# Detecting Deception in Disinformation Across Languages: The Role of Linguistic Markers

**Alba Pérez-Montero**[1] **Silvia Gargova**[2] **Elena Lloret**[1] **Paloma Moreda**[1]
[1]University of Alicante, Spain
[2]Big Data for Smart Society Institute (GATE), Bulgaria,
alba.perezm@ua.es, {elloret, moreda}@dlsi.ua.es
silvia.gargova@gate-ai.eu

## Abstract

The unstoppable proliferation of news driven by the rise of digital media has intensified the challenge of news verification. Natural Language Processing (NLP) offers solutions, primarily through content and context analysis. Recognizing the vital role of linguistic analysis, this paper presents a multilingual study of linguistic markers for automated deceptive disinformation detection across English, Spanish, and Bulgarian. We compiled datasets in these languages to extract and analyze both general and specific linguistic markers. We then performed feature selection using the *SelectKBest* algorithm, applying it to various classification models with different combinations of general and specific linguistic markers. The results show that Logistic Regression and Support Vector Machine classification models achieved F1-scores above 0.8 for English and Spanish. For Bulgarian, Random Forest yielded the best results with an F1-score of 0.73. While these markers demonstrate potential for transferability to other languages, results may vary due to inherent linguistic characteristics. This necessitates further experimentation, especially in low-resource languages like Bulgarian. These findings highlight the significant potential of our dataset and linguistic markers for multilingual deceptive news detection.

## 1 Introduction

Today's digital age has democratized information creation and dissemination, but it has also unleashed an unprecedented flow of content, blurring the lines between verified and deceptive information. The immense volume of online information and its impact in society make critical to develop robust automated systems for disinformation detection. This has prompted disinformation research from various perspectives, ranging from psychological analyses of deception to computational linguistics.

Specifically for written disinformation, a linguistic approach to analyzing news content offers a powerful solution, enabling quicker review by eliminating the need for factual data verification. This is precisely where Natural Language Processing (NLP) becomes a crucial tool for automating news processing. To effectively detect deceptive elements in written news, NLP must account for the fact that, although news is expected to objectively report factual events, linguistic studies as the one conducted by Tuchman (1998) demonstrate the inherent subjectivity of news reporting and how a communicator's psychological state is reflected in their linguistic choices. As Bajtin (1982) asserts, "every utterance (...) in any sphere of discursive communication, is individual and therefore may reflect the individuality of the speaker (or writer), i.e. it may possess an individual style".

Drawing on this, we address the subtask of deception detection in disinformation, as presented by Saquete et al. (2020), specifically through the lens of linguistic analysis. We explore the hypothesis that deceptive disinformation contains a unique linguistic "imprint" of deception, revealing how a sender's psychological state is manifested in their language. As defined by Yuan et al. (2024) "deceptive disinformation is intended to deliberately mislead readers or cause adverse effects". To this end, we perform a multilingual analysis. We first collect established general linguistic markers, then investigate how both these and our proposed subjective markers are represented in English, Spanish, and Bulgarian discourse, acknowledging their distinct linguistic families.

Building on this, our research focuses on three main objectives. First, we introduce a set of linguistic markers derived from part-of-speech (POS) tagging along with our designed subjectivity markers and readability indexes (Section 4.1). Second, we create a multilingual dataset for testing

*Proceedings of Recent Advances in Natural Language Processing*,pages 943–952
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-108

these markers in automated deceptive disinformation detection across Spanish, Bulgarian, and English (Section 3). This multilingual approach is vital for robust analysis and identifying potentially language-independent markers, addressing the English-centric bias in current research. Finally, we conduct a detailed analysis of these linguistic markers' discriminant potential for deception detection in disinformation (Section 4), presenting our findings (Section 5), conclusions (Section 6), and identifying limitations (Section 6).

## 2 Background

Researches on disinformation detection primarily employ two strategies: context-based and content-based analysis (Zhang and Ghorbani, 2020). Contextual approaches examine the platforms where disinformation originates and spreads, focusing on users' sharing habits. However, content-based strategies divide into fact verification and language analysis approaches. Automated fact-checking often compares online information with reliable, verified news, as highlighted by Kotonya and Toni (2020). Yet, the considerable volume of daily information complicates up-to-date fact-checking. Consequently, some researchers focus on language analysis to detect deceptive news by its writing style. Our research specifically addresses linguistic deception in disinformation, involving examining different types of linguistic markers to identify manipulative language within news discourse.

