Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1423–1430
Marseille, 11–16 May 2020

DecOp: A Multilingual and Multi-domain Corpus For Detecting Deception In Typed Text

Pasquale Capuozzo[1], Ivano Lauriola[2,3], Carlo Strapparava[3], Fabio Aiolli[2], Giuseppe Sartori[1]
[1] University of Padova, Department of General psychology
Via Venezia 8, 35131 - Padova, Italy
[2] University of Padova, Department of Mathematics
Via Trieste 63, 35121 - Padova, Italy
[3] Fondazione Bruno Kessler
Via Sommarive 18, 38123 - Trento, Italy
{pasquale.capuozzo, ivano.lauriola}@phd.unipd.it

Abstract

In recent years, the increasing interest in the development of automatic approaches for unmasking deception in online sources led to promising results. Nonetheless, among the others, two major issues remain still unsolved: the stability of classifiers performances across different domains and languages. Tackling these issues is challenging since labelled corpora involving multiple domains and compiled in more than one language are few in the scientific literature. For filling this gap, in this paper we introduce DecOp (Deceptive Opinions), a new language resource developed for automatic deception detection in cross-domain and cross-language scenarios. DecOp is composed of 5000 examples of both truthful and deceitful first-person opinions balanced both across five different domains and two languages and, to the best of our knowledge, is the largest corpus allowing cross-domain and cross-language comparisons in deceit detection tasks. In this paper, we describe the collection procedure of the DecOp corpus and his main characteristics. Moreover, the human performance on the DecOp test-set and preliminary experiments by means of machine learning models based on Transformer architecture are shown.
Keywords: deception, typed text, multilingual, multi-domain, corpus

1. Introduction

In the present day, the possibility of bumping into deceiving contents in online sources is higher than in the past because of the massive spreading of internet usage and the increasing amount of user-generated contents. The importance of this lies in the fact that previous findings in deception detection research showed that humans are ineffective in spotting deceit, with accuracy rates only slightly above the chance level (Bond Jr and DePaulo, 2006). Moreover, there are proofs that these poor performances are not influenced by factors such as age, sex, confidence, and experience and that professionals such as psychologists, detectives and judges are no more accurate than students and other citizens in this task (Aamodt and Custer, 2006). As a result, the possible dangers associated with the mentioned phenomena made necessary the development of new approaches for unmasking deceit not relying on human judgement.

Recently, there has been growing interest in the automatic detection of deception focused on language analysis (Hauch et al., 2015; Fitzpatrick et al., 2015). Among the others, some recent studies in this area have shown different promising applications by analysing the language of fake news (Conroy et al., 2015; Pérez-Rosas et al., 2017), court cases transcriptions (Fornaciari and Poesio, 2013; Yancheva and Rudzicz, 2013; Pérez-Rosas et al., 2015), deceptive product reviews (Ott et al., 2011; Fornaciari and Poesio, 2014; Kleinberg et al., 2018), cyber-crimes (Abozinadah et al., 2015; Mbaziira and Jones, 2016), autobiographical information (Levitan et al., 2018) and deceptive intentions regarding the future (Kleinberg et al., 2017).

Nonetheless, although the encouraging results achieved so far, there are at least two main unsolved issues in the automatic detection of deception.

A deceiving content can be created on a virtually unlimited range of subjects and thus the first issue refers to the generalization of the classification performances when dealing with data composed by different domains. Practically, previous findings showed a general decrease in classifiers' performance when training and test data belong to different domains and are significantly different in content (Krüger et al., 2017). In particular, previous studies about the automatic deception detection displayed that the effect of domain change determines a significant drop in the model performances in cross-domain classification experiments (Hernández-Castañeda et al., 2017; Pérez-Rosas and Mihalcea, 2014; Mihalcea and Strapparava, 2009). However, large labelled datasets which allow cross-domain comparisons are few in the scientific literature (Yao et al., 2017; Pérez-Rosas et al., 2017; Pérez-Rosas and Mihalcea, 2014; Mihalcea and Strapparava, 2009), making harder facing this issue.

