Masking domain-specific information for cross-domain deception detection

Javier Sánchez-Junquera[b,**], Luis Villaseñor-Pineda[a,d], Manuel Montes-y-Gómez[a], Paolo Rosso[b], Efstathios Stamatatos[c]

[a]*Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla 72840, Mexico*
[b]*PRHLT Research Center, Universitat Politècnica de València, Camino de Vera s/n, València 46022, Spain*
[c]*Department of Information and Communication Systems Engineering, University of the Aegean, Samos 83200, Greece*
[d]*Centre de Recherche en Linguistique Française GRAMMATICA (EA 4521), Université d'Artois, Arras 62000, France*

ABSTRACT

The facilities provided by social media and computer-mediated communication make easy the dissemination of deceptive behavior, after which different entities or people could be affected. The deception detection by supervised learning has been widely studied; however, the scenario in which there is one domain of interest and the labeled data is in another domain has received poor attention. This paper presents, to our knowledge, the first domain adaptation approach for cross-domain deception detection in texts. Our proposal consists in modifying original texts from the source and target domains in a form in which common content and style information is maintained, but domain-specific information is masked. In order to adequately select domain-specific terms to be masked, the proposed method uses unlabeled instances from both domains. Our experiments demonstrate that the masking technique is a good idea for detecting deception in cross-domain scenarios; and the performance could be further improved if unlabeled information from the target domain is considered.

1. Introduction

Over the years, human beings have found in deception a tool that provides either protection or another type of personal gain. Today, the presence of deception is becoming increasingly noticeable and harmful, e.g. due to the facilities provided by technology and the web. Deception refers to the attempt to create in another a belief which the communicator considers false (Vrij, 2000). For example, the fake service reviews that try to deliberately mislead customers; or lies that protect oneself from disapproval and manage others impressions outside boundaries of honesty (DePaulo et al., 2003). In many cases, the importance of catching liars is due to the undesirable consequences of deception in online reviews, trial hearings, predatory communication, among others (Ott et al., 2011; Pérez-Rosas et al., 2015; Rosso and Cagnina, 2017).

Text classification techniques have been extensively used to detect deception. For this approach it is necessary to acquire labeled data sets, which are traditionally constructed from manual labeling. Manual labeling is complex and expensive, especially in deception detection, due to the poor human skills as detectors and the need to design collection protocols for each domain of interest. Given this difficulty, it is essential to be able to use cross-domain solutions which employ labeled data from one domain for the classification of deception in other domain.

Previous work has shown that cross-domain approaches present a difficulty in detecting deception. The problem is that many cues to deception change from one domain to another due to the change in content, the consequences if the deceiver is getting caught lying, and the emotions experienced by the deceiver due to the topic (DePaulo et al., 2003; Vrij, 2008). For example, pronouns in one domain (e.g., essays on abortion) can be an indicator of deception while in another domain (e.g., reviews on hotels) they can describe truthful texts (Ott et al., 2011).

However, works such as (Feng et al., 2012; Pérez-Rosas and Mihalcea, 2014) have shown that both the style-related information and content words may be relevant for the detection of deception in cross-domain scenarios. These works have evaluated their proposals taking into account only characteristics of the source domain and ignoring those from the target domain. Hence, it might be possible to identify common characteristics (related to content or style) between the source and target domains, to obtain a more general representation of their texts.

[**]Corresponding author:
e-mail: `jjsjunquera@gmail.com` (Javier Sánchez-Junquera)

We propose, to our knowledge, the first domain-adaptation method for deception detection that uses information from source and target domains. This method is a contribution to both the deception detection task and the cross-domain problem. Our method is inspired by the text distortion approach successfully used in thematic text clustering and authorship attribution (Granados et al., 2011; Stamatatos, 2017b), but modified to be more suitable to deception detection and to be used as a domain adaptation approach. Its main idea is to transform original texts from the source and target domains by masking domain-specific terms. Source and target domains are observed to pick out the terms specific to only one of them. While the textual structure, the style-related information, and the common content words are maintained, the picked out domain-specific terms are masked obtaining a more general text representation. Our experiments show that the proposed method can improve the cross-domain classification between domains of online reviews or essays about controversial topics.

2. Related work

Computational works have shown important results in deception detection. First of all, such works confirm that human judges make more mistakes in detecting deception in comparison to automated methods (Ott et al., 2011). However, supervised learning studies are limited due to lack of appropriate corpora for deception detection. Furthermore, different kinds of features have been explored for the text representation in order to detect deception.