Most prior research on linguistic markers and dataset development for deception detection in disinformation has focused on English. To start with relevant previous studies on linguistic analysis on news classification, DePaulo et al. (2003) stands out as a meta-analysis incorporating over 150 markers for deception detection. As one of the pioneering studies in the field, it has been widely cited for its significant impact. On this line, Gravanis et al. (2019), in addition to a review of relevant prior studies, they propose a disinformation detection model that primarily leverages content analysis implemented with machine learning algorithms. Furthermore, they introduce "UNBiased," a novel corpus constructed to reduce bias in this classification task. On this research they mainly focus on three sets of linguistic markers proposed by Burgoon et al. (2003), Newman et al. (2003) and Zhou et al. (2004)). In the review made by Zhang and Ghorbani (2020), authors mention some of the most relevant datasets for disinformation detection in English. However, the lack of labeled datasets is the bottleneck for building an effective detection system for online misleading information (Shu et al., 2017), specially for languages other than English.

In Spanish, the analysis of linguistic markers for detecting linguistic deceptiveness has been approached through various linguistic proposals. Almela (2021) examined written statements and found that deceptive communication can result in more concise responses due to cognitive effort. They also identified a higher prominence of second and third-person pronouns as a key indicator of non-immediacy in deceptive statements, whereas Tretiakov et al. (2022) used deep learning techniques and showed strong capabilities for identifying false claims in Spanish, with certain models achieving an F1-score of 0.88. In Portuguese, Santos et al. (2020), in which they conduct an analysis of different types of linguistic markers for the detection of disinformation, including cohesion measures and applying readability indices. In their study they demonstrate that, in Portuguese, the set of markers they analyze can classify deceptive news with an accuracy exceeding 0.90.

In the context of Bulgarian, Temnikova et al. (2023) contributed by identifying 18 categories of linguistic markers for disinformation detection, also based on Zhou et al. (2004). They also developed a foundational dataset for disinformation research on Bulgarian social media. This study serves as a crucial reference for our multilingual analysis of disinformation.

Aligning with our multilingual approach, Yuan et al. (Yuan and Liu, 2024) quantitatively analyzed various features using clustering experiments to identify commonalities across English, Russian, and Chinese. Their findings revealed shared morphological markers but no consistent patterns in syntactic features or readability metrics. While an important step in multilingual disinformation study, their work's depth is limited by not applying complex markers or developing a classifier model.. Also focused on multilingual analysis, Krasitskii et al. (2024) approach on the applicability of Hungarian and Finnish resources for multilingual sentiment analysis. They use a Bulgarian dataset to evaluate pre-trained models like BERT (Bidirectional Encoder Representations from Transformers) and its multilingual variant, mBERT, achieving an accuracy between 0.8 and 0.9. Besides, Apostol et al.

(2025) concentrate their research on analyzing multiword discourse markers across languages. They make use of TED Talk transcripts in 10 languages, including Bulgarian, annotating fragments to train two cross-language machine learning models based on FastText (Bojanowski et al., 2017) and XLM-RoBERTa-Large (Conneau et al., 2019). While they reach an accuracy of up to 0.93 in Lithuanian, this drops to 0.5 for Bulgarian.

## 3 Datasets

To analyze the discriminatory capacity of the linguistic markers proposed in Section 4.1, we collected true and false news articles to construct datasets with similar characteristics in Spanish and Bulgarian. For English, we utilized a pre-existing dataset. Every dataset is balanced in real-deceptive news, as seen in Table 1. As previously noted by Antici et al. (2021) "the limited number of sentences in our corpus is symptomatic of our fine-grained annotation methodology, which is oriented to the collection of high-quality data". Additionally, to prevent biased results, we selected news articles with a similar word count, averaging 480-520 words in each dataset. This way, we construct a multilingual dataset comprising 480 texts in total.

| Dataset | Real News | Deceptive News | Total |
|---|---|---|---|
| Spanish | 80 | 80 | **160** |
| Bulgarian | 80 | 80 | **160** |
| English | 80 | 80 | **160** |
| **Grand Total** | **240** | **240** | **480** |

Table 1: Distribution of news by language and category.

### 3.1 Spanish dataset

To compile our Spanish news dataset, we referenced existing datasets by Bonet-Jover et al. (2023) and Posadas-Durán et al. (2019). Bonet-Jover et al. (2023) annotated news articles using *RUN-AS*, a fine-grained scheme based on journalistic techniques that classifies news and its essential parts as reliable or unreliable. In contrast, Posadas-Durán et al. (2019) labeled news as true if evidence suggested publication on reliable sites, and fake if sourced from websites specializing in deceptive content detection. This disparity highlights a lack of standardization in annotation methods within the field of disinformation deception detection. We began by extracting false news items and then search for parallel verified reports on the same events. This ensured that the differences between genuine and deceptive news examples in our corpus lay primarily in writing style, rather than topic.

