

Is Less Really More?

Fake News Detection with Limited Information

Zhaoyang Cao
Data Lab, EECS Department
Syracuse University
zycao@data.syr.edu

John Nguyen
Syracuse University
jnguye30@syr.edu

Reza Zafarani
Data Lab, EECS Department
Syracuse University
reza@data.syr.edu

ABSTRACT

The threat that online fake news and misinformation pose to democracy, justice, public confidence, and especially to vulnerable populations has led to a sharp increase in the need for fake news detection and intervention. Whether multi-modal or pure text-based, most existing fake news detection methods depend on textual analysis of entire articles. However, these fake news detection methods come with certain limitations. For instance, fake news detection methods that rely on full text can be computationally inefficient, demand large amounts of training data to achieve competitive accuracy, and may lack robustness across different datasets. This is because fake news datasets have strong variations in terms of the level and types of information they provide; where some can include large paragraphs of text with images and metadata, and others can be a few short sentences. Perhaps if one could only use minimal information to detect fake news, fake news detection methods could become more robust and resilient to the lack of information. We aim to overcome these limitations by detecting fake news using systematically selected, limited information that is both effective and capable of delivering robust, promising performance. We propose a framework called **SLIM** (**S**ystematically-selected **L**imited **I**nformation) for fake news detection. In **SLIM**, we quantify the amount of information by introducing information-theoretic measures. **SLIM** leverages limited information (e.g., a few named entities) to achieve performance in fake news detection comparable to that of state-of-the-art obtained using the full text, even when the dataset is sparse. Furthermore, by combining various types of limited information, **SLIM** can perform even better while significantly reducing the quantity of information required for training compared to state-of-the-art language model-based fake news detection techniques.

1 Introduction

The demand for fake news detection and intervention has grown rapidly due to the threat that false news poses to democracy, justice, and public confidence [14; 29; 47]. Among several fake news detection methodologies, research has shown that advanced pre-trained large language models and multi-modal frameworks perform significantly better than traditional machine learning and deep learning models. Language models perform better as they can learn contextual text representations during pretraining [17]. One example is

the work by Bhatt *et al.*, which proposes a Siamese network framework with multiple branches built on the BERT architecture, where each branch is tailored to process distinct types of textual information (such as article bodies, and social media comments). Using this enhanced sequence model, the framework can achieve a competitive performance on fake news detection [7]. It has also been shown that systems that combine topical distributions (e.g., from Latent Dirichlet Allocation) with text representations from large language models perform well on fake news detection [11]. A multi-modal example is **SAFE**, which identifies fake news using textual and visual modalities. **SAFE** analyzes the semantic and visual consistency between news articles (text) and their accompanying images. Harnessing multi-modal information, **SAFE** enhances the accuracy of fake news detection across different media formats [46]. These fake news detection techniques mostly rely on the textual analysis of the entire text as the primary signal for fake news identification, whether they are pure text-based or multi-modal. Despite the significant successes of the aforementioned methods, we still cannot neglect some drawbacks of the full text-based approaches. One of the primary concerns is computational efficiency. In addition, full text may not always be available in datasets used to train fake news detection models. The negative impact on efficiency is particularly important in various application scenarios, especially those requiring real-time responses. Consequently, *relying on limited information to detect fake news is a more competitive option in practical applications*, as it significantly decreases computational complexity while remaining robust and efficient to data sparsity constraints [25; 29].

However, merely reducing the amount of information is not sufficient, but such limited information should be strategically identified to maintain effectiveness, and one cannot simply rely on better machine learning techniques or better large language models. In particular, while large language models have achieved promising results on fake news detection, future language models, as noted by Tamkin *et al.*, might make it difficult or impossible to identify disinformation when only relying on the text body of the news article [35]. Research has shown that humans can be deceived by news produced by the GPT-2 and other language models and human detection is expected to become more challenging [31]. “Full-text”-based detection techniques would be insufficient as advanced language models mimic the real distribution of human text [31]. Hence, while language models have been widely proven to outperform other generic models in fake news detection, we cannot neglect that the

rapid growth of language models will hinder human detection. As a result, the difficulty of identifying disinformation motivates research to rely more heavily on other limited yet subtle information cues in fake news articles. Such subtle cues will play an essential role in detecting online malicious activity, as also noticed by other research studies [31; 35]. But what are examples of such limited information cues? By surveying the literature [4; 31; 35; 47], we categorize such limited cues into three broad types: (a) *keywords*, (b) *sequences*, and (c) *metadata*.

Researchers have explored improving fake news detection by harnessing such limited information cues [4; 39; 43]. However, these efforts face two key challenges: (1) the approaches primarily *integrate* these cues (as extra machine learning features) with existing “full-text”-based models, making it unclear how limited information alone contributes to fake news detection; and (2) the integrations are often ad-hoc and rely heavily on feature engineering, leaving open questions about which types (or quantities) of limited information are most beneficial. Our goal in this paper is to address these challenges.

This paper: Fake Detection with Limited Information. We aim to identify the means to utilize limited information for fake news detection through a systematic analysis of various ways of extracting information from limited information (e.g., keyword extraction and sequence tagging). To ensure that, in fact, less information is used, we propose information-theoretic measures to assess information quantity. Subsequently, we explore how various types of key limited information can be combined. We utilize this newly identified key information as input in a language model to assess its impact on the effectiveness of fake news detection and broadly investigate the following research direction (details can be found in Section 4): 1. We assess the impact of different types of limited information on fake news detection; 2. We study the influence of multiple modalities of limited information on fake news detection; and 3. We compare the performance of utilizing limited information state-of-the-art models. In sum, our major contributions are:

- To the best of our knowledge, this work is the first to propose various quantified strategies for using limited information for fake news detection.
- We identified the optimal combinations of utilizing limited information yielding the highest detection accuracy by integrating various key pieces of information. Examples include combining keywords with sequence tagging or keywords with metadata.
- We explored the viability of using limited data as a substitute for text body in the realm of fake news detection. All codes are publicly available.¹

Section 2 formally presents the related work. Section 3 describes the proposed architecture of the SLIM framework, followed by framework evaluation and experiments that address our research questions in Section 4. Section 5 concludes this research with directions for future work.

¹The code and data is available at <https://github.com/kappakant/SLIM>

2 Related Work

We categorize limited information into three main types: keywords, sequences (e.g., POS, NER annotations), and metadata (e.g., titles, authors). This categorization is both (2.1) theoretically grounded and (2.2) empirically validated, as we will present next.

2.1 Theoretical Justification

This systematic selection is supported by extensive research in computational linguistics and information retrieval, demonstrating that these information sources provide a comprehensive representation of textual data for downstream tasks (e.g., fake news detection).

First, the use of keywords is well-supported in computational linguistics and information retrieval for fake news detection. Keywords capture salient lexical features that are often indicative of deceptive or manipulative texts. For instance, Pérez-Rosas *et al.*, showed in their experiments that certain keyword patterns, including sensational phrases or exaggerated emotional expressions, are powerful indicators of fake news, with high classification accuracy [23]. Similarly, keyword-based retrieval, such as those described by Manning *et al.*, [26], has been foundational in identifying misinformation documents.

Sequence tags provide syntactic and semantic structure to text, which is useful for detecting inconsistencies in fake news. Sousa-Silva highlighted that fake news often contains anomalous syntactic patterns, such as inconsistent verb tenses, which can be effectively captured by POS tagging [32]. NER helps identify entities that are frequently manipulated or misrepresented in fake news [29].

