STOR 390 Final: Fake News Detectors

Anna Fetter

12/12/2024

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 as a newsource has laid a fertile foundation for the rapid spread 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% trust for national news organizations (Pew Research, 2024). Between this drop in public trust for news organizations and a rise in people getting their news from social media, the dangers of fake news are becoming more pervasive.

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 the process. Major social media platforms tackle fake news in varying ways — ranging from free speech absolutism to cautious censorship — but with little consistency across platforms. Governments struggle regulating fake news both because of the quantity and the private, often foreign, ownership of social media companies. As both real and fake news evolve, developing ethical fake news detectors is more challenging and more essential. However, the landscape of fake news detection is increasingly opaque, with social media companies restricting access to data by locking down APIs. While tools like Facebook's fastText remain publicly available, these restrictions hinder researchers' ability to develop and refine detection systems, raising concerns about transparency in the fight against misinformation. The goal of this paper is to explore the methods in *Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI by Ehtesham Hashmi et. al* and examine the moral considerations that need to be considered surrounding the implementation and further development of fake news detectors.

Summary of Methods

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 detection. The paper was published in IEEE in March 2024, which is a reputable journal that gives this paper legitimacy and trustworthiness. The stated goal of the paper is to advance the detection of fake news through regularization, optimization, and hyperparameter tuning.

The study applied established fake news detection methodologies to three publicly available datasets, refining these approaches with techniques like regularization, optimization, and hyperparameter tuning. Using binary classification, with 0 representing fake news and 1 representing real news, the study analyzed datasets WELFake, FakeNewsNet, and FakeNewsPrediction, totaling over 100,000 instances. Data preprocessing included lowercase conversion, tokenization, lemmatization, and text cleanup to enhance model performance. The study leveraged supervised and unsupervised fastText for word embeddings.

The researchers tested a range of machine learning (ML), deep learning (DL), and transformer-based models. ML models included decision trees, logistic regression, and support vector machines, while DL approaches used LSTM, BiLSTM, and CNN-LSTM architectures to capture sequential and contextual information.

Regularization techniques minimized overfitting and hyperparameter tuning optimized performance. Transformers like BERT, XLNet, and RoBERTa excelled at understanding contextual nuances in text. Additionally, explainable AI (XAI) algorithms, including LIME, provided insights into the key features influencing classification decisions.

Replication & Analysis of Methods

In my replication of the work produced in *Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI by Ehtesham Hashmi et. al*, I focused on training an SVM to predict fake news, implementing Facebook's open-source fastText, and using explainable AI LIME to understand why my SVM classified news entries the way it did. I solely used the publicly available WELFake dataset which contains over 70,000 entries. The following section traces through the steps I took and the output I got from each step. I completed this analysis using R instead of Python, a marked deviation from the original paper. I also needed to install Java to use the official fastText package from Facebook that's publically available on GitHub.

1. Load WELFake and preprocess the dataset.

I started with downloading WELFake off of Kaggle. Once I previewed the data, I began preprocessing the text to prepare the data for the models. I converted all text to lowercase, removed punctuation, stopwords and numbers, and performed word stemming. I used the snowballC and tm libraries available in R for this data preprocessing.

Table 1: Text Preprocessing Example

Original Text	Preprocessed Text
No comment is expected from Barack Obama Members of the #FYF911 or #FukYoFlag and #BlackLivesMatter movements called for the lynching and hanging of white people and cops. They encouraged others on a radio show Tuesday night to turn the tide and kill white people and cops to send a message about t	comment expected barack obama members fyf fukyoflag blacklivesmatter movements called lynching hanging white people cops encouraged others radio show tuesday night turn tide kill white people cops send message killing black people americaone fyoflag organizers called sunshine radio blog show hosted

2. Set the training and testing partition.

Once my data was preprocessed, I set a seed, shuffled the indices, and set the 80/20 training/testing partition. I also needed to make a corpus— the data science term that refers to a large collection of text used to train a model— and convert the textual input into a document term matrix, which counts the frequency of terms in a dataset. To improve efficiency, I removed terms that appeared in less than 5% of the dataset. This step made training my SVM more feasible using solely my MacBook and R.

