Survey of Fake News Datasets and Detection Methods in European and Asian Languages

Maaz Amjad[1], Sabur Butt[1], Alisa Zhila[2], Grigori Sidorov[1], Liliana Chanona-Hernandez[3], and Alexander Gelbukh[1]

[1]Instituto Politécnico Nacional, Centro de Investigación en Computación (CIC), Gustavo A. Madero, 07738 Mexico City, Mexico, maazamjad@phystech.edu, sbutt2021@cic.ipn.mx, gelbukh@gelbukh.com

[2] Ronin Institute for Independent Scholarship, United States, alisa.zhila@ronininstitute.org

[3]Instituto Politécnico Nacional, ESIMEZ, lchanonah2100tmp@alumnoguinda.mx

Abstract: The presence of fake news and "alternative facts" across the web is a global phenomenon that received considerable attention in recent years. Several researchers have made substantial efforts to automatically identify fake news articles based on linguistic features and neural network-based methods. However, automatic classification via machine and deep learning techniques demands a significant amount of annotated data. While several state-of-the-art datasets for the English language are available and commonly utilized for research, fake news detection in low-resource languages gained less attention. This study surveys the publicly available datasets of fake news in low/medium-resourced Asian and European languages. We also highlight the vacuum of datasets and methods in these languages. Moreover, we summarize the proposed methods and the metrics used to evaluate the classifiers in identifying fake news. This study is helpful for analysis of the available sources in the lower resource languages to solve fake news detection challenges.

Keywords: datasets, fake news, low resource languages, deep learning, machine learning, evaluation metrics.

AMS Subject Classification: 68T50 Natural language processing, 68T01 General topics in artificial intelligence

1 Introduction

The fake news phenomenon imposes devastating and havoc impact worldwide. It poses not only technical challenges for social media platforms but also a dramatic impact on everyday life. Rampant "*online*" fake news leads to "*offline*" societal

events (e.g., the PizzaGate[1]). For example, according to the United Kingdom Office of National Statistics, anti-vaccination misinformation online reduced vaccination coverage across England and Wales[2]. In another example, Financial Times reported that French regulators had fined Bloomberg C5M for publishing a fake press release[3]. Therefore, social media platforms and other organizations should gear up to battle the dissemination of fake news and take preventive measures to maintain a trustworthy news ecosystem.

Manual verification of news articles is troublesome. Traditionally, journalists are required to verify claims against written or spoken facts. This requires a substantial amount of time and resources. For example, in PolitiFact[4] employs at least two news editors to authenticate the news article. Additionally, the amount of data is exploding, worldwide and in all languages, making detection of deceiving and spin information difficult because of its fast dissemination and easy availability. This brings the need for constant monitoring of digital content employing automatic fake news detection.

Automatic fake news detection is aimed to assist in monitoring and analyzing of giant amounts of data, and to reduce human efforts and time resources. Multiple advanced techniques have been investigated to approach fake news detection such as traditional (linear and non-linear) Machine Learning/Deep (ML/DL), Data Mining (DM), and Natural Language Processing (NLP). However, the most well-known research has been focused around the resource-rich languages, in terms of availability of tools, size of datasets, and previous research, predominantly Western, such as English [1,2,3].

In this paper, we survey the available resources for fake news detection from the perspective of Asian and European lower-resource languages. First of all, we want to derive attention to the size of the fake news problem in the regions where millions speak a variety of low to medium-resource languages of people. Next, we show that substantial effort exists for these languages to solve the fake news problem. We also gave a systematic comparison of fake news definitions used in various studies. Further, we provide a detailed analysis to highlight the points, where more improvement or effort is needed to achieve more impactful results.

Our contributions can be summarized as follows:

- We provide the first review of recent studies in low and medium resources, particularly, Asian and European languages for automatic fake news detection;

1 https://www.rollingstone.com/feature/anatomy-of-a-fake-news-scandal-125877

2 https://www.scl.org/articles/12022-the-real-world-effects-of-fake-news-and-how-to-quantify-them

3 https://www.ft.com/content/b082851a-07c1-11ea-a984-fbbacad9e7dd

4 https://www.politifact.com/

- We categorize and summarize publicly available datasets in Asian and European languages;
- We study and compare the various definitions of "fake news" used in the surveyed works;
- We identify different general approaches to fake news detection and group the studies accordingly;
- We overview the metrics used for fake news detection evaluation in the surveyed studies and show to which extent the results can be compared across different works;
- Finally, we identify the main challenges around the fake news detection problem and highlight the promising pathways for further research to solve fake news detection problem in Asian and European languages.

We hope this work may serve as a useful reference for the sources available to develop fake news detection systems for low-resource languages.

The remaining paper is structured as follows. Section 2 presents and discusses various definitions of the term fake news. Section 3 describes and groups the datasets. Section 4 sheds light on experimental methodologies employed in the development of fake news detection systems. It also presents and compares the results. Further, Section 5 provides comparison of popular evaluation metrics. Finally, Section 6 concludes with a discussion and Section 7 outlines future opportunities.

2 Variations of Definitions of Fake News

Multiple definitions of fake news were proposed in [1,4,5]. In the study [4], the authors presented several definitions of *disinformation* elaborated by multiple researchers (in contrast to *misinformation*). The study [4] concluded that disinformation has a specific goal, which is to provide information that misleads the reader.

