DeFactify

RESEARCH DOCUMENT MAY 13, 2025

Ivet Kalcheva Version 2.0 (Final)

Contents

Introduction 4
Background 4
The Importance 4
Statement of the Problem 5
Research Objectives 5
Main Research Question 6
Methodology 7
Document Analysis 7
Literature Review 8
Data Collection and Processing 8
Results 9
Restoring Trust in the Age of Misinformation 11
Trust and AI: Opportunities and Challenges 11
Information Overload and the Need for Filtration 12
Implications for AI and Misinformation 12
The Emotional Component of Trust 12
Rebuilding the Islands: A Long-Term Vision 13
The Dual Role of Generative AI 13
Building Trust Through Long-Term Efforts 13
Conclusion 14
References 15

Summary

This research document investigates the effectiveness of machine learning techniques in detecting fake news by analysing linguistic and structural patterns, as well as source metadata.

Misinformation, often spread through social media and unverified news platforms, poses significant risks to public perception and democracy. Fake news is frequently used as a weapon to emotionally manipulate the public, influencing individuals and ideologies in both positive and negative ways and spreading false information and propaganda.

With the rise of AI, many journalistic websites risk using AI without adhering to established standards, which allows misinformation to proliferate as language models closely mimic internet speech. It is essential to provide people with user-friendly tools for verifying information; everyone can benefit from them no matter their age and demographic.

The goal of this study is to develop a model that can classify news as either real or fake. The end product will be a web-based tool designed to help users verify the authenticity of news content in real time.

Introduction

Background

Misinformation, manipulation, lies and deceit have existed throughout history. In the 13th century BC, *Ramesses the Great* spread falsehoods and propaganda to portray *the Battle of Kadesh* as a stunning victory for the Egyptians. He depicted himself defeating his enemies in battle through carvings on the walls of nearly all his temples. However, the treaty between the Egyptians and the Hittites reveals that the battle ended in a stalemate.

According to *Bounegru*, *Gray*, *Venturini* and *Mauri*, a lie becomes "fake news" when it is picked up by numerous blogs, shared by hundreds of websites, cross-posted on thousands of social media accounts and read by hundreds of thousands of people.

Today, the term "fake news" is often overused, particularly by some individuals during election seasons in various countries. A rough search on Google News for "fake news" returns about 5 million results. By 2018, the phrase had already been used around two million times on Twitter. In the future, "fake news" might be viewed as a relic of the tumultuous year of 2017 (if we're fortunate). However, the struggle against misinformation will not disappear. Companies and governments are beginning to take concrete actions, the effects of which will be felt for some time.

Both Google and Facebook have announced plans to hire many people to review content, enforce their terms of service and keep fake and illegal material off their platforms. Compared to previous years, many websites have been flagged or removed, indicating significant improvement in the right direction.

