 76.15 | 52.19 |
| LFC-Random | Random classes | 30.08 | **69.01** |
| GLAD | Random classes | 90.06 | 45.10 |

Bold indicates the highest value of the column

## 9 Discussion

One of the main goals of this paper was to evaluate the common practice of creating artificial data sets for the detection of fake reviews through crowdsourcing. This method is widely used in the literature, usually reporting successful results. Our results however suggest that this approach could be problematic.

First, the results of Experiments 2 and 4 (Sect. 6.2.3) show that the 'crowdsourced fake' reviews are intrinsically different from the genuinely fake reviews published on line, even when they are about exactly the same books.

Moreover, according to the degree of domain similarity between training and test set, we see a huge difference in the accuracy of the predictions. In Experiments 2 and 3 (Sects. 6.2.1 and 6.2.2), the accuracy ranges from 75.77% to 50.84%. While the first is a good performance, even comparing with the heuristic baseline of 69% (Table 9), the second drops to chance level. Such variation does not depend on the deceptiveness of the reviews, which is guaranteed in both cases, but on the degree of content similarity between training and test set.

This means that the success of the models trained in that way does not simply rely on the stylistic difference between the texts—which is supposed to convey information about their truthfulness—but also, to a non-insignificant extent, on the domain similarity between training and test set. This evidence has worrying implications for practical applications where, in principle, it is not possible to know in advance which kind of texts are deceptive. Therefore, even in case of success in the classification task, it should be clarified to which degree the models identify the deception, rather than content-related differences.

In order to avoid domain dependence problems, and any other bias possibly conveyed by the artificial production of deceptive texts, we examined a number of algorithms for probabilistic labeling, which rely on behavioral cues of deception and are agnostic with respect to the texts' content. We were also interested in verifying if the models trained with such labels, which use linguistic information for the predictions were able to overcome the performance of the same probabilistic labels, when compared to our gold standard.

The results in Table 10 shows that the models trained on labels relying on majority voting (MV and LFC-MV) struggle to improve the performance reached by the probabilistic labels themselves. Instead LFC-Random (iterated from random classes) and GLAD outperform their labels on all the metrics. But while the overall models' performance with GLAD labels is poor, with LFC-Random the accuracy of

**Table 10** Experiment 5

| Experimental design | | | | |
|---|---|---|---|---|
| Training set | DeRev with probabilistic labels | | | |
| Test set | DeRev 2018 gold standard | | | |
| Features | 147 linguistic, 3 behavioral | | | |
| Cross-validation | Tenfold | | | |
| **Performance** | | | | |
| | Accuracy (%) | Precision (%) | Recall (%) | F-measure (%) |
| MV | | | | |
| Model | 51.22 | 50.85 | **73.07** | **59.97** |
| Baseline | **52.58** | **51.89** | 70.62 | 59.83 |
| LFC-MV | | | | |
| Model | 51.74 | 51.13 | **78.74** | 62.00 |
| Baseline | **52.19** | **51.44** | 78.09 | **62.03** |
| LFC-Random | | | | |
| Model | **72.16** | **81.39** | **57.47** | **67.37** |
| Baseline | 69.01 | 77.37 | 53.74 | 63.43 |
| GLAD | | | | |
| Model | **45.43** | **47.39** | **82.99** | **60.33** |
| Baseline | 45.10 | 47.16 | 81.31 | 59.70 |

Bold indicates the highest value of the column

72.16% is not so far from that of the Experiment 2 (Sect. 6.2.1): in that experiment, the false reviews are crowdsourced, but the accuracy at 75.77% is probably affected by content related biases. In general, the outcomes suggests that, if the probabilistic annotation is reliable, then the following models can lead to results useful in terms of performance and immune to the biases pointed out in the previous experimental conditions.

Concerning the methods for the annotation, it turns out that majority voting, which does not account for the reliability of the voters (be they human coders or cues of any kind), is not effective. This is not the case for LFC and GLAD. However, while both method require the random initialization of the labels, GLAD also relies on an assumption about the distribution parameters of the dependent variable: the need to tune such parameters could explain the different performance between LFC and GLAD.