Additionally, we gathered random news articles to create a balanced, topic-independent dataset comprising 50% true and 50% deceptive news, totaling 160 articles. As a result, the Spanish dataset is half comprised of existing datasets and half handpicked.

### 3.2 Bulgarian dataset

To compile our Bulgarian dataset, we selected a sample from the *Bulgarian disinformation and Click-bait Corpus*, a collection of online news articles gathered over a defined period from distinct sources. These sources span a range of domains, including politics, interesting facts, and tips&tricks. The corpus was annotated by journalism students and is publicly available on HuggingFace[1]. It has also been used in prior research (Karadzhov et al., 2017).

First, we removed duplicate articles, which are common in the corpus due to the frequent reposting of identical or slightly modified content across multiple outlets. Duplicates were identified through title similarity. From the de-duplicated corpus, we sampled 80 legitimate (real) and 80 deceptive (fake) news articles, ensuring an equal class distribution to prevent bias in model training and evaluation.

To enhance diversity and reduce overfitting to specific temporal or topical patterns, the sampled articles were selected from different time periods and a broad range of sources.

### 3.3 English dataset

We incorporate an English-language dataset, which serves as a reference point due to the predominance of English in the development of linguistic analysis tools and their typically higher performance in this language. The dataset is derived from PolitiFact++ and GossipCop++ (Su et al., 2023), enhanced versions of the widely used FakeNewsNet corpus (Shu et al., 2018). These datasets combine human-written news articles sourced from the fact-checking websites PolitiFact [2] and GossipCop [3], along with automatically generated news content. For our purposes, we use only the human-written articles. From each dataset, we sample an equal

[1] https://huggingface.co/datasets/community-datasets/clickbait_news_bg

[2] https://www.politifact.com/

[3] https://www.gossipcop.com/

number of real and disinformation items, specifically, 40 real and 40 fake articles from PolitiFact++, and 40 real and 40 fake articles from GossipCop++, resulting in a balanced dataset across both sources (in total 80 real and 80 fake). This English dataset, previously used in similar tasks, serves as a baseline for our distinct analytical methodology.

## 4 Methodology

To approach the task of deceptive disinformation detection using linguistic markers we developed two pipelines: (a) feature extraction pipeline and (b) machine learning pipeline that integrates handcrafted linguistic markers and automatically extracted textual features. The goal is to assess the impact of linguistic markers on classification performance across multiple model architectures.

### 4.1 Feature extraction

To avoid biases associated with label imbalance, we constructed three perfectly balanced binary datasets, ensuring that each label comprises exactly 50% of the data. Every dataset consists of news articles labeled as either real or deceptive, with an equal number of instances per class.

**Data cleaning** Prior to feature extraction and model training, we performed manual text cleaning to remove noisy, non-content text that may bias feature extraction or downstream classification. In particular, we excluded messages that were advertisements and filtered out extraneous material often appended or interleaved in news content. This included unrelated news titles inserted mid-article, trailing promotional content, and templated strings such as *“Share on ...”*, *“Follow us on ...”*, and other similar directives. These steps were conducted outside the feature extraction and learning pipelines to ensure clean, semantically coherent input.

**Linguistic markers** To ensure consistency across experimental conditions, we conduct the extraction of handcrafted linguistic markers independently of the classification, as explained in Figure 1. Each marker is counted and normalized by the total number of tokens to accurately represent its frequency.

Given an input article, we first apply syntactic analysis using the Stanza NLP pipeline to segment the text into sentences and tokens. From these annotations, we compute basic textual statistics such as word count, sentence count, average word length, average sentence length, and the number of words exceeding six characters. Lexical diversity is measured through the count of unique words, while expressive punctuation is quantified via counts of question marks, exclamation points, commas, quotation marks, ellipses, and periods.

