Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2862–2871
Marseille, 11–16 May 2020

BanFakeNews: A Dataset for Detecting Fake News in Bangla

Md Zobaer Hossain[†♣], Md Ashraful Rahman[†♣], Md Saiful Islam[♣], Sudipta Kar[♠]
[♣] Shahjalal University of Science and Technology, Sylhet, Bangladesh
{zobaer37, ashraful54}@student.sust.edu, saiful-cse@sust.edu
[♠] University of Houston, Texas, USA
skar3@uh.edu

Abstract

Observing the damages that can be done by the rapid propagation of fake news in various sectors like politics and finance, automatic identification of fake news using linguistic analysis has drawn the attention of the research community. However, such methods are largely being developed for English where low resource languages remain out of the focus. But the risks spawned by fake and manipulative news are not confined by languages. In this work, we propose an annotated dataset of $\approx 50K$ news that can be used for building automated fake news detection systems for a low resource language like Bangla. Additionally, we provide an analysis of the dataset and develop a benchmark system with state of the art NLP techniques to identify Bangla fake news. To create this system, we explore traditional linguistic features and neural network based methods. We expect this dataset will be a valuable resource for building technologies to prevent the spreading of fake news and contribute in research with low resource languages. The dataset and source code are publicly available at https://github.com/Rowan1697/FakeNews.

Keywords: Fake news, Bangla, Low Resource Language

1. Introduction

Articles that can potentially mislead or deceive readers by providing fabricated information are known as fake news. Usually fake news are written and published with the intent to damage the reputation of an agency, entity, or person[1]. The popularity of social media (e.g. Facebook), easy access to online advertisement revenue, increased political divergence have been the reasons for the spread of fake news. Hostile government actors have also been implicated in generating and propagating fake news, particularly during elections and protests (TUFEKCI, 2018) to manipulate people for personal and organizational gains.

Impact of Fake news is creating havoc worldwide. During the 2016 US election, 25 percent of the Americans visited a fake news website in a six-week period of election which has been hypothesized as one of the issues that influenced the final results (Grave et al., 2018). In Bangladesh, the 2012 Ramu incident is an exemplary event where almost 25 thousand people participated in destroying the Buddhist temples on the basis of a Facebook post from a fake account (Manik, 2012). About 12 Buddhist temples and monasteries and 50 houses were destroyed by the angry mob. Fake news that contains blasphemy can easily repeat these types of incidents where people are very sentimental to their religions.

To tackle fake news there are some dedicated websites like www.politifact.com, www.factcheck.org, www.jaachai.com where they manually update potential fake news stories published in online media with logical and factual explanations behind the news being false. But these websites are not capable enough as they cannot respond quickly to any fake news event. Recently computational approaches are also being used to fight against the menace of fake news. Long et al. (2017) tried multi-perspective speaker profiles to detect fake news, where Yang et al. (2017) has used linguistic features to detect satirical news. Besides Karadzhov et al. (2017) has proposed a fully automated fact-checking model using external sources to check the claim of news stories. To detect fake news in social media, Dong et al. (2019) has used deep two-path semi-supervised learning. However, these works have been done only on news published in the English. As of now, around 341 millions of people in the world speak Bangla and it is the fifth language in the world in terms of number of speakers [2]. But to the best of our knowledge, there is no resource or computational approach to tackle the risk of fake news written in Bangla, which can negatively affect this large group of population.

In this paper, we aim at bridging the gap by creating resource for detecting fake news in Bangla. Our contributions can be summarized as follows:

- We publicly release an annotated dataset of $\approx 50K$ Bangla news that can be a key resource for building automated fake news detection systems.
- We develop a benchmark system for classifying fake news written in Bangla by investigating a wide rage of linguistic features. Additionally, we explore the feasibility of different neural architectures and pre-trained Transformers models in this problem.
- We present a thorough analysis of the methods and results and provide a comparison with human performance for detecting fake news in Bangla.

We expect this work will play a vital role in the development of fake news detection systems. In the rest of the paper, we briefly describe our dataset preparation methods, human performance & observation, and the development of fake news detection systems along with their performance.

† First and second authors contributed equally.

[1] https://en.wikipedia.org/w/index.php?title=Fake_news&oldid=921640983

[2] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

2. Related Works

Acknowledging the impact of fake news, researchers are trying different methodologies to find a quick and automatic solution to detect fake news in the recent past years. Previous works on satirical news detection mostly use Support Vector Machine (SVM) (Rubin et al., 2016; Burfoot and Baldwin, 2009; Ahmad et al., 2014). Focusing on predictive features an SVM model is proposed by Rubin et al. (2016) with 360 news articles collected from 2 satirical news sites (The Onion and The Beaverton) and 2 legitimate news sources (The Toronto Star and The New York Times) in 2015 and showed absurdity, grammar and punctuation marks are best for identifying satirical news. Leveraging neural networks, Yang et al. (2017) built a 4-level hierarchical network and utilized attention mechanism by using $\approx 16K$ satirical data (collected from 14 satirical news websites) and $\approx 160K$ true data and showed paragraph-level features are better than document level features in terms of the stylistic features. Some approaches focused on deception detection and utilised traditional machine learning models such as Naive Bayesian models (Oraby et al., 2017) and SVM (Ren and Zhang, 2016) to work with linguistic cues.

