Fake News Detection in Albanian: A feature-based Statistical Analysis for a Low-Resource Language

Elton Tata
Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Republic of North Macedonia

Jaumin Ajdari
Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Republic of North Macedonia

Nuhi Besimi
Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Republic of North Macedonia

ABSTRACT
Since fake information spreads so rapidly in digital media, detecting it has become a very important research problem. This study provides a statistical and feature-focused examination of fake news detection in Albanian.

A total of 54 manually engineered linguistic, lexical, punctuation, textual statistical, structural, and temporal features are analysed to capture key aspects of language. The analysis is conducted using a balanced dataset of 3,994 Albanian-language news articles. Descriptive statistics, correlation analysis, and feature importance measures are employed to identify the most significant indicators of fake news. The findings show that sensational language, reduced lexical diversity, inadequate source attribution, and discrepancies between headlines and content are important predictors of fake news.

The findings indicate that carefully designed, language-sensitive features can effectively identify disinformation patterns in Albanian news and provide clear insights into the detection methodology. This paper adds a framework for interpretable feature analysis that helps detect fake news in resource-poor linguistic environments. It also lays the foundation for future research on multilingual and disinformation-based topics.

General Terms
Machine Learning

Keywords
Fake News Detection, Low-Resource Languages, Feature Engineering, Statistical Text Analysis.

1. INTRODUCTION
Fake news detection has become a critical research problem due to the rapid spread of misleading content across digital news platforms and social media. Recent studies highlight that linguistic, stylistic, and structural cues remain among the most reliable indicators of disinformation, especially in settings where large-scale annotated data or computational resources are limited. This challenge is particularly acute for low-resource languages, where language-specific features and journalistic conventions differ significantly from those of high-resource languages such as English. Several recent papers have highlighted that feature-based and hybrid approaches remain competitive when datasets are moderate in size and interpretability is required. Canhasi et al. [1] presented a comprehensive dataset of fake news in Albanian and demonstrated that Albanian-specific linguistic and structural features capture patterns that are not adequately modeled by language-agnostic approaches. Subsequent studies have confirmed that hand-crafted features related to rhetorical structure, lexical choice, and stylistic consistency can provide strong discriminatory signals even without deep neural architectures. In general, feature-centric analyses have been shown to improve the robustness and transparency of fake news detection systems. Feature analysis studies [2–4] show that statistical text properties, punctuation usage, and lexical diversity are consistent indicators across domains, while low-resource linguistic studies [5] emphasize the importance of adapting features to the specific morphology and syntax of the language. These findings motivate a deeper investigation into which features matter most, rather than focusing exclusively on classifier accuracy. In this context, the present work focuses on a statistical and analytical examination of Albanian-specific features for fake news detection. Rather than proposing new classification architectures, the study systematically analyses linguistic, statistical, structural, and temporal features, assesses their relative importance, and provides interpretable insights to guide future research and implementation in low-resource language environments.

The following sections comprise the rest of this document: Section II provides a literature survey of the earlier researches and related to this work. Section III presents the dataset. Section IV present research methodology for the proposed study. Section V presents the statistical analysis. Section VI discusses the results, and Section VII presents the conclusions.

2. RELATED WORK
Early feature-based approaches to fake news detection emphasized linguistic and stylistic cues that distinguish fraudulent from factual writing. Studies such as the feature analysis work in [2] systematically examined lexical richness, sentiment polarity, and readability metrics, showing that fake news often exhibits exaggerated language and lower lexical diversity. These findings support the continued importance of statistical text features in disinformation detection. Rhetorical and discourse-level features have also been explored as distinguishing signals. Work in [6] showed that Rhetorical Structure Theory (RST) representations capture differences in writing style between fake and real news, particularly in professionally edited versus fabricated content. This is consistent with later feature-level analyses [7], which showed that discourse-level features complement rather than replace surface lexical cues. Several recent studies have explicitly focused on the importance and interpretability of features. The analysis in [8] investigated the correlations between hand-crafted features and fake news labels, concluding that punctuation intensity, repetition patterns, and headline-content mismatch are among the most influential predictors. Similarly, [9] provided a broad statistical breakdown of linguistic and structural features, emphasizing that interpretability is essential

for trust in automated detection systems. Low-resource linguistic contexts present additional challenges. The study in [10] showed that direct transfer of high-resource language models leads to degraded performance, mainly due to morphological and syntactic inconsistencies. Research in [11] and [12] confirmed that engineering language-specific features significantly improves detection reliability in such environments, even when deep learning models are available. Hybrid approaches that combine statistical features with vectorized text representations have been widely adopted. Works such as [13] and [14] demonstrated that combining TF-IDF representations with hand-crafted features allows models to capture both global lexical patterns and fine-grained linguistic signals. It is important to note that these studies report that performance gains are often attributed to feature design rather than model complexity. Recent surveys and comparative analyses [15–17] further reinforce that feature-based and hybrid methods remain competitive, especially when data sets are limited or computational efficiency is a priority. These studies note that, while transformer-based models dominate the standard, feature-oriented approaches offer advantages in transparency, controllability, and deployment feasibility. Finally, research [18 - 19] has shown that linguistic markers such as diacritics, sensational vocabulary, and attribution patterns are particularly informative. Together, these works motivate a focused statistical investigation of the behavior, correlation, and significance of features, the objective of the present study.