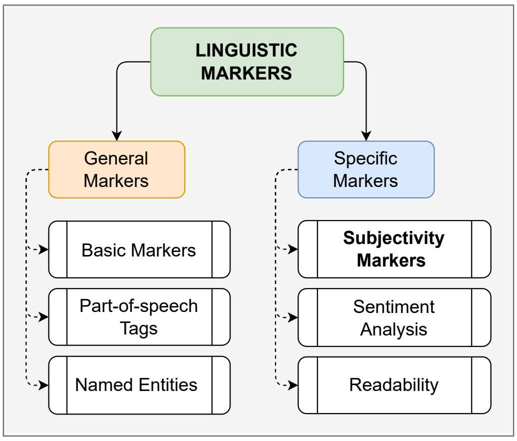

Figure 1: Diagram illustrating the types of markers used in our study.

We further extracted POS tag distributions and morphological features (e.g. mood, tense, person, voice) from the parsed tokens. Specific pronoun types, such as demonstrative, indefinite, and emphatic, were also counted due to their relevance in subjectivity and deception detection. Named entity recognition (NER) was used to quantify the number and type of named entities (PERSON, LOC, ORG), reflecting potential anchoring to real-world references. The total number of markers used is referenced in Table 2.

We also extracted higher-level stylistic categories inspired by prior linguistic research, such as Doubt, Certainty, Specificity, Non-specificity, Distancing, Participation, Expressivity and Polarity, described in Table 3.

**Sentiment** The sentiment detection for English and Spanish is carried out using the *pysentimiento*[4] library (Pérez et al., 2024), which uses transformer-based models (BETO for Spanish and BERT-base for English). These models output one of three sentiment labels: positive, negative, or neutral, along with their associated probabilities.

Due to the absence of dedicated sentiment analysis models for Bulgarian in the pysentimiento li-

[4] https://github.com/pysentimiento/pysentimiento

| Marker Type | # of Markers |
|---|---|
| **General Markers** | |
| Basic Markers | 6 |
| Part-of-speech Tag | 17 |
| Punctuation Type | 6 |
| Pronominal Type | 3 |
| Verbal Mood | 4 |
| Verbal Tense | 4 |
| Person | 3 |
| Voice | 2 |
| Named Entities | 4 |
| **Total General Markers** | **49** |
| **Specific Markers** | |
| Subjectivity Markers | 7 |
| Sentiment Analysis | 1 |
| Readability | 2 |
| **Total Specific Markers** | **10** |
| **Grand Total Markers** | **59** |

Table 2: Distribution of General and Specific Markers.

brary, we relied on a pre-trained emotion detection model for Bulgarian, as proposed by Temnikova et al. (2024). This model predicts fine-grained emotion categories, which are already grouped into coarse-grained sentiment classes—positive, negative, and neutral. In our work, we adopt these predefined groupings to map the model's output to the corresponding sentiment labels.

| Subjective Marker | Linguistic Elements |
|---|---|
| Doubt | *Questions, Conditional mood, Subjunctive mood, Interrogative pronouns.* |
| Certainty | *Imperative mood, Future tense.* |
| Specificity | *Named Entities, Demonstrative pronouns.* |
| Non-specificity | *Indefinite pronouns, Generalizing terms.* |
| Distancing | *Third person, Passive voice.* |
| Participation | *First and second person, Imperative mood.* |
| Expressivity | *Exclamation mark, Emphatic pronouns.* |

Table 3: Specific markers classification and linguistic elements conveying each type of subjective information.

**Readability** To assess textual readability, we extract the Flesch Reading Ease (FRE) score and the Brunet Index across all languages. For English and Spanish, we use the corresponding language-specific implementations provided by the `textstat` library[5]. Specifically, for Spanish, `textstat` applies the version of the FRE formula adapted by Fernández Huerta (1959), which is designed to reflect the syntactic and phonological structure of Spanish texts. The formula is defined as:

$$\text{FRE}_{\text{FH}} = 206.84 - (0.60 \times \text{SL}) - (1.02 \times \text{WL}),$$

where SL is the average sentence length (number of words per sentence) and WL is the average number of syllables per 100 words. Additionally, we extract the average number of syllables per word (ASW) for Spanish to capture lexical complexity.

For Bulgarian, where `textstat` does not support Flesch-based metrics, we employ an adaptation of the Flesch formula originally developed for Russian (Gordejeva et al., 2022), due to typological proximity:

$$\text{FRE}_{\text{bg}} = 208.7 - (2.6 \times \text{ASL}) - (39.2 \times \text{ASW}),$$

where ASL is the average sentence length and ASW is the average syllables per word, computed using regular expressions tailored to Bulgarian vowel clusters.

Lastly, we compute the Brunet Index (Brunet et al., 1978) across all languages to estimate lexical richness:

$$\text{BI} = W^{V^{-0.165}},$$

where $W$ is the total number of words and $V$ the number of unique word types.

