STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LXIX**, Number 1, 2024
DOI: 10.24193/subbi.2024.1.05

AUTOMATIC DETECTION OF VERBAL DECEPTION IN ROMANIAN WITH ARTIFICIAL INTELLIGENCE METHODS

MĂLINA CRUDU

Abstract. Automatic deception detection is an important task with several applications in both direct physical human communication, as well as in computer-mediated one. The objective of this paper is to study the nature of deceptive language. The primary goal of this study is to investigate deception in Romanian written communication. We created a number of artificial intelligence models (based on Support Vector Machine, Random Forest, and Artificial Neural Network) to detect dishonesty in a topic-specific corpus. To assess the efficiency of the Linguistic Inquiry and Word Count (LIWC) categories in Romanian, we conducted a comparison between multiple text representations based on LIWC, TF-IDF, and LSA. The results show that in the case of datasets with a common subject such as the one we used regarding friendship, text categorization is more successful using general text representations such as TF-IDF or LSA. The proposed approach achieves an accuracy of the classification of 91.3%, outperforming the similar approaches presented in the literature. These findings have implications in fields like linguistics and opinion mining, where research on this subject in languages other than English is necessary.

1. Introduction

Automated deception detection merges fields of research such as sociology, interpersonal psychology, communication studies, philosophy, and computational models of deception detection. The recognition of a misleading way of behaving is a task that has acquired expanding interest because of the quick development of deception in written sources, particularly the ones from cyberspace. Moreover, its applications in identifying potential harm for people and society make this challenge a relevant one and necessary to be resolved. In general, the benefits of solving this problem are reflected in domains such as

Received by the editors: 29 April 2024.

2010 *Mathematics Subject Classification.* 68T50.

1998 *CR Categories and Descriptors.* I.2.7 [**ARTIFICIAL INTELLIGENCE**]: Natural Language Processing –*Text analysis*.

Key words and phrases. Deception Detection, Text Classification, Natural Language Processing, Machine Learning.

business, jurisprudence, law enforcement, and national security. Text-based information of any structure, such as news articles, client surveys, political discourses, social media contents, witnesses' reports, and so on are at present used in deception research as they portray the ideal context of lying in genuine circumstances.

Considering the problem of automatic detection of deceptive language in Romanian written texts, relatively modest efforts, if any have been made, the spotlight being put on languages that are widely spoken such as English, Spanish, or Italian. Most of the previous work has focused on the psychological or social aspects of lying. They concentrated on deceiving and its relation to the cultural dimension of individualism/collectivism and not so much on the specificity of linguistic aspects of falsehood in Romanian.

Taking into account the introduced issue, there are various difficulties that can be noticed. One of these is the fact that the information provided does not present any additional data apart from the written language itself.

Researchers often take deception language as a whole, overlooking individual highlights that may distinguish one speaker from the others, and assuming that everyone lies in much the same way. Rather than comparing each individual sample of misleading language to its equivalent control text, the complete collection of "false" testimonies is compared against the set containing "true" claims. It is worth noting that the fundamental disadvantage of a corpus of "genuine" language is the difficulty in getting a sample of instances of language in which a speaker tells the truth for the purpose of comparison. Taking these factors into account, this study aims to analyze deception indications in written Romanian, which is a unique area of study, that has not been explored yet.

Because the implicit assumption about the homogeneity of language indications of deception contradicts earlier work from psychological and sociological disciplines and raises fundamental problems about the application of current deception tools on Romanian texts, our key research goals are:

RG1. Explore which language markers and indications are more successful in distinguishing deception given a piece of Romanian text on the topic of friendship.

RG2. Related to the previous research goal, we create and evaluate the effectiveness of a wide range of binary classifiers for predicting the truthfulness and deceptiveness of texts.

RG3. Determine whether or not the Latent Semantic Analysis (LSA) method is better suited for this task compared to other different data representations such as the ones based on Linguistic Inquiry and Word Count (LIWC) or term frequency-inverse document frequency (TF-IDF).

The paper is organized as follows. Work related to our problem is discussed in Section 2. The methodology described in Section 3 incorporates the preprocessing stage, text representations, and data analysis. Section 4 is concerned with the development of the classification models. Section 5 presents the experimental results and discussions. Conclusions and directions for further research are presented in Section 6.