Earlier works mainly focused on single-domain scenarios for which many of them propose traditional techniques based on simple text representations. Pérez-Rosas and Mihalcea (2014) and Ott et al. (2011) demonstrated that truthful and deceptive texts are separable, through word n-grams, psycholinguistic features from LIWC, and part-of-speech features. More sophisticated features, such as deep syntactic patterns (Feng et al., 2012), argumentative features (Cocarascu and Toni, 2016), and word embeddings (Ren and Ji, 2017), were also successfully evaluated. More recently, the character n-grams features have shown a good trade-off between simplicity and performance for detecting deception in online reviews and essays on controversial topics (Cagnina and Rosso, 2017; Sánchez-Junquera et al., 2018). All these works found that both content and style are important factors to distinguish deception from truth.

There are few works that have reported results on cross-domain deception classification. They merely evaluated how the performance decreases when their models are trained on a source domain, and no information from the target domain was observed (Li et al., 2014; Ren and Ji, 2017). These works showed interest in whether a relatively richer annotated domain could be used to train effective deception detection models for other domains, and how good the generalization ability of their models was. They suggested that the performance was affected because the target domain generally encoded some type of features different to the ones found in the source domain.

When the domain of labeled examples is different from that with the instances of interest (i.e., the cross-domain scenario), the results are affected by topic differences. This problem has been addressed in other classification tasks such as sentiment analysis and authorship attribution with domain adaptation approaches. On the one hand, a common idea in sentiment analysis is to search words from each domain that share a similar connotation (Pan et al., 2010); or to separate the vocabulary into general words (i.e., domain-independent features) and specific words (i.e, domain-specific features) for a different usage of those specific words from source domain (Tan et al., 2009; Wu et al., 2010). On the other hand, in authorship attribution, Stamatatos (2017b) proposes a text distortion method which masks the occurrences of the least frequent words of the language; thus, the algorithm compresses topic information and maintains textual structure related to personal style.

3. Masking domain-specific terms for deception detection

Masking techniques have been applied to different tasks. On the one hand, Granados et al. (2011) focused on masking frequent words to enhance performance in text clustering. On the other hand, based on the opposite perspective, Stamatatos (2017a) focused on masking the least frequent words to highlight style information that is used in authorship attribution. In our case, since both content and style information could be useful for detecting deception, we want to maintain both factors depending on the common information from source and target domains. For example, considering reviews about hotels and doctors as the source and target domain respectively, we could maintain function words (e.g., *the*, *my*) and common content words between the two domains (e.g., *staff*, *family*) and mask domain-specific words (e.g., *doctor*, *hotel*).

In this section, we present our domain adaptation approach. We first use the Frequently Co-occurring Entropy (FCE) to pick out domain-specific features and then we employ a distortion method to mask them.

3.1. Domain-specific terms filtering

In this work, we consider general terms as in (Pan et al., 2010): they should occur frequently and act similarly in both the source and target domains. Subsequently, domain-specific terms are those that do not satisfy this condition. In order to achieve a trade-off between frequency and similarity of terms, we use FCE, proposed in (Tan et al., 2009). The general formula is as follows:

$$FCE_w = \log\left(\frac{P_S(w) \times P_T(w)}{|P_S(w) - P_T(w)|}\right) \tag{1}$$

where $P_S(w)$ and $P_T(w)$ are the probabilities of the term w in the source and the target domain respectively[1]. In this work, as in (Tan et al., 2009; Wu et al., 2010), we compute $P_S(w)$ and $P_T(w)$ as follows:

$$P_S(w) = \frac{N_w^S + \alpha}{N^S + 2 \times \alpha} \quad \text{and} \quad P_T(w) = \frac{N_w^T + \alpha}{N^T + 2 \times \alpha} \tag{2}$$

[1] Defined as the probability of taking an instance from the corpus with the given term. No labeled data is necessary for this task.

Table 1. Examples of FCE results in Hotel and Doctor corpora.

w	N_w^{Hotel}	N_w^{Doctor}	FCE_w	Rank
my	987	351	2.50	1
ever	171	60	0.96	2
I	1340	402	−0.41	6
needs	35	35	−7.65	405
life	34	34	−7.69	409
helped	31	28	−7.74	425
spa	64	0	−25.15	2381
consultation	0	31	−26.67	10567
tests	0	9	−26.71	10605

where N_w^S and N^S are the number of instances where w occurs at least once and the total number of instances, respectively, in the source domain; and N_w^T and N^T are the number of instances where w occurs at least once and the total number of instances, respectively, in the target domain. We set $\alpha = 0.0001$ in order to overcome overflow, which appears for infrequent terms in a large corpus. On the other hand, β is included[2] to deal with the extreme case when $P_S(w) = P_T(w)$:

$$FCE_w = \log\left(\frac{P_S(w) \times P_T(w)}{|P_S(w) - P_T(w)| + \beta}\right) \tag{3}$$

Table 1 shows a simple example taking reviews on hotels and doctors as the source and the target domains respectively (details of the used corpora are given in Table 5). We can see that *my*, *ever*, and *I* could be considered as more general terms; *needs*, *life*, and *helped* are less frequent terms and are more related to the content of both domains; however, *spa*, *consultation*, and *tests* are infrequent in at least one domain or have dissimilar occurring probability.