In a similar study, researchers were investigating the ways the term "Fake news" was used [5]. The researchers categorized fake news into six types of news: fabrication, news satire, manipulation (e.g., editing pictures), advertising (e.g., ads depict as professional journalism), propaganda, and news parody. In the previous study [5], the scientists highlighted two popular themes among six types of news: the appropriateness and purpose of news articles.

The term "Fake news" has been defined from different perspectives. For example, satire can be defined as a news article that contains factually incorrect information. Nonetheless, the goal of this news article is not to deceive a reader by providing

unproven information but to highlight shameful, unethical, or otherwise "bad" attitudes. Finally, this brings up a new challenge to identify fake news because addressing this task demands clear definitions and examples to combat fake news on web-scale.

The year 2016 has been known as a *"post-truth"* era since it introduced recent advancements into traditional politics. In that view, Oxford Dictionary[5] announced *"post-truth"* as the word of the year 2016 shows that the sensitivity of fake news is a global problem. Similarly, Cambridge dictionary[6] called a news article fake news if it is propagated on the internet at a large scale to either use it as a joke or to influence public political ideologies.

Furthermore, in study [1], the authors classified fake news into three groups: serious fabrications, large scale hoaxes, and humorous fakes. The authors failed to provide specific reasons for using only these three categories. However, they shed light on the characteristics of each category and how to differentiate these three categories from each other. The same study also highlighted the limitations of datasets to perform fake news detection task. In addition to this, there is another type of fake news that is known as "clickbait", where the intent is to attract a consumer to click on a given link.

We propose the definition of fake news and fake news detection based on the previous works and analysis to define this term as follows:

- **Fake News:** Fake news is a factually incorrect news article and provides misleading information with the intent to deceive the readers making them believe it is true.
- **Fake News Detection:** For a given news article (unannotated) α, where $\alpha \in N$ (α is one news article out of N news article), an automatic fake news detection algorithm assigns score $S(\alpha) \in [0, 1]$ indicating the extent to which $S(\alpha)$ is assumed to be a fake news article.

For instance, if $S(\hat{\alpha})$ ¿ $S(\alpha)$, then it implies that $\hat{\alpha}$ has a higher tendency to be a fake news b b article. A threshold γ can be defined such that the prediction function $F: N \rightarrow$ [fake, not fake] is:

$$F(N) = \begin{cases} fake, & if \ \ S(\alpha) \in \gamma, \\ not\ fake, & \text{otherwise} \end{cases}$$

[5] https://en.oxforddictionaries.com/word-of-the-year/word-ofthe- year-2016

[6] https://dictionary.cambridge.org/us/dictionary/english/fake-news

3 Overview and Grouping of Fake News Datasets

In the era of artificial intelligence, data are essential assets to automate various computer-based tasks. In this view, automatic fake news detection is a striking but new area for the research community. However, fake news is a worldwide phenomenon and appears in all countries and in multiple languages. Several studies focusing on the English language achieved significant advancement and produced a few benchmark datasets in English.

Nevertheless, only limited sources in the form of datasets are available for poor resource-languages due to various reasons. The term "Fake news" is divided into many subcategories; this is why most publicly available datasets differ from one another and cannot be re-used in research with a slightly different focus within the broader "fake news" domain, as discussed in Introduction. Thirdly, data collection and annotation is a time-consuming and expensive task. Therefore, it is challenging to design new annotated datasets for fake news detection.

In this study, we primarily focus on non-English datasets available for automated fake news identification. Moreover, the inadequacy of fake news datasets is a major stumbling block, especially in automatically identifying fake news across multiple languages. We analyzed various datasets in related works that focused on assessing the integrity of news articles, Twitter postings, and YouTube comments.

We categorized the datasets into two sub-categories, (i) mainstream media articles datasets, (ii) social network posts datasets.

3.1 Mainstream Media Articles Datasets

Mainstream media articles datasets comprise on lengthy texts and news articles that can be seen in traditional e-newspapers, containing approximately 400 to 700 words. Below we describe three such datasets. The details of these datasets are presented in Tables 3 and 4.

3.1.1 Bend the Truth

In the recent study [6], a fake news dataset "Bend the Truth" is presented in the Urdu language that contains two types of news, (i) real news and (ii) fake news. The dataset covered five news topics, sports, entertainment, business, technology, and health.

The authors collected 500 real news from five different domains. Each domain is contributing 100 news in the proposed corpus. Since there are five categories of news, so in total, the dataset contains 500 real news. The real news in the Urdu language was manually collected from 16 news stream websites and four different countries from January 2018 to December 2018. These countries are the USA, UK, India, and Pakistan. The study provides a comprehensive description of real news collection and annotation methodology.

3.1.2 BanFakeNews

Study [7] introduces the fake news dataset BanFakeNews in Bangla language. Bangla language is the official language in Bangladesh and has more than 230 million native speakers and is spoken widely in Bangladesh and India.

The dataset contains three categories of news, (i) click baits, (ii) satirical, (iii) real, and fake news with their headlines. The dataset considered real and fake news in one category.

The dataset includes 242 topics that were further classified based on similar categories into 12 news domains (sports, politics, crime, technology, etc.). 48,678 real news were collected (till March 08, 2020) from 22 mainstream news websites in Bangladesh, each real news on average had 271.16 words. In contrast, only 1,299 fake news articles were collected from www.jaachai.com and www.bdfactcheck.com, and each fake news contained, on average, 276.36 words. The meta-data such as the source of the news article, publication time, news topic, and the relation between headline and article were provided for only 8500 news articles due to the task complexity. The assessment or labelling was done by undergraduate students who have a background in Computer Science and Engineering and Software Engineering who manually annotated the source of the news and tagged the relation between headline and article as "related" or "non-related."