To learn the fake news patterns Pérez-Rosas et al. (2018) also used a linear SVM classifier but to build the supervised learning model they used only linguistic features such as N-grams (Scholkopf and Smola, 2001), Punctuations, LIWC (Pennebaker et al., 2015) and conduct their evaluations using five-fold cross-validation. First they collected a dataset of 240 legitimate news from different mainstream news websites in the US then made another dataset containing fake versions of those legitimate news. To generate fake versions of the legitimate news items, they used the crowdsourcing via Amazon Mechanical Turk (AMT). However to find out the underlying features of fake news, neural networks are more useful (Shu et al., 2017) and for textual feature extraction, word embedding technique with deep neural networks are producing quality results. But it is noticeable that neural network based models requires large dataset. Liar (Wang, 2017) is a comparatively large dataset containing 12.8K human-labeled short statements collected from POLITIFACT.COM's API. They used both surface-level linguistic realization and deep neural network architecture.

Existing research towards clickbait detection involves hand-crafted linguistic features (Chakraborty et al., 2016; Biyani et al., 2016) and deep neural networks (Gairola et al., 2017; Rony et al., 2017) with datasets mostly containing clickbaits (headline and article of clickbait news) from different newspapers. Besides (Ciampaglia et al., 2015; Vlachos and Riedel, 2014) have proposed a fact-checking method through knowledge base. And Karadzhov et al. (2017) has proposed a fully automated fact-checking using external sources where their dataset contains 761 claims from snopes.com, which span various domains including politics, local news, and fun facts. Each of the claims is labeled as factually true (34%) or as a false rumor (66%).

Though researches focused on the English language have achieved significant advancement, very few works are available for different low resource languages like Indonesian, Bangla, Hindi. Pratiwi et al. (2017) created a dataset containing 250 pages of hoax and valid news articles and proposed a hoax detection model using the Naive Bayes classifier for the Indonesian language. In Chinese, a dataset of 50K news is used by Zhou et al. (2015) for their real-time news certification system on the Chinese micro-blogging website, Sina Weibo[3].

To the best of our knowledge, our work is the first publicly available news dataset in the Bangla for fake news detection. Throughout our literature review, we found that most of the works introduce a dataset suitable for their research approach and there is some dataset only focused on particular research topics like stance detection[4]. Since fake news related research for the Bangla are still in its early stage, we design our dataset in a diverse way so that it can be used in multiple lines of research. So we enrich our dataset with clickbaits, satirical, fake and true news with their headline, article, domain and other metadata which is explained briefly in the next section.

3. A New Dataset for Detecting Fake News Written in Bengali

To collect a set of authentic news, we select 22 most popular[5] and mainstream trusted news portals in Bangladesh. For collecting fake news we include the following types of news in our dataset.

- Misleading/False Context: Any news with unreliable information or contains facts that can mislead audiences.
- Clickbait: News that uses sensitive headlines to grab attention and drive click-throughs to the publisher's website.
- Satire/Parody: News stories that are intended for entertainment and parody.

we have collected news from popular websites that publish satire news in Bangla. While collecting satirical news from these sites we found that most of the sites have the exact same news. So after scraping news from these sites, we discarded the duplicates. We have collected the misleading or false context type of news from `www.jaachai.com` and `www.bdfactcheck.com`. These two websites provide a logical and informative explanation of fake news that is already published on other sites. So we have also collected the news that is mentioned on those two sites from the actual publishing sites and make sure that we avoid the duplicates. Clickbait is used to grab attention and drive click which eventually increases site visitors and generates revenue for them[6]. And we have found that most of the local or less popular sites usually do this. To collect clickbaits, we have gone through some of these sites and manually collect potential clickbait news from there. We call satire, clickbait, and false informative news all commonly as fake news

[3] `https://s.weibo.com/`

[4] `http://www.fakenewschallenge.org/`

[5] We used the Alexa rankings to determine the popularity (`www.alexa.com`)

[6] `www.webwise.ie/teachers/what-is-fake-news`

throughout the paper to avoid ambiguity. We have also collected the following meta-data along with the headlines and content:

- The domain of the published news site
- Publication time
- Category

From our dataset, we got 242 different categories as different publishers categorize the news in their own way. To generalize it, we took similar categories from different news to map into a single one. Finally, we categorize all news from the dataset into the 12 categories (Table 1).