3. DATASET

The quality of the training dataset is one of the key factors, which sets a limit on how well a machine learning system can perform. For fake news detection, besides a realistic distribution of the classes, the datasets should also represent at least reasonable examples from both fake and genuine news, and it is essential to have equally distributed classes in order not to create model bias or misleading statistical results. In addition, there should be a substantial volume of data to facilitate analytically meaningful analysis, and rich metadata should support temporal, structural and contextual research on misinformation patterns.

3.1 Data Collection and Sources

The dataset consists of a total of 3,994 Albanian news articles collected from diverse sources covering different political spectra, geographic origins, and journalistic standards. A manual and semi-automatic harvesting process was employed to collect the dataset, spanning a wide range of Albanian news media-industry constellations.

Articles were collected from a number of sources, including professional Albanian newspapers with established editorial policies, Kosovan news media aimed at the Albanian-speaking audience, print and online-only Albania language journals in North Macedonia, varying credibility news websites, and social-media-driven platforms.

The data source is also almost perfectly balanced, with 1,998 news articles and 1,996 fake articles (fabricated or manipulated or misleading information) respectively. This equal distribution of classes reduces the likelihood of classification bias and enables meaningful statistical analysis.

The topics span a wide domain from politics and governance, international relations, economics, and social and human-interest issues to sports and entertainment as well as health and science, with the focus on general stylistic and linguistic patterns rather than topic-specific cues.

Structured metadata is available for each article. The original headline is stored in the title field, the full article's body with original punctuation and paragraph structure can be found in the content field, publication_datetime includes the time of publication; source contains the publishing outlet or social network; and last but not least, the dataset includes a *label* field representing a binary annotation, where 0 denotes real news and 1 denotes fake news.

3.2. Data Preprocessing and Quality Control

Text preprocessing minimizes noise in the raw data and retains only semantically rich information, promoting model performance. A number of linguistically motivated transformations are applied to capture specific features of the Albanian language. To remove non-semantic artefacts introduced during digital publishing, URLs and hyperlinks are stripped from the text, as they do not contribute to the linguistic content of the articles.

Special character normalisation removes non-linguistic characters, such as extra punctuation marks, emojis not belonging to standard Albanian text and formatting signs published with the news articles as HTML extraction artefacts, while preserving rich Albanian special characters ‘ë’ and ‘ç’ that carry semantic meaning.

All text is then lowercased to ensure consistent matching, including special characters; consecutive whitespace is reduced to a single space character; leading and trailing whitespace is stripped; and quote and apostrophe characters are converted to their Unicode representation.

An Albanian-specific stopword removal strategy is applied, based on a selected set of frequently occurring particles with minimal discriminative value, such as “dhe”, “e”, “në”, “është”, “të”, “për”, “me”, and “nga”. Unlike standard stopword removal approaches, this process is selective and preserves words that may still carry semantic information useful for distinguishing fake from real news.

Duplicate detection and removal are performed to guarantee the uniqueness of entries in the dataset and to avoid data leakage between training and test sets. Exact matching is used to identify identical articles, while fuzzy matching based on TF-IDF similarity is applied to detect near-duplicates. Articles with similarity above 95% are considered duplicates and removed, resulting in 47 documents being discarded.

Quality assurance procedures include manual inspection of random samples from both classes, verification of metadata field consistency, preservation of Albanian diacritics, and statistical analysis of text length distributions to identify outliers. Articles containing fewer than 50 words were excluded, while unusually long articles exceeding 10,000 words were manually reviewed to ensure they represent legitimate journalistic content.

4. METHODOLOGY

The methodological approach is based on carefully designed feature engineering aimed at capturing linguistic patterns characteristic of Albanian-language news content while maintaining computational efficiency. This section describes the feature extraction process, the text vectorisation strategy, the procedures used for hyperparameter tuning, and the categorisation of news features.

4.1 Albanian-Specific Feature Engineering

Feature engineering is the process of turning unstructured text data into numbers that show patterns that are important for the classification task at hand. Deep learning models are made to learn representations from data on their own, but traditional machine learning methods can obtain a lot out of explicitly encoding domain knowledge through handcrafted features.
A total of 54 handcrafted features are extracted and grouped into five categories: 14 linguistic and lexical features, 12 punctuation pattern features, 15 textual statistical features, 8 structural features, and 5 temporal features. These feature groups capture discriminative cues that help distinguish between fake and real news content in the Albanian language.

4.1.1 Albanian Linguistic and Lexical features

The Albanian language exhibits distinctive linguistic characteristics that can be exploited to differentiate authentic news articles from fabricated or translated content. In this study, 14 linguistic and lexical features are extracted to capture language-specific patterns observed in Albanian news writing.

The Albanian character ratio is calculated by dividing the occurrences of characters ‘ë’ and ‘ç’ with total article characters. In professional Albanian written news, these characters occur with a statistical frequency consistent with regular use of orthography. A departure from this expected frequency, such as a complete absence, could be indicative of non-native authorship, machine translation, or low-quality content generation. A binary as_character(has_albanian_chars) is used to let us know whether the articles are void of any Albanian-specific characters.

For each sensationalist term, both its raw frequency and its relative frequency normalised by the total number of words in the article are computed to ensure comparability across articles of different lengths..