Finally, metadata plays a critical role in assessing credibility. Titles summarize the primary claim of a news article, and their linguistic features, such as clickbait patterns, have been studied by Kong *et al.*, in the context of fake news detection [18]. Author has been used by Castillo *et al.*, in their paper, demonstrating its importance in distinguishing reliable sources [8]. Together, these information sources—keywords, sequence tags, and metadata—form a comprehensive and robust foundation for fake news detection.

2.2 Empirical Justification

These three types of information have also been empirically validated, demonstrating their critical role in downstream tasks, such as fake news detection. In addition, these types can be combined in various capacities to form other types of limited information. We first review the related work on each type of information.

2.2.1 Keywords

Keywords are words that precisely and simply characterize an aspect of a subject stated in a document. They are crucial indicators of important textual information that spread among individuals [30]. Keywords can be extracted from textual documents using a variety of techniques, including statistical, rule-based, machine learning, or domain-specific approaches [30; 6]. However, to ensure that the extracted keywords are semantically consistent with the document, language model-based approaches that handle text to extract keywords can consider contextual information. As a result, the language models’ generated keywords might more accurately represent the content of the original text [12].

While keywords have been commonly used in fake news detection, systematic research on ways or how to use keywords is relatively lacking. Souza *et al.* proposed the Positive and Unlabeled Learning with the network-based Label Propagation (PU-LP) algorithm, which incorporates a keywords attention mechanism [9]. They employed Yake to extract keywords and then used these keywords in Graph Attention Neural Event Embedding (GNEE) to classify unlabeled nodes. Additionally, due to the unstructured texts of news on certain social media platforms, such as Twitter, Jayasiriwardene and Ganegoda utilized Core NLP and TF-IDF to extract keywords for more effective data collection for fake news detection. Additionally, to improve the precision and effectiveness of relevant news retrieval, they also used the WordNet lexical database to find synonyms and bigrams to generate proper key phrases [15].

2.2.2 Sequences

Sequence tagging, a fundamental task in natural language processing (NLP), involves the assignment of labels to individual tokens in a given sequence, such as words or subwords. These labels typically represent linguistic properties or semantic categories, facilitating various NLP tasks, including Part-Of-Speech tagging (POS), Named Entity Recognition (NER), and chunking. The significance of sequence tagging lies in its ability to discern syntactic roles, semantic entities, and even higher-order linguistic features by analyzing the sequential context of tokens. Furthermore, sequence tagging has great potential for detecting fake news. By leveraging its capacity to identify named entities and recognize linguistic patterns, sequence tagging can assist in the identification of fake information and misleading content [16; 33].

POS tagging: Some researchers have attempted to leverage sequence tagging methods for fake news detection. For instance, Balwant proposed an architecture that combines POS tag information from news articles using bidirectional long short-term memory (LSTM) and author profile information by convolutional neural network (CNN) [5]. His hybrid architecture showed high performance on the LIAR dataset. According to [21], certain POS tags are powerful indicators of emotional texts. For example, comparative adjectives (JJR) typically provide information or state facts, whereas superlative adjectives (JJS) are frequently used to express opinions. Positive text commonly features superlative adverbs (RBS) such as “most” and “best.” In addition, the choice of adjectives and adverbs can alter the meaning and semantics of a sentence. Pairing the same noun or verb with different adjectives or adverbs may result in different interpretations. However, such systematic combinations of POS tags in addition to how much and how often they are helpful have less been explored in research. The SLIM framework studied in this research will target such research gaps.

NER tagging: NER tags are also used for fake news detection. For instance, Al-Ash and Wibowo improved the BERT model by joining a NER and relational features classification (RFC) into a single formulation [1]. To improve generalization performance in joint learning, RFC and NER models shared the parameter layer in the BERT-joint framework. Shishah has introduced an approach to vector representation, which incorporates term frequency, inverse document frequency, and NERs [28]. However, the final results demonstrate that only term frequency yields the best performance when using an SVM classifier. This outcome may be due

to the absence of more advanced classifiers or the lack of a proper understanding of crucial information that might be useful in specific NERs. To address such issues, SLIM utilizes language models to extract varying percentages of keyword information and integrates them with proper sequence tags to detect fake news.

2.2.3 Metadata

Metadata is often used in fake news detection, where the common approach is to combine it (as extra features) with the full-text body and use it as input for fake news detection in *content-based fake news detection* [3; 19]. Content-based methods are often considered as the traditional approach to detect fake news, an area where researchers have made significant contributions [13; 22; 44; 47]. For instance, Wynne and Wint showed that highly accurate fake news classifiers can be trained using Gradient Boosting Classifiers and character n -grams as features in experiments [40]. Zhou *et al.* introduced the SAFE model, which investigates the multi-modal content (comprising textual and visual information) of news articles. Their case studies validate the effectiveness of the cross-modal relationship between both textual and visual features of news content [46].

A few studies have explored the role of metadata in fake news detection. For example, Elhadad *et al.* presented a novel approach to processing the entire textual content of news by extracting various textual features and a complex set of additional metadata-related features without dividing the news documents into sections [10]. They employ TF-IDF in the feature extraction phase. Similarly, Amine *et al.* utilized word embedding techniques and convolutional neural networks for feature extraction and compared various deep learning architectures applied to different metadata [2]. It is worth noting that past research did not consider the independent impact of metadata and it was always used as an add-on to improve fake news detection. Furthermore, metadata was often preprocessed using vectorizations such as TF-IDF or deep learning; hence, despite being a crucial and valuable limited piece of information, metadata is frequently underexplored.

Differing from existing works, SLIM explores various aspects of metadata, such as whether metadata can replace text and whether it can be augmented by other types of limited information, such as keywords (or sequence-tagging words), for detecting fake news.

2.2.4 Combining Various Types of Limited Information

Few studies have integrated the various types of limited information to tackle fake news detection. In a recent paper, Migyeong Yang *et al.* proposed a deep learning approach to debunk fake news about COVID-19 at its early stages [41]. They designed three embedding layers, the second of which is the Propagated Information Encoder (PIE). In this layer, they used NER tagging words and keywords to extract information for searching related YouTube videos. The text information from these videos, such as titles and descriptions, was then refined and used as input for this layer. Although their experiment, as a case study, successfully detected fake news on newly emerging and critical topics, it did not provide insights into where and how much limited information is necessary for fake news detection. Further-

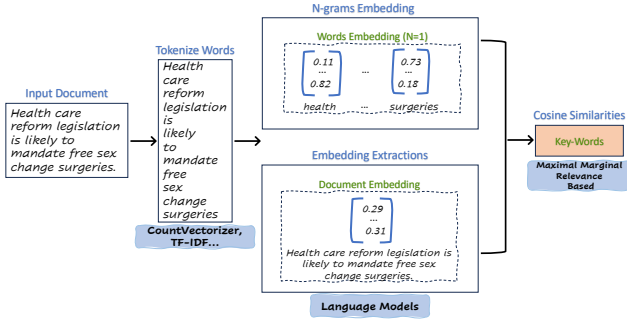


Figure 1: Extracting Keyword Information: the input is the body of the news under the proposed SLIM framework

more, they only used NER tagging words and keywords as a basis for searching videos rather than integrating these elements as the final set of features for detection.

3 The SLIM Framework

In the following subsections, we will first introduce the problem statement and framework formulation. Next, we will explain the approach to integrating information and performing the downstream task of fake news detection.

3.1 Problem Statement and Framework Formulation

Given an ordered set of the news article $A = \{w_1, w_2, \dots, w_p\}$, where w_i is the i th word, p is the total number of words in the article A . Our goal is to predict whether A is a fake news article ($\hat{y} = 0$) or a true one ($\hat{y} = 1$) by investigating its systematically-selected limited information.