Category	Authentic	Fake
Miscellaneous	2218	654
Entertainment	2636	106
Lifestyle	901	102
National	18708	99
International	6990	91
Politics	2941	90
Sports	6526	54
Crime	1072	42
Education	1115	30
Technology	843	29
Finance	1224	2
Editorial	3504	0

Table 1: Number of news in each category.

Human observation suggests that the source plays a key role in an article's credibility. Note that, by source here we mean one or more person or organization capable of providing verification of the claimed news. If there is no such source, then reporters or journalists are taken as the source of news. Besides, Long et al. (2017) has shown that adding speaker profiles along with document-level features improved the performance of fake news detection. So to make our dataset more resourceful, we include the source information as a meta-data for each news. Besides the source, we have also included the headline article relation in meta-data. "Related" and "Unrelated" tags are provided upon checking the relationships of the headline with the article. Since we have to go through each of the news to find out the source and headline-article relation so far we have managed to annotate only $\approx 8.5K$ data. All of the members of our data annotator team are undergraduate students of Computer Science and Engineering and Software Engineering department. Figure 1 is a sample from our dataset.

4. Human Baseline

Detecting fake news is a tough task. To see how good humans perform, we have conducted an experiment where we took 60 fake news and mixed them randomly with 90 authentic news and gave them to 5 human annotators who are undergraduate students(one Industrial and Production Engineering, one Chemistry and three Computer Science and Engineering students). They were told to read 150 news one by one and answer 2 questions for each news. Note

Label	1 (Authentic)
Domain	channelionline.com
Published Time	2018-09-19 18:15:40
Category	আন্তর্জাতিক
Source	বিবিসি
Headline-article relation	related
Headline	মুক্তি পেলেন নওয়াজ শরীফ
Article	পাকিস্তানের সাবেক প্রধানমন্ত্রী নওয়াজ শরীফকে মুক্তি দিয়েছে দেশটির উচ্চ আদালত। কথিত দুর্নীতির মামলায় ১০ বছরের সাজা পেয়ে দুই মাস কারাভোগের পর তিনি মুক্তি পান। বুধবার দেশটির আদালত পাকিস্তান মুসলিম লিগের শীর্ষ এ নেতাকে মুক্তির আদেশ দেন। আদালত একইসঙ্গে নওয়াজের মেয়ে মরিয়ম শরীফকেও মুক্তির আদেশ দিয়েছেন। তাদের আবেদনের পরিপ্রেক্ষিতে আদালত এই রায় দিয়েছেন বলে বিবিসি জানায়। পাকিস্তানের জাতীয় নির্বাচনের আগে দুর্নীতির মামলায় নওয়াজ শরীফকে ১০ বছর এবং তার মেয়ে মরিয়মকে ৭ বছরের কারাদণ্ড দেন আদালত।

Figure 1: Sample Data

	T	C	P	W	S
Authentic	48678	1479.14	41.20	271.16	21.15
Fake	1299	1428.19	44.13	276.36	23.62

Table 2: Distribution of mean of characters, punctuations, words, and sentences along with total number of authentic and fake news in dataset. T, C, P, W, S denotes total news count, characters, punctuations, words, sentences mean respectively.

that we only gave them the headlines and the articles. The first question is if the news is fake or authentic. Based on the answer of the first question they were provided another question with four options to select why they think the news is fake or authentic. Here the set of four options are different for fake and authentic. For creating these four options, we first took feedback from 10 people regarding what properties or methods they use to find out if any news is fake or authentic. Then we generalized their feedback into four options. The process of the human baseline experiment is shown in Fig. 2 using a flowchart.

From the first question, we got an estimate of how accurately humans can detect fake news. The F1-score for the fake class of the five annotators is 58%, 65%, 70%, 68%, and 63% respectively. The inter-annotator agreement, measured as Fleiss' Kappa(Fleiss, 1971) is 38.83% and mean pairwise Cohen's Kappa(Cohen, 1960) is 39.05% which indicates that our human annotators have given the same answer on nearly 39% percent of the questions. The second question helped us to find out what factors are crucial for humans to find the difference between fake and authentic news. If the news is fake, on an average 44% answer is 'The content is unrealistic' and 42% answer is 'Has no trustworthy source'. When news is true, on an average 62% answer is 'The Content is believable' and 21% answer is 'Source is reliable'. Here the source is a person or an organization who/that can provide the validation of the claimed news. From the feedbacks of annotators on overall answers, we found that they choose the option 'The Content is believable' because they read or hear similar news on their daily