Credibility-related lexical cues are also considered, focusing on reporting verbs and attribution markers commonly used in professional journalism, such as *sipas*, *raportohet*, and *konfirmohet*. These expressions typically signal source attribution and factual grounding, which are more prevalent in real news articles.

Finally, features capturing hedging and uncertainty language (e.g., *ndoshta*, *duket*, *besoj*) are included, as they may reflect varying levels of authorial commitment. Although such expressions are not inherently indicative of misinformation, their distribution differs between real and fake news and becomes more informative when analysed in combination with other linguistic features.

Table 1. Top Linguistic Features for Albanian Fake News Detection and Their Interpretation

No.	Feature Name	Feature Description	Indicative Behavior	Interpretation for Fake News Detection
1	Albanian Character Ratio	Ratio of Albanian-specific characters (“ë”, “ç”) to total characters	Lower in fake news	Indicates possible non-native authorship, translation artifacts, or low editorial quality
2	has_albanian_chars	Binary indicator of presence of “ë” or “ç”	Often absent in fake news	Suggests text normalization issues or foreign-language influence
3	Sensational Word Count	Absolute count of sensational keywords (e.g., *skandal*, *bomba*, *alarm*)	Higher in fake news	Reflects clickbait and emotionally charged language
4	Sensational Word Ratio	Sensational word count normalized by document length	Higher in fake news	Controls for article length while capturing sensationalism intensity
5	Credibility Marker Count	Count of attribution terms (e.g., *sipas*, *raportohet*, *konfirmohet*)	Lower in fake news	Indicates lack of source attribution and journalistic verification
6	Credibility Marker Ratio	Normalized ratio of credibility indicators	Lower in fake news	Shows weaker evidential grounding of claims
7	Hedging Word Count	Count of uncertainty expressions (e.g., *ndoshta*, *duket*)	Higher in fake news	Suggests speculative or weakly supported statements
8	Hedging Word Ratio	Hedging words normalized by text length	Higher in fake news	Highlights stylistic uncertainty independent of article size
9	Sensational Headline Tokens	Presence of sensational words in titles	More frequent in fake news	Captures headline-driven manipulation strategies

4.1.2 Punctuation Pattern Features

Punctuation pattern features capture stylistic cues related to emphasis and textual structure. These features include the number and ratio of exclamation marks, detection of repeated exclamation marks or question marks, use of ellipses, frequency of quotation marks as an indicator of attributed words, and patterns of capitalization measured by the ratio of words written entirely in capital letters.

Exclamation mark features consist of absolute counts, ratios normalised by document length, and binary indicators for multiple consecutive exclamation marks (e.g., "!!" or "!!!"). Fake news content tends to exploit exclamation marks to convey urgency or emotional intensity, while professional journalism uses them sparingly, usually within direct quotes.

Features related to question marks include the total number of question marks, their normalised ratio, and detection of repeated sequences such as "??" or "???". The use of ellipses, especially sequences of three consecutive periods ("..."), is also tracked, as they are commonly used in sensationalist writing to create suspense or imply unstated conclusions.

The frequency of quotation marks, including straight and curly quotation marks, as well as apostrophes, serves as an indirect indicator of the word being reported or attributed.

Capitalization patterns are analysed through the frequency and ratio of words written entirely in capital letters, defined as characters longer than two characters written entirely in capital letters. Such patterns are rare in professional journalism but appear more frequently in fake news-style writing and social media to simulate emphasis or emotional appeal.

4.1.3 Textual Statistical Features

Textual statistical features capture properties of news articles at the word and sentence levels that are not tied to specific vocabulary. A total of 15 statistical features are extracted to model these traits. Total word count, total character count, mean word length (calculated as the ratio of character count to word count), and text length variance are all examples of basic length-based measures. These statistics consistently vary between fake and genuine news articles, with fake news frequently displaying marginally shorter texts accompanied by longer or more complex words, possibly to create a misleading impression of authority.

Regular expressions are used to find sentence boundaries based on terminal punctuation (full stops, exclamation marks, and question marks). This lets us look at sentence-level structure. From this, you can figure out how many sentences there are, how long the average sentence is, and how much the length of the sentences varies. Real news articles usually have more balanced sentence structures with some variation, while extreme uniformity or too much variability is less common in professional journalism.

Lexical diversity is captured indirectly through repetition and length-based statistics. Fake news articles tend to reuse a limited vocabulary and rely on formulaic expressions, whereas real news exhibits more balanced and varied word usage. These patterns are reflected through sentence-level and word-length distribution features rather than explicit diversity ratios. Other statistical features are the number of paragraphs and the structure of those paragraphs when the boundaries are clear, the density of the content (comparing information-bearing words to the total word count), and the higher-order statistical moments (mean, variance, skewness, and kurtosis) of the word-length and sentence-length distributions. These features go beyond simple averages to show the style of text composition and supply extra signals that help tell the difference between fake and real news.