## 4.2 Experimental Setup

To investigate the impact of feature representations and model architectures on document classification performance, we implemented a modular pipeline architecture for each model, composed of three core components: preprocessing, feature selection, and classification. All components were implemented using the Pipeline utility from the *scikit-learn* library [6] to ensure consistency and reproducibility across all experiments.

**Preprocessing and Feature Integration** Each input instance consisted of a text document accompanied by a set of linguistic markers. Textual inputs were vectorized using either a bag-of-words (BoW) representation or a TF-IDF-based encoding, each limited to the 300 most frequent unigrams and bigrams. These textual representations were combined with two sets of linguistic markers, general and language-specific, in various combinations.

[5] https://github.com/shivam5992/textstat

[6] https://scikit-learn.org/stable/index.html

To ensure consistent scaling, all numeric features were standardized using z-score normalization. Text and non-text features were processed in parallel, enabling seamless integration of heterogeneous data modalities within a unified pipeline.

**Feature Selection** Following preprocessing, we applied univariate feature selection using mutual information to identify the most predictive features for the target label. The *SelectKBest* algorithm was employed to retain the top $k \in \{50, 100, 150\}$ features based on their mutual dependence with the class label. This step serves to reduce dimensionality and promote interpretability, which is particularly important when analyzing which lexical, syntactic, or discourse-level cues are most salient for the classification task.

**Classification Models** We selected four machine learning algorithms for evaluation: **Logistic Regression** with L1 regularization, a **Support Vector Machine** with a linear kernel, **Random Forest**, and **Gaussian Naive Bayes**.

These models were selected to (a) ensure robust performance on small datasets and (b) encompass a diverse range of learning paradigms, spanning linear to non-linear and probabilistic to ensemble-based approaches.

**Hyperparameter Tuning and Model Selection** Hyperparameters were optimized using grid search over a predefined parameter space, with performance evaluated via 5-fold stratified cross-validation. Stratification was used to preserve the class distribution across folds, mitigating the risk of performance inflation due to class imbalance. The hyperparameter grid included the regularization strength ($C$) for Logistic Regression and SVM, the number of estimators and tree depth for Random Forest, and the number of selected features ($k$) for all models.

Model selection is based on macro-averaged F1 score, appropriate for the balanced binary classification setup. We conduct the model evaluation separately for every language.

## 5 Results and Discussion

Our experimental results across English, Spanish, and Bulgarian support the hypothesis that deceptive news exhibits a measurable linguistic signature, which can be effectively captured using both general and language-specific linguistic markers. This section analyzes the predictive power of these features, their cross-linguistic consistency, and how model performance varies depending on the combination of features and classifiers. We evaluated the following configurations: (1) Model with Bag-of-Words (BoW), (2) Model with BoW and general features, (3) Model with BoW and language-specific features, (4) Model with BoW, general, and language-specific features, (5) Model with TF-IDF, (6) Model with TF-IDF and general features, (7) Model with TF-IDF and language-specific features, (8) Model with TF-IDF, general, and language-specific features, and (9) Model with general and language-specific features.

Table 4 reports only the results for the best-performing model–feature combinations for each language.[7]

### 5.1 Linguistic Features: Cross-Linguistic Consistency

The analysis revealed a strong cross-linguistic overlap in the types of features most predictive for deception detection. Specifically, basic text statistics such as number of words, and morphosyntactic features including POS tags (number of nouns, number of verbs, `pos_aux`), tense (past tense), mood (`mood_ind`), voice (passive voice), and person (`person_3`) were consistently selected across languages and models. These features provide structural cues that are not domain-specific and are robust across language families.

In addition, Named Entity (NE) markers such as number of NE, number of NE-PERSON , and number of NE-ORG were frequently selected, reflecting the relevance of referential specificity in deceptive texts. Features capturing readability (e.g. Brunet's Index, Flesch reading ease) and punctuation usage also appeared recurrently, suggesting that deceptive news often diverges in textual complexity and stylistic choices.