3.2. Text distortion methods

The main idea of the proposed method is to transform the original texts to a domain-abstract form where textual structure, related to a general style of deceivers or honest persons, is maintained while infrequent words, corresponding to domain-specific information, are masked. To this end, all the occurrences (in both training and test corpora) of domain-specific terms are replaced by symbols.

Let W_k be a set of k general terms. A text is tokenized and all $w \notin W_k$ will be masked according to a specific text distortion technique. We describe Distorted View with Multiple Asterisks (DV-MA) and Distorted View with Single Asterisks (DV-SA); two text distortion methods introduced by (Stamatatos, 2017a):

DV-MA: Every $w \notin W_k$ is masked by replacing each of its characters with an asterisk ($*$). Every digit in the text is replaced by the symbol #.

DV-SA: Every $w \notin W_k$ is masked by replacing each word occurrence with a single asterisk ($*$). Every sequence of digits in the text is replaced by a single symbol #.

We modify these methods by treating any token that includes punctuation marks in a special way. If the token is found to be domain-independent (e.g., commas and periods) then it is maintained. On the other hand, if it is found to be domain-specific (e.g., quotes, parentheses, or compound terms like *and/or*), it is replaced by the symbol @. Furthermore, to consider all the numeric details usually given by truthful communicators (Vogler and Pearl, 2018), we mask numerals (e.g., *one, two, three*, etc.) with a single symbol +.

Table 2. An example of transforming a doctor review, according to two distortion techniques, observing reviews on doctors and hotels. In these transformations $k = 400$.

My Neck/S-Lift procedure performed in March 2009 was handled in a very professional manner, and I was able to attend a social event three weeks after surgery.	
DV-MA	My @ ********* ********* in ***** #### was ******* in a very ************ ******, and I was able to ****** a ****** ***** +++++ ***** after *******.
DV-SA	My @ * * in * # was * in a very * *, and I was able to * a * * + * after *.

Table 3. An example of transforming an input text according to DV-MA using different values of k.

My Neck/S-Lift procedure performed in March 2009 was handled in a very professional manner, and I was able to attend a social event three weeks after surgery.	
k=0	** @ ********* ********* ** ***** #### *** ******* ** * **** ************ ******@ *** * *** **** ** ****** * ****** ***** +++++ ***** ***** *******@
k=400	My @ ********* ********* in ***** #### was ******* in a very ************ ******, and I was able to ****** a ****** ***** +++++ ***** after *******.
k=1000	My @ ********* ********* in ***** #### was handled in a very professional ******, and I was able to ****** a ****** ***** +++++ weeks after *******.

An example of transforming a sentence, according to these text distortion variants, is provided in Table 2. In this case, W_k includes the 400 most general terms from reviews on hotels (source domain) and reviews on doctors (target domain). Table 3 shows an example of transforming the same input text according to DV-MA algorithm using different values of k.

We can note that $k = 0$ means that every term is considered domain-specific, therefore, even punctuation marks will be masked. However, when $k = 400$, mainly function words (e.g., *My*, *in*, *a*, *The*, *I*, *after*) and some punctuation marks are maintained, because they are not associated to a particular domain. Finally, by expanding the set of general terms to $k = 1000$, content-related terms associated with both domains are also maintained (e.g., *handled*, *professional*, *weeks*). Note that the terms *Neck/S-Lift*, *2009*, and *three*, are always masked.

Table 4 shows another example of a sentence, taken from a hotel review, transformed according to DV-MA when two different target domains are considered. Observe how the set of domain-independent terms changes when the target domain concerns reviews on either doctors or restaurants. For example, *help* is a general term in reviews on both doctors and hotels, but not on restaurants; and *location* is a general term in reviews on both restaurants and hotels, but not on doctors.

[2] We take up on the values set in (Tan et al., 2009), i.e., $\alpha = \beta = 0.0001$.

Table 4. A hotel review is transformed according to DV-MA with $k = 1000$ for two different target domains. Highlighted (in yellow) the general terms depending on the target domain.

Hotel review	Target Domain	
	Doctors	Restaurants
Superb location and proximity to local attractions. Staff is always friendly and eager to help.	****** ******** *and* ********* *to* ***** ***********. *Staff is always friendly and eager to help.*	*Superb location and* ********* *to local* ***********. *Staff is always friendly and eager to* ****.