2. Related work

There are verbal signals to deceptive behavior that are part of the existing verbal lie detection methods utilized by professional lie catchers and scholars [16]. Automatic linguistic methods have been utilized to analyze the linguistic elements of the constitution of deceptive language in English generally.

Typically, specialists have used the word classes specified in the Linguistic Inquiry and Word Count, or LIWC [12], which is a linguistic examination tool that creates a taxonomy of words based on psychologically meaningful categories. It has been utilized to investigate matters such as personality, psychological acclimation to different changes, social judgments, tutoring dynamics, and mental health.

LIWC was for the first time used by Pennebaker's research group for several studies on the language of deception [9]. Through five different experiments, they collected a corpus of real and fake texts as part of their research. The factors that were considered to be relevant predictors in at least two of the experiments were: self-reference terms, references to others, exclusive words, negative emotion elements, and motion words. The justification behind the underperformance in a number of studies might be the fact that the verbal signals of deception in oral contact do not transpose in written communication and the other way around.

LIWC has been used for the examination of deception in written language. Research in this field has been addressed by computational linguists and a relevant example is represented by [8], who applied LIWC for post hoc analysis, evaluating many linguistic characteristics on a corpus of 100 fake and true statements on three contentious themes - the survey being similar to [9]. As an initial experiment, they used two ML classifiers: Naïve Bayes and Support Vector Machine. Both algorithms have been trained using word frequencies, like a Bag-of-Words model. They achieved an average classification accuracy of 70%, which is altogether higher than the 50% baseline. Based on this information, they computed a dominance score linked with a certain word class within the set of misleading texts as a measure of salience. The word coverage, or the linguistic item's weight in the corpus, was then calculated. Therefore, they determined some particular characteristics of deceptive texts.

In this strand of research, [10] used the same two ML classification algorithms. For their training, apart from drawing a comparison between lexically-based deception classifiers and a random guess baseline, the authors additionally assessed two more automatic approaches: genre recognition by analyzing the frequency distribution of parts of speech (POS) tags, and a text classification method which enables them to model both content and context with n-gram features. Their final goal was to identify fraudulent opinion spam, which is a fundamentally different challenge from the problem of identifying dishonest language. When it comes to detection, findings reveal that text classification based on n-grams is the best technique; nevertheless, combining LIWC features and n-gram features is the solution in order to achieve somewhat superior results.

Similar scientific endeavors as [8] were made by [1]. The importance of this paper comes from the novelty of exploring deception in the Spanish language and creating a comparison with similar studies that follow English as the main focus in order to uncover structural and lexical variations in the linguistic manifestation of deception in both languages. This paper describes an artificial intelligence model based on a Support Vector Machine (SVM) for detecting dishonesty in an ad hoc opinion corpus composed of various Spanish written communication texts. The questionnaire for the corpus compilation was designed similarly to that used by [8]. The created framework tests the effectiveness of the LIWC2001 categories in Spanish compared with a Bag-of-Words (BoW) model. The results emphasize the discriminatory power of the variables, the two first dimensions, linguistic and psychological processes, being the most relevant ones from the LIWC categories.

These investigations manage written language as utilized in asynchronous methods of communication, while Hancock and his research group investigated deceptive language in real-time computer-mediated communication (CMC), in which all members are online simultaneously using chat rooms. [6] explored dissimilarities between the transmitter's and the recipient's linguistic way across honest and deceptive communication in their initial research based on LIWC. They picked the elements thought to be important to the hypotheses for this study, which were word counts, pronouns, emotion words, sense terms, exclusive words, negations, and inquiry frequency. The findings revealed that when respondents lied, they were more chatty, using more words, more allusions to others, and more sense-related vocabulary.

3. Methodology

It is worth mentioning that during this study, we also created our own deception dataset of autobiographical narratives which was a non-topic-specific

dataset. All the experiments and data analysis we carried out were also done on that set of data but the results were not competitive further proving the difficulty of this classification problem, especially in the context of diverse narrative settings where a common subject of discussion is absent.

