Natural Language Engineering (2022), **28**, pp. 545–606
doi:10.1017/S1351324921000152

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Deception detection in text and its relation to the cultural dimension of individualism/collectivism

Katerina Papantoniou[1,2,*], Panagiotis Papadakos[1,2], Theodore Patkos[2], George Flouris[2], Ion Androutsopoulos[3] and Dimitris Plexousakis[1,2]

[1]Computer Science Department, University of Crete, Heraklion, Greece, [2]Institute of Computer Science, FORTH-ICS, Heraklion, Greece, and [3]Department of Informatics, Athens University of Economics and Business, Athens, Greece
*Corresponding author. E-mail: papanton@ics.forth.gr

(Received 16 April 2019; revised 24 May 2021; accepted 31 May 2021; first published online 2 July 2021)

Abstract

Automatic deception detection is a crucial task that has many applications both in direct physical and in computer-mediated human communication. Our focus is on automatic deception detection in text across cultures. In this context, we view culture through the prism of the individualism/collectivism dimension, and we approximate culture by using country as a proxy. Having as a starting point recent conclusions drawn from the social psychology discipline, we explore if differences in the usage of specific linguistic features of deception across cultures can be confirmed and attributed to cultural norms in respect to the individualism/collectivism divide. In addition, we investigate if a universal feature set for cross-cultural text deception detection tasks exists. We evaluate the predictive power of different feature sets and approaches. We create culture/language-aware classifiers by experimenting with a wide range of n-gram features from several levels of linguistic analysis, namely phonology, morphology and syntax, other linguistic cues like word and phoneme counts, pronouns use, etc., and token embeddings. We conducted our experiments over eleven data sets from five languages (English, Dutch, Russian, Spanish, and Romanian), from six countries (United States of America, Belgium, India, Russia, Mexico, and Romania), and we applied two classification methods, namely logistic regression and fine-tuned BERT models. The results showed that the undertaken task is fairly complex and demanding. Furthermore, there are indications that some linguistic cues of deception have cultural origins and are consistent in the context of diverse domains and data set settings for the same language. This is more evident for the usage of pronouns and the expression of sentiment in deceptive language. The results of this work show that the automatic deception detection across cultures and languages cannot be handled in unified manners and that such approaches should be augmented with knowledge about cultural differences and the domains of interest.

Keywords: Deception detection; Text classification; Language resources; Machine learning; Culture

1. Introduction

Automated deception detection builds on years of research in interpersonal psychology, philosophy, sociology, communication studies, and computational models of deception detection (Vrij 2008a; Granhag *et al.* 2014). Textual data of any form, such as consumer reviews, news articles, social media comments, political speeches, witnesses' reports, etc., are currently in the spotlight of deception research (Granhag *et al.* 2014). What contributed to this vivid interest is the enormous production of textual data and the advances in computational linguistics. In many cases, text is either the only available source for extracting deception cues or the most affordable and less intrusive one, compared to approaches based on magnetic resonance imaging (Lauterbur 1973) and electrodermal activity (Critchley and Nagai 2013). In this work, we exploit natural language processing (NLP) techniques and tools for automated text-based deception detection and focus on the relevant cultural and language factors.

As many studies suggest, deception is an act that depends on many factors such as personality (Fornaciari *et al.* 2013; Levitan *et al.* 2015), age (Sweeney and Ceci 2014), gender (Tilley *et al.* 2005; Toma et al. 2008; Fu *et al.* 2008), or culture (Taylor *et al.* 2014; Taylor *et al.* 2017; Leal *et al.* 2018). All these factors affect the way and the means one uses to deceive. The vast majority of works in automatic deception detection take an "one-size-fits-all" approach, failing to adapt the techniques based on such factors. Only recently, research efforts that take into account such parameters started to appear (Pérez-Rosas and Mihalcea 2014; Pérez-Rosas *et al.* 2014).

Culture and *language* are tightly interconnected since language is a means of expression, embodiment, and symbolization of cultural reality (Kramsch 2011) and as such differences among cultures are reflected in language usage. According to previous studies (Rotman 2012; Taylor *et al.* 2014; Taylor *et al.* 2017; Leal *et al.* 2018), this also applies to the expression of *deception* among people belonging to different cultures (a detailed analysis related to this point is provided in Section 2.2). The examination of the influence of cultural properties in deception detection is extremely important since differences in social norms may lead to misjudgments and misconceptions and consequently can impede fair treatment and justice (Jones and Newburn 2001; Taylor *et al.* 2014). The globalization of criminal activities that employ face-to-face communication (e.g., when illegally trafficking people across borders) or digital communication (e.g., phishing in e-mail or social media), as well as the increasing number of people passing interviews in customs and borders all over the world are only some scenarios that make the incorporation of cultural aspects in the research of deception detection a necessity. Since the implicit assumption made about the uniformity of linguistic indicators of deception comes in conflict with prior work from psychological and sociological disciplines, our three research goals are

(a) Can we verify the prior body of work which states that linguistic cues of deception are expressed differently, for example, are milder or stronger, across cultures due to different cultural norms? More specifically, we want to explore how the individualism/collectivism divide defines the usage of specific linguistic cues (Taylor *et al.* 2014; 2017). Individualism and collectivism constitute a well-known division of cultures, and concern the degree in which members of a culture value more individual over group goals and vice versa (Triandis *et al.* 1988). Since cultural boundaries are difficult to define precisely when collecting data, we use data sets from different countries assuming that they reflect at an aggregate level the dominant cultural aspects that relate to deception in each country. In other words, we use countries as proxies for cultures, following in that respect Hofstede (2001). We also experiment with data sets originating from different text genres (e.g., reviews about hotels and electronics, opinions about controversial topics, transcripts from radio programs, etc.).

(b) Explore which language indicators and cues are more effective to detect deception given a piece of text and identify if a *universal feature set*, that we could rely on for detection deception tasks exists. On top of that, we investigate the volatility of cues across different domains by keeping the individualism/collectivism and language factors steady, whenever we have appropriate data sets at our disposal.

(c) In conjunction with the previous research goal, we create and evaluate the performance of a wide range of binary classifiers for predicting the truthfulness and deceptiveness of text.

These three research goals have not been addressed before, at least from this point of view. Regarding the first goal, it is particularly useful to confirm some of the previously reported conclusions about deception and culture under the prism of individualism/collectivism with a larger number of samples and from populations beyond the closed environments of university campuses

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

and small communities used by the original studies. For the other two research goals, we aim at providing an efficient methodology for the deception detection task, exploring the boundaries and limitations of the options and tools currently available for different languages.

To answer our first and second research goals, we performed statistical tests on a set of linguistic cues of deception already proposed in bibliography, placing emphasis on those reported to differentiate across the individualism/collectivism divide. We conducted our analysis on datasets originating from six countries, namely United States of America, Belgium, India, Russia, Romania, and Mexico, which are seen as proxies of cultural features at an aggregate level. Regarding the third research goal, the intuition is to explore different approaches for deception detection, ranging from methodologies that require minimal linguistics tools for each language (such as word n-grams), to approaches that require deeper feature extraction (e.g., syntactic features obtained via language-specific parsers) or language models that require training on large corpora, either in separation or in combination. One of our challenges is the difficulty to collect and produce massive and representative deception detection data sets. This problem is amplified by the diversity of languages and cultures, combined with the limited linguistic tools for under-researched languages despite recent advances (Conneau *et al.* 2018; Alyafeai *et al.* 2020; Hu *et al.* 2020; Hedderich *et al.* 2020). To this end, we exploit various widely available related data sets for languages with adequate linguistic tools. We also create a new data set based on transcriptions from a radio game. For each language under research, we created classifiers using a wide range of n-gram features from several levels of linguistic analysis, namely, phonological, morphological, and syntactic, along with other linguistic cues of deception and token embeddings. We provide the results of the experiments from logistic regression classifiers, as well as fine-tuned BERT models. Regarding BERT, we have experimented with settings specific to each particular language, based on the corresponding monolingual models, as well as with a cross-language setting using the multilingual model (Devlin *et al.* 2019).

In the remainder of this paper, we first present the relevant background (Section 2), including both theoretical work and computational work relevant to deception and deception detection, with emphasis on the aspects of culture and language. We then proceed with the presentation of the data sets that we utilized (Section 3), the feature extraction process (Section 4), and the statistical evaluation of linguistic cues (Section 5). Subsequently, we present and discuss the classification schemes and the evaluation results, comparing them with related studies (Section 6). Finally, we conclude and provide some future directions for this work (Section 7).

2. Background

2.1 Deception in psychology and communication

Several theories back up the observation that people speak, write, and behave differently when they are lying than when they are telling the truth. Freud was the first who observed that the subconscious feelings of people about someone or something are reflected in how they behave and the word choices they make (Freud 1914). The most influential theory that connects specific linguistic cues with the truthfulness of a statement is the Undeutsch hypothesis (Undeutsch 1967; Undeutsch 1989). This hypothesis asserts that statements of real-life experiences derived from memory differ significantly in content and quality from fabricated ones, since the invention of a fictitious memory requires more cognitive creativity and control than remembering an actually experienced event.

On this basis, a great volume of research work examines which linguistic features are more suitable to distinguish a truthful from a deceptive statement. These linguistic features can be classified roughly into four categories: word counts, pronoun use, emotion words, and markers of cognitive complexity. The results for these dimensions have been contradictory and researchers seem to agree that cues are heavily *context-dependent*. More specifically, the importance of specific

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 1. Social psychology studies on within and across culture deception detection

Reference	Description	Within & Across culture accuracy (%)	
Bond *et al.* (1990)	Jordanian and US undergraduate students were videotaped while telling lies and truths for the examination of a deceiver's nonverbal behavior.	56	49
Bond and Atoum (2000)	American, Jordanian and Indian students, as well as an illiterate Indian sample were videotaped similarly to Bond *et al.* (1990).	54	51
Lewis and George (2008)	Study on deceptive computer-mediated communication between Spanish and US participants.	59	51

linguistic features tends to change based on many parameters such as the type of text, for example, dialogue, narrative (Picornell 2013), the medium of the communication, for example, face-to-face, computer-mediated (Zhou *et al.* 2004; Hancock *et al.* 2007; Zhou and Zhang 2008; Rubin 2010), deception type (Frank and Ekman 1997), how motivated the deceiver is (Frank and Ekman 1997), etc. There is also a volume of work that examines how the conditions that the experiments were performed in, for example, sanctioned, unsanctioned, influence the accuracy results, and the behavior of the participants (Feeley and deTurck 1998; Dunbar *et al.* 2015; Burgoon 2015).

Given the volatility of the results within even the context of a specific language, the implicit assumption made about the universality of deception cues can lead to false alarms or misses. Differences in social norms and etiquette, anxiety, and awkwardness that may stem from the language barrier (when speakers do not use their native languages) can distort judgments. A reasonable argument is that, since the world's languages differ in many ways, the linguistic cues which might have been identified as deceptive in one language might not been applicable to another. For example, a decrease in first person personal pronoun use is an indicator of deception in English (Hauch *et al.* 2015). What happens though in languages where personal pronoun use is not always overt such as in Italian, Spanish, Greek and Romanian (i.e., null subject languages)? In addition, modifiers (i.e., adjectives and adverbs), prepositions, verbs are also commonly examined cues. But not all languages use the same grammatical grammatical categories; for example, Russian and Polish have no articles (Newman *et al.* 2003; Zhou *et al.* 2004; Spence *et al.* 2012).

All psychology and communication studies that involve participants from different cultural groups, asking them to identify truth and fabrications within the same and different cultural group, conclude to the same result about the accuracy rate of predictions. More specifically, as Table 1 indicates, the accuracy rate in all the studies dropped to chance when judgments were made across cultures, whereas for within culture judgments, it was in line with the rest of the bibliography, that places accuracy to be typically slightly better than chance (DePaulo *et al.* 1985). Indeed, deception detection turns out to be a very challenging task for humans. It is indicative that even in studies that involve people who have worked for years at jobs that require training in deception detection, such as investigators or customs inspectors, the results are not significantly better (Ekman and O'Sullivan 1991). These results are usually attributed to *truth bias*, that is, the tendency of humans to actively believe or passively presume that another person is honest, despite even evidence to the contrary (DePaulo *et al.* 1985; Vrij 2008b). The further impairment in accuracy in across culture studies is attributed to the *norm violation model*. According to this model, people infer deception whenever the communicator violates what the receiver anticipates as being normative behavior, and this is evident in both verbal and nonverbal communication (Taylor *et al.* 2014).

2.2 Culture and language

The correlation and interrelation between cultural differences and language usage has been extensively studied in the past. The most influential theory is the Sapir–Whorf hypothesis that is also known as the theory of the linguistic relativity (Sapir 1921; Whorf 1956). This theory suggests that language influences cognition. Thus every human views the world by his/her own language. Although influential, the strong version of the Sapir–Whorf hypothesis has been heavily challenged (Deutscher 2010). However, neo-Whorfianism that is a milder strain of the Sapir–Whorf hypothesis is now an active research topic (West and Graham 2004; Boroditsky 2006), stating that language influences a speaker's view of the world but does not inescapably determine it.

Another view of the relationship between language and culture is the notion of *linguaculture* (or *languaculture*). The term was introduced by linguistic anthropologists Paul Friedrich (1989) and Michael Agar (1994). The central idea is that a language is culture bound and much more than a code to label objects found in the world (Shaules 2019).

Early studies (Haire *et al.* 1966; Whitely and England 1980) support that language and cultural values are correlated in the sense that the cross-cultural interactions that account for similarity in cultural beliefs (geographic proximity, migration, colonization) also produce linguistic similarity. Haire *et al.* (1966) found Belgian-French and Flemish-speakers held values similar to the countries (France and the Netherlands) with which they shared language, religion, and other aspects of cultural heritage. In such cases, parallel similarities of language and values can be seen because they are part of a common cultural heritage transmitted over several centuries.

2.3 Deception and culture

The *individualism/collectivism* dipole is one of the most viable constructs to differentiate cultures and express the degree to which people in a society are integrated into groups. In individualism, ties between individuals are loose and individuals are expected to take care of only themselves and their immediate families, whereas in collectivism ties in society are stronger. The individualism/collectivism construct strongly correlates with the distinction between high and low-context communication styles (Hall 1976). The low-context communication style, which is linked with more individualist cultures, states that messages are more explicit, direct, and the transmitter is more open and expresses true intentions. In contrast, in a high context communication messages are more implicit and indirect, so context and word choices are crucial in order for messages to be communicated correctly. The transmitter in this case tries to minimize the content of the verbal message and is reserved in order to maintain social harmony (Wrtz 2017). Some studies from the discipline of psychology examine the behavior of verbal and nonverbal cues of deception across different cultural groups based on these constructs (Taylor *et al.* 2014, 2017; Leal *et al.* 2018).

In the discipline of psychology, there is a recent work from Taylor *et al.* (2014, 2017) that comparatively examines deceptive lexical indicators among diverse cultural groups. More specifically, Taylor *et al.* (2014) conducted some preliminary experiments over 60 participants from four ethnicities, namely White British, Arabian, North African, and Pakistani. In Taylor *et al.* (2017), the authors present an extended research work, over 320 individuals from four ethnic groups, namely Black African, South Asian, White European, and White British, who were examined for estimating how the degree of the individualism and collectivism of each culture, influences the usage of specific linguistic indicators in deceptive and truthful verbal behavior. The participants were recruited from community and religious centers across North West England and were *self-assigned* to one of the groups. The task was to write one truthful and one deceptive statement about a personal experience, or an opinion and counter-opinion in English. In the study, the collectivist group (Black African and South Asian) decreased the *usage of pronouns* when lying and used more first-person and fewer third-person pronouns to distance the social group from the deceit. In contrast, the individualistic group (White European and White British) used fewer first-person and more third-person pronouns, to distance themselves from the deceit.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 2. Studies from social psychology discipline on the expression of sentiment in individualism and collectivism

Work	Description
Vrij *et al.* (2010)	Deception acts might emerge feelings of guilt, fear, or delight.
Markus and Kitayama (1991)	Collectivists, in order to avoid conflict and to protect social harmony, may be more engaged to friendly emotions rather than to more unattached emotions like anger.
Seiter *et al.* (2002)	Collectivists consider lying more socially acceptable behavior. To this end, emotions, as proposed by Vrij *et al.* (2010) might not emerge in the first place.
Matsumoto *et al.* (2008)	Individualism is connected with emotional expression, whereas collectivists are more probable to restraint their emotional expression.

In these works, Taylor stated the hypothesis that affect in deception is related to cultural differences. This hypothesis was based on previous related work that explored the relation between sentiment and deception across cultures, which is briefly summarized in Table 2. The results though refute the original hypothesis, showing that the use of *positive affect* while lying was consistent among all the cultural groups. More specifically, participants used more positive affect words and fewer words with negative sentiment when they were lying, compared to when they were truthful. Based on his findings, emotive language during deception may be a strategy for deceivers to maintain social harmony.

According to the same study, the use of *negations* is a linguistic indicator of deception in the collectivist group, but is unimportant for the individualist group. Negations have been studied a lot with respect to differences among cultures and the emotions they express. Stoytcheva *et al.* (2014) conclude that Asian languages speakers are more likely to use negations than English speakers, due to preference to the indirect style of communication. Moreover, Mafela (2013) states that for South African languages the indirect style of communication leads to the usage of negation constructs for the expression of positive meanings.

Contextual details is a cognition factor also examined in Taylor's works. According to the related literature, contextual details such as the spatial arrangement of people or objects, occur naturally when people describe existing events from their memory. The key finding of this study suggests that this is actually true for the relatively more individualistic participants, for example, European. For the collectivist groups though, spatial details were less important while experiencing the event at the first place and subsequently during recall. As a result, individualist cultures tend to provide fewer perceptual details and more social details when they are lying a trend that changes in collectivist cultures. Table 3 summarizes all the above findings.

It is important to mention that the discrepancies on linguistic cues between individualist and collectivist groups were not confirmed for all types of examined lies, namely lies about opinions and experiences. In more details, the analysis showed that *pronoun use* and *contextual embedding* (e.g., the "circumstances") varied when participants lied about experiences, but not when they lied about opinions. By contrast, the affect-related language of the participants varied when they lied about opinions, but not experiences. All the above findings indicate that it does not suffice to conceptualize liars as people motivated "to not get caught", since additional factors influence the way they lie, what they do not conceal, what they have to make up, who they want to protect, etc.

Leal *et al.* (2018) investigate if differences in low and high context culture communication styles can be incorrectly interpreted as cues of deceit in verbal communication. Through collective interviews, they studied British interviewees as a representatives of low-context cultures, and Chinese and Arabs as representatives of high-context cultures. The key findings of this work revealed that

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 3. Summary of differences in language use between truthful and deceptive statements across the four cultural groups examined in the work of Taylor *et al.* (2014); Taylor *et al.* (2017). Differences in pronoun usage and perceptual details were confirmed when participants lied about experiences, whereas affective language differences were confirmed when participants lied about opinions

Language indicator	White British (I)	White European (I)	Black African (C)	South Asian (C)
Positive affect	↑↑	↑↑	↑↑	↑↑
Negations	–	↑	–	↑
Perceptual details information	↓↓	↓	↑	↑↑
1st pers. pronouns	↓↓	↓	↑	↑↑
3rd pers. pronouns	↑↑	↑	↓	↓↓

(↑) more in deceptive, (↓) more in truthful, (–) no difference, (I) individualism, (C) collectivism, (↑↑, ↓↓) suggest larger differences between truthful and deceptive statements.

indeed differences between cultures are more prominent than differences between truth tellers and liars, and this can lead to communication errors.

2.4. Automated text-based deception detection

From a computational perspective, the task of deception detection that focuses on pursuing linguistic indicators in text is mainly approached as a classification task that exploits a wide range of features. In this respect, most research work combines psycholinguistic indicators drawn from prior work on deception (DePaulo *et al.* 1985; Porter and Yuille 1996; Newman *et al.* 2003) along with n-gram features (mainly word n-grams), in order to enhance predictive performance in a specific context. As already stated, the psycholinguistic indicators seem to have a strong discriminating power in most of the studies, although the quantitative predominance in truthful or deceptive texts is extremely sensitive to parameters, such as how motivated the deceiver is, the medium of communication and the overall context. The number of words that express negative and positive emotions, the number of pronouns, verbs, and adjectives, and the sentence length are among the most frequently used features.

Hirschberg *et al.* (2005) obtain *psycholinguistic* indicators by using the lexical categorization program LIWC (Pennebaker *et al.* 2001) along with other features to distinguish between deceptive and non-deceptive speech. In the work of Gîrlea *et al.* (2016), psycholinguistic deception and persuasion features were used for the identification of deceptive dialogues using as a data set dialogues taken from the party game Werewolf (also known as Mafia)[a]. For the extraction of the psycholinguistic features, the MPQA subjectivity lexicon[b] was used, as well as manually created lists. Various LIWC psycholinguistic, morphological, and n-gram features for tackling the problem of the automatic detection of deceptive opinion spam[c] are examined by Ott *et al.* (2011; 2013). These feature sets were tested in a linear Support Vector Machine (SVM) (Cortes and Vapnik 1995). In these two works, Ott *et al.* (2011 2013) provide two data sets with deceptive and truthful opinions, one with positive sentiment reviews (Ott *et al.* 2011) and one with negative sentiment (Ott *et al.* 2013). These data sets, either in isolation or combined, have been used as a gold standard in many works. Kleinberg *et al.* (2018) examined the hypothesis that the number of named entities is higher in truthful than in deceptive statements, by comparing the discriminative ability of named entities with a lexicon word count approach (LIWC) and a measure of sentence specificity. The results suggest that named entities may be a useful addition to existing approaches.

[a] https://en.wikipedia.org/wiki/Mafia_(party_game)
[b] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon
[c] Any fictitious opinion that has been deliberately written to sound authentic.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Feng *et al.* (2012) investigated how *syntactic* stylometry can help in text deception detection. The features were obtained from Context Free Grammar (CFG) parse trees and were tested over four different data sets, spanning from product reviews to essays. The results showed improved performance compared to several baselines that were based on shallower lexico-syntactic features.

Discourse and pragmatics have also been used for the task of deception detection. Rhetorical Structure Theory (RST) and Vector Space Modeling (VSM) are the two theoretical components that have been applied by Rubin and Vashchilko (2012) in order to set apart deceptive and truthful stories. The authors proposed a two-step approach: in the first step, they analyzed rhetorical structures, discourse constituent parts and their coherence relations, whereas in the second, they applied a vector space model to cluster the stories by discourse feature similarity. Pisarevskaya and Galitsky (2019) also explored the hypothesis that deception in text should be visible from its discourse structure. They formulated the task of deception detection as a classification task using discourse trees, based on RST. For evaluation reasons, they created a data set containing 2746 truthful and deceptive complaints about banks in English, where the proposed solution achieved a classification accuracy of 70%.

The motivation of Hernández-Castañeda *et al.* (2017) was to build a *domain-independent classifier* using SVM. The authors experimented with different feature sets: a continuous semantic space model represented by *Latent Dirichlet Allocation* (LDA) topics (Blei *et al.* 2003), a binary word-space model (Sahlgren 2006), and dictionary-based features in five diverse domains (reviews for books and hotels; opinions about abortion, death penalty, and best friend). The results revealed the difficulties of building a robust cross-domain classifier. More specifically, the average accuracy of 86% in the one-domain setting dropped to a range of 52% to 64% in a cross-domain setting, where a data set is kept for testing and the rest are used for training. LDA was also used by Jia *et al.* (2018) along with term frequency and word2vec (Mikolov *et al.* 2013) for the feature extraction step in a supervised approach to distinguish between fake and non-fake hotel and restaurant reviews. These different features types were examined both separately and in combination, while three classifiers were trained, namely logistic regression, SVM, and multilayer perceptron (MLP) (Rumelhart and McClelland 1987). The evaluation was performed using the Yelp filter data set[d] (Mukherjee *et al.* 2013a), and the experimental results showed that the combinations of LDA with logistic regression and LDA with MLP performed better with 81% accuracy. The work of Martinez-Torres and Toral (2019) focuses on how features may change influenced by the nature of the text in terms of content and polarity. The proposed method examines three different features types based on a bag of words representation. The first type uses all the words in a vocabulary (after a preprocessing step), the second one selects word features that are uniquely associated with each class (deceptive, truthful), while the third one further extends the classes to four, also adding the sentiment polarity factor. The data set of Ott *et al.* (2011 2013) was used for the evaluation of the six classifiers (i.e., k-NN, logistic regression, SVM, random forest, gradient boosting, and MLP) that were employed.

