STOR390 Midterm: Fake News Detectors

Anna Fetter

2024-10-25

introduction

Fake news is more than just a fear mongering buzzword: it's a reality in the digital age. The rapid growth of social media use, not just as a social outlet but as a news source, has laid a fertile foundation for the rapid growth of misinformation. Public trust in traditional news outlets is at an all time low and trust in social media is up. The gap in trust between social media and news outlets is especially close between young people, ages 18-29, with 52% saying they have a lot/some trust in social media outlets compared to 56% for national news organizations (Pew Research Center, 2024). Between this drop in public trust for news organizations and a rise in people relying on social media feeds for their news, the dangers of fake news are more pervasive than ever. There's also a societal concern around freedom of speech whenever the detection and deletion of fake news is brought up. Detecting and removing fake news from social media sites has been a project of major social media giants for years now, with no continuity between what each platform does with fake news. Major social media platforms have tackled fake news in varying ways — ranging from free speech absolutism to cautious censorship — but with little consistency across platforms. Social media is not a public square, like many believe, but a public court owned by tech giants. Governments struggle regulating fake news both because of the quantity and the private, often foreign, ownership of social media. As both real and fake news evolve, developing ethical fake news detectors becomes more challenging and essential.

introduction to the paper

Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI by Ehtesham Hashmi et. al explores and expands upon existing methods for fake news detectors. The stated goal of the paper is to advance the detection of fake news through regularization techniques, optimization techniques, and hyperparameter tuning. The authors utilized three publicly available datasets: WELFake, FakeNewsNet, and FakeNews Prediction. In the abstract, the researchers express concern over the widespread propagation of misinformation on social media. The researchers state that many individuals place unwavering support in social networks, without discerning the authenticity or origins of the information. To identify media-rich fake news, the researchers believe enhancing detection accuracy of fake news detectors is of the utmost importance to uphold the integrity of information systems in a rapidly changing cyberspace.

methods

The methodology of this study rely on applying established fake news detection methodologies to three publicly available datasets that differentiate between factual and fabricated news using binary classification. The researchers refined the existing detection methodologies using three techniques: regularization methods, optimization, and hyperparameter tuning. An extensive review of related work occurred first, so the researchers could eventually compare their results to previous results as well as understand the contemporary techniques other researchers were using to design, build, and implement fake news detectors.

In this binary classification study, 0 represents fake news and 1 represents real news. The research data set includes a large set of training and testing data. WELFake consists of 72,134 news articles with 35,028 of the

articles classified as fake news and 37,106 articles classified as real news. The creators of WELFake combined data from four already influential datasets including those from Kaggle, Reuters, and Buzzfeed Political. FakeNewsNet consists of 23,196 articles that relate to news content, social content, and spatiotemporal information. FakeNewsPrediction includes 3,164 articles. The datasets combined analyze over 100,000 instances.

The researchers performed extensive data preprocessing to improve the performance of the models. To preprocess the "text" column, which included all of the news comments, researchers made all letters lowercase, removed non-essential characters, tokenized words and sentences, used Python's ReGex library to filter and process other text elements, removed duplicate examples, and applied lemmatization. Lemmatization reduces words to their base or root forms, which improves consistency and improves the models' abilities to recognize similarities between different forms of the same word. The goal of this preprocessing was to enhance the performance of the models.

The researchers analyzed supervised and unsupervised FastText. FastText is an open-source library for word embeddings developed by Facebook AI Research(FAIR). FastText features an extensive 2 million words sourced from common crawl. FastText transforms words into vectors in continuous vector space.

Unsupervised FastText generates word vectors including subword information by breaking up the words into n-grams. N-grams are a series of 3-6 character word fragments. Word embeddings produce numerical representations to textual input. This helps unsupervised FastText models perform better by helping models recognize that words with similar subparts may be semantically related and also help the model have better performance outside of the training set. Researchers used the function *text_to_fasttest_embeddings* to generate embeddings for each word using *get_word_vector*. If the word is completely out of vocabulary, the function returns a zero vector.

Supervised FastText is used for text classification. Like unsupervised FastText, supervised FastText uses subword information. The key difference is that the supervised model is trained on a labeled training set where each text excerpt has an associated label. The model averages word vectors in a sequence to form the text representation that is used to predict the label. Supervised FastText can handle large datasets and large numbers of classes quickly and efficiently. The researchers found that supervised FastText consistently outperformed unsupervised FastText, proving the importance of labeled training data in text classification problems.