The other unsolved argument refers to the stability of linguistic clues of deception across different languages. In fact, previous research findings highlighted contrasting results in support of both the stability (Matsumoto et al., 2015b; Matsumoto et al., 2015a; Matsumoto and Hwang, 2015) and inconsistencies (Leal et al., 2018; Rungruangthum and Todd, 2017; DeCicco and Schafer, 2015) across different languages of verbal clues of deception. Moreover, so far most studies in the field of automatic deceit detection have primarily focused on the English language while only a few works (Pérez-Rosas and Mihalcea, 2014) focused on the automatic assessment of the effective-

ness of English-based linguistic clues of deception in predicting deceit in languages other than English. This point as also the need for further research in detecting deception across different languages has been supported by some authors (Rungruangthum and Todd, 2017; Matsumoto et al., 2015b; Spence et al., 2012). Nonetheless, also here the scarceness of large labelled datasets compiled including one or more languages other than English make complex tackling this research question.

If one considers the mentioned issues it becomes clear that the introduction of a labelled corpus involving multiple domains and compiled in more than one language can be beneficial for a deeper understanding of automatic deceit detection.

With the aim of filling this gap, in this paper we introduce DecOp (Deceptive Opinions), a new multilingual and multi-domain corpus for the automatic detection of deception in typed text. DecOp is composed of both truthful and deceptive first-person opinions about five different domains, is compiled in two different languages and has been developed for allowing several kinds of classification experiments. In particular, beyond the more usual within-domain classification task, DecOp allows cross-domain, author-based and cross-language classification tasks.

2. Related work

Recently, the scientific community showed a rising interest in the automatic detection of deception focused on the analysis of linguistic clues of deceit in typed text (Fitzpatrick et al., 2015; Hauch et al., 2015; Nunamaker et al., 2012).

One of the resulting contributions is the introduction of several datasets for studying different aspects and scenarios in which deceit can lead to serious consequences. For example, previous works have presented labelled corpora composed of both truthful and deceptive reviews regarding hotels (Ott et al., 2011; Ott et al., 2013), books (Fornaciari and Poesio, 2014) and restaurants (Li et al., 2015), as well as collections of trustworthy news and fact-checked fake news (Wang, 2017; Pérez-Rosas et al., 2017). Nonetheless, the just mentioned works tried to face deception-related practical issues, focusing on single and specific domains. However, as stated in the previous section, since the difference in content between training and test data leads to a decrease in classification performances, the need for multi-domain datasets is primary for assessing the models' generalization abilities.

Multi-domains datasets for automatic deceit detection has been recently introduced concerning fake products reviews (Yao et al., 2017) and fake news (Pérez-Rosas et al., 2017). After compiling the sets of data, the authors grouped them per domain thus allowing the possibility to test the effect of domain change on the classifiers' performance. Another example of multi-domain dataset was presented by (Mihalcea and Strapparava, 2009). Using the Amazon Mechanical Turk (AMT) service, the authors collected both truthful and deceptive statements regarding three different domains: abortion, death penalty and feelings about best friends. The final corpus consisted of 100 truthful and 100 deceptive statements for each considered domain. Compared to the just mentioned multi-domain corpora, with DecOp we aim at introducing a larger corpus, extending the number of domains and including an additional feature: the possibility of assessing the model performances in detecting deception across different languages.

Actually, multi-domain datasets composed of typewritten truthful and deceptive statements compiled with the same procedure in more than one language are rare in the scientific literature. To the best of our knowledge, only one example of a corpus with the described characteristics exists (Pérez-Rosas and Mihalcea, 2014). Following the methodology of (Mihalcea and Strapparava, 2009), the authors asked people from United States, India, and Mexico to provide both truthful and deceptive statements about abortion, death penalty and feelings about best friends. The US and Indian participants were recruited via AMT while participants from Mexico were enlisted via an ad hoc web page. The authors collected 100 truthful and 100 deceptive essays for each domain from both the US and Indian participants while fewer contributions were obtained from Mexican participants. The dataset introduced by (Pérez-Rosas and Mihalcea, 2014) represents a unique contribution to the study of typewritten deception across different domains and languages. However, beyond the small corpus size, the employed data collection procedure exhibits one crucial difference compared to our work: Indian participants performed the task in English while American and Mexican participants were asked to complete the task in their native language. With DecOp, we try to overcome the mentioned shortcomings by expanding the number of domains and by asking participants to perform the task in their native language in order to allow direct assessment of the language change on classifiers' performance.