Table 2. Textual Statistical Features Used in Analysis

No.	Feature Name	Feature Description	Indicative Behavior	Interpretation
1	Word Count	Total number of words per article	Lower in fake news	Shorter texts with compressed content
2	Character Count	Total number of characters	Lower in fake news	Less detailed textual content
3	Mean Word Length	Characters divided by word count	Higher in fake news	Artificial complexity or inflated wording
4	Text Length Variance	Variance of word count across sentences/articles	Higher in fake news	Inconsistent writing patterns
5	Sentence Count	Total number of sentences	Lower in fake news	Simplified narrative structure
6	Mean Sentence Length	Average words per sentence	Less balanced in fake news	Poor structural consistency
7	Sentence Length Variance	Variability of sentence length	Higher in fake news	Irregular sentence construction
8	Type–Token Ratio	Unique words / total words	Lower in fake news	Reduced lexical richness
9	Long Word Ratio	Words >10 characters	Higher in fake news	Over-elaborated lexical style
10	Short Word Ratio	Words <3 characters	Higher in fake news	Simplistic sentence construction
11	Paragraph Count	Number of paragraphs	Lower in fake news	Weak document structuring
12	Mean Paragraph Length	Average paragraph size	Less stable in fake news	Poor content organisation
13	Content Density	Informative words / total words	Lower in fake news	Low informational value
14	Word Length Skewness	Asymmetry of word length distribution	Higher in fake news	Stylistic irregularity
15	Sentence Length Kurtosis	Peakedness of sentence length distribution	Higher in fake news	Structural extremity

4.1.4 Structural and Compositional Features

Structural and compositional features capture dependencies between different parts of a news article and reflect its overall logical organisation. In this study, Eight features are derived to describe these properties.

Title-related features are extracted independently from the article body and include title length, title word count, and the presence of punctuation marks such as question marks and exclamation marks. Fake news headlines frequently rely on interrogative or exclamatory forms to provoke curiosity or emotional reactions, whereas professional news headlines tend to follow declarative, statement-orientated patterns.

Title–content relationship features model the alignment between the headline and the main body of the article. The title–content ratio measures whether a headline is disproportionately long or short relative to the article content. Fake news often exhibits extreme cases, such as highly sensational titles paired with minimal content or short clickbait titles followed by lengthy and unfocused text. Title–content similarity metrics further quantify the semantic consistency between the headline and the article body, helping to identify misleading or exaggerated headlines.

Content structure features describe the internal organisation of the article, including the number of paragraphs, their average length, and the distribution of information throughout the text. Most professional news articles use an inverted pyramid structure, which means that the most important information is at the top. In contrast, fake news may display irregular structures, such as delayed or fragmented presentation of key claims. These features offer information about narrative coherence and editorial discipline beyond surface-level statistics.

Table 3. Structural and Compositional Features

Feature Name	Description	Indicative Behavior in Fake News
Title Length	Number of characters or words in the headline	Often extreme (very long or very short)
Title Punctuation	Presence of "?" or "!" in the title	More frequent
Title–Content Ratio	Ratio between title length and content length	Highly variable
Title–Content Similarity	Semantic similarity between headline and body	Lower alignment
Paragraph Count	Number of paragraphs	Often fewer or irregular
Avg. Paragraph Length	Mean paragraph size	Higher variance
Paragraph Length Variance	Variability in paragraph sizes	Higher
Content Distribution	Placement of key information	Less front-loaded

4.1.5 Temporal Features

When publication timestamps are available, temporal features provide an additional discriminative signal for fake news detection. Five time-related attributes are extracted from the *publication_datetime* metadata. The hour of publication captures the time of day (values ranging from 0 to 23) at which an article is released and reflects differences in publishing behaviours between professional newsrooms and sources associated with fake news. Professional outlets tend to follow regular editorial schedules, whereas fake news is more frequently published at irregular hours.

The day-of-week feature encodes weekly publication patterns and allows the analysis of differences between weekday and weekend activity. In addition, two binary indicators are derived: is_weekend, marking articles published on Saturdays or Sundays, and is_night, indicating publications released between 22:00 and 05:00 local time, a period during which professional editorial activity is typically limited.

Although temporal features are relatively weak predictors when considered in isolation, they provide complementary contextual information when combined with linguistic, structural, and stylistic features, contributing to improved discrimination between real and fake news.

4.2 Text Representation Using TF-IDF

In addition to handcrafted features, textual content is represented using the TF-IDF (Term Frequency–Inverse Document Frequency) scheme in order to capture lexical usage patterns across the corpus. TF-IDF is employed as a standard and well-established representation that assigns higher weights to terms that are frequent within a document but relatively rare across the dataset, while down-weighting common terms with limited discriminative value.

This representation is not treated as a primary contribution of the study, but rather as a complementary baseline that enables joint analysis with Albanian-specific handcrafted features. TF-IDF supports the statistical examination of vocabulary-level signals and facilitates the assessment of how lexical patterns interact with higher-level linguistic, structural, and stylistic indicators in fake news detection.

4.3 Combined Feature Matrix Construction

To construct the final feature representation for each news article, horizontal concatenation is used to combine the TF-IDF text features with the engineered features described in Section 4.1. In particular, each article has 15,000 TF-IDF features from the vectoriser and 54 features made by hand, for a total of 15,054 features per instance.

Sparse matrix operations are used to efficiently combine these feature sources by turning the engineered features into dense vectors and stacking them horizontally with the TF-IDF sparse representations. This method allows for a single representation while still being efficient in terms of computation.

