DeFactify

DELIVERY DOCUMENT JUNE 6, 2025

Ivet Kalcheva Version 2.0 (Final)

Contents

Summary

This document introduces DeFactify, an AI project designed to detect fake news and fight misinformation. DeFactify analyses textual patterns in news articles to deliver answers to its users' questions – whether the news article is truthful or not. The document will explore the project's development, its societal impact, future possibilities and recommendations in depth.

Project Overview

Current State of the Project

DeFactify has progressed significantly through stakeholder, teachers and peers' feedback, transitioning from using a questionable and outdated dataset to building a more reliable and current one. The project is now equipped with a robust Stacking Classifier model that demonstrates a high accuracy rate of 94.15%, focusing on minimising false positives. While the dataset currently includes approximately 1,000 entries, it offers greater authenticity and relevance compared to the initial iteration. Efforts are ongoing to address dataset limitations, expand coverage and refine model performance to align with real-world complexities.

Domain Analysis

DeFactify operates in the field of misinformation detection, focusing on identifying fake news across various categories, including politics, health and technology. By analysing the textual content and titles of news articles, DeFactify detects deceptive patterns, providing a valuable resource to combat misinformation.

This initiative was made possible through an extensive **literature review** of previous studies on fake news detection, **document analysis** to identify differences and an examination of both fake and real news articles to gather data. Additionally, **model evaluation** was conducted to implement and assess the accuracy of various models on the dataset.