3. Data Collection Procedure

Based on previous approaches and existing data collection methodologies (Mihalcea and Strapparava, 2009; Pérez-Rosas and Mihalcea, 2014), for compiling DecOp we focused on first-person opinions on five different domains: Abortion (Abo), Cannabis legalization (CL), Euthanasia (Eut), Gay marriage (GM) and Policy on migrants (PoM). The rationale behind the selection of the domains is based on the assumption that the majority of people are likely to have their own point of view about these topics and thus they can easily express or deny it. On the other hand, for allowing a comparison between different languages, two different samples from both the US and Italy were considered. The two samples were asked to perform the task in their own language that is standard American English and Italian respectively.

All the participants were asked to type both truthful and deceptive first-person opinions about the above-mentioned topics in a free text response modality. The applied paradigm is based on an experimental ground-truth. This means that each participant provided a truthful or a deceptive opinion according to specific instructions. The truthful first-person opinions were generated by asking participants to provide in at least 4-5 lines their actual opinion about a given topic. Contrarily, for gathering deceptive opinions, participants were instructed to describe in at least 4-5 lines a fake opinion, not corresponding to their own opinion with

the main purpose to convince a hypothetical reader that the deceptive opinion represents their real point of view about the topic.
To maintain the balancement in the overall proportion between truthful and deceptive opinions, four Human Intelligence Tasks (HITs) were created. The HITs were balanced for ground-truth in a way that, overall, half of the first-person opinions gathered would have been deceptive and the other half truthful for each topic (Table 1). For instance, in the HIT 1, each participant typed his true opinion about Gay marriage, Euthanasia and Cannabis legalization while for Policy on migrants and Abortion they provided a deceptive opinion according to the instructions.

Domain	HIT1	HIT2	HIT3	HIT4
Abo	D	T	D	T
CL	T	T	D	D
Eut	T	D	T	D
GM	T	D	T	D
PoM	D	T	D	T

Table 1: HITs' structure description. T = Truthful opinion; D = Deceptive opinion; Abo = Abortion; CL = Cannabis Legalization; Eut = Euthanasia; GM = Gay Marriage; PoM = Policy on migrants.

The contributions from the US were collected through the Amazon Mechanical Turk (AMT) service, a popular crowdsourcing platform which allows to gather reliable data as those obtained via traditional methods (Buhrmester et al., 2016). The HITs were restricted to Turkers located in the US, whose approval rating was equal to or greater than 80%. The time allotted for each task was 20 minutes and, in order to avoid multiple entries from the same author, only a single submission per Turker was allowed. Each contribution was rewarded with 0.25$.
In order to gather comparable contributions in the Italian language, we followed as strictly as possible the same data collection procedure. Nonetheless, because only 2% of AMT workers are located in Italy (Difallah et al., 2018), we opted for the Google form service for collecting the Italian contributions. In line with the data collection procedure employed to gather the US contributions, four HITs were created respecting the same structure for both the domains and ground-truth. Participants were recruited on a volunteer basis by spreading the Google forms on social media and by e-mail. Furthermore, they have not received any monetary incentive for the HIT submission, no limitation in time was allotted for completing the task and only a single submission per participant was allowed.

4. Dataset description

In this section, the results of the data collection phase, the filtering procedure and a description of the main characteristics of DecOp are reported.
The four HITs implemented on AMT received a total of 727 submissions from Turkers located in the US. After a manual verification of the quality of each contribution, 227 were rejected because not in line with the instructions (i.e. unintelligible, unreasonably short, contained a description of the phenomenon instead of a first-person opinion etc.). After the manual check phase, the final collection consists of 500 HITs resulting in a total of 1250 truthful and 1250 deceptive opinions balanced across topics. The HITs were produced by 315 women and 185 men, with an average age of 37.7 ($\pm$ 13.2) years.
The four HITs built though the Google form service for the Italian language contributions received a total of 659 submissions. Following the same filtering procedure applied to the data collected from the US, manual verification of the quality of the contributions has been performed. Since 159 HITs were rejected because not in line with the instructions, the final collection consists of 500 responses, resulting in a total of 1250 truthful and 1250 deceptive opinions balanced across topics. The HITs were produced by 343 woman and 157 men, with an average age of 28.1 ($\pm$ 10.9) years.
The final version of the DecOp corpus was obtained by merging the data gathered from the two samples and its structure can be described as follow: for each participant, the five corresponding first-person opinions are provided with labels indicating their respective domain (Abo, CL, Eut, GM and PoM), their veracity condition (Truthful or Deceptive) and their language (standard American English or Italian). Moreover, the information about gender and age is provided for each participant. In Table 2 the DecOp corpus descriptive statistics are reported while Table 3 exhibits some examples of the first-person opinions gathered.