Fontanarava *et al.* (2017) proposed combining a large number of reviews along with reviewer features for the detection of fake reviews. Some of the features were newly introduced for the task inspired by relevant research in fake news. Features were fed to a random forest classifier which was evaluated on the Yelp filter data set. The results show that the combined features were beneficiary for the task studied.

Finally, various kinds of *embeddings* (e.g., token, node, character, document, etc.) and *deep learning* approaches have been applied to the deception detection task. One of the first works is that of Ren and Ji (2017) that employs a Bidirectional Long Short-Term Memory network (BiLSTM) (Graves *et al.* 2013) to learn document-level representations. A semi-supervised approach is employed in Yilmaz and Durahim (2018) for the detection of spam reviews, by using a combination of doc2vec (Le and Mikolov 2014) and node2vec (Grover and Leskovec 2016) embeddings. These embeddings are then fed into a logistic regression classifier to identify opinion

[d] http://odds.cs.stonybrook.edu/yelpchi-dataset/

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

spam. Zhang *et al.* (2018) proposed a deceptive review identification method that uses recurrent convolutional neural networks (Liang and Hu 2015) for opinion spam detection. The basic idea is that since truthful reviews have been written by people in the context of the real experience, while the deceptive ones are not, this contextual information can be exploited by the model. Aghakhani *et al.* (2018) adopted Generative Adversarial Networks (GANs) (Goodfellow *et al.* 2014) for the detection of deceptive reviews.

Non-English and multilanguage research

Without a doubt, the English language engrosses the majority of the research interest for the task of deception detection, due to the bigger pool of English speaking researchers, the interest of industry for commercial exploitation and the abundance of linguistic resources. However, analogous approaches have been utilized also in *other languages*.

In the work of Verhoeven and Daelemans (2014), the task of deception detection from text for the Dutch language is explored by using an SVM with unigram features. In the absence of any related data set, the authors proceeded with the construction of their own data set. SVMs have also been used for deception detection in opinions written in Spanish with the use of the Spanish version of the LIWC (Almela *et al.* 2012).

Similarly, in the work of Tsunomori *et al.* (2015), a dialogue corpus for the Japanese language is presented and subsequently a binary classification based on decision trees over this corpus is performed using acoustic/prosodic, lexical, and subject-dependent features. The comparison with a similar English corpus has shown interesting results. More specifically, while in the prosodic/acoustic features, there were no differences between the two languages, in lexical features the results were greatly different. In English, noise, third person pronoun, and features indicating the presence of "Yes" or "No" were effective. In Japanese the lexical features used in this research were largely ineffective; and only one lexical feature, the one that indicated the presence of a verb base form, proved effective.

For the Chinese language, one of the first studies is that of Zhou and Sung (2008) who examined the computer-mediated communication of Chinese players engaged in the Werewolf game. Having as starting point prior research for English, they ended up with a list of features (e.g., number of words, number of messages, average sentence length, average word length, total number of first-person and third person singular/plural pronouns) and they performed statistical analysis. Results revealed that, consistent with some studies for English speakers, the use of third person pronouns increased during deception. In Chinese though, there were no significant differences between the proportional use of first pronouns.

For spam detection in Arabic opinion texts an ensemble approach has been proposed by Saeed *et al.* (2019). A stacking ensemble classifier that combines a k-means classifier with a rule-based classifier outperforms the rest of the examined approaches. Both classifiers use content-based features, like n-grams. Given the lack of data sets for fake reviews in Arabic, the authors use for evaluation purposes the translated version of the data set of Ott *et al.* (2011, 2013). They also use this data set for the automatic labeling of a large data set of hotel reviews in Arabic (Elnagar *et al.* 2018). A supervised approach is also utilized for deceptive review detection in Persian (Basiri *et al.* 2019). In this work, POS tags, sentiment-based features, and metadata (e.g., number of positive/negative feedback, overall product score, review length, etc.) are exploited to construct and compare various classifiers (e.g., naive Bayes, SVMs, and decision trees). A data set with 3000 deceptive and truthful mobile reviews was gathered using customers reviews published in digikala.com. The labeling of the latter data set was performed by using a majority voting on the answers of 11 questions previously designed for spam detection by human annotators.

Last but not least, to the best of our knowledge, the only work toward the creation of *cross-cultural deception detection classifiers* is the work of Perez-Rosas et al. (2014; 2014). Similar to our work, country is used as a proxy for culture. Using crowdsourcing, the authors collected four deception data sets. Two of them are in English, originating from the United States and from

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 4. Overview of the used data sets. The corresponding columns are (a) data set, (b) culture, (c) language, (d) type, (e) origin, (f) collection process, (g) number of total, truthful, and deceptive documents, and (h) average length of words in truthful and deceptive documents. (T) stands for truthful and (D) for deceptive, while (I) stands for individualist cultures and (C) for collectivist cultures. Truthful documents tend to be longer than deceptive ones, except in Bluff and Russian collections

Dataset	Cult.	Lang.	Type	Origin	Process	#Docs			Average Length	
						T+D	T	D	T	D
OpSpam	I	en	reviews	written	turkers	1600	800	800	151.0	146.7
Boulder	I	en	reviews	written	turkers	1582	480	1102	114.9	93.9
DeRev	I	en	reviews	written	unsanctioned	236	118	118	128.5	125
Bluff	I	en	oral stories	transcript	sanctioned	267	89	178	173.4	207.0
EnglishUS	I	en	essays	written	turkers	600	300	300	79.5	62.5
CLiPS	I	nl	reviews	written	sanctioned	1298	649	649	143.8	133.7
EnglishIndia	C	en	essays	written	turkers	600	300	300	76.1	66.4
Russian	C	ru	essays	written	sanctioned	226	113	113	160.1	161.6
SpanishMexico	C	es	essays	written	sanctioned	346	172	174	95.6	64.8
Romanian	C	ro	essays	written	sanctioned	870	435	435	91.5	68.3
NativeEnglish	I	en	multi	multi	multi	4285	1787	2498	128.9	116.5

India, one in Spanish obtained from speakers from Mexico, and one in Romanian from people from Romania. Next, they built classifiers for each language using unigrams and psycholinguistic (based on LIWC) features. Then, they explored the detection of deception using training data originating from a different culture. To achieve this, they investigated two approaches. The first one is based on the translation of unigrams features, while the second one is based on the equivalent LIWC semantic categories. The performance, as expected, dropped in comparison with the within-culture classification and was similar for both approaches. The analysis for the psycholinguistic features showed that there are word classes in LIWC that only appear in some of the cultures, for example, classes related to time appear in English texts written by Indian people and in Spanish texts but not in the US data set. Lastly, they observed that deceivers in all cultures make use of negation, negative emotions, and references to others and that truth tellers use more optimism and friendship words, as well as references to themselves.

3. Data sets

We experimented with eleven data sets from six countries, namely United States, Belgium, India, Russia, Romania, and Mexico. We provide a detailed description of each data set below, while Table 4 provides some statistics and summarizes important information for each data set. We put much effort on the collection and the creation of the appropriate data sets. We wanted to experiment with fairly diverse cultures in terms of the degree of individualism/collectivism, having at the same time at our disposal basic linguistic tools and resources for the linguistic features extraction step.

In terms of the quantification of cultural diversity, we based our work on Hofstede's long-standing research on cultural differences (Hofstede 2001). Hofstede defined a framework that distinguishes six dimensions (power distance, individualism/collectivism, uncertainty avoidance,

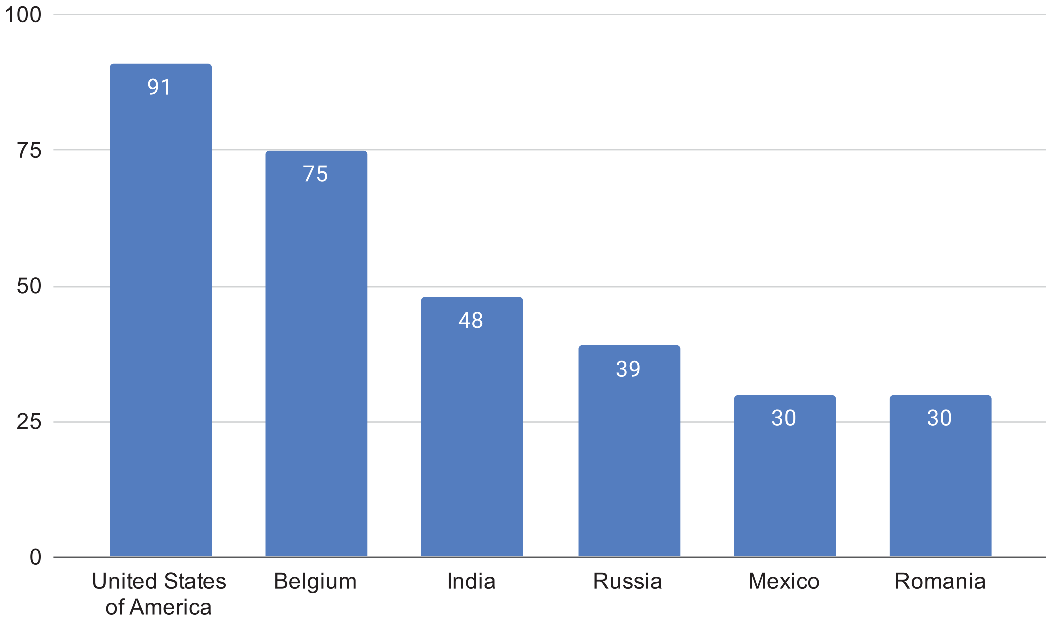

Figure 1. Differences between cultures along Hofstede's individualism dimension (source: https://www.hofstede-insights.com/product/compare-countries/).

masculinity/femininity, long-term/short-term orientation, and indulgence/restraint) along which cultures can be characterized. In his study, as in our work, country has been used as a proxy for culture. For each dimension, Hofstede's provides a score for each culture. Figure 1 depicts the cultural differences for the six aforementioned countries for the individualism dimension, which is the focus of our work. The individualism scores vary significantly, with United States possessing the highest one and both Mexico and Romania the lowest. We acknowledge that treating entire countries as single points along the individualism/collectivism dimension may be an over-simplification, especially for large countries. In the United States, for example, there is heterogeneity and diversity between regions (e.g., between Deep South and Mountain West) and even in the same region there may be different cultural backgrounds. However, the United States can be considered individualistic at an aggregate level, although there is a measurable variation on the value of this dimension (Vandello and Cohen 1999; Taras *et al.* 2016).

The creation of reliable and realistic ground truth data set for the deception detection task is considered a difficult task on its own (Fitzpatrick and Bachenko 2012). In our case, the selected corpora have been created using the traditional techniques for obtaining corpora for deception detection research, namely sanctioned and unsanctioned deception. Briefly, a sanctioned lie is a lie to satisfy the experimenter's instructions, for example, participants are given a topic, while an unsanctioned lie is a lie that is told without any explicit instruction or permission from the researcher, for example, diary studies and surveys in which participants recall lies already uttered. Crowdsourcing platforms, for example, Amazon Mechanical Turk[e], have also been used for the production of sanctioned content. In all sanctioned cases, a reward (e.g., a small payment) was given as a motivation. In addition, apart from the already existing data sets in the bibliography, we created a new data set (see Section 3.4) that concerns spoken text from transcripts of a radio game show.

[e] https://www.mturk.com

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

3.1 English – Deceptive Opinion Spam (OpSpam)

The OpSpam corpus[f] (Ott *et al.* 2011 2013) was created with the aim to constitute a benchmark for deceptive opinion spam detection and has been extensively used as such in subsequent research efforts. The authors approached the creation of the deceptive and truthful opinions in two distinct ways. First, they chose hotel reviews as their domain, due to the abundance of such opinions on the Web and focused on the 20 most popular hotels in Chicago and positive sentiment reviews. Deceptive opinions were collected by using Amazon Mechanical Turk. Quality was ensured by applying a number of filters, such as using highly rated turkers, located in the Unites States and allowing only one submission per turker. Based on these restrictions, 400 deceptive positive sentiment opinions were collected. Second, the truthful opinions were collected from TripAdvisor[g] for the same 20 hotels as thoroughly described in Ott *et al.* (2011). Only 5-star reviews were kept to collect reviews with positive sentiment, eliminating all non-English reviews, all reviews with less than 150 characters, and reviews of authors with no other reviews. This was an effort to eliminate possible spam from the online data. Then, 400 truthful comments were sampled to create a balanced data set. The same procedure was followed for negative sentiment reviews, by collecting 400 more deceptive opinions with negative sentiment through Amazon Mechanical Turk, and 400 truthful with 1 or 2 star reviews from various online sites. For more details, see Ott *et al.* (2013).

Human performance was assessed with the help of volunteers. They asked three untrained undergraduate university students to read and judge the truthfulness and deceptiveness of a subset of the acquired data sets. An observation from the results is that human deception detection performance is greater for negative (61%) rather than positive deceptive opinion spam (57%). But in both cases, automated classifiers outperform human performance.

In this work, we proceeded with the unification of these two data sets. The corpus contains:

- 400 truthful positive reviews from TripAdvisor (Ott *et al.* 2011),
- 400 deceptive positive reviews from Mechanical Turk (Ott *et al.* 2011),
- 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor, and Yelp (Ott *et al.* 2013),
- 400 deceptive negative reviews from Mechanical Turk (Ott *et al.* 2013).

3.2 English – Boulder Lies and Truth Corpus (Boulder)

Boulder Lies and Truth corpus[h] (Salvetti *et al.* 2016) was developed at the University of Colorado Boulder and contains approximately 1500 elicited English reviews of hotels and electronics for the purpose of studying deception in written language. Reviews were collected by crowdsourcing with Amazon Mechanical Turk. During data collection, a filter was used to accept US – only submissions (Salvetti 2014). The original corpus divides the reviews in three categories:

- Truthful: a review about an object known by the writer, reflecting the real sentiment of the writer toward the object of the review.
- Opposition: a review about an object known by the writer, reflecting the opposite sentiment of the writer toward the object of the review (i.e., if the writers liked the object they were asked to write a negative review, and the opposite if they did not like the object).
- Deceptive (i.e., fabricated): a review written about an object unknown to the writer, either positive or negative in sentiment.

This is one of the few available data sets that distinguish different types of deception (fabrications and lies). Since the data set was constructed via turkers, the creators of the data set took

[f] http://myleott.com/op-spam.html
[g] https://www.tripadvisor.com
[h] https://catalog.ldc.upenn.edu/LDC2014T24

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

extra care to minimize the inherent risks, mainly the tendency of turkers to speed up their work and maximize their economic benefit through cheating. More specifically, the creators implemented several methods to validate the elicited reviews, checking for plagiarism efforts and the intrinsic quality of the reviews. We unified the two subcategories of deception (fabrication and lie), since the focus of this work is to investigate deceptive cues without regard to the specific type of deception.

3.3 English – DeRev

The DeRev data set (Fornaciari and Poesio 2014) comprises deceptive and truthful opinions about books. The opinions have been posted on Amazon.com. This is a data set that provides "real life" examples on how language is used to express deceptive and genuine opinions, that is, this is an example of a corpus of unsanctioned deception. Without a doubt, manually detecting deceptive posts in this case is a very challenging task, since it is impossible to find definite proof that a review is truthful or not. For that reason a lot of heuristic criteria were employed and only a small subset of the collected data set that had high degree of confidence was accepted to be included in the gold standard data set. In more details, only 236 out of the 6819 reviews that were collected (118 deceptive and 118 truthful) constituted the final data set. The starting point for identifying the deceptive and genuine clues that define the heuristic criteria was a series of articles[i, j,k,l] with suggestions and advice about how to unmask a deceptive review in the Web, as well as specific incidents of fake reviews that have been disclosed. Such clues are the absence of information about the purchase of the reviewed book, the use of nicknames, reviews that have been posted for the same book in a short period of time, and a reference to a suspicious book (i.e., a book whose authors have been accused of purchasing reviews, or have admitted that they have done so). The truthfulness of the reviews was identified in a similar manner by reversing the cues. We performed a manual inspection, which confirmed that all of the 113 reviewers of the 236 reviews we used (excluding 8 reviewers whose accounts were no longer valid) had submitted at least one review marked by the platform as having been submitted in the United States. Hence, it is reasonable to assume that the vast majority of the reviewers were US-based.

3.4 English – Bluff The Listener (Bluff)

The "Wait Wait. . . Don't Tell Me!" is an hour-long weekly radio news panel game show produced by Chicago Public Media and National Public Radio (NPR)[m] that airs since 1998. One of the segments of this show is called "Bluff the Listener" in which a contestant listens to three thematically linked news reports from three panelists, one of which is truthful and the rest are fictitious. Most of the stories are humorous and somewhat beyond belief, for example, a class to teach your dog Yiddish. The listener must determine the truthful story in order to win a prize, whereas at the same time the panelist that is picked is awarded with a point to ensure the motivation for all the participants. An archive of transcripts of this show is available since 2007 in the official web page of the show. We used these transcripts and we managed to retrieve and annotate 178 deceptive and 89 truthful stories. Consequently, we collected the participant's replies to calculate the human success rate. Interestingly, the calculated rate was about 68%, which is quite high since in experimental studies of detecting deception, the accuracy of humans is typically only slightly better than

[i] http://www.guardian.co.uk/books/2012/sep/04/sock-puppetry-publish-be-damned

[j] https://www.moneytalksnews.com/3-tips-for-spotting-fake-product-reviews—from-someone-who-wrote-them/

[k] http://www.nytimes.com/2011/08/20/technology/finding-fake-reviews-online.html

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

[m] https://www.npr.org/programs/wait-wait-dont-tell-me

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

chance, mainly due to *truth bias* as previously mentioned. This might be attributed to the fact that the panelists of the show have remained almost the same, and as a result the listeners might have learned their patterns of deception over time. In addition, we have to stress that the intent of the panelists to deceive is intertwined with their intent to entertain and amuse their audience. Hence, it is interesting to examine if the linguistic cues of deception can be distorted by this double intent, and if they still suffice to discriminate between truth and deception even in this setting.

3.5 English/Spanish/Romanian – Cross-cultural deception

To the best of the authors' knowledge, this is the only available multicultural data set constructed for cross-cultural deception detection[n] (Pérez-Rosas and Mihalcea 2014; Pérez-Rosas *et al.* 2014). It covers four different languages, EnglishUS (English spoken in the US), EnglishIndia (English spoken by Indian people), SpanishMexico (Spanish spoken in Mexico), and Romanian, approximating culture with the country of origin of the data set. Each data set consists of short deceptive and truthful essays for three topics: opinions on abortion, opinions on death penalty, and feelings about a best friend. The two English data sets were collected from English speakers using Amazon Mechanical Turk with a location restriction to ensure that the contributors are from the country of interest (United States and India). The Spanish and Romanian data sets were collected from native Spanish and Romanian speakers using a web interface. The participants for Spanish and Romanian have been recruited through contacts of the paper's authors. For all data sets, the participants were asked first to provide their truthful responses, and then their deceptive ones. In this work, we use all the available individual data sets. We detected a number of spelling errors and some systematic punctuation problems in both English data sets, with the spelling problems to be more prevalent in the EnglishIndia data set. To this end, we decided to correct the punctuation errors, for example, "kill it.The person", in a preprocessing step in both data sets. Regarding the spelling errors, we found no correlation between the errors and the type of text (deceptive, truthful), and since the misspelled words were almost evenly distributed among both types of text, we did not proceed to any correction.

3.6. Dutch – CLiPS stylometry investigation (CLiPS)

CLiPS Stylometry Investigation (CSI) corpus (Verhoeven and Daelemans 2014) is a Dutch corpus containing documents of two genres namely essays and reviews. All documents were written by students of Linguistics & Literature at the University of Antwerp[o], taking Dutch proficiency courses for native speakers, between 2012 and 2014. It is a multipurpose corpus that serves in many stylometry tasks such as detection of age, gender, authorship, personality, sentiment, and deception, genre. The place that authors grew up is provided in the metadata. On this basis, it is known that only 11.2% of the participants grew up outside Belgium, with the majority of them (9.7% of the total authors) grown up in the neighboring country of the Netherlands.

This corpus, which concerns the review genre, contains 1298 (649 truthful and 649 deceptive) texts. All review texts in the corpus are written by the participants as a special assignment for their course. Notice that the participants did not know the purpose of the review task. For the collection of the reviews students were asked to write a convincing review, positive or negative, about a *fictional* product while the truthful reviews reflect the authors real opinion on an existing product. All the reviews were written about products from the same five categories: smartphones, musicians, food chains, books, and movies.

[n] http://web.eecs.umich.edu/~mihalcea/downloads.html#CrossCulturalDeception

[o] The city of Antwerp is the capital of Antwerp province in the Flemish Region of the Kingdom of Belgium.

3.7. Russian – Russian Deception Bank (Russian)

For the Russian language, we used the corpus of the rusProfilingLab[p] (Litvinova *et al.*, 2017). It contains truthful and deceptive narratives written by the same individuals on the same topic ("How I spent yesterday" etc.). To minimize the effect of the *observers paradox*[q], researchers did not explain the aim of the research to the participants. Participants that managed to deceive the trained psychologist who evaluated their responses were rewarded with a cinema ticket voucher. The corpus consists of 113 deceptive and 113 truthful texts, written by 113 individuals (46 males and 67 females) who were university students and native Russian speakers. Each corpus text is accompanied by various metadata such as gender, age, and results of a psychological test.

3.8. English – Native English (NativeEnglish)

Finally, we combined all the data sets that were created from native English speakers (i.e., OpSpam, Boulder, DeRev, Bluff, and EnglishUS) in one data set. The idea is to create one multidomain data set, big enough for training, where the input is provided by native speakers.

4. Features

In this section, we detail the feature selection and extraction processes. Furthermore, we explicitly define the features that we exploited for pinpointing differences between cultures.

4.1. Features extraction

We have experimented with three feature types along with their combinations, namely a plethora of linguistic cues (e.g., word counts, sentiment, etc.), various types of n-grams, and token embeddings. Linguistic indicators are extracted based on prior work, as already analyzed in Sections 2.3 and 2.4. Further, we have evaluated various types of n-grams in order to identify the most discriminative ones. The use of n-grams is among the earliest and more effective approaches for the task of deception detection. Ott *et al.* (2011) and Fornaciari *et al.* (2013) were among the first to use word n-grams for deception detection, while character n-grams and syntactic n-grams (defined below) have been used by Fusilier *et al.* (2015) and Feng *et al.* (2012), respectively. Lastly, due to the absence of a large training corpus, we tried to combine feature engineering and statistical models, in order to enhance the overall performance and get the best of both worlds. This approach is in line with recent research on deception detection that tries to leverage various types of features (Bhatt *et al.* 2018; Krishnamurthy *et al.* 2018; Siagian and Aritsugi 2020).

4.2. Linguistic cues

Table 5 presents the complete list of features for each language explored in this work. These features count specific cues in text, aiming to capture characteristics of deceptive and truthful language. These indicators have been conceptually divided into six categories, namely word counts, phoneme counts, pronoun use, sentiment, cognitive complexity, and relativity. The absence of a tick in Table 5 marks the inability to extract the specific feature, given the available linguistic tools and resources for each language while the "N/A" marks the nonexistence of the particular feature in the specific language, that is, articles in Russian.

Although we believe that most feature names are self-explanatory, we have to describe further the #hedges and #boosters features. Hedges is a term coined by the cognitive linguist George

[p] http://en.rusprofilinglab.ru/korpus-tekstov/russian-deception-bank

[q] According to Labov (1972), "the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation."