3.1. **Dataset.**

To study the distinction between true and deceptive statements, we used the only such data set, to the best of our knowledge, which is the deception dataset mentioned in [13], which covers four distinct cultures: the US, India, Mexico, and Romania. Each part of this dataset comprises short deceptive and truthful statements on three subjects of discussion: beliefs on abortion, views on capital punishment, and sentiments about a best friend. In this research, we used only the ones related to best friends as this topic is the most generic one and can replicate better how people lie on common topics.

The data extracted from the [13] dataset were gathered from native Romanian speakers using a web interface. The respondents have been enlisted through contacts of the paper's authors [11]. For the third subject (best friend), the participants in that study were first asked to meditate about their best friend and detail the motives behind their fellowship (incorporating facts and stories considered important for their relationship). Accordingly, for this situation, they were requested to express their true sentiments about how they felt about their best friend. Next, they were required to imagine an individual they could not stand, and depict their relationship with that person as if they were their best friend. In this subsequent case, they needed to lie about their emotions towards this individual.

In all cases, the instructions requested no less than 4 to 5 sentences and as numerous details as possible. Altogether, there were gathered 149 true and 149 false testimonies about best friends with an average of 78 words per statement. Furthermore, manual verification of the quality level of the input was made.

For ease of understanding and explanations, we decided to use a suggestive name for the dataset. As it is a topic-specific dataset, focused on the subject of best friends, the dataset will be from now on referred to as the BestFriend dataset. In Table 1, we included some examples from the BestFriend dataset, divided by class, which in this context is the level of truthfulness.

3.2. **Data preprocessing and representation.**

3.2.1. *Preprocessing.*

This stage is concerned with the preparation of deceptive and true texts before extracting relevant features. As part of the data preparation, several operations were performed. To begin with, we converted all the capital letters

TABLE 1. Dataset examples

Deceptive statement	**Truthful statement**
Mereu mă ajută când am nevoie. Dacă nu înțeleg ceva este foarte răbdătoare și îmi explică până la capăt. Nu este niciodată invidioasă pe mine. Ne înțelegem de minune.	Suntem cei mai buni prieteni deoarece ne putem spune orice în fata fără sa ne deranjeze, avem aceleași concepții și idei, ne ajutam la greu și petrecem la bine. Putem discuta o problema personala fără sa afle încă 10 oameni.

to lowercase and all punctuation marks were eliminated (they are always used in any correctly written text, but they do not carry any specific information required to train the model for this problem). For the next operation, we used the LIWC lexicon and a dictionary with Romanian words and their lemmas. We used either the word or its lemma if the word did not exist in the LIWC lexicon. After all the above-mentioned preprocessing had been done, we reconstructed the phrase with space as a separator between each word.

3.2.2. *Representation.*

Our study is based on a textual representation that is somehow different from the general models that are used in NLP, such as TF-IDF or BoW representations, but it preserves their intuition. This representation is based on the Linguistic Inquiry and Word Count lexicon.

Linguistic Inquiry and Word Count or LIWC, is a tool for textual examination where words are divided into psychologically relevant groups. The Romanian version of this lexicon incorporates 47,825 entries and is organized into 73 categories related to psychological processes. This taxonomy offers an effective technique for examining the emotional, cognitive, and structural components contained in language on a word-by-word basis. Words and word stems are classified in the LIWC internal lexicon along four broad dimensions: standard language processes, psychological processes, relativity, and personal concerns [4]. Each word or word stem is characterized by at least one of the 73 default word categories.

From all the categories, we chose the most relevant classes according to different studies that investigated a similar problem as the one stated in our research: [8], [7], [1] and [9]. These categories would be: **self-reference terms**, **references to others**, **negative emotion elements**, **motion words**, **belief-oriented vocabulary**, **words related to certainty**, **negation terms**,

sense terms and **positive emotion elements**. Table 2 contains examples of relevant instances of words that belong to the LIWC categories.