SLIM Variations: We have four variations of SLIM based on the types of inputs that each variation takes. Variations of the framework represent the different systematically selected features of information. For notation clarity, we define them as $\text{SLIM}_{\text{KEYWORD}}$, $\text{SLIM}_{\text{SEQUENCE}}$, $\text{SLIM}_{\text{METADATA}}$, and $\text{SLIM}_{\text{MULTIMODAL}}$. In the following sections, we will introduce the preprocessing steps required to build these variations.

3.1.1 $\text{SLIM}_{\text{KEYWORD}}$

The first variation $\text{SLIM}_{\text{KEYWORD}}$ takes keywords as input. The process of extracting keyword information is depicted in Figure 1. It includes varying percentages of keywords. To obtain the $\text{SLIM}_{\text{KEYWORD}}$, we first use BERT to obtain the document embedding $e_d \in \mathbb{R}^n$. Meanwhile, we use the N-grams for word embeddings. When $N = 1$, we can get word embedding e_{w_i} for an arbitrary i th word. Then we calculate the cosine similarity (denoted as S_{cosine} , given in Equation 2) between document embedding e_d and each e_{w_i} and retain the set of words with a cosine similarity greater than 0. We constrain the extraction process by the maximal marginal relevance (MMR) to avoid the redundancy of the sorting results and to ensure the correlation of the words (stated in Equation 3). The process of MMR is summarized in Algorithm 1. Formally, the final input $\text{SLIM}_{\text{KEYWORD}}$ is the set of keywords defined by

$$\text{SLIM}_{\text{KEYWORD}} = \{w_i | S_{\text{cosine}}(e_{w_i}, e_d) > 0\}, \quad (1)$$

Algorithm 1 MMR in $\text{SLIM}_{\text{KEYWORD}}$

```

1: Input:  $C, |A|, e_d, e_w, k$ 
2: Output:  $R$ 
3: Initialization:  $R = \emptyset, C = \text{SLIM}_{\text{KEYWORD}}$  (set of words that satisfy Equation 1)
4: while  $|R| < \lfloor |A| \cdot k \rfloor$  do
5:    $w^* = \arg \max_{w_i \in C} [\lambda \text{sim}(e_d, e_{w_i}) - (1 - \lambda) \max_{w_j} \text{sim}(e_{w_i}, e_{w_j})]$ 
6:    $R \leftarrow R \cup \{w^*\}, C \leftarrow C \setminus \{w^*\}$ 
7: end while
8: return  $R$ 

```

$$S_{\text{cosine}}(e_d, e_{w_i}) = \frac{e_d \cdot e_{w_i}}{\|e_d\| \cdot \|e_{w_i}\|}, \quad (2)$$

$$\text{MMR}(e_d, C, R) = \arg \max_{w_i \in C} [\lambda \text{sim}(e_d, e_{w_i}) - (1 - \lambda) \max_{w_j \in R} \text{sim}(e_{w_i}, e_{w_j})] \quad (3)$$

where e_d is the document embedding, C is the set of collected words, R is the returned result set, e_w is the word embedding, and sim refers to the cosine similarity S_{cosine} . At last, λ is the diversity and set to 0.5. Finally, k (in Algorithm 1) is the proportion of the desired number of words relative to the total number of words in the full text. By adjusting the value of k , we can derive keywords with the desired varying word counts.

3.1.2 $\text{SLIM}_{\text{SEQUENCE}}$

In $\text{SLIM}_{\text{SEQUENCE}}$, the framework uses both POS and NER tags as input; the input comprises sets of words from different sequence taggings. For POS tagging, we initially tokenize the news articles. Once we obtain the corresponding tokens, we employ the *pos.tag* function for POS tagging. We filter out adjectives and adverbs, storing them in a word set SLIM_{POS} . Finally, we perform a subset operation on SLIM_{POS} to extract varying proportions of words. Specifically, after obtaining SLIM_{POS} , we extract the top k proportion of words based on their indices, where k corresponds to the desired proportion of the total word count. For NER tagging, the tokenization process is similar to POS tagging. After obtaining tokens, we use the *ne.chunk* function to extract the filtered named entities, which are then stored in words set SLIM_{NER} . We do not perform additional operations and restrictions for NER words since the named entities in an article are generally not too many, such as a person, location, and the like.

3.1.3 $\text{SLIM}_{\text{METADATA}}$

The input to $\text{SLIM}_{\text{METADATA}}$ consists solely of metadata to explore whether metadata can replace lengthy texts as key information for fake news detection. The metadata contained in different datasets varies. In light of the aforementioned papers, we will focus on textual data such as **title** (which we denote as $\text{SLIM}_{\text{TITLE}}$) and **author** ($\text{SLIM}_{\text{AUTHOR}}$) rather than discrete data. **XLNet**_{base} is used as the encoder to generate embeddings for metadata and other types of information.

3.1.4 $\text{SLIM}_{\text{MULTIMODAL}}$

The input to $\text{SLIM}_{\text{MULTIMODAL}}$ involves integrations of different types of aforementioned inputs such as various percentages of keywords sets and NER words ($\text{SLIM}_{\text{KEYWORD}} \oplus \text{SLIM}_{\text{NER}}$),

as well as combinations of keywords sets and different types of metadata ($\text{SLIM}_{\text{KEYWORD}} \oplus \text{SLIM}_{\text{METADATA}}$). Formally,

$$\text{SLIM}_{\text{MULTIMODAL}}^I = \text{SLIM}_{\text{KEYWORD}} \oplus \text{SLIM}_{\text{NER}} \quad (4)$$

$$\text{SLIM}_{\text{MULTIMODAL}}^{II} = \text{SLIM}_{\text{KEYWORD}} \oplus \text{SLIM}_{\text{AUTHOR}} \quad (5)$$

$$\text{SLIM}_{\text{MULTIMODAL}}^{III} = \text{SLIM}_{\text{KEYWORD}} \oplus \text{SLIM}_{\text{TITLE}} \quad (6)$$

where $\oplus$ is the concatenation operator.

Framework: Given an input sequence x , we define its length as T (the number of words). During the pre-training phase, although we employ $\text{XLNet}_{\text{base}}$ as our pre-training model, the pre-training objective function is indeed crucial. This is because it facilitates a deeper understanding of the semantic and structural relationships inherent within the text. Throughout the pre-training process, this objective function enables the model to discern between distinct categories of keyword combinations (e.g., real news versus fake news), which gives the downstream classification tasks more robust features. The pre-training objective function, as defined in Equation 7, employs XLNet’s permutation language modeling to capture contextual information from the input.

$$\mathcal{F}(\theta) = \max_{\theta} \mathbb{E}_{z \sim \mathcal{Z}_T} \left[\sum_{t=1}^T \log p(x_{zt} \mid \mathbf{x}_{z_{<t}}; \theta) \right], \quad (7)$$

where in our case, $\mathbf{x}$ is the $\text{SLIM}_{\text{KEYWORD}}$ (and other defined inputs), $\mathcal{Z}_T$ represents the set of permutations of keywords set of length T . We use zt to represent the t_{th} element in $\mathcal{Z}_T$, and $z_{<t}$ to represent the 1_{st} to $t-1$ elements of $z \in \mathcal{Z}_T$. The likelihood function in equation 7 is defined as

$$p_{\theta}(X_{zt} = x \mid \mathbf{x}_{z_{<t}}) = \frac{\exp(e(x)^T g_{\theta}(\mathbf{x}_{z_{<t}}, z_t))}{\sum_{x'} \exp(e(x')^T g_{\theta}(\mathbf{x}_{z_{<t}}, z_t))}, \quad (8)$$

where g_{θ} is the two-stream self-attention model.

