UNIDECOR: A Unified Deception Corpus for Cross-Corpus Deception Detection

Aswathy Velutharambath[1,2] and **Roman Klinger**[1]
[1]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
[2]100 Worte Sprachanalyse GmbH, Heilbronn, Germany
aswathy.velutharambath@100Worte.de
roman.klinger@ims.uni-stuttgart.de

Abstract

Verbal deception has been studied in psychology, forensics, and computational linguistics for a variety of reasons, like understanding behaviour patterns, identifying false testimonies, and detecting deception in online communication. Varying motivations across research fields lead to differences in the domain choices to study and in the conceptualization of deception, making it hard to compare models and build robust deception detection systems for a given language. With this paper, we improve this situation by surveying available English deception datasets which include domains like social media reviews, court testimonials, opinion statements on specific topics, and deceptive dialogues from online strategy games. We consolidate these datasets into a single unified corpus. Based on this resource, we conduct a correlation analysis of linguistic cues of deception across datasets to understand the differences and perform cross-corpus modeling experiments which show that a cross-domain generalization is challenging to achieve. The unified deception corpus (UNIDECOR) can be obtained from https://www.ims.uni-stuttgart.de/data/unidecor.

1 Introduction

Deception detection has remained an area of vested interest in fields like psychology, forensics, law, and computational linguistics for a myriad of reasons like understanding behavioral patterns of lying (Newman et al., 2003; DePaulo and Morris, 2004), identifying fabricated information (Conroy et al., 2015), distinguishing false statements or testimonies (Şen et al., 2022) and detecting deception in online communication (Hancock, 2009). These are relevant tasks because of the truth bias, which is the inherent inclination of humans to actively believe or passively presume that a statement made by another person is true and accurate by default, without the need for evidence to substantiate this belief (Levine, 2014). While this facilitates efficient communication, it also makes people susceptible to deception, especially in online media where digital deception (Hancock, 2009) manifests in many forms like fake news, misleading advertisements, impersonation and scams. This warrants automatic deception detection systems that can accurately distinguish between truthful and deceptive discourse solely from textual data.

The task of automatic deception detection comes with several challenges. Deception or lying is a complex human behavior and its signals are faint in text. Moreover, it is sensitive to the communication context, interlocutors, and the stake involved (Ten Brinke and Porter, 2012; Salvetti et al., 2016). Most importantly, acquiring annotated data proves to be one of the major hurdles for deception studies. Traditional data annotation methods cannot be employed because human performance is shown to be worse than machines in differentiating truths and lies (Bond Jr. and DePaulo, 2006; Vrij, 2014). One way to collect accurate data is to get the labels at source by the person producing the text. Alternatively, they can be collected using the acquired knowledge that certain types of contents are deceptive. Across the literature, different strategies like crawling fake reviews (Yao et al., 2017), collecting text from users identified as suspicious (Fornaciari et al., 2020), using non-linguistic deception cues (Fornaciari and Poesio, 2014) and soliciting through crowd-sourcing (Ott et al., 2011, 2013; Salvetti et al., 2016) have been employed to obtain reliable annotations.

The diversity in the domains of interest, the medium of deceptive communication (spoken vs. written) and dataset creation procedures make it difficult to compare cues of deception across datasets and to understand their generalizability across different domains. With this study, we aim at mitigating this situation by conducting a comparative survey of publicly available textual deception datasets.

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 39–51

We contribute (1) a consolidated corpus in a unified format and (2) conduct experiments in which we evaluate models trained on one data set on all others. Our (3) results show that cross-corpus, particularly cross-domain, generalizability is limited, which motivates future work to develop robust deception detectors. We lay the foundation for such work with (4) additional correlation analyses of the linguistic cues of deception across datasets and verify their generalizability across domains.

2 Background & Related Work

Deception in communication is the act of intentionally causing another person to have a false belief that the deceiver knows or believes to be false (Zuckerman et al., 1981; Mahon, 2007; Hancock, 2009). Lies, exaggerations, omissions, and distortions are all different forms of deception (Turner et al., 1975; Metts, 1989). While the definition of deception varies across literature, they concur that it is intentional or deliberate (Mahon, 2007; Gupta et al., 2013).

2.1 Domains and Ground Truth

Deception research is spread across disciplines which contributed to a variety of domains and consequentially to a number of data collection methods. Domains include opinions statements on a specific topic (Pérez-Rosas and Mihalcea, 2014; Capuozzo et al., 2020; Lloyd et al., 2019), open domain statements (Pérez-Rosas and Mihalcea, 2015), online reviews (Ott et al., 2011, 2013; Fornaciari and Poesio, 2014; Yao et al., 2017), deceptive dialogues in strategic games like Mafiascum[1], Box of Lies and Diplomacy (de Ruiter and Kachergis, 2018; Soldner et al., 2019; Peskov et al., 2020; Skalicky et al., 2020) and court trials (Şen et al., 2022).

The ground truth generation strategies differ across datasets. While datasets of opinion statements (Pérez-Rosas and Mihalcea, 2014; Capuozzo et al., 2020; Lloyd et al., 2019), and online reviews (Ott et al., 2011, 2013; Fornaciari and Poesio, 2014; Yao et al., 2017) are collected in written form, interviews include both verbal and non-verbal content (Şen et al., 2022). Game-based corpora contain monologue (Skalicky et al., 2020) or dialogue data (Soldner et al., 2019; Peskov et al., 2020).

All of these resources contain instances that are labeled as truthful or deceptive. Only few studies employ the same procedure to generate both truthful and deceptive content (Salvetti et al., 2016; Skalicky et al., 2020); most resort to separate strategies for collecting them (Ott et al., 2011, 2013; Fornaciari et al., 2020). Instances labeled as deceptive are either solicited content or collected from a source identified as deceptive. Ott et al. (2011, 2013) crawled the truthful reviews from websites of interest and the deceptive ones were crowd-sourced through AMT[2], while Salvetti et al. (2016) solicited both via AMT. Yao et al. (2017) tracked fake review generation tasks from crowd-sourcing platforms to identify deceptive reviews and reviewers. For the datasets based on strategic games, the labels are assigned based on game rules. Opinion domain datasets contain stances on topics, like gay marriage and abortion, written by the same person, where the truthful labeled opinions align with the author's true opinion and deceptive ones align with the opposite (Pérez-Rosas and Mihalcea, 2014; Capuozzo et al., 2020).