3. Train the SVM.

I trained an SVM using a sample of 7000 observations from the training partition. While using the full testing set could have provided richer insights, anything above 7000 observations crashed my computer. I used a linear kernel like the authors suggested because LIME assumes linearity on local models. I used a cost of 1. The output of the SVM yielded 859 support vectors.

4. Use the SVM to predict the classes of the test set. Evaluate the accuracy.

After training the SVM, I used it to classify entries in the testing set. The model demonstrated strong performance, achieving an accuracy rate of 91%. Given that I only trained on a fraction of the training set, I was happy with the results. The accuracy rate would likely have been even higher if I could train on the full 50,000 observations in the training set. The confusion matrix also revealed a relative balance between false positives and false negatives.

Table 2: SVM Model Confusion Matrix for News Classification

	Predicted Fake (0)	Predicted Real (1)
Actual Fake (0)	6286	641
Actual Real (1)	639	6861

To align more closely with the metrics provided in *Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI by Ehtesham Hashmi et. al* I also analyzed precision, accuracy, recall, and the f1-score.

Table 3: SVM Model Performance Metrics

Metric	Score
Accuracy	0.9113
Precision	0.9075
Recall	0.9077
F1-Score	0.9076

5. Integrate Facebook's fastText

I integrated Facebook's fastText for word embeddings. I used the pre-trained word vectors available from Facebook's GitHub repository and mapped them to the preprocessed text. To get this up and running, I downloaded fastText to my computer and used the text package in R. FastText achieved a 97% accuracy rate on the testing set, performing better than my SVM.

Table 4: FastText Model Confusion Matrix for News Classification

	Predicted Fake (0)	Predicted Real (1)
Actual Fake (0)	6731	196
Actual Real (1)	194	7306

Table 5: FastText Model Performance Metrics

Metric	Score
Accuracy	0.9730
Precision	0.9717
Recall	0.9720
F1-Score	0.9718

6. Use LIME to create a visualization of the model's decisions.

Finally, I applied LIME (Local Interpretable Model-agnostic Explanations) to visualize the decisions made by the SVM. LIME enabled me to identify the most influential words that contributed to the model's

classification of a given news entry as fake or real. Since the SVM wasn't a default option for LIME, I also had to create an SVM wrapper function. I trained LIME on a small subset of only 1000 instances since anything above that would crash my computer. After using LIME to create an explanation function, I was able to generate a table that showed the words that carried the highest weights in the SVM classifying news entries as real or fake. With LIME, I was able to both see the words that held weight towards news classifying as real or fake within the testing and view misclassified instances.

Table 6: LIME Feature Importance Analysis

	Impact.Direction	Impact.Strength	Average.Impact	Times.Used
team	Real News	Very Strong	0.8672	1
business	Fake News	Very Strong	0.8628	1
risk	Real News	Very Strong	0.8615	1
source	Fake News	Very Strong	0.8597	1
meeting	Real News	Very Strong	0.8578	1
X.m	Fake News	Very Strong	0.8557	1
four	Real News	Very Strong	0.8536	1
chief	Fake News	Very Strong	0.8531	2
didn.t	Real News	Very Strong	0.8527	1
decision	Real News	Very Strong	0.8517	1

