Albanian Fake News Detection

ERCAN CANHASI, REXHEP SHIJAKU, and ERBLIN BERISHA, University of Prizren

Recent years have witnessed the vast increase of the phenomenon known as the fake news. Among the main reasons for this increase are the continuous growth of internet and social media usage and the real-time information dissemination opportunity offered by them. Deceiving, misleading content, such as the fake news, especially the type made by and for social media users, is becoming eminently hazardous. Hence, the fake news detection problem has become an important research topic. Despite the recent advances in fake news detection, the lack of fake news corpora for the under-resourced languages is compromising the development and the evaluation of existing approaches in these languages. To fill this huge gap, in this article, we investigate the issue of fake news detection for the Albanian language. In it, we present a new public dataset of labeled true and fake news in Albanian and perform an extensive analysis of machine learning methods for fake news detection. We performed a comprehensive feature engineering and feature selection experiments. In doing so, we explored the Albanian language-related feature categories such as the lexical, syntactic, lying-detection, and psycho-linguistic features. Each article was also modeled in four different ways: with the traditional bag-of-words (BoW) and with three distributed text representations using the state-of-the-art Word2Vec, FastText, and BERT methods. Additionally, we investigated the best combination of features and various types of classification methods. The conducted experiments and obtained results from evaluations are finally used to draw some conclusions. They shed light on the potentiality of the methods and the challenges that the Albanian fake news detection presents.

CCS Concepts: • **Computing methodologies → Information extraction;**

Additional Key Words and Phrases: Fake news, text categorization, natural language processing, machine learning, corpus construction

ACM Reference format:
Ercan Canhasi, Rexhep Shijaku, and Erblin Berisha. 2022. Albanian Fake News Detection. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 21, 5, Article 86 (November 2022), 24 pages.
https://doi.org/10.1145/3487288

1 INTRODUCTION

Social media is a fast data generating and disseminating platform where the huge number of users are interacting with massive number of items creating enormous volume of data. However, contrary to the traditional news sources such as orthodox printed newspapers or modern reliable online news portals, the credibility of contents circulating on social media platforms is questionable due to absence of rigorous editorial processes. On social media the deceptive content is not just easily shared but the readers of such media are much more prone to believe misinformation

Authors' address: E. Canhasi, R. Shijaku, and E. Berisha, Universty of Prizren, P.O. Box 1212, Prizren, R. Kosova, 2000; emails: ercan.canhasi@uni-prizren.com, {rexhepshijaku, berishaerblin}@gmail.com.

2375-4699/2022/11-ART86 $15.00
https://doi.org/10.1145/3487288

due to the lack of particular knowledge related to the topic or simply because of lack of interest for further questioning of a truthfulness of such a material [7]. The latter is even more evident on the web and with the social media users coming from the developing countries such as the Albanian-speaking ones.

Albanian is an Indo-European language that constitutes a subgroup of its own. It is on the same level as the Hellenic, Romance, Slavic, and Germanic subgroups. The language is characterized by a diverse vocabulary with many loan words. It has a writing system based on the Latin alphabet and is a morphologically rich language.

The articles written in Albanian, originally published on Albanian news portals and later posted on Facebook, are the primary focus of this work. Tackling the problem of detecting the fake news in Albanian requires a corpus of fake news. Hence, we first compiled the reference Albanian fake news corpus (Alb-Fake-News-Corpus) and later we trained various machine learning models using it.

The nature of the Albanian fake news, besides the other features, mainly shows the properties of a deceptive content known as a clickbait. This refers to specific content, usually news story, whose main goal is to coax its readers into visiting the offered link [44]. Most Albanian fake news is initially distributed over social media like Facebook and later but not so often finds its way onto mainstream web portals. Most of these fake materials that are posted on various Facebook pages are originally published on a very few common web news portals. Hence, identifying the initial sources of unreliable news can greatly simplify the task of creating the first fake news corpus in Albanian. Section 3 dives into details and describes the complete process behind the corpus creation.

The Albanian fake news stories that are initially seeded over Facebook pages share key linguistic characteristics that are carefully extracted and later used in machine learning stage.

The results of the Albanian fake news identification study that documents the performance of a fake news classifier are presented and discussed in this article.

1.1 Key Contributions and Organization of the Article

Despite the recent advances in fake news detection, the lack of available corpora especially of the one for under-resourced languages are compromising the development and evaluation of the existing approaches on these languages. To fill this huge gap, in this article, we investigate the issue of fake news detection in Albanian language. Inspired by previous initiatives for other languages, to the best of our knowledge, we introduce the first fake news reference corpus in Albanian. This corpus is composed of aligned true and fake news, which we processed to uncover some of their linguistic characteristics. We decided to make our corpus[1] publicly available to facilitate further research in already very sparse research field in Albanian. Later, we report experiments with supervised learning (employing various classification methods) on different sets of features (engineered features and modern text representations). We report excellent classification results alongside with proper answers for the following important research questions:

- Q1: What is the best strategy for compiling a novel fake news corpus for a low-resource language such as Albanian?
- Q2: What is the best and minimal set of features for successful Albanian fake news classification?
- Q3: What are the best current methods for automatic detection of Albanian fake news?
- Q4: Can the size of the texts influence the results of the classification?
- Q5: How can news sources and the news writing styles affect the classification results?

[1] github.com/rexshijaku/Alb-Fake-News-Corpus.

The remainder of this article is organized as follows: The next section presents the main related work in the area. Section 3 details the process of compiling the first ever Albanian fake news corpus (Alb-Fake-News-Corpus), where we also present our answer to the question Q1. Sections 4 and 5 report our experiments and the obtained results. In them, we also provide the answers to the rest of the research questions. Conclusions and guidelines for future work are presented in Section 6.

2 RELATED WORK

In this section the related work is summarized from three different aspects: (1) general fake news detection methodologies, (2) fake news datasets, and (3) feature extraction and selection.

2.1 General Fake News Detection Methodologies

The fake news detection problem has drawn a noticeable recognition in various research fields. A huge variety of methods for fake news detection in different domains and for various languages are proposed. Although there are very successful surveys of the different approaches to automatic detection of fake news and rumors that were proposed in the recent literature [3, 33, 65], in this section, we attempt to summarize the most relevant ones.

The composite method of linguistic cue approaches and network analysis approaches described in Reference [14] is among the first works of its kind. In this work, we presented a method for automatic deceptive information detection in online news sources. Our method employs the well-known vector space model [48]. Another similar methodology is used by Dadgar et al. [16] where they classify news into various groups using TF-IDF and SVM.

Using the real-word dataset to test their hypothesis, Jin et al. have employed the idea of conflicting viewpoints in social media to verify the news [28]. In a different way, using the correlation of publisher bias, news stance, and related user interactions as a foundation for their multi-relational fake news detection approach, authors in Reference [31] report competent results.

In Reference [52] authors approached the problem from a new angle by proposing an algorithm that takes into account the trust of the users. Originally, using the publisher and social emotion and merging them into an emotion-based fake news detection method is what Guo et al. proposed in Reference [24].

Buntain and Golbeck have proposed an automated system to detect fake news in popular Twitter threads [9]. They have applied this method to three publicly available datasets. In another work, the task-generic features have applied to tackle the detection of fake news [38]. Using the crowd signal for the problem and employing it in novel detective algorithm that performs Bayesian inference and jointly learns flagging accuracy of users over time is what Reference [57] contributed to the field. Authors in Reference [69] propose a **Similarity-Aware FakE (SAFE)** news detection method that investigates multi-modal (textual and visual) information of news articles.

Following the recent years' deep learning trends, Long et al. [31] have employed a novel algorithm based on attention-based long-short memory network for fake news detection problem. The performance of the method is tested on benchmark fake news detection datasets. In Reference [64] a novel gated graph neural network based on a set of explicit and latent features extracted from the textual information is presented. By doing so, a deep diffusive network model able to learn the representations of news articles, creators, and subjects is created and used in automatic fake news detection.

The logistic regression and Boolean crowdsourcing algorithm are two classification approaches used in Reference [54] to tackle the fake news detection. In another work, the data-mining algorithm for the problem of fake news detection, evaluation metrics, and datasets has been extensively presented [51]. In Reference [23] authors, as we did in this work, used a broad range of popular classification algorithms such as: Random Forests, SVM, Bounded Decision Trees, Stochastic Gradient

Descent, Gradient Boosting to train the fake news detection models. They report the best classification results when the Stochastic Gradient Descent method is used. Perez-Rosas et al. have formulated the fake news detection as a novel automated algorithm [41] trained on the combination of lexical, syntactic, and semantic information. Xinyu Zhou et al. recently proposed a theory driven model for fake news detection [68]. The method investigates news content at various levels: lexicon-level, syntax-level, semantic-level, and discourse-level. One of the rare semi-supervised formulation of fake news detection is one proposed by Guacho et al. in Reference [22]. Monti et al. have adapted the geometric deep learning-based model to detect fake news [36]. In another recent work, methodically similar to ours, authors for text modeling use standard text modeling known as bag-of-words and for classification they used 23 supervised artificial intelligence algorithms [39].

2.2 Feature Extraction and Selection

Feature extraction and common linguistic features identification have been extensively researched in linguistics and NLP fields. Even the first studies on computer-aided communication [10, 67] showed promising results. Nevertheless, a bit more complicated and focused works [8, 43] have been able to identify minimal set of linguistic features that correlates with deception. The cues such as the use of self reference, negative and swear words are the most common ones.

Enriching the linguistic feature set with signals from other sources such as credibility [35], psycho-linguistics [62], specific phrases [35] are just a few more recent ways of treating the problem of feature modeling. These and many other linguistic features have been successfully extracted and used in a huge number of methods for automatic fake news and rumor detection [53, 60, 66].

The most common categorization of features used in automatic fake news detection is as follows: lexical, syntactic, and semantic linguistic features.

Lexical features are usually used to model the character- and word-level signals. Besides using the common statistics such as the average length of words, percentage of various chars usage, and so on, on another level of abstraction the most straightforward approach is to use the most important content words or phrases (e.g., bi-grams and tri-grams) as features [12, 19, 47].