2.2 Automatic Deception Detection Methods

Several studies have explored the effectiveness of automatic methods to detect deception from textual data. These include feature-based classification methods with support vector machines (Ott et al., 2011; Pérez-Rosas and Mihalcea, 2014; Fornaciari and Poesio, 2014), logistic regression (de Ruiter and Kachergis, 2018), decision trees (Pérez-Rosas and Mihalcea, 2015), and random forests (Soldner et al., 2019; Pérez-Rosas and Mihalcea, 2015). Some studies also consider contextual information by using recurrent neural networks (Peskov et al., 2020) and transformer-based models (Capuozzo et al., 2020; Peskov et al., 2020; Fornaciari et al., 2021). Transformers are not always better – Peskov et al. (2020) show that BERT is en par with LSTMs while Fornaciari et al. (2021) showed that adding extra attention layers help to improve upon the previous state of the art.

Most works focused on modeling the concept of deception in one domain. An exception is Hernández-Castañeda et al. (2016) who report cross-domain classification results on OpSpam, DeRev2014, and CrossCultDe, but in an all-against-one setting, not in a one-against-one setup.

2.3 Linguistic Cues of Deception

To understand the phenomenon of deception better, previous studies have analyzed the linguistic

[1] https://www.mafiascum.net/

[2] Amazon's Mechanical Turk, https://www.mturk.com/

cues that characterize deceptive language in written statements, spoken conversations, and online communication (Newman et al., 2003; Bond and Lee, 2005) and demonstrated that a systematic analysis of these cues can prove valuable in automated deception detection specifically in computer-mediated communication (Zhou et al., 2004). Newman et al. (2003) noted that the use of fewer self-references in deceptive statements indicate that the liars are attempting to distance themselves from the lies. The use of exclusive words (e.g., *but*, *rather*) allow deceivers to introduce communicative ambiguity into the discourse. Hancock et al. (2007) noted that these cues are broadly associated with the number of words, use of pronouns, use of emotion words, and presence of markers of cognitive complexity. They also pointed out that these cues can manifest differently based on the type and medium of discourse; real-world vs. online or monologue vs. dialogue.

While these analyses have found application in machine learning models, there are more sets of features that have been used to automatically detect deception. These include n-grams (Fornaciari and Poesio, 2014; Fornaciari et al., 2020; Ott et al., 2011), part-of-speech tags (Lloyd et al., 2019; Fornaciari et al., 2020; Pérez-Rosas and Mihalcea, 2015), lexicon-based features, including the Linguistic Inquiry and Word Count (LIWC, Pennebaker et al., 2015) psychological categories, (Pérez-Rosas and Mihalcea, 2014; Yao et al., 2017) and production rules derived from syntactic context free grammar trees (Yao et al., 2017; Pérez-Rosas and Mihalcea, 2015). Duran et al. (2010), Swol et al. (2012) and Hauch et al. (2015) conducted extensive surveys and analyses of different linguistic cues of deception.

3 Unified Deception Dataset

As preparation for cross-corpus analysis of the concept of deception, we consolidate publicly available textual deception datasets into a unified format.[3] We now describe the included datasets.

Deceptive Opinion Spam (OPSPAM). Ott et al. (2011) describes *deceptive opinion spam* as fraudulent reviews written to sound authentic with the goal to deceive the reader. To study the nature of such reviews, they collected truthful reviews by crawling online review platforms like TripAdvisor[4] and crowd-sourced deceptive reviews via Amazon's Mechanical Turk (AMT). The initial OPSPAM dataset published by Ott et al. (2011) contains 400 truthful and 400 deceptive reviews with positive sentiments. Ott et al. (2013) extended the dataset to include reviews with negative sentiments. The complete OPSPAM dataset contains 1600 instances labeled for veracity and sentiment. It is available publicly with a Creative Commons Attribution-NonCommercial-ShareAlike license.[5]

Cross-cultural Deception (CROSSCULTDE). Pérez-Rosas and Mihalcea (2014) collected the CROSSCULTDE dataset to investigate deception in a cross-cultural setting. It consists of short essays on the topics of abortion, death penalty, and feelings about a best friend, collected from the United States, India, and Mexico. We take into account the data collected from the United States and India which are in English and consist of 100 deceptive and 100 truthful essays per topic per geographical region adding up to 1200 labeled instances. The dataset is available for download without mentioning any usage restrictions.[6]

Deception in Reviews (DEREV2014/2018). To investigate the phenomenon of sock puppetry, Fornaciari and Poesio (2014) collected DEREV2014, containing book reviews from *amazon.com* that were identified as authentic or fake using predefined linguistic cues. To overcome the shortcoming that these cues cannot be used while developing a deception classifier, Fornaciari et al. (2020) released the DEREV2018 dataset, in which they collect deceptive reviews based on *a priori* knowledge about authors who solicited fake reviews. Additionally, the authors crowd-sourced both truthful and deceptive reviews for the same books. The DEREV2014 dataset contains 118 reviews each with a truthful label and a deceptive label, while the DEREV2018 dataset includes 1552 reviews each collected from *amazon.com* and through crowdsourcing with a balanced distribution of truthful and deceptive reviews. The datasets overlap by 62 reviews. Both corpora are available for download.[7]

Open Domain Deception (OPENDOMAIN). Pérez-Rosas and Mihalcea (2015) study deception, gender, and age detection with an open domain

[3] We refer to our corpus as UNIDECOR: "Unified Deception Corpus". The scripts to download and convert the dataset can be found in the following repository: https://www.ims.uni-stuttgart.de/data/unidecor