## 10 Linguistic and error analysis

In order to carry out our analysis, we wondered if we should have taken into consideration some other variables, such as the positive or negative feedback contained in the reviews, and their own titles. The measure of the reviews'

positiveness is directly quantified by the number of stars that the reviewers attribute to the product they consider. This information is absent in CROWD-DEREV, but we collected this information from the reviews published on line. The Table 11 shows the stars' distribution in the DEREV 2018 gold standard. The false reviews have always very positive content, and only the 6.7% of the truthful ones receive negative reviews (one or two stars). Given the absence of negative reviews among the false ones, to find them could be a good indicator of truthfulness. However, such indicator appears in a small amount of cases. Also, we have to consider the possibility of finding negative fake reviews, which address the topic of the activities carried out against possible competitors, rather than to the direct advantage of the products. Since we were not able to estimate this kind of phenomenon, we preferred to not use this kind of information. For the same reason, we did not perform any evaluation of the reviews' sentiment, as it could be misleading in a realistic scenario.

We also examined the review titles, in order to verify their possible use for our task. We found out that most titles just repeat the book's title, or give a short positive (or rarely negative) comment. They did not seem to add any information not included in the body of the reviews, and also about the 9% of titles in DEREV are simply repeated. Therefore we did not use the titles in our analyses. Table 12 shows the most frequent titles in DEREV.

Table 13 is meant to provide some linguistic intuition about the most frequent expressions found in our corpora. More precisely, we show the most frequent tri-grams of lemmas (without punctuation), separately for the truthful and for the fake reviews of the DEREV 2018 gold standard, and for the mirroring-truthful and mirroring-fake reviews, in the case of CROWD-DEREV. There is clearly an overlap between the most frequent expressions in the 4 sets, but noticeably in the truthful reviews (first column) there is rare mention of title and author of the reviewed books, which in contrast appear in all other reviews categories. As expected, the single *n*-grams do not look particularly informative in themselves: the models mostly rely on the consideration of features which are weak signals, singularly taken. This also confirm how difficult is the task, and how it needs to be cleaned from possible biases.

However, we tried to clarify the linguistic differences between truthful and deceptive reviews, also carrying out an error analysis concerning the Experiment 5, in particular that where the model was trained with LFC-Random labels. Table 14 contains the most frequent lemmas collected from the false positive and false negative predictions of falsity. Due to the low frequency of bi-grams and tri-grams, for the error analysis we focused on single lemmas. Also in this case, the most frequent content words address the books title or author. This underlines the models' difficulty in evaluating words which, obviously, are not exclusive for deceptive or truthful texts: an example is the name 'John', and the content words which appear in titles of books that received a high number of fake reviews (e.g. 'box', 'lethal').

## 11 Conclusion

The outcomes of this paper may be summarized by the following points.

- To use machine learning methods to identify deceptive reviews of products published on e-commerce sites is not easy, because the ground truth about these reviews is unknown and unknowable. We were nevertheless able to create a reliable gold standard, of a sufficient size to support a complete experimental study of review classification. The models trained on such gold standard achieved an accuracy of over 93%, which could be considered an upper bound, reachable in ideal conditions.
- We trained a classifier discriminating n between two kinds of fake reviews: those published online and those created with crowdsourcing services. It turned out that these two categories of texts are substantially different from each other. This megenuine ones.
  genuine ones.
  genuine ones.
  genuine ones.
  ans that they are not interchangeable and the latter should not be considered an entirely reliable replacement for genuine ones.
- We applied the method, frequently proposed in literature, consisting in the creation of training sets constituted by truthful online and fake artificial reviews; furthermore we were able to test the performance against a real gold standard. Confirming the previous literature, we found that the results may be good.
  However, we also found a strong bias, usually overlooked, which depends on the degree of content coherence between the training and test sets. This implies that the accuracy of the predictions is somewhat affected by the *a priori* knowledge or the beliefs which lead the creation of the artificial texts. Since, in principle, it is not possible to know which kind of content will characterize the texts that will be classified, the creation of artificial data sets might suffer some uncontrolled bias, which will alter dangerously the evaluation of the real effectiveness of the models.
- Lastly, we examined Bayesian methods for the automatic annotation of unlabeled reviews. We tested a number of methods and we found that:
  - Compared to the gold standard, unsupervised Bayesian methods are proven effective for the annotation of texts whose class was unknown. In particular, the LFC algorithm reached a degree of correspondence with the gold standard of 69.01%: that is well above the chance threshold, which would be 54.70%, according to our Monte Carlo simulation.
  - Once annotated the reviews, it makes sense to use such classes for the training of further models relying on linguistic and/or behavioral features. In fact, even though the classes employed for the training are noisy, that is, considering the previous outcome, affected by a rate of error of about 30% on the gold standard, they allowed the training of models whose performance was higher than that of the LFC.
    With respect to this point, we were particularly careful to avoid any possible source of overfitting. In particular, we did not employ more than 150 features, which should be a prudential amount, given that our test set was constituted by 1552 reviews. This means that the size of our feature set was lower than the 10% of data set.