Syntactic features, which aim at modeling the contextual and functional signals such as the n-gram POS frequencies, or count of stop-words, have been employed, for instance, in References [46, 70]. Moreover, sentence complexity has been used as a signal for the reliability of the information [8].

Finally, Reference [46] has exploited features based on the novelty of words found in social media posts. Semantic features are often extracted by means of advanced NLP techniques. For example, sentiment analysis and opinion-mining approaches adopt features based on the opinions and emotions expressed in the text [11, 45].

Content features have been extensively studied and implemented for fake news detection. Clearly, the content plays a key role in discovering potential deception. However, such features present some downsides. First, textual cues to deception may achieve limited generalization capability in a real word application system. Second, they may lose in descriptiveness, as fake news is becoming more similar to proper news as far as the writing style [3]. Last, especially concerning rumors on social media, given the relative brevity of texts, other characteristics may prove to be more effective.

2.3 Fake News Corpora

Given the immaturity of the fake news detection research field, many efforts have been made to propose various novel fake news corpora. To the best of our knowledge, all of the corpora are mono-lingual, most of them are English, many are quite small and contain no more than a thousand labeled news. The detailed summary of the corpora in the literature is given in Table 1.

Table 1. Fake News Corpora in Different Languages

Reference work	Language	Number of true entities	Number of fake entities	Aligned	Specific time interval?	Metadata availability	Construction mode
Our corpus	Albanian	2,000	2,000	Yes	Yes (2020)	Link, Date, # of links, comments	Almost automatic
Renato M. Silva et al. [53]	Brazilian Portuguese	3,600	3,600	Yes	Yes (2016–2018)	Link, Date # of links	Semi-automatic
Verhoeven and Daelemans [59]	Dutch	270	270	No	Yes (2012–2013)	Demographics	Manual
Fornaciari and Poesio [21]	Italian	1,202	945	No	No	Timestamp	Manual
Zhang et al. [63]	Chinese	131	187	No	Yes (2001–2008)	None	Semi-automatic

The methodology used in corpora creation also varies from fully automatic [50] on one end to the completely manual [21, 42] on the other end of the spectrum. Our efforts in corpus creation were towards the utilization of the highly automatic methodology.

Despite their high costs, the manual exploring of the website for potential fake news extraction is the most commonly used method in many corpora creation efforts. Another very popular approach to corpora creation was the usage of the crowdsourcing [21] to collect the texts. Utilizing the Amazon Mechanical Turk or proprietary online platforms came with the cost of having to deal with issues of reliability of the collected data. Very few of the corpora report to contain the aligned instances of the fake to true news. The absence of meta-data for the corpora in literature is yet another very common limitation.

Next, we summarized some recent works on fake news corpus creation. One of the first works of this kind is the one from Pérez-Rosas and Mihalcea [42]. In it, authors proposed three datasets with 100 deceptive and 100 truthful sentences. Two datasets of satirical and true news for the specific set of domains such as civics, science, business, and entertainment, totaling 240 texts were presented by Rubin et al. in Reference [47].

Thorne et al. [55] have produced fake statements by purposely modifying Wikipedia sentences and then providing evidence for or against such claim in Wikipedia articles. The authors in Reference [44] by mainly focusing on Facebook publishers have collected clickbait data containing URLs from posts produced by nine verified Facebook publishers (three mainstream publishers and six hyperpartisan publishers). Later on, each post has been manually fact checked and annotated.

Two widely known English corpora are Emergent [20] and LIAR [61]. In Reference [20] authors have introduced the task of fact checking and produced a dataset of statements, containing both a veracity assessment and an analysis of the reason behind such assessment. Similarly, LIAR [61] contains short political statements, obtained through the website PolitiFact.com. Each statement is annotated with the author, the context, a veracity label, and a justification for such label.

Finally, the most comprehensive dataset of statements in terms of meta-data enrichment is the one of Shu et al. [50]. Authors have built a system for fake news identification, and besides providing an original dataset containing information about the content (textual and visual), they yield meta-data such as information about social context (i.e., users, network information, etc.) and characteristics of the spread evolution.

There are also some available datasets in Dutch [59], Chinese [63], Italian [21], and Brazilian Portuguese [53].

Recent years have witnessed a widespread increase of scientific community interest in preparing the annotated collections of fake news. Unfortunately, this interest is mainly limited to English

language and there are still a very few public corpora, especially for low-resource languages. The overview of different fake news corpora with non-English content is given in Table 1. The lack of such corpora limits the prospering of fake news detection for low-resource languages, particularly limits the experimentation with various feature extraction/selection and different machine learning workflows.

To fill this huge gap, we describe our achievements in building the novel reference corpus of aligned Albanian true and fake news. We strongly believe the corpus will seed initial research efforts in fake news/data identification, particularly for the Albanian language. Albanian, which is the native language of the authors of this article, is a resource-poor language in terms of **Natural Language Processing (NLP)**. Many previous works in this regard were mostly experimental and there are few generally available NLP resources that deal with the Albanian language. To the best of our knowledge, this is the first corpus of such nature for the Albanian language. Independently from most of the corpus building practiced cited in this section, we compiled our corpus with a hybrid methodology: Most of the steps are algorithmic and have very few manual steps, resulting in a weakly supervised strategy. We purposely preferred some manual steps to promote reliability. The complete workflow used in building the corpus is described in the next section.

3 THE ALB-FAKE-NEWS-CORPUS

Planning, designing, and finally creating a novel—in our case, pioneering—corpus with huge potential to be used as a reference point to other researches is a vastly demanding task. Some general language-agnostic, domain, and task-independent best practices and golden rules that anyone working on corpus creation should take into account are given in seminal work by Reference [27].

The additional more-deceptive content-oriented corpus compiling guidelines are proposed in References [47, 53]. The next few paragraphs elaborate the extensive list of these rules, which are also carefully taken into account while constructing the first ever corpus of fake news in Albanian.

According to cited works, one of the very first and hardest requirement to be met by a successfully compiled deceptive news corpus is the proper alignment of fake news and their corresponding deceptive versions. By doing so, the corpus designers will later aid the machine learning methods in finding the patterns among balanced sets of positive and negative instances. Another set of guidelines is about general feature balancing, mainly for avoiding the bias in learning. Few of the features from this category are: news should be of similar lengths, they should match the same time interval, and articles should have same meta-data information associated.

We have followed the above steps and directions to create our corpus for the final purpose of fake news classification. For such purpose, our corpus is composed of aligned true and fake news written in Albanian. Alignment in this context does not mean pairing each fake article with the reciprocal true one, which exactly denies the fake article. Undoubtedly, that kind of arrangement is preferred, but, since it is highly resource-expensive, we choose an alternative and very common approach. The alignment procedure we used in this work is based on topic similarity; which simply means that a fake news story and its corresponding true story are matched if they share the same or similar topic. For topic modeling, we experimented with some well-known models such as TF-IDF/VSM [49] and LSI [18].

For the alignment, we mean that, for each fake news, we collected corresponding true news, which, if not explicitly denying the fake news, is topically related (which is the most common case).

Let us suppose one is given a hypothetical article A described by the title: “Can you believe that Messi was close to death?” where the content in a highly general way expresses some non-verifiable information on Messi, being close to his friend named Death. Finding the news

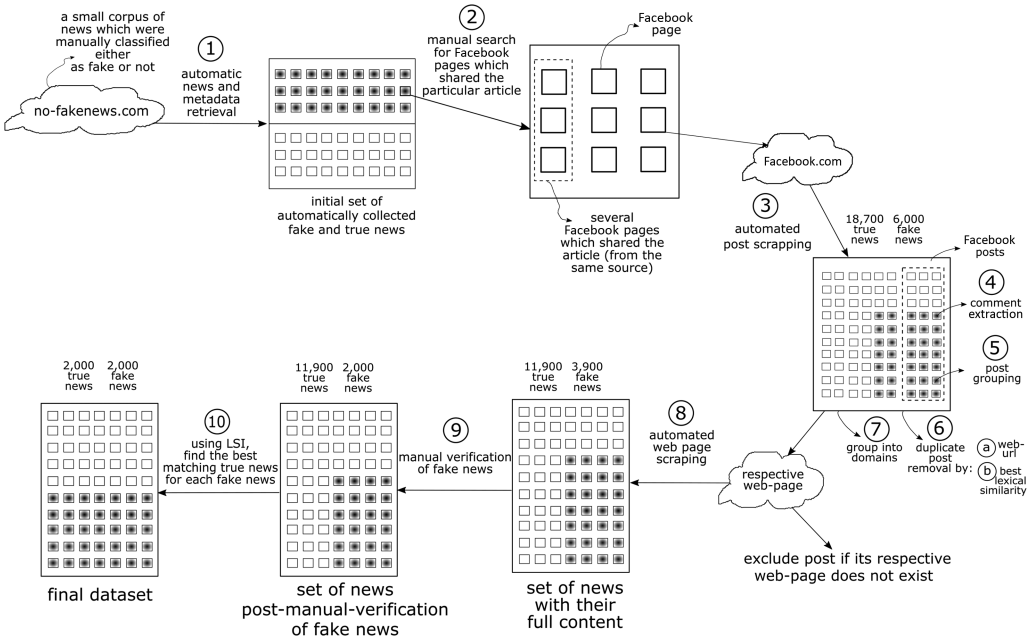

Fig. 1. Process of building the Alb-Fake-News-Corpus.

article(s) explicitly denying the article *A* would be so hard that it was out of our abilities and our sparse resources. Therefore, we adopted most widely practiced method of finding topically related true news. Hence, as a corresponding true article of the article *A*, we extracted another topically related/similar news *B* that is true, since it is obtained from highly reliable news source. The news article *B* would potentially report on "Messi being close to next championship!" since the topical similarity would require such similarity. To find the appropriate texts to compose the corpus was a challenging task. We searched the web for the available fake news, which were manually checked to guarantee that they had deceptive content. The manual verification was important to ensure the data quality and, therefore, the reliability of the resulting corpus. The selected fake news were then used in a semi-automatic process to look for their corresponding true versions on the web. The availability of the deceptive news and their corresponding true versions is very important for machine learning tasks (which require positive and negative instances for the learning success) and linguistic investigations, which look for textual patterns and their contexts of usage for language description.