Fake News Detection: Finally, we will conduct the downstream task, which is fake news detection. Building upon the aforementioned inputs, we will directly load the pre-trained weights of $\text{XLNet}_{\text{base}}$ model and fine-tune it using our defined SLIM variants. The loss function in the fine-tuning stage of the SLIM framework is the cross entropy loss.

$$\mathcal{L}_{\text{SLIM}}(\theta) = \mathcal{L}_{\text{CE}}(\theta) = -\frac{1}{N} \sum_{i=1}^N \sum_{c \in \mathcal{C}} y_{i,c} \log p(y_{i,c} = 1 \mid x_i, \theta), \quad (9)$$

where N is the sample size, and y is the label for the input words set. The parameter θ is updated by:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{m_t}{\sqrt{v_t} + \epsilon}, \quad (10)$$

where η is the learning rate and set to $5 \cdot 10^{-5}$, m_t is the momentum estimate, v_t represents the squared gradient estimate, and ϵ is the stability constant and set to $1 \cdot 10^{-8}$. For the prediction, the optimization target is to minimize the cross-entropy loss between the predicted logits from the fine-tuned $\text{XLNet}_{\text{base}}$ model and the ground-truth labels of the fake news detection task under the Adam optimizer. At last, the predicted label is obtained by applying the argmax function to the logits, selecting the class with the highest predicted probability as the output. Mathematically,

$$\hat{y} = \text{argmax}_i(\mathbf{z}_i), \quad (11)$$

where $\hat{y}$ is the predicted label and $\mathbf{z}_i$ is the logits computed by the final layer of the SLIM framework.

3.2 Quantifying Limited Information

In order to better quantify and compare the information density of limited information with that of the full text, we employed two methods. The first method targets information density: we have proposed a method based on Shannon entropy, which we refer to as *normalized Shannon entropy* for fake news detection. The second method explores the relationship of average token counts, which not only provides a more intuitive representation of the difference in information volume between inputs but also illustrates that fewer tokens correspond to reduced costs for future, more extensive commercial language models.

3.2.1 Normalized Shannon Entropy

In information theory, Shannon entropy [27] measures the average uncertainty of information and is defined as:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x), \quad (12)$$

where $p(x)$ is the probability of x in the distribution $\mathcal{X}$. In the context of a news article $A = \{w_1, w_2, \dots, w_p\}$, $p(x)$ is modeled as the relative frequency of each word w_i in the article. Words that appear more often in the article have a higher probability of being randomly chosen from the article. Thus, we define the significance level of a word w as:

$$\text{sig}(w) = \frac{f_w^{\mathcal{T}}}{|\mathcal{T}|}, \quad (13)$$

where $f_w^{\mathcal{T}}$ is the word frequency of w within the original full text $\mathcal{T}$, and $|\mathcal{T}|$ represents the total number of words in the full text $\mathcal{T}$. Thus, we can represent the information density by calculating the information score $S_{\text{normalized}}$ under normalized Shannon entropy of an arbitrary article A as:

$$S_{\text{normalized}} = \sum_{w \in A} \frac{H(w)}{\text{sig}(w)} \quad (14)$$

Mathematical Interpretation By dividing Shannon entropy by significance level, we can obtain the average information uncertainty per unit of the significance level. This ratio helps to numericalize the information density of each unit of importance. Additionally, when the range of significance levels is broad (e.g., some words are very frequent while others are rare), dividing Shannon entropy by significance level helps to mitigate the scale effect, making the measure of information density more consistent.

3.2.2 Average Token Counts

Tokens serve as the building blocks of the original text, enabling the model to process and generate natural language in a structured way [38]. A fixed tokenizer aims to maintain a consistent informational value for each token, so a reduction in token count generally conveys less information and diminishes the expression of information. Hence, we calculated the average token count for different types of inputs and compared them with the token count of the full text to verify that our inputs are sparser.

Table 1: Performance comparison of datasets on the SLIM, CapsNet, MisROBERTA, and selected DocEmb models. The percentage of keywords used in comparisons for both types is 25%. The best performance is highlighted in bold, and the second best is underlined.

Method	Dataset	
	ReCOVery	Fake_And_Real_News
DocEmb_TFIDF BiLSTM	89.56±0.0025	92.26±0.0032
DocEmb_TFIDF BiGRU	90.54±0.0017	92.60±0.0028
DocEmb_BERT BiLSTM	90.27±0.0033	93.05±0.0026
DocEmb_BERT BiGRU	90.13±0.0014	93.07±0.0051
MisROBERTA	91.35±0.0066	97.34±0.0076
BiLSTM_CapsNet	95.49±0.0134	95.56±0.0091
SLIM	95.55±0.0046	97.60±0.0031
SLIM _{KEYWORD}	92.86±0.0070	92.76±0.0016
SLIM ^{III} _{MULTIMODAL}	93.72±0.0074	93.72±0.0049

3.2.3 Information Density Comparisons

To compare whether the different types of input we designed in section 3.1 indeed contain limited and less information, we calculated the information density of each type using the average token count and proposed normalized Shannon entropy score. The results on the ReCOVery dataset, presented in Figure 2a and 2b separately, reveal the following: the title exhibits the lowest normalized Shannon score (91.03) and count of tokens (15.88) due to its inherent conciseness as part of the metadata. NER words, as an effective representation for identifying and classifying key entities, also show a low score of 354.87, which is 10% of the full text, and token counts of 88.39, 8.59% of the full text. Additionally, both POS words and keywords, with the default 10% proportion, demonstrate significantly lower Shannon scores and token counts compared to the full text. It is noteworthy that both information density evaluation metrics for POS words do not exhibit a linear increase as the percentage rises. The figures of normalized Shannon entropy score and average token count for the remaining two datasets are presented in Appendix B.

3.3 Fake News Detection

Ultimately, we will conduct the downstream task, which is fake news detection. Building upon the aforementioned framework formulation, we will employ language models for fake news detection, as language model-based approaches currently yield the best performance for detecting fake news.

3.3.1 Base configurations

We will use **XLNet**_{base} as the encoder to generate the corresponding embeddings of the input information [42]. We use Adam in the optimization process. For the prediction phase, we apply the **argmax** function to the logits from XLNet to obtain the final prediction label. Mathematically,

$$\hat{y} = \operatorname{argmax}_i(\mathbf{z}_i), \quad (15)$$

where $\hat{y}$ is the predicted label and $\mathbf{z}_i$ is the logits computed by the final layer of the SLIM framework.

Table 2: Dataset statistics

Dataset	Labels	Train	Validation	Test
RECOVERY	Truth	966	278	120
	Fake	487	114	64
FAKE_REAL_NEWS	Truth	1143	557	597
	Fake	1154	592	551

4 Experimental Results

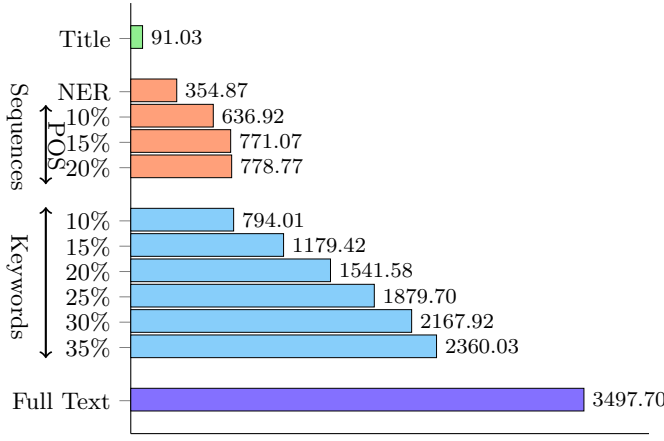
In this section, we will introduce the experimental setup, including preprocessing and datasets. Subsequently, we conducted extensive experiments to address the following five research questions, RQ1 through RQ5. The research questions are as follows: **RQ1**: How does SLIM compare to other baselines? **RQ2**: How effective are keywords for fake news detection? **RQ3**: How effective are sequences for fake news detection? **RQ4**: How effective is metadata for fake news detection? **RQ5**: Can multiple modalities of limited information enhance fake news detection?