**Table 11** Stars for reviews in DeRev 2018 gold standard

| Stars | Truthful reviews | Fake reviews |
|---|---|---|
| ***** | 486 (62.63%) | 775 (99.87%) |
| *** | 173 (22.29%) | 1 (0.13%) |
| ** | 65 (8.38%) | 0 |
| * | 23 (2.96%) | 0 |
| | 29 (3.74%) | 0 |

**Table 12** The most frequent reviews' titles in DeRev

| Title | Frequency |
|---|---|
| 'Box' | 20 |
| 'Call of the Wild' | 20 |
| 'The Call of the Wild' | 19 |
| 'Great book' | 13 |
| 'A Dangerous Fortune' | 13 |
| 'Great Book' | 12 |
| 'Great Read!' | 11 |
| 'Death of a Serpent' | 10 |
| 'The Neverending Story' | 10 |
| 'The Mormon Candidate' | 10 |
| 'Awesome!' | 9 |
| 'Interesting' | 9 |
| 'Bad Doctor' | 9 |
| 'Disappointing' | 8 |
| 'Great' | 8 |
| 'Christmas for Joshua' | 8 |
| 'Loved it!' | 8 |
| 'Infinite Exposure' | 7 |
| 'Great Read' | 7 |
| 'Of Human Bondage' | 7 |
| 'Saving Rachel' | 7 |

Moreover, even though the use of the behavioral cues of deception was fair, we trained models with and without them. In both cases, the outcome was higher than the LFC thresholds even though, as expected, the behavioral cues improved the performance of about 2 percent points. Therefore, even simple linguistic features can give additional information which can be exploited in order to go beyond the accuracy of the unsupervised methods alone.

– A limitation of our finding is given by the necessity of applying methods which take into account the reliability of the cues employed for the annotation (or of the coders, in case of handcrafted annotations). Majority voting, which does not evaluate the coders' skills, turns out to be not

**Table 13** The most frequent reviews' titles in DeRev

| DeRev 2018 gold standard | | | | Crowd-DeRev | | | |
|---|---|---|---|---|---|---|---|
| True | Freq. | False | Freq. | Mirroring true | Freq. | Mirroring false | Freq. |
| This book be | 164 | This book be | 125 | It be a | 199 | This book be | 198 |
| It be a | 127 | John <unk> 's | 87 | This book be | 168 | It be a | 175 |
| One of the | 112 | I have read | 86 | The book be | 168 | The book be | 141 |
| Read this book | 108 | It be a | 85 | One of the | 164 | One of the | 116 |
| The book be | 101 | This be a | 72 | Yes yes yes | 140 | <unk> <unk> <unk> | 116 |
| Of the book | 96 | I do n't | 67 | Of the book | 113 | Of the book | 115 |
| I don't | 82 | John <unk> be | 64 | Be one of | 110 | Read this book | 114 |
| Be one of | 81 | Read this book | 63 | Read this book | 99 | A lot of | 100 |
| This be a | 79 | Twist and turn | 58 | This be a | 98 | Be one of | 94 |
| Call of the | 77 | Be go to | 54 | A lot of | 76 | This be a | 78 |
| Of the wild | 75 | In this book | 52 | Be a book | 71 | I have read | 77 |
| I have read | 67 | By john <unk> | 50 | Of the wild | 68 | In this book | 76 |
| The call of | 58 | Dr . box | 50 | The story be | 67 | In 5 month | 73 |
| <unk> and <unk> | 54 | Put it down | 50 | Call of the | 66 | John <unk> be | 72 |
| This be the | 50 | This be the | 48 | Read the book | 65 | I want to | 69 |
| Of the <unk> | 50 | A lot of | 47 | <unk> <unk> <unk> | 65 | Recommend this book | 62 |
| But it be | 50 | If you be | 47 | The story of | 63 | How i sell | 61 |
| A lot of | 49 | The book be | 46 | Be a very | 62 | You want to | 59 |
| Be able to | 48 | John <unk> have | 46 | Of the good | 59 | I sell 1 | 57 |
| In the book | 48 | Be a great | 46 | Be a great | 59 | Be a very | 57 |
| Of the story | 44 | Mr . Locke | 45 | Book be a | 58 | Sell 1 million | 55 |
| The story be | 43 | You do n't | 42 | The call of | 58 | Be a great | 55 |
| Part of the | 42 | And i be | 41 | In this book | 57 | Of the good | 54 |
| There be a | 42 | Can't wait | 41 | I have read | 54 | Book be a | 54 |
| Some of the | 40 | One of the | 40 | Of this book | 51 | The story be | 54 |