TABLE 2. LIWC categories and relevant examples

LIWC category	**Examples**
self-reference terms	"eu", "îmi", "înșine", "mi-"
references to others	"însăți", "își", "l-", "le"
negative emotion elements	"panicat", "neliniștit", "smiorcăi", "amărât"
motion words	"fugi", "prăbușire", "împiedicat"
belief-oriented vocabulary	"bănui","gândire", "reține"
words related to certainty	"bineînțeles", "categoric", "iminent"
negation terms	"fără", "n-aș", "nicăieri"
sense terms	"privitor", "pălăvrăgea", "înșfăca"
positive emotion elements	"acceptat", "mulțumire", "valoros"

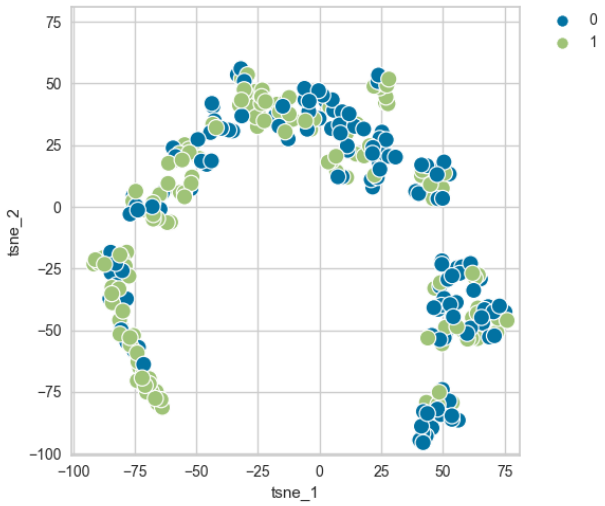
These categories are used as they are considered related to deception, for instance, an increased sense terms linguistic variable indicates deception, as liars attempt to create a detailed story. To give another example, there is less self-reference in false narratives and more frequent references to third parties and objective elements. This suggests impersonality as the liar tries to increase the narrative distance [2].

Moreover, we added two more relevant values in the feature vector of each text, more specifically **the number of words** presented in each text and the **Type-Token Ratio** (TTR). The TTR is defined as the ratio of unique tokens (types) divided by the total number of tokens. We added these measurements as it is generally considered that liars tend to produce more words during deceptive discussions [7] and the deceptive narratives are expressed with a higher syntactical simplicity [3], thus a lower Type-Token Ratio that evaluates a person's verbal diversification and asses textual richness.

For each text that was provided, we created a count vector with eleven values in which the first nine represents a category of words, more specifically the number of words that fit into this category and are included in the text and the last two represent the number of words and the TTR.

3.3. **Data analysis.**

Before proceeding to create the machine learning models we wanted to study the difficulty of the classification task that we are trying to solve and in order to do that we used a number of techniques.



0: Deceptive class 1: Truthful class

FIGURE 1. t-SNE algorithm applied on the BestFriend Dataset

3.3.1. *Investigate difficulty of the classification.*

To determine the difficulty of the classification task, we employed **t-distributed Stochastic Neighbor Embedding** (t-SNE) algorithm. It computes a non-linear dimensionality reduction which allows us to separate and visualize data that cannot be separated by a straight line. After we ran different tests, with different parametrizations (such as different perplexity and number of iterations) of the algorithm, we concluded that there is a tendency towards clearer shapes as the perplexity value increases. Applying the t-SNE algorithm, we obtained a semicircle shape (Figure 1), the graphic having a tendency to differentiate the two classes at the opposite poles of the figure.

3.3.2. *Feature Relevance.*

As we previously mentioned in the RG1 goal, we want to explore the quality of linguistic markers in deception detection. In order to do that, another approach that we considered was the investigation of the relevance of the extracted features.

Pearson Correlation Coefficient

More precisely, we computed the Pearson Correlation Coefficient [5] to examine the relationships between all features within the dataset. Additionally,

we also calculated the coefficients between the features and the labels to assess their predictive power. Given the values of the correlations between each feature and the set of labels, computing the Pearson Correlation Coefficient on the BestFriend Dataset, provided us with values between -0.183703 and 0.350826. By sorting the values of the coefficient we deduced that the most relevant features are: self-reference words, the number of words, belief-oriented vocabulary, sense terms, vocabulary related to positive emotions, references to others, and lastly, motion words. When it comes to features, there is a fairly strong positive relationship (correlation coefficient with a value above 0.7) between the number of words and the following features: self-reference terms, references to others, and belief-oriented vocabulary. This is expected as the dataset focuses on autobiographical stories, relationships with people, and opinions on them. Moreover, a moderate positive correlation (value above 0.6) exists between the number of words and both words related to certainty and positive emotion elements. This would suggest that people talk more when they experience positive emotions but also when they are or try to emulate a sense of certainty.