This hybrid feature space combines different types of information that work well together. TF-IDF features pick up on patterns in the data at the word and phrase level, while handcrafted features encode indicators that are linguistically and structurally motivated for writing news in Albanian. These representations work together to let downstream statistical analysis look at how much data-driven and knowledge-driven features contribute and interact with each other without using vocabulary alone.

Because the combined feature set has different types of features with different numeric ranges, feature scaling was looked into. But since tree-based learning algorithms don't change when monotonic transformations are applied, explicit rescaling of engineered features wasn't necessary and didn't change any of the analyses that came after. This property makes it possible to directly look at how important features are and how they behave statistically across groups of features.

5. STATISTICAL ANALYSIS

This section presents a statistical analysis of the engineered feature set described in Section 4, with the goal of understanding how different groups of features behave across real and fake news articles and which characteristics contribute most strongly to class separation. Rather than focusing on model architecture or optimisation, the analysis emphasises descriptive statistics, comparative trends, and interpretability of the extracted features.

5.1 Descriptive Statistics (Mean and Standard Deviation)

Table 4 presents the mean values and standard deviations of key linguistic features across the data and stratified by news authenticity.

Table 4: Descriptive Statistics of Linguistic Features

Feature	Overall Mean	Real Mean	Fake Mean	Overall Std	Real Std	Fake Std
Number of Tokens	207.73	286.40	128.99	249.98	313.97	118.29
Words w/o Punctuation	184.15	253.12	115.10	221.24	277.76	105.87
Number of Characters	1249.60	1742.23	756.47	1510.46	1897.48	689.92
Number of Verbs	31.47	42.87	20.06	38.89	48.64	19.96
Number of Nouns	26.09	36.91	15.25	32.38	40.75	14.20
Number of Adjectives	8.42	11.97	4.86	10.69	13.26	5.25
Number of Adverbs	5.20	7.25	3.14	7.24	9.02	3.88
Average Word Length	5.00	5.11	4.89	0.39	0.36	0.39
Words in Upper Case	24.83	31.78	17.87	30.39	39.61	13.48

The analysis reveals significant differences between real and fake news articles. Real news articles demonstrate consistently higher mean values in almost all linguistic metrics, with particularly pronounced differences in content length (1742 vs. 756 characters), number of tokens (286 vs. 129 tokens), and use of nouns (37 vs. 15 nouns). These findings suggest that fake news articles in Albanian tend to be shorter on average and less linguistically elaborate than their authentic counterparts.

5.2 Variance-Based Analysis

Analysis of variance reveals critical insights into the homogeneity of each news category. Table 5 presents the variance comparisons for the main features.

Table 5: Variance Comparison Between Real and Fake News

Feature	Overall Variance	Real Variance	Fake Variance	Ratio (R/F)
Number of Tokens	62,469.28	98,527.75	13,991.51	7.04
Number of Characters	2,281,542.57	3,591,013.77	475,527.60	7.55
Content Word Count	62,536.32	98,293.71	13,951.52	7.05
Number of Verbs	1,512.43	2,365.84	398.40	5.94
Number of Nouns	1,048.46	1,660.56	201.64	8.23

One interesting observation is that the big variance in true news articles is much higher than in fake news. The variance ratio was significantly higher (about 6-8 times) for real news; that is also an indication of more diversity in article length and linguistic complexity for journalism. Fake news articles, on the other hand, have more regular structures, which could be attributed to templated or formulaic content production in disinformation.

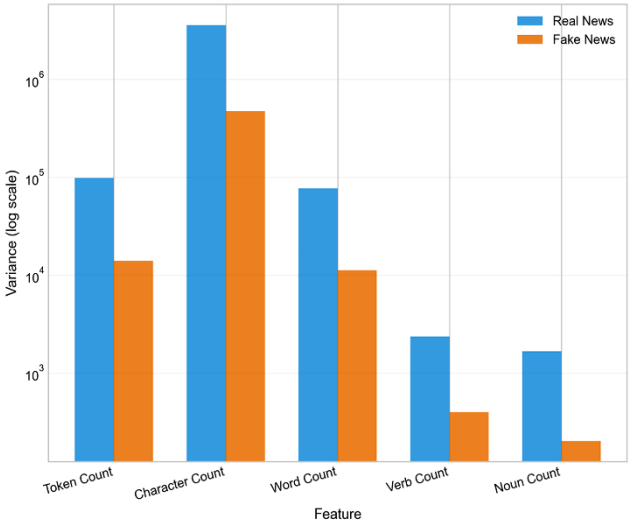

Figure 1: Variance comparison of selected textual features between fake and real news articles

5.3 Distributional Characteristics of Textual Features

The distributional properties of articles offer additional information about structural differences between news categories.