[4] https://www.tripadvisor.com/

[5] https://myleott.com/op-spam.html

[6] https://web.eecs.umich.edu/~mihalcea/downloads.html

[7] https://fornaciari.netlify.app/

Dataset	Domain	Truthful	Deceptive	Total	TC	SC
Bluff the listener (BLUFF)	game	251 (33.3%)	502 (66.7%)	753	241.66	11.5
Diplomacy dataset (DIPLOMACY)	game	16402 (94.9%)	887 (5.1%)	17289	24.53	1.7
Mafiascum dataset (MAFIASCUM)	game	7439 (76.9%)	2237 (23.1%)	9676	4690.69	362.8
Multimodal Decep. in Dialogues (BOXOFLIES)	game	101 (20.2%)	400 (79.8%)	501	12.2	1.6
Miami University Decep. Detection Db. (MU3D)	interview	160 (50.0%)	160 (50.0%)	320	131.7	5.7
Real-life trial data (TRIAL)	interview	60 (49.6%)	61 (50.4%)	121	79.85	3.9
Cross-cultural deception (CROSSCULTDE)	opinion	600 (50.0%)	600 (50.0%)	1200	80.0	4.5
Deceptive Opinion (DECOP)	opinion	1250 (50.0%)	1250 (50.0%)	2500	65.56	4.0
Boulder Lies and Truth Corpus (BLTC)	review	1041 (69.8%)	451 (30.2%)	1492	116.92	6.5
Deception in reviews (DEREV2014)	review	118 (50.0%)	118 (50.0%)	236	145.22	6.7
Deception in reviews (DEREV2018)	review	1552 (50.0%)	1552 (50.0%)	3104	176.6	8.1
Deceptive opinion spam (OPSPAM)	review	800 (50.0%)	800 (50.0%)	1600	170.5	9.5
Online deceptive reviews (ONLINEDE)	review	101431 (85.9%)	16694 (14.1%)	118125	171.5	7.2
Open Domain Deception (OPENDOMAIN)	statement	3584 (50.0%)	3584 (50.0%)	7168	9.33	1.0
		134789 (82.1%)	29296 (17.9%)	164085	436.88	31.05

Table 1: Datasets included in our unified corpus (UNIDECOR), together with statistical information. TC: average token count; SC: average sentence count.

dataset acquired via AMT. Workers were asked to contribute seven true and seven plausible deceptive statements without a restriction of domain, each in a single sentence. The balanced dataset consists of 7168 annotated instances with additional demographic information. The data set is made available without specifying usage restrictions.[6]

Real-life Trial Data (TRIAL). To study real-life high-stake deception scenarios, Pérez-Rosas et al. (2015) collected videos of trial hearings from publicly available sources like "The Innocence Project" website[8]. The dataset contains multimodal information with annotations for non-verbal behavior like facial displays and gestures in addition to crowdsourced transcriptions. It contains 60 truthful and 61 deceptive reviews.This corpus is made available without specifying any usage restrictions.[6]

Boulder Lies and Truth Corpus (BLTC). Salvetti et al. (2016) built a balanced dataset containing reviews elicited via AMT for the domains of electronic appliances and hotels. The crowdworkers were instructed to write fake or real reviews, with positive or negative sentiment, about objects that they were familiar with or not. Unlike other datasets which limited the labeling to truthful vs. deceptive, this dataset distinguished between fake and deceptive reviews, where the former are fabricated opinions about an unknown object while the latter was a false review of a known object. The corpus contains 1492 reviews, out of which 451 are truthful and the rest is labeled as fake or deceptive. It is available through the LDC.[9]

Online Deceptive Reviews (ONLINEDE). To address the bottleneck that large realistic data for deception detection do not exist, Yao et al. (2017) created the ONLINEDE corpus containing manipulated reviews posted online. They employed the automatic deception detection framework outlined by Fayazi et al. (2015) to identify deceptive reviewers and reviews from social media manipulation campaigns. It contains more than 100K labeled reviews with ≈10000 deceptive instances, covering more than 30 domains. The dataset is available for research purposes from the authors.

Mafiascum Dataset (MAFIASCUM). This dataset published by de Ruiter and Kachergis (2018) contains a collection of more than 700 games of Mafia, an online strategy game played on the Internet forum MAFIASCUM[10]. Here, players are assigned deceptive or non-deceptive roles randomly, which serve as annotations of the instances. Each of the 9000 documents contain all messages written by a single user in a specific game. The average token count in the instances (4690.69) is therefore considerably higher than in other corpora. The authors have made the dataset publicly available along with the code used for analyses.[11]

Miami University Deception Detection Database (MU3D). To investigate the role of gender and race in deception studies, Lloyd et al. (2019) created MU3D. It is a collection of interview videos where participants were instructed to talk truthfully or deceptively about their relationship with a person

[8] http://www.innocenceproject.org/

[9] Linguistic Data Consortium, https://catalog.ldc.upenn.edu/LDC2014T24

[10] https://www.mafiascum.net/

[11] https://bitbucket.org/bopjesvla/thesis/src/master/

whom they liked or disliked. The 80 participants, each belonging to a different gender and ethical background contributed to a positive truth, a negative truth, a positive lie and a negative lie, counting to 160 truthful and 160 deceptive interview content. The transcriptions of these videos along with demographic information, valency, and veracity annotations are made available for research purposes with a Creative Commons Attribution-NonCommercial-NoDerivs license.[12]

Multimodal Deception in Dialogues (BOXOFLIES). To explore deception in conversational dialogue, Soldner et al. (2019) collected the BOXOFLIES dataset which is based on the "Box of Lies" game, a segment on "The Tonight Show Starring Jimmy Fallon" where two celebrity guests take turns describing the contents of a box but are allowed to lie. The opposing player must decide if they believe the description or not. The collected dataset contained 25 videos of the game, transcribed and annotated for non-verbal cues of deception and the veracity of the describer. We exported the statements containing veracity label from the dataset using ELAN[13], a tool used to create and modify annotations for audio and video data. The dataset is available for download without specifying any usage restrictions.[6]

Diplomacy Dataset (DIPLOMACY). To study deception in a conversational context specifically in long-lasting relationships, Peskov et al. (2020) employed the negotiation-based online game DIPLOMACY. The players use deception as a strategy to convince other players to form alliances, for which they use a chat interface. Contrary to other deception datasets, DIPLOMACY contains an additional label for perceived truthfulness of an instance. The intended and perceived truthfulness of each message was annotated by the sender and the receiver respectively. Out of more than 13k messages less than 5% are labeled as intended or perceived lie, resulting in an imbalanced dataset. We use the dataset made available through ConvoKit.[14]

Deceptive Opinion (DECOP). To study deception in multi-domain and multi-lingual settings, Capuozzo et al. (2020), following the method described by Pérez-Rosas and Mihalcea (2014), collected truthful and deceptive opinion statements on five different topics, namely abortion, cannabis legalization, euthanasia, gay marriage, and policies on migrants. The experiment was conducted for English and Italian, from which we include the English instances in UNIDECOR. They consist of 2500 opinions statements with balanced labels. This dataset can be obtained from the authors.