Table 5. Statistics of the datasets. The number of deceptive (D) and truthful (T) instances, the average vocabulary size (per instance), as well as the average length of instances (either characters or words) are given.

		Instances		Vocabulary		Length(ch)		Length(w)	
Type	Domain	T	D	T	D	T	D	T	D
Spam	Hotel	800	800	101	95	821	791	172	164
	Doctor	200	356	66	75	465	593	97	119
	Restaurant	200	200	97	89	762	709	160	146
Controversial	Abortion	100	100	64	50	499	359	101	73
	Best Friend	100	100	51	40	337	266	72	57
	Death Penalty	100	100	60	54	463	395	93	78

4. Experiments

4.1. Datasets

We use benchmark datasets in English that include two genres: reviews (opinion spam) and essays (controversial opinions). The former comprises three domains, namely Hotel, Restaurant, and Doctor. The latter also comprises three domains, namely Abortion, Death Penalty, and Best Friend. Table 5 shows the statistics of the six datasets.

The datasets of reviews are parts of those collected by Li et al. (2014). The truthful reviews were mined from a set of real customers and the deceptive ones were collected by crowdsourcing. For each domain, *turkers* were asked to describe a fake experience as if it had been real.

All essays were also collected using crowd-sourcing. For Abortion and Death Penalty, participants were asked to express both their personal opinion and the opposite on that topic, imagining that they were taking part in a debate. In the Best Friend domain, participants were asked to write about their best friend and describe the detailed reasons for their friendship. Subsequently, they were asked to think about a person they could not tolerate and describe her/him as if s/he was their best friend (Pérez-Rosas and Mihalcea, 2014).

4.2. Experimental setup

Preprocessing: We convert all words to lowercase letters and do not remove any character (e.g., symbol, punctuation mark, number or delimiter).

Text Representation: The proposed method uses two parameters: k indicates the top general terms which will not be masked and n is the order (length) of character n-grams that represent the masked texts. We empirically select the values of k and n by performing grid search for each pair of source-target domains: $k \in \{0, 100, 200, ..., 1000\}$ and $n \in \{3, 4, 5, 6, 7\}$. Except for Figure 4, all reported results were using $n = 4$ and the best value for k for each case. After the masking stage, we represent the transformed texts without removing any character n-gram feature and use a binary[3] weighting scheme.

Classifier: We use the Naïve Bayes (NB) classifier. Similar performance has been obtained based on Support Vector Machines, thus we only report results for NB.

Evaluation: We use 80% of the unlabeled target domain instances and all the source domain instances for picking out domain-specific terms in an unsupervised manner (information about the class, D or T, of each instance is not used). Then, we apply masking in all the texts of both training and test sets. We train the classifier using only the source domain instances and we apply the learned model to the unobserved (20%) target domain instances. In all the experiments, to avoid over-fitting, we randomly select 80% (for the masking process) and 20% (as the test set) unlabeled instances from target domain creating two disjoint subsets; we repeat this procedure 10 times ensuring that each instance was classified two times. The results reported in all experiments are average results of these 10 individual results. We use F_1 as the evaluation measure.

Baseline: Our baseline method is based on the same text representation and classifier but without applying any distortion method. It does not use any information from the target domain.

4.3. Results and discussion

Figure 1 shows the average and standard deviation of F_1 in cross-domain deception detection for all pairs of source and target domains in reviews and essays: the blue (DV-MA) and red (DV-SA) bars indicate the results of the domain adaptation by the proposed approach, and the green bars indicate results of our baseline. The Figure also shows a line chart with the F_1 results in the single-domain scenario for each target domain (using the same representation); e.g., above the bars of Dr->H and Rest->H (results of DV-MA, DV-SA, and baseline respectively), a line chart indicates that we obtained $F_1 = 0.89$ when both training and test instances come from the Hotel domain.

The two variants of masking domain-specific terms, i.e. DV-MA and DV-SA, do not show significant differences in F_1. However, DV-MA tends to perform slightly better. From Figure 1 we can note that the proposed method always improves the performance of the baseline (in average by 14%). We suppose that the cases in which the proposed approach is only slightly better than the baseline are due to the similarity between the specific source and target domains and the little descriptive power of the source domain patterns over the target domain. Despite the fact that the proposed method demonstrates the usefulness of exploiting information from both domains and masking the domain-specific information, the differences of obtaind results with respect to those of single domain cases indicate that there is a lot of space for improvement.

Surprisingly, our method achieved higher F_1 than the single-domain evaluation in the Death Penalty corpus. We guess the reason is the difficulty of finding relevant patterns in this corpus, so the information obtained from other domains improve the performance. Similar behavior with these controversial topics can be found in (Pérez-Rosas and Mihalcea, 2014).