4.1 Experimental Setup

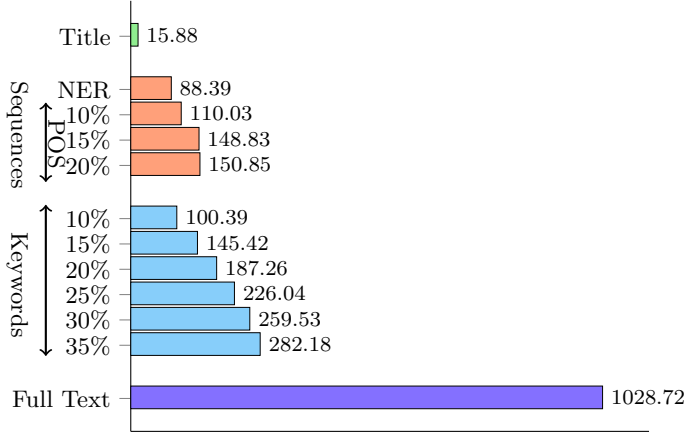
For each experiment, we conducted five trials to obtain the average accuracy. During data preprocessing, paragraph separators ‘\n’ were removed, and all text was converted to lowercase to ensure consistency.

4.1.1 Dataset

Our experiments are conducted on two public benchmark datasets of fake news detection: ReCOVery [45], and Fake-And-Real-News [20]. The division of training, validation, and testing sets in the ReCOVery are in the same way as the articles from which they are derived. The training, validation, and testing sets are divided in a ratio of 50% : 25% : 25% in the Fake-And-Real-News dataset. The basic statistics of the datasets and detailed source descriptions of these datasets are in Table 2 and Appendix A.



(a) Representation of information density by average normalized Shannon entropy ($\bar{S}_{normalized}$) on the ReCOVeRY dataset. The Title shows the lowest normalized Shannon score (91.03). NER words also have lower Shannon scores (~10% of full text density). Keywords and POS words at 10% threshold show significantly lower Shannon scores than full text



(b) Representation of information density by the average count of tokens on the ReCOVeRY dataset. The title and NER words have the lowest average token count among all types. All inputs, including sequences and keywords at different percentages, have much lower token counts than full text, with the highest reaching only ~27% of full text length.

Figure 2: Representation of information density by average normalized Shannon entropy (a) and the average count of tokens (b) on the ReCOVeRY dataset

4.1.2 Metadata Selection

The metadata we selected to use in our work contains textual data only. To be specific, `title`, `author` are selected in the ReCOVeRY dataset. Meanwhile, `title` is selected in the Fake_And_Real_News dataset. Additionally, only `author` is selected in the ReCOVeRY dataset.

4.1.3 Evaluation Metrics

We report accuracy, macro- F_1 , and AUC. We also conduct statistical significance comparisons between different experimental groups. We use ** to represent p -values below 0.01 and use * to represent p -values between 0.01 and 0.05 for two groups. The absence of asterisks indicates that there is no statistically significant difference between the two experimental groups.

RQ1: How does SLIM compare to other baselines?

In this section, we present a comprehensive comparative analysis between our proposed SLIM framework against various state-of-the-art models, including different deep learning models and large language models. The baseline models we employed are described as follows.

► **DocEmb**: DocEmb was proposed by Truică and Apostol [37]. Instead of relying on handcrafted features or complex deep learning architectures, the approach utilizes pre-trained document embeddings to capture the semantic meaning of news articles. These embeddings are then fed into models of neural network architecture. Based on the combinations with good performance presented in their paper, we utilize 4 different combinations in our work: 2 vectorization methods (TF-IDF, BERT) combined with 2 downstream neural network models (BiLSTM and BiGRU).

► **BiLSTM_Capsnet**: BiLSTM_Capsnet was proposed by Sridhar and Sanagavarapu [34]. The framework uses a multi-task learning architecture. The architecture’s subtasks include modeling the article contents, and the shared common task is determining whether or not the article is fake. The BiLSTM network is used to model the subtasks, and CapsNet serves as the common meta classifier.

► **MisROBERTA**: MisROBERTA was proposed by Truică and Apostol [36]. The model incorporates various techniques, such as data augmentation and adversarial training, to improve its robustness in detecting misleading content.

We first conducted experiments and obtained our baseline results of the datasets under the SLIM framework. The baseline entails using only the full-text body as input to build the XLNet model for prediction accuracy. The results of the SLIM baseline are presented in Table 3. We observed that the full text exerts heterogeneous impacts, however, the prediction accuracy for all datasets exceeded 93%.

The comparison of the performance of different baselines is shown in Table 1. The results illustrate that, compared to other baseline models, the SLIM achieved the highest accuracy in both the ReCOVeRY and Fake_And_Real_News dataset. Meanwhile, by using only keywords with half the information density of the full text, we are able to achieve impressive accuracy. Not only does this performance closely approach some state-of-the-art fake news detection models (e.g., MisROBERTA), but it also surpasses many of the latest deep learning and language model-based approaches (e.g., DocEmb). Moreover, when we combine keywords with the title (which always has the lowest information density), the accuracy is further improved.

RQ2: How effective are keywords for fake news detection?

Subsequently, we primarily investigated the impact of limited yet effective information (except metadata) mentioned in the first two phases of section 3 on the SLIM_{KEYWORD} framework. Initially, we explored the effect of keywords on fake news detection. We extracted keyword sets from different datasets using the methodology outlined in section 3.1.1. Additionally, for each dataset, we attempted to extract the maximum percentage of keywords feasible (rounded down using the floor function). We set the default, i.e., the minimum percentage of keywords, to be 10% of the original full text. Then, for each dataset, we gradually increased the per-

Table 3: Performance comparison of datasets of the SLIM (full-text) baseline frameworks: The performance of all datasets in fake news detection using the SLIM framework exceeded 93%.

