# DeFactify

DELIVERY DOCUMENT JUNE 1, 2025


Ivet Kalcheva Version 1.0 (Final)

# Contents

# Summary

This document introduces DeFactify, an AI-powered tool designed to detect fake news and address the widespread issue of misinformation. DeFactify analyses textual patterns in news articles to deliver meaningful insights and recommendations. The document will explore the project's development, its societal impact and future possibilities.

# Project Overview

## Current State of the Project

DeFactify has progressed significantly, transitioning from using a questionable and outdated dataset to building a more reliable and current one.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 minimizing 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 domain of misinformation detection, targeting fake news across categories like politics, health, technology, etc. By analysing news articles' textual content and titles, it identifies deceptive patterns, offering a resource to combat misinformation for stakeholders such as media platforms, policymakers and the public.

## Short-Term Possibilities

- Enhance dataset diversity to improve model robustness (on sources and categories for broader knowledge).
- Validate model predictions with real-time data scraping for ongoing relevance.
- Develop user interfaces to simplify model interpretation for non-technical stakeholders.
- Explore more techniques and language models to improve performance on nuanced cases.

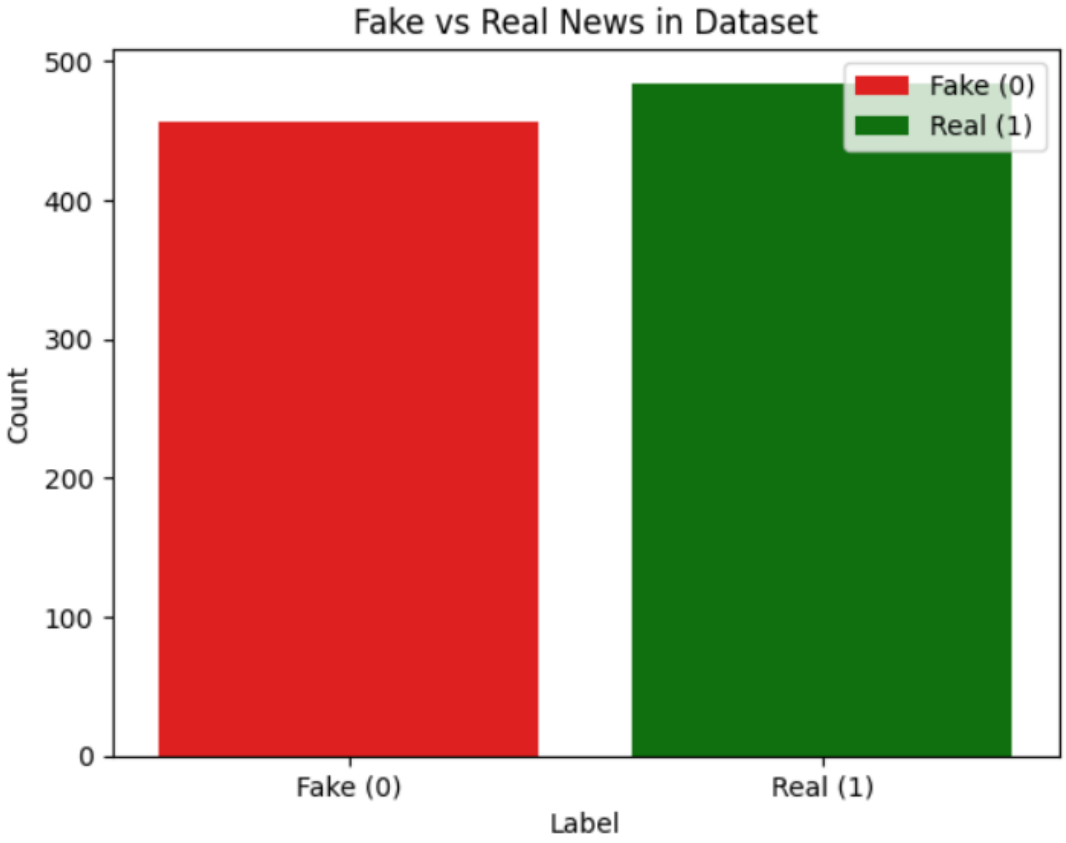
## Long-Term Possibilities

- Collaborate with global fact-checking organizations to standardize detection mechanisms.
- Integrate DeFactify into social media platforms and search engines to prevent the spread of fake news.
- Expand into multilingual and multimedia fake news detection.

# Data Analysis and Insights

## Data Collection

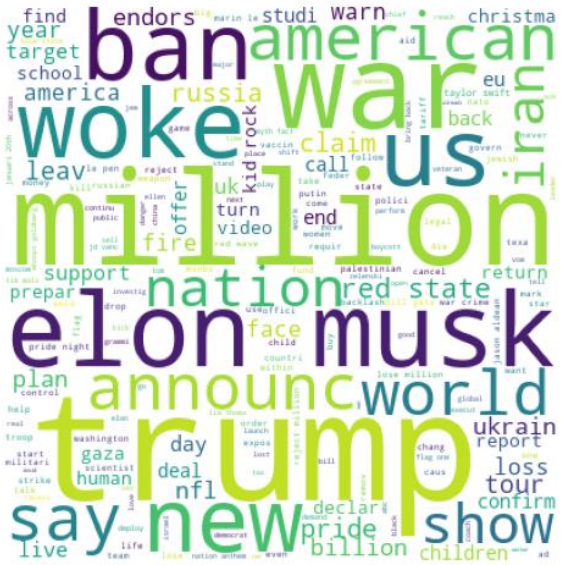
For iteration 0, an initial dataset containing 30,016 entries of fake and real news from 2016–2017 was used. However, its authenticity could not be guaranteed, as the dataset's origin was unclear, and it was both outdated and potentially unreliable. To address these concerns, a new dataset was created 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)