	EN	IT
writers	500	500
opinions	2500	2500
sentence count	11219	6302
word count	181016	137709
unique words	7177	10193

Table 2: Dataset statistics.

DecOp has been specifically designed to allow different analyses and classification tasks. In the following, we expose the DecOp capabilities and possible use-cases for machine learning methods. In detail, the deigned tasks are:

Within-topic In the simplest scenario, opinions come from the same single topic, reducing issues related to open domains. Analyses restricted to a single domain provide a robust baseline for other experiments, and allow to understand, to analyzed, and to compare different topics.

Cross-topic Real-world applications, such as fake-reviews detection, consists of texts and opinions from multiple, virtually infinite, domains. A Multi-domain dataset allows the development of models able to generalize when applied to new data, and the portability towards domains with a lack of available training data. Moreover, cross-topic analyses allow us to estimate and to quantify the drop in performance when the training is performed on different topics. However, DecOP is not an open domain dataset. As introduced in the data collection procedure, five different topics have been included in the corpus to tackle this problem.

TRUTHFUL	DECEPTIVE
DOMAIN: ABORTION	
IT	
Penso che ogni donna dovrebbe avere diritto di scegliere se portare avanti o meno una gravidanza. In ogni caso non mi schiero completamente a favore dell'aborto, perchè ci sono circostanze in cui viene comunque interrotta una gravidanza nonostante potrebbero esserci soluzioni alternative, che non comportino né un cambiamento di vita drastico per la donna, né la perdita di una vita.	L'aborto è una cosa inumana e non capisco come possa essere legale. Fortunatamente esistono gli obiettori di coscienza che decidono al posto di quelle sciagurate che hanno pensato bene di rimanere incinta e poi se ne pentono e vogliono uccidere il bambino. Che poi, cosa costa portare a termine la gravidanza e dare in adozione il bambino?
EN	
While I am morally torn on the issue, I believe that ultimately it is a woman's body and she should be able to do with it as she pleases. I belive people should not dehumanize the fetus tough, to make themselves feel better. The decision about laws regarding this issue should be left up to the states to decide. To combat this problem, birth control should be easily accessible.	Abortion is the termination of a life and should not be allowed. If a fetus has made it to the point of being able to survive "on its own" outside its mother's body, what right do we have to cut its life short. If the mother's life is in danger, she already chose that she was willing to sacrifice her life to have a child when she consented to procreating.

Table 3: The table shows some examples of the gathered opinions for both the Italian (IT) and standard American English (EN) languages included in the DecOp corpus. The domain considered for the example is Abortion.

Author-based Different writers have different behaviors, different writing styles, and they can express the same concept in different ways. The writing style information can dramatically improve the effectiveness of a machine learning model by comparing the target opinion against known data, such as past opinions, texts, and reviews. In a real-world scenario, historical data can be easily retrieved and used for this purpose. Some examples are comments on social networks or online reviews. To this end, the introduced corpus contains up to 5 opinions per writer, allowing author-based analyses.

Cross-language As noted already, currently it's unclear if verbal clues indicative of deceit can be assumed to be stable regardless of the language employed from the sender. Since DecOp was compiled with the same procedure in more than one language, it provides the possibility of assessing the classifiers' performances on different languages in the detection of deceit.

The available data has been split into training and test sets. The training set consists of opinions from 400 writers per language, whereas the test set contains the remaining 100 writers. The test writers have been uniformly sampled from the 4 HITs. As a consequence, the labels distribution is preserved. We release the DecOp corpus and this training/test split, and the code to performs basic and useful operations[1]. The same split has been used in our experimental assessment, described in Section 5.

5. Experimental assessment

This section describes a wide set of experiments carried out to assess the DecOp potentialities, and to provide baselines for future analyses.

Two main sets of experiments have been conducted. In the first set, a human evaluation has been performed on a binary classification task, where participants have been asked to categorize both truthful and deceptive opinions in a cross-topic setting. Secondly, popular machine learning models based on Transformer architecture (Vaswani et al., 2017) have been used to recognize deceptive opinions in within-topic, cross-topic and author-based scenarios.