Based on the authors' previous experiences with the Albanian web and social media, this section summarizes and argues the initial educated guesses regarding the nature of content, the way in which articles are created, published, shared, and interacted with. By doing so, we establish the foundation for compiling the first Albanian fake news dataset.

Locating the fake news is a much more challenging task than finding true news [3, 33]. We postulate that the vast majority of fake materials that are posted on various Facebook pages are originally published on very few common web news portals. Hence, identifying the initial sources of unreliable news can greatly simplify the task of creating the first fake news corpus in Albanian.

Figure 1 outlines the sequence of steps that were undertaken in corpus creation. The starting point of this process was a set of fake and true news that was manually and meticulously classified

to their genuine belonging categories by a site that was investigating and reporting fake news to the public (no-fakenews.com,[2] with more than 100 articles).

After automatically fetching this set of news (1), we initially formulated a bag of distinct domains that were seemingly sharing deceptive content and those that were producing veritable news. In this bag, we included some prestigious news agencies that have not been a part of the initial corpus. Subsequently, we manually hit Facebook search querying all obtained news by using their web urls, titles, and web domain names (2). As a result, we confirmed our initial hypothesis and found that fake articles from the same domains were shared by multiple Facebook pages, also a single Facebook page was sharing fake articles from various domains. We observed that reliable content creators usually have a single page on this social media (often official and verified, with a lot of contact information as well as consistent content) in which they share their news from a single web-based source. However, dishonest authors use multiple Facebook accounts to reach their audience. These accounts do not provide any contact information, neither are verified nor have a consistent name with the web domains from where they share their news. Diving deeper allowed us to find evidence that most of these accounts were previously creating content on totally different topics, however, now, based on our inspections, we assume that when they reached a certain point, apparently they changed ownership and were acquired by their current owners. These analyses served as an instrument by which we made a clear distinction between reliable and deceptive sources. At the end of this process, dozens of Facebook pages were produced. The ratio between the Facebook pages and their corresponding domains, was, respectively, for deceptive sources 1:2 and for reliable sources 1:1. Following this, we picked 25 Facebook pages and scraped their posts (3) that were shared over the recent six-month period.

Afterwards, we merged all posts into a single collection (24,700 in total), in this stage, we applied: (4) comment extraction, (5) post grouping, (6) duplicate post removal, and (7) grouping into domains. Initial grouping (5) was applied with a purpose to eliminate same posts that were referring to the same web-resource or article, and to make an aggregation of responses (such as reactions and comments) for the same posts. Those that were likely to be same and to point to different sources were also eliminated (6) using cosine similarity measure. Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space [49]. In our case, these vectors represent news titles.

The excluded duplicates comprised a considerable part of the set, more concretely, we spot that fake news were strongly predisposed to be reproduced, more than one-third of fake news that we gathered in the previous step were removed in this phase. Needless to say, after duplicate removal, we always keep exactly one copy of the identified copies.

In the last step of this stage (7) posts were grouped into their domains. Once we constructed a single page scraper (which was capable of using domain specific XPaths—in other words, it was able to extract text from documents having different HTML structures), we returned back to the web to automatically retrieve the complement part of each remained article (8). At this point, we noticed that a lot of posts from deceptive Facebook pages were missing their corresponding web page, hence, it was obvious to conclude that these articles were used temporarily. For the reason that these articles lack the necessary information for the subsequent steps, we excluded them from the corpus.

Since the true news were fetched from the most reliable online news portals, we decided to skip the manual verification of them. On the contrary, we manually inspected a set of 3,900 unique articles that were gathered from deceptive sources, and after this process 2,000 remained as fake (9).

[2] no-fakenews.com/.

Table 2. Basic Statistics of the Alb-Fake-News-Corpus

Description	True news	Fake news
Average number of tokens	**275.6**	114.6
Average number of unique words	**146.5**	74.1
Average length of words (in characters)	**5.2**	4.9
Average number of sentences	**5.1**	2.8
Average length of sentences (in words)	**54.9**	40.6
Average number of verbs (normalized)	58.2	**65.3**
Average number of nouns (normalized)	**66.8**	65.8
Average number of adjectives (normalized)	**15.2**	14.7
Average number of adverbs (normalized)	7.2	**7.8**
Average number of pronouns (normalized)	7.1	**9.5**

Given the absence of any institutional/financial support, authors decided to do the manual annotation themselves instead relying on any voluntary help, mainly because of the sensitivity of the annotation process. Each news article was voted by at least two out of three annotators. If agreement on class was reached, then the news article was included into final set of fake news. Each news article, especially the potentially fake ones, where verified in as many ways as possible.

During this examination, we gained a more rigorous understanding about the fake news posted by Albanian portal. A distinguishing feature for most fake posts is a curiosity gap they intend to create.

This is planned to be achieved by introducing sensitive topics that are known and relevant for the community, and it is often done by using distorted, exaggerated, or overgeneralized headlines that are intriguing and lead to sites where the promised answers are not fully addressed. In short, generally these posts overpromise and the resources where the user is led under-deliver.

All in all, the nature of Albanian fake news, besides the other features, mainly shows the properties of a deceptive content known as a clickbait. This refers to "content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web or Facebook page."

Finally, we found the best matching true news for our group of fake news (10). For news alignment the LDA topic modeling based approach was used. The same approach is also used to treat the problem of news similarity calculation.

In summary, starting from 100 news, we collected nearly 24,700 news that after processing were reduced to 4,000 in total, equally balanced to 2,000 of fake and 2,000 true news. In this way, we compiled a dataset as a benchmark of our further analysis and the first of its kind in the Albanian language.

We show in Table 2 summary statistics of some general NLP features. It is obvious that the true news are much larger than fake news on many dimensions, such as the number of tokens, words, terms, and characters. This can greatly compromise classification results, as already pointed out in previous work [48, 53]. Table 3 present two aligned examples of fake and true news from the proposed corpus.

We can see in Table 2 that only on number of verbs, adverbs, and pronouns fake news are in the lead, while on the other POS types and rest of statistics the real news contained higher number of occurrences. Although there is previous work on POS tagging for Albanian language [29, 30], for the POS tagging, we used an internal IP protected model of Gjirafa, Inc. The model was developed under the funding of Gjirafa, Inc., and as such it is a highly accurate (0.97 F1 score) deep neural network-based, very efficient, and efficacious model. It consists of 58 carefully defined tags: NLE, ADJP, VAUXP1, VAUXS2, PFS, PRTA, VAUXP2, PRTQ, INTJ, VAUXS1, SFS, VIMP, PREF, PINT,

VP2, ABBR, ADVQ, NUMD, VMOD, VP1, VAUXP3, VS2, ADJMP, PRTR, PCL2, PRTC, APST, VAUXS3, VS1, PPOS, ADJFP, PPER, PRT, PRTV, PDEM, NUMC, PRTN, PIND, PREL, PRTS, ADJMS, VPCP, NMP, ADJFS, VP3, CNJS, PCL1, NFP, NP, ADV, CNJC, PNTE, VS3, NMS, PNTS, PRP, NFS, ART.

It is important to highlight that the Alb-Fake-News-Corpus is publicly available.[3]

4 EXPERIMENTS

In this section, we describe the conducted experiments and discuss the obtained results. We carefully designed a broad set of experiments to answer the research questions initially defined at the end of Section 1. In doing so, we performed experiments using:

(1) The extensive types of features obtained via the various feature engineering and selection methods;
(2) The set of traditional and few very recent classification methods;
(3) The various versions of our original fake news corpus, such as truncated and aligned versions.

To show the significance and usability of the compiled Albanian fake news corpus, we utilized the corpus in automatic fake news detection. To detect fake content, we needed to identify a set of discriminating features that distinguish deceptive writing from regular writing. Hence, in Section 4.1, we first present the methodology used for feature extraction and selection. After determining these features, in Section 4.2, we describe the outputs of experiments with the supervised learning models trained to classify the Albanian fake news.

4.1 Feature Extraction and Selection

Fake news detection on traditional news media mainly relies on news content. The availability of additional information in social media context can be used to further improve the fake news detection. Since there are no any previous researches on Albanian fake news detection upon which we could build up, in this work, we also aspire to experiment with a very broad set of various categories and types of relevant features. The rest of this subsection present the details on extraction and representation of useful features from the Albanian text content.

Social media news content features represent the meta-information related to a news stories. A list of such set used in this work includes:

- Web source: the original web portal considered as the first publisher of the news article;
- Facebook page source: the Facebook page that has lately been used to publish the article first disseminated on a web portal from the previous point;
- Facebook news title: the short title text that summarizes the news story; in the Albanian fake news context found to be remarkably important source of signals;
- Original news title: the short title text that aims to catch the attention of readers and describes the main topic of the article. Highly valuable source of fake news signals such as clickbait title words;
- Body Text: the main text that elaborates the details of the news story; in the Albanian fake news detection used as the main source of features engineering for classification training.

From these raw content sources, various kinds of features can be derived. The features we derived and used in this work are mostly linguistic, psycholinguistic, and content-based. Each of which are described in more detail below.

[3] github.com/rexshijaku/Alb-Fake-News-Corpus.