Experiments	ReCOVery			Fake_And_Real_News		
	Accuracy	Macro- F_1	AUC	Accuracy	Macro- F_1	AUC
SLIM	95.55 $\pm$ 0.0046	94.71	95.53	97.60 $\pm$ 0.0031	97.60	97.62

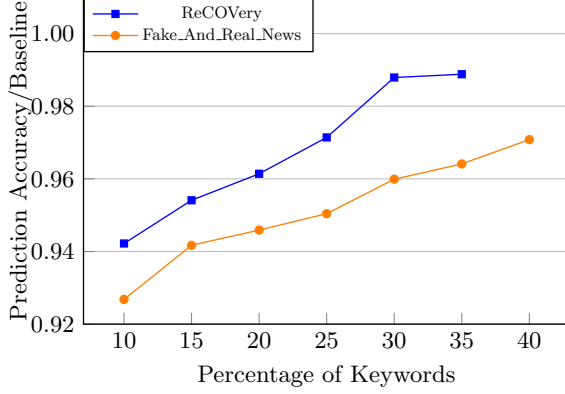


Figure 3: Performance comparison of datasets of the SLIM_{KEYWORD} frameworks. All datasets achieve an accuracy ratio of over 96% when we extract 30% of the keywords, among which the ReCOVery datasets showed an approximately 99% accuracy ratio.

centage of keywords extracted by 5% for experimentation. The results are depicted in Figure 3. From Figure 3, we set the y-axis as the prediction accuracy divided by the baseline accuracy (referred to as the *accuracy ratio*), as this provides a more intuitive way to visualize the impact of keywords on detection from both the graphical and numerical perspectives. The following figures utilize this y-axis configuration. In summary, for all datasets, there is an overall trend of increasing accuracy ratio as the percentage of keywords increases. Across all datasets except for the Fake_And_Real_News dataset, once the extracted keywords reach 30% of the text, we observe that the accuracy ratio reaches approximately 99%. This indicates that comparable and good performance can be achieved by extracting only 30% of the full text, significantly reducing computational inefficiency and enhancing scalability for large datasets. This finding implies that keyword extraction can effectively filter out irrelevant words and information in fake news detection.

RQ3: How effective are sequences for fake news detection?

Within the SLIM_{SEQUENCE} framework, we also explored the impact of POS tagging words and NER tagging words on fake news detection. For POS tagging words, adjectives and adverbs are particularly powerful for enhancing fake news detection, given their frequent usage in texts to express authors’ opinions and emotions. Therefore, we adopted a similar approach to extracting the percentage of POS tagging words as with keywords. As for NER words, since the occurrence of named entities in texts is not typically abundant, we did not impose any percentage limitations during extraction. Our experiments demonstrated that, across the two datasets, NER tagging words accounted for approximately 10%, which is consistent with our default minimum

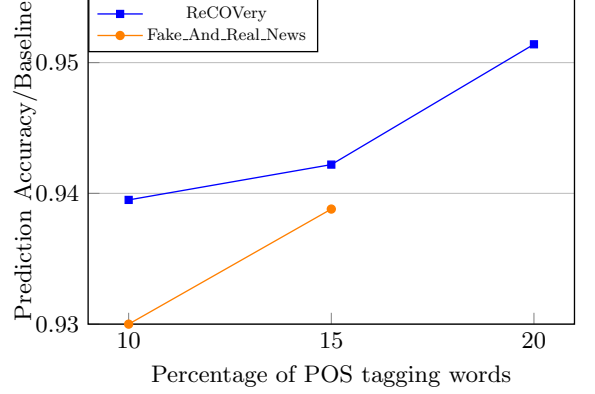


Figure 4: Performance comparison of datasets of the SLIM_{SEQUENCE} frameworks in POS tagging words. The percentage of POS tagging words (primarily adjectives and adverbs) that can be extracted from the full text is approximately 10% to 20%. However, using a small number of POS tagging words can achieve an accuracy ratio of 94%.

percentage. The results of SLIM_{SEQUENCE} framework regarding POS tagging and NER tagging are presented in Figure 4 and Table 4, respectively. We can observe from Figure 4 that the maximum percentage of POS tagging words that can be extracted from the Fake_And_Real_News datasets is 15%. Meanwhile, the ReCOVery dataset allows for the extraction of up to 20% of the POS tagging words from the original text. As a result, POS tagging shows an overall increasing trend across all datasets, where the accuracy ratio increases as the percentage of POS tagging words increases. However, compared to the performance of keywords, the accuracy ratio of POS tagging words remains around 94%. Secondly, regarding the NER tagging words performance, in the ReCOVery dataset, NER tagging words achieve an 86.82% accuracy (which is significantly lower than the baseline accuracy) and an accuracy ratio of 93%. The prediction accuracy for the Fake_And_Real_News dataset is 90.08%, with p -value between 0.01 and 0.05, indicating a significant decrease compared to the baseline.

RQ4: How effective is metadata for fake news detection?

In practical scenarios, we often observe a partial overlap between the information contained in metadata (such as **title**) and the content of the text body [24]. As a result, the overlapped information is redundantly utilized during tokenization, leading to reduced efficiency and increased consumption of embedding resources. Therefore, we aim to mitigate the drawbacks mentioned before. As metadata usually contains the minimum of information needed to distinguish an article, we aim to explore whether fake news detection can be achieved only through metadata, replacing the need for the full-text body. We exclusively use metadata as the

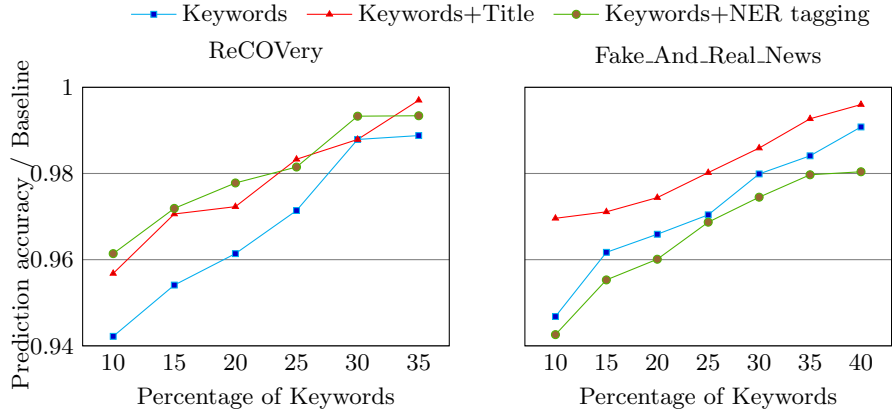


Figure 5: Performance comparison of datasets of the $\text{SLIM}_{\text{MULTIMODAL}}$ frameworks. Generally, the integration of different types of limited information improves fake news detection accuracy compared to using only keywords ($\text{SLIM}_{\text{KEYWORD}}$). In the Fake_And_Real_News dataset, the performance of keywords and NER words shows an approximately 0.5% decline compared to using only keywords.

Table 4: Performance comparison of datasets of the $\text{SLIM}_{\text{SEQUENCE}}$ frameworks in NER tagging words. The performance of NER words exhibits heterogeneous effects across different datasets

Dataset	SLIM_{NER}		
	Accuracy	Macro- F_1	AUC
RECOVERY	$86.82^{**} \pm 0.0078$	83.78	83.29
FAKE_AND_REAL_NEWS	$90.08^{*} \pm 0.0092$	90.08	90.14

Table 5: Performance of the metadata-only framework ($\text{SLIM}_{\text{METADATA}}$). Metadata cannot substitute text, yielding results significantly lower to the results obtained using text alone.

Dataset	$\text{SLIM}_{\text{METADATA}}$		
	Accuracy	Macro- F_1	AUC
RECOVERY (TITLE)	$82.25^{**} \pm 0.0066$	78.14	77.80
RECOVERY (AUTHOR)	$76.99^{**} \pm 0.0071$	74.54	77.62
FAKE_AND_REAL_NEWS (TITLE)	$85.21^{**} \pm 0.0034$	85.19	85.42

input, feeding it directly into the $\text{SLIM}_{\text{METADATA}}$ framework to obtain the results. To be more precise, for the ReCOVery dataset, its metadata includes both **author** and **title**. Therefore, we input these two pieces of metadata separately to obtain the results. However, for the Fake_And_Real_News dataset, its metadata only includes the **title**. Hence, the input is the **title**. The results are in Table 5.