The Figure 1 depicts the amount of data used in the within-topic, cross-topic, and author-based scenarios. All experiments have been repeated 5 times, and the average results have been collected.

5.1. Human Performance

In this section, an assessment of human performance in detecting deceptive opinions in the typed text on the DecOp test-set is performed. This procedure is necessary for at least two reasons. First of all, previous studies revealed that humans are weak lie detectors, with accuracy rates only slightly above the chance level (Bond Jr and DePaulo, 2006; Aamodt and Custer, 2006) and this is one of the main reasons why automated classification approaches in detecting deceit are becoming so popular. Thus, low human performances in this task would support the validity of our corpus. On the other hand, this validation procedure will provide a precise baseline against which the automated classification model performances may be compared.

Relying again on the AMT service, 100 Turkers located in the US (gender: F = 78, M = 22; age: M = 36,8, SD = 10,5) and 100 located in Italy (gender: F = 74, M = 26; age: M = 26,6, SD = 7,9) were recruited. The task was restricted to the Turkers whose approval rating was equal to or greater than 80%. The time allotted for each task was 20 minutes and only a single submission per Turker was allowed. Each submission was rewarded with 0.10$.

Each Turker was asked to classify 10 opinions randomly extracted from the DecOp test-set as truthful or deceptive in their respective language in a binary response modality. The overall classification performance was computed separately for the US participants and the Italian ones.

As expected, results confirmed the scarce efficiency of hu-

[1]The DecOp corpus is available for research purposes upon request.

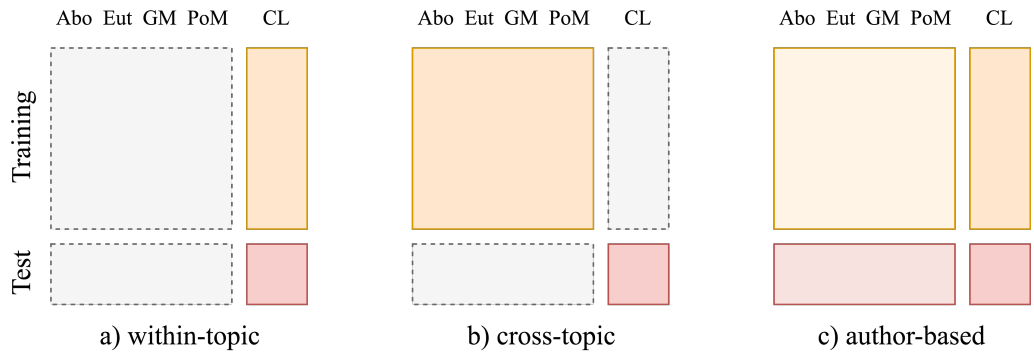

Figure 1: Tasks configuration. Given a target domain (CL in the specific case): a) the model is trained with all opinions from the same domain; b) the training set is extended to opinions from different domains; c) the target domain drives the classification, opinions from other domains inject the author's information.

mans in detecting deception with an overall accuracy of 57,9% and 58% for the Italian and the US participants respectively. Interestingly, a deeper examination of the poor performance achieved by both the groups revealed a common human judgemental bias affecting their abilities in unmasking deceit: the so-called truth-bias. Among the cognitive biases affecting deceit detection, the truth-bias is the most documented (Bond Jr and DePaulo, 2006; Levine et al., 1999; Zuckerman et al., 1981) and can be defined as the propensity of judging more often messages as truthful than deceptive, regardless of their actual veracity (Levine, 2014). In this case, both groups showed an unbalanced veracity judgement towards classifying opinions as truthful. Indeed, concerning the overall performance, the 79,3% and 62,9% of the opinions were classified as truthful from the Italian and the US participants respectively.

5.2. Exploring the corpus with Transformers

In the second evaluation phase, machine learning models have been used to assess the DecOp corpus in various tasks, providing strong baselines for future developments.
Thanks to its meaningful and astonishing results in the recent literature, the Transformer (Vaswani et al., 2017) encoder has been used in this paper.
In detail, a reduced version of the Transformer has been used. The architecture consists of an embedding layer to provide an initial representations of words and sequences, and a stacked sequence of transformer blocks.
Amongst a plethora of available word embeddings, a pre-trained FastText (Bojanowski et al., 2017) has been used to encode tokenized textual sequences and to fed the Transformer. In short, FastText is a popular word-embedding which merges word-level and character-level information. The selection of the word-embedding depended on two key points:

- FastText showed excellent performance on several tasks compared with other methods, such as word-2-vec.
- Recently, pre-trained FastText models have been released for 157 different languages (Grave et al., 2018). Most important, the same pre-training procedure has been applied to the models, allowing a multi-language comparison.