Table 3. True and Fake News: Examples from the Corpus

Fake	True
Title: Kjo qe beri Kosovari ne Zvicer do iu habise; Body: Një qytetar i ka befa suar mërgimtarët në Zvicër, pasi që ai ka shkuar atje nga Kosova me veturën e tij Golf 4, Shtype reklamen dhe shiko videon e ketij shkrimiKëtë vit mërgimtarët janë të paktë në numër që po vijnë në Kosovë, nga shtetet perëndimore. Pandemia koronavirus ka bërë që ata të mos vijnë, pasi që po detyrohen të karantinohen dy javë, pas kthimit nga pushimet në Kosovë.E pasi që mërgimtarët nuk po vijnë, një qytetar nga Kosova i ka be fasuar ata duke shkuar në Zvicër me veturën e tij Golf 4. Një mërgimtarë i cili e ka parë rrugëve të Zvicrës kosovarin me tabela KS, ka publikuar një video në grupin 'Marakli t'kerreve'.Shtype reklamen dhe shiko videon e ketij shkrimi	Title: Kosovari shkon me Golf 4 në Zvicër, filmohet nga shqiptarët në Zurich Body: Një qytetar i ka befasuar mërgimtarët nëZvicër, pasi që ai ka shkuar atje nga Kosova me veturën e tij Golf 4.Këtë vit mërgimtarët janë të paktë në numër qëpo vijnë në Kosovë, nga shtetet perëndimore. Pandemia koronavirus ka bërë qëata të mos vijnë, pasi që po detyrohen të karantinohen dy javë, pas kthimit ngapushimet në Kosovë.E pasi që mërgimtarët nuk po vijnë, njëqytetar nga Kosova i ka befasuar ata duke shkuar në Zvicër me veturën e tijGolf 4.Një mërgimtarë i cili e ka parë rrugëve tëZvicrës kosovarin me tabela KS, ka publikuar një video në grupin 'Maraklit'kerreve'.Video besohet se është bërë në kantonin eZurichit.Raste si kjo janë të pakta, pasi që mungesa evizave ka bërë që kosovarët të mos mund të shkojnë lirshëm në shtetetperëndimore.Ndërsa e bukura e rastit të djeshëm, ishte se kosovari ka shkuar me veturë Golf 4, veturë jo fort komode për rrugë të largëta pasi që edhe është e vjetër si model.
Title: U Përfol Për Lidhje Me Capital T Dhe Me Politikanin Nga Kosova, Zbulohet E Vërteta E Adrola Dushit Body: U Përfol Për Lidhje Me Capital T Dhe Me Politikanin Nga Kosova, Zbulohet E Vërteta E Adrola Dushit;U përfol për lidhje me Capital T dhe me politikanin nga Kosova, zbulohet e vërteta e Adrola DushitSht1pe v1deon e me poshtme p3r vazhd1min e plot3 t'lajmit:JA DHE VIDEO:Një foto e Adrola Dushit me Capital T bëri bujë në rrjetet sociale, pasi dyshja dukeshin shumë të afërt me njëri-tjetrin. Shoqëruar me një emoji zemre nga Capital T, ky postim u pëlqye nga mbi 100 mijë vetë për të mos folur për mijëra komentet që kanë shpërthyer në profilin e reperit ja dhe v1deo sht1pe rek1am dhe sh1ko v1deon	Title: Adrola Dushi ka treguar rreth lidhjes me politikanin nga Kosova Body: Një imazh e Adrola Dushit me Capital T bëri bu jë dje në rrjetet sociale, pasi dyshja dukes hin shumë të afërt me njëri-tjetrin. Shoqëruar me një emoji zemre nga Capital T, ky pos tim u pëlqye nga mbi 100 mijë vetë për të mos folur për mijëra komentet që kanë shpërt.hyer në profilin e reperit.Ky imazh erdhi menjëherë pas fjalëve se modelja po shijon një roma ncë me kandidatin për deputet në Kuvendin e Kosovës, Albin Gashi, i cili është djali i avokatit të njohur Tomë Gashi.E pyetur për këtë, Adrola Dushi tha ek skluz.ivisht për "iconstyle.al", se nuk është i vërtetë lajmi i roma ncës me djalin nga Kosova.Modelja bukuroshe mohoi lid hjen me Gashin, për të cilën mediat shkruanin se është konsoliduar dhe se dys.hja po shkon drejt martesës.Dhe tani që u siguruam që Adrola nuk është në një lid hje me Albin Gashin, le të shijojmë imazhin e ëmbël me Capital T.Një pjesë e mirë e fansave dëshirojnë t'i shohin bashkë, por ka shumë nga ata që mendojnë se janë thjesht shokë!Nuk e dimë nëse po kurdisin ndonjë projekt ata të dy, por marrëd.hënia e tyre shoqërore është parë edhe në ditëlindjen e modeles, ku Capital T ishte i pranishëm!E pyetur për Capital T, Adrola zgjodhi të mos flasë!Dys hja ia arriti qëllimit dje që të mbushë kryetitujt e mediave, por për më tepër informacion, me sa duket do t'ia lëmë kohës ta tregojë!

Table 4. Overview of Features Used

Category - Subcategory	Number of Features	Description (number of features)
Lexical features - Character based	120	Total # of characters (1), percentage of digits (1), percentage of letters (1), percentage of uppercase letters (1), frequency of character unigram (36), most common char bigrams (40) and tri-grams. (40)
Lexical features - Digits, special characters, punctuation chars	40	Frequency of digits (0–9) (10), special characters (e.g., %, &) and punctuation (30).
Lexical features - Word	22	Total # of words (1), average # of words per sentence (1), total # of out-of-vocabulary words (1), average # of out-of-vocabulary words (1), most frequent word uni-/bi-/ tri-grams (6*3).
Lexical features—POS related	60	Average number of POS types per sentence (1), POS type to unique words ratio (1), total # of stopwords (1), average # of stopwords per sentence, and most frequent POS uni-/bi-/ tri-grams (32+16+8).
Lexical features—Sentence and article level	4	# of sentences (1), average length of sentences (1), body starts with title (1), cosine similarity of title to body (1).
Lying-detection feature set	15	Vocabulary complexity: total # of syllables (1), average # of syllables per word (1), total # of large words (1), average number of large words per sentence (1), grammatical complexity: number of short sentences (1), number of long sentences(1), flesh-Kincaid grade level (1), number of conjunctions (1); Uncertainty: number of words express certainty (1), number of tentative words (1), modal verbs (1); Specificity and expressiveness: rate of adjectives and adverbs (2), number of affective terms (1); verbal non-immediacy (1).
Psycholinguistic features	6	The # of positive and negative emotions (2); The average number of positive and negative emotions per sentence (2) emotions to number of words ratios (2).
	Total: 267	

Since fake news stories are intentionally created for financial or political gain rather than to report objective claims, they often contain opinionated and inflammatory language, crafted as *clickbait* [12, 44]. Thus, it is reasonable to primarily exploit linguistic features that capture the different writing styles and sensational headlines to detect fake news. The summary of the complete set of features used in this work is given in Table 4.

Lexical/Syntactic features (char, word, POS, and sentence level): These features include clues from many levels aiming to model the author's lexicon-related writing style, particularly the vocabulary and character choices used in article writing. The char level is the largest feature set, consisting of 120 features, producing the densest representation mainly because of the lower level of representation. The char level can be further enriched with the digits, special characters, and punctuation chars producing 160 features in total. Each author organizes sentences differently. Syntactic features represent an author's sentence-level style. These features include frequency

of function words, punctuation, and **parts-of-speech (POS)** tagging. We use a list of Albanian function words that consists of determiners, conjunctions, prepositions, pronouns, auxiliary verbs, modals, qualifiers, and question words. The word-level features among the common signals such as the total and average number of words, include also such a statistic for out-of-vocabulary words. The most frequent n-grams compared to char level were much sparser, therefore, we decreased the number of top features used. The next level of abstraction regarding the lexical features was aimed on POS level. On the highest level, namely, sentence level, among others, we additionally modeled the title to content relations.

Lying-detection feature set: Our feature set includes features that were known to be effective in detecting lying-type deception in computer-mediated communications and typed documents [25]. These features are: (1) Quantity (number of syllables, number of words, number of sentences); (2) Vocabulary Complexity (number of big words, number of syllables per word); (3) Grammatical Complexity (number of short sentences, number of long sentences, Flesh-Kincaid grade level, average number of words per sentence, number of conjunctions); (4) Uncertainty (Number of words express certainty, number of tentative words, modal verbs); (5) Specificity and Expressiveness (rate of adjectives and adverbs, number of affective terms); (6) Verbal Non-immediacy (self-references, number of first-, second-, and third-person pronoun usage). Certainty: following is the partial list of words expressing certainty: absolutisht, i.e., sigurt (absolutely sure), pa dyshim (no doubt), njëqind përqind, i.e., sigurt (hundred percent certain), i.e., bindur se (convinced that), Shanset/Probabiliteti/Mundesitë janë që (Chances/Odds/Options are that), padyshim serioz (Without doubt), Me siguri jo/po (Certainly not/yes), and so on. Tentative words: next is the part of the list of words expressing tentative language: Ndoshta (Perhaps), Sugjero/propozo (Suggest), Trego (Indicate), Beso (Believe), Një numer i (A number of), Mund të ketë qenë (May have been), Varet nga (Depending on), Nuk ka gjasa/shumë pak e mundur (Unlikely), Duhet (Should), and so on.

Psycholinguistic features: Using the linguistic behavior in text-based communication research is long known as a beneficial to deceptive psychological mechanisms. Tools like **Linguistic Inquiry and Word Count (LIWC)** have been widely used in studying various relationships between psychology and linguistics [15]. Since there are not any previous works on Albanian psycholinguistic analysis, we consulted few experts working on similar fields. As a result, we identified 200 archetypal emotions and grouped them in two sets. One as a set of positive and the second containing the negative emotions. A short list of positive emotions includes: Lumtur, Gëzim, Lezet, Pëlqej, Eksituar, Patriotizëm, Pasion, and so on. The negative list includes: Pandjenjë, Apati, Dëshpërim, Pakënaqësi, Ksenofobi, Trullosur, and so on.

4.2 Text Representation Models

Besides the conventional way of representing the training instances as a set of manually crafted features, we additionally represented each fake news in four different more content oriented ways. First with the classic **bag-of-words (BoW)** [49] and then with three state-of-the-art distributed text representations using Word2Vec [34], FastText [2], and BERT [17] techniques. We used the **TF-IDF (term frequency-inverse document frequency)** scoring methods to score the weights of the tokens of each news, which is also known as a BoW representation. The pre-trained words vector as numeric representations are trained using the Word2Vec, FastText, and BERT. When available, we used pre-trained models, otherwise, we trained models with Albanian language documents from a large set of news crawled from the Albanian web in the past two years.