Bluff the Listener (BLUFF). To study humorous deception with no malicious intent, Skalicky et al. (2020) compiled the BLUFF dataset. It contains data from the "Bluff the Listener" game which is part of the radio show "Wait. . . Don't Tell Me". It is a variation of the game "Two Truths and a Lie" in which a panelist tells three stories, two of which are true, and one of which is false. This corpus published by Skalicky et al. (2020) contains 753 humorous stories collected from 251 episodes broadcast from 2010 to 2019. The authors downloaded the transcripts from News-Bank[15], a curated repository containing current and archived media. One-third of the stories are truthful while two-thirds are fabricated, counting to 251 truthful and 502 deceptive stories. The dataset is publicly available and can be downloaded via the OSF platform.[16]

Aggregation. We consolidate the datasets into one unified corpus in which each instance is assigned a binary label indicating if it is truthful or deceptive. We retain annotation dimensions that are available for more than one dataset (age, gender, country, and sentiment). More details on the aggregation process and a sample entry from the corpus are available in Appendix A. Table 1 provides an overview of the corpora, including size, label distribution, token and sentence counts[17], along with the domain. The datasets vary greatly in its size, but the distribution of labels is mostly comparable, except for BLTC, ONLINEDE and DIPLOMACY with comparably high counts for truthful instances.

4 Similarity Analysis

The datasets included in UNIDECOR come from a variety of domains and differ markedly in terms of the method of collection. At the same time, datasets from the same domains also have differences (e.g., solicited reviews vs. actual reviews). To understand the differences of datasets better, we explore the similarity values between these datasets

[12] https://sc.lib.miamioh.edu/handle/2374.MIA/6067

[13] https://tla.mpi.nl/tools/tla-tools/elan/download/

[14] https://convokit.cornell.edu/documentation/diplomacy.html

[15] www.newsbank.com

[16] https://osf.io/download/mupd9

[17] Using NLTK's wordpunct_tokenize and sent_tokenize

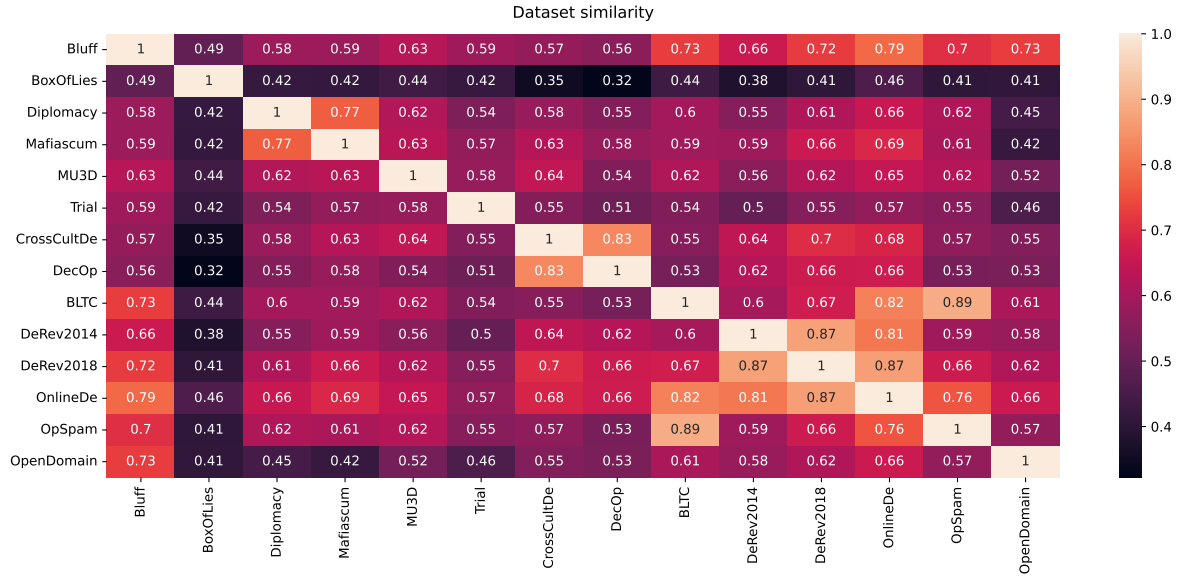

Figure 1: Similarity values, according to the measure proposed by Li and Dunn (2022), between all pairs of datasets.

using the corpus-similarity measure defined by Li and Dunn (2022), which uses word unigram frequencies and character trigram frequencies of the datasets to calculate the Spearman's $\rho \in [-1; 1]$.[18]

Figure 1 shows a symmetrical matrix of similarity scores for dataset pairs. The correlation values could in principle be negative, but we do not observe any such values because all corpora are in the same language and have a high degree of term and character frequency overlap.

The heatmap reflects the domains of datasets. For instance, BLTC, OPSPAM, as well as DEREV2014 and DEREV2018 from the review domain have similarity scores of 0.89 and .87, respectively. The opinion statement datasets CROSSCULTDE and DECOP exhibit a high similarity score of 0.83. Similarly, MAFIASCUM and DIPLOMACY show relatively high similarity (0.77), despite differences in the game rules.

Datasets obtained under specific conditions within the same domain are assigned a lower similarity score. For instance, BOXOFLIES, which is a game that takes place in an in-person setting, differs from the online game datasets (.42 with DIPLOMACY and MAFIASCUM). We also observe similarity across domains, e.g., BLUFF is more similar to reviews than games, presumably due to its monologue setting instead of dialogue.

[18]We use the Python implementation https://github.com/jonathandunn/corpus_similarity

5 Linguistic Correlation Analysis

To understand the generalizability of linguistic cues across different dataset, we conduct a correlation analysis, similar to previous studies that focused on isolated or smaller numbers of corpora (Pérez-Rosas and Mihalcea, 2015; Skalicky et al., 2020)

5.1 Method

We aim at identifying frequently used features which are general across domains. We build our analysis on the "Linguistic Inquiry and Word Count" (LIWC22[19], Pennebaker et al., 2015) and Flesch-Kincaid (Kincaid et al., 1975) and Gunning Fog (Robert, 1968) readability scores as measures of complexity or sophistication of language.[20]

We use point-biserial correlation[21] (Glass and Hopkins, 1996) to measure the relation between deception labels (discrete) and a score assigned by LIWC or readability measurement (continuous). The correlation value ranges from -1 to $+1$.

5.2 Results

Table 2 lists the features which show at least a weak correlation (> 0.15) with $p \leq 0.05$ for at least three datasets. The positive and negative correlation values correspond to the strength of association with truth and deception respectively.

[19]https://www.liwc.app/

[20]https://pypi.org/project/readability/.