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 5. The list of used features. Features that are examined in Taylor's work (2014, 2017) are marked with an asterisk (∗). The dot (•) marks nonnormalized features. Absence of a tick marks the inability to extract this specific feature for this particular language. The N/A indicates that this feature is not applicable for this particular language

Features	English	Dutch	Russian	Spanish	Romanian
1. Linguistic cues					
a. Word counts					
Average word length •	✓	✓	✓	✓	✓
#adjectives & #adverbs	✓	✓	✓	✓	✓
#articles	✓	✓	N/A	✓	✓
#boosters	✓				
#filled pauses	✓				
#function words	✓	✓	✓	✓	✓
#hedges	✓				
#lemmas •	✓	✓	✓	✓	✓
#negations∗	✓	✓	✓	✓	✓
#prepositions	✓	✓	✓	✓	✓
#punctuation marks •	✓	✓	✓	✓	✓
#vague words	✓				
#verbs	✓	✓	✓	✓	✓
#words •	✓	✓	✓	✓	✓
b. Phoneme counts					
#fricatives	✓	✓	✓	✓	✓
#nasals	✓	✓	✓	✓	✓
#plosives	✓	✓	✓	✓	✓
c. Pronoun use					
#total pronouns∗	✓	✓	✓	✓	✓
#1st person pronouns∗	✓	✓	✓	✓	✓
#1st person pronouns (singular)	✓	✓	✓	✓	✓
#1st person pronouns (plural)	✓	✓	✓	✓	✓
#3rd person pronouns∗	✓	✓	✓	✓	✓
#demonstrative pronouns	✓	✓	✓	✓	✓
#indefinite pronouns	✓	✓	✓	✓	✓
d. Sentiment					
positive|negative -SentiWordNet∗	✓				
positive|negative -MPQA∗	✓				
positive|negative -FBS∗	✓				
sentiment-ANEW∗	✓				

Table 5. Continued

Features	English	Dutch	Russian	Spanish	Romanian
positive|negative -Spanish lexicon∗				✓	
positive|negative -VU-sentiment lexicon∗		✓			
positive|negative -RuSentiLex∗			✓		
positive|negative -RoSentiLex∗					✓
e. Cognitive complexity					
mean sentence length ●	✓	✓	✓	✓	✓
mean preverb length ●	✓				
#conjunction words	✓	✓	✓	✓	✓
#subordinate clauses	✓				
f. Relativity					
#exclusion words	✓				
#modal verbs	✓				
#motion verbs	✓	✓	✓		
#spatial words∗	✓	✓	✓	✓	✓
#verbs in future tense	✓		✓	✓	
#verbs in past tense	✓	✓	✓	✓	✓
#verbs in present tense	✓	✓	✓	✓	✓
2. N-grams *(N=uni,bi,tri,uni+bi,bi+tri,un+bi+tri)*					
Phoneme n-grams	✓	✓	✓	✓	✓
Character n-grams	✓	✓	✓	✓	✓
Word n-grams	✓	✓	✓	✓	✓
POS n-grams	✓	✓	✓	✓	✓
Syntactic n-grams	✓				
3. BERT Embeddings					
BERT embeddings	✓	✓	✓	✓	✓
4. Mixture					
N-grams & linguistic cues	✓	✓	✓	✓	✓
BERT embeddings and linguistic cues	✓	✓	✓	✓	✓

Lakoff (1973) to describe words expressing some feeling of doubt or hesitancy (e.g., guess, wonder, reckon etc.). On the contrary, boosters are words that express confidence (e.g., certainly, apparently, apparent, always). Both are believed to correlate either positively or negatively with deception and thus are frequently used in related research work (Bachenko *et al.* 2008). Regarding the important feature of pronouns, we consider first person pronouns in singular and plural form, for example, I versus we, mine versus ours, etc., third person pronouns, for example, they,

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

indefinite pronouns, for example, someone, anyone, etc., demonstrative pronouns (e.g., this, that, etc.), and the total number of pronouns. The linguistic tools used for the extraction of the features, for example, POS taggers, named entity recognition tools, etc., are shown in Table 8. Some of the features were extracted with handcrafted lists authored or modified by us. Such features include filled pauses (e.g., ah, hmm etc.), motion verbs, hedge words, boosters, etc.

Table 7 lists the sentiment analysis tools used for each language. We exploited, whenever possible, language-specific sentiment lexicons used in the bibliography and avoided the simple solution of automatically translating sentiment lexicons from American English. Related research (Mohammad *et al.* 2016) has shown that mistranslation (e.g., positive words translated as having neutral sentiment in the target language), cultural differences, and different sense distributions may lead to errors and may insert noise when translating sentiment lexicons. Analogously, we maintained the same practice for the rest of the features. When this was not feasible, we proceeded with just the translation of linguistic resources (mostly for the Russian language). For the *#spatial words* feature that counts the number of spatial references in text, we followed a two-step process. We employed a combination of a named entity recognizer (NER) tool (see Table 8) and spatial lexicons for each language. The lexicons, principally gathered by us, contain spatially related words (e.g., under, nearby, etc.) for each language, while the named entity recognizer extracts location related entities from the corpora (e.g., Chicago, etc.). In the case of the English language, the existence of a spatial word in the text was computed using a dependency parse, in order to reduce false positives. The final value of this feature is the sum of the two values (spatial words and location named entities). For Romanian, we had to train our own classifier based on Conditional Random Fields (CRFs) (Lafferty *et al.* 2001; Finkel *et al.* 2005) by using as training corpus the RONEC (Dumitrescu and Avram 2020b) corpus, an open resource that contains annotated named entities for 5127 sentences in Romanian.

The values of the features were normalized depending on their type. For example, the *#nasals* feature was normalized by dividing with the total number of characters in the document, while the *#prepositions* with the number of tokens in the document. The features *#words*, *#lemmas*, *#punctuaction marks*, *average word length*, *mean sentence length*, and *mean preverb length* were left nonnormalized. For each sentiment lexicon, except for ANEW, we computed the score by applying the following formula to each document d of $|d|$ tokens and each sentiment s (positive|negative):

$$\text{sentiment\_score}(d, s) = \frac{\sum_{w \in d} \text{sentiment\_strength}(w, s)}{|d|}$$

The *sentiment_strength* for SentiWordNet is a value in the interval [0,1] while for the rest sentiment resources the values are either 0 or 1.

For the ANEW lingustic resource (Bradley and Lang 1999) that rates words in terms of pleasure (affective valence), arousal, and dominance with values from 0 to 10, we only considered the normalized valence rating that expresses the degree of positivity or negativity of a word. The applying formula in this case is

$$\text{ANEW-sentiment\_score}(d) = \frac{\sum_{w \in d} (\text{ANEW\_valence}(w) - 5)}{|d| \cdot 5}$$

Lastly, we included phoneme-related features in our analysis. Our hypothesis was that phonological features, captured by phonemes for text, will be more discriminative in spoken data sets, since the deceiver will put extra care to sound more truthful to the receiver, even subconsciously. This hypothesis is in line with an increasing volume of work that investigates the existence of non-arbitrary relations between phonological representation and semantics. This phenomenon is known as *phonological iconicity* and links a word's form with the emotion it expresses (Nastase *et al.* 2007; Schmidtke *et al.* 2014). Table 6 summarizes such representative works.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 6. Phoneme connection to sentiment in phonological iconicity studies

Work	Description
Fnagy (1961)	This early study on Hungarian poems showed that sonorants (e.g., /l/, /m/) occur more often in tender, but plosives (e.g., /k/, /t/) more often in aggressive poems.
Taylor and Taylor (1965)	Evidence that pleasantness relations are language specific.
Zajonc *et al.* (1989)	Passages about Hell from Miltons “Paradise Lost” were found to contain significantly more front vowels and hard consonants than passages about Eden while the latter contained more medium back vowels.
Whissell (1999)	The analysis of phonemes in different sources (song lyrics, poetry, word lists, advertisements) shows that plosives correlate with unpleasant words.
Auracher *et al.* (2010)	Multilingual analysis on poems found that plosive sounds are more likely to express a pleasant mood, whereas a relatively high frequency of nasal sounds indicates an unpleasant mood. Universality is claimed since the authors found the same trend independently of the language.
Kraxenberger and Menninghaus (2016)	This work failed to reproduce the results of Auracher *et al.* (2010).
Papantoniou and Konstantopoulos (2016)	The analysis of names of movie fictional characters showed among other findings the connection of the nasals with negative sentiment.

Table 7. Sentiment lexicons used for each language

en	ANEW (Bradley and Lang 1999)	Normative emotional ratings for 3188 words. It provides values in respect to pleasure, arousal, and dominance of each term.
	FBS (Hu and Liu 2004)	6786 words (2006 positive and 4783 negative).
	MPQA (Riloff and Wiebe 2003; Wilson *et al.* 2005)	Each word annotated for intensity (strong, weak). 6885 words (2718 positive and 4912 negative).
	SentiWordNet (Baccianella *et al.* 2010)	All WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness.
nl	VU-sentiment-lexicon (Maks *et al.* 2014)	9237 words (3314 positive and 5923 negative).
ru	RuSentiLex (Loukachevitch and Levchik 2016)	Lexicon generated through semi-automatic techniques, which contains 16,057 words (10,227 negative, 3770 positive, 1747 neutral and 291 either positive or negative based on context).
es	Spanish Sentiment Lexicon (Pérez-Rosas *et al.* 2012)	It provides two polarity lexicons: a. an automatically generated with 2496 concepts and b. a semi-automatically generated with 1347 concepts. We employed the semi-automatically generated lexicon since it is the one with largest reported accuracy of approximately 90%.
ro	RoSentiLex	We translated the MPQA lexicon by using a bilingual Romanian-English dictionary (Mihalcea 2014).

http://www.cs.uic.edu/liub/FBS/opinion-lexicon-English.rar
http://mpqa.cs.pitt.edu/#subj_lexicon
https://sentiwordnet.isti.cnr.it
https://github.com/opener-project/VU-sentiment-lexicon
http://www.labinform.ru/pub/rusentilex/index.htm

Table 8. Linguistic tools used on each language for the extraction of features

Tool	English	Dutch	Russian	Spanish	Romanian
Phonemes	Espeak-ng	Espeak-ng	Espeak-ng	Espeak-ng	Espeak-ng
Lemmatizer	Stanford CoreNLP	–	OpenNLP	–	–
Stemmer	Stanford CoreNLP	snowball	snowball	Stanford CoreNLP	snowball
POS Tagger	Stanford CoreNLP	OpenNLP	TreeTagger	Stanford CoreNLP	TreeTagger
NER	Stanford CoreNLP	OpenNLP	DBPedia Spotlight	Stanford CoreNLP	custom
Syntactic Parser	Stanford CoreNLP	–	–	–	–

Table 9. Examples of n-gram features

	Phoneme	Character	Word	POS	Syntactic
unigram	l' ɛvəl [level]	h, o, t, e, l	abortion, new	EX, IN, NNP, NNPS	ccomp, xcomp
bigram	nj' uː l' ɛvəl [new level]	ho, ot, te, el	new hotel	VBN DT, VBG RP	root-nsubj, root-aux
trigram	nj' uː l' ɛvəl ɑt [new level at]	hot, ote, tel	a business trip	NN NN VBZ	aux-advmod-root

4.3. N-grams

We have evaluated several variations of n-grams from various levels of linguistic analysis to encode linguistic information. Given the diversity of the data sets, we used different types of n-grams to identify those that are more effective in discriminating deceptive and truthful content. For each n-gram type and for each data set, we extracted unigrams, bigrams, trigrams, unigrams+bigrams, bigrams+trigrams, and unigrams+bigrams+trigrams. Some examples are shown in Table 9.

- *Phoneme n-grams:* These features were extracted from the phonetic representation of texts derived by applying the spelling-to-phoneme module of the espeak-ng speech synthesizer (see Table 8). We examined phoneme n-grams at the level of words.
- *Character n-grams:* Consecutive characters that can also belong to different words.
- *Word n-grams:* We examined versions with and without stemming and stopword removal.
- *POS n-grams:* POS n-grams are contiguous part-of-speech tag sequences, such as adjective-noun-verb, noun-verb-adverb, and so on, that provide shallow grammatical information. We extracted POS n-grams using the appropriate POS-tagger for each language (see Table 8).
- *Syntactic n-grams:* syntactic n-grams (sn-grams) are constructed by following all the possible paths in dependency trees and keeping the labels of the dependencies (arcs) along the paths. We used Stanford's CoreNLP syntactic parser for the construction of dependency trees for the English data sets (see Table 8).

4.4. BERT embeddings

Regarding token embeddings, we used the contextualized embeddings from the BERT (Devlin *et al.* 2019) model. BERT, which stands for Bidirectional Encoder Representations from Transformers, is a language model based on a stack of transformer encoder layers pretrained on a large unlabeled cross-domain corpus using masked language modeling and next-sentence prediction objectives. Since its introduction, BERT has achieved state-of-the-art results in many NLP tasks. In most cases, the best results are obtained by adding a shallow task-specific layer (e.g., a linear classifier) on top of a pretrained BERT model, and *fine-tuning* (further training) the pretrained BERT model jointly with the task-specific layer on a labeled task-specific data set.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 10. BERT pretrained models used for each language

Lang.	Name	Description
en	bert-base-uncased (Devlin *et al.* 2019)	12-layer, 768-hidden, 12-heads, 110M parameters, 30K wordpieces
nl	BERTje cased (de Vries *et al.* 2019)	12-layer, 768-hidden, 12-heads, 110M parameters, 30K wordpieces
ru	RuBERT cased (Kuratov and Arkhipov 2019)	12-layer, 768-hidden, 12-heads, 180M parameters, 120K wordpieces
es	BETO uncased (Caete *et al.* 2020)	12-layer, 1024-hidden, 16-heads, 110M parameters, 30K wordpieces
ro	bert-base-romanian-cased (Dumitrescu and Avram 2020a)	12-layer, 768-hidden, 12-heads, 1M parameters, 50K wordpieces
multi	bert-base-multilingual-cased	12-layer, 768-hidden, 12-heads, 110M parameters, 110K wordpieces

https://github.com/google-research/bert
https://github.com/wietsedv/bertje
http://docs.deeppavlov.ai/en/master/features/models/bert.html
https://github.com/dccuchile/beto
https://github.com/dumitrescustefan/Romanian-Transformers

In effect, each encoder layer of BERT builds token embeddings (dense vectors, each representing a particular token of the input text). The token embeddings of each encoder layer are revised by the next stacked encoder layer. A special classification embedding ([CLS]) is also included in the output of each layer, to represent the entire input text. In classification tasks, typically the [CLS] embedding of the top-most encoder layer is passed on to the task-specific classifier, which in our case decides if the input text is deceptive or not. We explore this approach in Section 6.2. We note that BERT uses a WordPiece tokenizer[r] (Schuster and Nakajima 2012), which segments the input text in tokens corresponding to character sequences (possibly entire words, but also subwords or even single characters) that are frequent in the large corpus BERT is pretrained on. We also note that BERT's token embeddings are context-aware, that is, different occurrences of the same token receive different embeddings when surrounded by different contexts. In Table 10, we provide details about the used BERT models. We exploit pretrained models on each language, as well as the multilingual BERT model, which is pretrained over Wikipedia in 104 languages.

5. Statistical evaluation of linguistic cues

In this section, we conduct a statistical analysis of the linguistic cues (see Section 4.2) per data set. In more details, we conduct a Mann–Whitney U test to identify the statistically significant linguistic features of each data set (the NativeEnglish data set is the unified data set of all native English speakers data sets). Afterward, we apply a multiple logistic regression (MLR) analysis over the statistically important features of each data set. This test shows the distinguishing strength of the important linguistic features. We discuss the results for each data set/culture and try to provide some cross-cultural observations.

5.1. *Statistical significance analysis*

Since we cannot make any assumption about the distribution of the feature values in each data set, we performed the nonparametric Mann–Whitney U test (two-tailed) with a 99% confidence

[r] Consult also https://huggingface.co/transformers/master/tokenizer_summary.html

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

interval and $\alpha = 0.01$. The null hypothesis (H0) to be refuted is that there is no statistically significant difference between the mean rank of a feature for texts belonging to the deceptive class and the mean rank of the same feature for texts belonging to the truthful class. The results are available in the Appendix Tables 31 and 32. Below we summarize the main observations.

1. No statistically significant features were found in the Russian collection and as a result we ignore this data set in the rest of this analysis. This is probably due to the inappropriateness of the selected features and/or the shortage of language resources for the Russian language, or even because of the intrinsic properties and peculiarities of the data set itself. This suggests that we cannot come to any conclusion about how the linguistic features are used in this data set and compare it with the rest.
2. Statistically significant differences were found in most of the data sets for the features: *#lemmas*, *#words*, and *#punctuation*. In more details:
   - The importance of *#lemmas* is observed in most of the data sets. A large number of lemmas seems to be a signal for truthful texts in most of the examined data sets, with the exception of the DevRev and Bluff data sets, where a large number of lemmas is a signal for deceptive texts. These two data sets are quite distinct from the rest, since the former is an example of unsanctioned deception, while the latter concerns transcriptions of spoken data with notably stylistic elements like humor and paralogism. Although, we cannot characterize it as a universal feature, since it is not observed in the Russian data set, it is a *language-agnostic* cue that seems to be employed across most cultures.
   - The same observations hold also for the feature *#words*, with the exception that it is not statistically significant for the OpSpam dataset.
   - Regarding the *#punctuation* feature, it is rather important for all data sets except for Bluff and DeRev. Since Bluff is a data set created from transcripts, the transcription process might shadow the intonation and emotional status of the original agent with the idiosyncrasies of the transcriber/s, for example, there are almost zero exclamations. Furthermore, the use of punctuation, except in DeRev and Bluff, is an indication of truthful text.
3. An observation of possibly cultural origin is the fact that *sentiment-related features*, positive or negative, are notably important for the individualist cultures (US and Dutch). The expression of more positive sentiment vocabulary is linked with the deceptive texts, while negative sentiment is linked to truthful text, except in the EnglishUS case, where the negative sentiment is related to the deceitful texts. For the collectivistic cultures that are more engaged in the high context communication style, sentiment-related features are not distinguishing. As explained earlier, the effort to restrain sentiment and keep generally friendly feelings toward the others in order to protect social harmony might be responsible for this difference. Our findings contradict Taylor's results and are in agreement with his original hypothesis and related studies like Seiter *et al.* (2002) (see Section 2.3).
4. Another important finding of our experiments is that in almost all data sets, the formulation of sentences in *past tense* is correlated with truth, while in *present tense* with deception, independently of the individualistic score of the corresponding culture. This can be attributed to the process of recalling information in the case of truthful reviews or opinions. In the case of deception, present tense might be used due to preference to simpler forms, since the deceiver is in an already overloaded mental state. In the US data sets, the only exceptions are the Bluff and the OpSpam data sets, where we observe the opposite. However, in the OpSpam data set, these two features are not statistically significant.
5. Furthermore, the *#modal verbs* is important in the US data sets. Specifically, an increased usage of modal verbs usually denotes a deceptive text.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

6. Another cross-cultural observation correlated with the degree of individualism is the *#spatial words* feature. Specifically, for the data sets where this feature is important, we observe a difference in the frequency of spatial details for the deceptive texts in the collectivist data sets and the truthful texts in the individualistic ones. In detail, more spatial features are linked with deception for the Romanian and SpanishMexico data sets, while their frequency is balanced in the case of Dutch and diverges to truthful text for the NativeEnglish data set. These observations are in agreement with Taylor (see Table 3). On top of that, discrepancies in the quantity of spatial details have also been found in different modalities (Qin *et al.* 2005). More specifically, deceivers had significantly fewer spatial details than truth-tellers in audio but more in text. This signifies how sensitive this linguistic cue is not only across cultures but also when other parameters such as context or modality vary.
7. Regarding the *#pronouns*, our results show mixed indications about their usage that do not fully agree with Taylor. Notice though that we had only limited tool functionality for pronoun extraction (i.e., no tools for Dutch and SpanishMexico). As a result, we created our own lists for English and used translations for the other languages. Generally, pronouns in various forms seem to be important in most data sets. *Third person pronouns* are correlated with deceptive texts mainly in EnglishUS and less in the Romanian and EnglishIndia data sets, all of which belong to the same cross-cultural opinion deception detection data set and to truthful ones in the Boulder data set (with a rather small difference though). This is in partial agreement with Taylor's results, where third-person pronouns are linked with deception in collectivist languages. Regarding *first-person pronouns*, the observations show mixed results. They are linked with both truthful and deceptive text, in the latter case though only for individualistic data sets (i.e., Bluff and OpSpam). Exploring the use of singular and plural forms sheds a bit more light, since the plural form is linked with truthful text in both collectivistic and individualistic cultures, except in Dutch where the plural form slightly prevails for deceptive. Finally, *indefinite* and *demonstrative* pronouns are rarely important.
8. The *#nasal* feature that counts the occurrences of /m/, /n/ and in some languages /ŋ/ in texts is rather important for the highly collective SpanishMexico and Romanian data sets. It prevails in truthful texts while we observe the opposite in the individualistic NativeEnglish. This is an interesting observation that enriches the relevant research around nasals. Generally, there are various studies (see Table 6) that claim a relation between the occurrence of consonants and the emotion of words based on the physiology of articulation for various languages. Most of the studies link nasals with sadness and plosives with happiness, although there are other studies contradicting these results (see Table 6). Furthermore, nasals have been connected with different semantic classes like iconic mappings, size, and affect as shown by Schmidtke *et al.* (2014). Finally, notice that plosives are not statistically significant in our results. We believe that this is a direction that needs further research with larger data sets and more languages.
9. Finally, the *#filled pauses* feature, which was incorporated to showcase differences between written and oral deception cues, does not provide any remarkable insight.

A collateral resulting observation is that most of the distinguishing features do not require complex extraction processes but only surface processing like counts on the token level.

EnglishUS and EnglishIndia data sets comparison

The EnglishUS and EnglishIndia data sets are ideal candidates to examine individualism-based discrepancies in linguistic deception cues by keeping the language factor the same. These two data sets are part of the Cross-Cultural Deception data set (see Section 3.5) and were created using the same methodology. Both contain opinions on the same topics (death penalty, abortion, best friend), in the same language, and come from two cultures with a large difference in terms of the individualism index score (91 vs. 48). Initially, to explore differences in the authors writing competence we computed the Flesch reading-ease score[s] (Kincaid *et al.* 1975) on both data sets. The scores are similar (63.0 for the EnglishIndia data set and 63.6 for EnglishUS data set) and correspond to the same grade level. Notice though that based on Tables 31 and 32 in the Appendix, the native speakers use larger sentences and more subordinate clauses. A possible explanation is that since Indians are not native speakers of English, they might lack in language expressivity and use English similarly whether they are telling the truth or lying.

A crucial observation though is the limited number of statistically important features in the case of the EnglishIndia (only 3) compared to the EnglishUS (15). Furthermore, the pronoun usage differs a lot between the two data sets. In more details, the individualist group employs more *#1st person pronouns* in truthful text, while in the case of the EnglishIndia first person pronouns are not important. In the case of *#3rd person pronouns,* both data sets use the same amount of pronouns with a similar behavior. As already mentioned, this might be a difference of cultural origin, since individualist group deceivers try to distance themselves from the deceit, while in the collectivist group deceivers aim to distance their group from the deceit. Finally, we notice again the importance of the sentiment cues for the native English speakers and their insignificance in the EnglishIndia data set, which correlates to our previous observations. For the remaining features, it is risky to make any concluding statements in relation to cultural discrepancies.

5.2. Multiple Logistic Regression (MLR) analysis

To further examine the discriminative ability of the linguistic features and explore their relationship, we conducted a *multiple logistic regression* (MLR) analysis on the resulting significant features from the Mann–Whitney U test. The null hypothesis is that there is no relationship between the features and the probability of a text to be deceptive. In other words, all the coefficients of the features are considered equal to zero for the dependent variable.

Since MLR presupposes uncorrelated independent variables, we keep only the most significant feature for any set of correlated and dependent features for each data set and manually filter out the rest. For example, we keep only the single most important positive or negative sentiment feature per data set (e.g., in English where we use various lexicons). Also in the case of features that are compositions of more refined features, we keep the most refined ones when all of them are important, for example, we keep the feature pair *#first person pronouns (singular)* and *#first person pronouns (plural)* instead of the more general *#first person pronouns*. Overall, we cannot guarantee that there is no correlation between the features.