Relief Algorithms

Correspondingly to what we expressed in the latter paragraph, we wanted to deepen our analysis and we also applied feature selection using Relief algorithms [15]. Relief calculates a score for each feature expressing the relevance of that feature for the output label. The scores are used to rank and choose top-scoring features for feature selection. For our analysis, we looked at features based on their relief scores, prioritizing those with higher values such as self-references, insight vocabulary, the number of words, sense terms, positive emotion words, words related to certainty, negative emotion terms, and lastly, the TTR. Comparing these results with the ones obtained via the Pearson Coefficient, we can conclude that both algorithms found relevant five categories: self-references, belief-oriented vocabulary, sense terms, positive emotion terms, and the number of words.

4. Developing the Classification Model

After pre-processing and feature extraction, we wanted to evaluate three different classifiers: Support Vector Machine(SVM), Random Forest (RF), and Artificial Neural Network (ANN).

To develop the deception classifiers of the first two above-mentioned classification algorithms, we used the Scikit-learn (Sklearn) library. For all the classifiers we used the default parametrization given by the library. More precisely, for RF the criterion is set on *gini* and the $n\_estimators$ is 100. For SVM we did not modify the kernel function from the default value of Radial Basis

Function Kernel or rbf. For the ANN we employed the Keras deep learning framework for constructing our neural network architecture. The model comprises two Dense layers: the first layer consists of 64 neurons with a rectified linear unit (ReLU) activation and the second layer is a single neuron output layer with a sigmoid activation function. The model was compiled using the Adam optimizer and the binary cross-entropy loss function to optimize the network's performance in the binary classification task. Additionally, we utilized accuracy as the evaluation metric. The training process involved fitting the classifier to the training data using a $batch\_size$ of 32 and training for 25 epochs, with a validation split of 0.2.

For the models' training and testing, we did the experiments using 5-fold Stratified Cross-Validation. The 5-fold Stratified Cross-Validation ensures that each fold is then used once as validation while the four remaining folds form the training set and that each is made by maintaining the percentage of samples for each class. This way we have a division of 80% of the data being used in the training process and 20% for testing. The next section presents the results we obtained by our three classifiers during a number of different experimental setups.

5. Results and discussions

Consistent with what we expressed in the **RG2** research goal, we created and evaluated a number of binary classifiers along with linguistic models. The results that we obtained and a conclusive discussion based on the outcomes of our experiments are presented in the current section.

5.1. **Experimental results.**

To implement our machine learning models, we employed Python 3.7 and the Windows operating system. The tables in this section summarize the results of our experiments in terms of accuracy and F1-Score in the testing step for the three various classifiers that have been utilized on the linguistic models. These metrics are expressed in the form of confidence intervals (CI) that have a 95% confidence level. For these calculations the next formula was used:

$$CI = \bar{x} \pm \frac{z * \sigma}{\sqrt{n}}$$

where:

- $\bar{x}$ is the mean of the testing accuracies
- n is the sample size
- σ is the standard deviation of the testing accuracies
- z is the confidence coefficient, which is 1.96 for a 95% confidence level

For the experiments that we conducted, we used different linguistic models that will be presented in the following.

Experiment 1. LIWC-based model: 11 features

The first linguistic model uses in the feature representation all the initial attributes that were selected, 9 characteristics computed based on the LIWC lexicon (the number of: **self-reference terms**, **references to others**, **negative emotion elements**, **motion words**, **belief-oriented vocabulary**, **words related to certainty**, **negation terms**, **sense terms**, **positive emotion elements**), to which we added the **number of words** and the **TTR**. The results obtained for the LIWC-based model, using all three classifiers, can be consulted in Table 3. Along with this experiment we tried to use all of the 73 categories of the LIWC lexicon, along with the two features added by us (the number of words and the TTR), but the results were extremely similar to the ones obtained via our linguistic model with only 11 features.