[21]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pointbiserialr.html

Features	Datasets													
	BLTC	BLUFF	BOXOFLIES	CROSSCULTDE	DECOP	DEREV2014	DEREV2018	DIPLOMACY	MAFIASCUM	MU3D	ONLINEDE	OPENDOMAIN	OPSPAM	TRIAL
Analytic	**.13**	−.04	**.12**	.01	.02	**−.25**	**.23**	**.02**	**−.02**	**.14**	**.10**	**.05**	**.15**	**.25**
Authentic	.03	−.05	.00	**.28**	**.22**	**.28**	**−.05**	**−.03**	−.02	.07	.00	**−.04**	**−.09**	−.09
BigWords	.02	.00	**.18**	.04	**.05**	**−.21**	**.24**	.01	−.01	**.18**	**−.01**	**.03**	**−.08**	.09
Clout	.00	.00	.02	**−.11**	**−.28**	**−.45**	.00	**.02**	**.02**	.03	**−.05**	.01	**.10**	**.26**
Cognition	**−.08**	**.17**	−.05	.02	**.07**	−.06	**−.13**	−.01	−.01	**−.17**	.00	**−.09**	**−.06**	**−.28**
GunningFog	**.18**	**−.21**	**.12**	**.21**	**.25**	.01	**.13**	**−.09**	**−.03**	−.04	**.13**	.02	.02	.06
Kincaid	**.18**	**−.21**	**.14**	**.2**	**.24**	.01	**.13**	**−.08**	**−.03**	−.04	**.13**	**.03**	.02	.06
Linguistic	**−.07**	**.10**	**−.15**	.04	**.10**	**.29**	**−.14**	**−.02**	**−.03**	**−.16**	**−.05**	**−.05**	**−.18**	−.08
Period	.01	**−.07**	.02	**−.11**	**−.18**	**.26**	**−.07**	.00	.00	.03	**.01**	**.03**	**.24**	−.06
Physical	.02	.03	**.15**	−.04	**−.16**	**−.25**	**.06**	.00	**.03**	.04	**−.15**	−.01	−.01	.06
WC	**.18**	**−.21**	.04	**.22**	**.25**	.02	**.13**	**−.10**	.01	−.04	**.13**	−.02	.02	.06
auxverb	**−.08**	**.12**	−.06	**−.08**	**−.09**	**.22**	**−.12**	−.01	**.02**	**−.15**	.00	**.03**	**−.08**	**−.21**
focusfuture	**−.09**	**.09**	−.02	−.04	**−.08**	**−.17**	**−.2**	−.01	.02	−.04	**.01**	**−.04**	**−.16**	.08
function	−.05	**.13**	−.03	.00	**.10**	**.25**	**−.06**	**−.04**	**−.03**	**−.15**	**−.03**	**−.05**	**−.23**	**−.23**
i	**−.06**	**−.15**	−.07	**.13**	**−.3**	**.39**	**−.16**	**−.05**	.02	−.01	**−.12**	**−.04**	**−.33**	−.13
shehe	.01	**−.11**	−.03	**−.15**	.00	**−.17**	**−.07**	.00	**−.04**	**−.14**	**.04**	**−.04**	−.01	**−.18**
verb	**−.11**	.07	**−.09**	**−.06**	**−.07**	**.16**	**−.26**	**−.02**	.00	**−.14**	**−.07**	−.01	**−.16**	−.14
you	**−.10**	**.17**	−.03	−.05	**−.07**	**−.19**	**−.23**	.01	**.03**	−.08	**−.05**	**−.05**	.01	−.05

Table 2: Point-biserial correlation between the deception labels and linguistic features (LIWC categories + readability). We only show features with a correlation coefficient of $\geq .15$ and $p \leq .05$ for at least three datasets. Correlation scores with $p \leq .05$ are shown in bold.

Deceptive language is argued to have fewer self-references ("i") and more references to others ("shehe", "you"), as liars attempt to distance themselves from their lies (Newman et al., 2003; DePaulo et al., 2003). Our analysis supports this hypothesis in the categories "shehe" and "you" for a substantial number of data sets. Contrary to our expectation, however, in 8 out of 14 datasets the category "i" is seen to correlate with deception and not with truth, with an exception of CROSSCULTDE ($\rho = .13$) and DEREV2018 (.39).

Studies have attributed less cognitive complexity in language to deceptive communication (Newman et al., 2003; DePaulo et al., 2003). Liars use fewer words related to cognitive concepts (e.g., *think*, *believe*), which should correspond to a positive correlation value for the category "Cognition" in LIWC. However, our analysis corroborates this observation only in BLUFF ($\rho = .17$) and DECOP ($\rho = .07$).

In general, we found no consistent linguistic cues across domains and datasets in our analysis. This might be because deception is highly sensitive to the goal of a lie and the stakes involved, which is not consistent across the domains under consideration.

6 Deception Detection Experiments

The correlation analysis in the previous section showed that deception cues do barely generalize across domains. This analysis might be limited by the choice of categories, which motivates us to conduct cross-corpus modeling experiments.

6.1 Experimental Setup

In the within-corpus setup, we fine-tune and evaluate ROBERTa models (Liu et al., 2019) on the same dataset via 10-fold cross-validation. In the cross-corpus setting, we train on one corpus and test on the other. To ensure comparability between these experiments, we perform 10-fold cross-validation in both settings: we also evaluate 10 times on the same corpus subsets in the cross-corpus setup. This is not strictly required but ensures comparability.