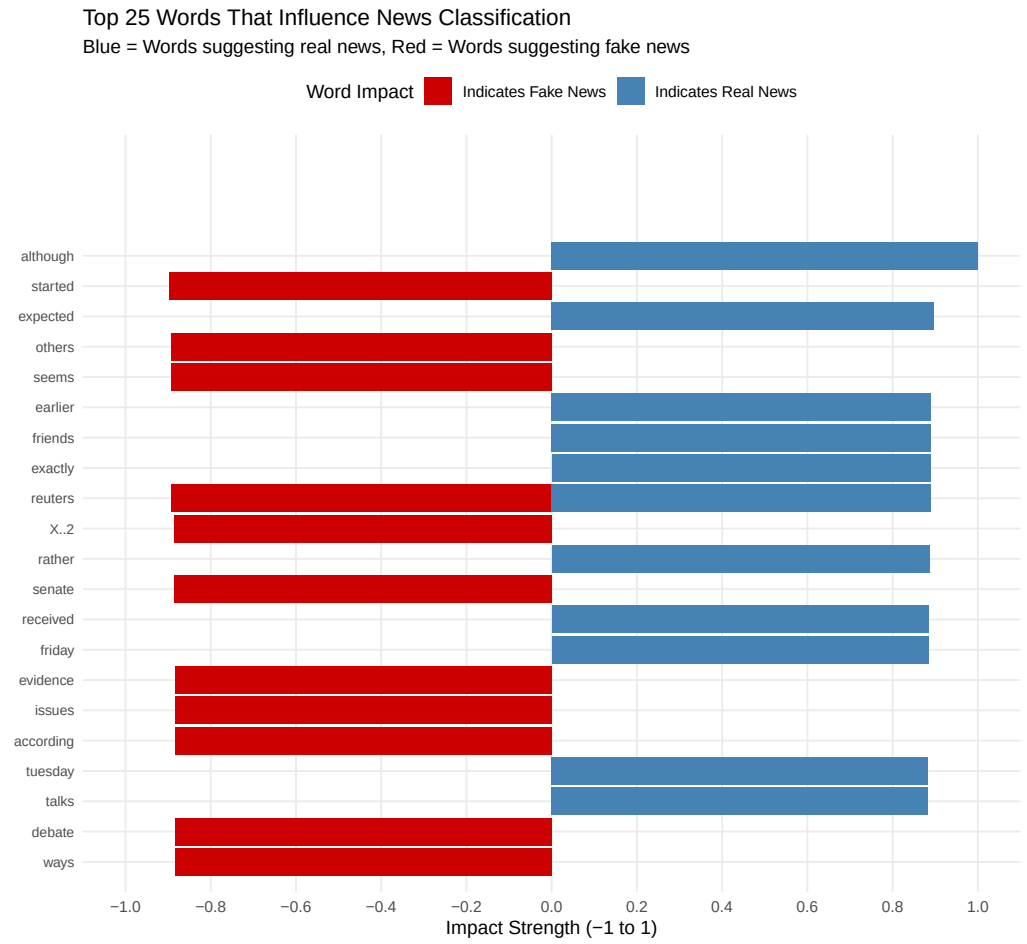

Table 7: Examples of Misclassified News Articles

Type of Error	Article Text
Fake News Classified as Real	Videos Israel Tracked 'Anti-Government' Journalists On Facebook Netanyahu thinks the new channel doesn't have enough government supervision and is too critical of his government and policies. An Isr...
Fake News Classified as Real	During last week's Berlin InnoTrans trade show, French company Alstom unveiled the world's first zero-emissions hydrogen-powered train. According to The Local, the Coradia iLint hydrogen train... ...
Real News Classified as Fake	(Reuters) - Bill Clinton faced down protesters for 10 minutes at a presidential campaign rally in Philadelphia for his wife, Hillary Clinton, over their criticisms that a 1994 crime bill he approved w...
Real News Classified as Fake	Europe is struggling with the greatest mass-movement of people across its borders since World War II. And the crisis has exposed the limitations and gaps in its immigration system. A Greek riot polic...

Critique of Methodology

After applying some of the methods presented in *Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI*, I'm left with very few critiques on the authors' methods. They were incredibly thorough in the sheer number of machine learning and deep learning models they used. They ran their simulations on significantly more data than I did, leading to better accuracy. Additionally, they were able to stack fastText on top of their SVM, and I used these separately. Their research focused much more on tuning than implementation.