The feedback collected from students, teachers and stakeholder was instrumental throughout the project, helping it reaches its current stage. The user interface played a significant role in demonstrating and showcasing the model's capabilities.

It complied with legal standards; I only scraped news from publicly available and reliable sources. Although I attempted to scrape from sites with restrictions, those efforts were unsuccessful, leading me to believe I was following regulations.

Short-Term Possibilities

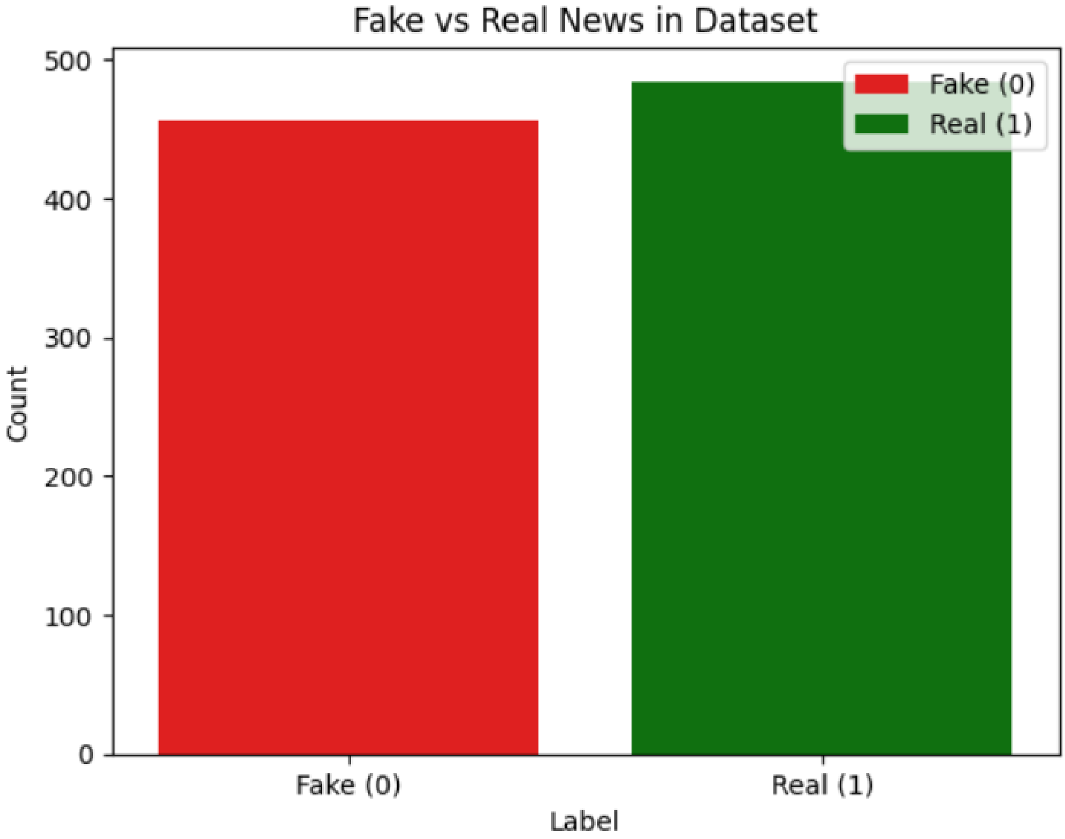
1. **Enhance Dataset Diversity:**
   - ✦ Collaborate with news agencies and organizations to access diverse, high-quality datasets.
   - ✦ Include data from forums, blogs and less formal platforms to test robustness in varied contexts.
   - ✦ Ensure regional and cultural inclusivity in the dataset to handle globally relevant nuances.
2. **Automate Dataset Expansion with APIs:**
   - ✦ Leverage APIs from news platforms, social media or RSS feeds for continuous dataset enrichment.
   - ✦ Use natural language processing (NLP) to extract key information from unstructured data sources like PDF reports or images.
3. **Real-Time Data Validation:**
   - ✦ Use real-time web scraping and APIs to cross-reference news against verified sources.
   - ✦ Integrate mechanisms for flagging discrepancies between real-time data and predictions.
4. **User-Friendly Interfaces:**
   - ✦ Develop dashboards with visual insights into prediction confidence, model reasoning and detected biases.
   - ✦ Offer customization options for stakeholders, such as adjusting detection sensitivity or filtering by topics of interest.
5. **Explore Advanced Models:**
   - ✦ Research techniques for enhancing explainability, such as attention heatmaps or decision pathways, to explain why an article might be fake or truthful.
   - ✦ Test ensemble methods combining linguistic, sentiment and statistical features for enhanced accuracy.
   - ✦ Include the ability to detect satire as an optional feature, distinguishing it from malicious misinformation.