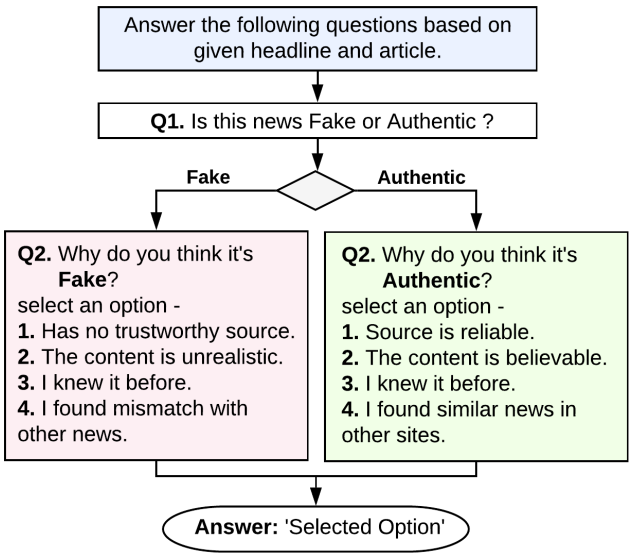

Figure 2: Process of human baseline experiment

life and the content has nothing to disbelieve it. If the content looks fishy to them then they look for the source of the news. So the experiment shows that source of the news is important. After the experiment we had a follow-up interview with the human participants to find out why they make mistakes in detecting fake news. Most significant reasons are as follows:

- **Disguise:** If any fake news is represented just like true news that is the fake news contains enriched information such as strong references, scientific facts and statistics in detail then human mistake it as true news.
- **Trending:** People tend to believe any news when they notice a lot of newspapers are reporting/have reported the same news.
- **Source:** When a dubious news doesn't contain any reference (an entity who provided the information) then humans take it as false news.
- **Satire:** Humans can detect almost all types of satire news but sometimes some true news also sounds like satire and humans take them as fake news.

These analysis indicates that the source plays an influential role in fake news detection. So In our dataset, we manually annotated the source of the news so that in future we can use the source as a momentous feature for detecting fake news.

5. Methodologies

In this section, we describe the systems we develop to classify fake news written in Bangla. Our approaches include traditional linguistic features and multiple neural networks based models.

5.1. Traditional Linguistic Features

- **Lexical Features:** Due to the strong performance in various text classification tasks, we extract word n-grams (n=1,2,3) and character n-grams (n=3,4,5) from the news articles. As the weighting scheme we use the term frequency-inverse document frequency (TF-IDF).
- **Syntactic Features:** Syntactic structure of texts is often beneficial for understanding particular patterns in documents which eventually help classification problems. So we tag the words of the news articles with their *Parts of Speech* (POS) tags using (Loper and Bird, 2002). We use the normalized frequency of different POS tags (*Adjective, Noun, Verb, Demonstrative, Adverb, Pronoun, Conjunction, Particle, Quantifier, Postposition*) as a feature set for each document.
- **Semantic Features:** Distributed representations of word and sub-word tokens have shown effectiveness in text classification problems by providing semantic information. So we experiment with pre-trained word embedding, where we represent an article by the mean and the standard deviation of the vector representations of the words in it. We experiment with the Bangla 300 dimensional word vectors pre-trained[7] with Fasttext (Grave et al., 2018) on *Wikipedia*[8] and *Common Crawl*[9], where we have a coverage of 55.21%. Additionally, we experiment with another set of pre-trained 100 dimensional word vectors trained on $\approx 20K$ Bangla news by Ahmad and Amin (2016) with Word2Vec (Mikolov et al., 2013), where we have a coverage of 53.95%. We will call it *News Embedding* throughout the rest of this paper.
- **Metadata and Punctuation (MP):** We observed higher presence of some punctuation symbols like '!' in the fake news. So we use the punctuation frequency as features. Additionally we use some meta information like the lengths of the headline and the body of news articles as features.

  We have found that the publishing sites of fake news are less popular than the sites of true news. So we used the Alexa Ranking[10] of the sites which are designed to estimate the popularity of websites as a feature. We didn't find the rank of some of the news sites so we annotate these with maximum rank from other sites. And we used the normalized value of ranks as a feature in experiments.

5.2. Neural Network Models

Neural networks are demonstrating impressive performance in a wide range of text classification and generation tasks. Given a large amount of training data, such models typically achieve higher accuracy than linguistic feature based methods. Hence, we experiment with several neural network models that have been used as benchmark models in different text classification tasks.

[7] https://fasttext.cc/docs/en/crawl-vectors.html
[8] http://wikipedia.org
[9] https://commoncrawl.org/
[10] https://www.alexa.com

Convolutional Neural Network (CNN): Convolutional networks have shown effectiveness in classifying short and long texts in varieties of problems (Kim, 2014; Shrestha et al., 2017). So we experiment on classifying fake news using a CNN model similar to (Kim, 2014). We use kernels having lengths from 1 to 4 and empirically use 256 kernels for each kernel length. As a pooling layer, we experiment with global max pool and average pool. We use ReLU (Agarap, 2018) as the activation function inside the network.