The Importance

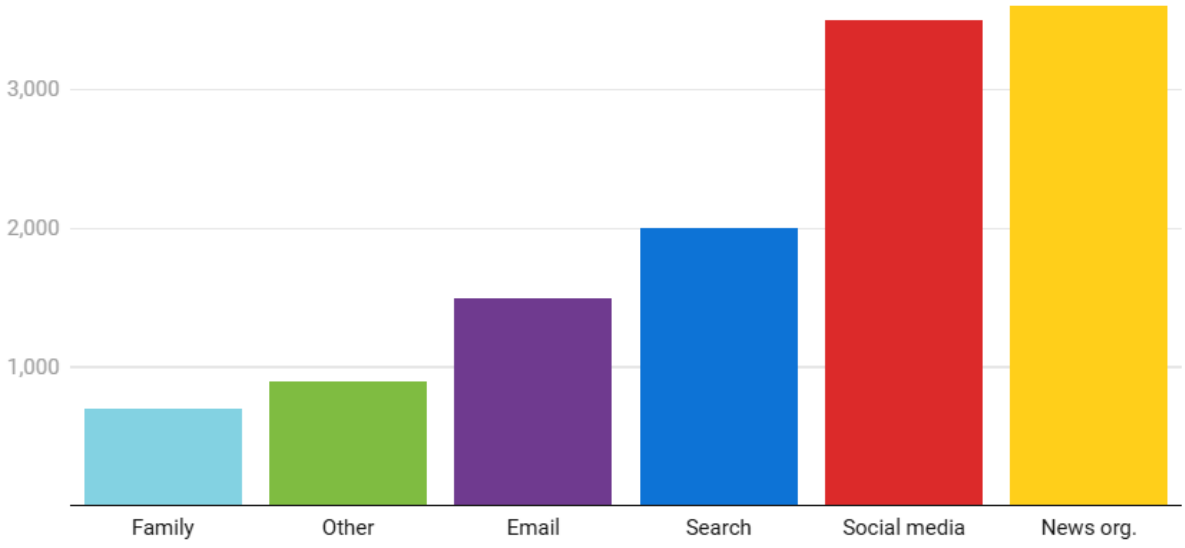
You might be wondering about the importance of tools for detecting fake news. The importance cannot be overstated. For instance, during one of the elections in the USA, there were false rumours circulating about a pizzeria involved in illegal activities. Despite efforts to remove related posts, the pizzeria ended up receiving an unexpected visit from a man, armed with a rifle, who believed the conspiracy and decided to take matters into his own hands. This situation highlights one of the worst possible outcomes associated with fake news.

In today's world, it is becoming increasingly difficult to discern what is real and what is not. Automated systems for detecting fake news can be essential for both individuals and organizations in maintaining the integrity of information. With AI-driven solutions, users can quickly assess the reliability of news articles, which helps mitigate the impact of clickbait, disinformation campaigns, and emotionally manipulative content.

Statement of the Problem

Before starting the project, I did not really use any news-checking applications. I tended to believe most of them on social media without any questions, which made me susceptible to misinformation. I think many of us share this problem, regardless of our age.

Figure 1: Where people get online news in the US, 2017



Source: Source: Pew Research Center, "How Americans Encounter, Recall, and Act Upon Digital News," February 9, 2017. • Get the data • Created with Datawrapper

Many of the tools available are often complicated, outdated or limited in their scope. Additionally, human bias and inconsistency can further undermine their reliability, often categorizing them as either right- or left-wing. This research indicates that machine learning models, trained on diverse datasets with annotated labels, can outperform both manual and rule-based systems in terms of efficiency and accuracy.

Research Objectives

This research project is designed to thoroughly evaluate and comprehend the process of categorising news articles into two distinct classifications: "Fake" and "Truthful." The study will delve into the linguistic features and structural elements that set apart fake news from genuine reporting. Additionally, it will examine the impact of source credibility, accompanying metadata and the distinctions of writing style on the perceived trustworthiness of news articles. Through this analysis, the goal is to uncover the underlying patterns that influence public perception and the reliability of information in the media landscape.

Main Research Question

What linguistic and structural features distinguish fake news from real news and how can machine learning be used to detect and explain these differences effectively?

Subquestions

- What are the most predictive linguistic cues for fake news?
- How does source credibility affect article classification?
- What machine learning models are most suitable for this task?
- What are the ethical challenges involved in building such a system?
- How can explanation mechanisms improve user trust in AI predictions?

By answering these questions, this research will build a strong case for adopting AI-driven news detection solutions to effectively combat with the misinformation.

Methodology

Document Analysis

Manual analysis of articles labelled as fake and real to identify common patterns in:

- Use of highly emotional and sensationalist language.
- Content originating from unfamiliar or non-credible news sources.
- Lack of clear authorship with many articles skipping the journalist's name.
- Websites that attempt to imitate legitimate domains or use misleading web addresses.
- Low-quality or unprofessional images, including AI-generated visuals with unnatural or symbolic elements.
- Language designed to provoke hostility or outrage toward individuals, groups or organizations.
- Frequent use of exaggerated or attention-grabbing terms such as "*million*", "*war*" and "*woke*", and references to public figures like "*Trump*" or "*Elon Musk*".
- Emphasis on subjective opinions, viral social media narratives and emotionally charged words such as "*support*", "*satire*" and "*true*", often prioritizing controversy over verified facts.
- Overuse of punctuation, including excessive exclamation marks and capital letters (e.g., "*UNBELIEVABLE!!*").
- Presence of spelling, grammatical and punctuation errors.
- Headlines structured as clickbait, designed to attract clicks rather than accurately represent the article content.