Most notably, a suite of subjective features — especially `DIST` (Distancing), `PART` (Participation), `ESPE` (Specificity), and `DUD` (Doubt) — were highly predictive and consistently selected across all three languages when included. These features appear to encode the affective and cognitive stance of the writer, which aligns with psychological theo-

[7] The Appendix provides a comprehensive overview of the selected features for the model configurations in Table 4: https://drive.google.com/file/d/1U7qwieIZ3xqul0lOgcyGpLPNPVKpFlZi/view?usp=sharing

| **Model** | **Feature Combination** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|---|---|
| ***English*** | | | | | |
| Logistic Regression | TF-IDF + general + specific | 0.88 | 0.88 | 0.88 | **0.88** |
| SVM | TF-IDF + general | 0.78 | 0.76 | 0.81 | 0.79 |
| Random Forest | BoW + general | 0.84 | 0.76 | 1.00 | 0.86 |
| Naive Bayes | BoW + general + specific | 0.78 | 0.71 | 0.94 | 0.81 |
| ***Spanish*** | | | | | |
| Logistic Regression | BoW + general | 0.75 | 0.70 | 0.88 | 0.78 |
| SVM | BoW + general | 0.81 | 0.78 | 0.88 | **0.82** |
| SVM | BoW + specific | 0.81 | 0.78 | 0.88 | **0.82** |
| Random Forest | BoW + general | 0.78 | 0.76 | 0.81 | 0.79 |
| Naive Bayes | TF-IDF | 0.75 | 0.72 | 0.81 | 0.76 |
| ***Bulgarian*** | | | | | |
| Logistic Regression | BoW + general | 0.69 | 0.69 | 0.69 | 0.69 |
| SVM | TF-IDF + specific | 0.69 | 0.71 | 0.63 | 0.67 |
| Random Forest | BoW + specific | 0.72 | 0.71 | 0.75 | **0.73** |
| *Naive Bayes* | *general + specific* | *0.63* | *0.58* | *0.88* | *0.70* |

Table 4: The table presents the results for the best-performing model–feature combinations across all languages.

ries of deception suggesting greater cognitive load and emotional distancing in deceptive narratives.

### 5.2 Impact of Feature Combinations

The integration of general and specific linguistic markers frequently improved model performance, though not universally. In English, ogistic Regression with TF-IDF combined with general and specific markers reached the highest F1-score (0.88), a substantial improvement over using TF-IDF alone (F1 = 0.78). Similarly, in Spanish, SVM with BoW + general or BoW + specific achieved top F1-scores (0.82), indicating the complementary value of the linguistic markers.

However, in some cases, combinations including specific markers were highly effective. For instance, in Bulgarian, Random Forest with BoW + specific markers outperformed combinations involving general markers (F1 = 0.73 vs. 0.71), suggesting that specific stylistic cues alone can provide strong discriminatory power in lower-resource or morphologically rich languages.

The choice between BoW and TF-IDF also played a critical role and was language- and model-dependent. While TF-IDF generally provided stronger performance in English, BoW proved more effective in several Spanish and Bulgarian configurations, particularly when combined with linguistic markers. These variations underscore the importance of tailoring the feature representation to language-specific properties and the classifier used.

### 5.3 Impact of Classifier Choice

No single classifier consistently outperformed others across all languages. Logistic Regression was the top performer in English, especially when paired with TF-IDF and full marker sets (F1 = 0.88), and showed robust performance in Spanish (F1 = 0.78). SVM achieved the highest scores in Spanish (F1 = 0.82) and competitive results in the other languages. Random Forest excelled in Bulgarian (F1 = 0.73), suggesting that ensemble methods may be particularly effective in morphologically complex languages where interaction effects between features are more intricate.

Naive Bayes, while rarely achieving the top F1 scores, demonstrated high recall in English (Recall = 0.94 with BoW + general + specific), indicating its potential utility in settings where minimizing false negatives (i.e., undetected disinformation) is critical.

### 5.4 Language-Specific Variability

Overall performance varied significantly across the three languages, with the highest scores achieved in English (F1 up to 0.88), followed by Spanish (up to 0.82), and Bulgarian (up to 0.73). This gradient may reflect differences in NLP tool maturity, corpus size and quality, and typological features of the languages. English, as a high-resource language, benefits from more refined preprocessing tools and well-defined annotation schemes, which likely contribute to the stronger performance. Bulgarian, in contrast, showed more variability in optimal model/feature pairings, highlighting the need for specialized handling in low- or medium-resource language contexts.

### 5.5 Implications

These findings confirm the cross-linguistic validity of deception markers, especially those tied to syntax, referentiality, and subjectivity. Yet, effective detection still requires language-specific adaptation in feature design and model choice. While linguistic features offer interpretability and robustness across domains, their utility is maximized when tailored to the target language and task.