In Tables 11 and 12, we present the results of the MLR analysis, reporting the features with p-value < 0.1, for the native English and the cross-language cases, respectively. For each feature in the table, we report the corresponding coefficient, the standard error, the z-statistic (Wald z-statistic), and the p-value. Higher coefficient values increase the odds of having a deceptive text in the presence of this specific feature, while lower values increase the odds of having a truthful text. The Wald (or z-value) is the regression coefficient divided by its standard error. The larger magnitude (i.e., either too positive or too negative) indicates that the corresponding regression coefficient is not 0 and the corresponding feature matters. Generally, there are no features participating in all functions both in the context of the native English data sets and across different languages and cultures, an indication of how distinct the feature sets are both within and across cultures. Among different languages is difficult to conclude how the characteristics of each language (e.g., pronoun-drop languages) and/or the different extraction processes (e.g., sentiment lexicons) affect the analysis. A more thorough analysis is safer to be performed in the context of the same language though. Below, we report some observations from this analysis.

[s] This reading-ease score is based on the average length of the words and sentences.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 11. Multiple logistic regression analysis on linguistic features for each US data set. SE stands for standard error. We show in bold *p*-values < 0.01. Positive or negative estimate values indicate features associated with deceptive or truthful text, respectively

OpSpam	Estimate	SE	Wald	*p*-value
#1st person pronouns (singular)	24.411	4.116	5.931	**∼ 0**
#boosters	19.456	9.293	2.094	0.036
#hedges	12.461	4.76	2.618	**0.009**
positive sentiment-MPQA	11.056	1.684	6.564	**∼ 0**
#nasals	9.773	5.747	1.7	0.089
#fricatives	8.432	4.611	1.829	0.067
#verbs	8.324	2.814	2.958	**0.003**
mean preverb length	0.192	0.034	5.628	**∼ 0**
mean sentence length	0.121	0.018	6.546	**∼ 0**
#punctuation marks	–1.155	0.109	–10.594	**∼ 0**
#spatial words	–2.138	0.794	–2 691	**0.007**
intercept	–4.749	1.189	–3 995	**∼ 0**
Boulder	Estimate	SE	Wald	*p*-value
#modal verbs	8.84	2 984	2 963	**0 003**
positive sentiment-MPQA	5.088	1 366	3 723	**∼ 0**
#lemmas	–0.014	0 009	–1 665	0 096
#punctuation marks	–0.16	0 081	–1 979	0 048
#articles	–6.273	1 946	–3 224	**0 001**
DeRev	Estimate	SE	Wald	*p*-value
#lemmas	0.05	0 021	2 433	0 015
#words	–0.025	0 01	–2 453	0 014
#pronouns	–13.319	4 077	–3 267	**0 001**
#articles	–18.203	5 177	–3 516	**∼ 0**
EnglishUS	Estimate	SE	Wald	*p*-value
#3rd person pronouns	13.5	3.421	3.946	**∼ 0**
#verbs	6.532	3.38	1.933	0.053
intercept	2.838	1.686	1.683	0.092
#words	–0.026	0.012	–2.08	0.038
#punctuation marks	–0.462	0.182	–2.533	0.011
#1st person pronouns (plural)	–8.493	4.207	–2.019	0.043
#prepositions	–9.008	3.276	–2.75	**0.006**
#indefinite pronouns	–10.642	3.762	–2.829	**0.005**

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 11. Continued

OpSpam	Estimate	SE	Wald	*p*-value
#1st person pronouns (singular)	–22.886	3.982	-5.748	**∼ 0**
sentiment-ANEW	16.621	9.272	1.793	0.073
#function words	–13.126	4.825	–2.72	**0.007**
#demonstrative pronouns	–26.514	13.569	–1.954	0.051
Bluff	Estimate	SE	Wald	*p*-value
sentiment-ANEW	16.621	9.272	1.793	0.073
#function words	–13.126	4.825	–2.72	**0.007**
#demonstrative pronouns	–26.514	13.569	–1.954	0.051

Native English data set observations

For the native English language data sets shown in Table 11, we observe high coefficients for the various types of pronouns (especially of the first person ones). Although there is no clear indication about their direction, most of the times they are associated with truthful text. The only exceptions are the *#1st person pronouns (singular)* in the case of OpSpam, the *#demonstrative pronouns* in the case of Boulder, and the *#3rd person pronouns* in the case of EnglishUS. Additionally, we can observe the importance of sentiment, as already noted in the statistical analysis, especially of positive sentiment as captured by MPQA, which highly discriminates deceptive texts in the OpSpam and Boulder data sets. Finally, the *#punctuation marks* feature is correlated with truthful text in many data sets, although with a lower coefficient.

Notice that in the results there are a number of features with high coefficients that appear to be strong only in one data set, for example, the #boosters, #nasals, and #hedges features that are extremely distinguishing only in the OpSpam collection. This observation indicates differences and variations among various data sets/domains, in accordance with previous considerations in the literature on how some features can capture the idiosyncrasies of a whole domain or only of a particular use case. In the case of the OpSpam, such features might be representative of the online reviews domain or might reflect how mechanical turkers fabricate sanctioned lies (Mukherjee *et al.* 2013b).

Regarding the *#spatial details* feature, for which we made some interesting observations in the previous statistical analysis, we observe that they are important for discriminating truthful text only in OpSpam. The observation that fake reviews in OpSpam include less spatial language has already been pointed out by Ott *et al.* (2011; 2013) for reviews with both positive and negative sentiment. This is not the case in situations where owners bribe customers in return for positive reviews or when owners ask their employees to write reviews (Li *et al.* 2014).

Finally, regarding the *#lemmas* and *#words*, they were not found to be important in this analysis. The same holds for the used tenses in most data sets with no clear direction.

Per culture and cross-cultural observations

Table 12 reports the per culture and cross-cultural observations. Although the resulting feature sets are quite distinct, we observe some similarities around the usage of pronouns. Again pronouns have very large coefficients in most data sets, and the usage of *#1st person pronouns* is correlated with truthful text for the individualistic native English and collectivist Romanian speakers, while the usage of *#3rd person pronouns* is correlated with deception in collectivist EnglishIndia and Romanian data sets. Positive sentiment, as already discussed, prevails in native English speakers for deceptive text, while sentiment features do not play any major role in other

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 12. Multiple logistic regression analysis for each data set across cultures. The Russian dataset is absent since no significant features were found in the Mann–Whitney test. SE stands for standard error. We show in bold *p*-values < 0.01. Positive or negative estimate values indicate features associated with deceptive or truthful text, respectively

USA – NativeEnglish	Estimate	SE	Wald	*p*-value
#boosters	8.338	3.86	2.16	0.031
#plosives	5.024	1.916	2.623	**0.009**
positive sentiment-MPQA	3.747	0.864	4.338	**∼ 0**
#verbs	3.351	1.181	2.837	**0.005**
mean preverb length	0.104	0.015	6.749	**∼ 0**
#lemmas	0.009	0.004	2.122	0.034
#words	−0.005	0.002	−2.479	0.013
#punctuation marks	-0.329	0.042	−7.755	**∼ 0**
#articles	−2.406	1.197	−2.01	0.044
#1st person pronouns (plural)	−11.399	2.154	−5.293	**∼ 0**
#exclusion words	−13.679	4.115	−3.324	**∼ 0**

Belgium – Dutch	Estimate	SE	Wald	*p*-value
#verbs	2 614	1 202	2 175	0.031
average word length	0 321	0 113	2 853	**0.004**
#lemmas	0 029	0 009	3 28	**0.001**
#words	−0 016	0 005	−3 542	**∼ 0**
mean sentence length	−0 052	0 017	−3 055	**0.002**
#verbs in present tense	−0 848	0 307	−2 765	**0.006**
intercept	−2 657	1 114	−2 385	0.017

India – EnglishIndia	Estimate	SE	Wald	*p*-value
#negations	15.602	4.21	3.706	**∼ 0**
#3rd person pronouns	6.917	2.224	3.11	**0.002**

Mexico – SpanishMexico	Estimate	SE	Wald	*p*-value
intercept	7.728	2.259	3.421	**∼ 0**
#verbs in past tense	−4.775	2.361	−2.022	0.043
#conjunction words	−10.95	4.465	−2.453	0.014
#prepositions	−12.67	3.939	−3.216	**0.001**
#nasals	−23.77	7.4	−3.213	**0.001**

Romania – Romanian	Estimate	SE	Wald	*p*-value
#3rd person pronouns	19.107	5.094	3.75	**∼ 0**
#fricatives	7.056	3.904	1.807	0.071
intercept	2.48	0.628	3.947	**∼ 0**
#spatial words	−1.529	0.743	−2.06	0.039
#1st person pronouns (plural)	−12.692	6.205	−2.046	0.041
#1st person pronouns (singular)	−21.014	8.111	−2.591	**0.01**

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

cultures. Additionally, the *#lemmas* and *#words* do not seem to discriminate between the different classes of text, and the usage of tenses plays a mixed and not significant role. *#nasals* appear to correlate with deceptive text in native English speakers, while it is the most discriminative feature for truthful text for Spanish. For the EnglishIndia data set, by far the most distinguishing feature is the *#negations*. This is a finding that agrees with the relevant bibliography in relation to the significance of negations in South Asian languages (see also Section 2.3). A final observation is the absence of features correlated with truthful and deceptive text in the similarly created EnglishIndia and SpanishMexico data sets, respectively.

6. Classification

In this section, we evaluate the predictive performance of different feature sets and approaches for the deception detection task. First, we present and discuss the results of logistic regression, then the results of fine-tuning a neural network approach based on the state-of-the-art BERT model, and finally we provide a comparison with other related works. As a general principle, and given the plethora of different types of neural networks and machine learning algorithms in general, this work does not focus on optimizing the performance of the machine learning algorithms to the specific data sets. Our focus is to explore, given the limited size of training data, which are the most discriminative types of features in each domain and language, and, in succession, if the combination of features is beneficial to the task of deception detection.

We split the data sets into training, testing, and validation subsets with a 70-20-10 ratio. We report the results on test sets, while validation subsets were used for fine-tuning the hyperparameters of the algorithms. In all cases, we report *Recall*, *Precision*, *F-measure,* and *Accuracy*. These statistics were calculated according to the following definitions:

$$Precision\ (P){:}\frac{tp}{tp+fp} \qquad F_1{:}2\cdot\frac{Precision\cdot Recall}{Precision+Recall}$$

$$Recall\ (R){:}\frac{tp}{tp+fn} \qquad Accuracy\ (Accu.){:}\frac{tp+tn}{tp+tn+fp+fn}$$

where a true positive (tp) and a true negative (tn) occurs when the model correctly predicts the positive class or negative class, respectively, while a false positive (fp) and a false negative (fn) when the model incorrectly predicts the positive and negative class, respectively.

6.1. Logistic regression experiments

Logistic regression has been widely applied in numerous NLP tasks, among which deception detection from text (Fuller *et al.* 2009; Popoola 2017). We experimented with several logistic regression models, including one based on linguistic features (i.e., *linguistic*), various n-grams features (*phoneme-gram*, *character-gram*, *word-gram*, *POS-gram*, and *syntactic-gram*), and the *linguistic+* model that represents the most performant model that combines linguistic features with any of the n-gram features. For our experiments, we used two implementations of logistic regression of Weka (Hall *et al.* 2009) *simple logistic* (Landwehr *et al.* 2005; Sumner *et al.* 2005) and *logistic* (Le Cessie and Van Houwelingen 1992). The *simple logistic* has a built-in attribute selection mechanism based on LogitBoost (Friedman *et al.* 2000), while the *logistic* aims to fit a model that uses all attributes. In all cases, we have two mutually exclusive classes (deceptive, truthful), and we use a classification threshold of 0.5. In the case of n-grams, a preprocessing step selects the highest occurring 1000 n-gram features, while when the attribute selection is set on the *CfsSubsetEval* evaluator of Weka is used. The *CfsSubsetEval* evaluator estimates the predictive power of *subsets* of features.

In the following tables (Tables 13, 14, 15, 16, 17, 18, 19, 21, 22, and 23), we present the logistic regression results. We group the native English data sets and seek for differences across them,

Table 13. Results for the OpSpam dataset

Type	Best Setup	R	P	F1	AUC	Accu.
Linguistic	*SimpLog*	0.67	0.73	0.70	0.80	0.71
Phoneme-gram	(1,3), *SimpLog*	0.76	0.79	0.77	0.86	0.78
Character-gram	(1,3), *SimpLog*	0.78	0.78	0.78	0.86	0.78
Word-gram	(1,2), *SimpLog*, *stem*	0.81	0.82	0.81	0.90	0.82
POS-gram	(3,3), *Log*, *attrsel*	0.72	0.70	0.71	0.78	0.71
Syntactic-gram	(1,2), *SimpLog*	0.76	0.70	0.73	0.77	0.72
Linguistic+	*Word*, (1,1), *SimpLog*, *stop*, *lowercase*	**0.88**	**0.85**	**0.86**	**0.91**	**0.86**
Majority baseline						0.50

since they are written in the same language and we assume the same culture for the authors (see Section 6.1.1 and Tables 13, 14, 15, 16, and 17). Then we proceeded with cross-domain data set experiments for native English data sets, by iteratively keeping each native English data set as a testing set and using the rest as the training set (see Section 6.1.2 and Table 20). Lastly, in Section 6.1.3, we present cross-culture experiments. We report only the best performed experimental set-up in the test set based on the accuracy value for each feature type. The measures *Precision*, *Recall,* and *F1* refer to the deceptive class while in all cases we report a majority baseline that classifies all instances in the most frequent class. We also report AUC (surface area under a ROC curve) measure (Hanley and Mcneil 1982). AUC value shows the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance. Consequently, the closer the AUC is to 1, the better the performance of the classifier (Ling *et al.* 2003). The description of the experimental set-up uses the following notation:

(a,b):	all n-grams of size in [a,b], with $a \geq b$ and $a, b \in [1,3]$ (e.g., (1,2) denotes all unigrams and bigrams).
stem:	word stemming.
attrsel:	attribute selection.
stop:	stopwords removal.
lowercase:	lowercase conversion.
SimpLog|*Log*:	Weka algorithm (*simple logistic* or *logistic*).

6.1.1 Native English data set experiments

Tables 13, 14, and 15 present the results for the US data sets that concern the online reviews domains (i.e., datasets OpSpam, DeRev, and Boulder). Each data set consists of reviews about a particular product category or service, with the exception of the Boulder data set, which covers the wider domain of hotels and electronic products.

In the OpSpam data set (see Section 3.1), the best performance is achieved with the combination of linguistic cues with the word-gram (unigram) configuration (86% accuracy). The other configurations, although not as performant, managed to overshadow the majority baseline (see Table 13). Additionally, the second best performance of the word unigram approach showcases the importance of the word textual content in this collection.

In the DeRev data set (see Section 3.3), the word unigram configuration offers exceptional performance (accuracy of 1.00%). The rest configurations achieve much lower performances. However, as in the case of the OpSpam data set, the performance in all the configurations is a

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 14. Results on the test set for the DeRev dataset

Type	Best Setup	R	P	F1	AUC	Accu.
Linguistic	*Log,attrsel*	0.72	0.74	0.71	0.84	0.72
Phoneme-gram	(1,2), *Log*	0.83	0.79	0.81	0.88	0.80
Character-gram	(1,2), *Log*	0.91	0.84	0.87	0.90	0.87
Word-gram	(1,1), *Log*, *stop*, *stem*	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
POS-gram	(3,3), *Log*	0.83	0.68	0.75	0.76	0.72
Syntactic-gram	(3,3),*Log,attrsel*	0.65	0.60	0.62	0.64	0.61
Linguistic+	*Word*, (1,1), *Log*, *stop*, *stem*	1.00	0.92	0.96	0.98	0.96
Majority baseline						0.49

Table 15. Results on the test set for the Boulder dataset

Type	Best Setup	R	P	F1	AUC	Accu.
Linguistic	*Log*, *attrsel*	0.71	0.69	0.62	0.60	0.70
Phoneme-gram	(2,3), *SimpLog*	0.97	0.70	0.81	0.50	0.68
Character-gram	(3,3), *SimpLog*	**0.99**	0.70	**0.82**	0.53	0.70
Word-gram	(1,3), *Log*, *stop*, *stem*, *attrsel*	0.92	0.73	0.81	0.65	0.71
POS-gram	(2,3), *Log*, *attrsel*	0.90	**0.75**	**0.82**	**0.67**	**0.73**
Syntactic-gram	(1,3), *SimpLog*	0.98	0.70	**0.82**	0.65	0.70
Linguistic+	*SN*, (1,2), *Log*, *attrsel*	0.87	0.73	0.79	0.60	0.68
Majority baseline						0.70

lot better than the majority baseline. Since we were puzzled with the 1.00 value in all measures for the unigram configuration, we ran some additional experiments. Our results show that in this specific data set, there are words that appear only in one class. For example, the word "Stephen" is connected only with the truthful class and the words "thriller", "Marshall", "faith", and "Alan" only with the deceptive class. After thoroughly checking how this collection was created, we found that the above observation is a result of how this data set was constructed. Specifically, the authors have used different items (i.e., books) for the deceptive and truthful cases, and as a result the classifiers learn to identify the different items. To this end, the performance of the linguistic, POS-gram, and syntactic-gram configurations are more representative for this data set, since they are more resilient to this issue.

The Boulder data set is quite challenging, since it includes two domains under the generic genre of online reviews (hotels and electronics), and two types of deception, that is, lies and fabrications (see Section 3.2). Given the above, we observe that the performance of all classifiers is much lower and close to the majority baseline (as shown in Table 15). The best accuracy is provided by the POS (bigrams+trigrams) configuration that achieves with a value of 73%, followed closely by the rest. Notice also the poor performance of the AUC measures, which is an important observation since the data set is not balanced.

The results for the EnglishUS data set, which is based on deceptive and truthful essays about opinions and feelings (see Section 3.5), are presented in Table 16. In this data set, the linguistic

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 16. Results on the test set for the EnglishUS dataset

Type	Best Setup	R	P	F1	AUC	Accu.
Linguistic	*SimpLog*	**0.62**	**0.74**	**0.67**	**0.74**	**0.70**
Phoneme-gram	(1,1), *SimpLog*	0.67	0.63	0.65	0.70	0.64
Character-gram	(1,1), *Log*, *attrsel*	0.57	0.59	0.58	0.60	0.58
Word-gram	(1,3), *Log*	0.67	0.60	0.63	0.66	0.61
POS-gram	(1,1), *Log*, *attrsel*	0.68	0.66	0.67	0.70	0.67
Syntactic-gram	(1,1), *Log*, *attrsel*	0.70	0.60	0.65	0.70	0.62
Linguistic+	*Word*, (1, 1), *SimpLog.stem*	0.65	0.68	0.67	0.71	0.68
Majority baseline						0.50

Table 17. Results on the test set for the Bluff dataset

Type	Best Setup	R	P	F1	AUC	Accu.
Linguistic	*Log*	0.60	0.75	0.67	0.60	0.60
Phoneme-gram	(1,1), *SimpLog*	**0.74**	0.65	0.69	0.49	0.56
Character-gram	(1,1), *SimpLog*	0.69	0.71	0.70	0.56	0.60
Word-gram	(1,2), *Log*	0.46	0.64	0.53	0.58	0.46
POS-gram	(1,2), *SimpLog*	0.71	0.71	0.71	0.63	0.61
Syntactic-gram	(1,2), *SimpLog*	0.63	0.73	0.68	**0.64**	0.60
Linguistic+	*POS*, (1,3), *Log*	0.69	**0.77**	**0.73**	0.62	**0.65**
Majority baseline						0.69

model offers the best performance (71% accuracy). The combination of linguistic cues with word unigrams, the POS-gram (unigrams), and the phoneme-gram (unigrams) configurations provide lower but relative close performance.

Lastly, Table 17 contains the results for the Bluff data set, which is the only data set that originates from spoken data and is multidomain (see Section 3.4). All the configurations are equal or below the major baseline which is 69%. Notice that this is a small unbalanced data set with most configurations having a low AUC performance. The inclusion of features that elicit humorous patterns could possibly improve the performance of the classifiers, since an integral characteristic of this data set is humor, a feature that we do not examine in this work.

Tables 18 and 19 present the top ten features in terms of their estimate value for each class, for the configuration with the best performance. We observe that in one-domain data sets, the content in the form of word grams is prevalent and implicitly express deceptive patterns. This is the case for the OpSpam and DeRev data sets. For example, spatial details and verbs in past tense (i.e., told, renovated, updated, based, returned) are associated with the truthful class while positive words (e.g., amazing, luxury, intriguing) are related to deceptive class. In the rest data sets that consist of different topics (i.e., two in the Boulder, three in the EnglishUS, and multiple in the Bluff data set), the best performance is achieved with the use of linguistic cues, more abstract types of n-grams such as POS-grams or with the combination of linguistic cues with n-grams. We also observe the

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 18. A list with top ten discriminating deceptive features for each native English dataset. The features are listed by decreasing estimate value as calculated by the logistic regression algorithm

OpSpam	**DeRev**	**Boulder**	**EnglishUS**	**Bluff**
word-gram+ling.	word-gram	POS-gram	ling.	POS-gram+ling.
#1st person pron.(sing.)	gettingbookreview	intercept	#boosters	#1st person pron. (pl.)
#boosters	intrigu	NNS IN RB	#3rd person pron.	negative -FB
#hedges	thriller	VB PRP MD	#nasals	#conjuctions
positive sent.-MPQA	money	TO VB PRP	negative -FB	sent.-ANEW
forever	bibl	VB VBG	#spatial words	#1st person pron.
millennium	advic	PRP MD RB	intercept	#negations
lacking	amaz	NN IN		#1st person pron. (sing.)
regency	approach	RB DT		#plosives
grand	humor	NN MD		#demonstrative pron.
luxury	addict	DT NN		#verbs in future tense

Table 19. A list with top ten discriminating truthful features for each native English dataset. The features are listed by decreasing estimate value as calculated by the logistic regression algorithm

OpSpam	**DeRev**	**Boulder**	**EnglishUS**	**Bluff**
word-gram+ling.	word-gram	POS-gram	ling.	POS-gram+ling.
priceline	intercept	NNS PRP RB	#1st person pron. (sing.)	#boosters
#spatial words	still	CD NNP	#prepositions	#hedges
separate	told	IN DT DT	#conjunction words	#indefinite pron.
returned	stand	CD	#indefinite pron.	#nasals
doormen	cour	RB VBG DT	#motion verbs	positive -MPQA
updated	sad	IN JJ IN	#words	negative -FBS
renovated	pros	NNS RB RB		#modal verbs
fault	master	VBG NN IN		#exclusion words
based	free	NN LRB		#total pronouns
note	unlik	VB CD		positive -SentiWordNet

existence of the feature “priceline” in the OpSpam list. This refers to one of the sites from which the truthful reviews were collected (e.g., Yelp, Priceline, TripAdvisor, Expedia etc.). However, since this resembles the problem in the DeRev data set in which particular features mostly are associated with one class we checked that a very small percentage of truthful reviews contain such reference. As a closing remark, we would like to showcase the rather stable performance of the linguistic models in all data sets (except maybe in the case of the Bluff data set in which the performance of all models is hindered). As a result, the linguistic cues can be considered as a valuable information for such classification models that in many cases can provide complementary information and improve the performance of other content or noncontent-based models.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

6.1.2 Cross data set experiments for US data sets

In this part, we examine the performance of the classifiers when they are trained on different data sets than those on which they are evaluated. In more details, we used every native English data set once as a testing set for evaluating a model trained over the rest native English data sets.

The setting of these experiments results in highly heterogeneous data sets not only in terms of thematic but also in terms of the collection processes, the type of text (e.g., review, essay), the deception type, etc. These discrepancies seem to be reflected to the results (see Table 20). Overall, the results show that the increased training size, with instances that are connected with the notion of deception but in different context, and without sharing many other properties, are not beneficial for the task. Note also that the configuration in these experiments results in unbalanced data sets both in training and testing sets so the comparison is fairly demanding.