We use the English ROBERTa-base, with 12 layers, 768 hidden-states, 12 heads and 125M parameters as available in the HuggingFace implementation (Wolf et al., 2020). We finetune with default hyperparameters for 6 epochs using the Auto Model for Sequence Classification. [22]

[22] https://huggingface.co/transformers/v3.0.2/model_doc/auto.html

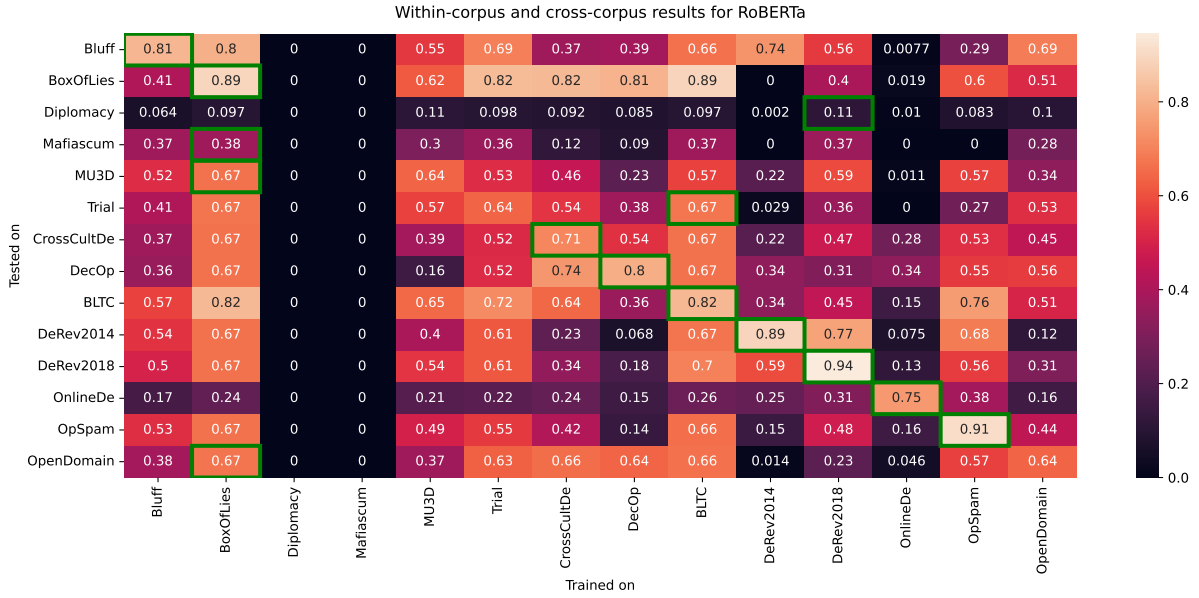

Figure 2: Performance of RoBERTa models with F_1 measure on the deception label. The best model on each test set is highlighted with a green box.

6.2 Results

The heatmap in Figure 2 shows the results as F_1 measure for the deception label (Appendix B shows results for both labels). The diagonal corresponds to within-corpus experiments. For most datasets, the model shows better performance in the within-corpus setting than in the cross-corpus evaluation. This is not the case for MU3D, TRIAL, and OPENDOMAIN, but the difference is negligible (0.04).

Models on datasets from the same domain or which are otherwise similar (§ 4) show comparably better results in the cross-corpus setting. For instance, training on OPSPAM and testing on BLTC achieves an F_1 score of 0.76 on the deception label. Training on BLTC and testing on OPSPAM is however not as good (0.66). Similar observations can be made for DEREV2014 and DEREV2018, and CROSSCULTDE and DECOP.

The heatmap shows the lowest performance for MAFIASCUM and DIPLOMACY, with an F_1=0. We assume that this is a result of the imbalanced label distribution in DIPLOMACY and the long documents in MAFIASCUM (see Table 1). Similarly, the exceptionally good results on the BOXOFLIES test set are due to the bias towards the deceptive label (see appendix for F_1 score on truth label).

Note that previous work reported other evaluation measures than F_1, which makes this dramatically low performance difficult to compare. Our evaluation with accuracy (shown in the appendix in Figure 4) appears to be more positive with .77 and .95.

From the sub-par results on cross-corpus experiments, we conclude that generalization across domains and dissimilar datasets is challenging, even with pre-trained language models with rich contextual information. In our future work, we plan to use this dataset to train models that can capture domain-independent cues of deception, which can presumably generalize better across datasets.

7 Conclusion & Future Work

Different scientific disciplines have contributed to the creation of deception datasets for textual communication in a variety of domains. In this study, we present a comprehensive survey of deception datasets in English available for research and compile them into a unified deception dataset. We are not aware of any previous work that considered a comparably large amount of corpora and evaluated models between all of them. Some of the evaluation results are encouraging, but particularly between dissimilar domains, the generalization is limited and requires future research.

The RoBERTa-based classification experiments and linguistic correlation analysis of deception cues demonstrate that it is indeed challenging to generalize the concept of deception across datasets, or domains. In the classification experiment results, the wildly diverging F_1 scores can be attributed to

the complexity of the task as well as to the limitations of the approach employed. In future work, we plan to explore the reasons for this variability across datasets further.

Additionally, we acknowledge the need to address the issue of biased models, such as the ones trained on MAFIASCUM, ONLINEDE, and DIPLOMACY, which tends to favor truthful labels owing to the label imbalance in these datasets, resulting in an F_1 score of 0. To overcome this challenge, we could employ techniques like oversampling to rectify the class imbalance and improve the reliability and effectiveness of our approach.

The goal of our future work is to create robust deception detection models that work reliably across corpora and domains. This includes understanding differences in the concept as it represents itself in these data and understanding differences in linguistic realization.

Our UNIDECOR dataset serves as a valuable resource for future research enabling standardized data comparison, transfer learning, and domain adaptation experiments.

Acknowledgments

We thank Kai Sassenberg for fruitful discussions regarding the concept of deception. Roman Klinger's research is partially funded by the German Research Council (DFG), projects KL 2869/1-2 and KL 2869/5-1.

Limitations

The goal of the current study was to unify the resources available for deception and report observations on cross-corpus and within-corpus analyses. While reporting the baseline performance using RoBERTa, we did not perform any optimization specific to the datasets. Hence, better results might be reported in the papers which handle the datasets or domains in isolation.

Ethical Considerations

The datasets used in this research are publicly available resources from previous studies. We have taken appropriate steps to ensure that we do not violate any license terms or intellectual property rights. Also, proper attribution is given to the original sources of the data. Deception is a sensitive topic, and non-anonymous data should not be used. To the best of our knowledge, all data sets that we considered have been compiled or collected according to such standards.

The performance of deception detection systems is not perfect, making them unsuitable for examining the utterances of individuals due to the threat of incorrect predictions. Even if automatic systems might reach a close-to-perfect performance, we consider their practical application to analyze and profile people unethical. However, there might be use cases, for instance in forensics, that can be considered ethical from a utilitaristic perspective.

Given the ethical implications of employing automated deception detection systems on individual, non-anonymous statements, we propose utilizing the resources collected and models developed on anonymous data. Any data analysis that could lead back to its origin must only be conducted with the data creator's informed consent and knowledge of potential consequences.

We consider the research in this paper to be fundamental, with the goal of better understanding human communication.