Long Short Term Memory: Due to the capability of capturing sequential information in an efficient manner, Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks are one of the most widely used models in text classification and generation problems. Specifically, bidirectional LSTM (Bi-LSTM) have shown impressive performance by capturing sequential information from the both directions in texts. Moreover, attention mechanism has been seen as a strong pooling technique for classification tasks when used with Bi-LSTM.
In this work, we experiment with a Bi-LSTM model having attention on top which is similar to (Zhou et al., 2016). We use 256 LSTM units. We use two layers of Bi-LSTM in the network.

5.3. Pre-trained Language Model

Recently pre-trained language models like OpenAI GPT (Radford et al., 2018), BERT (Devlin et al., 2018), ULMFiT (Howard and Ruder, 2018) have made a breakthrough in a wide range of NLP tasks. Specifically BERT and its variant models have shown superior performance in the GLUE benchmark for Natural Language Understanding (NLU) (Wang et al., 2019). To evaluate the scope of such a language model in our work, we use the multilingual BERT model to classify news documents. We use the pre-trained model weights and implementation publicly distributed by HuggingFace's Transformers (Wolf et al., 2019).

6. Experimental Setup

Data Pre-processing: We perform several pre-processing techniques like text normalization and stop words, punctuation removal from the data. Here we use Bangla stop words from Stopwords ISO [11]. We observe better validation performance by such pre-processing of the data.
Evaluation Metric: We use Micro-F1 scores to evaluate different methods. As the dataset is imbalanced, we also report the precision (P), recall (R), and F1 score for the minority class (*fake*).
Baselines: We compare our experimental results with a majority baseline and a random baseline. The majority baseline assigns the most frequent class label (*authentic news*) to every article, where the other baseline randomly tags an article as *authentic* or *fake*. We report the mean of precision, recall, and F1-score of 10 random baseline experiments in Table 3. The Standard Deviation(SD) of precision, recall, and F1-score in both overall and fake class is less than 10^{-2} except the recall of Fake class which is 0.027.

Experiments: With the linguistic features, we experiment on training a Linear Support Vector Machine (SVM) (Hearst, 1998), Random Forest (RF) (Liaw and Wiener, 2002) and a Logistic Regression (LR) (McCullagh and Nelder, 1989) model. We split our dataset for training and testing in a 70:30 train-to-test ratio. We tune the penalty parameter (C) based on the validation results.
For BiLSTM, CNN, and BERT based experiments, the hyper-parameters are Optimizer: Adam Optimizer (Kingma and Ba, 2015), Learning rate: 0.00002, Batch size: 32. Hidden size for BERT model is 756 while in CNN and BiLSTM it is 256. For CNN, we use the kernel lengths of 1 to 4 and zero right paddings in the experiment. The train and test dataset is kept at a 70:30 ratio. And In training, we use 10% of the test data as validation data. We use 50 epochs for each experiment and put a checkpoint on F1 score of fake class using validation data. And finally we report the result using our test dataset on the best scoring model from training.
In BERT, For fine tuning our dataset we use the sequence classification model by HuggingFace's Transformers. And we use the BERT's pre-trained multilingual cased model which is trained on 104 different languages[12].

7. Results and Analysis

We report our results in Table 3. Overall performance of every experiment is almost the same. Most of the cases we achieve almost perfect Precision, Recall and F1. But the results of Precision, Recall, and F1-Score of fake class vary in experiments to experiments. In our dataset for experiments, the number of authentic news is 37.47 times higher than the number of fake news which could be the reason behind such variance in results of the overall and fake class. To evaluate the performance of different models we will use the precision, recall, F1-Score of the fake class in the rest of the section.
Fig. 3 shows that experiment with linguistic features with SVM, RF and LR. Here SVM outperforms LR and RF by a quite margin except result of the news embedding. In the case of news embedding RF scores 55% of F1-score and here SVM, LR scores 46%, 53% of F1-score respectively. For most of the features RF performs better than the LR model. We report the result of SVM on Table 3 since it has performed better than the others. Table 3 shows that lexical features perform better than other linguistic features, majority & random baselines, and neural network models as well. It is also observed that the F1-score of fake class in the SVM model decreases while increasing the number of grams in both word and character n-grams.
The result of POS tag does not improve over the random baselines and the F1-score of this feature indicates that it cannot separate fake news from authentic news. Again F1-Score of MP, Word Embedding(Fasttext), Word Embedding(News) are better than the POS tags but fall behind the lexical features and neural network models. However, these features outperform the majority, and random baselines. We got our best result when incorporating all linguistic features with SVM. It scores 91% F1-score.