effective for reliable annotation. This finding could be relevant as majority voting, thanks to its simplicity, is still a popular algorihtm for annotation in case of multiple coders.

– Our guess is that the annotated data sets released by Yelp and Amazon might have been created following, at least in part, a methodological approach similar to that applied in this study. In fact in those cases the problem is exactly the same: to label reviews whose class is unknown. More importantly, compared to the mixed, artificial training sets examined above, the models trained with probabilistic labels could be less impressive, but more reliable, and closer to the performance which could be expected in real applications.

**Table 14** The most frequent lemmas from false positive and false negative predictions of falsity, from Experiment 5—LFC-Random labels

| False positive | | False negative | |
|---|---|---|---|
| Lemma | Freq. | Lemma | Freq. |
| By | 30 | John | 300 |
| @Card@ | 29 | Locke | 118 |
| At | 23 | Creed | 117 |
| Man | 21 | @Card@ | 112 |
| CLassic | 14 | Your | 99 |
| John | 14 | At | 72 |
| Your | 13 | By | 62 |
| But | 12 | Box | 53 |
| Ever | 12 | Year | 39 |
| Year | 11 | But | 34 |
| Movie | 11 | Ever | 31 |
| As | 10 | Twist | 31 |
| Time | 7 | Man | 23 |
| War | 7 | Lethal | 23 |
| Indian | 7 | Most | 22 |
| Young | 6 | Work | 19 |
| History | 5 | As | 17 |
| Brown | 5 | Only | 11 |
| American | 5 | Than | 10 |
| Its | 5 | Still | 10 |
| Still | 5 | Its | 9 |
| Most | 5 | Time | 8 |
| Version | 4 | Version | 6 |
| Century | 4 | Movie | 6 |
| Only | 4 | Classic | 4 |

- Therefore, in absence of a gold standard, the best practice that we would recommend is as follows:
  - to use crowdsourcing services only in absence of any heuristic cue of deception. In fact, the crowsourced texts do not reproduce the false ones published on line and therefore the classifier's performance on the field is unpredictable;
  - to rely on the cues of deception if they are avialble, and to determine the silver standard through the application of algorithms which jointly evaluate the probabilistic labels and the cues effectiveness.

**Acknowledgements** Leticia Cagnina thanks CONICET for the continued financial support. This work was funded by MINECO/FEDER (Grant No. SomEMBED TIN2015-71147-C2-1-P). The work of Paolo Rosso was partially funded by the MISMIS-FAKEnHATE Spanish MICINN research project (PGC2018-096212-B-C31). Massimo Poesio was in part supported by the UK Economic and Social Research Council (Grant Number ES/M010236/1).

## References


Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54–61.

Banerjee, S., & Chua, A. Y. (2014). *Applauses in hotel reviews: Genuine or deceptive?* In: Science and Information Conference (SAI), 2014 (pp. 938–942). New York: IEEE.

Bhargava, R., Baoni, A., & Sharma, Y. (2018). Composite sequential modeling for identifying fake reviews. *Journal of Intelligent Systems,*. https://doi.org/10.1515/jisys-2017-0501.

Bickel, P. J., & Doksum, K. A. (2015). *Mathematical statistics: Basic ideas and selected topics* (2nd ed., Vol. 1). Boca Raton: Chapman and Hall/CRC Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In: *Proceedings of the eleventh annual conference on computational learning theory* (pp. 92–100). New York: ACM.