The performance for the linguistic-only setting has an average accuracy of 50%. These result show that there is no overlap between the distinguishing features across the data sets that can lead to an effective feature set, as has already been revealed in the MLR analysis (see Table 12). In the case of the All -Bluff data set, the linguistic cues only configuration has the lowest accuracy of 33% which is quite below the random chance. After a closer inspection, we observed that the classifier identifies only the truthful texts (recall that the Bluff data set has an analogy 2:1 in favour of the deceptive class). This could be explained by the reversed direction of important features such as the *#words* compared to the rest of the data sets. Moreover, there are features that are statistically significant only in this data set and not on the training collection, for example, the negative sentiment FBS, and vice versa, for example, *#demonstrative pronouns*. The interested reader can find the details in Table 31 in the Appendix and in Table 11.

Similarly, n-gram configurations are close to randomness in most of the cases. However, topic relatedness seems to have a small positive impact on the results for the All -OpSpam and the All -Boulder data sets, as expected, since the Boulder data set contains hotels and electronics reviews, and the OpSpam data set also concerns hotels. The high recall values for the deceptive class in some of the classifiers depict the low coverage and the differences between the data sets.

The POS-grams and the syntactic-grams settings that are less content dependent, fail to detect morphological and syntactical patterns of deception, respectively, across the data sets. This could be attributed to the fact that such n-gram patterns might not be discriminating across different data sets and due to the fact that such types of n-grams can be implicitly influenced from the unrelated content. Overall and as future work, we plan to remove strongly domain-specific attributes from the feature space, in order for the training model to rely more on function words and content independent notions. In this direction, a hint for a possible improvement is given in Tables 21 and 22, where the most performant models include functions words, auxiliary verbs, and so on.

6.1.3 Per culture experiments

For this series of experiments, we grouped data sets based on the culture of the participants. Specifically, we experiment with individualistic data sets from the US and Belgium (Hofstede's individualistic scores of 91 and 75) and the collectivist data sets from India, Russia, Mexico, and Romania (individualistic scores of 48, 39, 30, and 30, respectively). For the United States culture, we used the unified NativeEnglish data set. This data set is unbalanced in favor of the deceptive class due to the Boulder and Bluff data sets and consists of 4285 texts in total (2498 deceptive and 1787 truthful). The results are presented in Table 23. We also measured the accuracy of pairs of of the available n-gram feature types, to check if different types of n-grams can provide different signals of deception. The results show only minor improvements for some languages. We provide the results in the Appendix (see Table 33).

Generally and despite the fact that it is safer to examine results in a per data set basis, it is evident that the word and phoneme-grams set-ups prevail in comparison with the rest of the

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 20. Cross-data set results for US data sets

Type	Best Setup	R	P	F1	AUC	Accu.
All-OpSpam						
Linguistic	*SimpLog*	**0.91**	0.51	0.65	0.52	0.52
Phoneme-gram	(1,1), *Log*	0.67	0.55	0.60	0.58	0.56
Character-gram	(1,1), *Log*	0.87	**0.58**	**0.70**	0.68	**0.62**
Word-gram	(1,2), *SimpLog*, *stem*	0.86	0.57	0.69	0.68	0.61
POS-gram	(1,1), *SimpLog*	0.90	**0.58**	**0.70**	**0.70**	**0.62**
Syntactic-gram	(1,1), *Log*	0.90	0.53	0.67	0.62	0.55
Linguistic+	*Word*, (1,3), *Log*, *lowercase*	0.68	0.55	0.61	0.56	0.56
Majority baseline						0.50
All-DeRev						
Linguistic	*SimpLog*	0.83	**0.56**	0.67	0.55	0.58
Phoneme-gram	(2,3), *Log*	0.79	**0.56**	0.65	0.61	0.58
Character-gram	(1,2), *SimpLog*	0.89	0.53	0.66	0.59	0.55
Word-gram	(2,2), *Log*, *stem*	0.78	**0.56**	0.65	**0.64**	**0.59**
POS-gram	(1,3), *SimpLog*	**0.90**	0.55	**0.68**	**0.64**	0.58
Syntactic-gram	(1,3), *Log*	0.66	0.55	0.60	0.57	0.56
Linguistic+	*Word*, (1,1), *Log*, *attrsel*, *lowercase*	0.80	0.55	0.65	0.52	0.57
Majority baseline						0.50
All-Boulder						
Linguistic	*Log*, *attrsel*	0.70	0.70	0.70	0.53	0.58
Phoneme-gram	(3,3), *Log*, *attrsel*	**0.90**	0.70	**0.79**	0.52	**0.67**
Character-gram	(1,1), *Log*, *attrsel*	0.69	**0.74**	0.71	**0.58**	0.61
Word-gram	(2,3), *Log*, *attrsel*, *lowercase*	0.79	0.72	0.75	0.54	0.64
POS-gram	(1,2), *Log*, *attrsel*	0.62	0.72	0.67	0.54	0.57
Syntactic-gram	(1,2), *SimpLog*	0.64	0.73	0.68	0.56	0.58
Linguistic+	*Phoneme*, (1,1), *Log*	0.61	0.70	0.65	0.49	0.54
Majority baseline						0.70
All-EnglishUS						
Linguistic	*SimpLog*	**0.98**	0.50	0.66	0.57	0.51
Phoneme-gram	(1,1), *SimpLog*	0.95	0.52	**0.67**	0.54	0.53
Character-gram	(1,1), *Log*	0.89	**0.54**	**0.67**	0.58	0.56
Word-gram	(1,3), *Log*, *stop*	0.86	0.53	0.66	0.52	0.56
POS-gram	(1,2), *Log*	0.73	0.53	0.61	0.55	0.54
Syntactic-gram	(1,2), *Log*	0.76	0.53	0.62	0.52	0.54

Table 20. Continued

Type	Best Setup	R	P	F1	AUC	Accu.
Linguistic+	*Character*, (1,2), *Log*	0.76	0.55	0.64	**0.59**	**0.57**
Majority baseline						0.50
All-Bluff						
Linguistic	*SimpLog*	0.13	0.50	0.20	0.43	0.33
Phoneme-gram	(3,3), *Log*, *attrsel*	0.88	0.66	0.75	0.48	0.61
Character-gram	(1,1), *Log*, *attrsel*	0.63	0.66	0.65	0.45	0.54
Word-gram	(3,3), *Log*, *stop*, *stem*, *attrsel*	**0.96**	0.67	**0.79**	0.51	0.66
POS-gram	(3,3), *Log*, *attrsel*	0.60	0.69	0.64	0.55	0.54
Syntactic-gram	(1,2), *Log*	0.59	0.68	0.63	0.51	0.54
Linguistic+	*POS*, (1,3), *SimpLog*	0.84	**0.71**	0.77	**0.59**	**0.67**
Majority baseline						0.67

setups. Even when the best accuracy is achieved through a combination of feature types, word, and phoneme n-grams belong to the combination. This is the case for the native English data set and the Romanian data set (see Tables 23 and 33, respectively). Overall, for all the examined data sets, the classifiers surpass the baseline by a lot.

The most perplexing result was the performance of the linguistic cues in the EnglishIndia and EnglishUS data sets (results presented in Tables 16 and 23) that are part of the cross-cultural dataset (see Section 3.5). These data sets have similar sizes, cover the same domains, and were created through an almost identical process. However, we observe that while the feature sets of the EnglishUS achieve accuracy of 71%, and the accuracy drops to 54% in the EnglishIndia. This is surprising, especially for same genre data sets that use the same language (i.e., EnglishUS and EnglishIndia). To ensure that this difference is not a product of the somewhat poor quality of text in the EnglishIndia data set (due to the orthographic problems), we made corrections in both data sets and we repeated the experiments. However, since the differences in the results were minor, it is difficult to identify the cause of this behavior. One hypothesis is that this difference in the performance of the feature sets may be attributed to the different expression of deception between these two cultures, given the fact that almost all other factors are stable. The second hypothesis is that since most Indians are non-native speakers of English, they use the language in the same way while being truthful or deceptive. This hypothesis is also supported by the fact that there are very few statistically important features for EnglishIndia, for example, *#negations* and the *#3rd person pronoun*. As a result, the classifiers cannot identify the two classes and exhibit a behavior closer to randomness. Notice that we might be noticing implications from both hypotheses, since the *#3rd person pronoun* is also important while deceiving for the collectivistic Romanian.

Lastly, to get a visual insight over the above results we present the most valuable features for the configuration that achieved the best accuracy in the logistic regression experiments for all the examined data sets (see Tables 24 and 25). The features are listed by decreasing estimate value. Most of the cases include morphological and semantic information that has been explicitly defined in linguistic cues (e.g., the use of pronouns as in "my room," tenses, spatial details, polarized words, etc.). As a result, the combination of such n-gram features with linguistic cues do not work in synergy. Moreover, notice the contribution of two features for discriminating deception in the SpanishMexico; the word "mi mejor" and the word "en" both attributed to the deceptive class. A similar behavior with a small resulting feature set is also evident in the Russian data set.

Table 21. A list with top ten discriminating deceptive features for each of the five cross-dataset cases. The features are listed by decreasing estimate value as calculated by the logistic regression algorithm

All -OpSpam	**All -DeRev**	**All -Boulder**	**All -EnglishUS**	**All -Bluff**
POS	word-gram	phoneme-gram	character+ling.	POS-gram+ling.
#	a luxur	wɛn aɪ askt	#boosters	#total pronouns
	[a luxury]	*[when we asked]*		
intercept	rock hotel	jɔː fɜːst bʊk	#3rd person pronouns	positive -MPQA
	[Rock hotel]	*[your first book]*		
EX	chicago hotel	ɪt wɒz an	#nasals	#verbs
	[Chicago hotel]	*[it was an]*		
JJR	chicago and	wiː faɪnəli gɒt	negative -FB	NN CC PRP
	[Chicago and]	*[we finally got]*		
RP	from our	maɪ ɹ uːm aɪ	#spatial words	NNP NNS
		[my room I]		
PDT	husband and	nɛkst taɪm wiː	intercept	DT NNP IN
		[next time we]		
PRP	you would	nɒt biː steɪɪŋ		IN NNP NNP
		[not be staying]		
VB	care of	maɪ steɪ at at	#indefinite pronouns	NNP CD
		[may stay at]		
WRB	with our	ðə dʒeɪmz ɪn	#negations	NNP POS
		[the check in]		
VBG	though	and aɪ ɹ iːsəntli	#prepositions	NNP VBZ
		[and I recently]		

6.1.4 Discussion on features

Among all the variations of n-grams tested in this work, word n-grams achieve the best results across almost all the data sets. The results for the other types of n-grams seem to be a little lower and to fluctuate in a per data set basis. More content-based n-gram types such character-grams and phoneme-grams have an adequate performance while the other variations that bear more abstract and generalized linguistic information, such as POS n-grams and syntactic n-grams achieve lower performance. However, POS-gram seem to perform quite better than the syntactic n-grams. The difference in accuracy decreases in cross-domain experiments in which semantic information is more diverse, and as already discussed, linguistic indications of deception change from one domain to another. Lastly, stemming, stopwords removal, and lowercase conversion are generally beneficiary, so it is a preprocessing step that must be examined. The experimental results show that the discriminative power of linguistic markers of deception is overly better than random baseline and the expected human performance (according to literature slightly better than chance, see Section 2) especially in one domain scenarios (see Tables 13, 14, 15, 16, and 17). More specifically, linguistic markers of deception are struggling in cross-domain settings (see Section 6.1.2). This confirms that linguistic markers of deception vary considerably and are extremely

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 22. A list with top ten discriminating truthful features for each of the five cross-dataset cases. The features are listed by decreasing estimate value as calculated by the logistic regression algorithm

All -OpSpam	**All -DeRev**	**All -Boulder**	**All -EnglishUS**	**All -Bluff**
POS	word-gram	POS-gram	character+ling.	POS-gram+ling.
WP$	next day	əkɹ ɒs ðə stɹ iːt	#exclusion words	PRP VBP IN
		[across the street]		
LS	michigan avenu	wiː hav steɪd	#3rd person pronouns	DT NNP NNP
	[Michigan avenue]	*[we have stayed]*		
LRB	the swissotel	wiː wɜː ɒn	#1st person pronouns (pl.)	#punctuation marks
		[we were on]		
$	michigan ave	wiː steɪd hiə	#articles	NN MD
	[Michigan Ave.]	*[we stayed here]*		
JJS	bed and	ɐ kɔːnə ɹ uːm	#conjuctions	NNP IN NNP
		[a corner room]		
FW	was look	and wɒz təʊld	intercept	PRP VBP NN
		[and was told]		
NNS RB RB	the elev	naɪts at ðə	#demonstrative pronouns	IN NN VBZ
	[the elevator]	*[nights at the]*		
WDT	a coupl	dʒʌst fɹ ɒm ɐ	#1st person pronouns (sing.)	PRP$ JJS
	[a couple]	*[just from a]*		
VBP	was excel	gɒt ɐ gɹ eɪt	#function words	intercept
	[was excellent]	*[got a great]*		
RBR	never sta	kɒfi ɪn ðə	positive -FBS	#1st person pron. (pl.)
	[never stayed]	*[coffee in the]*		

sensitive even within the same culture, let alone across different cultures (see Table 23). Different domains, individual differences, even the way the texts were collected seem to influence the behavior of linguistic markers and indicate how complex the deception detection task is. In the native English case in which the employed feature set is richer and in general the linguistic markers are more well-studied, we can observe better results. This might signal that there are opportunities for enhancement.

Lastly, the combination of linguistic features with n-gram variations does not enhance the performance in a decisive way in most of our experiments. N-grams and more often word-grams or phoneme-grams in an indirect way can capture information that has been explicitly encoded in the linguistic cues. However, there are cases when this combination can improve the performance of the classifier. In such cases, the resulting feature space succeeds to blend content with the most valuable linguistic markers.

6.2 BERT experiments

In these experiments, we use BERT (Devlin *et al.* 2019) with a task-specific linear classification layer on top, using the sigmoid activation function, as an alternative to the logistic regression

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 23. Per culture results

Type	Best Set-up	R	P	F1	AUC	Accu.
Individualistic						
USA-NativeEnglish						
Linguistic	*SimpLog*	0.79	0.64	0.71	0.65	0.62
Phoneme-gram	(1,1), *SimpLog*	0.81	0.67	0.73	0.68	0.65
Character-gram	(1,3), *SimpLog*	0.82	0.70	0.76	0.71	0.69
Word-gram	(1,2), *SimpLog*, *stop*, *lowercase*	0.84	**0.73**	**0.78**	**0.79**	**0.72**
POS-gram	(1,1), *Log*	0.85	0.65	0.73	0.66	0.64
Syntactic-gram	(2,3), *Log*, *attrsel*	0.86	0.67	0.75	0.71	0.67
Linguistic+	*Word*, (1,2), *SimpLog*, *stop*, *lowercase*	**0.82**	**0.73**	0.77	**0.79**	**0.72**
Majority baseline						0.58
Belgium-CLiPS						
Linguistic	*Log*	0.64	0.60	0.61	0.69	0.60
Phoneme-gram	(1,1), *SimpLog*	0.70	0.75	0.72	0.81	0.73
Character-gram	(1,3), *Log*, *attrsel*	0.71	0.74	0.73	0.80	0.73
Word-gram	(1,1), *Log*, *stop*, *attrsel*	**0.74**	**0.80**	**0.77**	**0.83**	**0.78**
POS-gram	(3,3), *Log*	0.48	0.50	0.49	0.51	0.50
Linguistic+	*Word*, (1,1), *Log*, *stem*, *stop*, *attrsel*	**0.74**	0.77	0.76	**0.83**	0.76
Majority baseline						0.50
Collectivistic						
India-EnglishIndia						
Linguistic	*SimpLog*	0.60	0.54	0.57	0.60	0.54
Phoneme-gram	(2,3), *SimpLog*	0.70	0.53	0.60	0.57	0.60
Character-gram	(1,2), *Log*, *attrsel*	**0.72**	0.59	**0.65**	0.61	**0.61**
Word-gram	(1,2), *Log*, *stem*	0.67	**0.60**	0.63	**0.63**	**0.61**
POS-gram	(2,2), *SimpLog*	0.60	**0.60**	0.60	0.62	0.60
Syntactic-gram	(2,3), *Log*	0.67	0.57	0.61	0.59	0.58
Linguistic+	*Word*, (3,3), *SimpLog*, *lowercase*	0.58	0.56	0.57	0.58	0.56
Majority baseline						0.50
Russia-Russian						
Linguistic	*Log*, *attrsel*	0.50	0.50	0.50	0.50	0.50
Phoneme-gram	(1,2), *Log*, *attrsel*	0.82	**0.60**	**0.70**	**0.68**	**0.64**
Character-gram	(2,2), *Log*, *attrsel*	0.55	0.50	0.52	0.49	0.50
Word-gram	(1,2), *SimpLog*	**0.86**	0.59	**0.70**	0.63	**0.64**

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 23. Continued

Type	Best Set-up	R	P	F1	AUC	Accu.
POS-gram	(2,2), *SimpLog*	0.68	**0.60**	0.64	0.55	0.61
Linguistic+	*POS*, (2,2), *SimpLog*	0.45	0.59	0.51	0.54	0.57
Majority baseline						0.50
Mexico-SpanishMexico						
Linguistic	*Log*	0.62	0.60	0.61	0.67	0.60
Phoneme-gram	(1,3), *Log*	**0.82**	**0.70**	**0.76**	**0.79**	**0.74**
Character-gram	(1,1), *Log*	0.79	0.60	0.68	0.62	0.63
Word-gram	(1,3), *SimpLog*, *lowercase*	**0.82**	**0.70**	**0.76**	**0.79**	**0.74**
POS-gram	(1,3), *SimpLog*	0.65	0.63	0.64	0.62	0.63
Linguistic	*Word*, (1,1), *Log*, *stem*, *stop*, *attrsel*	0.62	0.64	0.63	0.65	0.63
Majority baseline						0.50
Romania-Romanian						
Linguistic	*SimpLog*	0.62	0.64	0.63	0.67	0.64
Phoneme-gram	(1,2), *Log*, *attrsel*	0.59	0.68	0.63	0.69	0.66
Character-gram	(1,2), *SimpLog*	**0.66**	0.59	0.62	0.60	0.62
Word-gram	(1,3), *Log*, *stem*, *attrsel*	0.61	0.67	0.64	0.72	0.65
POS-gram	(1,1), *Log*, *attrsel*	0.62	0.59	0.61	0.60	0.64
Linguistic	*Phoneme*, (1,2), *Log*, *attrsel*	0.61	**0.72**	**0.66**	**0.70**	**0.68**
Majority baseline						0.50

classifiers of the previous experiments[t]. As already discussed in Section 4.4, BERT is already pretrained on a very large unlabeled corpus. Here it is further trained ('fine-tuned') jointly with the task-specific classifier on deception detection data sets to learn to predict if a text is deceptive or not. BERT produces context-aware embeddings for the tokens of the input text, and also an embedding for a special classification token ([CLS]), intended to represent the content of the entire input text. Here the input to the task-specific linear classifier is the embedding of the [CLS] token. We do not 'freeze' any BERT layers during fine-tuning, that is, the weights of all the neural layers of BERT are updated when fine-tuning on the deception detection data sets, which is the approach that typically produces the best results in most NLP tasks. We use *categorical cross entropy* as the loss function during fine-tuning and *AdamW* as the optimizer (Loshchilov and Hutter 2019). Finally, we exploit monolingual BERT models for each language (see Table 10), as well as the multilingual mBERT model. The BERT limitation of processing texts up to 512 wordpieces does not affect us, since the average length of the input texts of our experiments is below this boundary (see Table 4). However, due to batching and GPU memory restrictions, the upper bound of the used text length was 200 wordpieces, so there is some loss of information due to text truncation, though it is limited overall. More specifically, the truncation affects 5.6% of the total number of texts of all the data sets used in our experiments (506 texts

[t] For our experiments, we used the python libraries tensorflow 2.2.0, keras 2.3.1, and the bert-for-tf2 0.14.4 implementa-tionof google-research/bert, over an AMD Radeon VII card and the ROCm 3.7 platform.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 24. A list with 10 discriminating deceptive features for each dataset. The features are listed by decreasing estimate value as calculated by the logistic regression algorithm. In square brackets is the English translation

Native English	**CLiPS**	**EnglishIndia**	**Russian**	**SpanishMexico**	**Romanian**
word-gram	word-gram	word-gram	word-gram	word-gram	phoneme-gram+ling.
how to	7	then it		intercept	#3rd person pron.
	[7]	*[then it]*	*[woke up]*		
luxury	clichés	intercept			#1st person pron. (plural)
	clichés		*[him]*		
month	Potter	lie	intercept		#negations
	[Potter]	*[lie]*			
cleaning	ober	person he			#prepositions
	[waiter]	*[person he]*			
i needed	menukaart	very good			#fricatives
[I needed]	*[menu]*	*[very good]*			
turned	vegetarisch	difficult			#function words
	[vegetarian]	*[difficult]*			
all i	cappuccino	a punishment			dor' esk s' a
[all I]	*[cappuccino]*	*[a punishm]*			*[I want to]*
seemed to	horrorfilm	she never			#conjunctions
	[horror movie]	*[she never]*			
intercept	centrum	i would			' ' intelidʒ' enta
	[centre]	*[I would]*			*[intelligence]*
be staying	opslagruimte	and not			intercept
	[storage area]	*[and not]*			

out of a total of 8971). The effect of truncation is more severe in the Bluff, OpSpam, and Russian data sets, where 41% (109 out of 267), 21% (332 out of 1600), and 29% (65 out of 226) of the texts were truncated, respectively; the average text length of the three data sets is 190, 148, and 160 wordpieces, respectively. In the other data sets, the percentage of truncated texts was much smaller (10% or lower). We note that valuable signals may be lost when truncating long texts, and this is a limitation of our BERT experiments, especially those on Bluff and OpSpam, where truncation was more frequent. For example, truthful texts may be longer, and truncating them may hide this signal, or vice versa. Deceptive parts of long documents may also be lost when truncating. In such cases, models capable of processing longer texts can be considered, such as hierarchical RNNs (Chalkidis *et al.* 2019; Jain *et al.* 2019) or multi instance learning as in (Jain *et al.* 2019). No truncation was necessary in our logistic regression experiments, but long texts may still be a problem, at least in principle. For example, if only a few small parts of a long document are deceptive, features that average over the entire text of the document may not capture the deceptive parts. We leave a fuller investigation of this issue for future work.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 25. A list with 10 discriminating truthful features for each dataset. The features are listed by decreasing estimate value as calculated by the logistic regression algorithm. In square brackets is the English translation

Native English	**CLiPS**	**EnglishIndia**	**Russian**	**SpanishMexico**	**Romanian**
word-gram	word-gram	word-gram	word-gram	word-gram	phoneme-gram+ling.
my best	fastfoodketen	he should	-	mi mejor	kl' as
	[fast food chain]	*[he should]*	*[somewhere]*	*[my best]*	*[class]*
rate	Harry	help me		en	' un^{j} ' alt
	[Harry]	*[help me]*	*[in the door]*	*[in]*	*[another]*
river	Parijs	the girl			tʃ' e f' atʃe
	[Paris]	*[the girl]*	*[regular]*		*[what he is doing]*
michigan	schitterend	to him			n' oj ' ɔamen^{j}' ' ^{j}
[Michigan]	*[splendid]*	*[to him]*	*[then]*		*[we people]*
returned	effecten	always there			#total pron.
	[effects]	*[always there]*	*[in the evening]*		
elevator	Katniss	in our			#1st person pron.
	[Katniss]	*[in our]*	*[me]*		
stayed here	sla	are do			omului
	[lettuce]	*[are doing]*			*[human]*
i believe	2011	name			b' ut ' in
[I believe]	*[2011]*	*[name]*			*[but in]*
location	Woody	els			#nasals
	[Woody]	*[else]*			
	broodjes	so abort			resp' ekt
	[sandwiches]	*[so abortion]*			*[respect]*

In addition, we combined BERT with linguistic features. To this end, we concatenate the embedding of the [CLS] token with the linguistic features and pass the resulting vector to the task-specific classifier. In this case, the classifier is a multilayer perceptron with one hidden layer, consisting of 128 neurons with ReLU activations. The MLP also includes *layer normalization* (Ba *et al.* 2016) and a *dropout* layer (Srivastava *et al.* 2014) to avoid overfitting. Hyperparameters were tuned by random sampling 60 combinations of values and keeping the combination that gave the minimum validation loss. Early stopping with patience 4 was used on the validation loss to adjust the number of epochs (the max number of epochs was set to 20). The tuned hyperparameters were the following: *learning rate* (1e-5, 1.5e-5, 2e-5, 2.5e-5, 3e-5, 3.5e-5, 4e-5), *batch size* (16, 32), *dropout rate* (0.0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45), *max token length* (125, 150, 175, 200, and average training text length in tokens), and the used *randomness seeds* (12, 42, and a random number between 1 and 100).