Literature Review

An in-depth review of current and past research on fake news detection was undertaken, with a particular focus on:

- **Natural Language Processing (NLP) Techniques**, which refers to how computers can understand human writing (like English) in a way that lets them analyse the meaning of text. Such as identify keywords or phrases that are common in fake vs real news, analysing sentence structure or writing style and recognizing exaggeration, sensationalism or emotionally charged content.
- **Supervised Machine Learning Approaches** is a type of artificial intelligence where we teach a computer to make decisions based on examples. I label the collected data as either real or fake, then feed it into the model for training. This allows the model to make predictions about new, unlabelled news articles. For this process, I have chosen to use the **Naive Bayes** algorithm (in Iteration 1) and **K-NN** (in Iteration 0). **Decision Trees** can be also used to integrate author-based credibility as a feature – articles with clearly listed authors in most cases end up being more reputable than the anonymous ones. By Iteration 3, I decided to implement a stacking classifier as the final approach. This model uses a Support Vector Machine (SVM) as the meta-classifier with Naive Bayes and Random Forest as the base classifiers. This ensemble method combines the strengths of multiple algorithms to achieve improved accuracy and robustness in predictions.
- **Publicly Available Datasets from Previous Studies** – refer to collections of real and fake news articles that have already been compiled and labelled by other researchers. As I discussed with my teachers, I initially struggled to find the right dataset. When I finally found one, it was older – about 6 to 7 years old – but it helped guide me in the right direction in terms of what I need for the ideal dataset. However, since it was old and outdated, I went and collected news articles myself, which were relevant and more up-to-date for more accurate results.

Data Collection and Processing

Scraping and cleaning articles from various online sources involves several steps. First, we extract titles, content and metadata, such as publication date, author, category and source. Next, we eliminate duplicates and empty entries. Additionally, we convert categories into numerical data and transform titles and content into a matrix that indicates the frequency of each word mentioned in the document, focusing on both true and fake news.

Results

To summarize the previously spread-out answers to the questions:

- Articles labelled as real often included clear sourcing and journalist names, presented balanced, fact-based narratives and avoided emotionally manipulative language and speculative claims.
- People are more likely to be convinced that an article is truthful if a journalist is named as the author. This indicates that someone has invested time, research and hard work into delivering the information, compared to an anonymous article.
- Regarding the machine learning model, this is so far from what my understandings are:

	Algorithm	Key Insight	Limitations	Accuracy
Iteration 0	K-Nearest Neighbours (K-NN)	Used with TF-IDF and performed overall very well.	Struggles with satire; requires high memory for larger datasets.	83.51%
Iteration 1	Naive Bayes	Faster, better at distinguishing subtle cues in language.	Struggles with excessive punctuation.	90.43%
Iteration 2	Combined Model	Combined strengths of Naive Bayes (text features) and Decision Tree (listed journalist).	Slight dip in accuracy from Naive Bayes alone; majority vote loss.	89.89%
Iteration 3	Random Forest	Handles noisy features and identifies feature importance effectively.	May not generalize well for smaller datasets.	92.55%
Iteration 3	SVM	Excellent for high-dimensional data, captures non-linear relationships.	Requires careful tuning of hyperparameters.	93.62%
Iteration 3	Stacking Classifier	Combines strengths of multiple models; SVM used as meta-model.	Computationally intensive; dependent on base model diversity.	94.15%