The hyper-parameters of the Transformer, i.e. the final architecture, have been preliminarily selected on a development set (10%) extracted from the training data. The resulting configuration is summarized in the Table 4, and it has been shared by all experiments.

Field	Value
dim. word emb.	300
encoding layers	2
attention heads	5
max seq length	256
learnable param.	800K
optimizer	Adam
learning rate	1e-5

Table 4: Transformer configuration.

Popular and effective pre-trained models, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), or ALBERT (Lan et al., 2019), have not been taken into account for multiple reasons, which are:

- The amount of available data can be insufficient to fine-tune the classification task. Although this problem can partially solved, for instance by injecting a domain adaption step before fine-tuning (Garg et al., 2019), this aspect beyond the scope of this work.
- Despite the effectiveness and the benefits of pre-trained language models, there is a lack of quality models for the Italian language, making a possible cross-language comparison unfair. It is easy to see that word embeddings are affected by the same issue. However, after a preliminary investigation, we argue that this gap is acceptable.
- This part of the work aim at providing a baseline for most of the tasks that can be analyzed with the DecOp corpus. Further models, extensions, ensembles, or specific techniques are outside the scope of this paper.

	language	Abo	CL	Eut	GM	PoM
within topic	EN	$0.656_{\pm 0.060}$	$0.630_{\pm 0.055}$	$0.676_{\pm 0.014}$	$0.620_{\pm 0.087}$	$0.676_{\pm 0.067}$
	IT	$0.656_{\pm 0.026}$	$0.688_{\pm 0.041}$	$0.684_{\pm 0.030}$	$0.664_{\pm 0.089}$	$0.732_{\pm 0.077}$
cross topic	EN	$0.720_{\pm 0.014}$	$0.726_{\pm 0.043}$	$0.692_{\pm 0.012}$	$0.710_{\pm 0.023}$	$0.758_{\pm 0.015}$
	IT	$0.818_{\pm 0.012}$	$0.818_{\pm 0.012}$	$0.788_{\pm 0.042}$	$0.816_{\pm 0.026}$	$0.772_{\pm 0.024}$
author-based	EN	$0.873_{\pm 0.005}$	$0.767_{\pm 0.019}$	$0.782_{\pm 0.085}$	$0.883_{\pm 0.015}$	$0.896_{\pm 0.014}$
	IT	$0.901_{\pm 0.016}$	$0.873_{\pm 0.027}$	$0.891_{\pm 0.019}$	$0.848_{\pm 0.020}$	$0.877_{\pm 0.010}$

Table 5: Average test accuracy scores and standard deviation computed on within-topic, cross-topic, and author-based tasks.

Three sets of experiments have been conducted, where the Transformer has been used in within-topic, cross-topic, and author-based classification tasks. In the within-topic setting, the Transformer has been trained and evaluated on opinions of a given individual target domain. This experiments provides us a useful baseline to better understand the task and its complexity. However, as introduced in the previous section, this setting is restrictive and it does not reflect a complex open domain scenario.
The cross-topic evaluation represents a natural extension of the previous experiment, which takes into account the multiplicity of domains. In this setting, the Transformer has been trained on 4 of the 5 available domains and it has been evaluated on the remaining target domain.

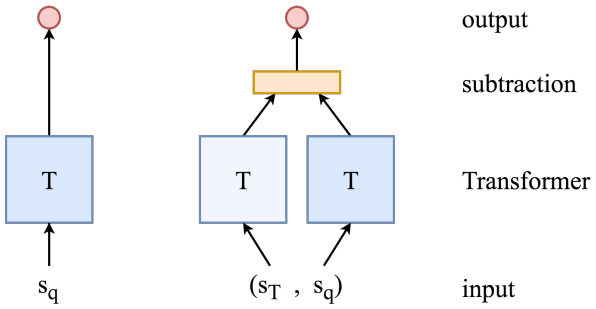

Figure 2: Left: the single-input architecture used to solve cross-topic and within-topic tasks. Right: a simple siamese network to catch authors patterns and writing styles. s_T represents a truthful sequence (opinion), whereas s_q is the questioned sequence.