TABLE 3. Testing accuracy and F1-Score for the LIWC-based model on the BestFriend dataset

Classifier	**Accuracy (CI%)**	**F1-Score (CI%)**
SVM	0.658±0.099	0.622±0.126
RF	0.715±0.065	0.707±0.055
ANN	0.731±0.034	0.735±0.072

During the tests, we took into consideration the analysis we conducted on the dataset and the results presented in Subsection 3.3, and we created some simplified linguist models, with only the features that were found to be qualitative attributes in our study. We tried several set-ups such as using only the features found relevant by either the Pearson Coefficient or by the Relief Algorithms and also tried creating a model with features from both. Given this context, we retrained and tested these leaner models, however the results we obtained showed a decrease compared to the initial LIWC-based model with 11 features.

Experiment 2. TF-IDF model

To draw a comparison between the linguistic model based on the LIWC lexicon and general representations used in Natural Language Processing, we trained the classifiers that use a TF-IDF representation. This representation was obtained by utilizing the $TfidfVectorizer$ with $smooth\_idf$ on $True$ to prevent zero divisions and the min_df on 0.001 to ignore terms that have a document frequency strictly lower than the given threshold.

Experiment 3. LSA model

As previously stated in our **RG3** goal we wanted to draw a comparison between a TF-IDF representation and a LSA one. Latent Semantic Analysis or LSA is a technique that learns latent topics by decomposing or factorizing the document-term matrix such as the TF-IDF matrix using a mathematical technique known as Singular Value Decomposition or SVD. The purpose of Latent Semantic Analysis is to reduce the dimensionality of the corpus vector space while detecting higher-order patterns within it.

For the LSA representation, the TF-IDF vectors were mapped by calling $TruncatedSVD$ with a number of 300 topics that have a variance of 99%. The topic value was chosen in regard to the variance ratio graph that we plotted for the dataset based on the TF-IDF representations and we chose the value that presented the highest value. The plot can be visualized in Figure 2 and was created by calculating the total variance ratio as the sum of the variances explained by each of the selected components for all the possible values (from one to the total length of the vocabulary).

For the TF-IDF and LSI representations we experimented with various token N-gram sizes (from 1-gram to 5-grams), but we decided to utilize a smaller subset, just from unigrams and 2-grams, as the discriminating capability of the other values as types of N-grams proved to be extremely limited, and as a result, these findings were omitted. Additionally, we also experimented with Principal Component Analysis (PCA) representations but found results similar to those obtained with LSA.

TABLE 4. Testing accuracy and F1-Score for the TF-IDF and LSA models on the BestFriend dataset

Classifier	**N-grams**	**TF-IDF representation**		**LSA representation**	
		Accuracy	F1-Score	Accuracy	F1-Score
SVM	1-gram	0.785±0.082	0.791±0.095	0.789±0.086	0.797±0.08
	2-grams	0.718±0.092	0.726±0.074	0.678±0.137	0.669±0.153
RF	1-gram	0.755±0.068	0.753±0.082	0.678±0.038	0.675±0.062
	2-grams	0.715±0.021	0.665±0.056	0.668±0.088	0.673±0.087
ANN	1-gram	0.913±0.069	0.914±0.064	0.89±0.141	0.88±0.165
	2-grams	0.896±0.15	0.904±0.127	0.849±0.185	0.837±0.215

The values in the tables confirm that the classification task is solved more successfully in the case of the neural network-based classifier across all of the linguistic models. This suggests the potential of neural network architectures in similar classification tasks such as deception detection in legal contexts as courtroom cases would represent. Secondly, despite employing LSA (Latent