Overall, the results validate the study's premise: *deceptive disinformation can be reliably identified through a combination of general linguistic cues and subjectivity markers*. When modeled appropriately, these markers remain effective across diverse languages, providing a strong basis for developing multilingual and cross-lingual detection systems.

## 6 Conclusions and future work

This work presents a multilingual analysis of general and specific linguistic markers for the automatic detection of deception in disinformation across English, Bulgarian, and Spanish. Our approach is based on the premise that news inherently contains a subjective component, enabling us to develop a set of specific subjectivity-related markers. Analyzing both general and these specific markers, which incorporate subtasks like sentiment analysis and readability indexes, allows us to first characterize the linguistic style of news and then train an optimized model for automated deception detection in disinformation. The use of traditional machine learning methods is justified by the need to ensure the transparency and explainability of the process, which allows for necessary adjustments at any point during the experimentation along with resource optimization. We build a Silver Standard Corpus (SSC) for deceptive news detection, which is an initial, single-expert annotated dataset, recognizing that while it provides a high-accuracy baseline, it will need further multi-expert refinement to become a Gold Standard Corpus (GSC), as discussed in the study by Chowdhury and Lavelli (2011). Additionally, we present a new dataset for Bulgarian and Spanish, contributing to the study of deception in disinformation in languages other than English. Our multilingual approach allows for a comparative analysis of deceptive linguistic characteristics across these three distinct language families.

In the future, we intend to explore the proposed specific markers more deeply to build a robust methodology for analyzing deceptive language through the lens of subjectivity. Additionally, we'll focus on enhancing the semi-automatic annotation of specific markers, potentially using large language models to gain deeper semantic and pragmatic insights. Furthermore, our ultimate goal is to enlarge the presented datasets, either by collecting more news articles or by applying data augmentation techniques. This expansion aims to create a robust, gold-standard resource for disinformation analysis and to incorporate new languages, promoting wider collaboration in deception detection research. Consequently, enlarging the dataset would enable the use of transformers, allowing for a direct comparison with the performance of traditional machine learning techniques.

## Limitations

Despite the promising results, our study has several limitations. First, the relatively small size of the datasets, comprising only 160 examples per dataset, may restrict the generalization of our findings and limit the robustness of the trained models. Second, our approach relies heavily on external NLP libraries such as Stanza for preprocessing and feature extraction. Any inaccuracies or inconsistencies within these tools could propagate errors into our pipelines and affect overall performance. Lastly, the datasets used in this study differ in their labeling methodologies: some categorize news articles based on the credibility of their sources, while others depend on comprehensive fact-checking annotations. This discrepancy poses challenges for direct comparison and may introduce noise into the classification task, potentially impacting model evaluation and cross-dataset applicability.

## Acknowledgments

Alba Pérez-Montero is supported by a grant for the recruitment of predoctoral research staff (CIACIF/2023/106) from the Fondo Social Europeo Plus of Generalitat Valenciana - European Social Fund Plus of the Generalitat Valenciana. The research work is part of the R&D projects "CORTEX: Conscious Text Generation" (PID2021-123956OB-I00), funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe"; "CLEAR.TEXT" (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and "European Union NextGenerationEU/PRTR"; "VIVES" project (2022/TL22/00215334) as subproject of "ILENIA" project (grant number 2022/TL22/00215337) from the Ministerio de Transformación Digital y por el Plan de Recuperación, Transformación y Resiliencia – Financiado por la Unión Europea – NextGenerationEU; NextGenerationEU within the framework of the project Desarrollo Modelos ALIA; and the project "NL4DISMIS"(CIPROM/2021/021)" funded by the Generalitat Valenciana.

Silvia Gargova is supported by GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under grant agreement no. 857155, the programme "Research, Innovation and Digitalization for Smart Transformation" 2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01, and the BROD project, funded by the Digital Europe programme of the European Union under grant agreement no. 101083730.

## References

Ángela Almela. 2021. A Corpus-Based Study of Linguistic Deception in Spanish. *Applied Sciences*, 11(19):8817.

Francesco Antici, Luca Bolognini, Matteo Antonio Inajetovic, Bogdan Ivasiuk, Andrea Galassi, and Federico Ruggeri. 2021. Subjectivita: An italian corpus for subjectivity detection in newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 40–52. Springer.

Elena-Simona Apostol, Ciprian-Octavian Truică, Mariana Damova, Purificação Silvano, Giedre Valunaite Oleškeviciene, Chaya Liebeskind, Dimitar Trajanov, Anna Baczkowska, Emma Angela Montecchiari, and Christian Chiarcos. 2025. Multiword discourse markers across languages: A linguistic and computational perspective. *International Journal of Applied Linguistics*.