We discover that, from Table 5, utilizing only textual metadata (**title** and **author** in this work) as input for the fake news detection results in a statistically significant decrease in prediction accuracy compared to the baseline (which uses the full-text body as the input) performance. Specifically, in the ReCOVery and Fake_And_Real_News dataset, when using metadata alone as a single input for detection under the $\text{SLIM}_{\text{METADATA}}$ framework, the accuracy generally decreases by approximately 10% compared to the baseline. Without considering any text, we could not achieve the same level of accuracy by exclusively using metadata for fake news detection. However, if aiming for a relatively good level of accuracy, we can use metadata or selectively combine less

information of full text for future fake news detection.

RQ5: Can multiple modalities of limited information enhance fake news detection?

In this section of the experiment, we aim to investigate whether combining different pieces of limited key information can enhance the performance of the $\text{SLIM}_{\text{MULTIMODAL}}$ framework. Initially, for each dataset, we combined their respective percentages of keywords and NER tagging words. As mentioned in the methodology, we concatenated these two distinct word sets together to form a composite input for the encoder. The final results are depicted in Figure 5. Additionally, we sought to integrate keyword information with metadata to assess whether metadata could serve as additional information to enhance the performance. The results are also presented in Figure 5.

The results in Figure 5 lead us to the following conclusions. Firstly, in the ReCOVery dataset, we found that the integration of limited information $\text{SLIM}_{\text{MULTIMODAL}}$: keywords + title, keywords + NER tagging words) improves detection performance compared to using only keywords for fake news detection. Furthermore, we observed that NER words have a greater impact on fake news detection than metadata (**title**). Finally, in the Fake_And_Real_News dataset, metadata can still be experimentally verified as useful for improving accuracy when combined with keywords. However, it is worthily noted that the heterogeneous effects of NER tagging exist, such that combining keywords with NER words results in a slight accuracy reduction of approximately 0.5% compared to the $\text{SLIM}_{\text{KEYWORD}}$.

5 Conclusion and Future Work

In this work, we systematically investigated the viability of limited-information strategies for fake news detection using the SLIM framework. We investigated and conducted extensive experiments with different types of information strategies: keyword extraction, sequence tagging, and textual metadata. Our empirical analysis demonstrates that strategic keyword extraction preserves critical information even under severe sparsity constraints: retaining merely 30% of full-text keywords achieves a near-perfect accuracy ratio

(99%) across multiple benchmarks. Linguistic tagging experiments further revealed that limited syntactic-semantic representations suffice for detection. Constrained POS and NER tagging sets independently achieved a 92% accuracy ratio. While metadata exhibited diminished standalone performance, its complementary role in the multimodal framework proved statistically significant. Our systematic evaluation of multi-modality limited information demonstrates that multi-view fusion of keywords, named entities, or contextual titles achieves substantial performance increase: not only does this combination surpass single-modality keyword analysis, but it also consistently outperforms state-of-the-art neural network approaches across two benchmark datasets. Our findings substantiate that strategically selected information subsets can achieve accuracy parity with full-text analysis, establishing an efficiency-optimized framework for fake news detection and providing guidelines for sparse-data environments where full-text acquisition is impractical. Future work will focus on enhancing robustness through syntactic-semantic augmentation techniques, including controlled paraphrase generation and dependency shuffling.

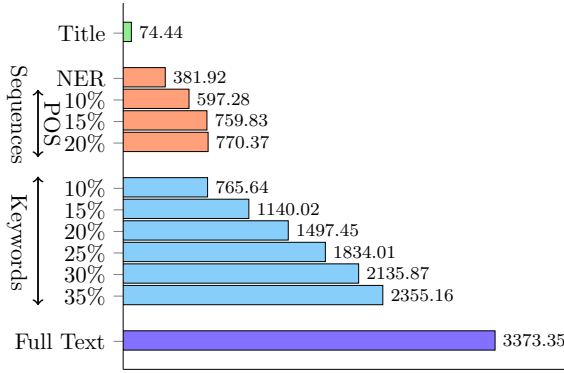
6 Acknowledgements

This research was supported in part by the National Science Foundation under award No. 2241070.

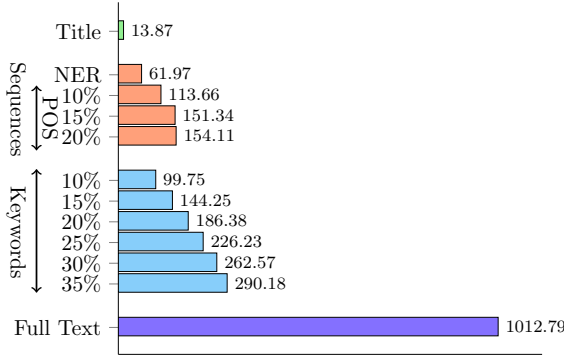
7 REFERENCES

- [1] H. S. Al-Ash and W. C. Wibowo. Fake news identification characteristics using named entity recognition and phrase detection. In *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 12–17. IEEE, 2018.
- [2] B. M. Amine, A. Drif, and S. Giordano. Merging deep learning model for fake news detection. In *2019 International Conference on Advanced Electrical Engineering (ICAEE)*, pages 1–4. IEEE, 2019.
- [3] W. Antoun, F. Baly, R. Achour, A. Hussein, and H. Hajj. State of the art models for fake news detection tasks. In *2020 IEEE international conference on informatics, IoT, and enabling technologies (ICIOT)*, pages 519–524. IEEE, 2020.
- [4] S. Arora, S. Wu, E. Liu, and C. Ré. Metadata shaping: A simple approach for knowledge-enhanced language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1733–1745, 2022.
- [5] M. K. Balwant. Bidirectional lstm based on pos tags and cnn architecture for fake news detection. In *2019 10th International conference on computing, communication and networking technologies (ICCCNT)*, pages 1–6. IEEE, 2019.
- [6] S. K. Bharti and K. S. Babu. Automatic keyword extraction for text summarization: A survey. *arXiv preprint arXiv:1704.03242*, 2017.
- [7] S. Bhatt, N. Goenka, S. Kalra, and Y. Sharma. Fake news detection: Experiments and approaches beyond linguistic features. In *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2021, Volume 2*, pages 113–128. Springer, 2022.
- [8] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011.
- [9] M. C. de Souza, M. P. S. Gôlo, A. M. G. Jorge, E. C. F. de Amorim, R. N. T. Campos, R. M. Marcacini, and S. O. Rezende. Keywords attention for fake news detection using few positive labels. *Information Sciences*, 663:120300, 2024.
- [10] M. K. Elhadad, K. F. Li, and F. Gebali. A novel approach for selecting hybrid features from online news textual metadata for fake news detection. In *3PGCIC-2019 14*, pages 914–925. Springer, 2020.
- [11] A. Gautam, V. Venkatesh, and S. Masud. Fake news detection system using xlnet model with topic distributions: Constraint@ aaai2021 shared task. In *International Workshop on Combating On line Hostile Posts in Regional Languages during Emergency Situation*, pages 189–200. Springer, 2021.
- [12] M. Grootendorst. Keybert: Minimal keyword extraction with bert., 2020.
- [13] G. B. Guacho, S. Abdali, N. Shah, and E. E. Papalexakis. Semi-supervised content-based detection of misinformation via tensor embeddings. In *2018 IEEE/ACM ASONAM*, pages 322–325. IEEE, 2018.
- [14] L. Harriss and K. Raymer. Online information and fake news. *Parliamentary Office of Science and Technology, POSTnote*, 559, 2017.
- [15] T. D. Jayasiriwardene and G. U. Ganegoda. Keyword extraction from tweets using nlp tools for collecting relevant news. In *2020 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pages 129–135. IEEE, 2020.
- [16] J. Kapusta, M. Drlik, and M. Munk. Using of n-grams from morphological tags for fake news classification. *PeerJ Computer Science*, 7:e624, 2021.
- [17] J. Y. Khan, M. T. I. Khondaker, S. Afroz, G. Uddin, and A. Iqbal. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4:100032, 2021.
- [18] S. H. Kong, L. M. Tan, K. H. Gan, and N. H. Samsudin. Fake news detection using deep learning. In *2020 IEEE 10th symposium on computer applications & industrial electronics (ISCAIE)*, pages 102–107. IEEE, 2020.
- [19] V. M. Krešňáková, M. Sarnovský, and P. Butka. Deep learning methods for fake news detection. In *2019 IEEE 19th international symposium on Computational Intelligence and informatics and 7th IEEE international conference on recent achievements in mechatronics, automation, computer sciences and robotics (CINTI-MACRo)*, pages 000143–000148. IEEE, 2019.
- [20] G. McIntire. https://github.com/GeorgeMcIntire/fake_real_news_dataset.