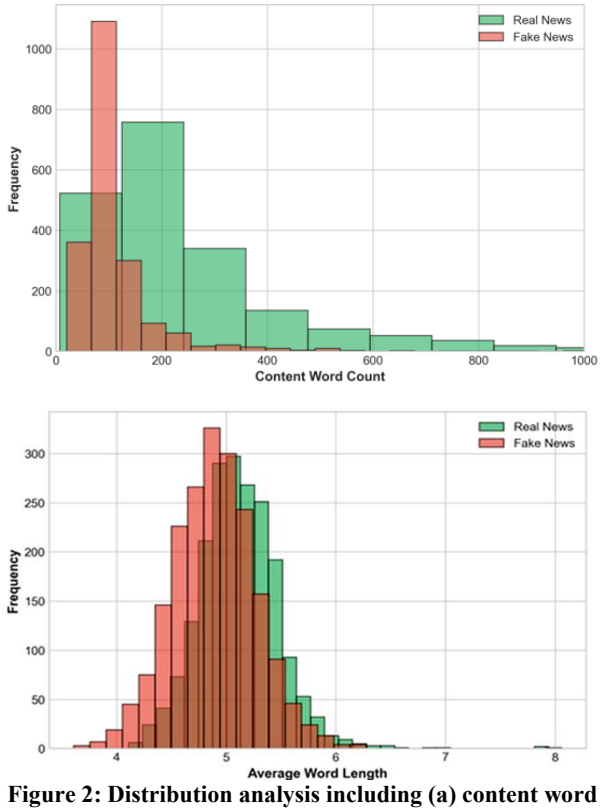

Figure 2: Distribution analysis including (a) content word count histogram, (b) average word length distribution

5.4 Correlation Analysis of Handcrafted Features

Understanding the correlation structure between features is essential for effective feature engineering and model interpretation.

Table 6: Correlation between Selected Textual Features and the News Class Label

Feature	Correlation with News Class (Fake vs Real)	Interpretation
Noun Count	-0.334	Moderate negative
Adjective Count	-0.333	Moderate negative
Character Count	-0.326	Moderate negative
Token Count	-0.315	Moderate negative
Word Count	-0.312	Moderate negative
Verb Count	-0.293	Weak negative
Adverb Count	-0.284	Weak negative
Avg Word Length	-0.283	Weak negative
Uppercase Words	-0.229	Weak negative
Modal Verbs	-0.181	Weak negative

Negative correlations across all features indicate that higher values of these linguistic metrics are associated with authentic news (label = 0), while lower values correspond to fake news (label = 1). The strongest predictors include the number of nouns (r = -0.334), the number of adjectives (r = -0.333), and the number of characters (r = -0.326).

Inter-feature correlations: The correlation matrix reveals a high correlation between features related to length (signs, characters, verbs, nouns), with correlation coefficients exceeding 0.9 in numerous instances. This multicollinearity informed the feature selection strategy and supports the use of gradient boosting models, which effectively handle correlated features.

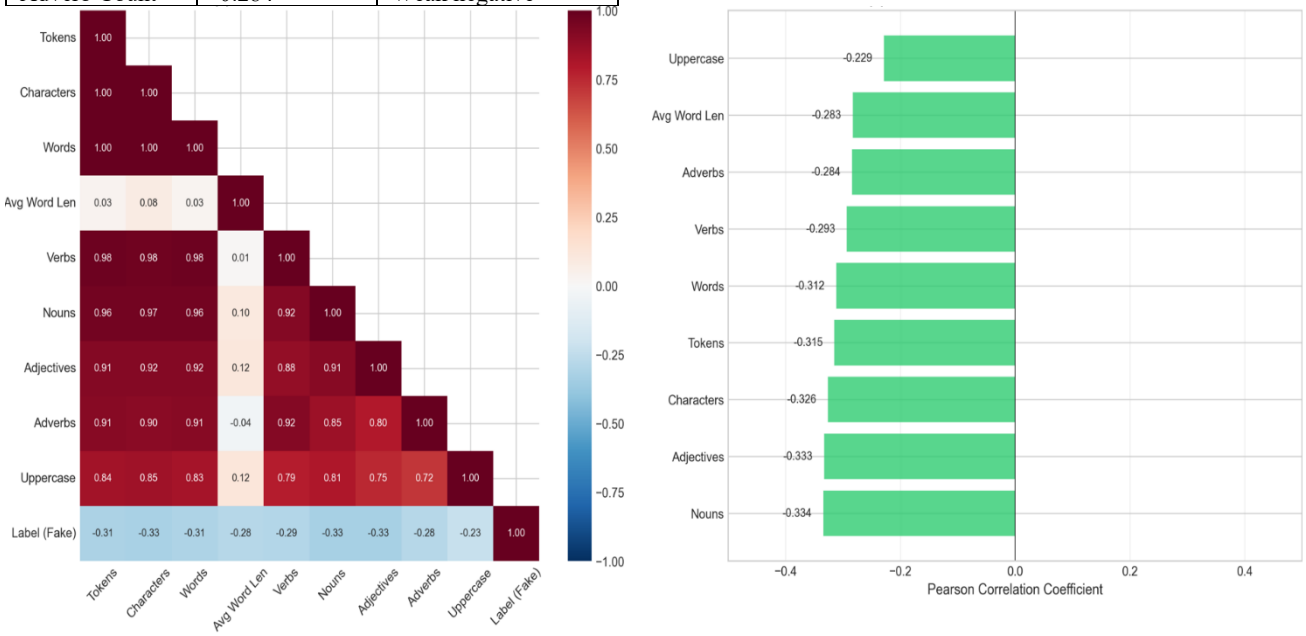

Figure 3. Correlation heatmap of selected textual features and their relationship with the fake news label, illustrating moderate negative correlations and strong inter-feature dependencies.