3.1.3 Persian Stance Dataset

A recent study [8] contributed the first stance detection dataset in the Persian language to study article-claim stance and headline-claim stance classification tasks. The study created a web-based tool (the stance detection system)[7] to collect claims and news articles. Stance detection is defined as investigating what other mainstream news organizations publish about a piece of news, i.e., understanding what these organizations claim about that specific news [9] and on Twitter [10]. The authors defined a claim as news published by another news agency, and that claim was used to check the stance of the body of the news article. The same web-based tool was used to annotate the news article's stance against the claim made and to find the integrity of each claim. All the claims were collected and created from rumors and news headlines using two Iranian websites (Fakenews and Shayeaat).

The study provides 2,124 news articles and textual claims (news headlines) to the stance detection system that annotates the news article's stance against the claim into four groups, agree, disagree, discuss, and unrelated. The research reported that Fakenews and Shayeaat assemble rumors (headlines) from different sources and manually check the news articles' credibility.

[7] https://github.com/majidzarharan/persian-stanceclassification

3.1.4 Arabic Fake Rumours Dataset

A recent study [11] analyzed fake news in the context of rumor detection. The study presented a corpus in the Arabic language for the automatic fake rumor detection task. It considered rumor detection as a binary classification problem. The authors focused on three Arab celebrities, Fifi Abdu (an Egyptian dancer), Abdelaziz Bouteflika (the former Algerian president), and Adel Imam (an Egyptian comedian).

To retrieve the YouTube comments, YouTube API was used to collect the comments associated with these three personalities' death rumors. The authors considered YouTube comments as potential rumors to address automatic fake new detection tasks. Furthermore, it is essential to mention that the term "fake news" authors didn't mean "news articles".

The researchers used keywords associated with rumors to data-mine fake stories (comments) related to these personalities' death. For example, the study used keywords such as "*Algerian president dies*", "*yes death*," "*Bouteflika death*," to retrieve rumors related to the death of the Algerian president. Similarly, they mined comments associated with the death of an Egyptian comedian using the keywords "*adel imam dies*", "*Adel die*," and "*Allah yerhamo*." Likewise, the Egyptian dancer's death rumours were collected using "*Fifi died*", "*True news*", and "*Allah yarhemak*". If one of these keywords appeared in a comment, then the comment was tagged as a rumor. In the contrary case, the comment was labelled as the no-rumor if the comment contained the celebrity's name and did not mention death. In the end, 343 rumors and 3092 non-rumors were included in the final detest.

3.1.5 Czech, Polish, and Slovak Fact-checking Datasets

The study [12] presented datasets in three languages: Czech, Polish, and Slovak to address fact-checking tasks in West Slavic Languages. The Czech dataset contained 9082 claims of politicians that were annotated by expert annotators in four classes: (i) False, (ii) True, (iii) Unverifiable, and (iv) Misleading. Likewise, the dataset also contained 2835 politicians claims in Polish and 12554 politicians claims in Slovak language. However, the authors did not mention whether the claims of the same politicians were used in three languages. The authors downloaded the claims from websites in April 2018.

3.1.6 DANFEVER (Danish) Dataset

A new dataset in Danish has been proposed [13] for the claim detection task. The dataset contained 6,407 claims in Danish language that are manually annotated into three classes: (i) Supported claims, (ii) Refuted, and (iii) NotEnoughInfo claims. Different sources, such as Danish Wikipedia and Den Store Danske have been used for claims generation.

3.1.7 FactCorp (Dutch) Dataset

The author in [14] proposes to investigate fact-checks from a corpus linguistic approach. This study aims to understand and learn more about the extent and substance of factchecks, additionally more about that how science appears (incorrectly) in the news and how to behave from the science communication perspective. A FactCorp contains a 116 million words corpus and 1974 fact-checks reported from three different Dutch newspapers. The author of this study has done different analyses as a result, including keyword, qualitative content elements, and rhetorical moves analysis. According to these analyses, they show that FactCorp allows a wealth of possible applications, emphasizing the need to develop such resources.

In the study [15], the researcher argues that network analysis's persistent disregard for conflict leads to enormous conclusions on heated arguments. The researcher in this study introduces a method for incorporating negative user-to-user contact into online arguments by analyzing signed networks with negative and positive relationships. The 'black Pete' debate on Twitter is analyzed on the annual Dutch celebration in this study. The dataset containing 430,000 tweets is used, and ML and NLP-based solutions are applied to identify the stance of users in online debate and the interaction between users. The results demonstrate that some groups are targeting each other, while others appear to be scattered across Twitter.

3.1.8 DEAP-FAKED (Estonian)

Recently, hoaxes and fake news spreading on social media have attracted more attention, especially in politics and healthcare (COIVD-19). For the detection of fake news on social media platforms, a Deep-Facked framework has been proposed in [16]. A deep-Facked approach is the combination of NLP-based and GNN-based techniques. Two different publicly available databases containing articles from the healthcare, politics, business, and technology domain are used in the Deep-Facked approach.

3.1.9 Cresci-2017 (Finnish)

The goal of this study is to investigate the influence of bots on Finnish politics Twitter, using a dataset of accounts that follow important Finnish politicians before the 2019 parliamentary election.