Tables 26 and 27 present the results for these experiments. The former presents the results for each native English data set, while the latter for the cross-culture data sets. For the US culture, we used the unified data set NativeEnglish. We explored both the BERT model alone and the

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 26. Results on fine tuning BERT model for US datasets

Experiment	R	P	F1	Accu.
$\text{OpSpam}_{bert,en}$	0.94	0.89	0.91	**0.90**
$\text{OpSpam}_{bert+linguistic,en}$	0.84	0.96	0.90	**0.90**
$\text{Boulder}_{bert,en}$	0.70	0.89	0.79	**0.67**
$\text{Boulder}_{bert+linguistic,en}$	0.70	0.88	0.78	**0.67**
$\text{DeRev}_{bert,en}$	0.89	1	0.94	0.94
$\text{DeRev}_{bert+linguistic,en}$	0.96	0.96	0.96	**0.96**
$\text{EnglishUS}_{bert,en}$	0.59	0.88	0.71	0.74
$\text{EnglishUS}_{bert+linguistic,en}$	0.64	0.87	0.74	**0.76**
$\text{Bluff}_{bert,en}$	0.93	0.85	0.89	**0.83**
$\text{Bluff}_{bert+lingustic,en}$	0.88	0.82	0.85	0.77

BERT model augmented with the whole list of linguistic cues of deception studied in this work. For the native English cases, we used the BERT model for the English language, while for the per culture experiments, we experimented with both the language monolingual models and the multilingual version of the BERT model. The data set subscript declares the experimental set-up, for example, $_{bert+linguistic,en}$ uses the English language BERT model along with the linguistic cues. Exactly what types of linguistic or world knowledge BERT-like models manage to capture (or not) and the extent to which they actually rely on each type of captured knowledge is the topic of much current research (Rogers *et al.* 2020). It has been reported that the layers of BERT probably capture different types of linguistic information, like surface features at the bottom, syntactic features in the middle and semantic features at the top (Jawahar *et al.* 2019). Fine-tuning seems to allow retaining the most relevant types of information to the end task, in our case deception detection.

Overall, the experiments show similar, and in some cases improved results, compared to the logistic regression ones and the available related work (see Section 6.3). As shown in Table 26, this is the case for the OpSpam, Boulder, and EnglishUS data sets, while the performance drops a bit in the case of the DeRev data set for the plain BERT model (the excellent 98% accuracy drops to 94% for the plain BERT model, rising again to 96% for the combined BERT with the linguistic features). An interesting point is that for the Bluff data set, the plain BERT model offers better performance to the logistic classifier (83% accuracy compared to 75%), which drops to 77% when combined with the linguistic features. This is the only case where the addition of the linguistic features drops the performance of the classifier. The reason might be that the plain BERT model possibly manages to capture humor, which is an internal feature of this data set and a feature not captured by the linguistic features.

Regarding the per culture data sets shown in Table 27 and compared to the logistic regression experiments, there are clear gains in the accuracy of most of the models for the NativeEnglish, EnglishIndia, and CLiPS data sets. However, this is not the case for the SpanishMexico and the Russian data sets. Especially in the case of the peculiar Russian data set, out of the four experimental set-ups, only the BERT alone set-up with the dedicated Russian BERT model slightly surpassed the statistical random baseline of 50%. Recall that similar low performance is not only evident in our logistic experiments but also in the related work. The low performance in the case of BERT, where there are no feature extraction steps that can propagate misfires of the used tools or a problematic handling from our side, showcases that this is an intrinsically problematic collection.

A rather important finding is the contribution of the linguistic features. The addition of the linguistic features to the BERT models leads to better performance in many of the experiments, such

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 27. Per culture results for: a. the fine-tuned BERT model b. the fine-tuned BERT model along with the linguistic features. Results are reported both for the monolingual and the multilingual BERT models

Experiment	R	P	F1	Accu.
Individualist				
USA-NativeEnglish				
NativeEnglish$_{bert,en}$	0.79	**0.79**	0.79	0.75
NativeEnglish$_{bert+linguistic,en}$	**0.86**	0.77	**0.81**	**0.77**
NativeEnglish$_{bert,multi}$	0.82	0.75	0.78	0.73
NativeEnglish$_{bert+linguistic,multi}$	0.81	0.76	0.78	0.74
Belgium-CLiPS				
CLiPS$_{bert,nl}$	0.73	0.78	0.75	0.74
CLiPS$_{bert+linguistic,nl}$	**0.77**	0.78	0.77	0.77
CLiPS$_{bert,multi}$	0.74	**0.86**	**0.80**	**0.80**
CLiPS$_{bert+linguistic,multi}$	0.68	0.82	0.75	0.75
Collectivist				
India-EnglishIndia				
EnglishIndia$_{bert,en}$	0.39	**0.70**	0.50	0.62
EnglishIndia$_{bert+linguistic,en}$	**0.66**	0.63	**0.64**	**0.70**
EnglishIndia$_{bert,multi}$	0.64	0.64	**0.64**	0.64
EnglishIndia$_{bert+linguistic,multi}$	0.19	0.50	0.27	0.51
Russia-Russian				
Russian$_{bert,ru}$	0.32	**0.62**	0.42	**0.56**
Russian$_{bert+linguistic,ru}$	**0.64**	0.48	**0.55**	0.48
Russian$_{bert,multi}$	**0.64**	0.44	0.52	0.42
Russian$_{bert+linguistic,multi}$	0.52	0.43	0.47	0.42
Mexico-SpanishMexico				
SpanishMexico$_{bert,es}$	0.42	**0.76**	0.54	0.70
SpanishMexico$_{bert+linguistic,es}$	**0.65**	0.65	0.65	0.70
SpanishMexico$_{bert,multi}$	0.39	0.63	0.48	0.65
SpanishMexico$_{bert+linguistic,multi}$	0.61	0.70	**0.66**	**0.73**
Romania-Romanian				
Romanian$_{bert,ro}$	**0.74**	0.67	0.70	0.68
Romanian$_{bert+linguistic,ro}$	0.65	**0.83**	**0.73**	0.69
Romanian$_{bert,multi}$	0.60	0.62	0.61	0.61
Romanian$_{bert+linguistic,multi}$	0.69	0.71	0.70	**0.71**

Table 28. Comparison between the monolingual BERT models and the multilingual model. We report the average accuracy of the monolingual BERT model among the BERT-only and the BERT+linguistic setups and for the mBERT model respectively. With bold font we mark the best accuracy. St. sign. stands for statistical significance. We performed a 1-tailed *z*-test with a 99% confidence interval and $\alpha = 0.01$

Dataset	Avg. accu. BERT	Avg. accu. mBERT	St. sign.	1-tailed probability
NativeEnglish	**76.0**	73.5	yes	0.039
CLiPS	75.5	**77.5**	no	0.115
EnglishIndia	**66.0**	57.5	yes	0.001
Russian	**52.0**	42.0	no	0.016
SpanishMexico	**70.0**	69.0	no	0.388
Romanian	68.5	**66.0**	no	0.133

Table 29. Comparison for the multilingual model when a model is trained over one language and then tested on another one. With bold are the accuracy values over 60%. Rom. is Romanian, SpanMex. is SpanisMexico

Training/Testing	Rom	SpanMex	Rus	EnIn	CLiPS	NatEn	EnUS	NatEn \ EnUS
Rom	–	**65.6**	47.3	54.6	49.6	52.2	**61.5**	50.7
SpanMex	**64.8**	–	49.5	57.8	50.6	52.8	**60.3**	51.5
Rus	46.3	45.9	–	49.0	51.0	46.2	49.6	45.6
EnIn	**61.0**	**67.0**	50.0	–	50.2	49.2	**68.0**	53.7
CLiPS	44.4	47.1	52.2	48.5	–	54.0	48.6	54.8
NatEn	**60.5**	**67.0**	47.7	**63.5**	50.2	–	N/A	N/A
EnUS	**63.4**	**69.9**	50.0	**60.0**	52.0	N/A	–	48.5
NatEn \ EnUS	56 0	59 2	49 5	50 3	49 8	N/A	51.0	–

as in the case of EnglishUS, DeRev, NativeEnglish, SpanishMexico, Romanian, and EnglishIndia data sets. This showcases their importance compared to the corresponding logistic regression experiments, where the linguistic cues improved the n-grams approaches only in the case of the SpanishMexico and the Russian data set. The linguistic features seem to work better when combined with the BERT classifier, which might be the result of the model learning nonlinear combinations of the features. As already mentioned, in the case of the DeRev data set, the addition of the linguistics cues greatly improves the performance of the classifier, leading to almost excellent performance. Even though we have not made explicit experiments to identify that are the helpful linguistic cues in the case of BERT models, we can speculate that they are phoneme related features, for example, *#fricatives, #plosives, #nasals*, and the punctuation feature. These are significant features, which either the BERT models cannot capture or exploiting their explicit counts seems to be more effective (see Tables 31 and 32 in the Appendix).

Table 28 provides a comparison between the monolingual BERT models and mBERT. In particular, monolingual BERT models seem to perform better, except in Dutch and Romanian. Despite the lower performance of the mBERT model, the difference is not prohibitive.

6.2.1 Cross-language experiments

In this section, we proceed with cross-language experiments due to the adequate performance of the mBERT model. The idea of the experiment is to fine-tune a BERT model over one

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

language and test its performance over another language, trying to exploit similarities in the morphological, semantic and syntactic information encoded in BERT layers, across cultures. Our main focus is on cultures that are close in terms of the individualism dimension, thus could possibly share similar deceptive patterns that BERT can recognize. We are also interested in cross-cultural experiments to evaluate to what extent BERT can distinguish between deceptive and truthful texts in a crosslingual setting. Finally, we have also added the EnglishUS data set to experiment with same domain and alike collection procedure but cross-language data sets (i.e., Romanian, SpanishMexico, EnglishIndia, and EnglishUS). We also performed experiments with the NativeEnglish minus the EnglishUS collection to explore the effectiveness of a large training dataset to a different domain (EnglishUS) and to different cultures (Romanian, SpanishMexico, EnglishIndia). For each experiment, we trained a model over the 80% of a language-specific data set, validated the model over the rest 20% of the same data set, and then tested the performance of the model over the other data sets. Notice that these experiments are not applicable for the NativeEnglish and EnglishUS data sets, since the former is a superset of the latter.

For most of the experiments, the results are close to randomness. For example, this is the case when Russian and Dutch (CLiPS) are used either as testing or training sets with any other language and when the combined NativeEnglish data set is used for testing on any other language. For the Russian language this is quite expected given the performance in the monolingual experiments. However, on the Dutch data set, the situation is different, since the fine-tuned BERT model manages to distinguish between deceptive and truthful texts in the monolingual setting but when the mBERT is trained on the Dutch data set, it does not perform well on the other data sets.

The Romanian, SpanishMexico, EnglishUS, and EnglishIndia data sets that are part of the Cross-Cultural Deception data set (see Section 6.3) show a different behavior. A model trained on one data set offers an accuracy between 60% and 70% on the other set using the mBERT, with SpanishMexico exhibiting the best performance when is is used as testing set for the EnglishUS trained model. This indicates that the domain is an important factor that alleviates the discrepancies in terms of culture and language in the crosslingual mBERT setting. A reasonable explanation might be vocabulary memorization or lexical overlap, which occurs when word pieces are present during fine-tuning and in the language of the testing set. However, according to Pires *et al.* (2019), mBERT has the ability to learn even deeper multilingual representations.

Another important observation is the performance whenever the NativeEnglish is used as training set. The domain similarity is rather small in this case, since NativeEnglish is a largely diverse data set. The results show that mBERT can possibly reveal connections in a zero-shot transfer learning setting when the training size is quite adequate. This has been observed also in other tasks, like the multilingual and multicultural irony detection in the work of Ghanem *et al.* (2020). In this case instead of the mBERT model, the authors applied an alignment of monolingual word embedding spaces in an unsupervised way. Zero-shot transfer learning for specific tasks based on mBERT is also the focus of other recent approaches (Pires *et al.* 2019) (Libovický *et al.* 2019) that show promising results. Removing the EnglishUS data set from the NativeEnglish data set reduces considerably the performance in the Romanian, SpanishMexico, and EnglishIndia datasets, showcasing the importance of domain even for cross-lingual data sets. Notice though that for the SpanishMexico and the Romanian data sets, the performance is greater than that of a random classifier, indicating cues of the zero-shot transfer connection hypothesis at least for this data set. On the other hand, the random performance for the EnglishUS and EnglishIndia data sets that have the same language with the trained model and which additionally belong to the same domain with the SpanishMexico and Romanian data sets, showcases that it is difficult to generalize.

6.3. *Comparison with other works*

Table 30 provides an overall comparison between our best experimental set-up and results, with those presented in other studies on the same corpora. The comparison was based on the accuracy

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 30. Comparison with other works on the same corpora. Bold values denote models studied in this work and the best scores. Accu. stands for Accuracy and St. Sign. marks cases where a statistically significant difference between this work and related work was found

Work	Accu. (%)	St. Sign.
OpSpam 1600 samples		
BERT	**0.90**	–
BERT+Linguistic	**0.90**	–
BERT (Kennedy *et al.* 2019)	**0.90**	no
RCNN (Zhang *et al.* 2018)	0.88	no
Psycholinguistic+word bigrams (Ott *et al.* 2011; 2013)	0.87	no
Human performance (Ott *et al.* 2011; 2013)	0.59	yes
DeRev 236 samples		
Word n-gram	**1.00**	–
LDA+word space model (Hernández-Castañeda *et al.* 2017)	0.95	yes
Various numeric features, for example, length of reviews, frequency of n-grams etc. (Fornaciari and Poesio 2014)	0.76	yes
EnglishUS 600 samples		
BERT+Linguistic	0.76	–
Character 5-grams (Sánchez-Junquera *et al.* 2018)	0.73	no
LDA+word space model (Hernández-Castañeda *et al.* 2017)	**0.85**	yes
LIWC (Pérez-Rosas and Mihalcea 2014; Pérez-Rosas *et al.* 2014)	0.69	yes
Syntax+words (Feng *et al.* 2012)	0.78	no
Words (Mihalcea and Strapparava 2009)	0.71	no
Bluff 267 samples		
BERT	**0.76**	–
Human performance	0.69	no
CLiPS 1298 samples		
BERT	**0.80**	–
Unigrams (Verhoeven and Daelemans 2014)	0.72	yes
EnglishIndia 600 samples		
BERT+Linguistic	**0.70**	–
LIWC (Pérez-Rosas and Mihalcea 2014; Pérez-Rosas *et al.* 2014)	0.66	yes

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 30. Continued

Work	Accu. (%)	St. Sign.
Russian 226 samples		
Word n-gram	0.64	–
POS tags+POS tags bigrams features (Pisarevskaya *et al.* 2017)	0.57	no
Rocchio classification (Litvinova *et al.* 2017)	**0.68**	no
SpanishMexico 346 samples		
Phoneme-gram	**0.74**	–
LIWC (Pérez-Rosas and Mihalcea 2014; Pérez-Rosas *et al.* 2014)	0.68	no
Romanian 870 samples		
BERT+Linguistic	**0.71**	–
LIWC (Pérez-Rosas *et al.* 2014)	0.64	yes

scores reported in those studies. In addition, we report human accuracy whenever it is available. For comparison purposes, we set a p-value of 0.01 and performed a 1-tailed z-test evaluating if the differences between two proportions are statistically significant. By comparing absolute numbers only, the comparison is not so straightforward and cannot easily lead to conclusions, since the studies employed different model validation techniques and set difference research goals.

To the best of our knowledge, the only computational work that addresses cross-cultural deception detection is the work of Pérez-Rosas *et al.* (2014). In that work the authors build separate deception classifiers for each examined culture and report a performance ranging between 60% and 70%. Then they build cross-cultural classifiers by applying two alternative approaches. The first one was through the translation of unigrams and the second one by using equivalent LIWC semantic categories for each language. Both approaches resulted in lower performances. All the approaches were tested on the Cross-Cultural Deception data set, which was created by the authors (Pérez-Rosas and Mihalcea 2014; Pérez-Rosas *et al.* 2014), and which we also used in this work (see Section 6.3). The treatment is different since each sub-domain data set (death penalty, abortion, best friend) is separately examined. However, since average scores are also reported we compare this work with those scores. In addition, since the EnglishUS data set has been extensively used in other works in the same way, we also report the average accuracy for these cases.

The comparison in Table 30 shows that BERT outperforms other approaches in most of the cases. BERT's performance is mostly surpassed in the relatively smaller sized data sets, indicating the need for fine-tuning BERT over a large number of training samples. In particular, BERT achieves state-of-the-art performance for the OpSpam data set, that is the gold standard for opinion spam detection. In addition, for the CLiPS data set, the BERT model outperforms the other models studied in this work, as well as another unigram approach in the bibliography (Verhoeven and Daelemans 2014). For the Cross-Cultural Deception data set (see Section 3.5), BERT outruns other approaches that are based on feature engineering for the Romanian and the EnglishIndia datasets. In the case of SpanishMexico data set, the combination of linguistic cues with word n-grams seems to have a strong discriminative power and in the EnglishUS data set the combination of latent Dirichlet allocation topics (LDA) with a word-space model achieves the highest accuracy. Lastly, in comparison with human judgments, for the two data sets that we have numbers (i.e., OpSpam and Bluff), the automatic detection approaches significantly outperform human performance with respect to the accuracy measure.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

7. Conclusions

This study explores the task of automated text-based deception detection within cultures by taking into consideration cultural and language factors, as well as limitations in NLP tools and resources for the examined cases. Our aim is to add a larger scale computational approach in a series of recent interdisciplinary works that examine the connection between culture and deceptive language. Culture is a factor that is usually ignored in automatic deception detection approaches, which simplistically assume the same deception patterns across cultures. To this end, we experimented with data sets representing six cultures, using countries as culture proxies (United States, Belgium, India, Russia, Mexico, and Romania), written in five languages (English, Dutch, Russian, Spanish, and Romanian). The data sets cover diverse genres, ranging from reviews of products and services to opinions in the form of short essays and even transcripts from a radio game show. To the best of our knowledge, this is the first effort to examine in parallel and in a computational manner, multiple and diverse cultures for the highly demanding deception detection task in text.

We aimed at exploring to what extent conclusions drawn from the social psychology field about the connection of deception and culture can be confirmed in our study. The basic notion demonstrated by these studies is that specific linguistic cues to deception do not appear consistently across all cultures, for example, they change direction, or are milder or stronger between truthful or deceptive texts. Our main focus was to investigate if these differences can be attributed to cultural norm differences and especially to the individualism/collectivism divide. The most closely related work is that of Taylor (Taylor *et al.* 2014; Taylor *et al.* 2017) from the field of social psychology that studies the above considerations for four linguistic cues of deception, namely negations, positive affect, pronouns usage, and spatial details in texts from individualistic and collectivist cultures. Having as starting point Taylor's work, we performed a study with similar objectives over a larger feature set that we created that also covers the previously mentioned ones.

The outcome of our statistical analysis demonstrates that indeed there are great differences in the usage of pronouns between different cultural groups. In accordance with Taylor's work, people from individualistic cultures employ more third person and less first person pronouns to distance themselves from the deceit when they are deceptive, whereas in the collectivism group this trend is milder, signaling the effort of the deceiver to distance the group from the deceit. Regarding the expression of sentiment in deceptive language across cultures, the original work of Taylor hypothesized that different cultures will use sentiment differently while deceiving, a hypothesis that was not supported by the results of his research. The basis for this hypothesis is the observation that in high-context languages, which are related with collectivist cultures, people tend to restrain their sentiment. Our experiments support the original hypothesis of Taylor, since we observe an increased usage of positive language in deceptive texts for individualistic cultures (mostly in the US data sets), which is not observed in more collectivist cultures. In fact, by examining the statistical significant features and the resulting feature sets from the MLR analysis, we notice that generally, there are fewer discriminating deception cues in the high-context cultures. This might be attributed to the fact that the bibliography overwhelmingly focuses on individualistic cultures and to a lesser degree on collectivist cultures, leading to a smaller variation in deceptive cues for the latter. Additionally, it might indicate that during deception, high-context cultures use other communication channels on top of the verbal ones, a hypothesis that needs further research. Moreover, in affirmation of the above considerations, we observed that the strongly distinguishing features are different for each culture. The most characteristic examples are the *#negations* for the EnglishIndia data set and the phoneme-related features for the SpanishMexico and Romanian datasets (*#nasals* and *#fricatives*). Both types of features have been related to the implicit expression of sentiment in previous studies. However, there is a need for a more thorough analysis, in order for such observations to be understood and generalized in other cultures. In relation to spatial details differences, we found that in the cross-cultural deception task, the collectivist groups increased the spatial details vocabulary. The exact opposite holds for the individualist groups, who used more spatial details while being truthful. This result is in accordance with Taylor's work.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

These findings can be analyzed in conjunction with our second research goal which was to investigate the existence of a universal feature set that is reliable enough to provide a satisfactory performance across cultures and languages. Our analysis showed the absence of such a feature set. On top of this, our experiments inside the same culture (United States of America) and over different genres revealed how volatile and sensitive the deception cues are. The more characteristic example is the Bluff data set in which deception and humor are employed at the same time and the examined linguistic features have the reversed direction. Furthermore, another variable in the examined data sets is the type of deception. The examined data sets contain multiple types such as falsifications, oppositions, and exaggerations to name a few. In addition, the data collection extraction process varies from user-generated content (e.g., posts in TripAdvisor, Amazon reviews), crowd-sourced workers, volunteers in controlled environments, and finally cases outside computer-mediated communication (the transcriptions from the Bluff the Listener show). Despite this diversity, we have to note that some features seem to have a broader impact. This is the case for the length of texts (*#lemmas* and *#words* features), where deceptive texts tend to be shorter. This was observed independently of the culture and the domain with only one exception, that of the Bluff data set. This is in accordance with previous studies, attributing this behavior to the reduction of cognitive/memory load (Burgoon 2015) during the deception act.

Our third goal was to work toward the creation of culture/language-aware classifiers. We experimented with varying approaches and examined if we can employ specific models and approaches in a uniform manner across cultures and languages. We explored two classification schemes; logistic regression and fine-tuning BERT. Moreover, the experimentation with the logistic regression classifiers demonstrated the superiority of word and phoneme n-grams over all the others n-gram variations (character, POS, and syntactic). Our findings show that the linguistic cues, even when combined with n-grams, lag behind the single or combined n-gram features, whenever models are trained for a specific domain and language (although their performance surpasses the baselines). In more details, shallow features, like the various n-grams approaches, seem to be pretty important for capturing the domain of a data set, while the linguistic features perform worse. This is the case at least for the native English data sets, where we conducted experiments over various genres and found that the shallow features perform better, even across domains. On the other hand, the linguistic cues seem to be important for the collectivist cultures, especially when combined with swallow features (e.g., in Russian, SpanishMexico, and Romanian data sets). The fine tuning of the BERT models, although costly in terms of tuning the hyperparameters, performed rather well. Particularly, in some data sets (the NativeEnglish, CLiPS, and EnglishIndia data sets), we report state-of-the-art performance. However, the most important conclusion is that the combination of BERT with linguistic markers of deception is beneficial, since it enhances the performance. This is probably due to the addition of linguistic information that BERT is unable to infer, such as phoneme-related information. Indeed, phonemes play an important role in all individual parts of this study. The experimentation with the multilingual embeddings of mBERT, as a case of zero-shot transfer learning, showed promising results that can possibly be improved by incorporating culture specific knowledge or by taking advantage of cultural and language similarities for the least resourced languages. Finally, we observed the importance of domain-specific deception cues across languages, which can be identified by mBERT. Given the promising results of mBERT, other recently introduced multilingual representations may be applied. Alternatives include, for example, MUSE (Chidambaram *et al.* 2019; Yang *et al.* 2020), LASER (Artetxe and Schwenk 2019), and LaBSE (Feng *et al.* 2020). XLM (Conneau and Lample 2019) and its XLM-R extension (Conneau *et al.* 2020) have been reported to obtain state-of-the-art performance in zero-shot cross-lingual transfer scenarios, making them appropriate for low resource languages (Hu *et al.* 2020).