[11] https://github.com/stopwords-iso

[12] https://github.com/google-research/bert

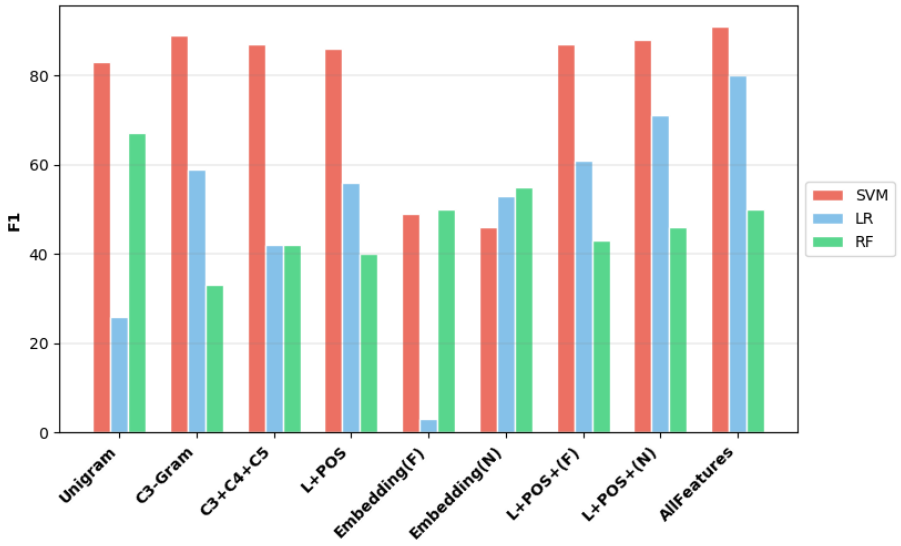

Figure 3: Comparison between SVM, LR and RF on linguistic features

	Overall			Fake Class		
	P	**R**	**F1**	**P**	**R**	**F1**
Baselines						
Majority	0.97	1.00	0.99	0.00	0.00	0.00
Random	0.97	0.50	0.66	0.03	0.50	**0.05**
Traditional Linguistic Features						
Unigram (U)	0.99	0.99	0.99	0.99	0.71	0.83
Bigram (B)	0.98	0.99	0.99	0.97	0.42	0.59
Trigram (T)	0.98	0.99	0.98	0.74	0.31	0.44
U+B+T	0.99	0.99	0.99	0.98	0.68	0.80
C3-gram(C3)	0.99	0.99	0.99	0.98	0.82	0.89
C4-gram(C4)	0.99	0.99	0.99	0.99	0.78	0.87
C5-gram(C5)	0.99	0.99	0.99	1.00	0.74	0.85
C3+C4+C5	0.99	0.99	0.99	1.00	0.77	0.87
All Lexical(L)	0.99	0.99	0.99	1.00	0.76	0.86
POS tag	0.97	1.00	0.98	1.00	0.00	0.01
L+POS	0.99	0.99	0.99	0.99	0.76	0.86
Embedding(F)	0.98	0.99	0.99	0.94	0.33	0.49
Embedding(N)	0.98	0.99	0.99	0.84	0.32	0.46
L+POS+E(F)	0.99	0.99	0.99	0.99	0.77	0.87
L+POS+E(N)	0.99	0.99	0.99	0.98	0.79	0.88
MP	0.97	0.99	0.98	0.94	0.15	0.27
L+POS+E(F)+MP	0.99	0.99	0.99	0.99	0.84	**0.91**
L+POS+E(N)+MP	0.99	0.99	0.99	0.98	0.84	**0.91**
All Features	0.99	0.99	0.99	0.98	0.84	**0.91**
Neural Network Models						
CNN	0.98	1.00	0.99	0.79	0.41	**0.59**
LSTM	0.99	0.99	0.99	0.69	0.44	0.53
BERT	0.99	1.00	0.99	0.80	0.60	**0.68**

Table 3: Results of experiments with Traditional Linguistic Features (SVM) and Neural Networks with test set. P and R denote precision and recall, respectively. 'F' and 'N' abbreviate 'Fastext' and 'News', respectively.

Neural Network models have shown better results in different text classification problems (Lai et al., 2015; Joulin et al., 2016). But in our experiments, we found that F1-Score of fake class in neural networks cannot outperform the linear classifiers. In CNN, experiment with average pooling, global max technique scores 59% and 54% F1-Score respectively. Generally an attention model with LSTM performs better than CNN (Yang et al., 2016; Chen et al., 2016). In our case, the F1-Score of BiLSTM with attention cannot improve over the CNN model. The best results of neural networks based experiments came from BERT, whose F1-Score is 68%. Though neural network models perform better than the majority and random baselines still it falls behind the performance of the SVM model.

To compare the performance of our fake news detection system with humans, We tested the same dataset that we used for the human baseline with our two best models. The model with SVM and character 3-gram scores 41% F1-score where it scores 38% with SVM and all linguistic features. Here, the performance of our best models drops significantly. While checking individual output we found that our model cannot separate the misleading or false contents from authentic contents but it can separate the satire news. In our test set, 87.45% news is satire and the rest of them are misleading news and clickbait. On the other hand, the dataset we used for the human baseline contains 23% satire news. That is why the performance of our models drops with the human baseline dataset. The average F1-score of humans on the human baseline dataset is 64.8%. Comparing with the result of humans performance it can be said that human performance is better than our models. But, we have also observed that humans also cannot perfectly separate the misleading or false contents, which indicates that both computational approaches and humans are not capable enough to separate the fake news that contains misleading or false information.