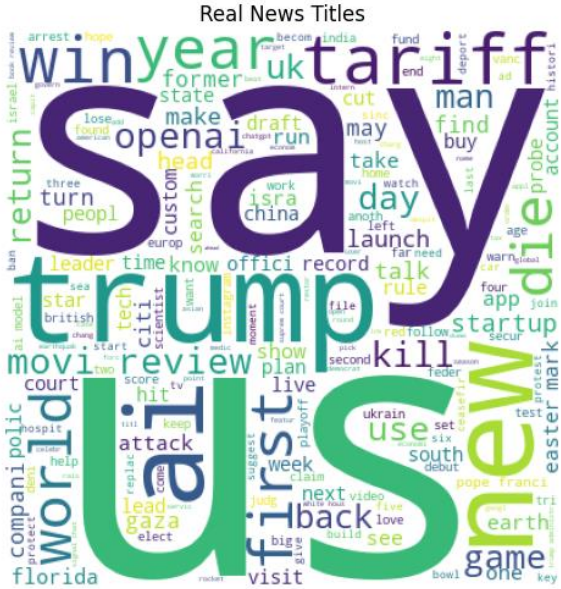
- **Fake News Characteristics:** Use of emotionally charged language and exaggerated claims.
- **Real News Characteristics:** Neutral tone and reliance on factual descriptions.
- **Visualization Tools:** Word clouds and frequency charts highlighted distinct term patterns in each class.

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

The limited size and scope of the dataset could inadvertently introduce biases, potentially skewing model predictions. For instance, underrepresentation of certain topics, sources, languages, or regions might lead to systematic errors in classification. These biases can result in some articles being unfairly labelled, which undermines the system's credibility and inclusivity. To mitigate these biases, a twofold approach is essential. First, expanding the dataset to include diverse topics, sources, languages, and regions is crucial for ensuring balanced representation. This expansion will make the system more robust and capable of handling 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. These audits should involve stakeholders from diverse backgrounds to ensure a comprehensive evaluation.

## Transparency and Interpretability

Transparent models foster trust and empower users to make informed decisions. While DeFactify currently achieves transparency primarily through its documentation and accompanying notebook, the long-term goal is to develop a system that leverages Explainable AI. Such a system would provide not only the likelihood of an article being true or fake but also the reasoning behind its predictions, which is critical for building user confidence. Future enhancements aim to include advanced visualization tools to illustrate the influence of specific words or phrases, as well as an intuitive user interface that enables stakeholders to easily verify and contextualize results. Currently, the project’s transparency relies on its comprehensive documentation and research, offering users insights into the methodology and decision-making process. By making the system's workings understandable, DeFactify lays the groundwork for future advancements in explainability.

# Societal Impact

DeFactify's societal implications are significant, particularly concerning freedom of speech and the potential for unintended consequences. Mislabelling legitimate news as fake can damage the credibility of journalists or media outlets, potentially silencing important voices. This risk emphasizes the importance of carefully evaluating the system's outputs and avoiding over-reliance on automation. Over time, misclassifications could erode public trust in media and technology, especially if the system becomes widely adopted without proper checks and balances. To mitigate these risks, it is essential 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 key to reducing dependence on automated systems. Furthermore, ensuring that the system operates as an assistive tool – rather than as the final arbiter of truth – can help safeguard freedom of speech and foster a healthier information ecosystem.

# Automation and Human Oversight

While DeFactify leverages advanced AI to detect patterns in fake news, it is not a substitute for human expertise. The interaction between automation and human oversight is critical for achieving reliable outcomes. DeFactify's strength lies in its ability to process vast amounts of data quickly and identify potential fake news patterns at scale. However, complex or borderline cases often require a nuanced understanding of context, language, and intent that only humans can provide. Integrating human review into the system ensures that final decisions consider broader contexts and ethical considerations. For example, human fact-checkers can verify borderline cases flagged by the model, providing an additional layer of scrutiny. This collaboration between machine predictions and human expertise minimizes the risk of misclassification, reinforces accountability, and enhances the reliability of the system's outputs. Ultimately, this balanced approach ensures that DeFactify remains a valuable tool while respecting the complexity of real-world information assessment.

# Final Recommendations

## For Stakeholders

- **Adopt With Caution:** DeFactify is an effective supplementary tool but requires validation for critical decisions.
- **Expand Dataset:** Broaden the dataset to reduce biases and enhance model generalizability.
- **Foster Transparency:** Develop tools to explain model reasoning and build stakeholder trust.
- **Collaborate Broadly:** Engage with fact-checking organizations and policymakers to refine use cases and ethical deployment.

## Use Case Suitability

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

- **Short Term:** Effective for identifying blatant fake news patterns.
- **Long Term:** Potential for integration into broader misinformation prevention systems, contingent on dataset expansion and ethical safeguards.

## Consequences of Use

- **Positive:** Enhances the ability to flag and address fake news efficiently.
- **Negative:** Risk of reputational damage if legitimate news is misclassified. Appropriate human oversight is essential.

# Conclusion

DeFactify represents a promising step toward mitigating the spread of misinformation. While its current performance is strong, continuous improvement and careful deployment are critical to its success. By integrating ethical considerations, expanding datasets and fostering collaboration, DeFactify can serve as a pivotal tool in the fight against fake news.