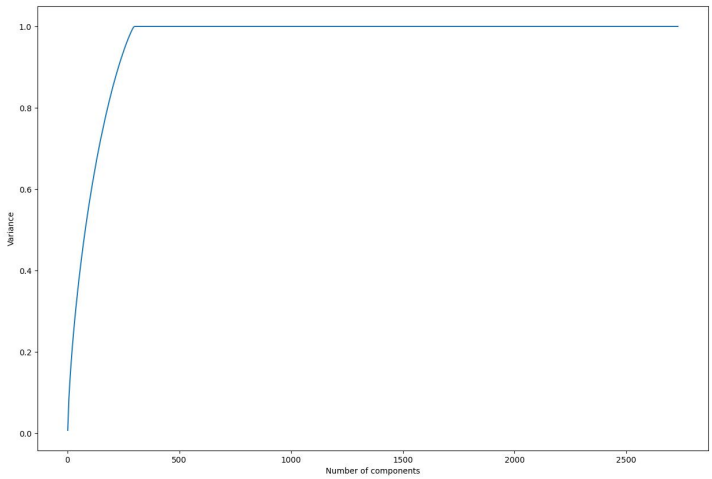


FIGURE 2. Graphic of the coherence score depending on the number of topics for the Best Friend dataset

Semantic Analysis) dimensionality reduction, we did not observe improvements in the studied metrics. This indicates that LSA may not enhance the performance of the classifiers in this context nor capture the semantic complexity of the specificities of deceit. Furthermore, increasing the size of N-grams did not result in improved performance metrics. This finding suggests that simply expanding N-grams may not necessarily enhance the classifier's performance as it might introduce data sparsity. This might also be a result of the fact that Romanian, even though it generally follows a Subject-Verb-Object (SVO) word order, is more flexible than English in terms of word order variation meaning that two n-grams could contain the same words but in a different order. Another conclusion that we draw is related to the linguistic models that we designed, especially the ones that do not use LIWC as a base for feature vector creation. Even though TF-IDF and LSA are considered to be general models, they were able to achieve in most cases better results than the models trained with psychologically relevant attributes. This made us conclude that a major part of the deceptive process in the case of topic-specific statements is not related to which category of words we use, but which terms we utilize.

5.2. **Comparison with SOTA.**

Even though our study is considered to be a novelty, from the point of view of experimenting on Romanian datasets, we tried to draw a comparison between the results that we obtained and the performance of other similar approaches 5, even though they are not all implemented on a topic specific dataset or they are following different languages. Considering this aspect, we chose a sample of researches to compare their results with the ones we achieved.

Firstly, [14] created a new open-domain deception dataset that also includes demographic data such as gender and age. Even though the methods that obtained the best performance are not similar to the ones conducted by us, and the dataset has a somewhat autobiographical topic, the authors tried several sets of features, including semantic features based on the LIWC lexicon. This approach had only an accuracy of 60.21% compared to the maximum of 69.50% obtained via part-of-speech (POS) tags.

Next, a more similar approach to ours in terms of the classification algorithms that have been used, the selected features, and the data utilized for research is presented in [13]. Even though the study presents experiments made on a cross-cultural dataset, a comparison deserves to be done as our experiments were made on the Romanian version of the dataset collected for the mentioned paper.

From all the research we evaluated, [1] is the closest one to our approach in terms of methodology, dataset, and target language. The research is the exploration of a non-English language, more precisely on Spanish written communication. They have designed an automatic classifier based on SVM and the dataset is created similarly to the one mentioned in [8]. We consider this comparison to be the most relevant one as it is done based on a language close to Romanian, the topic of the dataset is the same as ours and the methodology is similar.

6. Conclusions and Future Work

Although many artificial intelligence models for automatic deception detection were implemented, most of them were for English texts, the Romanian Language being somehow neglected. In this paper, we researched an important Natural Language Processing task, analyzed a topic-specific dataset, and investigated automatic methods for the identification of deceptive language in written Romanian statements on the topic of friendship, using several representations for their training such as the LIWC psycho-linguistic categories, TF-IDF and LSA. By comparing different algorithms and evaluating their output we achieved a 91.3% accuracy in terms of detecting deception, which

TABLE 5. Comparison between our models and relevant research

Dataset	Features	Classifier	Classification performance
Open Domain Deception Dataset [14]	All categories from the LIWC lexicon	SVM	Accuracy 60.21%
Open Domain Deception Dataset [14]	POS tags	SVM	Accuracy 69.50%
Best Friend Spanish Dataset [1]	All categories from the LIWC lexicon	SVM	F1-Score 84.5%
Best Friend English Dataset [13]	Linguistic categories from the LIWC lexicon	SVM	Accuracy 75.98%
Best Friend Romanian Dataset	**TF-IDF unigram representation**	**ANN**	**Accuracy 91.3% F1-Score 91.4%**
Best Friend Romanian Dataset	LSA unigram representation	ANN	Accuracy 89%
Best Friend Romanian Dataset	Selected categories LIWC representation	ANN	Accuracy 73.1%

represents competitive results that outperform similar methodologies we used as state-of-the-art approaches for this task.