Long-Term Possibilities

1. **Collaboration with Fact-Checking Organizations:**
   - ✦ Partner with global entities like FactCheck.org, Snopes or regional equivalents to validate data.
2. **Integration with Major Platforms:**
   - ✦ Embed DeFactify as a plugin or API for platforms like Facebook, Twitter, and Google Search.
   - ✦ Develop browser extensions or mobile apps for instant article verification during browsing.
3. **Multilingual and Multimedia Detection:**
   - ✦ Expand NLP capabilities to support multiple languages and dialects for wider applicability.
   - ✦ Incorporate image and video verification using deepfake detection and metadata analysis.
4. **Cloud Migration and Scalability:**
   - ✦ Shift to scalable cloud platforms like AWS, Azure, or Google Cloud for better performance and availability.
   - ✦ Enable distributed training and inference to handle large-scale data efficiently.
5. **Ethical and Legal Compliance:**
   - ✦ Collaborate with legal experts to ensure DeFactify adheres to regional laws and privacy standards.
   - ✦ Develop ethical guidelines for its use to prevent misuse, such as censorship or biased detection.
   - ✦ Foster a community of users and developers contributing to dataset enrichment, feature suggestions, and testing.

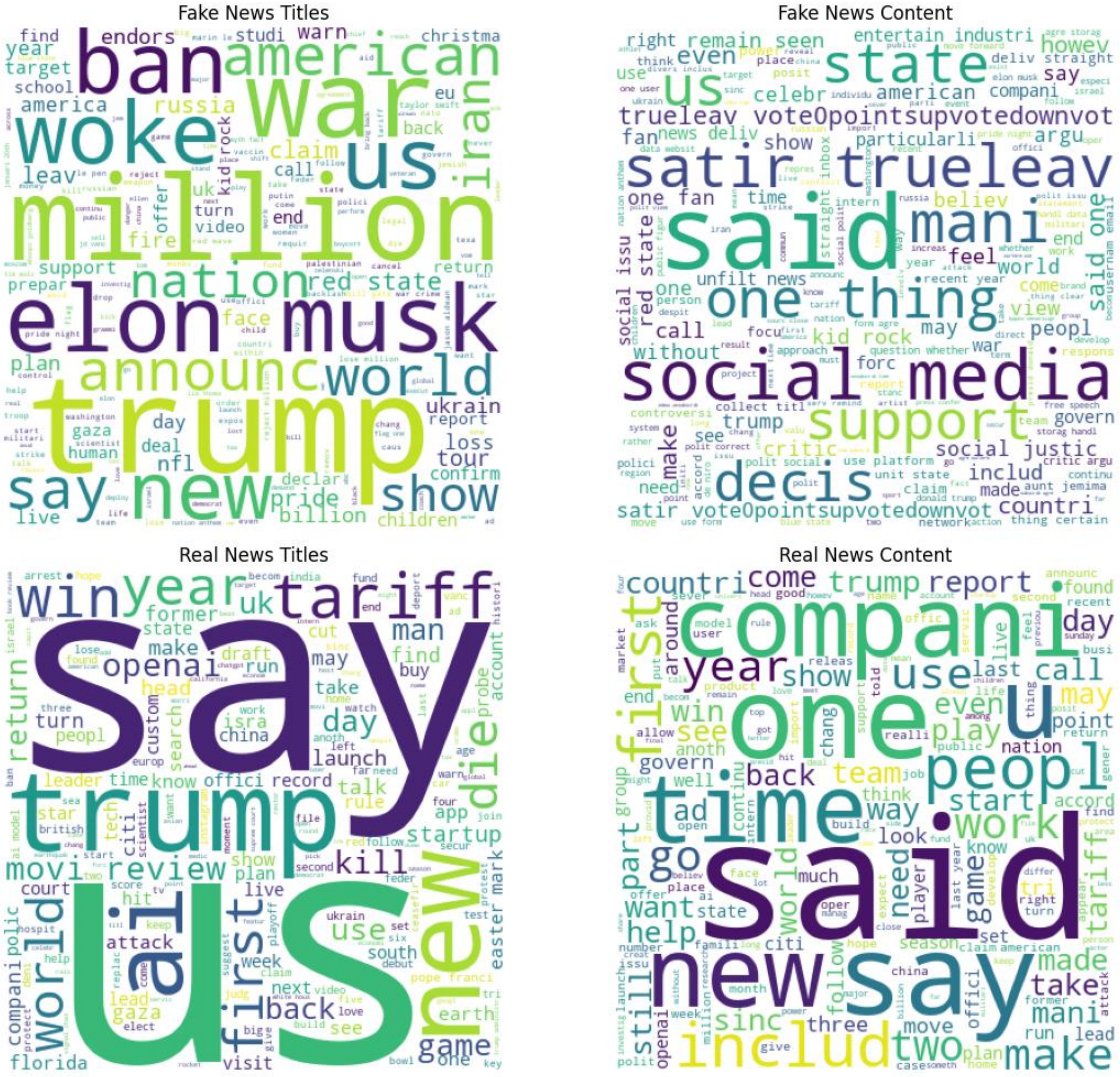
Data Collection

For iteration 0, an initial dataset containing 30,016 entries of fake and real news from 2016–2017 was used. However, its authenticity could not be guaranteed, as the dataset's origin was unclear, and it was both outdated and potentially unreliable. To address these concerns, a new dataset was created by scraping recent news articles from trusted sources over the past months. While the new dataset is smaller – comprising approximately 1,000 rows – it is significantly more authentic and relevant, providing sufficient diversity for initial analysis.



Exploratory Data Analysis (EDA)

- **Fake News Characteristics:** Use of emotionally charged language and exaggerated claims.
- **Real News Characteristics:** Neutral tone and reliance on factual descriptions.
- **Visualization Tools:** Word clouds and frequency charts highlighted distinct term patterns in each class.