	Algorithm	Key Insight	Limitations	Accuracy
Iteration 3	AdaBoost with Naive Bayes	Reasonable performance for boosting weaker learners like Naive Bayes.	Limited improvement over standalone Naive Bayes.	88.30%
Iteration 3	AdaBoost with SVM	Attempted to boost SVM but resulted in overfitting and instability.	Poor performance due to boosting already strong learner.	43.09%

- Developing a fake news detection system raises several ethical challenges. First, biases in training data can lead to unfair classifications, especially if the content reflects specific political or cultural leanings. There is also the risk of violating freedom of speech by mislabeling controversial or satirical content as fake. Transparency is essential – users need clear explanations for why content is flagged. Mistakes like false positives can damage reputations, while false negatives may let misinformation spread. Additionally, scraping data must respect privacy and copyright laws. As fake news creators evolve, models must be updated to stay effective. Lastly, while AI tools are powerful, they should support – not replace – human judgment and media literacy.
- Explanation mechanisms improve user trust in AI predictions by making the decision-making process transparent and understandable. When users can see why a news article was classified as fake or real – such as highlighting specific words, sentence structures or metadata that influenced the result – they are more likely to trust the system's output. These explanations help users feel in control and informed, reducing the "black box" effect often associated with AI. Clear, human-readable insights also allow users to critically evaluate the reasoning, making the AI feel more like a supportive tool rather than an unquestionable authority.

Restoring Trust in the Age of Misinformation

One of the greatest challenges in combating fake news is not just identifying false information but also rebuilding public trust in the sources of news and information. Misinformation damages more than just individual understanding; it erodes societal cohesion. Once trust is broken – through repeated exposure to lies, conspiracy theories or emotionally manipulative content – it becomes incredibly difficult to repair. People often retreat into echo chambers or abandon faith in media altogether, which can lead to further polarization and disengagement from critical discourse.

Trust, as the Dutch saying goes, "komt te voet en gaat te paard" – it arrives on foot but leaves on horseback. Restoring it within the news media landscape requires a multi-faceted approach. Technological tools, transparent systems and an emphasis on digital literacy are essential. Artificial intelligence, while not a panacea, can play a pivotal role when it is developed and implemented with transparency, explainability and user empowerment as core principles.

Trust and AI: Opportunities and Challenges

Questions such as "*How can we trust AI models?*" and "*Do they risk adding another layer of distrust?*" underscore the complexity of integrating AI into the fight against misinformation. The paradox is clear: while AI can enhance detection of falsehoods, its opaque nature can also breed scepticism. To avoid this, AI systems must be designed with a focus on transparency.

When users understand how a fake news detection model operates – what features it examines and why a specific article is flagged – they are more likely to accept its judgments. Features such as clear visual indicators, confidence scores and side-by-side comparisons of flagged content and reliable sources can demystify the process. This not only improves the perceived reliability of the AI system but also helps bridge the gap between scepticism and trust.

However, as James Gleick writes in *The Information: A History, a Theory, a Flood* ". In an ocean of information, we will begin to search for islands we trust". These "islands" symbolise spaces where consistency, credibility and clarity thrive amidst chaos. AI should not aim to dominate or dictate these islands but rather assist users in navigating toward them, providing tools to discern the truth without imposing it.

Information Overload and the Need for Filtration

One of Gleick's key observations is that in an era of abundant information, the problem shifts from access to filtration. We no longer need help finding information – we need help distinguishing the meaningful from the irrelevant, the true from the false. This has profound implications for the fight against misinformation. If misinformation serves as a pollutant in the ocean of information, then trust is the filtration system.

However, this filtration system is not purely cognitive; it is also emotional and social. People gravitate toward sources that align with their values and provide a sense of safety or validation. In this way, misinformation can thrive by exploiting emotional biases, creating a perception of trustworthiness even when the content is demonstrably false.