- [21] A. Pak, P. Paroubek, et al. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.
- [22] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li, and J. Liu. Content based fake news detection using knowledge graphs. In *The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I 17*, pages 669–683. Springer, 2018.
- [23] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.
- [24] A. Piotrkowicz, V. Dimitrova, and K. Markert. Automatic extraction of news values from headline text. In *Proceedings of the student research workshop at the 15th conference of the European chapter of the association for computational linguistics (EACL SRW 2017)*, pages 64–74. Association for Computational Linguistics, 2017.
- [25] S. Raza and C. Ding. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362, 2022.
- [26] H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.
- [27] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- [28] W. Shishah. Fake news detection using bert model with joint learning. *Arabian Journal for Science and Engineering*, 46(9):9115–9127, 2021.
- [29] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.
- [30] S. Siddiqi and A. Sharan. Keyword and keyphrase extraction techniques: a literature review. *International Journal of Computer Applications*, 109(2), 2015.
- [31] I. Solaiman, M. Brundage, J. Clark, A. Askill, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.
- [32] R. Sousa-Silva. Fighting the fake: A forensic linguistic analysis to fake news detection. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 35(6):2409–2433, 2022.
- [33] M. A. Spalenza, E. de Oliveira, L. Lusquino-Filho, P. M. Lima, and F. M. França. Using ner+ ml to automatically detect fake news. In *International Conference on Intelligent Systems Design and Applications*, pages 1176–1187. Springer, 2020.
- [34] S. Sridhar and S. Sanagavarapu. Fake news detection and analysis using multitask learning with bilstm capsule model. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 905–911, 2021.
- [35] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*, 2021.
- [36] C.-O. Truică and E.-S. Apostol. Misrobæarta: transformers versus misinformation. *Mathematics*, 10(4):569, 2022.
- [37] C.-O. Truică and E.-S. Apostol. It’s all in the embedding! fake news detection using document embeddings. *Mathematics*, 11(3):508, 2023.
- [38] A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [39] S. Vincent, R. Sumner, A. Dowek, C. Blundell, E. Preston, C. Bayliss, C. Oakley, and C. Scarton. Personalised language modelling of screen characters using rich metadata annotations. *arXiv preprint arXiv:2303.16618*, 2023.
- [40] H. E. Wynne and Z. Z. Wint. Content based fake news detection using n-gram models. In *Proceedings of the 21st international conference on information integration and web-based applications & services*, pages 669–673, 2019.
- [41] M. Yang, C. Park, J. Kang, D. Lee, D. Choi, and J. Han. Fighting against fake news on newly-emerging crisis: A case study of covid-19. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 718–721, 2024.
- [42] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.
- [43] Y. Zhang, Z. Shen, Y. Dong, K. Wang, and J. Han. Match: Metadata-aware text classification in a large hierarchy. In *Proceedings of the Web Conference 2021*, pages 3246–3257, 2021.
- [44] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2):1–25, 2020.
- [45] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani. Recovery: A multimodal repository for covid-19 news credibility research. In *CIKM-ACM*, pages 3205–3212, 2020.
- [46] X. Zhou, J. Wu, and R. Zafarani. : Similarity-aware multi-modal fake news detection. In *PAKDD*, pages 354–367. Springer, 2020.
- [47] X. Zhou and R. Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.



(a) Representation of information density by average normalized Shannon entropy on the Fake_And_Real_News dataset. The title yields the lowest score of 74.44. NER words exhibit a relatively lower Shannon score, capturing 11% of the information density of the full text. Similarly, both keywords and POS words, when sampled at the default 10%, demonstrate significantly lower scores compared to the full text



(b) Representation of information density by the average count of tokens on the Fake_And_Real_News dataset. The title and NER words maintain the lowest token counts among all input types. Across all inputs, the token counts remain significantly lower than those of the full text, with the highest reaching about 29% of the full text.

Figure 6: Representation of information density by average normalized Shannon entropy (a) and the average count of tokens (b) on the Fake_And_Real_News dataset.

APPENDIX

A Dataset

A.1 ReCOVery Dataset

The ReCOVery dataset is a repository that has been built to make it easier to conduct research on countering COVID-19-related information. After conducting a thorough search and investigation of around 2,000 news publishers, 60 were found to have extremely high or low levels of credibility by the authors of the dataset. The repository includes 2,029 news pieces about the coronavirus that were published between January and May 2020, as well as 140,820 tweets that show how these stories were shared on the Twitter social network. ReCOVery has a wide collection of news articles, social media posts, images, videos, and audio recordings pertaining to COVID-19. The dataset covers various themes and topics related to the pandemic, including public health guidance, government policies, scientific research, and societal impacts. Additionally, ReCOVery includes metadata such as publication dates, image, country, sources, and con-

textual information [45].

Descriptions of the variables: **label**: news label (1 = real, 0 = fake); **text**: content of the news; **title** and **author**.

A.2 Fake And Real News Dataset

The Fake_And_Real_News Dataset comprises two distinct components sourced through different methods. The first part consists of 13,000 articles labeled as "fake news," obtained from a dataset released by Kaggle during the 2016 election cycle. For the second part, To gather these, the author turned to All Sides, a platform hosting news and opinion pieces spanning the political spectrum. With articles categorized by topic and political leaning, All Sides facilitated web scraping from diverse media outlets, including prominent names like the New York Times, WSJ, Bloomberg, NPR, and the Guardian. Finally, a total of 5,279 real news articles published in 2015 or 2016 were successfully scraped. The dataset was meticulously constructed to ensure balance, with an equal number of fake and real articles, resulting in a null accuracy of 50%. The finalized dataset encompasses 10,558 articles, complete with headlines, full-body text, and corresponding labels denoting their authenticity (real or fake) [20]. The dataset is publicly available in the provided GitHub repository.

Descriptions of the variables: **label**: news label; **text**: content of the news; and **title**.

B Information Density Comparisons

This section presents the remaining graphs for the representation of information density and average token count respectively for the Fake_And_Real_News Dataset. Notably, the difference between full-text and keywords is significantly lower in the average token graphs compared to the normalized Shannon entropy graph. Generally, keyword subsequences naturally prioritize words carrying the most information, as reflected by the 10% keyword category achieving the highest information density per percentage of full text across all categories. A consistent trend across all datasets is that, at the default 10% threshold, NER words yield the lowest scores (apart from the title), followed by POS words. Furthermore, when selecting only 25% of the keywords from the full text, the information density, measured by normalized Shannon entropy, is reduced by nearly half. Despite this reduction, we still achieve a comparable level of accuracy.