References

Gary D. Bond and Adrienne Y. Lee. 2005. Language of lies in prison: linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, 19(3):313–329.

Charles F. Bond Jr. and Bella M. DePaulo. 2006. Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3):214–234.

Pasquale Capuozzo, Ivano Lauriola, Carlo Strapparava, Fabio Aiolli, and Giuseppe Sartori. 2020. DecOp: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1423–1430, Marseille, France. European Language Resources Association.

Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Bob de Ruiter and George Kachergis. 2018. The mafiascum dataset: A large text corpus for deception detection. *ArXiv*, abs/1811.07851.

Bella M. DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin*, 129(1):74–118.

Bella M. DePaulo and Wendy L. Morris. 2004. *Discerning lies from truths: behavioural cues to deception and the indirect pathway of intuition*, page 15–40. Cambridge University Press.

Nicholas D. Duran, Charles Hall, Philip M. McCarthy, and Danielle S. McNamara. 2010. The linguistic correlates of conversational deception: Comparing natural language processing technologies. *Applied Psycholinguistics*, 31(3):439–462.

Amir Fayazi, Kyumin Lee, James Caverlee, and Anna Squicciarini. 2015. Uncovering crowdsourced manipulation of online reviews. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 233–242, New York, NY, USA. Association for Computing Machinery.

Tommaso Fornaciari, Federico Bianchi, Massimo Poesio, and Dirk Hovy. 2021. BERTective: Language models and contextual information for deception detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2699–2708, Online. Association for Computational Linguistics.

Tommaso Fornaciari, Leticia Cagnina, Paolo Rosso, and Massimo Poesio. 2020. Fake opinion detection: how similar are crowdsourced datasets to real data? *Language Resources and Evaluation*, pages 1–40.

Tommaso Fornaciari and Massimo Poesio. 2014. Identifying fake Amazon reviews as learning from crowds. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 279–287, Gothenburg, Sweden. Association for Computational Linguistics.

G.V. Glass and K.D. Hopkins. 1996. *Statistical Methods in Education and Psychology*. Allyn and Bacon.

Swati Gupta, Kayo Sakamoto, and Andrew Ortony. 2013. Telling it like it isn't: A comprehensive approach to analyzing verbal deception. Online.

Jeffrey T. Hancock. 2009. Digital deception: Why, when and how people lie online. In *Oxford Handbook of Internet Psychology*. Oxford University Press.

Jeffrey T. Hancock, Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. 2007. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23.

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

Ángel Hernández-Castañeda, Hiram Calvo, Alexander Gelbukh, and Jorge J. García Flores. 2016. Cross-domain deception detection using support vector networks. *Soft Computing*, 21(3):585–595.

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report 8-75, University of Central Florida, Institute for Simulation and Training.

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

Haipeng Li and Jonathan Dunn. 2022. Corpus similarity measures remain robust across diverse languages. *Lingua*, 275:103377.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach.

Paige E. Lloyd, Jason C. Deska, Kurt Hugenberg, Allen R. McConnell, Brandon T. Humphrey, and Jonathan W. Kunstman. 2019. Miami university deception detection database. *Behavior Research Methods*, 51:429–439.

James Edwin Mahon. 2007. A definition of deceiving. *International Journal of Applied Philosophy*, 21(2):181–194.

Sandra Metts. 1989. An exploratory investigation of deception in close relationships. *Journal of Social and Personal Relationships*, 6(2):159–179.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675.

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501, Atlanta, Georgia. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 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*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

James W. Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 59–66, New York, NY, USA. Association for Computing Machinery.

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

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1125, Lisbon, Portugal. Association for Computational Linguistics.

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3811–3854, Online. Association for Computational Linguistics.

Gunning Robert. 1968. *The Technique of Clear Writing*. McGraw-Hill, New York.

Franco Salvetti, John B. Lowe, and James H. Martin. 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'16)*, pages 3510–3517, Portorož, Slovenia. European Language Resources Association (ELRA).

Stephen Cameron Skalicky, Nicholas D. Duran, and Scott Andrew Crossley. 2020. Please, please, just tell me: The linguistic features of humorous deception. *Dialogue Discourse*, 11:128–149.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. 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)*, pages 1768–1777, Minneapolis, Minnesota. Association for Computational Linguistics.

Lyn M. Van Swol, Michael T. Braun, and Deepak Malhotra. 2012. Evidence for the pinocchio effect: Linguistic differences between lies, deception by omissions, and truths. *Discourse Processes*, 49(2):79–106.

Leanne Ten Brinke and Stephen Porter. 2012. Cry me a river: identifying the behavioral consequences of extremely high-stakes interpersonal deception. *Law and Human Behavior*, 36(6):469–477.

Ronny E. Turner, Charles Edgley, and Glen Olmstead. 1975. Information control in conversations: Honesty is not always the best policy. *The Kansas Journal of Sociology*, 11(1):69–89.

Aldert Vrij. 2014. *14. Detecting lies and deceit: Pitfalls and opportunities in nonverbal and verbal lie detection*, pages 321–346. De Gruyter Mouton, Berlin, Boston.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenlin Yao, Zeyu Dai, Ruihong Huang, and James Caverlee. 2017. Online deception detection refueled by real world data collection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 793–802, Varna, Bulgaria. INCOMA Ltd.

Lina Zhou, Judee K. Burgoon, Jay F. Nunamaker, and Doug Twitchell. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation*, 13(1):81–106.

Miron Zuckerman, Bella M. DePaulo, and Robert Rosenthal. 1981. Verbal and nonverbal communication of deception. In Leonard Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 14, pages 1–59. Academic Press.

M. Umut Şen, Verónica Pérez-Rosas, Berrin Yanikoglu, Mohamed Abouelenien, Mihai Burzo, and Rada Mihalcea. 2022. Multimodal deception detection using real-life trial data. *IEEE Transactions on Affective Computing*, 13(1):306–319.

Appendix

A Details on the Aggregated Dataset

All datasets included in the unified collection contains one binary label indicating whether an instance is truthful or deceptive, the naming convention for which has been normalized retaining the original label for backward compatibility. However, some datasets like Salvetti et al. (2016) and Peskov et al. (2020) include an additional dimension for deception, where the former differentiates between lying about a known object and lying about an unknown object, and the latter contains annotations on the *perceived truthfulness* of the statement in addition to the actual intention. For providing a unified format, we map both these deceptive instances in Salvetti et al. (2016) to one label and since the perceived truthfulness is an independent annotation, we do not take this label into account.