In this social media life, opinion mining and sentiment analyses are important tasks, e.g. when stipulating fake and hoax news. In this study, the author [17] addressed this deficiency by presenting a 27,000-sentence data set that was annotated with sentiment polarity by three native annotators separately. They used the same three annotators throughout the data set, which gives the unique opportunity to study annotator behavior across time. Furthermore, they examine their inter-annotator agreement and present two baselines to verify the utility of the dataset.

A new dataset in the Finnish language on rumor detection is presented [18]. In this study, the author assesses two different models based on LSTM and two models based on BERT. Because the models were trained on tiny and biased corpora, these findings suggest that additional work is needed for pre-trained models in the Finnish language.

3.1.10 Fake.Br Corpus (Portuguese)

The study [19] proposed a news corpus for fake news detection in Brazilian Portuguese (PT). The dataset contained 7,200 news, which was manually labelled and contained an equal number of fake news (3,600) and true news (3,600) articles. The news articles were retrieved from January of 2016-2018.

3.1.11 Partelet Corpus of Propaganda Texts (Hungarian)

A digital Hungarian language database of communist propaganda text named as Partelet has been presented in [20]. This paper serves two purposes: first, to provide a general overview of the corpus compilation method and basic statistical data, and second, to demonstrate the dataset utility using two case studies. Results illustrate that the proposed corpus offers a unique potential for doing research on Hungarian propaganda speech as well as assessing changes in this language using computer-assisted approaches over 35 years.

Recent developments in the field of semantic encoding demonstrate significant progress and call attention to such strategies. These approaches' main purpose is to convert human-written natural language text into a semantic vector space. The train and execution of a semantic encoder for the Hungarian language are discussed in this study [21]. Since Hungarian is not a commonly spoken language, the number of linguistic available resources is restricted. Although the method described here is used with the Hungarian language, it may be used in any small or medium-sized language.

3.1.12 Spanish Fake News Corpus (Spanish)

The Study [22] introduced the first Spanish corpus to investigate and analyze the style-based fake news detection in the Spanish language. The dataset included an overlap of distinct news topics and classes containing true news (491) and fake news (480). The news was manually tagged and obtained from January to July of 2018 from several websites.

3.1.13 Fake News Polarization (Italian)

The study [23] aims at disseminating fake news on Facebook pages. The dataset consisted of 333,547 news officials and 51,535 fake news on Facebook posts which were further divided into "entities" (i.e., news topics). The data was collected in July-December 2016 exclusively by means of the Facebook Graph API.

3.1.14 CT-FAN-21 Corpus (Bulgarian, Turkish, Spanish)

The research [24] investigated into misleading news articles in European languages including Spanish, Turkish and Bulgarian. They tested out their CT-FAN-21 corpus on 900 trained and 354 test articles submitted by 27 teams for Task 3A, 20 teams for task 3B assigned for 1) 3A; topical domain detection of news articles and 2) 3B; multi-class fake news detection.

3.1.15 FakeDeS: Spanish dataset for Fake news

Datasets for fake news in Spanish are available though not in abundance [22,25,26]. In 2021 IberLeF released [25] the second iteration of the fake news challenge named "FakeDeS". The first edition was released [26] in 2020 named as "MEX-A3T". The second edition of the dataset used "MEX-A3T" dataset as the training set and created a new test dataset with data related to COVID-19. The topic distribution of the dataset comprised of science, society, health, politics, entertainment, education, economy and sport.

The dataset was compiled using fact-checking websites and newspapers. The second edition of the dataset has 970 (491 True, 480 Fake) training files and 572 (286 True, 286 Fake) test files, while the first edition contains 676 (338 True, 338 Fake) training files and 295 (153 True, 142 Fake) test files.

3.1.16 Fake News Dataset for Slovak

A dataset in the Slovak language is presented [27] with a focus on home news, world news, and economic news. However, in this paper, we discuss the extension of this dataset introduced in the paper [28] with deep learning baselines. The data was obtained from multiple news sources targeted at a specific domain of Slovak home news. The targeted news was annotated with labels 0 (Fake News) or 1 (True News) using konspiratori.sk (database for news credibility) at the initial stage and then manually verifying it. The final distribution shaped into 11,410 (training), 3,803 (validation), and 3,804 (test) articles respectively.

Table 1

Mainstream / Social Media Articles Datasets in Asian languages.

		Fake News Datasets			
Name	Language	Size	Main Input	Task	Annotation
Bend the Truth	Urdu	Real: 500 Fake: 400	News Articles	Fake News Detection	Professional Journalists
BanFake News	Bangla	Real: 48678 Fake: 1299	News Articles	Fake News Detection	Trained annotators
Persian Stance	Persian	Articles: 2124 Claims: 600	News Articles	Stance Detection	Trained annotators
Arabic Rumours	Arabic	Rumours: 343 Non-rumours: 3092	YouTube Comments	Fake Rumours Detection	Trained annotators

4 Comparison of Methods for Fake News Detection

Several studies have been conducted to understand and investigate multiple ways to automatically differentiate fake news from real news. This study analyzes the essential research in which we find work related to fake news detection tasks in Asia. Table 5 shows the proposed techniques in low-resource Asian languages. We limited the scope of this study by only analyzing the methods used in Asian languages, and would like to work on methods for European Languages in our future work. To the best of our knowledge, no prior research has been done to analyze automatic fake news detection systems in lowresource Asian languages. We categorize them into two subsections: Non-Neural Network Techniques and Neural Network Techniques.