Cagnina, L. C., & Rosso, P. (2017). Detecting deceptive opinions: Intra and cross-domain classification using an efficient representation. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *25*(Suppl. 2), 151–174. https://doi.org/10.1142/S0218488517400165.

Cardoso, E. F., Silva, R. M., & Almeida, T. A. (2018). Towards automatic filtering of fake reviews. *Neurocomputing*, *309*, 106–116. https://doi.org/10.1016/j.neucom.2018.04.074.

Carpenter, B. (2008). *Multilevel bayesian models of categorical data annotation.* Retrieved from http://lingpipe.files.wordpress.com/2008/11/carp-bayesian-multilevel-annotation.pdf.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Costa, P. T., & MacCrae, R. R. (1992). Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual. Psychological Assessment Resources.

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, *28*(1), 20–28.

Dempster, A. P., Laird, N. M., & 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.

Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In: *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 213–220). New York: ACM.

Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Exploiting burstiness in reviews for review spammer detection. In: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (Vol. 13, pp. 175–184).

Feng, S., Banerjee, R., & Choi, Y. (2012). Syntactic stylometry for deception detection. In: *Proceedings of the 50th annual meeting of the association for computational linguistics* (Vol. 2: Short Papers, pp. 171–175). Jeju Island: Association for Computational Linguistics.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Fornaciari, T., & Poesio, M. (2013). Automatic deception detection in Italian court cases. *Artificial intelligence and law*, *21*(3), 303–340. https://doi.org/10.1007/s10506-013-9140-4.

Fornaciari, T., & Poesio, M. (2014). Identifying fake amazon reviews as learning from crowds. In: *Proceedings of the 14th conference of the European chapter of the Association for Computational Linguistics* (pp. 279–287). Gothenburg: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/E14-1030.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.*, Analytical methods for social research Cambridge: Cambridge University Press.

Graves, A., Jaitly, N., & Mohamed, A. R. (2013). Hybrid speech recognition with deep bidirectional LSTM. In: *2013 IEEE workshop on automatic speech recognition and understanding (ASRU)* (pp. 273–278). New York: IEEE.

Hernández-Castañeda, Á., & Calvo, H. (2017). Deceptive text detection using continuous semantic space models. *Intelligent Data Analysis*, *21*(3), 679–695.

Hernández Fusilier, D., Guzmán, R., Móntes y Gomez, M., & Rosso, P. (2013). Using pu-learning to detect deceptive opinion spam. In: *Proc. of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 38–45).

Hernández Fusilier, D., Montes-y Gómez, M., Rosso, P., & Cabrera, R. G. (2015). Detecting positive and negative deceptive opinions using pu-learning. *Information Processing & Management*, *51*(4), 433–443.

Hovy, D. (2016). The enemy in your own camp: How well can we detect statistically-generated fake reviews–an adversarial study. In: *The 54th annual meeting of the association for computational linguistics* (p 351).

Jelinek, F., Lafferty, J. D., & Mercer, R. L. (1992). Basic methods of probabilistic context free grammars. *Speech recognition and understanding* (pp. 345–360). New York: Springer.

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In: *Proceedings of the 2008 international conference on web search and data mining* (pp. 219–230). New York: ACM.

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support vector machines in R. *Journal of Statistical Software*, *15*(9), 1–28.

Kim, S., Lee, S., Park, D., & Kang, J. (2017). Constructing and evaluating a novel crowdsourcing-based paraphrased opinion spam dataset. In: *Proceedings of the 26th international conference on world wide web* (pp. 827–836). Geneva: International World Wide Web Conferences Steering Committee.

Li, F., Huang, M., Yang, Y., & Zhu, X. (2011). Learning to identify review spam. *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, *22*(3), 2488–2493.

Li, H., Chen, Z., Liu, B., Wei, X., & Shao, J. (2014a). Spotting fake reviews via collective positive-unlabeled learning. In: *2014 IEEE international conference on data mining (ICDM)* (pp. 899–904). New York: IEEE.

Li, H., Fei, G., Wang, S., Liu, B., Shao, W., Mukherjee, A., & Shao, J. (2017). Bimodal distribution and co-bursting in review spam detection. In: *Proceedings of the 26th international conference on world wide web* (pp. 1063–1072). Geneva: International World Wide Web Conferences Steering Committee.

