IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 6.252

Volume 2, Issue 1, December 2022

# Claim Analyzer : Evaluating Credibility of Data

**Ranjana Shende[1], Neel Karkade[2], Parshik Rakhade[3], Shrutik Thool[4], Vijay Dubey[5], Vishal Nikode[6]**

Professor, Department of Computer Science & Engineering[1,]

Students, Department of Computer Science & Engineering[2,3,4 5,6]

G. H. Raisoni Institute of Engineering and Technology, Nagpur, Maharashtra, India

An Autonomous Institute Affiliated to Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur, Maharashtra, India

Accredited by NAAC With A+ Grade

**Abstract**: *The current state of news in the era of computers The news environment in the modern computer age The news environment in the modern computer age The news environment in the modern computerage.The news environment in the modern computer age Social media platforms have taken the role ofthe antiquated traditional print media as a part of the news ecosystem in the current computer era.False news is spread at an astounding velocity and scale because social media platforms allow us to consume news much more rapidly and with less restricted editing. According to recent research, several efficient methods for spotting fake news encrypt social context-level data and news content sequential neural networks using a unidirectional examination of the text sequence. In order to represent the pertinent information of false news and improve classification performance while capturing semantic and long-distance connections in sentences, a bidirectional training strategy is a necessity. Claim Analyzer is only a model for evaluating the veracity of claims made online The statements might either be True or False. Fake news is a kind of propaganda in which false information is knowingly disseminated via news organizations and/or social media platforms. It is crucial to create methods of spotting false news information since its spread can have detrimental effects, such as influencing elections and widening political rifts. BERT is intended to simultaneously condition on both left and right context in all layers in order to pre-train deep bidirectional representations from unlabeled text. Therefore, state-of-the-art models for a variety of tasks, including question answering and language inference, may be created using just one extra output layer to fine-tune the pre-trained BERT model without making significant task-specific architecture adjustments. By merging several parallel blocks of the single-layer deep Convolutional Neural Network (CNN)with the BERT and variable kernel sizes and filters, we propose a BERT-based (Bidirectional Encoder Representations from Transformers) deep learning technique. The biggest obstacle to natural language understanding is ambiguity, which may be handled with the help of this combination. When applied to huge datasets, our suggested model provides88% accuracy.*

**Keywords:** Fake news Detector, Principle Component Analysis, ERT, FNC(Fake news challenges), LSTM(Long Short term memory), Machine Learning

## I. INTRODUCTION

We are in an informational digital era. We can access thousands, if not millions, of materials online with a single click. But can we rely on all of this data? Regrettably, no. In recent years, false news has grown (or perhaps there has been an explosion or pandemic). What, however, is false news? Propaganda, a hoax, and/or disinformation that is deliberately disseminated and published as legitimate news, frequently on social media, with the objective to deceive for political or financial benefit are all examples of fake news. Eye-catching headlines and visuals are frequently used in fake news to boost sharing and views. In recent years, the phrase "fake news" has developed into something more general and comprehensive, embracing items that may have Online resources are abundant, and this is unlikely to change. We must continue to be cautious about what information is reliable and what is not as more items become available online. It is our responsibility to be responsible enough to distinguish between authentic and fraudulent information.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
IJARSCT
Impact Factor: 6.252
Volume 2, Issue 1, December 2022

## II. LITERATURE SURVEY

Society and individuals are negatively influenced both politically and socially by the widespread increase of fake news either way generated by humans or machines. In the era of social networks, the quick rotation of news makes it challenging to evaluate its reliability promptly. Therefore, automated fake news detection tools have become a crucial requirement. To address the aforementioned issue, a hybrid Neural Network architecture, that combines the capabilities of CNN and LSTM, is used with two different dimensionality reduction approaches, Principle Component Analysis (PCA) and Chi-Square. This work proposed to employ the dimensionality reduction techniques to reduce the dimensionality of the feature vectors before passing them to the classifier.

To develop the reasoning, this work acquired a dataset from the Fake News Challenges (FNC) website which has four types of stances: agree, disagree, discuss, and unrelated. The nonlinear features are fed to PCA and chi-square which provides more contextual features for fake news detection. The motivation of this research is to determine the relative stance of a news article towards its headline. The proposed model improves results by ~4% and ~20% in terms of Accuracy and F1−score . The experimental results show that PCA outperforms than Chi-square and state-of-the-art methods with 97.8% accuracy.