4.1 Features for Fake News Detection

There are two main methods to tackle the fake news detection task (i) analyzing the content of the news article and (ii) analyzing the context of the news article. In the first method, a recent study comprehends the fake news detection phenomenon; it reveals that fake news tends to spread faster than real news [2].

In contrast, in the second method, linguistic features differentiate fake news articles from real news articles, i.e., discussing typical patterns. For example, in recent studies, linguistic features have been used to perform automatic fake news identification task [3, 6-8, 11, 29]. It is essential to highlight that most of the studies on fake news detection lack concrete guidelines on what features are necessary for the task. This is significant to know because these studies use specific data and feature sets to train classifiers. Moreover, the studies also lack details about why fake news is classified as fake news and the classifiers' decision behind classifying fake news articles.

Table 2
Mainstream / Social Media Articles Datasets in European languages

Fake News Datasets					
Name	Language	Size	Main Input	Task	Annotation
Czech fact-checking	Czech	True: 5669 False: 1222 Unverifiable: 1343 Misleading: 848	Claims of politicians	Fact-Checking Detection	Trained annotators
Polish fact-checking	Polish	True: 1761 False: 648 Unverifiable: 113 Misleading: 313	Claims of politicians	Fact-Checking Detection	Trained annotators
Slovak fact-checking	Slovak	True: 7987 False: 1670 Unverifiable: 1751 Misleading: 1146	Claims of politicians	Fact-Checking Detection	Trained annotators
	Slovak	True: 9979	News articles	Fake news	

Slovak Fake news		Fake: 9048		Detection	konspiratori.sk annotators
DANFEVER	Danish	Supported Claims: 3,124 Refuted Claims: 2,156 Notenoughinfo Claims: 1,127	Text annotated as claims	Claim Verification	Trained annotators
FactCorp	Dutch	Fact-Checks: 1,974	Dutch news	Fact-Checking	Trained annotators
Deep-Faked	Estonian	True: 9,129 Fake: 5,058	News article	Fake News Detection	Trained annotators
Cresci-2017	Finnish	Bots Account: 3000 Genuine Account: 3000	Claims of politicians	Identify the bot account	Trained annotators
Suomi24	Finnish	Sentences: 27,000	Social website	Fake news detection	Trained annotators
Fake.Br Corpus	Portuguese	True: 3600 Fake: 3600	News article	Fake News Detection	Trained annotators
Spanish Fake News Corpus	Spanish	True: 491 Fake: 480	News website	Fake News Detection	Trained annotators
FakeDeS	Spanish	True: 777 Fake: 766	News articles Fact-Checking Websites	Fake news Detection	Trained annotators
Fake News Polarization	Italian	Official: 333,547 Fake: 51,535	Facebook	Fake News Detection	Trained annotators
Partelet	Hungarian	Text Tokens: 13,185,200	Partelet journel	Propaganda Detection	Trained annotators
HoaxItlay	Italian	News: 37k	Twitter streaming API	Fact-checking/Disinformation	Trained annotators
CT-FAN-21 Corpus	Bulgarian, Turkish, Spanish	False:111 True: 65 Partially False: 138 Others: 40	News Articles	Fact-Checking Detection	Trained annotators

4.2 Non-Neural Network Techniques

Most studies used linguistic features to adders to the automatic fake news detection task. Researchers have been using linguistic features such as N-grams, syntactic features such as POS tags, and semantic features like text entailment and metadata (the headline's lengths and the body of news articles) to implement fake news classification on the benchmark datasets.

A recent study [11] presented a fake news corpus in the Arabic language. The study focused on fake news detection in their dataset using three machine learning classifiers. The experiments were performed with the train test split ratio 70/30, respectively using N-grams features, namely, word N-grams where N varies from uni-gram to tri-gram, with term frequency-inverse document frequency (TF-IDF) weighting scheme. Three supervised machine learning algorithms have been used, such as Decision Tree (DT), Support Vector Machine (SVM) with linear kernel, and

Multinomial Naıve Bayes (MNB) classifiers. The study reported that the SVM achieved the highest accuracy of 0.95 compared to other classifiers in classifying rumors in YouTube comments.

A similar study [6] on fake news detection in Bangla language, used linguistic features such as word N-grams (n=1,2,3) and character N-grams (N=3,4,5) along with the normalized frequency of different POS tags. The study removed stop words and punctuation in the pre-processing phase. Additionally, the research utilized metadata (the headline's lengths and the body of news articles) and punctuation frequency as features. Furthermore, to convert words into vectors, the study used TFIDF as the frequency weighting scheme. For the classification, linguistic features were supplied into a linear Support Vector Machine (SVM), Random Forest (RF), and a Logistic Regression (LR) model. For the experiments, the split data ratio was 70/30 train-to-test, respectively. The SVM model outperformed other classifiers and achieved 0.89 F1-score and 0.90 F1-score using character 3-gram weighted frequencies and all linguistic features.

Likewise, research [7] investigated automatic fake news detection in the Urdu language. The study classified news articles using combinations of different N-gram types (words, characters, and functional words). It showed that the combinations provide better results than N-grams of a single type. The experiments used five N-gram frequency weighting schemes (TFIDF, normalized, log-entropy, binary, TF) and seven different machine learning classifiers. The study provided a comprehensive analysis of different feature sets used in the experiments. Lexical features with N-gram size 1 to 3 obtained better results compared with 4,5,6. Finally, the study reported that AdaBoost outperformed other classifiers by getting 0.86 F1-real and 0.90 F1-fake scores. The authors also reported a balanced accuracy of 0.88.