In the author-based classification task, the model has been jointly trained with two different sources, which are the questioned opinion and the writer information. The questioned opinion is a training opinion which can be, as in the case of the previous experiments, either truthful or deceptive. The writer information is an extra source that helps the model to understand the author's writing style, and to reduce issues related to author's bias. An intuitive example of author bias is the the prolixity.
Several methods can be developed to inject writer's information into the system. Here, a simple siamese network has been used, whose input consists of pairs of opinions belonging to the same writer. The first element is a truthful opinion, whereas the second opinion can be either truthful or deceptive. The output of the network is the label of the questioned opinion.
The siamese network used in this work consists of a pair of Transformers, one for each input opinion. The configuration of the transformers is the same used in the previous experiments (see Table 4). The representation extracted on the top of the two transformer is merged to produce a joint representation for the input pair. A simple subtraction layer has been used for this purpose. Then, a classification layer has been used to produce the output. Figure 2 depicts the siamese network. Note that the author-based setting requires at least one truthful opinion in inference.

6. Results

Results of the machine learning models applied to within-topic, cross-topic, and author-based tasks are exposed in Table 5. There are three main results that deserve to be mentioned:

- Despite the wave of results from the literature (as discussed in section 1.), cross-topic models are more accurate than within-topic models. This aspect depends on two points. Firstly, cross-topic models have been trained on much more data (See Figure 1, in cross topic experiments the available opinions are 4 times more). Secondly, word-embeddings have been pretrained on open-domain corpora, and Transformer architectures, contrary to dictionary-based approaches such as n-grams, are able to catch semantic information rather than content information.

- The author-based approach improves the performance by a huge gap (5-10%), showing that the writer's information, if available, is fundamental for this task.

- Models trained on IT tasks systematically reach better performance than the EN counterpart. The recognition of deceptive opinions is an hard task in which the language plays a key role. Different words facets and tenses, typical of the IT language, may be better caught by FastText, which mixes word-level and character-level information. This aspect opens future research directions.

7. Conclusions

This paper introduces the DecOp corpus, a new language resource for the automatic detection of deception in typed text. DecOp is composed of both truthful and deceptive narratives about five different domains and has been gathered in two different languages: standard America English and Italian. With an overall size of 5000 examples balanced both across the five domains and the two languages considered, to the best of our knowledge, DecOp is the largest

corpus allowing cross-domain and cross-language classification tasks for the automatic detection of deceit in typewritten text. The resource is specifically designed for tackling two main unsolved issues related to automatic deceit detection: assessing if the model performances hold when training and test data belong to different domains and evaluating the stability of verbal clues of deception across different languages.

In order to assess the corpus potentialities, two main sets of experiments have been performed. The first part of the experimental assessment displayed poor classification performances of *human* judgement in detecting deceit on the DecOp test-set, highlighting the truth-bias as a possible explanation of these results. It is interesting to notice that although both the US and Italian participants exhibited comparable classification performances, the impact of the truth-bias seem to be more pronounced in the Italian samples than in the US ones (79.3% and 62,9% of opinions classified as truthful respectively). Overall, these results confirm previous findings of the weak human abilities in detecting deception (Bond Jr and DePaulo, 2006; Aamodt and Custer, 2006), thus providing some support to the validity of the DecOp corpus.

The second part of the experimental assessment showed that machine learning models based on Transformer architecture outperform the human performances on the same DecOp test-set, achieving promising results in within-topic, cross-topic and author-based scenarios.

In summary, with the introduction of the DecOp corpus, we aim at stimulating further research in the automatic deceit detection field with the goal of improving the state of the art results in multiple-domain and multilingual contexts.

8. References

Aamodt, M. G. and Custer, H. (2006). Who can best catch a liar? *Forensic Examiner*, 15(1).

Abozinadah, E. A., Mbaziira, A. V., and Jones, J. (2015). Detection of abusive accounts with arabic tweets. *Int. J. Knowl. Eng.-IACSIT*, 1(2):113–119.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bond Jr, C. F. and DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2016). Amazon's mechanical turk: A new source of inexpensive, yet high-quality data?

Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

DeCicco, A. J. and Schafer, J. R. (2015). Grammatical differences between truthful and deceptive narratives. *Applied Psychology in Criminal Justice*, 11(2).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Difallah, D., Filatova, E., and Ipeirotis, P. (2018). Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 135–143. ACM.