## III. METHODOLOGY

The data collection and experiments that were conducted in the study are described in this section.

### 3.1 Data Collection and Description of Dataset

Kaggle and Beautiful Soup have been used to collect data. Utilizing the real-world fake news dataset, we conducted in-depth experiments for this article. It is made up of two files: test and train.csv. csv: A test dataset without labels. It is a collection of the fake and actual news that was spread during the 2016 U.S. General Presidential Election; we can see the instances with the class labels in the corresponding fake news dataset.

### 3.2. Preprocessed Dataset Description

Two unsupervised learning tasks were taught by the researchers for the pre-training BERT algorithm. Masked LM is how the first task is referred to. This operates by anticipating those tokens that have been randomly masking 15% of a text. Predicting the next sentence is the second task (NSP). Tasks like Question Answering and Natural Language Inference serve as motivation for this. For these objectives, models must faithfully represent the connections between phrases. They pre-train for a binarized prediction task that can be easily produced from any corpus in a single language in order to address this.

### 3.3. Creating an LSTM Model

First, we will split our dataset to an 70:15:15 train:test ratio. Text data must be converted to a vector format so that it can be analysed by computers if we want our model to make predictions based on it. Wiki-words-250 by Tensorflow employs a Word2Vec Skip-Gram architecture. By guessing the context from an input word, skip-gram is trained.

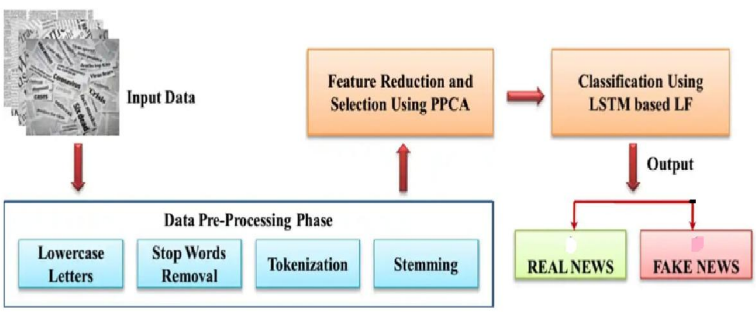


**Figure:** Block diagram of cliam analyzer

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
IJARSCT
Impact Factor: 6.252

### 3.4 Introducing BERT

BERT model consists of:

- **Inp**ut layer: This will represent the sentence that will be passed into the model).
- **The bert_preprocess layer:** Here we pass in our input to preprocess the text.
- **The bert_encoder layer:** Here we pass the preprocessed tokens into the BERT encoder.

Dropout layer with a rate of 0.2. The pooled_output of the BERT encoder is passed into it (more on this below)
Dense layers with 10 and 1 neurons respectively. The first one will use a ReLU activation function, and the second will use sigmoid.

### 3.5 Final Dataframe

Then, with an 80:20 train:test ratio, we may divide the data used to train and test our model. With Tensorflow-hub, we can now import the BERT pre-processor and encoder. Our neural network can now be developed. The model must be functional, with each layer's output acting as an argument to the one below it. The model includes:
One input layer (which will serve to represent each sentence fed into the model);

- The bert preprocess layer: We provide our input here so that the text can be processed.
- The BERT encoder is accessed through the bert encoder layer, which receives the preprocessed tokens.

One dropout layer with a 0.2 dropout rate. It receives the BERT encoder's pooled output (more on this below) 2 dense layers, each containing 10 and 1 neurons. The first will employ a ReLU activation function, and the second a sigmoid activation function.
As you can see, the dropout layer will receive the "pooled output". This number represents how the text is organised overall. It is the representation of the [CLS] token outputs.

## IV. RESULT AND DISCUSSION

In this model we are getting 88 % accuracy with BERT model

Model performance

```
[ ] # load weights of best model
    path = 'c1_fakenews_weights.pt'
    model.load_state_dict(torch.load(path))

    <All keys matched successfully>

[ ] with torch.no_grad():
      preds = model(test_seq, test_mask)
      preds = preds.detach().cpu().numpy()

    preds = np.argmax(preds, axis = 1)
    print(classification_report(test_y, preds))
```

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.85 | 0.92 | 0.88 | 3213 |
| 1 | 0.92 | 0.85 | 0.88 | 3522 |
| accuracy | | | 0.88 | 6735 |
| macro avg | 0.88 | 0.88 | 0.88 | 6735 |
| weighted avg | 0.88 | 0.88 | 0.88 | 6735 |