[3] *tf* and *tf-idf* were also tested, but obtained slightly lower results.

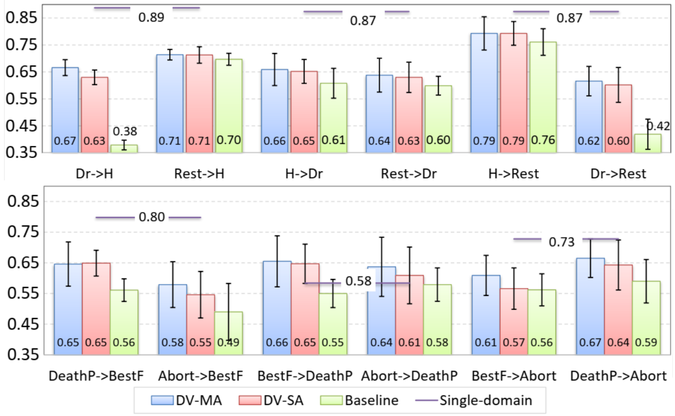

Fig. 1. Evaluation results of the proposed approach and our baseline; the three bars show the average of F_1 and the standard deviation for the cross-domain problem (e.g., Dr->H means that Doctor is the source domain and Hotel is the target domain). For each target domain, a line chart indicates the single-domain performance.

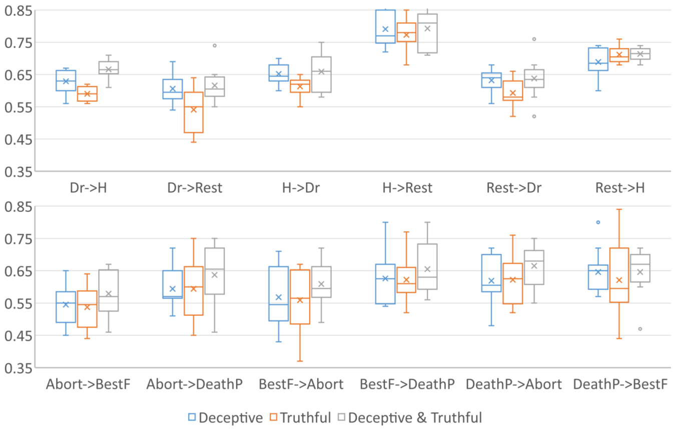

Fig. 2. Results of DV-MA when the unlabeled instances of the target domain are exclusively deceptive, truthful, or belong to any of those classes.

4.3.1. *Sensitivity to the distribution of observed instances*

In order to filter out domain-specific terms, this work uses FCE, which does not require labeled instances. Therefore, this work can assume that there are no labeled instances from the target domain. On the other hand, deceptive and truthful instances have many terms with dissimilar distribution. In this section, we try to answer the question: is the classification accuracy affected if the majority of the observed instances in the target domain belong to a certain class? To answer this question, we compare (see Figure 2) the results of evaluating the proposed approach based on DV-MA by observing deceptive instances exclusively, truthful instances exclusively, or an equal number of instances of these two classes.

The results of Figure 2 do not consistently indicate whether or not it is better that the set of observed instances of the target domain is balanced with respect to the classes. In general, it can be noted that the results are comparable. This suggests that our method is robust to the distribution of target domain instances over the classes and the selection of domain-specific terms is not affected when more deceptive/truthful instances are included in the unlabeled data.

Table 6. Average and standard deviation F_1 results using different strategies for selecting the terms to be masked.

		Baseline	**Most frequent words in English**	**FCE**
Unlabeled data from both domains				✓
Masking technique			✓	✓
Source	**Target**			
Hotel	Restaurant	0.761 ± (0.050)	**0.779** ± (0.034)	**0.793*** ± (0.062)
	Doctor	0.608 ± (0.055)	**0.645** ± (0.044)	**0.659*** ± (0.060)
Restaurant	Hotel	0.697 ± (0.022)	**0.726** ± (0.023)	0.714 ± (0.020)
	Doctor	**0.599** ± (0.035)	0.596 ± (0.055)	**0.638*** ± (0.063)
Doctor	Restaurant	0.419 ± (0.056)	**0.554** ± (0.049)	**0.616*** ± (0.055)
	Hotel	0.379 ± (0.018)	**0.540** ± (0.026)	**0.666*** ± (0.030)
Abortion	Best Friend	0.490 ± (0.093)	**0.579** ± (0.080)	0.579 ± (0.075)
	Death Penalty	0.579 ± (0.055)	**0.647** ± (0.064)	0.637 ± (0.096)
Death Penalty	Abortion	0.590 ± (0.071)	**0.640** ± (0.058)	**0.665*** ± (0.063)
	Best Friend	0.561 ± (0.037)	**0.645** ± (0.084)	**0.646*** ± (0.072)
Best Friend	Abortion	**0.562** ± (0.053)	0.544 ± (0.075)	**0.609*** ± (0.065)
	Death Penalty	0.550 ± (0.046)	**0.594** ± (0.085)	**0.655*** ± (0.083)