4.3 Preprocessing

In the experiments with the engineered features, we applied the Z-score normalization using information from the training examples. Before generating the feature vectors with BoW, Word2Vec,

(a) True news.

(b) Fake news.

Fig. 2. Word clouds representing the relative frequency of the tokens.

FastText, and BERT all instances of our dataset were converted to lowercase. After that, we tokenize the documents based on whitespaces and punctuation marks. Figure 2 presents word clouds to visually summarize the relative frequency of tokens obtained after the preprocessing. The fake news word cloud (b) consists of sensational expressions, while the true news word cloud (a) contains mostly descriptive and unimpressive terms.

4.4 Classification Methods

We trained the various Albanian fake news classification models mainly using the classic methods and few techniques that are more recent. The full list of methods is:

(1) **Logistic Regression (LR)** [26];
(2) Naive Bayes [32];
(3) **Support vector machines (SVM)** [4];
(4) **Decision trees (DT)** [6];
(5) **Random forest (RF)** [5];
(6) KNN classification [1];
(7) XGBoost [13];

Since the well-known Python library Scikit-learn [40] includes implementation of most of these methods, we used it primarily for those methods but also for many different data science-related tasks such as feature selection, evaluation, and so on.

Considering that this work's goal is not to discuss the parameter and/or hyper-parameter optimization but rather to evaluate the quality of proposed corpus classification results for the most of the methods, we set their parameters to the default values.

4.5 Performance Metrics

To compare the results, we employed the following well-known performance measures for spam and other misleading content [37, 53]:

- **Fake news caught rate (FCR)** or recall: proportion of fake news correctly identified (the higher, the better);
- **Fake news precision rate (FPR):** proportion of news classified as fake and that truly belong to the fake class;
- **Legitimate news blocked rate (LBR)** or false positive rate: proportion of legitimate news incorrectly labeled as fake news (the lower, the better);
- F_m: harmonic average of the FCR and FPR;

Table 5. Results Obtained by Each Method in the Experiments with the Lexical/Linguistic, Lying, and Psycho-linguistic Features

	FCR	FPR	LBR	F_m
MNB	0.551 (7)	0.701 (7)	0.256 (6)	0.617 (7)
DT	0.790 (6)	0.813 (5)	0.199 (4)	0.801 (6)
RF	0.874 (3)	0.825 (4)	0.202 (5)	0.849 (4)
SVM	0.888 (2)	0.868 (2)	0.148 (3)	0.878 (2)
KKN	0.848 (4)	0.774 (6)	0.270 (7)	0.809 (5)
LR	0.841 (5)	0.868 (3)	**0.139** (1)	0.855 (3)
XGBoost	**0.911** (1)	**0.873** (1)	0.144 (2)	**0.892** (1)

(a) Complete set of features; Full texts; 267 in total

	FCR	FPR	LBR	F_m
MNB	0.530 (7)	0.576 (7)	0.403 (7)	0.552 (7)
DT	0.770 (6)	0.744 (5)	0.274 (4)	0.757 (5)
RF	**0.882** (1)	0.777 (3)	0.261 (3)	0.826 (2)
SVM	0.824 (4)	**0.818** (1)	**0.190** (1)	0.821 (3)
KKN	0.831 (3)	0.750 (4)	0.285 (5)	0.788 (4)
LR	0.807 (5)	0.685 (6)	0.383 (6)	0.741 (6)
XGBoost	0.861 (2)	0.802 (2)	0.219 (2)	**0.831** (1)

(b) Minimum set of most relevant features; Full texts; 9 in total

	FCR	FPR	LBR	F_m
MNB	0.689 (7)	0.705 (7)	0.287 (7)	0.697 (7)
DT	0.824 (5)	0.824 (4)	0.175 (4)	0.824 (4)
RF	0.872 (3)	0.840 (3)	0.165 (3)	0.856 (3)
SVM	0.873 (2)	0.863 (2)	0.138 (2)	0.868 (2)
KKN	0.802 (6)	0.785 (5)	0.219 (5)	0.793 (6)
LR	0.841 (4)	0.779 (6)	0.238 (6)	0.808 (5)
XGBoost	**0.924** (1)	**0.904** (1)	**0.097** (1)	**0.914** (1)

(c) Complete set of features; Truncated texts, 267 in total

	FCR	FPR	LBR	F_m
MNB	0.490 (7)	0.578 (7)	0.398 (7)	0.530 (7)
DT	0.824 (5)	0.840 (4)	0.175 (3)	0.832 (4)
RF	0.895 (2)	0.854 (2)	0.170 (2)	0.874 (2)
SVM	0.877 (3)	0.847 (3)	0.177 (4)	0.861 (3)
KKN	0.837 (4)	0.786 (5)	0.254 (5)	0.810 (5)
LR	0.766 (6)	0.737 (6)	0.304 (6)	0.751 (6)
XGBoost	**0.903** (1)	**0.872** (1)	**0.148** (1)	**0.887** (1)

(d) Minimum set of most relevant features; Truncated texts, 9 in total

The bold values indicate the best scores. The number in parentheses in each table slot shows the ranking of each method on a specific dataset.

5 RESULTS

We invested a huge amount of time on experimenting with various linguistic-based and deceptive-oriented feature categories as well as with the different text representation methods. The experiments were carefully arranged to also find legitimate answers for the open research questions presented at the end of Section 1.

As referred in the literature, the legal news are often longer than fake news [53]. Hence, to evaluate the hypothesis that the classifiers can be biased by the size of the text, and to answer the question Q4 from Section 1.1, we run the same set of experiments on full texts and on the truncated ones. If this hypothesis is true, then conclusions based on the results obtained with the full texts may be wrong, because the classifiers can present overestimated performance. Next, we report the results of experiments regarding various setting scenarios.

5.1 Results Obtained with the Engineered Features: Various Linguistic-based and Deceptive-oriented Feature Categories

The initial set of experiments was conducted on a whole set of features and with full length texts. The results of this first set of experiments are shown in Table 5. Few methods obtained an F-measure close to 0.9, which indicates that the full set of fake news-oriented features are sufficiently relevant to facilitate the automatic detection of close to 90% of the fake news (FCR). XGBoost obtained the best result for all the four performance measures. It was able to detect more than 90% of fake news with the price of wrong blocking 1.44% of true news. MNB achieved the worst results.

We run the same experiments on complete set of features but in this case, we used the truncated version of real news texts. Truncation was done on length of 50 words, which is the fake news average length in words.

As it can be seen in Tables 5(a) and 5(c), the results from experiments on complete set of features and on truncated version of real news are at least 10% better compared to the same experiments on full length texts. These results do not confirm, at least in this case, our hypothesis that the classifiers are biased by the length of the text.

Additionally, we calculated feature importance with forests of trees method to find the minimal size of relevant features to optimize text classification error. By doing so, we identified the following minimal feature set consisting of next 13 features given in decreasing order of significance: number of nouns, number of syllables, number of punctuation, the frequency of Albanian letter "ë", the frequency of character ".", the frequency of word "ua", the frequency of suffix "uar", number of uppercase characters, number of complex words, number of digits, the frequency of suffix "ës", the average length of sentences and the number of out-of-vocabulary words.

Authors were quite delighted to see that many of the top features are thoroughly aligned with their expectations. For instance, the number of Albanian letter "ë" signals that news published in reliable sources will contain higher number of it, while the news originating from unreliable, suspicious sources usually contain it at lower frequency. The difference in usage of this character derives from the fact that the standard PC/laptop keyboards do not contain character "ë". Hence, the rigorous writers and serious journalists will usually endeavor to use it grammatically correct, while the fake news copywriters will simply use character "e" instead. Another group of impressive examples are features "uar", "ës", and "ua". They are strongly connected to two major Albanian language dialects, namely, Tosk and Gheg. The Albanian adjective "vonuar" (Tosk) with its stem "von" and suffix "uar" in more relaxed usage takes the form of "vonum" (Gheg) ("stem:von+suffix:um"). The fake news would usually contain lower frequency of suffixes such as "uar" or "ës" features.

As it can be verified from Tables 5(b) and 5(d), the same pattern of results can be observed with reduced set of features on the both version of lengths.

With an additional experiment, we tried to answer the question of how accurate would be the classifier based only on a single feature, i.e., the length of the article. Table 6(a) verifies the expected outcome that classification based merely on the length of the articles is of lower quality. The main reason for such effect is the average length of fake news being much shorter than the one of the true news. By additionally truncating the length of the articles first to 100 (Table 6(b)), later to 50 (Table 6(c)) words in length, we show that our classifiers accuracy degrades to almost random baseline classifier quality.

Yet another intriguing research question would be to ask how likely is it that proposed fake news classifier is in fact an Albanian dialect classifier. By explicitly using only the subset of features closely related to dialect differences, we run a set of experiments. Results summarized in Table 6(d) clearly indicate that the classifier based on this set of dialectic features is significantly less accurate when compared to the best results obtained with full set of features. To further elaborate on this point, we run an additional set of experiments by testing our XGBoost-trained model with four sets of texts: Tosk true news, Tosk fake news, Gheg true news (topically related Gheg texts that are not fake news), and Gheg fake news. The total number of news used in this set of experiments are 80, each subset consists of 20 manually and diligently labeled news. As reported in Table 6(e) where Tosk true and fake news are used and in Table 6(f) where similar is done with Gheg news, our classifier is able to recognize fake news regardless of the dialect used in the articles.

5.2 Results Obtained with Features Generated by Text Representation Techniques Using Full Texts

The results for BoW and other more advanced text representation techniques are condensed in Table 7. In each subtable, the highest scores are emphasized in bold.