Although this work is focused on deception detection from text using style-based features and without being concerned with a particular domain, we plan to consider additional features that have been used in other domains and other related work. Specifically, we aim to incorporate features used in discourse-level analysis, such as rhetorical relationships (Rubin *et al.* 2015;

Karimi and Tang 2019; Pisarevskaya and Galitsky 2019), other properties of deception like acceptability, believability, the reception (Jankowski 2018) of a deceptive piece of text (e.g., number of likes or dislikes), and/or source-based features such as the credibility of the medium or author using stylometric approaches (Potthast *et al.* 2018; Baly *et al.* 2018). Such features are used extensively in fake news detection (Zhou and Zafarani 2020). We also plan to examine the correlation of such features with the perceiver's culture (Seiter *et al.* 2002; Mealy *et al.* 2007).

We also plan to study deception detection under the prism of culture over other languages and cultures, for example, Portuguese (Monteiro *et al.* 2018), German (Vogel and Jiang 2019), Arabic[u], and Italian (Fornaciari and Poesio 2012; Capuozzo *et al.* 2020). We are also interested in exploring different contexts, for example, fake news (Pérez-Rosas *et al.* 2017), modalities, for example, spoken dialogues, as well as employing other state-of-the-art deep learning approaches, for example, XLNet (Yang *et al.* 2019), RoBERTa (Liu *et al.* 2019), and DistilBERT (Sanh *et al.* 2019).

Additionally, we plan to extend the Bluff the Listener data set with new episodes of this game show, in order to further examine the linguistic cues of deception and humor and how they correlate and to enrich the community with relevant gold data sets for nonstudied languages, for example, Greek. Moreover, we plan to investigate the role of phonemes and its relation with the expression of sentiment and incorporate and study phonemes embeddings (Haque *et al.* 2019). Finally, we will apply and evaluate our models in real-life applications. This will hopefully add more evidence to the generality of our conclusions and eventually lead to further performance improvements and reliable practical applications.

Acknowledgments. We would like to thank the anonymous reviewers, Yannis Tzitzikas and Stasinos Konstantopoulos for their insightful feedback which helped us to improve this paper.

Financial support. This work was supported by the a. Foundation for Research and Technology - Hellas (FORTH) and b. the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat forResearch and Technology (GSRT), under grant agreement No 4195.

References

Agar M. (1994). *Language Shock: Understanding the Culture of Conversation*. New York: William Morrow and Company, Inc.

Aghakhani H., **Machiry A.**, **Nilizadeh S.**, **Kruegel C. and Vigna G.** (2018) Detecting deceptive reviews using generative adversarial networks. In 2018 IEEE Security and Privacy Workshops (SPW), pp. 89–95.

Almela A., **Valencia-Garca R. and Cantos P.** (2012). Seeing through deception: A computational approach to deceit detection in written communication. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, EACL 2012, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 15–22.

Alyafeai Z., **AlShaibani M.S. and Ahmad I.** (2020). *A Survey on Transfer Learning in Natural Language Processing*.

Artetxe M. and Schwenk H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics* **7**, 597–610.

Auracher J., **Albers S.**, **Zhai Y.**, **Gareeva G. and Stavniychuk T.** (2010). P is for happiness, N is for sadness: Universals in sound iconicity to detect emotions in poetry. *Discourse Processes* **48**(1), 1–25.

Ba J.L., **Kiros J.R. and Hinton G.E.** (2016). *Layer Normalization*.

Baccianella S., **Esuli A. and Sebastiani F.** (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta: European Language Resources Association (ELRA).

Bachenko J., **Fitzpatrick E. and Schonwetter M.** (2008). Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK: Coling 2008 Organizing Committee, pp. 41–48.

Baly R., **Karadzhov G.**, **Alexandrov D.**, **Glass J. and Nakov P.** (2018). Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3528–3539.

[u] https://www.autoritas.net/APDA/task-description/

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Basiri M.E., Safarian N. and Farsani H.K. (2019). A supervised framework for review spam detection in the persian language. In *2019 5th International Conference on Web Research (ICWR)*, pp. 203–207.

Bhatt G., Sharma A., Sharma S., Nagpal A., Raman B. and Mittal A. (2018). Combining neural, statistical and external features for fake news stance identification. In *Companion Proceedings of the The Web Conference 2018*, WWW 2018. Republic and Canton of Geneva: Switzerland. International World Wide Web Conferences Steering Committee, pp. 1353–1357.

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

Bond C.F. and Atoum A. (2000). International deception. *Personality and Social Psychology Bulletin* **26**, 385–395.

Bond C.F., Atoum A., Mahmoud A. and Bonser R.N. (1990). Lie detection across cultures. *Journal of Nonverbal Behavior*, **14**, 189–204.

Boroditsky L. (2006). *Linguistic Relativity*.

Bradley M. and Lang P. (1999). Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. Technical report. Gainesville, FL: UF Center for the Study of Emotion and Attention.

Burgoon J.K. (2015). When is deceptive message production more effortful than truth-telling? A baker's dozen of moderators. *Frontiers in Psychology* **6**, 1965.

Caete J., Chaperon G., Fuentes R. and Prez J. (2020). Spanish pre-trained BERT model and evaluation data. In *to appear in PML4DC at ICLR 2020*.

Capuozzo P., Lauriola I., Strapparava C., Aiolli F. and Sartori G. (2020). DecOp: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 1423–1430.

Chalkidis I., Androutsopoulos I. and Aletras N. (2019). Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4317–4323.

Chidambaram M., Yang Y., Cer D., Yuan S., Sung Y., Strope B. and Kurzweil R. (2019). Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Florence, Italy: Association for Computational Linguistics, pp. 250–259.

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L. and Stoyanov V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics, pp. 8440–8451.

Conneau A. and Lample G. (2019). Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., pp. 7059–7069.

Conneau A., Rinott R., Lample G., Williams A., Bowman S., Schwenk H. and Stoyanov V. (2018). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, pp. 2475–2485.

Cortes C. and Vapnik V. (1995). Support-vector networks. In Machine Learning, pp. 273–297.

Critchley H. and Nagai Y. (2013). *Electrodermal Activity (EDA)*. New York, NY: Springer New York, pp. 666–669.

de Vries W., van Cranenburgh A., Bisazza A., Caselli T., van Noord G. and Nissim M. (2019). *BERTje: A Dutch BERT Model*.

DePaulo B., Stone J. and Lassiter D. (1985). Deceiving and detecting deceit. In The Self and Social Life, pp. 323–370.

Deutscher G. (2010). *Through the Language Glass: Why the World Looks Different in Other Languages*. New York: Metropolitan Books/Henry Holt and Co.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Dumitrescu S. and Avram A.-M. (2020a). *RomanianBERT*.

Dumitrescu S.D. and Avram A.-M. (2020b). Introducing RONEC – The Romanian named entity corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4436–4443.

Dunbar N., Jensen M., Burgoon J., Kelley K., Harrison K., Adame B. and Bernard D. (2015). Effects of veracity, modality, and sanctioning on credibility assessment during mediated and unmediated interviews. *Communication Research* **42**, 649–674.

Ekman P. and O'Sullivan M. (1991). Who can catch a liar? *The American Psychologist* **46**, 913–920.

Elnagar A., Lulu L. and Einea O. (2018). An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis. *Procedia Computer Science* **142**, 182–189.

Feeley T.H. and deTurck M.A. (1998). The behavioral correlates of sanctioned and unsanctioned deceptive communication. *Journal of Nonverbal Behavior* **22**, 189–204.

Feng F., Yang Y., Cer D., Arivazhagan N. and Wang W. (2020). *Language-Agnostic BERT Sentence Embedding*.

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

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Finkel J.R., Grenager T. and Manning C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005).* Ann Arbor, MI: Association for Computational Linguistics, pp. 363–370.

Fitzpatrick E. and Bachenko J. (2012). Building a data collection for deception research. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, Avignon, France: Association for Computational Linguistics, pp. 31–38.

Fnagy I. (1961). Communication in poetry. *WORD* **17**, 194–218.

Fontanarava J., Pasi G. and Viviani M. (2017). Feature analysis for fake review detection through supervised classification. In 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 658–666.

Fornaciari T., Celli F. and Poesio M. (2013). The effect of personality type on deceptive communication style. In *EISIC*. IEEE, pp. 1–6.

Fornaciari T. and Poesio M. (2012). DeCour: A corpus of DEceptive statements in Italian COURts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012).* Istanbul, Turkey: European Language Resources Association (ELRA), pp. 1585–1590.

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*. Association for Computational Linguistics, pp. 279–287.

Frank M.G. and Ekman P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology* **72**, 1429–1439.

Freud S. (1914). *The Psychopathology of Everyday Life*. Macmillan Company.

Friedman J., Hastie T. and Tibshirani R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics* **28**, 337–407.

Friedrich P. (1989). *The Language Parallax: Linguistic Relativism and Poetic indeterminacy*. Austin: University of Texas Press.

Fu G., Evans A., Wang L. and Lee K. (2008). Lying in the name of the collective good: A developmental study. *Developmental Science* **11**, 495–503.

Fuller C.M., Biros D.P. and Wilson R.L. (2009). Decision support for determining veracity via linguistic-based cues. *Decision Support System* **46**, 695–703.

Fusilier D.H., Montes-y Gómez M., Rosso P. and Cabrera R.G. (2015). Detection of opinion spam with character n-grams. In Gelbukh A. (ed.) *Computational Linguistics and Intelligent Text Processing*, Cham: Springer International Publishing, pp. 285–294.

Ghanem B., Karoui J., Benamara F., Rosso P. and Moriceau V. (2020). Irony detection in a multilingual context. In Jose J.M., Yilmaz E., Magalhães J., Castells P., Ferro N., Silva M.J. and Martins F. (eds.) *Advances in Information Retrieval.* Cham:Springer International Publishing, pp. 141–149.

Grlea C., Girju R. and Amir E. (2016). Psycholinguistic features for deceptive role detection in Werewolf. In NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12–17, 2016, pp. 417–422.

Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A. and Bengio Y. (2014). Generative adversarial nets. In Ghahramani Z., Welling M., Cortes C., Lawrence N. and Weinberger K.Q. (eds.), *Advances in Neural Information Processing Systems*, volume **27**. Curran Associates, Inc., pp. 2672–2680

Granhag P.A., Vrij A. and Verschuere B. (2014). *Detecting Deception: Current Challenges and Cognitive Approaches.* John Wiley & Sons, Ltd.

Graves A., Jaitly N. and Mohamed A. (2013). Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278.

Grover A. and Leskovec J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2016, New York, NY, USA: Association for Computing Machinery, pp. 855–864.

Haire M., Porter L.W. and Ghiselli E.E. (1966). *Managerial Thinking: An International Study.* New York: Wiley.

Hall E. 1976. *Beyond Culture*. Anchor Books.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. and Witten I.H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations* **11**, 10–18.

Hancock J., Curry L.E., Goorha S. and Woodworth M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes* **45**, 1–23.

Hanley J. and Mcneil B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

Haque A., Guo M., Verma P. and Fei-Fei L. (2019). *Audio-Linguistic Embeddings for Spoken Sentences.*

Hauch V., Blandn-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, 307–342.

Hedderich M.A., Lange L., Adel H., Strtgen J. and Klakow D. (2020). *A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios.*

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

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Hirschberg J., Benus S., Brenier J., Enos F., Hoffman S., Gilman S., Girand C., Graciarena M., Kathol A., Michaelis L., Pellom B.L., Shriberg E. and Stolcke A. (2005). Distinguishing deceptive from non-deceptive speech. In INTERSPEECH-2005, pp. 1833–1836.

Hofstede G.H. (2001). *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations, 2nd and enlarged edition*. Thousand Oaks, CA: Sage.

Hu J., Ruder S., Siddhant A., Neubig G., Firat O. and Johnson M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In DaumÉ III H. and Singh A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119. *Proceedings of Machine Learning Research*, Virtual. PMLR, pp. 4411–4421.

Hu M. and Liu B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2004. New York, NY, USA: ACM, pp. 168–177.

Jain N., Kumar A., Singh S., Singh C. and Tripathi S. (2019). Deceptive reviews detection using deep learning techniques. In Métais E., Meziane F., Vadera S., Sugumaran V. and Saraee M. (eds.), *Natural Language Processing and Information Systems*. Cham: Springer International Publishing, pp. 79–91.

Jankowski N. (2018). Researching fake news: A selective examination of empirical studies. *Javnost – The Public* **25**, 248–255.

Jawahar G., Sagot B. and Seddah D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics,, pp. 3651–3657.

Jia S., Zhang X., Wang X. and Liu Y. (2018). Fake reviews detection based on LDA. In *2018 4th International Conference on Information Management (ICIM)*, pp. 280–283.

Jones T. and Newburn T. (2001). Widening access: Improving police relations with hard to reach groups. *Police Research Series* 138.

Karimi H. and Tang J. (2019). Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3432–3442. Minneapolis, Minnesota: Association for Computational Linguistics.

Kennedy S., Walsh N., Sloka K., McCarren A. and Foster J. (2019). Fact or factitious? Contextualized opinion spam detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, pp. 344–350.

Kincaid J.P., Fishburne Jr., R.P., Rogers R.L. and Chissom B.S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Article, Institute for Simulation and Training, University of Central Florida.

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

Kramsch C. (2011). The symbolic dimensions of the intercultural. *Language Teaching* **44**, 354–367.

Kraxenberger M. and Menninghaus W. (2016). Mimological reveries? Disconfirming the hypothesis of phono-emotional iconicity in poetry. *Frontiers in Psychology* **7**, 1779.

Krishnamurthy G., Majumder N., Poria S. and Cambria E. (2018). A deep learning approach for multimodal deception detection. In *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2018)*.

Kuratov Y. and Arkhipov M. (2019). Adaptation of deep bidirectional multilingual transformers for russian language. *CoRR*, abs/1905.07213.

Lafferty J.D., McCallum A. and Pereira F.C.N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML 2001, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 282–289

Lakoff G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic* **2**, 458–508.

Landwehr N., Hall M. and Frank E. (2005). Logistic model trees. *Machine Learning* **59**, 161–205.

Lauterbur P.C. (1973). Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature* **242**, 190–191.

Le Q. and Mikolov T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning – Volume 32*, ICML 2014, II–1188–II–1196. JMLR.org.

Le Cessie S. and Van Houwelingen J. (1992). Ridge estimators in logistic regression. *Applied Statistics* **41**, 191–201.

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**, 192–213.

Levitan S.I., Levine M., Hirschberg J., Nishmar C., Guozhen A. and Rosenberg A. (2015). Individual differences in deception and deception detection. In *Proceedings of COGNITIVE 2015*.

Lewis C.C. and George J.F. (2008). Cross-cultural deception in social networking sites and face-to-face communication. *Computers in Human Behavior* 24, 2945–2964.

Li J., **Ott M.**, **Cardie C. and 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)*. Association for Computational Linguistics, pp. 1566–1576.

Libovický J., **Rosa R. and Fraser A.** (2019). How language-neutral is multilingual BERT? *arXiv e-prints*, arXiv:1911.03310.

Ling C.X., **Huang J. and Zhang H.** (2003). AUC: A better measure than accuracy in comparing learning algorithms. In Xiang Y. and Chaib-draa B. (eds.), *Advances in Artificial Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 329–341.

Litvinova O., **Seredin P.**, **Litvinova T. and Lyell J.** (2017). Deception detection in Russian texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3–7, 2017, Student Research Workshop*, pp. 43–52.

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. *CoRR*, abs/1907.11692.

Loshchilov I. and Hutter F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Loukachevitch N. and Levchik A. (2016). Creating a general russian sentiment lexicon. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).

Mafela M.J. (2013). Cultural diversity and the element of negation. Intercultural Communication Studies.

Maks I., **Izquierdo R.**, **Frontini F.**, **Agerri R.**, **Vossen P. and Azpeitia A.** (2014). Generating polarity lexicons with WordNet propagation in 5 languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 1155–1161. Reykjavik, Iceland: European Language Resources Association (ELRA).

Markus H.R. and Kitayama S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review* **98**, 224–253.

Martinez-Torres M. and Toral S. (2019). A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. *Tourism Management* **75**, 393–403.

Matsumoto D., **Yoo S.**, **Fontaine J.**, **Anguas-Wong A.**, **Arriola M.**, **Ataca B.**, **Bond M.**, **Boratav H.**, **Breugelmans S.**, **Cabecinhas R.**, **Chae J.**, **Chin W.**, **Comunian A.**, **Degere D.**, **Djunaidi A.**, **Fok H.**, **Friedlmeier W.**, **Ghosh A.**, **Glamcevski M. and Granskaya J.** (2008). Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism. *Journal of Cross-Cultural Psychology* **39**, 55–74.

Mealy M., **Stephan W. and Carolina Urrutia I.** (2007). The acceptability of lies: A comparison of Ecuadorians and Euro-Americans. *International Journal of Intercultural Relations* **31**, 689–702.

Mihalcea R. (2014). Romanian-english dictionary. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Mihalcea R. and Strapparava C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2–7 August 2009, Singapore, Short Papers*. The Association for Computer Linguistics, pp. 309–312.

Mikolov T., **Sutskever I.**, **Chen K.**, **Corrado G. and Dean J.** (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 2*, NIPS 2013. Red Hook, NY, USA: Curran Associates Inc., pp. 3111–3119

Liang M. and Hu X. (2015). Recurrent convolutional neural network for object recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3367-33-75.

Mohammad S.M., **Salameh M. and Kiritchenko S.** (2016). Sentiment lexicons for Arabic social media. In *Proceedings of 10th edition of the the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Monteiro R.A., **Santos R.L.S.**, **Pardo T.A.S.**, **de Almeida T.A.**, **Ruiz E.E.S. and Vale O.A.** (2018). Contributions to the study of fake news in Portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language – 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, pp. 324–334.

Mukherjee A., **Venkataraman V.**, **Liu B. and Glance N.S.** (2013a). What Yelp fake review filter might be doing? In Kiciman E., Ellison N.B., Hogan B., Resnick P. and Soboroff I. (eds.), *ICWSM*. The AAAI Press.

Mukherjee A., **Venkataraman V.V.**, **Liu B. and Glance N.S.** (2013b). Fake review detection: Classification and analysis of real and pseudo reviews. Technical report, University of Illinois at Chicago.

Nastase V., **Sokolova M. and Shirabad J.S.** (2007). Do happy words sound happy? A study of the relation between form and meaning for English words expressing emotions. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)*.

Newman M.L., **Pennebaker J.W.**, **Berry D.S. and Richards J.M.** (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 29, 665–675.

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*. Association for Computational Linguistics, pp. 497–501.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

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*, HLT 2011. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 309–319.

Papantoniou K. and Konstantopoulos S. (2016). Unravelling names of fictional characters. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 2154–2163.

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

Pérez-Rosas V., Banea C. and Mihalcea R. (2012). Learning sentiment lexicons in spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 3077–3081. ACL Anthology Identifier: L12-1645.

Pérez-Rosas V., Bologa C., Burzo M. and Mihalcea R. (2014). *Deception Detection Within and Across Cultures*. Cham: Springer International Publishing, pp. 157–175.

Pérez-Rosas V., Kleinberg B., Lefevre A. and Mihalcea R. (2017). Automatic detection of fake news. *CoRR*, abs/1708.07104.

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)*, pp. 440–445. MD: Association for Computational Linguistics.

Picornell I. (2013). Analysing deception in written witness statements. *Linguistic Evidence in Security, Law and Intelligence* **1**, 41–50.

Pires T., Schlinger E. and Garrette D. (2019). *How multilingual is multilingual BERT?*

Pisarevskaya D. and Galitsky B. (2019). An anatomy of a lie: Discourse patterns in customer complaints deception dataset. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*, Dialogue 2019, pp. 513–531.

Pisarevskaya D., Litvinova T. and Litvinova O. (2017). Deception detection for the Russian language: Lexical and syntactic parameters. In *Proceedings of the 1st Workshop on Natural Language Processing and Information Retrieval associated with RANLP 2017*. Varna, Bulgaria: INCOMA Inc., pp. 1–10.

Popoola O. (2017). Using rhetorical structure theory for detection of fake online reviews. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*. Association for Computational Linguistics, pp. 58–63.

Porter S. and Yuille J.C. (1996). The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior* **20**(4), 443–458.

Potthast M., Kiesel J., Reinartz K., Bevendorff J. and Stein B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 231–240. Melbourne, Australia: Association for Computational Linguistics.

Qin T., Burgoon J.K., Blair J.P. and Nunamaker J.F. (2005). Modality effects in deception detection and applications in automatic-deception-detection. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pp. 23b–23b.

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

Riloff E. and Wiebe J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2003, pp. 105–112. Stroudsburg, PA, USA: Association for Computational Linguistics.

Rogers A., Kovaleva O. and Rumshisky A. (2020). A primer in BERTology: What we know about how BERT works.

Rotman L. (2012). How culture influences the telling and detection of lies: Differences between low- and high-context individuals. Master's thesis, Lancaster University, Twente University, Enschede, the Netherlands. https://essay.utwente.nl/62456/1/Rotman

Rubin V.L. (2010). On deception and deception detection: Content analysis of computer-mediated stated beliefs. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem – Volume 47*, ASIS&T 2010. Silver Springs, MD, USA: American Society for Information Science, pp. 32:1–32:10.

Rubin V.L., Conroy N. and Chen Y. (2015). Towards news verification: Deception detection methods for news discourse. In *Proceedings of the Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, Hawaii International Conference on System Sciences (HICSS 48)*.

Rubin V.L. and Vashchilko T. (2012). Identification of truth and deception in text: Application of vector space model to rhetorical structure theory. In *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection*, EACL 2012. Association for Computational Linguistics, pp. 97–106.

Rumelhart D.E. and McClelland J.L. (1987). *Learning Internal Representations by Error Propagation*. MIT Press, pp. 318–362.

Saeed R.M., Rady S. and Gharib T.F. (2019). An ensemble approach for spam detection in arabic opinion texts. Journal of King Saud University – Computer and Information Sciences.

Sahlgren M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, University of Stockholm. http://eprints.sics.se/437/1/TheWordSpaceModel.pdf

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Salvetti F. (2014). *Detecting Deception in Text: A Corpus-Driven Approach*. PhD thesis, University of Colorado Boulder, Department of Computer Science.

Salvetti F., Lowe J.B. and Martin J.H. (2016). A tangled web: The faint signals of deception in text – boulder lies and truth corpus (BLT-C). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).

Sánchez-Junquera J., Villaseñor-Pineda L., Montes-y Gómez M. and Rosso P. (2018). Character n-grams for detecting deceptive controversial opinions. In Bellot P., Trabelsi C., Mothe J., Murtagh F., Nie J.Y., Soulier L., SanJuan E., Cappellato L. and Ferro N. (eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Cham: Springer International Publishing, pp. 135–140.

Sanh V., Debut L., Chaumond J. and Wolf T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Sapir E. (1921). *Language, An Introduction to the Study of Speech*. Brace, NY.

Schmidtke D.S., Conrad M. and Jacobs A.M. (2014). Phonological iconicity. *Frontiers in Psychology* **5**, 80.

Schuster M. and Nakajima K. (2012). Japanese and Korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 5149–5152.

Seiter J.S., Bruschke J. and Bai C. (2002). The acceptability of deception as a function of perceivers' culture, deceiver's intention, and deceiver-deceived relationship. *Western Journal of Communication* **66**, 158–180.

Shaules J. (2019). *The Language and Culture Debate*. Singapore: Springer Singapore, pp. 105–120.

Siagian A.H.A.M. and Aritsugi M. (2020). Robustness of word and character n-gram combinations in detecting deceptive and truthful opinions. *Journal of Data and Information Quality* 12.