Previous study [8] on stance classification in the Persian language used three machine learning classifiers. The study reported that two feature types, such as bag-of-words representation (BoW) and TFIDF, were used. Eventually, the study showed that Random Forest achieves an accuracy of 0.69 in recognizing the stance of headline-claim.

Study [30] used term frequency (TF) weighting scheme and Naive Bayes classifier. The study reported that Native Bayes obtained 0.78 accuracy to identify hoax news using the Indonesian language.

4.3 Neural Network Techniques

In recent studies, Deep Learning techniques have been widely used in different tasks such as text classification and generation tasks. These techniques, namely, Neural Networks, achieved significant results and showed impressive performance in solving various NLP-related tasks. Different neural network architectures such as the Convolutional Neural Network, Recurrent Neural Network, and Transformer all

need much data to learn hidden patterns. These techniques obtain better results than linguistic feature-based methods.

The study [6] used semantic features to differentiate fake articles from real news articles. The experiments were conducted based on two types of word embeddings (vector representations of each news article) Fast text word embeddings [31] (300-dimensional word vectors) and Word2Vec [32] (100dimensional word vectors). The research used 256 different kernels, having to vary in size lengths from 1 to 4. The global max pool and the average pool were used in the pooling layer. For the activation function, ReLU [18] activation function was used.

In the prior study [8], the study focused on both tasks, headline-claim stance classification and articleclaim stance classification. The research was based on deep learning techniques, particularly the stack LSTM architecture using pre-trained 300-dimensional word embedding. All the experiments were performed with the deep learning library Keras[8]. 100-word embedding features are fed to two LSTMs to consider word sequences. The neural network has three dense layers, and each layer contained 300 neurons. In the last layer of the neural network, the softmax activation function is used to obtain the final output. The headline text was fed as input to the neural network. In addition to this, the study investigated two tasks, headline-claim stance detection, and article-claim stance detection. Thus, the authors reported that stackLSTM did not perform well in recognizing the headline-claim (in this task, the Random Forest classifier outperformed deep learning). However, the study illustrated that stackLSTM exceeded other techniques by obtaining 0.72 accuracy in finding the article-claim stance.

We observed that Deep Learning methods are not so prominent in addressing the automatic fake news detection task, especially in European and Asian low-resource languages, for several reasons.

First of all, the inadequacy of available sources in the form of datasets. Secondly, creating datasets is a time-taking task, but it requires financial support, which is a challenging part most of the time. Thirdly, the research community in Asian and European low-resource languages is minimal. Finally, the available datasets are small in size, therefore, not sufficient to train Deep Learning techniques to tackle automatically identifying fake news. Table 3 and Table 4 show the size of the available datasets in Asian and European languages for differentiating fake news.

5 Popular Evaluation Metrics

This section explains various evaluation approaches and metrics used to assess the performance of different fake news detection systems. Fake News detection is

[8] https://keras.io

almost universally approached as a classification task, either as a binary classification (more often) or a multi-label task. A binary classification problem has the goal to classify the instances of a given set into two categories. For example, in fake news detection as a binary classification task, the goal is to differentiate fake news from real news. However, if a problem is concerned with more than two groups, it is a multi-class task. For example, detecting a stance between a news headline and the whole text of the news article is a multi-class task since there are more than two labels involved, such as *agree disagree*, *unrelated*, etc.

In general, studies on fake news detection used different metrics to evaluate the performance of the presented methods, creating some inconvenience compared to the results among different works. For example, we observe studies reporting precision, recall, accuracy, F1 score, and ***ROC-AUC*** to evaluate the performance of various models trained on balanced datasets. In contrast to the balanced datasets, multiple studies have reported precision, recall, along with the F1 score for the fake class and ROC-AUC to examine the overall system quality and evaluate the model performance. In addition to this, some studies also calculated ***Micro-F1*** scores for highly unbalanced datasets. Finally, we noticed that most of the studies used the F1 score to measure the model's performance since most of the datasets are unbalanced.

Discussion and Conclusions

This work provides the first overview of various publicly available datasets for automatic fake news detection in Asian and European languages. Most of which are poor resource languages, providing comparative statistics on their sizes, grouping them by the length and source of content news, and surveying dataset annotation procedures. We have also surveyed the approaches used in the studies on fake news detection in low-resource languages and grouped them into studies using traditional machine learning and neural network approaches. We note that working on Asian languages with more resources, notably, Chinese, demonstrates wider adoption of neural networks and achieves better results with those. Finally, we provide a brief overview of the evaluation metrics used to report fake news detection performance. It is important to note that due to a large variety of metrics available, some studies choose to report different metrics than others, which leads to difficulties in comparison among studies.

Although low-resource languages have limited resources and a plethora of challenges, these languages lack expert-based fact-checking websites, i.e., PolitiFact[9] or FactCheck[10], which provide the services of fact-checking. However, tackling fake news tasks in low-resource languages can decrease the detrimental consequences of fake news globally. Multiple studies reported fact extraction [33] and relation extraction [34] in English, but this research still needs attention in low resource languages. Unless these techniques are enhanced, robust Knowledge Bases

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

[10] https://www.factcheck.org/

(KB) cannot be created for fact-checking, to eliminate fundamental issues like redundancy [35], invalidity [36], conflicts [37], unreliability [38] and incompleteness [39,40] in fake news. Style detection in fake news identifies the intent of the content and the style of text changes across languages and domains. The textual style also evolves with time, and hence more attention should be put to create solutions needed for style-based fake news detection in low-resource languages. While targeting the fake news, it is also important to analyse the check worthiness [41] of the news, which can be analysed by the potential of influence [42], user reputation [43], historical likelihood of the topic and title verification [44] of the content. This improved the efficiency of fake news that can have a mass impact on society and unfortunately, most of these topics need emphasis by the research community of the low resource languages.