Besides the specific set of fake news-oriented features, we experimented with different text representations such as BoW, Word2Vec, FastText, and BERT. This section presents the results of these experiments. Initially, we used the full text of the documents. For BoW, we performed the following sub-experiments: (1) No common NLP pre-processing is applied and BOW is used in

Table 6. Results Obtained by Each Method in the Experiments with the Length and Dialect Classifiers

	FCR	FPR	LBR	F_m
MNB	**1.000** (1)	0.506 (7)	1.000 (7)	0.672 (7)
DT	0.681 (5)	0.688 (6)	0.317 (6)	0.684 (6)
RF	0.673 (6)	0.707 (5)	0.286 (4)	0.690 (5)
SVM	0.737 (4)	0.807 (2)	0.181 (2)	**0.771** (1)
KKN	0.812 (2)	0.734 (4)	0.303 (5)	0.771 (2)
LR	0.659 (7)	**0.834** (1)	**0.135** (1)	0.736 (4)
XGBoost	0.752 (3)	0.753 (3)	0.253 (3)	0.752 (3)

(a) News classifier trained only on the article length as a feature. Full texts.

	FCR	FPR	LBR	F_m
MNB	**1.000** (1)	0.506 (7)	1.000 (7)	0.672 (7)
DT	0.793 (3)	0.713 (6)	0.327 (6)	0.751 (6)
RF	0.779 (5)	0.728 (5)	0.299 (5)	0.752 (5)
SVM	0.753 (7)	**0.789** (1)	**0.207** (1)	0.770 (3)
KKN	0.819 (2)	0.738 (4)	0.298 (4)	**0.777** (1)
LR	0.779 (4)	0.771 (2)	0.237 (2)	0.775 (2)
XGBoost	0.768 (6)	0.753 (3)	0.258 (3)	0.760 (4)

(b) Article length as a feature. Texts truncated to 100 words length.

	FCR	FPR	LBR	F_m
MNB	**1.000** (1)	0.506 (6)	1.000 (6)	0.672 (6)
DT	0.984 (4)	0.521 (2)	0.929 (3)	**0.681** (1)
RF	0.986 (3)	0.521 (3)	0.931 (4)	0.681 (2)
SVM	0.981 (5)	0.519 (5)	0.933 (5)	0.679 (5)
KKN	0.981 (6)	**0.522** (1)	**0.923** (1)	0.681 (3)
LR	1.000 (2)	0.506 (7)	1.000 (7)	0.672 (7)
XGBoost	0.981 (7)	0.521 (4)	0.924 (2)	0.681 (4)

(c) Article length as a feature. Texts truncated to 50 words length.

	FCR	FPR	LBR	F_m
MNB	0.683 (7)	0.558 (7)	0.556 (7)	0.614 (7)
DT	0.721 (5)	0.671 (6)	0.363 (6)	0.695 (6)
RF	0.817 (2)	0.736 (5)	0.301 (5)	0.774 (4)
SVM	0.777 (4)	0.803 (2)	0.196 (2)	**0.790** (1)
KKN	**0.822** (1)	0.745 (4)	0.289 (4)	0.781 (2)
LR	0.696 (6)	**0.828** (1)	**0.148** (1)	0.756 (5)
XGBoost	0.796 (3)	0.760 (3)	0.258 (3)	0.778 (3)

(d) Dialect classifier results.

	FCR	FPR	LBR	F_m
XGBoost	0.900	0.857	0.150	0.878

(e) Dialect classifier results. Tosk true and fake news classification results.

	FCR	FPR	LBR	F_m
XGBoost	0.800	0.762	0.250	0.780

(f) Dialect classifier results. Gheg true and fake news classification results.

The number in parentheses in each table slot shows the ranking of each method on a specific dataset.

Table 7(a); (2) Both stopword removal and stemming was applied in Table 7(b); and (3) Feature ranking was used for selecting the best 2,000 features shown in Table 7(c).

Unexpectedly, the results showed that stop-word removal and stemming did not significantly improve the performance of the classification methods. The only reasonable explanation is that in fake news classification task the stopwords and various word forms play an important role, which is not true for many other domains [56, 58]. The MNB in plain BoW experiments shows the precision of about 94% fake news detection and it wrongly blocks the true news on 15% of cases. As far as the feature selection is concerned, the results from Table 7(c) clearly imply that feature selection does not improve results.

The scores in the experiments with BoW Table 7(a) were also better than those obtained with the engineered features in Table 5(a). However, the dimensionality of the BoW-based representation is much higher than the dimensionality of the representation based on engineered features, even more when the minimal set of 13 features is considered. Hence, in low computational resource settings, a fake news classifier based on the latter can be trained and used in production as a very fast, scalable, and reliable inference model.

When it comes to experiments with different text representation models, analysis becomes even harder. As an illustration, let us analyze the best F-measure obtained with BoW, which is 0.909, compared with Word2Vec, FastText, and BERT best results, 0.912, 0.904, 0.924, respectively. The differences are not significant, and the possible explanations are as follows: The advanced SOTA language models for the Albanian are trained on relatively clear, well-written, low-noise, and standard Albanian language. On the other side, the language used in fake news not only contains noisy data such as abbreviations, slang, and misspelled words, but it is often written in non-standard Albanian dialect known as northern Albanian or Gheg. One of the best solutions to this issue would be to train the word embedding models on real-world noisy texts.

Table 7. Scores Obtained by Each Method in the Experiments with the Full Texts

	FCR	FPR	LBR	F_m
MNB	0.949 (2)	0.867 (3)	0.157 (4)	0.906 (2)
DT	0.850 (6)	0.832 (6)	0.184 (6)	0.841 (6)
RF	0.866 (7)	0.867 (4)	0.143 (3)	0.866 (5)
SVM	0.938 (4)	0.883 (2)	0.134 (2)	**0.909** (1)
KKN	**0.962** (1)	0.695 (7)	0.454 (7)	0.807 (7)
LR	0.912 (5)	**0.888** (1)	**0.124** (1)	0.900 (4)
XGBoost	0.939 (3)	0.866 (5)	0.157 (5)	0.901 (3)

(a) BoW

	FCR	FPR	LBR	F_m
MNB	0.926 (4)	0.871 (3)	0.148 (2)	0.898 (3)
DT	0.771 (7)	0.843 (5)	0.155 (4)	0.805 (7)
RF	0.910 (5)	0.851 (4)	0.172 (5)	0.879 (5)
SVM	**0.968** (1)	0.874 (2)	0.150 (3)	**0.919** (1)
KKN	0.886 (6)	0.826 (7)	0.201 (7)	0.855 (6)
LR	0.947 (3)	**0.886** (1)	**0.131** (1)	0.916 (2)
XGBoost	0.949 (2)	0.842 (6)	0.191 (6)	0.892 (4)

(b) BoW - stopwords and stemming

	FCR	FPR	LBR	F_m
MNB	0.912 (4)	**0.900** (1)	**0.108** (1)	**0.906** (1)
DT	0.843 (7)	0.831 (6)	0.184 (6)	0.837 (7)
RF	0.915 (3)	0.850 (5)	0.174 (5)	0.881 (5)
SVM	0.936 (2)	0.874 (3)	0.145 (3)	0.904 (2)
KKN	0.854 (6)	0.830 (7)	0.188 (7)	0.842 (6)
LR	0.910 (5)	0.881 (2)	0.133 (2)	0.895 (4)
XGBoost	**0.938** (1)	0.858 (4)	0.167 (4)	0.896 (3)

(c) BoW - feature selection

	FCR	FPR	LBR	F_m
MNB	0.936 (2)	0.889 (2)	0.138 (2)	**0.912** (1)
DT	0.833 (7)	0.836 (6)	0.187 (6)	0.834 (6)
RF	0.896 (6)	0.864 (5)	0.166 (4)	0.879 (5)
SVM	0.923 (3)	0.884 (3)	0.145 (3)	0.905 (2)
KKN	0.914 (5)	0.768 (7)	0.326 (7)	0.830 (7)
LR	0.916 (4)	**0.901** (1)	**0.133** (1)	0.903 (4)
XGBoost	**0.944** (1)	0.867 (4)	0.167 (5)	0.904 (3)

(d) Word2vec

	FCR	FPR	LBR	F_m
MNB	0.927 (3)	0.858 (3)	0.142 (3)	0.892 (3)
DT	0.801 (7)	0.825 (6)	0.162 (5)	0.813 (7)
RF	0.878 (6)	0.849 (4)	0.147 (4)	0.863 (5)
SVM	**0.943** (1)	0.868 (2)	0.132 (2)	**0.904** (1)
KKN	0.914 (5)	0.750 (7)	0.318 (7)	0.821 (6)
LR	0.920 (4)	**0.877** (1)	**0.117** (1)	0.898 (2)
XGBoost	0.934 (2)	0.844 (5)	0.164 (6)	0.887 (4)

(e) FastText

	FCR	FPR	LBR	F_m
MNB	0.948 (3)	0.879 (3)	0.162 (3)	0.911 (3)
DT	0.820 (7)	0.845 (6)	0.182 (5)	0.832 (7)
RF	0.898 (6)	0.869 (4)	0.167 (4)	0.883 (5)
SVM	**0.962** (1)	0.888 (2)	0.152 (2)	**0.924** (1)
KKN	0.934 (5)	0.770 (7)	0.338 (7)	0.841 (6)
LR	0.940 (4)	**0.897** (1)	**0.138** (1)	0.918 (2)
XGBoost	0.954 (2)	0.864 (5)	0.183 (6)	0.907 (4)

(f) BERT

Regarding the classification methods, it is clear that XGBoost obtained the best score in most of the experiments with the BoW-based representation, being able to detect, on average, 97% of fake news with the price of wrongly blocking, on average, 6% of true news. In the experiments with the distributive text representation techniques (Word2Vec and FastText), RF achieved the best results. However, DT and MNB obtained the worst FCR and F-measure in all the experiments.

5.3 Results Obtained with Features Generated by Text Representation Techniques Using Truncated Texts

The results of experiments explained in this section are given in Table 8. The experiments are the same with ones given in the previous section with a single difference: The full text length of news content in this case is limited to 50 words in length.