Fake News Titles

Fake News Content

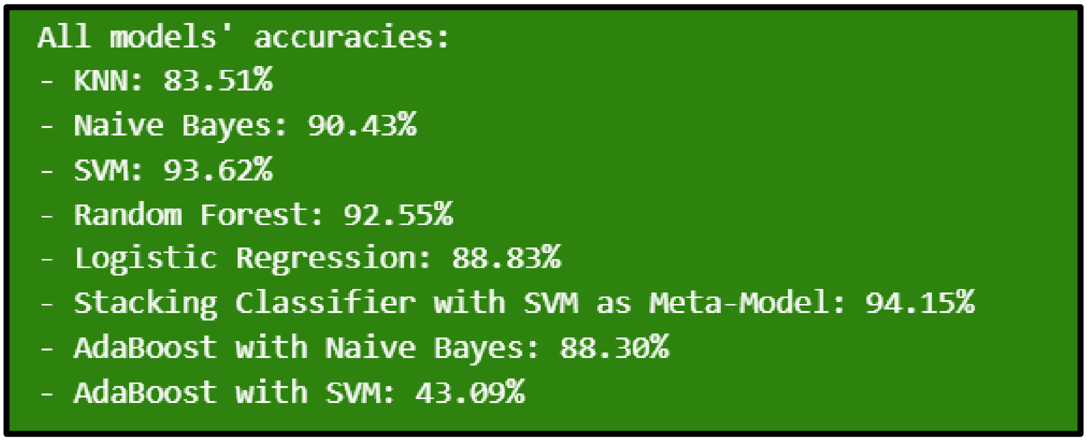
Real News Titles

Real News Content

Machine Learning Model

Model Performance

- **Best Model:** Stacking Classifier with SVM meta-model (accuracy: 94.15%, low false positives).
- **Other Models:** Simpler models (e.g., KNN, Logistic Regression) and certain ensemble methods (e.g., AdaBoost with SVM) were unsuitable due to higher false positive rates.



```
All models' accuracies:
- KNN: 83.51%
- Naive Bayes: 90.43%
- SVM: 93.62%
- Random Forest: 92.55%
- Logistic Regression: 88.83%
- Stacking Classifier with SVM as Meta-Model: 94.15%
- AdaBoost with Naive Bayes: 88.30%
- AdaBoost with SVM: 43.09%
```

Challenges

- **False Positives:** Labelling fake news as real can cause significant misinformation.
- **Borderline Cases:** Some misclassifications necessitate human oversight to ensure accuracy.

Explainable AI

Importance of Explainability

Explainable AI ensures that DeFactify's predictions are interpretable and trustworthy. Providing clear, understandable insights into why a news article is classified as fake or real is crucial for building confidence among users and stakeholders. This is achieved through detailed documentation provided in the notebook and research document, as well as a user-friendly interface designed for intuitive model testing.

Implementation in DeFactify

- **User Interface:** A user-friendly interface facilitates easy model interaction and understanding for non-technical users.
- **Documentation:** Comprehensive documentation ensures that users can understand the model's logic and assumptions, enhancing trust and accountability.

Benefits of Explainability

- **Trust:** By providing clear explanations, stakeholders can better understand and rely on the system's outputs, reducing scepticism.
- **Ethical Deployment:** Transparency in the model's workings ensures responsible use, fostering accountability and compliance with ethical standards.

Ethical Considerations

Bias and Fairness

Although this may seem like an obvious consideration, it is important to address biases and fairness. The limited size and scope of the dataset can inadvertently introduce biases, potentially skewing model predictions. For example, underrepresentation of certain topics, sources, languages or regions may lead to systematic errors in classification. These biases can result in some articles being unfairly labelled, undermining the system's credibility and inclusivity.

To mitigate these biases, a twofold approach is essential. First, it is crucial to expand the dataset to include a diverse range of topics, sources, languages and regions, ensuring balanced representation. This expansion will make the system more robust and better equipped to handle varied contexts. Second, conducting regular fairness audits can help identify and address biases in both the data and the model's predictions, thereby promoting equity and inclusivity.

Transparency and Interpretability

Second in importance, but not the least, is the transparency of models, which fosters trust and empowers users to make informed decisions. Currently, DeFactify achieves transparency primarily through its documentation and accompanying notebook. However, the long-term goal is to develop a system that utilizes Explainable AI. This system would not only provide the likelihood of an article being true or false but also explain the reasoning behind its predictions, which is essential for building user confidence.

Future enhancements aim to include advanced visualization tools to illustrate the impact of specific words or phrases, along with an intuitive user interface that allows stakeholders to easily verify and contextualize results. At present, the project's transparency relies on its comprehensive documentation and research, offering users insights into the methodology and decision-making process. By making the workings of the system understandable, DeFactify lays the groundwork for future improvements in explainability.

Societal Impact

DeFactify has significant societal implications, particularly regarding freedom of speech and the potential for unintended consequences. Mislabelling legitimate news as fake can damage the credibility of journalists and media outlets, potentially silencing important voices. This risk underscores the need to carefully evaluate the system's outputs and avoid over-reliance on automation. Over time, misclassifications could erode public trust in media and technology, especially if the system is widely adopted without proper checks and balances.

To mitigate these risks, it is crucial to encourage critical thinking among users. DeFactify's predictions should be viewed as one tool among many for evaluating information. Users must consider additional sources and context to reach informed conclusions. Promoting media literacy and critical analysis is essential for reducing dependence on automated systems. Furthermore, ensuring that the system acts as an assistive tool rather than the final arbiter of truth can help protect freedom of speech and foster a healthier information ecosystem.

Automation and Human Oversight

DeFactify utilizes advanced AI to identify patterns in fake news, but it should not replace human expertise. The interaction between automation and human oversight is essential for achieving reliable results. The strength of DeFactify lies in its ability to process large volumes of data quickly and detect potential fake news patterns on a large scale. However, complex or borderline cases often require a nuanced understanding of context, language and intent that only humans can provide.