Final result:

```
[ ] #@title Default title text
    # testing on unseen data
    unseen_news_text = ["Donald Trump Sends Out Embarrassing New Year's Eve Message; This is Disturbing",   # Fake
                        "Trump administration issues new rules on U.S. visa waivers",                       # True
                        "WATCH: George W. Bush Calls Out Trump For Supporting White Supremacy",            # Fake
                        "U.S. lawmakers question businessman at 2016 Trump Tower meeting: sources"         # True

                        ]

    # tokenize and encode sequences in the test set
    MAX_LENGHT = 15
    tokens_unseen = tokenizer.batch_encode_plus(
        unseen_news_text,
        max_length = MAX_LENGHT,
        pad_to_max_length=True,
        truncation=True
    )

    unseen_seq = torch.tensor(tokens_unseen['input_ids'])
    unseen_mask = torch.tensor(tokens_unseen['attention_mask'])

    with torch.no_grad():
      preds = model(unseen_seq, unseen_mask)
      preds = preds.detach().cpu().numpy()

    preds = np.argmax(preds, axis = 1)
    preds

    array([1, 0, 1, 0])
```

## V. CONCLUSION

In this study, we compared the effectiveness of various uni-directional pre-training machine learning models. According to our analysis, the accuracy of the real-world fake news dataset is not up to par. Additionally, research towards a bidirectional training model—a more potent feature extractor—was prioritised. This feature inspired us to develop BERT, a pre-trained word embedding model based on a bidirectional transformer encoder. Compared to GloVe, BERT extracts features more effectively and produces results that are useful for NLP tasks. The accuracy of the experiments carried out utilising the BERT-based machine learning approach was 88%. A subset of machine learning called deep learning learns to describe the world as a nested hierarchy of concepts, which gives it immense power and flexibility. The ability of deep learning to perform feature engineering on its own is one of its key advantages over other machine learning techniques. In order to enable quicker learning, a deep learning algorithm will examine the data in search of features that correlate and integrate them.

Future work will involve developing a hybrid technique that applies to both the binary and multi-class real-world fake news datasets and combines content, context, and temporal level information from news stories. For multi-label datasets that spread in a graph, this hybrid approach can be useful in identifying instances of fake news. We will continue to examine the issue of false news from the perspective of various echo chambers found in social media data, which can be thought of as a collection of people who have the same position on any given social issue.

## REFERENCES

**[1].** COVID-19 Fake News Detection on Social Media by Khondoker Mirazul, Jahid Reza, Swarna Saha , Humayra Akter International Conference Paper Date of Publication: 6 June 2022Publisher: IC4ME2

**[2].** Stance Detection in Fake News : A Combined Feature Representation by Bilal Ghanem ,Francisco Rangel , Paolo Rosso Published in: Fact Extraction and VERification (FEVER)Date of Publication: 12 November 2021 Publisher: Association for Computational Linguistics

**[3].** Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM) by Gyu Sang Choi, Byung-

Won On Published in: IEEE Access (Volume: 8)Date of Publication: 26 August 2021Publisher: IEEE

**[4].** A Survey on Stance Detection for Mis- and Disinformation Identification byMomchil Hardalov, Arnav Arora, Preslav Nakov, Isabelle AugensteinInternational Conference on Data Mining Workshops (ICDMW))Published in: arXiv (2103.00242v3)Date of Publication: 8 May 2022 Publisher: Association for Computational Linguistics

**[5].** S. A. García, G. G. García, M. S. Prieto, A. J. M. Guerrero, and C. R. Jiménez, "The impact of term fake news on the scientific community scientific performance and mapping in web of science," Social Sciences, vol. 9, no. 5, 2020.

**[6].** S. Kogan, T. J. Moskowitz, and M. Niessner, "Fake News: Evidence from Financial Markets," 2019

**[7].** Kaggle - Churn Modelling Classification Data Set

**[8].** A. Robb, "Anatomy of a fake news scandal," Rolling Stone, vol. 1301, pp. 28–33, 2021

**[9].** F. T. Asr and M. Taboada, "Misinfotext: a collection of news articles, with false and true labels," 2019.)

**[10].** H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, "exBAKE: automatic fake news detection model based on bidirectional encoder representations from transformers (bert)," Applied Sciences, vol. 9, no. 19, 2020.

**[11].** W. Y. Wang, Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection, Association for Computational Linguistics, Stroudsburg, PA, USA, 2017.