8. Conclusion

In this paper, we present the first labeled dataset on Bangla Fake news. Here the evaluation of linear classifiers and neural network based models suggest that linear classifiers with traditional linguistic features can perform better than the neural network based models. Another key finding is

that character-level features are more significant than the word-level features. In addition to that, it is also found that the use of punctuations in fake news is more frequent than authentic news, and most of the time fake news is found on the least popular sites.

However, since character level features have shown better results so we will incorporate the character level features in neural network models such as (Kim et al., 2016; Hwang and Sung, 2017). From human observation, we found that the source can also play a key role in fake news detection. We will also include this feature in our future experiments. Besides we will also continue to expand our dataset. We have manually annotated around 8.5K news. We will continue this process to reach the 50K mark. We hope our dataset will provide opportunities for other researchers to use computational approaches to fake news detection.

9. Acknowledgements

We would like to thank data annotators and resource team who helped us during human baseline experiment. Also we thank Natural Language Processing Group, Dept. of CSE, SUST for their valuable comments and discussions with us.

10. Bibliographical References

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *CoRR*, abs/1803.08375.

Ahmad, A. and Amin, M. R. (2016). Bengali word embeddings and it's application in solving document classification problem. In *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pages 425–430. IEEE.

Ahmad, T., Akhtar, H., Chopra, A., and Akhtar, M. W. (2014). Satire detection from web documents using machine learning methods. In *2014 International Conference on Soft Computing and Machine Intelligence*, pages 102–105. IEEE.

Biyani, P., Tsioutsiouliklis, K., and Blackmer, J. (2016). " 8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Burfoot, C. and Baldwin, T. (2009). Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 161–164.

Chakraborty, A., Paranjape, B., Kakarla, S., and Ganguly, N. (2016). Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16. IEEE.

Chen, H., Sun, M., Tu, C., Lin, Y., and Liu, Z. (2016). Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1650–1659, Austin, Texas, November. Association for Computational Linguistics.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. (2015). Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Dong, X., Victor, U., Chowdhury, S., and Qian, L. (2019). Deep two-path semi-supervised learning for fake news detection. *CoRR*, abs/1906.05659.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gairola, S., Lal, Y. K., Kumar, V., and Khattar, D. (2017). A neural clickbait detection engine. *arXiv preprint arXiv:1710.01507*.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hearst, M. A. (1998). Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, July.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hwang, K. and Sung, W. (2017). Character-level language modeling with hierarchical recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5720–5724. IEEE.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Karadzhov, G., Nakov, P., Màrquez, L., Barrón-Cedeño, A., and Koychev, I. (2017). Fully automated fact checking using external sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 344–353, Varna, Bulgaria, September. INCOMA Ltd.

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2016). Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Long, Y., Lu, Q., Xiang, R., Li, M., and Huang, C.-R. (2017). Fake news detection through multi-perspective

speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 252–256, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Manik, J. A. (2012). A hazy picture appears. `https://www.thedailystar.net/news-detail-252212`, October. [Online; posted 03-October-2012].

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall / CRC, London.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Oraby, S., Reed, L., Compton, R., Riloff, E., Walker, M., and Whittaker, S. (2017). And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. *arXiv preprint arXiv:1709.05295*.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Pratiwi, I. Y. R., Asmara, R. A., and Rahutomo, F. (2017). Study of hoax news detection using naïve bayes classifier in indonesian language. In *2017 11th International Conference on Information & Communication Technology and System (ICTS)*, pages 73–78. IEEE.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Ren, Y. and Zhang, Y. (2016). Deceptive opinion spam detection using neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 140–150, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Rony, M. M. U., Hassan, N., and Yousuf, M. (2017). Diving deep into clickbaits: Who use them to what extents in which topics with what effects? In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 232–239. ACM.

Rubin, V., Conroy, N., Chen, Y., and Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California, June. Association for Computational Linguistics.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Shrestha, P., Sierra, S., González, F., Montes, M., Rosso, P., and Solorio, T. (2017). Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain, April. Association for Computational Linguistics.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

TUFEKCI, Z. (2018). It's the (democracy-poisoning) golden age of free speech. `https://www.wired.com/story/free-speech-issue-tech-turmoil-new-censorship/?CNDID=50121752`, January. [Online; posted 16-January-2018].