Incorporating human review into the system ensures that final decisions take into account broader contexts and ethical considerations. For instance, human fact-checkers can verify borderline cases flagged by the model, adding an extra layer of scrutiny. This collaboration between machine predictions and human expertise reduces the risk of misclassification, reinforces accountability and enhances the reliability of the system's outputs. Ultimately, this balanced approach allows DeFactify to remain a valuable tool while acknowledging the complexities of evaluating real-world information.

Data Privacy and Security

An often overlooked yet vital ethical consideration is data privacy and security. DeFactify's operations involve collecting, analysing and validating large volumes of data, which may include sensitive or personally identifiable information (PII). If mishandled, this data can lead to privacy violations, misuse or even legal repercussions, severely impacting the system's credibility and user trust.

To mitigate these risks, DeFactify must prioritize privacy through a comprehensive approach. First, all data collection processes should adhere strictly to ethical standards, ensuring that no data is sourced from unauthorized or restricted platforms. Second, implementing robust encryption and anonymization techniques is essential to safeguard any sensitive information processed by the system. These measures will ensure that personal details are protected and cannot be traced back to individuals.

Additionally, transparency in data handling practices is crucial. Clear and accessible documentation should outline how data is collected, stored and used, ensuring compliance with global data protection regulations like GDPR and CCPA. Regular privacy audits can further enhance confidence by identifying and addressing potential vulnerabilities.

By embedding privacy and security into its design, DeFactify not only meets legal requirements but also reinforces its commitment to ethical integrity, fostering a trustworthy environment for its users and stakeholders.

Final Recommendations

For Stakeholders

- **Adopt With Caution:** While DeFactify is an effective tool for combating misinformation, it should be viewed as a supplementary system rather than a standalone solution. Stakeholders must validate the model's predictions, particularly in high-stakes situations, to ensure reliability and accuracy. Combining automated analysis with expert human oversight can significantly reduce errors and improve decision-making.
- **Expand Dataset:** Broader datasets are critical to mitigating biases and ensuring that the model can generalize effectively across diverse contexts. This includes incorporating multilingual data, underrepresented regions and niche topics to improve the system's inclusivity and robustness.
- **Foster Transparency:** Transparency is vital for gaining stakeholder trust. DeFactify should continue developing tools that make its reasoning process clear, such as explanation modules that show why specific articles were classified as fake or truthful. These tools can empower users to better understand and interpret the system's outputs, making them active participants in the decision-making process.
- **Collaborate Broadly:** Engaging with fact-checking organizations, journalists and policymakers can refine DeFactify's use cases and ensure its ethical deployment. These collaborations can also facilitate the development of standardized guidelines for misinformation detection, helping to align the tool with broader societal goals.

Use Case Suitability

DeFactify is suitable for organizations seeking to combat misinformation in a structured manner. In its current state:

- **Short Term:** It is particularly effective at identifying blatant patterns of fake news, such as articles with fabricated sources or sensationalist headlines. This makes it suitable for organizations looking for immediate support in reducing obvious misinformation threats.

- **Long Term:** With further advancements, DeFactify has the potential to become an integral part of comprehensive misinformation prevention systems. These systems could integrate data from multiple sources, incorporate advanced verification techniques and apply ethical safeguards to ensure fair and responsible deployment. Expanding its scope to detect nuanced misinformation, such as satire or manipulated media, could make it indispensable in the fight against fake news.

Consequences of Use

- **Positive:** DeFactify provides a powerful mechanism for flagging and addressing fake news efficiently. It enables organizations to act quickly, reducing the spread of misinformation and protecting the public from its harmful effects. The tool also promotes media literacy by encouraging users to critically evaluate news content.
- **Negative:** There is a risk of reputational damage if legitimate news is misclassified as fake. This could lead to distrust among users and stakeholders, potentially undermining the system's credibility. To mitigate this, human oversight must be integrated into the workflow, particularly for cases that are ambiguous or context sensitive.

Conclusion

DeFactify is a promising tool in the effort to reduce the spread of misinformation. Although its current performance is promising, ongoing improvements and careful implementation are essential for its continued success. By incorporating ethical considerations, expanding its datasets and encouraging collaboration, DeFactify can become a key resource in the battle against fake news.

References

- ✦ **Grammarly.** (n.d.). Writing assistance software. https://www.grammarly.com