As for future work, although the classification algorithms provided positive results, there are a number of improvements that can be mentioned. One development can be made in terms of the datasets that are used as there is a lack of data for this type of task, especially in less studied languages such as Romanian. Furthermore, datasets that explore different types of deception and contexts where people lie would be helpful in creating more accurate textual lie detectors.

Moreover, even though every language has its individuality and the deception process should be, from a certain point, particular for the language, the proposed approach could be extended to different languages. This direction could be the one of studying possible structural and lexical dissimilarities between the linguistic manifestation of deceit in languages from different families (i.e. Romance languages and Germanic languages).

Additionally, more features can be added to the classification algorithm for an improvement in deception detection. We plan to explore further the implication of affect and the possible inclusion of automatic emotion analysis into the identification of deceptive language. Moreover, different representations of

the texts are to be considered, as word embeddings are a very versatile method in problems of text analysis and classification.

Finally, another advancement that can be made is in the algorithms that we studied. Given the fact that the best results were obtained with an artificial neural network, a dive into some Deep Learning architectures would bring a new perspective on deception detection. For example, an architecture based on Transformers might help with the difficulty of the classification task and bring better outcomes.

References

[1] Ángela Almela, Rafael Valencia-García, and Pascual Cantos. Seeing through deception: A computational approach to deceit detection in written communication. In Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari, editors, *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 15–22, Avignon, France, April 2012. Association for Computational Linguistics.

[2] Luigi Anolli, Michela Balconi, and Maria Ciceri. Linguistic styles in deceptive communication: Dubitative ambiguity and elliptic eluding in packaged lies. *Social Behavior and Personality: an international journal*, 31:687–710, 01 2003.

[3] Jeffrey S. Bedwell, Shaun Gallagher, Shannon N. Whitten, and Stephen M. Fiore. Linguistic correlates of self in deceptive oral autobiographical narratives. *Consciousness and cognition*, 20(3):547–555, 2011.

[4] Diana Paula Dudău and Florin Alin Sava. Performing multilingual analysis with linguistic inquiry and word count 2015 (liwc2015). an equivalence study of four languages. *Frontiers in Psychology*, 12:570568, 2021.

[5] David Freedman, Robert Pisani, and Roger Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[6] Jeffrey T. Hancock, Lauren E. Curry, Saurabh Goorha, and Michael T. Woodworth. Lies in conversation: An examination of deception using automated linguistic analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26, 2004.

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

[8] Rada Mihalcea and Carlo Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li, editors, *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[9] Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675, 2003. PMID: 15272998.

[10] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[11] Katerina Papantoniou, Panagiotis Papadakos, Theodore Patkos, Giorgos Flouris, Ion Androutsopoulos, and Dimitris Plexousakis. Deception detection in text and its relation to the cultural dimension of individualism/collectivism. *CoRR*, abs/2105.12530, 2021.
[12] James W. Pennebaker and Martha E. Francis. *Linguistic Inquiry and Word Count*. Lawrence Erlbaum Associates, Incorporated, 1999.
[13] Verónica Pérez-Rosas and Rada Mihalcea. Cross-cultural deception detection. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
[14] Verónica Pérez-Rosas and Rada Mihalcea. Experiments in open domain deception detection. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1125, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
[15] Marko Robnik-Sikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53:23–69, 10 2003.
[16] Aldert Vrij, Pär Anders Granhag, and Stephen Porter. Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, 11(3):89–121, 2010. PMID: 26168416.

Department of Computer Science, Faculty of Mathematics and Computer Science, Babeş -Bolyai University, Mihail Kogălniceanu 1, 400084, Cluj-Napoca, Romania

Email address: malina.crudu@stud.ubbcluj.ro