The researchers explored various types of machine learning (ML), deep learning (DL), and transformer-based models. For supervised ML-based models, researchers included decision trees (DT), support vector machines (SVM), logistic regression (LR), and random forest (RF) with boosting methods extreme gradient boosting (XGBoost), and Categorical Boosting (CatBoost). Researchers implemented DL models of long short-term memory (LTSM) and its variant BiLSTM, gated recurrent unit (GRU), and the hybrid convolutional neural network LSTM (CNN-LSTM). RNN-based models handle sequential data well, while LSTM models capture long-term dependencies. BiLSTM processes the data both forward and backwards, which enhances its abilities in complex sequential tasks. Regularization and hyperparameter training were especially important for the researchers in DL models. Through the use of kernel L2 regularization, models are encouraged to adopt smaller values for the weight of the model. This regularization both minimizes the likelihood of overfitting and protects the model's ability to generalize. To hyperparameter tune for DL models, researchers methodically adjusted the model's through targeted optimization over 10 epochs.The researchers employed text classification transformers BERT, XLNet, and RoBERTa. Transformer is a Natural Language Processing (NPL) system. The architecture of transformers helps them capture intricate contextual information. The transformers excel in understanding the context of a word by looking at the words that come before and after the word.

Since DL-based models operate as black boxes, the researchers also employ explainable AI (XAI) algorithms to find the words that contribute the most to the classification of a sentence as fake news. The researchers specifically use LIME to interpret multiple black box DL models.

summary of results

The researchers used four standard metrics to assess performance: accuracy, precision, recall, and F1 score. The researchers provided tables to show the results of ML and DL models using supervised and unsupervised FastText.

In unsupervised FastText models, the SVM classifier performed best across all three data sets in both accuracy and F1 scores with scores of 0.99 for all four performance metrics. The researchers found that the SVM classifier effectively handles high-dimensional data, creates clear decision boundaries, and navigates complex non-linear relationships which makes SVM a strong performer in fake news detector tasks. The other ML classifiers provided inconsistent results, especially compared to DL-based models that consistently maintain their performance. Researchers also discovered DL and ML algorithms stacked with supervised FastText outperformed their unsupervised FastText counterparts.

The transformer based models performed well; all models produced precision, recall, accuracy, and F1 scores above 0.95 on all data sets, which is better than some of its ML and DL counterparts with scores as low as 0.78 and most ranging in the low to mid .90s.

The researchers concluded that the deep learning model CNN-LSTM exceeded the performance of all other learning architectures presented in their research. The methods in this study improved the accuracy of fake news detection when compared to existing state-of-the-art research.The researchers compared their results to the baselines using WELFake and FakeNewsNet dataset; the accuracy of their methods was 0.99 accuracy compared to the baseline of 0.97.

Researchers used interpretability modeling, specifically LDA and local interpretable model-agnostic explantations (LIME), to make the complex learning processes of these detectors more clear. LIME provides local explanations, pointed to the words that were most influential to the decision. Examples of these words in some of the research tests include “disqualified”, “Obamacare”, “reuters”, “tv”, “president”, “thing”, “latest”, “revelation”, “email”, “Hillary”, “realDonaldTrump”, “FraudNewsCNN” and “point”. Each of these words has a weight attached to them, signifying how much is swayed the decision.

moral concerns

The most prominent normative concern when discussing the implementation of fake news detectors is balancing freedom of expression versus protecting from the harm produced from fake news. Many democratic societies, like the United States, value freedom of expression. In the United States, it’s against the 1st Amendment for government actors to restrict speech, except for a few edge cases of unprotected speech. However, since many of these fake news detectors are or will be deployed on social media platforms, private companies get to make many of the final calls on how strict they want their fake news detectors to be, if they even want to use them at all. Private corporations suppressing free speech could be a serious normative concern in its own right. Both suppression of speech and fake news have fascist roots. Falsely flagging fake news as real (false positives) and real news as fake (false negatives) may produce negative societal impacts. The normative concern when building fake news detectors is striking a balance between the two extremes of free speech absolutism and strict censorship of unpopular or fake news.

As revealed in this study through the use of explainable AI, words like “president” and “Obamacare” are more susceptible to being flagged as fake news. This poses a normative question: how can risk around political or election misinformation be mitigated without suppressing these topics by making the models over classify potentially contoversial news as fake?

conclusion

Fake news detection is more critical than ever as misinformation posts on social media evolve and shape public perception of reality. This study highlights the advancements in fake news detectors, combining advanced machine learning and deep learning models with FastText and explainable AI to improve accuracy

and crack some of the reasoning behind the black box models. The 0.99 accuracy, precision, recall, and F1 scores are quite good, but as fake news continues to evolve, there will soon be a need to refine these models even more. The use of LIME revealed that words relating to politics are more likely to flag news as fake. Ethical concerns arise in the creation and implementation of these detectors: where does society want to land on the spectrum of free speech absolutism and complete censorship? Who is responsible for this decision on speech? The contemporary challenge is to effectively detect fake news using fair and accurate detectors to protect democratic values.

references

E. Hashmi, S. Y. Yayilgan, M. M. Yamin, S. Ali and M. Abomhara, "Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI," in IEEE Access, vol. 12, pp. 44462-44480, 2024, doi: 10.1109/ACCESS.2024.3381038. https://ieeexplore.ieee.org/document/
Pew Research Center, "Republicans, young adults now nearly as likely to trust info from social media as from national news outlets," Pew Research Center, Oct. 16, 2024. [Online]. Available: https://www.pewresearch.org/short-reads/2024/10/16/republicans-young-adults-now-nearly-as-likely-to-trust-info-from-social-media-as-from-national-news-outlets