We also note that one of the crucial difficulties faced while assembling the fake news datasets was finding the datasets focusing solely on fake news detection. In many cases, datasets are purposed for multiple tasks that are only indirectly related to the fake news detection problem, particularly, the datasets annotated for rumor detection, stance detection, and differentiating between fake news sub-classes satire.

Future Opportunities and Research Directions

Future research on fake news detection might extend datasets' explanations in most major languages used in Natural Language Processing to study fake news from various perspectives. We also want to track attention to explainable machine learning algorithms to solve automatically fake news detection. The explainable algorithm can point out important features and the classifiers' decision behind classifying news articles as fake or not fake. This can significantly improve the performance of existing fake news detection systems. We also want to investigate the performance of the machine learning algorithm trained on one dataset and test it on a different dataset across languages. For example, we want to study whether training on stance detection dataset and testing on rumor detection can provide better performance. In future research, we also want to analyze the methods used in European languages.

Acknowledgment

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of the CONACYT, Mexico and grants 20211784, 20211884, and 20211178 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico.

References

[1] Rubin, V., Chen, Y. & Conroy, N. Deception detection for news: three types of fakes. Proceed. of the Association for Information Science and Technology, V.25, N.1, 2015, pp.1-4.

[2] Lazer, D., Baum, M., Benkler, Y., Berinsky, A., Greenhill, K., Menczer, F., Metzger, M., Nyhan, B., Pennycook, G., Rothschild, D. & Schudson M. The science of fake news. Science. V.359, N.6380, 2018, pp.1094-1096.

[3] Perez-Rosas, V., Kleinberg, B., Lefevre, A. & Mihalcea, R. Automatic detection of fake news. ArXiv Preprint ArXiv:1708.07104, 2017.

[4] Fallis, D. A functional analysis of disinformation. Int. Conf. 2014 Proceedings, 2014, pp.621-627.

[5] Tandoc Jr, E., Lim, Z. & Ling, R. Defining "fake news" A typology of scholarly definitions. Digit. Journal., V.6, N.2, 2018, pp.137-153.

[6] Amjad, M., Sidorov, G., Zhila, A., Gómez-Adorno, H., Voronkov, I., & Gelbukh, A. Bent the truth: A benchmark dataset for fake news detection in Urdu and its evaluation. J. Intell. Fuzzy Syst., V.39, 2020, pp.2457-2469.

[7] Hossain, M., Rahman, M., Islam, M. & Kar, S. Banfakenews: A dataset for detecting fake news in Bangla. ArXiv Preprint ArXiv:2004.08789, 2020.

[8] Zarharan, M., Ahangar, S., Rezvaninejad, F., Bidhendi, M., Jalali, S., Eetemadi, S., Pilehvar, M. & Minaei-Bidgoli, B. Persian Stance Classification Dataset, 2019.

[9] Pomerleau, D. & Rao, D. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. Fake News Challenge, 2017.

[10] Sadr, M., Mousavi Chelak, A., Ziaei, S. & Tanha, J. A predictive model based on machine learning methods to recognize fake persian news on twitter. Int. J. Nonlinear Anal. Appl., V.11, 2020, pp.119-128.

[11] Alkhair, M., Meftouh, K., Smaïli, K., & Othman, N. An Arabic corpus of fake news: Collection, analysis and classification. Int. Conf. on Arabic Language Processing, 2019, pp.292-302.

[12] Přibáň, P., Hercig, T., & Steinberger, J. Machine learning approach to fact-checking in West Slavic languages. Proceed. of the Int. Conf. on Recent Advances in Natural Language Processing, 2019, pp.973-979.

[13] Nørregaard, J. & Derczynski, L. DANFEVER: Claim verification dataset for Danish. Proceed. of the 23rd Nordic Conference on Computational Linguistics, 2021, pp.422-428.

[14] Meulen, M. & Reijnierse, W. FactCorp: A Corpus of Dutch Fact-checks and its Multiple Usages. Proceedings Of The 12th Language Resources And Evaluation Conference, 2020, pp. 1286-1292

[15] Keuchenius, A., Tornberg, P. & Uitermark, J. Why it is important to consider negative ties when studying polarized¨ debates: A signed network analysis of a Dutch cultural controversy on Twitter, PloS One, 2021.

[16] Mayank, M., Sharma, S. & Sharma, R. DEAP-FAKED: Knowledge Graph based Approach for Fake News Detection. ArXiv Preprint ArXiv:2107.10648, 2021.

[17] Linden, K., Jauhiainen, T. & Hardwick, S. FinnSentiment–A Finnish Social Media Corpus for Sentiment Polarity Annotation. ArXiv Preprint ArXiv:2012.02613, 2020.

[18] Hämäläinen, M., Alnajjar, K., Partanen, N., & Rueter, J. Never guess what I heard... Rumor Detection in Finnish News: a Dataset and a Baseline.ArXiv Preprint ArXiv:2106.03389, 2021.