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

Srivastava N., Hinton G., Krizhevsky A., Sutskever I. and Salakhutdinov R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**, 1929–1958.

Stoytcheva S., Cohen D. and Blake C. (2014). Exploring cultural differences in language usage: The case of negation. *Proceedings of the American Society for Information Science and Technology* 51, 1–4.

Sumner M., Frank E. and Hall M. (2005). Speeding up logistic model tree induction. In *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Springer, pp. 675–683.

Sweeney C.D. and Ceci S.J. (2014). Deception detection, transmission, and modality in age and sex. *Frontiers in Psychology* **5**, 590.

Taras V., Steel P. and Kirkman B.L. (2016). Does country equate with culture? beyond geography in the search for cultural boundaries. *Management International Review* **56**, 455–487.

Taylor I.K. and Taylor M.M. (1965). Another look at phonetic symbolism. *Psychological Bulletin* **64**, 413–427.

Taylor P.J., Larner S., Conchie S.M. and Menacere T. (2017). Culture moderates changes in linguistic self-presentation and detail provision when deceiving others. *Royal Society Open Science* **4**(6), 170128.

Taylor P.J., Larner S., Conchie S.M. and van der Zee S. (2014). *Cross-Cultural Deception Detection*, Chapter 8. Wiley-Blackwell, pp. 175–201.

Tilley P., George J. and Marett K. (2005). Gender differences in deception and its detection under varying electronic media conditions. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pp. 24b–24b.

Toma C., Hancock J. and Ellison N. (2008). Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin* **34**, 1023–1036.

Triandis H.C., Bontempo R., Villareal M.J., Asai M. and Lucca N. (1988). Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships. *Journal of Personality and Social Psychology* **54**, 323–338.

Tsunomori Y., Neubig G., Sakti S., Toda T. and Nakamura S. (2015). *An Analysis Towards Dialogue-Based Deception Detection*. Cham: Springer International Publishing, pp. 177–187.

Undeutsch U. (1967). Beurteilung der glaubhaftigkeit von aussagen [evaluation of statement credibility/ statement validity assessment]. *Forensische Psychologie* **11**, 26–181.

Undeutsch U. (1989). *The Development of Statement Reality Analysis*. Dordrecht: Springer Netherlands, pp. 101–119.

Vandello J. and Cohen D. (1999). Patterns of individualism and collectivism across the united states. *Journal of Personality and Social Psychology* **77**, 279–292.

Verhoeven B. and Daelemans W. (2014). CLiPS stylometry investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Vogel I. and Jiang P. (2019). Fake news detection with the new german dataset "GermanFakeNC". In Doucet A., Isaac A., Golub K., Aalberg T. and Jatowt A. (eds.), *Digital Libraries for Open Knowledge*. Cham: Springer International Publishing, pp. 288–295.

Vrij A. (2008a). *Detecting Lies and Deceit: Pitfalls and Opportunities*. Wiley Series in the Psychology of Crime, Policing and Law. Wiley.

Vrij A. (2008b). *Detecting Lies and Deceit: Pitfalls and Opportunities*. New York, NY, USA: John Wiley & Sons Ltd.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Vrij A., Granhag P.A. and Porter S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest* **11**, 89–121.

West J. and Graham J. (2004). A linguistic-based measure of cultural distance and its relationship to managerial values. *Management International Review* **44**, 239–260.

Whissell C. (1999). Phonosymbolism and the emotional nature of sounds: Evidence of the preferential use of particular phonemes in texts of differing emotional tone. *Perceptual and Motor Skills* **89**, 19–48.

Whitely W. and England G.W. (1980). Variability in common dimensions of managerial values due to value orientation and country differences. *Personnel Psychology* **33**, 77–89.

Whorf B.L. (1956). *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf.* Cambridge: Technology Press of Massachusetts Institute of Technology.

Wilson T., Wiebe J. and Hoffmann P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT 2005. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 347–354.

Wrtz E. (2017). Intercultural communication on web sites: A cross-cultural analysis of web sites from high-context cultures and low-context cultures. *Journal of Computer-Mediated Communication* **11**, 274–299.

Yang Y., Cer D., Ahmad A., Guo M., Law J., Constant N., Hernandez Abrego G., Yuan S., Tar C., Sung Y.-H., Strope B., and **Kurzweil R.** (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Online. Association for Computational Linguistics, pp. 87–94.

Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R.R. and Le Q.V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *In Advances in Neural Information Processing Systems*, volume **32**. Curran Associates, Inc., pp. 5753–5763.

Yilmaz C.M. and Durahim A.O. (2018). SPR2EP: A semi-supervised spam review detection framework. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM 2018. IEEE Press, pp. 306–313.

Zajonc R., Murphy S. and Inglehart M. (1989). Feeling and facial efference: implications of the vascular theory of emotion. *Psychological Review* **96**, 395–416.

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

Zhou L., Burgoon J.K., Nunamaker J.F. and Twitchell D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation* **13**, 81–106.

Zhou L. and Sung Y. (2008). Cues to deception in online chinese groups. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, p. 146.

Zhou L. and Zhang D. (2008). Following linguistic footprints: Automatic deception detection in online communication. *Communications of the ACM* **51**, 119–122.

Zhou X. and Zafarani R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys* 53.

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Appendix A. Mann-Whitney U test

Table 31. *p*-value for linguistic cues for the native English datasets based on Mann-Whitney U test. The numbers in brackets denote the mean for truthful and the mean for deceptive texts respectively. With bold font *p*-values < 0.01

Linguistic cue	OpSpam	Boulder	DeRev	EnglishUS	Bluff
avg. word length	0.093 [4.533 4.509]	0 026 [4.43 4.403]	**∼ 0** [4.582 4.796]	0.674 [4.437 4.426]	0.034 [4.784 4.866]
#adj. and #adv.	0.337 [0.167 0.165]	0.122 [0.171 0.176]	**∼ 0** [0.164 0.141]	0.466 [0.155 0.153]	0.506 [0.126 0.125]
#articles	**0.006** [0.12 0.117]	**0.001** [0.121 0.116]	**∼ 0** [0.127 0.113]	0.923 [0.1 0.1]	0.302 [0.106 0.103]
#boosters	**∼ 0** [0.005 0.006]	0.094 [0.005 0.006]	0.087 [0.007 0.009]	0.27 [0.008 0.01]	0.184 [0.007 0.005]
#filled pauses	0.965 [0.001 0]	0.97 [0.002 0.001]	0.91 [0.008 0]	∼ 1	∼ 1
#function words	**∼ 0** [0.279 0.29]	0.07 [0.293 0.296]	**∼ 0** [0.294 0.244]	0.021 [0.318 0.329]	**0.001** [0.243 0.227]
#hedges	**∼ 0** [0.013 0.016]	0.329 [0.012 0.014]	0.027 [0.013 0.008]	0.729 [0.015 0.016]	0.512 [0.012 0.012]
#lemmas	0.04 [89.237 84.71]	**∼ 0** [70.015 59.974]	**0.001** [73.593 76.983]	**∼ 0** [51.163 42.283]	**∼ 0** [107.393 125.129]
#negations	0.065 [0.017 0.016]	0.031 [0.017 0.02]	0.202 [0.015 0.011]	0.029 [0.023 0.029]	0.075 [0.013 0.011]
#prepositions	0.43 [0.106 0.106]	0.502 [0.099 0.097]	**0.008** [0.107 0.12]	**∼ 0** [0.111 0.093]	0.982 [0.111 0.111]
#punctuation marks	**∼ 0** [2.45 1.964]	**∼ 0** [2.095 1.92]	0.023 [2.597 2.817]	**∼ 0** [1.829 1.568]	0.735 [3.503 3.505]
#vague words	0.564 [0.481 0.46]	0.593 [0.383 0.359]	0.475 [0.525 0.542]	0.658 [0.207 0.18]	0.285 [0.382 0.5]
#verbs	**∼ 0** [0.173 0.184]	0.014 [0.183 0.187]	0.092 [0.179 0.171]	**0.002** [0.206 0.216]	0.412 [0.185 0.181]
#words	0.294 [150.109 144.7]	**∼ 0** [114.206 93.446]	**0.003** [126.11 122.051]	**∼ 0** [79.177 62.263]	**∼ 0** [169.393 201.331]
#fricatives	**0.001** [0.135 0.138]	0.108 [0.139 0.138]	0.043 [0.134 0.139]	0.884 [0.144 0.145]	0.139 [0.134 0.132]
#nasals	**0.001** [0.081 0.083]	0.313 [0.082 0.083]	0.204 [0.083 0.085]	0.369 [0.089 0.091]	0.468 [0.086 0.085]
#plosives	0.165 [0.104 0.105]	0.621 [0.111 0.112]	0.958 [0.103 0.103]	0.876 [0.097 0.097]	0.84 [0.108 0.107]
#pronouns	**∼ 0** [0.066 0.076]	0.024 [0.076 0.08]	**0.002** [0.077 0.057]	0.977 [0.093 0.095]	0.912 [0.057 0.057]
#1st person pron.	**∼ 0** [0.04 0.051]	0.076 [0.04 0.043]	**∼ 0** [0.037 0.017]	**∼ 0** [0.049 0.03]	0.024 [0.009 0.013]

Table 31. Continued

#3rd person pron.	0.427 [0.02 0.02]	0.038 [0.029 0.027]	0.055 [0.03 0.022]	∼ **0** [0.036 0.052]	0.068 [0.03 0.034]
#1st person pron. (singular)	∼ **0** [0.023 0.037]	0.055 [0.032 0.035]	∼ **0** [0.035 0.013]	∼ **0** [0.03 0.018]	0.055 [0.006 0.009]
#1st person pron. (plural)	∼ **0** [0.017 0.014]	0.28 [0.008 0.007]	0.149 [0.002 0.004]	**0.004** [0.02 0.012]	0.402 [0.003 0.004]
#demonstrative pron.	0.484 [0.016 0.017]	0.02 [0.022 0.024]	0.331 [0.033 0.029]	0.612 [0.024 0.024]	**0.01** [0.019 0.015]
#indefinite pron.	0.07 [0.019 0.02]	0.577 [0.022 0.022]	0.458 [0.022 0.022]	**0.001** [0.035 0.027]	0.672 [0.019 0.018]
pos. SentiWordNet	0.389 [0.382 0.374]	**0.005** [0.392 0.42]	0.544 [0.384 0.359]	0.763 [0.401 0.408]	0.251 [0.316 0.306]
neg. SentiWordNet	0.172 [0.256 0.252]	0.745 [0.276 0.281]	0.356 [0.229 0.217]	0.045 [0.3 0.32]	0.374 [0.249 0.244]
pos. MPQA	**0.001** [0.08 0.088]	∼ **0** [0.075 0.087]	0.048 [0.087 0.095]	0.283 [0.084 0.087]	0.27 [0.058 0.056]
neg. MPQA	0.032 [0.032 0.03]	0.442 [0.036 0.038]	0.937 [0.039 0.04]	**0.008** [0.067 0.079]	0.222 [0.048 0.042]
pos. FBS	0.188 [0.06 0.063]	0.038 [0.057 0.063]	0.856 [0.058 0.056]	0.495 [0.04 0.045]	0.071 [0.029 0.032]
neg. FBS	0.716 [0.021 0.022]	0.688 [0.026 0.027]	0.669 [0.027 0.028]	0.054 [0.053 0.062]	0.023 [0.034 0.028]
sentiment-ANEW	0.056 [0.045 0.048]	0.861 [0.04 0.04]	0.076 [0.041 0.046]	0.904 [0.021 0.021]	∼ **0** [19.606 23.958]
mean sentence length	∼ **0** [17.474 17.686]	**0.003** [17.569 16.596]	∼ **0** [19.186 22.119]	∼ **0** [17.59 15.1]	0.799 [21.124 21.039]
mean preverb length	∼ **0** [5.077 5.652]	0.02 [5.586 5.281]	∼ **0** [5.762 7.569]	0.374 [5.25 5.029]	0.277 [7.195 7.772]
#conjunction words	0.073 [0.045 0.043]	0.902 [0.042 0.042]	0.153 [0.043 0.046]	0.164 [0.036 0.034]	0.06 [0.03 0.034]
#subordinate clauses	∼ **0** [0.47 0.558]	0.952 [0.553 0.548]	0.012 [0.542 0.675]	∼ **0** [0.742 0.606]	0.918 [0.75 0.758]
#exclusion words	∼ **0** [0.007 0.006]	0.386 [0.006 0.006]	0.126 [0.007 0.004]	0.017 [0.005 0.003]	0.51 [0.005 0.005]
#modal verbs	∼ **0** [0.015 0.018]	**0.002** [0.016 0.021]	**0.001** [0.021 0.025]	0.429 [0.044 0.048]	0.03 [0.021 0.015]
#motion verbs	0.113 [0.103 0.097]	0.152 [0.088 0.083]	0.805 [0.209 0.189]	0.153 [0.047 0.039]	0.234 [0.088 0.102]
#spatial words	∼ **0** [0.194 0.178]	0.878 [0.165 0.166]	0.287 [0.205 0.201]	0.867 [0.176 0.174]	0.483 [0.212 0.207]
#verbs in future tense	0.745 [0.034 0.035]	0.673 [0.036 0.039]	**0.004** [0.037 0.048]	0.744 [0.041 0.045]	0.709 [0.03 0.031]
#verbs in past tense	0.077 [0.508 0.515]	∼ **0** [0.403 0.355]	∼ **0** [0.354 0.172]	∼ **0** [0.169 0.127]	∼ **0** [0.321 0.408]
#verbs in present tense	0.1 [0.452 0.446]	**0.002** [0.556 0.597]	∼ **0** [0.603 0.773]	**0.001** [0.786 0.826]	∼ **0** [0.643 0.556]

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 32. *p*-value for linguistic cues for datasets from individualistic cultures based on Mann-Whitney U test. The numbers in brackets denote the first number the mean for truthful and the second number the mean for deceptive texts. With bold font *p*-values < 0.01. With N/A are marked those features that are not applicable for the specific language

Linguistic cue	NativeEnglish	Dutch	India	Russian	Mexico	Romania
avg. word length	0.052 [4.505 4.491]	**∼ 0** [8.577 8.694]	0.373 [4.466 4.445]	0.525 [4.339 4.3]	0.486 [4.149 4.134]	0.613 [4.151 4.167]
#adj. and #adv.	0.824 [0.164 0.164]	0.047 [0.339 0.347]	0.884 [0.143 0.145]	0.391 [0.052 0.047]	0.046 [0.126 0.139]	0.634 [0.145 0.146]
#articles	**∼ 0** [0.117 0.113]	0.022 [0.231 0.237]	0.823 [0.095 0.096]	N/A	0.839 [0.102 0.104]	0.181 [0.035 0.037]
#boosters	**0.002** [0.006 0.007]	–	0.409 [0.011 0.013]	–	–	–
#filled pauses	0.943 [0.002 0]	–	∼ 1	–	–	–
#function words	0.044 [0.288 0.291]	0.815 [0.402 0.403]	0.1 [0.306 0.313]	0.62 [0.181 0.175]	0.339 [0.202 0.206]	0.339 [0.138 0.141]
#hedges	0.032 [0.013 0.015]	–	0.27 [0.013 0.014]	–	–	–
#lemmas	**∼ 0** [77.553 71.217]	**0.01** [102.835 98.21]	**0.001** [48.613 43.307]	0.37 [105.77 109.15]	**∼ 0** [66.523 48.649]	**∼ 0** [62.301 49.874]
#negations	0.985 [0.018 0.018]	0.04 [0.016 0.017]	**∼ 0** [0.02 0.026]	0.707 [0.018 0.019]	0.538 [0.021 0.022]	0.494 [0.024 0.026]
#prepositions	**0.002** [0.105 0.102]	0.086 [0.165 0.16]	0.407 [0.108 0.106]	0.761 [0.201 0.198]	**0.007** [0.124 0.112]	0.279 [0.086 0.084]
#punctuation marks	**∼ 0** [2.312 2.047]	**∼ 0** [2.664 2.421]	0.012 [1.532 1.432]	0.341 [2.384 2.387]	**∼ 0** [2.719 2.181]	**∼ 0** [2.267 1.992]
#vague words	0.582 [0.407 0.389]	–	0.349 [0.24 0.157]	–	–	–
#verbs	**∼ 0** [0.182 0.188]	**0.003** [0.281 0.291]	0.145 [0.199 0.205]	0.222 [0.173 0.17]	0.052 [0.189 0.196]	0.083 [0.173 0.177]
#words	**∼ 0** [127.933 115.155]	**∼ 0** [166.598 153.798]	**0.002** [74.757 65.46]	0.339 [192.354 196.867]	**∼ 0** [106.43 72.328]	**∼ 0** [100.223 74.97]
#fricatives	0.574 [0.138 0.138]	0.214 [0.056 0.056]	0.094 [0.144 0.148]	0.565 [0.104 0.105]	0.973 [0.149 0.149]	**0.002** [0.087 0.093]
#nasals	0.012 [0.083 0.084]	0.081 [0.059 0.058]	0.491 [0.097 0.095]	0.683 [0.094 0.093]	**∼ 0** [0.111 0.104]	**∼ 0** [0.097 0.09]
#plosives	**∼ 0** [0.105 0.107]	0.083 [0.086 0.085]	0.839 [0.097 0.097]	0.181 [0.164 0.167]	0.693 [0.139 0.14]	0.575 [0.196 0.194]
#pronouns	**∼ 0** [0.074 0.078]	0.2 [0.171 0.166]	0.839 [0.087 0.093]	0.575 [0.153 0.147]	0.624 [0.1 0.102]	0.198 [0.075 0.072]
#1st person pron.	0.256 [0.04 0.04]	0.099 [0.044 0.041]	0.072 [0.043 0.038]	0.416 [0.057 0.054]	0.351 [0.06 0.057]	**∼ 0** [0.015 0.01]
#3rd person pron.	0.076 [0.026 0.028]	0.067 [0.098 0.093]	**0.002** [0.043 0.056]	0.367 [0.015 0.013]	0.192 [0.036 0.041]	**0.004** [0.013 0.017]
#1st person pron. (singular)	**∼ 0** [0.027 0.031]	0.021 [0.043 0.039]	0.067 [0.03 0.028]	0.863 [0.027 0.028]	**∼ 0** [0.013 0.008]	**0.001** [0.006 0.004]

https://doi.org/10.1017/S1351324921000152 Published online by Cambridge University Press

Table 32. Continued

Linguistic cue	NativeEnglish	Dutch	India	Russian	Mexico	Romania
#1st person pron. (plural)	**∼ 0** [0.013 0.009]	0.021 [0.002 0.004]	0.087 [0.013 0.009]	0.834 [0.012 0.011]	**0.001** [0.012 0.007]	**∼ 0** [0.009 0.005]
#demonstrative pron.	0.058 [0.02 0.021]	0.358 [0.01 0.009]	0.246 [0.019 0.018]	0.196 [0.006 0.006]	0.212 [0.003 0.003]	0.994 [0.007 0.007]
#indefinite pron.	0.285 [0.023 0.022]	0.215 [0.007 0.008]	0.919 [0.033 0.034]	0.772 [0.001 0.001]	0.45 [0.011 0.012]	**0.004** [0.011 0.009]
pos. SentiWordNet	0.195 [0.385 0.393]	–	0.087 [0.418 0.451]	–	–	–
neg. SentiWordNet	0.558 [0.267 0.271]	–	0.875 [0.313 0.316]	–	–	–
pos. MPQA	**∼ 0** [0.079 0.086]	–	0.968 [0.093 0.093]	–	–	–
neg. MPQA	0.791 [0.04 0.041]	–	0.571 [0.084 0.084]	–	–	–
pos. FBS	**0.006** [0.054 0.059]	–	0.172 [0.046 0.05]	–	–	–
neg. FBS	0.783 [0.029 0.03]	–	0.708 [0.066 0.067]	–	–	–
sentiment-ANEW	0.318 [0.039 0.04]	–	0.71 [0.021 0.024]	–	–	–
pos. sentiment	–	0.035 [0.109 0.114]	–	0.744 [0.055 0.053]	0.141 [0.028 0.038]	0.063 [0.076 0.072]
neg. sentiment	–	0.699 [0.118 0.116]	–	0.806 [0.02 0.019]	0.167 [0.039 0.036]	0.439 [0.09 0.094]
mean sentence length	0.598 [17.814 17.343]	**∼ 0** [19.533 18.31]	0.205 [16.177 15.24]	0.64 [13.956 13.23]	**∼ 0** [39.134 29.471]	**∼ 0** [22.782 20.048]
mean preverb length	**∼ 0** [5.394 5.655]	–	0.047 [6.001 5.299]	–	–	–
#conjunction words	0.325 [0.042 0.041]	0.329 [0.11 0.108]	0.158 [0.032 0.029]	0.254 [0.066 0.063]	**∼ 0** [0.077 0.066]	**0.001** [0.05 0.045]
#subordinate clauses	**0.001** [0.556 0.579]	–	0.453 [0.476 0.506]	–	–	–
#exclusion words	**∼ 0** [0.006 0.005]	–	0.602 [0.004 0.005]	–	–	–
#modal verbs	**∼ 0** [0.021 0.023]	–	0.2 [0.051 0.056]	–	–	–
#motion verbs	0.041 [0.096 0.088]	0.212 [0.042 0.046]	0.831 [0.038 0.037]	0.46 [0.05 0.04]	–	–
#spatial words	**∼ 0** [0.185 0.175]	**0.001** [0.102 0.109]	0.227 [0.21 0.202]	0.28 [0.094 0.088]	**0.001** [0.06 0.071]	**∼ 0** [0.228 0.271]
#verbs in future tense	0.26 [0.035 0.038]	–	0.626 [0.054 0.06]	0.814 [0.203 0.202]	0.653 [0.032 0.038]	–
#verbs in past tense	**∼ 0** [0.404 0.374]	0.02 [0.27 0.293]	0.691 [0.161 0.157]	0.746 [0.667 0.671]	**∼ 0** [0.126 0.072]	0.789 [0.107 0.115]
#verbs in present tense	**0.001** [0.555 0.582]	**∼ 0** [0.552 0.517]	0.556 [0.781 0.777]	0.452 [0.149 0.142]	**∼ 0** [0.841 0.89]	0.021 [0.711 0.734]

Table 33. Results per culture of logistic regression experiments on various feature types, including combinations of pairs. The accuracy measure is reported and the bold font marks the pair with the best achieved performance. Best n-gram row indicates the best accuracy for the no-paired configuration

	US	Dutch	India	Russia	Mexico	Romania
Single feature types						
Linguistic	0.62	0.60	0.54	0.50	0.60	0.64
Phoneme	0.65	0.73	0.60	0.64	**0.74**	0.66
Character	0.69	0.73	0.61	0.50	0.63	0.62
Word	**0.72**	**0.78**	0.61	**0.64**	**0.74**	0.65
POS	0.64	0.50	0.60	0.61	0.63	0.64
SN	0.67	–	0.58	–	–	–
Pairs of n-grams						
Phoneme + Word	0.70	**0.78**	**0.66**	0.55	**0.74**	0.63
Phoneme + POS	0.68	0.76	0.64	0.46	**0.74**	**0.70**
Phoneme + SN	0.68	–	0.60	–	–	–
Character + Phoneme	0.66	0.75	0.62	0.60	0.55	0.65
Character + Word	**0.72**	**0.78**	0.63	0.57	0.62	0.63
Character + POS	0.67	0.73	0.61	0.46	0.58	0.67
Character + SN	0.69	–	0.63	–	–	–
Word + POS	**0.72**	**0.78**	0.64	0.60	**0.74**	0.63
Word + SN	**0.72**	–	0.63	–	–	–
POS + SN	0.68	–	0.51	–	–	–
Best linguistic + n-gram model						
Linguistic+ Best n-gram	**0.72**	0.76	0.56	0.57	0.63	0.68

Cite this article: Papantoniou K, Papadakos P, Patkos T, Flouris G, Androutsopoulos I and Plexousakis D (2022). Deception detection in text and its relation to the cultural dimension of individualism/collectivism. *Natural Language Engineering* **28**, 545–606. https://doi.org/10.1017/S1351324921000152