5.5 Statistical Significance and Reliability Analysis

A systematic comparison between fake and real news articles reveals statistically significant differences in all features examined.

Table 7 presents the results of independent samples t-tests.

Feature	**t-statistic**	**p-value**	**Cohen's d**	**Sig.**
Number of Nouns	22.42	< 0.001	0.71	***
Number of Adjectives	22.29	< 0.001	0.7	***
Number of Characters	21.81	< 0.001	0.69	***
Number of Tokens	20.96	< 0.001	0.66	***
Words w/o Punctuation	20.75	< 0.001	0.66	***
Number of Verbs	19.38	< 0.001	0.61	***
Number of Adverbs	18.72	< 0.001	0.59	***
Average Word Length	18.63	< 0.001	0.59	***
Words in Upper Case	14.85	< 0.001	0.47	***
Number of Modal Verbs	11.62	< 0.001	0.37	***

Note: *** indicates p < 0.001; Effect sizes interpreted as small (0.2), medium (0.5), large (0.8)

Table 7 reports the results of the statistical significance test between fake and real news articles. All features examined show statistically significant differences ($p < 0.001$), with Cohen's d values indicating medium to large effect sizes. These results confirm that the observed differences are robust and provide strong empirical support for the conclusions drawn in this study.

5.6 Feature Importance Interpretation

A combined examination of statistical effect sizes, correlation coefficients, and variance comparisons provides an additional interpretative perspective on the relative contribution of individual features. Rather than introducing a new predictive model, this section synthesises the previously reported quantitative findings to identify the most discriminative indicators. Across all statistical analyses, noun count, adjective count, character count, and token count consistently exhibit the largest effect sizes and strongest correlations with the fake news label. These features also demonstrate substantial variance differences between real and fake news categories, indicating both statistical and practical significance. In contrast, temporal and certain stylistic features show weaker correlation coefficients and smaller effect sizes, suggesting a complementary rather than primary role in discrimination. The convergence of mean comparisons, variance ratios, correlation analysis, and effect size estimation highlights the central importance of linguistic richness and structural complexity in distinguishing authentic Albanian news content. The consistency across independent statistical measures further supports the robustness and interpretability of the feature-based approach.

6. DISCUSION

6.1 Feature Group-Level Observations

The engineered features are analyzed by grouping them into linguistic–lexical, punctuation, textual statistical, structural–compositional, and temporal categories. For each group, distributions and summary statistics were examined separately for real and fake news articles.

Linguistic and lexical features show clear differences between the two classes. Fake news articles tend to exhibit lower usage of Albanian-specific characters, higher frequencies of sensationalist vocabulary, and reduced presence of credibility markers such as attribution verbs. Real news articles, by contrast, display more consistent use of diacritics, richer lexical diversity, and more frequent source-related expressions.

Textual statistical features further highlight stylistic differences. Fake news articles often demonstrate higher repetition, lower lexical diversity, and irregular sentence-length distributions. Real news articles show more stable sentence structure and greater balance between short and long sentences, reflecting conventional journalistic writing practices.

Structural and compositional features reveal differences in how information is organized. Fake news headlines more frequently rely on exclamatory or interrogative forms and exhibit weaker alignment with article content. In contrast, real news articles tend to follow more consistent headline–content relationships and a clearer internal structure, often resembling the inverted pyramid model.

Temporal features alone are weak discriminators; however, they provide complementary signals when analyzed alongside linguistic and structural features. Differences in publication timing suggest that fake news is more likely to be published outside typical newsroom schedules, although this effect is dataset-dependent.

6.2 Comparative Contribution of Feature Groups

To better understand the relative contribution of different feature groups, their indicative behaviour across classes is analysed rather than relying solely on absolute performance metrics. This approach supports interpretability and avoids overfitting conclusions to a specific classifier.

Across the dataset, linguistic and lexical features emerge as the most informative group, followed by textual statistical features. Structural features contribute additional discriminative power, particularly in identifying headline manipulation patterns. Temporal features contribute marginally but enhance robustness when combined with other feature types.

Table 8. Summary of Feature Groups and Their Statistical Behaviour

Feature Group	Number of Features	Behaviour (Fake vs Real)	Value
Linguistic–Lexical	14	Sensational tone, weak attribution **vs.** rich vocabulary and source attribution	Strong
Punctuation	12	Emphasis marks and capitalization **vs.** controlled punctuation use	Strong
Textual Statistical	15	Repetition and irregular length **vs.** balanced structure and lexical diversity	Strong
Structural	8	Clickbait titles, weak alignment **vs.** headline–content consistency	Medium
Temporal	5	Off-hour/weekend publishing **vs.** regular newsroom timing	Weak

6.3 Interpretation and Implications

The statistical analysis confirms that fake news in Albanian exhibits systematic linguistic, stylistic, and structural differences from real news. These differences are not driven by topic alone but reflect broader patterns of authorship, editorial practice, and content manipulation.

Importantly, the dominance of interpretable linguistic and statistical features supports the use of feature-based machine learning approaches for low-resource languages, where transparency and explainability are essential. The findings also motivate further research into feature importance analysis and cross-language validation of Albanian-specific indicators.