Li, H., Liu, B., Mukherjee, A., & Shao, J. (2014b). Spotting fake reviews using positive-unlabeled learning. *Computación y Sistemas*, *18*(3), 467–475.

Li, J., Ott, M., Cardie, C., & Hovy, E. H. (2014c). *Towards a general rule for identifying deceptive opinion spam*. In: ACL (Vol. 1, pp. 1566–1576).

Lin, C. H., Hsu, P. Y., Cheng, M. S., Lei, H. T., & Hsu, M. C. (2017). Identifying deceptive review comments with rumor and lie theories. In: *International conference in swarm intelligence* (pp. 412–420). New York: Springer.

Liu, B., Dai, Y., Li, X., Lee, W. S., & Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. In: *Third IEEE international conference on data mining* (pp. 179–186). New York: IEEE.

Liu, B., Lee, W. S., Yu, P. S., & Li, X. (2002). Partially supervised classification of text documents. *ICML*, *2*, 387–394.

Martens, D., & Maalej, W. (2019). Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering*,. https://doi.org/10.1007/s10664-019-09706-9.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781.

Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., & Ghosh, R. (2013a). Spotting opinion spammers using behavioral footprints. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 632–640) New York: ACM.

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. S. (2013b). What yelp fake review filter might be doing? In: *Proceedings of the seventh international AAAI conference on weblogs and social media.*

Negri, M., Bentivogli, L., Mehdad, Y., Giampiccolo, D., & Marchetti, A. (2011). Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In: *Proceedings of the conference on empirical methods in natural language processing* (pp. 670–679). Stroudsburg: Association for Computational Linguistics.

Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative deceptive opinion spam. In: *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 497–501).

Ott, M., Choi, Y., Cardie, C., & Hancock, J. (2011). Finding deceptive opinion spam by any stretch of the imagination. In: *Proceedings of the 49th Annual meeting of the association for computational linguistics: human language technologies* (pp. 309–319). Portland, Oregon: Association for Computational Linguistics.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count (LIWC): LIWC2001*. Mahwah: Lawrence Erlbaum Associates.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

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

Ren, Y., & Ji, D. (2017). Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, *385*, 213–224.

Rout, J. K., Dalmia, A., Choo, K. K. R., Bakshi, S., & Jena, S. K. (2017). Revisiting semi-supervised learning for online deceptive review detection. *IEEE Access*, *5*(1), 1319–1327.

Saini, M., & Sharan, A. (2017). Ensemble learning to find deceptive reviews using personality traits and reviews specific features. *Journal of Digital Information Management*, *12*(2), 84–94.

Salloum, W., Edwards, E., Ghaffarzadegan, S., Suendermann-Oeft, D., & Miller, M. (2017). Crowdsourced continuous improvement of medical speech recognition. In: *The AAAI-17 workshop on crowdsourcing, deep learning, and artificial intelligence agents.*

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of international conference on new methods in language processing*. Retrieved from http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.

Shehnepoor, S., Salehi, M., Farahbakhsh, R., & Crespi, N. (2017). Netspam: A network-based spam detection framework for reviews in online social media. *IEEE Transactions on Information Forensics and Security*, *12*(7), 1585–1595.

Skeppstedt, M., Peldszus, A., & Stede, M. (2018). More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In: *Proceedings of the 5th workshop on argument mining* (pp. 155–163).

Strapparava, C., & Mihalcea, R. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. In: *Proceedings of the 47th annual meeting of the association for computational linguistics and the 4th international joint conference on natural language processing.*

Streitfeld, D. (August 25th, 2012). The best book reviews money can buy. *The New York Times*.

Whitehill, J., Wu, T., Bergsma, F., Movellan, J. R., & Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems* (pp. 2035–2043). Cambridge: MIT Press.

Xie, S., Wang, G., Lin, S., & Yu, P. S. (2012). Review spam detection via temporal pattern discovery. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp 823–831). New York: ACM.

Yang, Y., & 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* (pp. 42–49). New York: ACM.

Zhang, W., Bu, C., Yoshida, T., & Zhang, S. (2016). Cospa: A co-training approach for spam review identification with support vector machine. *Information*, *7*(1), 12.

Zhang, W., Du, Y., Yoshida, T., & Wang, Q. (2018). DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing & Management*, *54*(4), 576–592.

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