4.3.2. *The contribution of observing the target data*

The proposed method differs from the baseline in two aspects. First, unlabeled data from the target domain are observed; second, a masking technique is applied. In this section we try to clarify if the improvement of our method over the baseline is due to the data observed from the target domain, the masking technique, or both. To this end, we compare the performance of the masking method using DV-MA without access to target domain data and the proposed method using DV-MA with access to target domain data. The former does not depend on the target domain and masks the less frequent words of the English language[4]. The latter extracts domain-specific terms by applying FCE to the source and target domains.

Table 6 compares the results of these two methods with the baseline. The third and fourth columns are compared by printing in bold the highest value for each pair of domains and the fifth column shows, in bold and asterisk, those cases in which our method obtains the best results by observing unlabeled data from the target domain. We can see that by masking frequent words of the language our method improves the baseline (in 10 cases out of 12) by 9% in average. Although in this way domain adaptation is not actually performed since no information from the target domain is used, it can be concluded that the masking technique itself is useful to enhance performance in cross-domain deception detection. Furthermore, the results improve even more (in 9 cases out of 12) by 5% in average when information from the target domain is used and the terms to be masked are picked out accordingly. Therefore, if it is possible to observe unlabeled information from the target domain, the performance is enhanced. If, on the other hand, such information is not available, the masking technique is still useful.

4.3.3. *Presence of masks in discriminatory features*

Granados et al. (2011) concluded that in cases the textual structure was not maintained, the performance of clustering decreased. In a similar way, one may suspect that, even by transforming the original texts, the most discriminatory features are

[4] We extract the most frequent words of the BNC corpus (https://www.kilgarriff.co.uk/bnc-readme.html). For each pair of domains, we report the higher F_1 varying $k \in \{0, 100, 200, ..., 500, 1000, 2000, ..., 5000\}$ following the practice of Stamatatos (2017b).

Table 7. Features with high information gain in Hotel and Restaurant domains, with $k = 400$ and $n = 4$. The underscore symbol indicates a blank space in char n-grams. Examples of sentences where these char n-grams occur are highlighted (in yellow).

Class	4-gram	Examples of information captured from original texts	
		Hotel (source domain)	Restaurant (target domain)
True	#_**	*with 2 bathrooms*	*for lunch 2 days later*
	*...	*for the romantic couple...*	*or the salmon...*
	_$##	*only $15 per day*	*to $10 and steaks closer to $30*
	_(**	*in a (dark) corner*	*very (very) few places*
	mall	*The room was very small*	*the small plates*
Deception	_my_	*I made my reservation at*	*on my next visit*
	I _wi	*and I will choose other*	*I will be back again*
	I_wa	*I was able to relax*	*but I was pleasantly*
	anyo	*if anyone carried my bags*	*to anyone looking for*
	_rec	*I'd only recommend this*	*I would recommend this*

n-grams that do not include the masking characters (i.e., *, @, + , #). That way, it would not be necessary to maintain the domain-specific terms (either original or masked) in the representation (i.e., they might be removed). However, a deeper look in the most discriminatory character n-grams makes possible to note that there are many of them that include masking symbols.

Table 7 shows some character n-grams with high information gain for the hotel and restaurant domains as well as examples of sentences in which they occur. As it can be seen, truthful reviews on hotels or restaurants are characterized by providing numerical information, and explanatory phrases enclosed in parentheses. Thanks to the masking technique, the proposed method is not distracted by specific numbers or what was the particular clarification given. In general, it captures an abstract type of information commonly used by real customers.

4.3.4. Effect of the parameters' values

In the previous experiments, we empirically select the values of k by performing grid search for each pair of domains. Figure 3 shows boxplots with the distribution of F_1 with all pairs of review (opinion spam) domains on one hand, and essay (controversial opinion) domains on the other. In these cases, we used DV-MA with character 4-gram features and varying $k \in \{0, 100, 200, ..., 1000\}$. For the baseline, we used character 4-gram features too and the Figure shows the average of F_1 since the baseline does not depend on k.