In addition to truth labels, datasets also contain additional annotations like demographic information related to the author, sentiment, valency of the instance and perceived truthfulness. We retain only those annotation dimensions which are available for more than one dataset which are age, gender, country, and sentiment

The unified dataset includes corpora that are available for research purposes which are downloadable from source, made available directly by the creators, or obtained from a consortium like the Linguistic Data Consortium. We provide a script to automatically download all datasets if they are available for download, which otherwise provides instructions on how to obtain them. Once all datasets are populated in their respective folders, a second script is used to generate the unified dataset in json format. You can find the repository with instructions to obtain the aggregated UNIDECOR, Unified Deception Corpus, at https://www.ims.uni-stuttgart.de/data/unidecor.

The following entry shows an example instance from the corpus.


```
{
        "source": "OPEN_DOMAIN",
        "text_ID": "119_f_t_1",
        "text": "Thad cochran has been in the us senate since before the internet
            was
        invented.",
        "participant_ID": "NA",
        "age": "20",
        "sentiment": "NA",
        "language": "EN",
        "gender": "Female",
        "country": "United States",
        "original_label": "truth",
        "truth_label": "T",
        "topic_name": "statement",
        "domain": "opinion",
        "mode": "written",
        "split": null,
        "fold": null
}
```

B Additional Experimental Results

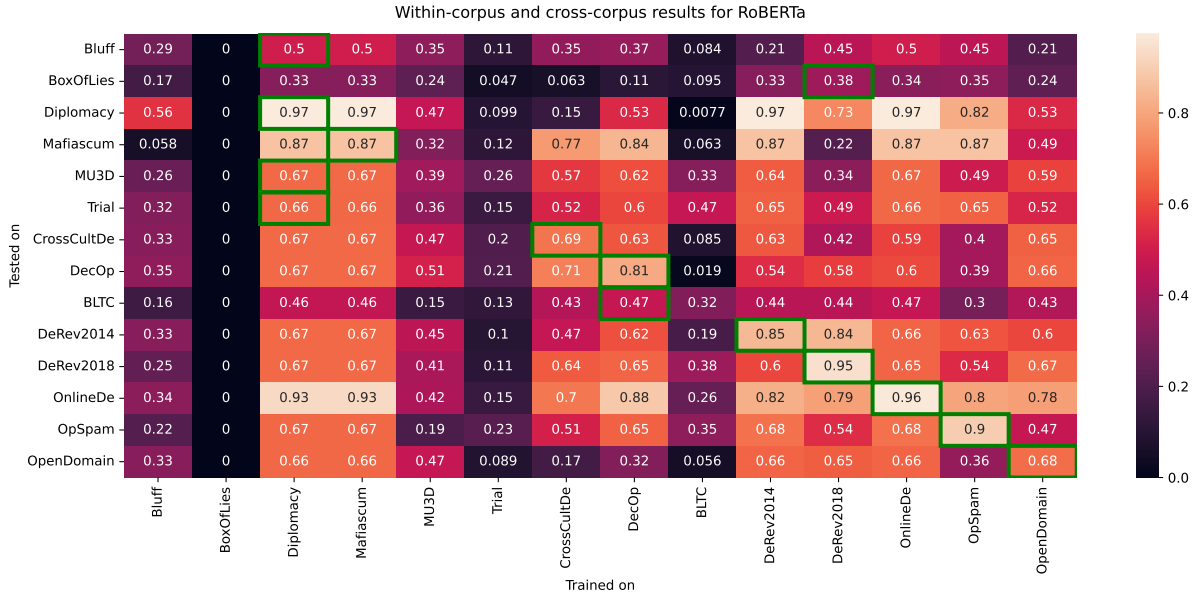

Figure 3: A heatmap representing the performance of RoBERTa model with the F_1 measure on the truth label across different datasets. Figure 2 in the main paper analogously shows the results for the deception category.

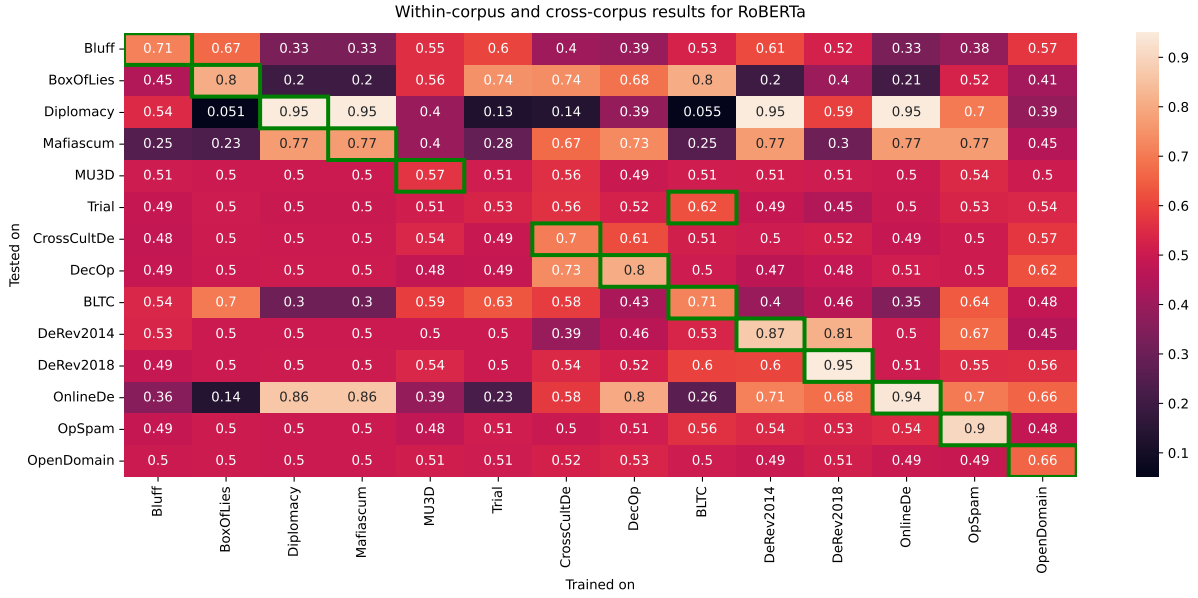

Figure 4: A heatmap representing the accuracy of RoBERTa model different datasets. As the categories of truth and deception and mutual exclusive in all our datasets, this corresponds to a micro-average of the results shown in Figure 2 and 3.