I would be interested to see these researchers apply their work to different datasets and expand into different mediums. Currently, the authors are only examining text-based input, and many of their methods such as fastText and LIME, are only made to examine text-based input. I would be interested in seeing if their findings around model tuning and how different models perform could be generalized into fake news detection for photo and video format, or even textual input with emojis. Given how long some of these models already took to run, I would be curious how researchers would accommodate for much larger file sizes for non-textual input.

Addressing Normative Concerns

This novel analysis of these methods strengthens the case for integrating fake news detectors into social media platforms. The implementation of the basics of fake news detectors through training an SVM, implementing Facebook's pretrained fasttext model, and using explainable AI LIME proved to be less complex than what is presented in mainstream culture about the complexity of machine learning. Implementing fake news detectors seems to be very feasible, at least for text based input. The novel analysis presented proves that with just the text content, not even the username or geographical data of the input, SVM and fastText predicted at rates over 90%. During my personal implementation, I could only train the SVM on 7,000 entries- 10% of the total data in WELFake- and still achieved 91% accuracy. With a larger dataset and more precise tuning, the authors of *Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI by Ehtesham Hashmi et. al* proved how feasible accuracy rates between 97-99% are for these detectors. For the small percentage of real news that is classified as fake, traditional media outlets could fill the gap in investigating these leeds.

While technically feasible, implementing fake news detectors raises complex ethical questions about informed consent, user privacy, and broader societal implications.

Utilitarian philosopher JS Mill's principle, *volenti non fit injuria* —"no injury is done to the willing"— offers a lens to examine the ethical implications of user consent in the context of the harm principle(Stanford Encyclopedia of Philosophy, 2016). JS Mill's harm principle states that personal autonomy is checked right at the point where it causes objective harm to another moral agent. There is a risk of harm when implementing fake news detectors, including censorship of marginalized or radical online communities whose content falls outside the bounds of "normal content". Following a strict utilitarian view, most social media companies have already evaded this harm because they already have consent from their users to use their posts to train AI models. Twitter's new clause in its privacy policy which became effective on November 15, 2024 reads: "We may use the information we collect and publicly available information to help train our machine learning or artificial intelligence models for the purposes outlined in this policy"(Twitter, 2024). While this is a step towards consent, social media giants often rely on tacit or coerced consent, leaving their users unaware of the extent of what they're consenting to.

Tacit consent is the implied consent users grant by using a good or service. This could look like users automatically granting a social media platform access to their Tweets to train models on by default. To turn off this feature, users would have to comb through their settings to turn it off. Even when consent is explicit — picture the popup window that pops up every few months to accept a platform's revised terms and conditions when a user opens the app — these platforms intentionally design this user flow to be coercive so that user's accept the conditions quickly without reading all the changes. According to German philosopher Immanuel Kant's categorical imperative, which emphasizes treating individuals as ends in themselves rather than as means to an end, failing to secure fully informed, explicit consent violates the moral obligation to respect user autonomy.

Viewing the implementation of fake news detectors solely through the lens of user consent is insufficient due to the disagreements around consent from philosophers themselves. This lens on consent only considers the consent of those who post the content, not the consumers of the content. Users of these platforms also tacitly consent to not knowing what's real or what's fake on social media platforms. Under this argument, the implementation of fake news detectors is justifiable because it weighs in favor of both the users consenting through terms of service and protects the consumers of the content from fake news.

Privacy law in the United States is predicated on the belief that there "exists no reasonable expectation of privacy" in a public place. Social media has proven to be a legal minefield over the past decade when it comes to digital applications of existing privacy laws. Posts on social media are presumed automatically to be public and published. But when it comes to user data and demographic information that is stored within the app, this raises much more debate around user privacy.