Following the same pattern as in the previous set of experiments, removing stopwords, applying stemming, and performing feature selection did not significantly improve the results with the truncated news. The best F_m score with BoW was 0.858, it increased to 0.863 after applying stopwords removal and stemming, and it decreased to 0.854 after applying feature selection. Similar result variations can be observed with other performance measures. The results in the experiments with Word2Vec are roughly similar to those obtained in experiments with BoW. The results obtained in this set of experiments simply strengthened our previous hypothesis that the word embedding models when used on our problem were generating the vectors of low quality. The main reason for this, in our understanding, is the Albanian fake news domain's enormity and the lack of pre-trained models' ability to treat the noisy content.

Table 8. Scores Obtained by Each Method in the Experiments with the Truncated Texts

	FCR	FPR	LBR	F_m
MNB	0.782 (6)	**0.875** (1)	**0.114** (1)	0.826 (4)
DT	0.761 (7)	0.780 (5)	0.221 (4)	0.771 (7)
RF	0.802 (5)	0.793 (4)	0.215 (3)	0.797 (5)
SVM	0.903 (3)	0.818 (2)	0.207 (2)	**0.858** (1)
KKN	0.810 (4)	0.764 (7)	0.258 (6)	0.786 (6)
LR	0.905 (2)	0.806 (3)	0.224 (5)	0.853 (2)
XGBoost	**0.910** (1)	0.772 (6)	0.276 (7)	0.835 (3)

(a) BoW

	FCR	FPR	LBR	F_m
MNB	0.784 (6)	**0.863** (1)	**0.128** (1)	0.822 (4)
DT	0.722 (7)	0.768 (5)	0.224 (3)	0.744 (7)
RF	0.838 (4)	0.784 (4)	0.237 (5)	0.810 (5)
SVM	0.926 (2)	0.807 (3)	0.227 (4)	**0.863** (1)
KKN	0.805 (5)	0.760 (6)	0.261 (6)	0.782 (6)
LR	0.897 (3)	0.823 (2)	0.199 (2)	0.858 (2)
XGBoost	**0.928** (1)	0.743 (7)	0.330 (7)	0.825 (3)

(b) BoW stopwords and stemming

	FCR	FPR	LBR	F_m
MNB	0.782 (6)	**0.874** (1)	**0.116** (1)	0.826 (5)
DT	0.763 (7)	0.788 (5)	0.210 (3)	0.775 (6)
RF	0.846 (4)	0.814 (2)	0.199 (2)	0.830 (4)
SVM	0.913 (2)	0.802 (4)	0.232 (5)	**0.854** (1)
KKN	0.805 (5)	0.656 (7)	0.434 (7)	0.723 (7)
LR	0.892 (3)	0.810 (3)	0.215 (4)	0.849 (2)
XGBoost	**0.920** (1)	0.778 (6)	0.269 (6)	0.843 (3)

(c) BoW - feature selection

	FCR	FPR	LBR	F_m
MNB	0.779 (6)	**0.871** (1)	**0.112** (1)	0.822 (4)
DT	0.758 (7)	0.781 (5)	0.212 (3)	0.769 (6)
RF	0.821 (4)	0.800 (4)	0.204 (2)	0.810 (5)
SVM	0.905 (2)	0.806 (2)	0.216 (4)	**0.853** (1)
KKN	0.804 (5)	0.706 (7)	0.342 (7)	0.751 (7)
LR	0.895 (3)	0.804 (3)	0.216 (5)	0.847 (2)
XGBoost	**0.911** (1)	0.772 (6)	0.269 (6)	0.836 (3)

(d) Word2vec

In the previous section, we also showed that the results obtained with the content features and full length texts were slightly higher than those obtained with the engineered features. On contrary, when content features are extracted form truncated texts, engineered feature-based classification experiments show better results. For instance, the best F_m scores from Table 5 (0.892, 0.831, 0.914, 0.887) are much better than the ones from Table 8 (0.858, 0.863, 0.854, 0.853).

The results show another stimulating property: Across the various sub-experiments, regardless of the features involved, the best methods are the same when they are sorted by scores. Truncating the length of content, in our understanding, brings stability into training processes, hence results show the described property.

We observed the difference between the results obtained with the full texts and the truncated texts, which was not more than 5% difference on F_m. This result to some extent confirms our hypothesis that the classifiers are biased by the size of the text, especially when contentbased features are concerned. Hence, we recommend using the truncated version of true news so classification model will be trained with minimized effect of length biases.

5.4 How Do News Sources Affect the Classification Results?

The initial set of fake news extracted from unreliable sources contained 4K items. During the verification process the set was further filtered to 2K of manually verified fake F_{NR} and 2K of real news R_{NR}. Consequently, we identified a set of news whose content is not fake, although they originate from unreliable sources and hence show the regular fake news features. To measure the effect of the unreliable sources and the writing style typical for this kind of publisher, we run the same set of classification experiments on these newly obtained training sets. Table 9 summarizes the output of the experiment.

The obtained F1 measures from experiments with engineered features are on average much lower than one obtained with text modeling. The latter as well are greatly lower than results from previous experiments. The most logical explanation is that: (1) when merely content is taken into account, the source reliability and the writing style of the article do not affect, to a great extent, the classification results; (2) oppositely, when manually engineered features are used, the effect is greater, hence the classification scores drop drastically.

Table 9. Results Obtained by Each Method in the Experiments with Fake and Real News Originating from Unreliable Sources F_NR vs R_NR

	FCR	FPR	LBR	F_m
MNB	0.453 (2)	0.490 (7)	0.396 (7)	0.471 (4)
DT	0.442 (3)	0.494 (6)	0.381 (5)	0.467 (5)
RF	**0.525** (1)	0.528 (5)	0.395 (6)	**0.527** (1)
SVM	0.356 (6)	**0.669** (1)	**0.148** (1)	0.464 (6)
KKN	0.356 (7)	0.669 (2)	0.148 (2)	0.464 (7)
LR	0.434 (4)	0.623 (4)	0.221 (3)	0.512 (2)
XGBoost	0.416 (5)	0.624 (3)	0.211 (4)	0.499 (3)

(a) F_NR vs R_NR Features set; Full texts; 267 in total

	FCR	FPR	LBR	F_m
MNB	0.158 (7)	**1.000** (1)	**0.000** (1)	0.273 (7)
DT	0.526 (4)	0.769 (7)	0.273 (6)	0.625 (5)
RF	0.526 (5)	1.000 (2)	0.000 (2)	0.690 (4)
SVM	0.684 (2)	0.929 (4)	0.091 (4)	**0.788** (1)
KKN	**0.737** (1)	0.778 (6)	0.364 (7)	0.757 (2)
LR	0.368 (6)	1.000 (3)	0.000 (3)	0.538 (6)
XGBoost	0.632 (3)	0.923 (5)	0.091 (5)	0.750 (3)

(b) F_NR vs R_NR Text modeling; BOW; normalized

Hereby, the results of this set of experiments clearly show that it is the news/articles source and hence the used writing style that plays the most significant role in fake news classification success.

6 CONCLUSIONS

Fake news detection-related research is of big benefit in general for global society and in particular for societies speaking low-resource languages. Although, this is an issue that humanity is facing for a long time and in various forms, nowadays this is even more emphasized due to the continuous internet and social media usage expansion. In this article, we presented an extensive evaluation of a novel Albanian fake news corpus to find the most relevant minimum set of features or combination of features and the most appropriate machine learning methods to be used for the automatic detection of fake news. We carefully designed the broad set of experiments for answering the following research questions:

- Q1: What is the best strategy for compiling the novel fake news dataset for a low-resource language such as Albanian?

  The answer for this question starts with the detailed description given in Section 3, enriched with experiments and results presented in Sections 4 and 5, respectively. Particularly, compared to the previous work that has been done in this field, we presented a novel strategy for discovering and identifying the fake news., which as a process proved to be an arduous task in the past, both when it was performed manually as well as automatically. By starting from a small set of news and rapidly reaching dozens of thousands news we showed the correctness of our strategy. Conducting this process over social media came with its advantages. First, we needed a single scraping tool, which was used on social media, to fetch news from various sources by which we also were able to obtain very distinct features related to these news. Second, we constructed a simple single page scraper that used domain-specific XPaths and was able to fully collect news residing in varied websites. This was highly more efficient than creating complex crawlers for each website. Another important part of this process was the solution for the problem of alignment of true to fake news. We based it on LSA/LSI topic modeling, which was also used to calculate news similarities;
- Q2: What is the best and minimal set of features for successful Albanian fake news classification? The various experiments were conducted mainly experimenting with the lexical/linguistic, lying, and psycho-linguistic features on one side and features generated by text representation techniques (BoW, Word2Vec, FastText, BERT) on the other side. Experiments show that there are no significant differences between using the engineered set of features and the results obtained using the features extracted from language models. Nevertheless, working on finding the minimal set of relevant features revealed highly valuable insights regarding the corpus, Albanian fake content, features, and their effect on classification;

- Q3: What are the best current methods for automatic detection of the Albanian fake news? To answer this question, we compared the performance of the following widely used machine learning methods: LR, SVM, RF, Bagging, DT, MNB, and XGBoost. None of these methods was superior to the others in all experiments. However, the methods that obtained the best results in most of the evaluated scenarios were XGBoost, LR, SVM, and RF. MNB and DT, in general, obtained the lowest results;
- Q4: Can the size of the texts influence the results of the classification? Many previous woks noted that the average length of true news is larger than the length of typical fake news. Hence, to investigate whether this stands for Albanian language and web, we performed experiments with full texts and with truncated ones. The output of this set of experiments does not show significant increase of classification results. Nevertheless, results still weakly signal a length bias, hence, we postulate that the truncated texts are closer to expected results for the real-world applications;
- Q5: How do news sources affect the classification results? The results of this set of experiments clearly show that it is the news/articles source and hence the writing style used in them that plays the most significant role in fake news classification success.