7. CONCLUSION

This paper presents a focused statistical and analytical study of hand-crafted features for fake news detection in the Albanian language. This paper does not present a new classification architecture or emphasise performance metrics; rather, it focuses on understanding which linguistic, structural, and stylistic features are most important for distinguishing fake news from real news in a low-resource linguistic environment. Through a detailed examination of Albanian-specific linguistic features, punctuation patterns, textual statistical properties,

structural composition features, and temporal features, the analysis demonstrates that reliable indicators of news authenticity are rooted in language use and journalistic practice rather than in increasingly complex model architectures. Features pertaining to sensational vocabulary, source attribution, lexical diversity, title-content alignment, and punctuation consistently manifest as reliable indicators, thereby affirming the significance of meticulously designed, language-specific features in the analysis of fake news.

Statistical analysis shows that these features exhibit consistent and interpretable behaviour across classes, supporting their suitability for transparent and explainable fake news detection systems. In particular, the results reinforce the view that hand-crafted features remain valid in scenarios where datasets are moderate in size, interpretability is required, or computational resources are limited, common features are common to resource-constrained languages such as Albanian.

Overall, this paper contributes to a systematic feature-level perspective on fake news research in Albanian, complementing existing model-based studies. The findings may inform future work on feature selection, feature importance analysis, and hybrid systems that combine statistical representations with learnt text embeddings. More broadly, the study highlights that careful feature analysis is a crucial step toward building reliable, interpretable, and implementable systems for fake news detection in low-resource linguistic environments.

8. REFERENCES

[1] Canhasi, E., Shijaku, R., & Berisha, E. (2022). *Albanian fake news detection*. Transactions on Asian and Low-Resource Language Information Processing, 21(5), 1–24.

[2] Leung, J., Vatsalan, D., & Arachchilage, N. (2023). *Feature analysis of fake news: Improving fake news detection in social media*. Journal of Cyber Security Technology, 7(4), 224–241.

[3] Agarwal, A., Agarwal, B., & Harjule, P. (2022). *Understanding the role of feature engineering in fake news detection*. In Soft Computing: Theories and Applications (SoCTA 2021) (pp. 769–789). Springer.

[4] Petrou, N., Christodoulou, C., Anastasiou, A., Pallis, G., & Dikaiakos, M. D. (2023). *A multiple change-point detection framework on linguistic characteristics of real versus fake news articles*. Scientific Reports, 13, 6086.

[5] Gereme, F., Zhu, W., Ayall, T., & Alemu, D. (2021). *Combating fake news in "low-resource" languages: Amharic fake news detection accompanied by resource crafting*. Information, 12(1), 20.

[6] Singh, V. K., Ghosh, I., & Sonagara, D. (2021). *Detecting fake news stories via multimodal analysis*. Journal of the Association for Information Science and Technology, 72(1), 3–17.

[7] Sousa-Silva, R. (2022). *Fighting the fake: A forensic linguistic analysis to fake news detection*. International Journal for the Semiotics of Law, 35(6), 2409–2433.

[8] Kasseropoulos, D. P., & Tjortjis, C. (2021). *An approach utilizing linguistic features for fake news detection*. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 646–658). Springer.

[9] Almarashy, A. H. J., Feizi-Derakhshi, M. R., & Salehpour, P. (2023). *Enhancing fake news detection by multi-feature classification*. IEEE Access, 11, 139601–139613.

[10] Haider, S., Luceri, L., Deb, A., Badawy, A., Peng, N., & Ferrara, E. (2023). *Detecting social media manipulation in low-resource languages*. In Companion Proceedings of the ACM Web Conference 2023 (pp. 1358–1364).

[11] Abdedaiem, A., Dahou, A. H., & Cheragui, M. A. (2023). *Fake news detection in low-resource languages using SetFit framework*. Inteligencia Artificial, 26(72), 178–201.

[12] De, A., Bandyopadhyay, D., Gain, B., & Ekbal, A. (2021). A transformer-based approach to multilingual fake news detection in low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(1), 1-20.

[13] Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., & Vilares, J. (2021). *Sentiment analysis for fake news detection*. Electronics, 10(11), 1348.

[14] Tajrian, M., Rahman, A., Kabir, M. A., & Islam, M. R. (2023). *A review of methodologies for fake news analysis*. IEEE Access, 11, 73879–73893.

[15] Hamed, S. K., Ab Aziz, M. J., & Yaakub, M. R. (2023). *A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion*. Heliyon, 9(10).

[16] Shu, K., Zhou, X., Wang, S., Zafarani, R., & Liu, H. (2019). *The role of user profiles for fake news detection*. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 436–439).

[17] Szabó Nagy, K., Kapusta, J., & Munk, M. (2023). *Feature extraction from unstructured texts as a combination of the morphological and the syntactic analysis and its usage in fake news classification tasks*. Neural Computing and Applications, 35, 22055–22067.

[18] Mohamed, A. O., Eltayeb, I. A. I., Ruslan, R. A., & Jeffry, N. E. (2025). *Fake news detection: A review of conventional and state-of-the-art approaches*. Journal of Applied Sciences Technology and Computing, 2(1), 43–54.

[19] Alghamdi, J., Lin, Y., & Luo, S. (2024). *Fake news detection in low-resource languages: A novel hybrid summarization approach*. Knowledge-Based Systems, 296, 111884.