Fitzpatrick, E., Bachenko, J., and Fornaciari, T. (2015). Automatic detection of verbal deception. *Synthesis Lectures on Human Language Technologies*, 8(3):1–119.

Fornaciari, T. and Poesio, M. (2013). Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21(3):303–340.

Fornaciari, T. and 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*, pages 279–287.

Garg, S., Vu, T., and Moschitti, A. (2019). Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Hauch, V., Blandón-Gitlin, I., Masip, J., and Sporer, S. L. (2015). Are computers effective lie detectors? a meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review*, 19(4):307–342.

Hernández-Castañeda, Á., Calvo, H., Gelbukh, A., and Flores, J. J. G. (2017). Cross-domain deception detection using support vector networks. *Soft Computing*, 21(3):585–595.

Kleinberg, B., Nahari, G., Arntz, A., and Verschuere, B. (2017). An investigation on the detectability of deceptive intent about flying through verbal deception detection. *Collabra: Psychology*, 3(1).

Kleinberg, B., Mozes, M., Arntz, A., and Verschuere, B. (2018). Using named entities for computer-automated verbal deception detection. *Journal of forensic sciences*, 63(3):714–723.

Krüger, K. R., Lukowiak, A., Sonntag, J., Warzecha, S., and Stede, M. (2017). Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, 23(5):687–707.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Leal, S., Vrij, A., Vernham, Z., Dalton, G., Jupe, L., Harvey, A., and Nahari, G. (2018). Cross-cultural verbal deception. *Legal and Criminological Psychology*, 23(2):192–213.

Levine, T. R., Park, H. S., and McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communications Monographs*, 66(2):125–144.

Levine, T. R. (2014). Truth-default theory (tdt) a theory of human deception and deception detection. *Journal of Language and Social Psychology*, 33(4):378–392.

Levitan, S. I., Maredia, A., and Hirschberg, J. (2018). Linguistic cues to deception and perceived deception in in-

terview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1941–1950.

Li, H., Chen, Z., Mukherjee, A., Liu, B., and Shao, J. (2015). Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In *ninth international AAAI conference on web and social Media*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Matsumoto, D. and Hwang, H. C. (2015). Differences in word usage by truth tellers and liars in written statements and an investigative interview after a mock crime. *Journal of Investigative Psychology and Offender Profiling*, 12(2):199–216.

Matsumoto, D., Hwang, H. C., and Sandoval, V. A. (2015a). Cross-language applicability of linguistic features associated with veracity and deception. *Journal of Police and Criminal Psychology*, 30(4):229–241.

Matsumoto, D., Hwang, H. C., and Sandoval, V. A. (2015b). Ethnic similarities and differences in linguistic indicators of veracity and lying in a moderately high stakes scenario. *Journal of Police and Criminal Psychology*, 30(1):15–26.

Mbaziira, A. and Jones, J. (2016). A text-based deception detection model for cybercrime. In *Int. Conf. Technol. Manag*.

Mihalcea, R. and Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312. Association for Computational Linguistics.

Nunamaker, J. F., Burgoon, J. K., Twyman, N. W., Proudfoot, J. G., Schuetzler, R., and Giboney, J. S. (2012). Establishing a foundation for automated human credibility screening. In *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on*, pages 202–211. IEEE.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (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-Volume 1*, pages 309–319. Association for Computational Linguistics.

Ott, M., Cardie, C., and 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*, pages 497–501.

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

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., and Burzo, M. (2015). Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 59–66. ACM.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2017). Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Rungruangthum, M. and Todd, R. W. (2017). Differences in language used by deceivers and truth-tellers in thai online chat. *Journal of the Southeast Asian Linguistics Society*, 10(2):90–114.

Spence, K., Villar, G., and Arciuli, J. (2012). Markers of deception in italian speech. *Frontiers in psychology*, 3:453.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Yancheva, M. and Rudzicz, F. (2013). Automatic detection of deception in child-produced speech using syntactic complexity features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 944–953.

Yao, W., Dai, Z., Huang, R., and Caverlee, J. (2017). Online deception detection refueled by real world data collection. *arXiv preprint arXiv:1707.09406*.

Zuckerman, M., DePaulo, B. M., and Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In *Advances in experimental social psychology*, volume 14, pages 1–59. Elsevier.