Implications for AI and Misinformation

Gleick's perspective pushes us to ask deeper questions about the role of AI in this landscape. Can AI act as a reliable filtration system? Can it help users identify islands of trust without becoming another authoritarian gatekeeper? The answers lie in how AI systems are designed and deployed:

- **Enhancing Transparency:** AI must serve as a clear and open filter, showing users how and why certain pieces of information are classified as credible or false. This aligns with Gleick's emphasis on clarity as a cornerstone of trust.
- **Empowering User Agency:** Instead of dictating what to believe, AI tools should enable users to navigate the information ocean more effectively. By presenting evidence, comparing sources and offering explanations, AI can help users make informed decisions.
- **Building New Islands:** AI can also aid in the creation of new "islands" of trust by identifying and amplifying high-quality, consistent and well-sourced journalism. It can assist in spotlighting underrepresented voices that adhere to rigorous standards of credibility.

The Emotional Component of Trust

Gleick's work also highlights the emotional dimensions of information. Misinformation often succeeds not because it is factual but because it is compelling – it tells a story that resonates emotionally. Restoring trust, therefore, requires not only intellectual rigor but also emotional intelligence. AI systems must be attuned to these dynamics, ensuring that their interactions with users feel respectful and empathetic rather than cold or condescending.

Rebuilding the Islands: A Long-Term Vision

James Gleick's ideology suggests that rebuilding trust in the age of misinformation is not merely about stopping falsehoods but about creating spaces – both literal and metaphorical – where truth can thrive. These spaces are not static; they require ongoing maintenance and adaptation. Collaboration among technologists, educators, journalists and policymakers will be essential in cultivating these islands of trust. Furthermore, the public must be encouraged to engage critically and actively in this process, becoming co-creators of these trusted spaces.

The Dual Role of Generative AI

Generative AI further complicates the landscape of misinformation. Its ability to produce convincing fake articles, images and even videos present a significant threat. At the same time, these tools can be used to detect and counter misinformation by identifying patterns of deception or generating truthful rebuttals to misleading claims. Balancing this dual role requires responsible innovation and stringent ethical oversight.

Building Trust Through Long-Term Efforts

Restoring trust is not a short-term attempt. It requires sustained efforts, including:

- **Digital Literacy Campaigns:** Educating the public about recognizing fake news and understanding AI tools.
- **Transparency in AI Models:** Openly sharing the inner workings of detection systems and maintaining explainability.
- **Collaboration with Journalistic Standards:** Aligning AI tools with the practices and values of credible journalism.

By adopting these approaches, we can work toward not only reducing the spread of misinformation but also repairing the trust fractured by it. The long-term vision is clear: to rebuild the confidence people once had in reliable journalism, aided by AI-supported tools, continuous education and ethical practices. Only then can we move toward an informed society that values the truth as a shared foundation.

Conclusion

This research document highlights the growing importance of reliable and accessible fake news detection tools in a world overwhelmed by digital information. By analysing linguistic patterns, source credibility and structural cues, machine learning models demonstrate strong potential in classifying news as real or fake with considerable accuracy.

However, building such systems goes beyond technical performance – it involves ethical responsibility, transparency and user trust. Clear explanation mechanisms and regularly updated models are crucial for making AI-driven tools both effective and trustworthy. Ultimately, the goal is not to replace human judgment, but to empower users with smart, intuitive tools that help them navigate the complex media landscape more critically and confidently.

References

- **BBC** (2018, January 18). *Why we fall for fake news*. BBC News. https://www.bbc.com/news/blogs-trending-42724320
- **Brookings Institution** (2017, December 18). *How to combat fake news and disinformation*. https://www.brookings.edu/articles/how-to-combat-fake-news-and-disinformation/
- **NOAA** (n.d.). *Fake news worksheet*. NOAA AOML Library. https://www.aoml.noaa.gov/general/lib/lib1/nhclib/Fake-News-WorksheetProQuest.pdf
- **PolitiFact** (n.d.). *Get Bad News* [Interactive Game]. https://www.getbadnews.com/en
- **Wikipedia** (2024, April 20). *Fake news*. Wikipedia. https://en.wikipedia.org/wiki/Fake_news
- **Gleick, J.** (2011). *The Information: A History, a Theory, a Flood. New York.*