As can be seen, the performance varies with different values of k. We conclude that it is always important to mask a set of terms ($k > 0$), possibly because two different domains have at least some terms with dissimilar distribution. At the same time, performance in general decreases with $k > 600$ for the examined corpora, indicating that terms with relatively low FCE score are actually distractful and it is better to mask them. Interestingly, with almost all pairs of domains evaluated, the proposed method improved the performance of the baseline for a wide range of values of k ($50 < k < 600$).

Similarly, Figure 4 shows boxplots with the distribution of F_1 of the proposed approach based on DV-MA with $k = 400$ and the baseline for various n-gram lengths. We can note that similar performance is obtained for all examined n values, which further proves the robustness of the proposed method.

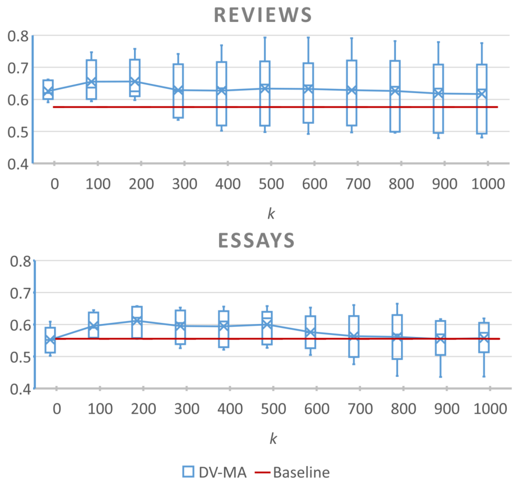

Fig. 3. F_1 **of DV-MA (varying k values) and baseline models.**

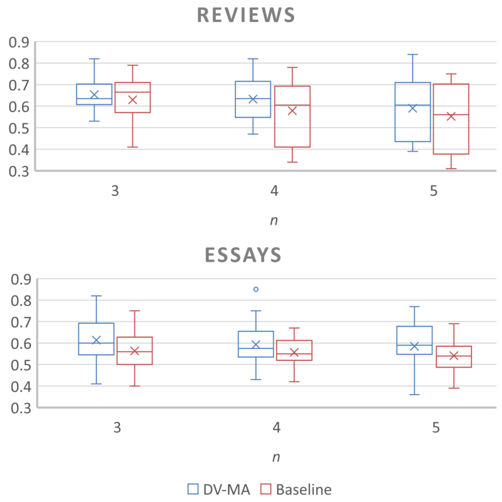

Fig. 4. F_1 **of DV-MA (varying n values) and baseline models.**

4.3.5. Comparison to other works

In previous works, there are cross-domain deception detection results reported for the reviews corpora we used. Reported results on the essays corpora refer to a different evaluation setup using two source domains (Pérez-Rosas and Mihalcea, 2014).

Table 8 shows the cross-domain deception detection results reported by (Cagnina and Rosso, 2017) with the same versions of the review datasets we used in this work. Cagnina and Rosso proposed an efficient representation for this task and trained their model using only the source domain. The proposed method is also trained using only the source domain, however, the terms to be masked are selected by observing unlabeled data from the target domain. To get a fair comparison with Cagnina and Rosso, we show in Table 8 the obtained results of our method in two cases: when no target domain information is used (the most frequent words of language are masked) and when unlabeled data from the target domain are used (based on FCE).

The third and fourth columns are compared by printing in bold the highest value for each pair of domains, and it is pos-

Table 8. Comparison of the performance of the proposed approach (either with or without access to target domain information) to the results reported by Cagnina and Rosso (2017) on review datasets.

		Cagnina and Rosso (2017)	**Most frequent words in English**	**FCE**
Unlabeled data from both domains				✓
Masking technique			✓	✓
Source	**Target**			
Hotel	Restaurant	0.64	**0.779** ± (0.034)	**0.793*** ± (0.062)
	Doctor	0.50	**0.645** ± (0.044)	**0.659*** ± (0.060)
Restaurant	Hotel	0.66	**0.726** ± (0.023)	0.714 ± (0.020)
	Doctor	0.50	**0.596** ± (0.055)	**0.638*** ± (0.063)
Doctor	Restaurant	**0.57**	0.554 ± (0.049)	**0.616*** ± (0.055)
	Hotel	0.42	**0.540** ± (0.026)	**0.666*** ± (0.030)

sible to note that in five out of six cases, the F_1 reported by Cagnina and Rosso is improved by our method when no target domain information is used (the most frequent words of the language are masked). The fifth column shows, in bold and asterisk, those cases in which observing unlabeled data from the target domain, our method obtains a higher score than indicated in the two previous columns.

Finally, it is important to point out that other cross-domain results have been reported by Li et al. (2014) and Ren and Ji (2017). However, these authors used the original versions of the three review (opinion spam) datasets, which contain more instances; therefore, their results cannot be directly compared with the ones obtained in this study[5].