Mijail Bajtin. 1982. *Estética de la creación verbal*. Siglo XXI Editores.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Alba Bonet-Jover, Robiert Sepúlveda-Torres, Estela Saquete, and Patricio Martínez Barco. 2023. Annotating reliability to enhance disinformation detection: annotation scheme, resource and evaluation. *Procesamiento del Lenguaje Natural*, 70:15–26.

Étienne Brunet et al. 1978. *Le vocabulaire de Jean Giraudoux structure et évolution*. Slatkine.

Judee K. Burgoon, J. P. Blair, Tiantian Qin, and Jay F. Nunamaker. 2003. Detecting deception through linguistic analysis. In *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics*, ISI'03, page 91–101, Berlin, Heidelberg. Springer-Verlag.

Md Faisal Mahbub Chowdhury and Alberto Lavelli. 2011. Assessing the practical usability of an automatically annotated corpus. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 101–109.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

José Fernández Huerta. 1959. Medidas sencillas de lecturabilidad. *Consigna*, 214:29–32.

Jelizaveta Gordejeva, Richard Zowalla, Monika Pobiruchin, and Martin Wiesner. 2022. Readability of English, German, and Russian Disease-Related Wikipedia Pages: Automated Computational Analysis. *Journal of Medical Internet Research*, 24(5):e36835.

Georgios Gravanis, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. 2019. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017. We built a fake news / click bait filter: What happened next will blow your mind! In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 334–343, Varna, Bulgaria. INCOMA Ltd.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mikhail Krasitskii, Olga Kolesnikova, Liliana Chanona Hernandez, Grigori Sidorov, and Alexander Gelbukh. 2024. Multilingual approaches to sentiment analysis of texts in linguistically diverse languages: A case study of finnish, hungarian, and bulgarian. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 49–58.

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. PMID: 15272998.

Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, and Jesús Jaime Moreno Escobar. 2019. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Juan Manuel Pérez, Mariela Rajngewerc, Juan Carlos Giudici, Damián A. Furman, Franco Luque, Laura Alonso Alemany, and María Vanina Martínez. 2024. pysentimiento: A python toolkit for opinion mining and social nlp tasks.

Roney Santos, Gabriela Pedro, Sidney Leal, Oto Vale, Thiago Pardo, Kalina Bontcheva, and Carolina Scarton. 2020. Measuring the impact of readability features in fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1404–1413, Marseille, France. European Language Resources Association.

Estela Saquete, David Tomás, Paloma Moreda, Patricio Martínez-Barco, and Manuel Palomar. 2020. Fighting post-truth using natural language processing: A review and open challenges. *Expert Systems with Applications*, 141:112943.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Kai Shu, Suhang Wang, and Huan Liu. 2017. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 8.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models.

Irina Temnikova, Silvia Gargova, Ruslana Margova, Veneta Kireva, Ivo Dzhumerov, Tsvetelina Stefanova, and Hristiana Krasteva. 2023. New bulgarian resources for studying deception and detecting disinformation.

Irina Temnikova, Iva Marinova, Silvia Gargova, Ruslana Margova, Alexander Komarov, Tsvetelina Stefanova, Veneta Kireva, Dimana Vyatrova, Nevena Grigorova, Yordan Mandevski, and Stefan Minkov. 2024. SM-FEEL-BG - the first Bulgarian datasets and classifiers for detecting feelings, emotions, and sentiments of Bulgarian social media text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14954–14966, Torino, Italia. ELRA and ICCL.

Arsenii Tretiakov, Alejandro Martín, and David Camacho. 2022. Detection of false information in spanish using machine learning techniques. In *Intelligent Data Engineering and Automated Learning – IDEAL 2022*, pages 42–53, Cham. Springer International Publishing.

Gaye Tuchman. 1998. La objetividad como ritual estratégico: un análisis de las nociones de objetividad de los periodistas. *CIC. Cuadernos de Información y Comunicación*, (4):199.

L Yuan, H Jiang, H Shen, L Shi, and N Cheng. 2024. Sustainable development of information dissemination: A review of current fake news detection research and practice. systems 2023, 11, 458.

Wei Yuan and Haitao Liu. 2024. Finding common features in multilingual fake news: a quantitative clustering approach. *Digital Scholarship in the Humanities*, 39(2):790–804.

Xichen Zhang and Ali A. Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Inf. Process. Manage.*, 57(2).

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.