The future work will be mainly shaped by the new questions or shortcomings we already identified in this work. First, training the modern language models on more noisy Albanian texts, using the other Albanian dialects could be of huge benefit. By doing so, one would be able to better utilize those models in obtaining superior and more robust classification results. Enriching the Albanian fake news corpus with comments from social media could provide additional signals for more successful classification settings. We would also like to experiment with (1) ensembles of predictions using different sets of features and (2) stacking of classifiers trained with different sets of features.

Declaration of Competing Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Credit authorship contribution statement *Ercan Canhasi:* Conceptualization, Methodology, Implementation/Coding, Investigation, Writing - original draft, Experimentation, Validation, Data curating, Supervision, Visualization. *Rexhep Shijaku*: Manual and Automatic corpus creation/implementation, Data extraction/scraping, Writing Section 3, Validation, Data curating, Visualization, Proofreading *Erblin Berisha*: Validation, Data curating, Visualization.

REFERENCES

[1] Naomi S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *Amer. Statist.* 46, 3 (1992), 175–185.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Ling.* 5 (2017), 135–146.

[3] Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Inf. Sci.* 497 (2019), 38–55.

[4] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*. 144–152.

[5] L. Breiman. 2001. Random forests. *Mach. Learn.* 45 (10 2001), 5–32. DOI:https://doi.org/10.1023/A:1010950718922

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. 2017. *Classification and Regression Trees*. Wiley interdisciplinary reviews: data mining and knowledge discovery. DOI:https://doi.org/10.1201/9781315139470

[7] E. Merkley, P. J. Loewen, T. Owen, D. Ruths, L. Teichmann, O. Zhilin, and A. Bridgman. 2015. The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harv. Kenn. School (HKS) Misinf. Rev.* (2015).

[8] Erica J. Briscoe, D. Scott Appling, and Heather Hayes. 2014. Cues to deception in social media communications. In *Proceedings of the 47th Hawaii International Conference on System Sciences*. IEEE, 1435–1443.

[9] Cody Buntain and Jennifer Golbeck. 2017. Automatically identifying fake news in popular Twitter threads. In *Proceedings of the IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 208–215.

[10] Judee K. Burgoon, J. Pete Blair, Tiantian Qin, and Jay F. Nunamaker. 2003. Detecting deception through linguistic analysis. In *Proceedings of the International Conference on Intelligence and Security Informatics*. Springer, 91–101.
[11] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*. 675–684.
[12] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 9–16.
[13] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.
[14] Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proc. Assoc. Inf. Sci. Technol.* 52, 1 (2015), 1–4.
[15] Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. 2015. Survey of review spam detection using machine learning techniques. *J. Big Data* 2, 1 (2015), 23.
[16] Seyyed Mohammad Hossein Dadgar, Mohammad Shirzad Araghi, and Morteza Mastery Farahani. 2016. A novel text mining approach based on TF-IDF and support vector machine for news classification. In *Proceedings of the IEEE International Conference on Engineering and Technology (ICETECH)*. IEEE, 112–116.
[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[18] Susan T. Dumais. 2004. Latent semantic analysis. *Ann. Rev. Inf. Sci. Technol.* 38, 1 (2004), 188–230.
[19] Vanessa Wei Feng and Graeme Hirst. 2013. Detecting deceptive opinions with profile compatibility. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*. 338–346.
[20] William Ferreira and Andreas Vlachos. 2016. Emergent: A novel data-set for stance classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1163–1168.
[21] Tommaso Fornaciari and Massimo Poesio. 2013. Automatic deception detection in Italian court cases. *Artif. Intell. Law* 21, 3 (2013), 303–340.
[22] E. E. Papalexakis, G. B. Guacho, and S. Abdali. 2018. Semi-supervised content-based fake news detection using tensor embeddings and label propagation. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 322–325.
[23] Shlok Gilda. 2017. Evaluating machine learning algorithms for fake news detection. In *Proceedings of the IEEE 15th Student Conference on Research and Development (SCOReD)*. IEEE, 110–115.
[24] Chuan Guo, Juan Cao, Xueyao Zhang, Kai Shu, and Miao Yu. 2019. Exploiting emotions for fake news detection on social media. *arXiv preprint arXiv:1903.01728* (2019).
[25] Jeffrey T. Hancock, Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. 2007. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discour. Process.* 45, 1 (2007), 1–23.
[26] David W. Hosmer Jr, Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. Vol. 398. John Wiley & Sons.
[27] Eduard Hovy and Julia Lavid. 2010. Towards a "science" of corpus annotation: A new methodological challenge for corpus linguistics. *Int. J. Translat.* 22, 1 (2010), 13–36.
[28] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*. AAAI Press, 2972–2978.
[29] Besim Kabashi and Thomas Proisl. 2018. Albanian part-of-speech tagging: Gold standard and evaluation. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*.
[30] M. Biba, J. Kanerva, S. Rönnqvist, F. Ginter and N. Kote. 2019. Morphological tagging and lemmatization of Albanian: A manually annotated corpus and neural models. *arXiv preprint arXiv* (2019).
[31] Yunfei Long. 2017. Fake news detection through multi-perspective speaker profiles. Proceedings of the eighth international joint conference on natural language processing (volume 2: Short papers). 2017
[32] Christopher D. Manning, Hinrich Schütze, and Prabhakar Raghavan. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
[33] Priyanka Meel and Dinesh Kumar Vishwakarma. 2020. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications* 153 (2020), 112986.
[34] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 3111–3119.

[35] Tanushree Mitra, Graham P. Wright, and Eric Gilbert. 2017. A parsimonious language model of social media credibility across disparate events. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*. 126–145.

[36] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673* (2019).

[37] Renato Moraes Silva, Tulio Alberto, Tiago Almeida, and Akebo Yamakami. 2017. Towards filtering undesired short text messages using an online learning approach with semantic indexing. *Exp. Syst. Applic.* 83 (04 2017). DOI:https://doi.org/10.1016/j.eswa.2017.04.055

[38] Alex Olivieri, Shaban Shabani, Maria Sokhn, and Philippe Cudré-Mauroux. 2019. Creating task-generic features for fake news detection. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

[39] Feyza Altunbey Ozbay and Bilal Alatas. 2020. Fake news detection within online social media using supervised artificial intelligence algorithms. *Phys. A: Statist. Mech. Applic.* 540 (2020), 123174.

[40] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg et al. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12 (2011), 2825–2830.

[41] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104* (2017).

[42] Verónica Pérez-Rosas and Rada Mihalcea. 2014. Cross-cultural deception detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. 440–445.

[43] Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1120–1125.

[44] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *Proceedings of the European Conference on Information Retrieval*. Springer, 810–817.

[45] Vahed Qazvinian, Emily Rosengren, Dragomir Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1589–1599.

[46] Yumeng Qin, Dominik Wurzer, Victor Lavrenko, and Cunchen Tang. 2016. Spotting rumors via novelty detection. *arXiv preprint arXiv:1611.06322* (2016).

[47] Victoria L. Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? Using satirical cues to detect potentially misleading news. In *Proceedings of the 2nd Workshop on Computational Approaches to Deception Detection*. 7–17.

[48] Victoria L. Rubin, Niall J. Conroy, and Yimin Chen. 2015. Towards news verification: Deception detection methods for news discourse. In *Proceedings of the Hawaii International Conference on System Sciences*. 5–8.

[49] G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM* 18, 11 (Nov. 1975), 613–620. DOI:https://doi.org/10.1145/361219.361220.

[50] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* 8, 3 (2020), 171–188.

[51] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. CoRR abs/1708.01967 (2017).

[52] Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding user profiles on social media for fake news detection. In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 430–435.

[53] Renato M. Silva, Roney L. S. Santos, Tiago A. Almeida, and Thiago A. S. Pardo. 2020. Towards automatically filtering fake news in Portuguese. *Exp. Syst. Applic.* 146 (2020), 113199.

[54] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506* (2017).

[55] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355* (2018).

[56] Michal Toman, Roman Tesar, and Karel Jezek. 2006. Influence of word normalization on text classification. *Proc. InSciT* 4 (2006), 354–358.

[57] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Proceedings of the the Web Conference*. 517–524.

[58] Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Inf. Process. Manag.* 50, 1 (2014), 104–112.

[59] Ben Verhoeven and Walter Daelemans. 2014. CLiPS stylometry investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *Proceedings of the Language Resources and Evaluation Conference*. 3081–3085.

[60] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 647–653.

[61] William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).

[62] Aleksander Wawer and Grzegorz Wojdyga. 2019. Fact checking or psycholinguistics: How to distinguish fake and true. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[63] Hu Zhang, Hong-ye Tan, Jia-heng Zheng et al. 2009. Deception detection based on SVM for Chinese text in CMC. In *Proceedings of the 6th International Conference on Information Technology: New Generations*. IEEE, 481–486.

[64] Jiawei Zhang, Bowen Dong, and S. Yu Philip. 2020. FakeDetector: Effective fake news detection with deep diffusive neural network. In *Proceedings of the IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1826–1829.

[65] Xichen Zhang and Ali A. Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Inf. Process. Manag.* 57, 2 (2020), 102025.

[66] Lina Zhou, Judee K. Burgoon, Douglas P. Twitchell, Tiantian Qin, and Jay F. Nunamaker Jr. 2004. A comparison of classification methods for predicting deception in computer-mediated communication. *J. Manag. Inf. Syst.* 20, 4 (2004), 139–166.

[67] Lina Zhou, Douglas P. Twitchell, Tiantian Qin, Judee K. Burgoon, and Jay F. Nunamaker. 2003. An exploratory study into deception detection in text-based computer-mediated communication. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*. IEEE.

[68] Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. 2020. Fake news early detection: A theory-driven model. *Dig. Threats: Res. Pract.* 1, 2 (2020), 1–25.

[69] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-aware multi-modal fake news detection. *arXiv preprint arXiv:2003.04981* (2020).

[70] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2016. Learning reporting dynamics during breaking news for rumour detection in social media. *arXiv preprint arXiv:1610.07363* (2016).

accepted 17 September 2021