5. Conclusions

This paper is a contribution to the cross-domain deception detection, a doubly challenging task due to the cross-domain problems and the difficulty at detecting deception. The proposed method improves the cross-domain classification performance in which labeled instances from the target domain are not given. The suitability of our method is due to we apply a text distortion technique that transforms original texts in a form in which distractful information is masked. We demonstrate that the masking technique is a good idea for detecting deception in cross-domain scenarios. Moreover, the performance is further improved if we consider *unlabeled* information from the target domain in order to pick out the terms to be masked. The method is robust to the distribution of the classes in the unlabeled data that is observed and to the parameter *n* (length of the n-grams used as features).

To our knowledge, this is the first domain adaptation approach that combines information from the source and target domain for a better text representation in the deception detection task. More data are needed to study more carefully how *k* depends on specific corpora characteristics.

Acknowledgments

The work of Javier Sánchez-Junquera and Paolo Rosso was funded by the MISMIS-FAKEnHATE Spanish MICINN research project (PGC2018-096212-B-C31). This work was partially supported also by CONACyT under grants CB-2015-01-257383, the Thematic Networks program, and the scholarship CONACyT-Mexico 613411.

References

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, 151–174.

Cocarascu, O., Toni, F., 2016. Detecting deceptive reviews using Argumentation. Proceedings of the 1st International Workshop on AI for Privacy and Security - PrAISe '16 , 1–8.

DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K., Cooper, H., 2003. Cues to deception. Psychological bulletin 129, 74.

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: Short Papers-Volume 2, pp. 171–175.

Granados, A., Cebrian, M., Camacho, D., de Borja Rodriguez, F., 2011. Reducing the loss of information through annealing text distortion. IEEE Transactions on Knowledge and Data Engineering 23, 1090–1102.

Li, J., Ott, M., Cardie, C., Hovy, E., 2014. Towards a general rule for identifying deceptive opinion spam, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1566–1576.

Ott, M., Choi, Y., Cardie, C., Hancock, J.T., 2011. Finding deceptive opinion spam by any stretch of the imagination, in: Proceedings 49th Annual Meeting of the Association for Computational Linguistics:HLT - Volume 1, pp. 309–319.

Pan, S.J., Ni, X., Sun, J.T., Yang, Q., Chen, Z., 2010. Cross-domain sentiment classification via spectral feature alignment, in: Proceedings of the 19th International Conference on World Wide Web, ACM, New York, NY, USA. pp. 751–760.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Burzo, M., 2015. Deception detection using real-life trial data, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, New York, NY, USA. pp. 59–66.

Pérez-Rosas, V., Mihalcea, R., 2014. Cross-cultural deception detection, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 440–445.

Ren, Y., Ji, D., 2017. Neural networks for deceptive opinion spam detection: An empirical study. Information Sciences 385-386, 213–224.

Rosso, P., Cagnina, L.C., 2017. Deception Detection and Opinion Spam. Springer International Publishing, Cham. chapter 8. pp. 155–171.

Sánchez-Junquera, J., Villaseñor-Pineda, L., Montes-y-Gómez, M., Rosso, P., 2018. Character n-grams for detecting deceptive controversial opinions, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - Proceedings of the 9th International Conference of the CLEF Association. LNCS, vol. 11018, Springer-Verlag, pp. 135–140.

Stamatatos, E., 2017a. Authorship attribution using text distortion, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 1138–1149.

Stamatatos, E., 2017b. Masking topic-related information to enhance authorship attribution. Journal of the Association for Information Science and Technology 69, 461–473.

Tan, S., Cheng, X., Wang, Y., Xu, H., 2009. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. European Conference on Information Retrieval , 337–349.

Vogler, N., Pearl, L., 2018. Using linguistically-defined specific details to detect deception across domains. Natural Language Engineering 1, 1–27.

Vrij, A., 2000. Detecting lies and deceit: the psychology of lying and implications for professional practice. Wiley series in psychology of crime, policing and law, Wiley.

Vrij, A., 2008. Detecting lies and deceit: pitfalls and opportunities. Wiley Series in the Psychology of Crime, Policing and Law, Wiley.

Wu, Q., Tan, S., Duan, M., Cheng, X., 2010. A two-stage algorithm for domain adaptation with application to sentiment transfer problems, in: Information Retrieval Technology. Springer Berlin Heidelberg, pp. 443–453.

[5] The original versions are currently unavailable. Li et al. (2014) reported 0.784 (H->Rest) and 0.679 (H->Dr), whereas Ren and Ji (2017) reported 0.826 (H->Rest) and 0.676 (H->Dr).