Vlachos, A. and Riedel, S. (2014). Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA, June. Association for Computational Linguistics.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Yang, F., Mukherjee, A., and Dragut, E. (2017). Satirical news detection and analysis using attention mechanism and linguistic features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1979–1989, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zhou, X., Cao, J., Jin, Z., Xie, F., Su, Y., Chu, D., Cao, X.,

and Zhang, J. (2015). Real-time news cer tification system on sina weibo. In *Proceedings of the 24th International Conference on World Wide Web*, pages 983–988. ACM.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany, August. Association for Computational Linguistics.

Appendix

A News Sources

Domain	Rank	#News
Authentic News		
kalerkantho.com	2040	4491
jagonews24.com	1771	4426
banglanews24.com	3539	4035
banglatribune.com	14505	3696
jugantor.com	1422	2835
dhakatimes24.com	34756	2654
ittefaq.com.bd	6300	2589
somoynews.tv	3214	2552
dailynayadiganta.com	9678	2371
bangla.bdnews24.com	3329	2365
prothomalo.com	470	2350
bd24live.com	6989	2335
risingbd.com	11162	2220
dailyjanakantha.com	24403	1531
bd-pratidin.com	2141	1421
channelionline.com	8878	1401
samakal.com	8698	1372
independent24.com	216950	1220
rtnn.net	1921350	1149
bangla.thereport24.com	219278	859
mzamin.com	8715	785
bhorerkagoj.net	60018	21
Satire News		
channeldhaka.news	249033	436
earki.com	2226986	291
motikontho.wordpress.com	7291230	195
bengalbeats.com	707465	192
sarcasmnews.fun	N\A	14
ctnews7	7484275.0	4
Prothombarta.news	193562	1
TheReport24.com	219278	1
aparadhchokh24bd.com	4242479	1
shadhinbangla24	1115208	1
Clickbait		
bengaliviralnews.com	1884782	48
gonews24.com	83988	10
lastnewsbd.com	77165	3
bangladeshonline24.com	3023578	2
news.zoombangla.com	5770	2
prothombarta.news	193562	2
prothombhor.net	2831785	2
somoyerkonthosor.com	72800	2
agooannews.com	N\A	1
aparadhchokh24bd.com	4242479	1
banglanews24.com	3527	1
bdjournal365.com	2395727	1
bdsangbad.com	5272008	1
bdtype.com	309252	1
bn.mtnews24.com	63562	1
daily-bangladesh.com	6961	1
dkpapers.com	216868	1
sangbadprotidin24.com	1489949	1
sonalinews.com	37689	1
Fake		
banglainsider.com	23339	3
bd-pratidin.com	2141	3
bengaliviralnews.com	1884782	3
notunshokal.com	3012897	3
alokitobangladesh.com	84832	2
bangla.dhakatribune.com	15159	2
banglanews24.com	3527	2
dailyinqilab.com	9970	2
dailysangram.com	84689	2
ittefaq.com.bd	6285	2
jugantor.com	1420	2
kalerkantho.com	2031	2
shadhinbangla24.com.bd	1268056	2
somewhereinblog.net	50584	2
sylhettoday24.news	94555	2
bangla.bdnews24.com	3332	1
bangla24.com.bd	6652978	1
bangladeshbani24.com	N\A	1
banglatribune.com	14517	1
bd-journal.com	8334	1
bd24live.com	6989	1
bd24report.com	23915	1
bdhotnews.com	8415177	1
bengali.oneindia.com	1150	1
bn.banglafact.com	2911410	1
bn.bdcrictime.com	57569	1
bn.mtnews24.com	63562	1
channelionline.com	336950	1
city24news.com	147311	1
coxsbazarnews.com	130112	1
dailyamadernandail.com	3304326	1
dailyjanakantha.com	24403	1
dailynayadiganta.com	9689	1
dailysatkhira.com	738489	1
deshebideshe.com	28828	1
dhakajournals.com	3625884	1
dhakatimes24.com	35,273	1
ekushey-tv.com	21121	1
famousnews24.com	469905	1

Domain	Rank	#News
gonews24.com	83988	1
jagonews24.com	1771	1
keuamaremairala.com	10045257	1
mujibsenanews.com	2420840	1
nirapadnews.com	158811	1
ppbd.news	29471	1
priyo.com	38831	1
ekusherbangladesh.com.bd	124277	1
probashkotha.com	654309	1
prothombhor.net	2831785	1
protissobi.com	4943617	1
rtvonline.com	12281	1
sharebusiness24.com	566064	1
sharenews24.com	98406	1
shawdeshbhumi.com	N\A	1
snpsports24.com	1113116	1
somoyerkonthosor.com	72800	1
sylhetprotidin24.com	3086838	1
timesbangla.in	2675956	1
awarenessbulletin.blog-spot.com	N\A	1
bdexclusivenews.blog-spot.com	N\A	1
bangla.24livenews-paper.com	6725	1
ourevergreen-bangladesh.com	436458	1

Table 4: Detailed statistics of the collected news with the domain URL and Alexa ranking(as of 08 March 2020).