[19] Monteiro, R., Santos, R., Pardo, T., De Almeida, T., Ruiz, E. & Vale, O. Contributions to the study of fake news in Portuguese: New corpus and automatic detection results. Int.Conf. on Computational Processing of the Portuguese Language, 2018, pp.324-334.

[20] Kmetty, Z., Vincze, V., Demszky, D., Ring, O., Nagy, B. & Szabo, M. P´art´elet: A Hungarian corpus of propaganda´ texts from the Hungarian socialist era. Proceedings Of The 12th Language Resources And Evaluation Conference, 2020, pp. 2381-2388.

[21] Kantor, A., Kiss, A. & Grad-Gyenge, L. Semantic Encoder Tasks for the Hungarian.´

[22] Posadas-Durán, J., Gómez-Adorno, H., Sidorov, G. & Escobar, J. Detection of fake news in a new corpus for the Spanish language. J. Intell. Fuzzy Syst., V.36, N.5, 2019, pp.4869-4876.

[23] Vicario, M., Quattrociocchi, W., Scala, A. & Zollo, F. Polarization and fake news: Early warning of potential misinformation targets. ACM Transactions On The Web (TWEB), V.13, 2019, pp.1-22.

[24] Shahi, G., Struß, J. & Mandl, T. Overview of the CLEF-2021 CheckThat! lab task 3 on fake news detection. Working Notes Of CLEF, 2021.

[25] Gómez-Adorno, H., Posadas-Durán, J., Bel-Enguix, G. & Capetillo, C. Overview of FakeDeS at IberLEF 2021: Fake News´ Detection in Spanish Shared Task. Proces. Leng. Nat., V.67, 2021, pp.223-231.

[26] Aragón, M. E., Jarquín-Vásquez, H. J., Montes-y-Gómez, M., Escalante, H. J., Pineda, L. V., Gómez-Adorno, H., Posadas-Durán, J. & Bel-Enguix, G. Overview of MEX-A3T at IberLEF 2020: Fake news and aggressiveness analysis in Mexican Spanish. IberLEF@ SEPLN, 2020, pp.222-235.

[27] Sarnovsky, M., Maslej-Kresnova & Hrabovska, N. Annotated dataset for the fake news classification in Slovak language. 18th Int. Conf. on Emerging ELearning Technologies and Applications (ICETA), 2020, pp.574-579.

[28] Ivancová, K., Sarnovský, M., & Maslej-Krcšňáková, V. Fake news detection in Slovak language using deep learning techniques. 2021 IEEE 19th World

Symposium on Applied Machine Intelligence and Informatics (SAMI), 2021, pp.000255000260.

[29] Patwa, P., Bhardwaj, M., Guptha, V., Kumari, G., Sharma, S., Pykl, S., Das, A., Ekbal, A., Akhtar, M.S. & Chakraborty, T. Overview of constraint 2021 shared tasks: Detecting English covid-19 fake news and Hindi hostile posts. In International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer, Cham, 2021, pp. 42-53.

[30] Pratiwi, I., Asmara, R. & Rahutomo, F. Study of hoax news detection using naive bayes classifier in Indonesian language. 2017 11th Int. Conf. on ICTS, 2017, pp.73-78.

[31] Grave, E., Bojanowski, P., Gupta, P., Joulin, A. & Mikolov, T. Learning word vectors for 157 languages. ArXiv Preprint ArXiv:1802.06893, 2018.

[32] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process., 2013, pp.3111-3119.

[33] Thorne, J., Vlachos, A., Christodoulopoulos, C. & Mittal, A. Fever: A large-scale dataset for fact extraction and verification. ArXiv Preprint ArXiv:1803.05355, 2018.

[34] Yu, B., Zhang, Z., Liu, T., Wang, B., Li, S. & Li, Q. Beyond word attention: Using segment attention in neural relation extraction. IJCAI, 2019, pp.5401-5407.

[35] Altowim, Y., Kalashnikov, D. & Mehrotra, S. Progressive approach to relational entity resolution. Proceed. of the VLDB Endowment, V.7, N.11, 2014, pp.999-1010.

[36] Hoffart, J., Suchanek, F., Berberich, K. & Weikum, G. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artif. Intell., V.194, 2013, pp.28-61.

[37] Kang, B. & Deng, Y. The maximum Deng entropy. IEEE Access, V.7, 2019, pp.120758-120765.

[38] Ye, J. & Skiena, S. Mediarank: Computational ranking of online news sources. Proceed. of the 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, 2019, pp.2469-2477.

[39] Kazemi, S. & Poole, D. Simple embedding for link prediction in knowledge graphs. ArXiv Preprint ArXiv:1802.04868, 2018.

[40] Shi, B. & Weninger, T. Open-world knowledge graph completion. Proceed. of the AAAI Conf. on Artif. Intell., V.32, 2018.

[41] Hassan, N., Arslan, F., Li, C. & Tremayne, M. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. Proceed. of

the 23rd ACM SIGKDD Int. Conf. on Data. Min. Knowl. Discov., 2017, pp.1803-1812.

[42] Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. Science, V.359, N.6380, 2018, pp.1146-1151.

[43] Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. Commun. ACM., V.59, N.7, 2016, pp.96-104.

[44] Ghanem, B., Rosso, P. & Rangel, F. Stance detection in fake news a combined feature representation. Proceed. of the First Workshop on Fact Extraction And VERification (FEVER), 2018, pp.66-71.