Automated Evaluation of Factual Accuracy in Online Discourse

Jacquelyn Garcia jag053@ucsd.edu

Amelia Lei aclei@ucsd.edu

Ali Arsanjani arsanjani@google.com

Sam Lau lau@ucsd.edu

Abstract

In a new age where rapid advancements in artificial intelligence and an overwhelming amount of online information are prevalent, distinguishing fact from fiction has become increasingly difficult. With many fact-checking and content verification programs being scaled back or eliminated by major tech companies, there exists a critical gap in public tools designed to assess credibility. This project addresses that gap by developing a machine learning-driven framework for automated factuality evaluation. Leveraging a suite of machine learning and large language models, our team aims to build an ensemble system that evaluates on multiple factuality factors to generate a credibility score for any user-provided article or text. By empowering users to independently verify the reliability of online content, this project seeks to contribute toward a more informed digital information ecosystem.

Code: https://github.com/JacquelynGarcia/DSC180A-Q1Project

1 Introduction

Over the last decade, the decline of fact-based web has become one of the central crises of the digital era. The infrastructure once designed to protect truth online through fact-checking partnerships and content moderation systems has undergone a systematic dismantling. Major platforms like Meta and Google, once known for promoting accurate information online, have stepped back from their fact checking roles, cutting funding and support for groups that helped maintain public trust. This pullback reflects a broader political and economic shift, as growing mistrust, conflicting regulations, and the labeling of content moderation as censorship have blurred the line between truth and falsehood.

At the same time, the rapid rise of new technologies has fundamentally increased the scale of misinformation. Artificial intelligence and generative models have made it easier than ever to create realistic but deceptive content. Meanwhile, algorithmic recommendation systems continue to amplify polarizing material for engagement. However, these same technologies also enable us to create automated systems capable of analyzing text, detecting linguistic bias, and evaluating credibility through machine learning at a scale much larger than human fact-checkers.

In this environment, traditional approaches to misinformation mitigation such as manual fact checking and centralized labeling systems are becoming unsustainable. The collapse of these systems has exposed the need for a new approach, one that leverages a suite of machine learning models. Our team aims to develop a framework that evaluates articles and online text across multiple factuality factors, frequency heuristic, echo chamber, sensationalism, and credibility to generate an interpretable credibility score. By transforming factuality assessment into an automated and explainable process, this project seeks to contribute towards a more transparent and informed digital landscape.

1.1 Related Work

Early research on misinformation detection focused on identifying and verifying check-worthy claims in political text. ClaimBuster (Hassan et al. 2017) was one of the first end-to-end systems to detect factual claims in real time. This system combined linguistic and contextual features to prioritize statements for human fact-checkers. The LIAR dataset (Wang 2017) and it's extension LIAR-PLUS (Alhindi, Petridis and Muresan 2018) introduced a large, labeled dataset of political statements that were annotated with truth ratings, speaker metadata, and contextual variables. These datasets formed the benchmark for supervised truth classification models.

To improve robustness, later studies adopted ensemble techniques to capture more complex feature interactions. Ahmed, Traore and Saad (2017) proposed one of the first comprehensive ensemble-based frameworks for fake news detection. Their model combined n-gram and term-frequency features with multiple machine learning classifiers to automatically detect deceptive content in both news articles and online reviews.

With the rise of large language models, misinformation detection has entered a retrieval-augmented generation (RAG) era. Yue et al. (2024) introduced a new framework that leverages dense retrieval models to collect evidence passages relevant to a claim and then generates both supporting and contrasting arguments conditioned on those retrieved documents. By training the model under a contrastive learning objective, the system learns to distinguish between credible and non-credible evidence.

1.2 Data

This project makes use of the LIAR-PLUS dataset (Alhindi, Petridis and Muresan 2018), an extended version of the original LIAR corpus (Wang 2017), which contains over 12,800 short political statements fact-checked by *PolitiFact*. Each entry is then annotated with one of six categorical truth ratings, ranging from pants-on-fire to true, and is accompanied by rich metadata. This includes the speaker's name, occupation, political party, state, subject, and justification text. The metadata fields allow factuality to be modeled through the use of linguistics and socio-political standpoints.

The dataset comes distributed in three predefined splits including training, validation, and testing. Our team ensured that each subset underwent preprocessing to standardize formatting and prepare for feature extraction. We standardized column names, removed redundant index columns, and all text fields were stripped of whitespace and normalized. Suffixes, such as ".json" from IDs, were removed and string variables were converted to lowercase during tokenization. The cleaned dataset then became an optimal foundation for generating stronger factuality features.

The frequency heuristic model captured linguistic and style repetitions within statements. Using TF-IDF and count-vector representations, features like average word frequency, buzzword score, and repetition score were computed. These features were then combined into a feature matrix used to estimate a probability-based frequency heuristic score which reflected the density of highly frequent or exaggerated language.

The echo chamber model examined the concentration of political party representation across topic categories. By grouping subjects by party and calculating dominance ratios, topics heavily associated with a single party received higher echo chamber values. The scores were then categorized on a four level scale.

The sensationalism model assessed emotional intensity and rhetorical hyperbole. A weighted mapping of six truth labels were extracted using TextBlob. For each statement, the model calculated the number of exclamation marks, all-caps tokens, and sensational keywords. Polarity and subjectivity metrics were then added as numerical inputs. This allowed us to then categorize statements into low, medium, or high sensationalism.

Finally, the credibility model used linguistic cues and speaker attributes. Truth labels were mapped into three ordinal classes. Each statement's subjectivity was computed with TextBlob and paired with two party encodeding and expertise level gathered from the professional roles metadata. After we vectorized, standardized, and concatenated within a ColumnTransformer, we produced a credibility score for each statement.

Together, these data transformations turn the LIAR-PLUS corpus (Alhindi, Petridis and Muresan 2018) into a multi-dimensional feature space representing linguistic, social, and contextual aspects of factuality. Our goal is to use this preprocessing pipeline as the foundation for an ensemble system designed to evaluate and interpret factual accuracy across online discourse.

References

Ahmed, Hadeer, Issa Traore, and Sherif Saad. 2017. "Detecting opinion spams and fake news using text classification." *Security and Privacy* 1, p. e9. [Link]

Alhindi, Tariq, Savvas Petridis, and Smaranda Muresan. 2018. "Where is Your Evidence: Improving Fact-checking by Justification Modeling." In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium Association for Computational Linguistics. [Link]

Hassan, Naeemul, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak et al. 2017. "Claimbuster: The first-ever end-to-end fact-checking system." *Proceedings of the VLDB Endowment* 10 (12): 1945–1948

Wang, William Yang. 2017. "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection." *arXiv preprint arXiv:1705.00648*

Yue, Zhenrui, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. "Retrieval augmented fact verification by synthesizing contrastive arguments." *arXiv preprint arXiv:2406.09815*