Text-based models, like those used in the novel analysis with SVM and fastText, focus solely on the content of the posts without requiring additional personal data such as usernames or geographical data. Since there is minimal use of sensitive personal data, other than that of which the user willingly posted in the content of the post, there is minimal privacy risk. However, now social media news is often in photo or video content on platforms like Instagram and TikTok. Simply transcribing this content into plain text form would not be enough. Fake news detectors would take in video and photo content. Detecting fake news in these mediums requires the models to analyze more sensitive data. Facial recognition technology could be implemented and social media giants could use additional metadata such as location tags, timestamps, and device information. Video platforms also have additional behavioral data from the content consumers such as watchtime and demographics who are heavily engaging with the content. This enhanced data collection as social media companies evolve brings up extreme privacy risks for users in general, not just with the implementation of fake news detectors. Many social media giants would use this information to place targeted advertisements and improve the consumer's feed, even if the company chose not to implement a fake news detector so this risk is not unique.

Open-source fake news detectors promote transparency and accountability, which are crucial features when it comes to gaining the public trust necessary to combat the spread of misinformation. By making the models publicly accessible, researchers, developers, and the public can scrutinize how fake new detectors work and

why certain content may get flagged as fake. This openness helps build trust and can help uncover hiddean biases or unjustified censorship of certain content.

However, training these models requires lots of user data. While the data might not be super sensitive in text-based models, more complex photo and video models or those that utilize user metadata pose significant security risks. To protect user privacy and security, many social media giants make it very hard if not impossible to view or download this type of data. This may make it more difficult for researchers and developers to develop open-source models. If researchers do get access to this type of training data, they should employ techniques to protect the users by anonymizing datasets, stripping out metadata that could compromise user identities, and utilizing differential privacy techniques. Still, this could be challenging as more of these characteristics might lead to models being able to better predict what news is fake, since certain internet personalities are known for spreading fake news. Like many other types of machine learning models, training and implementing fake news detectors presents a valid conflict between a perfectly transparent open-source model and user privacy and consent.

Even with weighing the risks to user consent and privacy, fakes news detectors are still a net good. Existing tools such as SVM and fastText yield remarkable accuracy, thus minimizing the risk of inaccurately classifying real news as fake. As with concerns around freedom of expression, these social media platforms themselves get to choose if they want to use them because their terms and conditions rule the terms of their apps. The resources that these tech giants have leave them with no excuse when it comes to implementing these fake news detectors to keep their platforms safer.

Conclusion

Fake news detectors are essential in fighting the dissemination of fake news online. Through my replication of methods outlines in *Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI*, I showed the feasibility of using SVM and fastText to classify text-based fake news with accuracy rates surpassing 90%, even with minimal tuning and computational contraints. This replication confirms the potential of fake news to acheive results comparable to the original study. The research presented in *Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI* shows the promise of fake news detectors built using a wide variety of fake news detectors. The research sets important benchmarks for what a well-trained model can acheive in fake news detection.

This novel analysis and the accompanying moral debate also highlighted the importance of not only the use of fake news detectors, but the importance of open-source fake news detectors to foster public trust and ensure transparency. At the same time, the moral considerations brought up the importance of user privacy and consent. As fake news detectors continue to develop to accomodate multimedia formats, these considerations will further be amplified. Balancing the pros of transparency and the risks to privacy highlight the continued monitoring of the implementation of fake news detectors.

Despite these challenges, the benefits of fake news detectors far outweigh the risks. Their ability to identify misinformation at scale provides an important line of defense against the dangers of fake news, making them an invaluable resource in the digital age. Social media platforms have the resources to deploy these tools, and they should take responsibility for creating a safer, more trustworthy online environment.

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

Stanford Encyclopedia of Philosophy, "John Stuart Mill: Moral, Social, and Political Philosophy," Stanford Encyclopedia of Philosophy, Apr. 12, 2016. [Online]. Available: https://plato.stanford.edu/entries/mill-moral-political/

Twitter, "Terms of Service," Twitter, Nov. 15, 2024. [Online]. Available: https://twitter.com/en/tos