Detecting Misinformation in Fitness Content

CS4991 Capstone Report, 2025

Evan Yuan
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
hdn6fv@virginia.edu

ABSTRACT
The spread of misinformation within social media fitness culture has contributed to harmful consequences, including the glorification of unrealistic body standards and the normalization of unsafe practices, such as the use of performance-enhancing drugs (PEDs). To address this growing concern, I propose creating a machine-learning detection tool capable of identifying misleading fitness content. Using Natural Language Processing (NLP) techniques, my model will be able to analyze captions, comments, and keywords to classify posts as either "misleading" or "non-misleading". My tool will use supervised learning algorithms, starting with a logistic regression baseline and progressing to more advanced models such as neural networks and random forests. I anticipate my tool assisting gym-goers in navigating fitness advice more safely, while also inspiring future integration into social media platforms to combat misinformation and promote a healthier fitness culture.

1. INTRODUCTION

From 2000 to 2019, the use of non-medical anabolic steroids has nearly tripled, according to the Australian National Drug Strategy Household Survey of 2019, a trend that is growing and alarming within the modern fitness culture (Sport Integrity Australia, 2019). While steroid use for performance enhancement is not a new phenomenon, the motivations driving their use in the modern era are deeply concerning. Steroids were commonly used by professional bodybuilders and athletes in the 1980s and 1990s to give them an advantage in their respective sport. Today, more than ever, young individuals are turning to these illegal substances in pursuit of extreme physical ideals, striving to achieve the lean, muscular physiques glorified across social media. This shift reflects not only changing societal standards but also the dangerous influence of misinformation in shaping behaviors around fitness and body image.

Social media has fueled this trend, with fitness influencers often promoting unrealistic body standards while being secretive of their own usages of PEDs. Many influencers fail to disclose their reliance on PEDs, creating a false narrative that their physiques are attainable naturally. The lack of transparency also overlooks the severe health risks associated with PED use, including liver damage, heart issues, and high blood pressure (Cleveland Clinic, 2023). While some influencers provide genuine and evidence-based advice, many promote misleading advice, including superficial workouts and questionable supplement endorsements. This misinformation is particularly harmful to younger audiences, who are more impressionable and vulnerable to adopting unsafe practices in pursuit of idealized body and fitness standards.

2. RELATED WORKS

The combat against misinformation on social media has been widely studied, with several studies providing inspiration for my fitness misinformation detection tool. The work by Taherdoost (2023), demonstrates how machine learning tools are used to improve content recommendations, detect harmful content, and analyze user behavior. His research showcases the role of Natural Language Processing (NLP) in detecting sentiment, classifying posts, and identifying misinformation patterns. This study provides a foundation for applying similar techniques to detect misinformation in fitness-related posts.

Research by Di Sotto and Viviani (2022) provides an in-depth analysis of health misinformation detection within social media platforms using machine learning and data science techniques. Their study explores the key challenges associated with distinguishing genuine and informative content from misleading health information and examines NLP-based methodologies for identifying deceptive claims. Given the similarities between fitness misinformation and health-related misinformation, their findings provide practical strategies for building a classification system that flags misleading fitness claims.

Khanam, et al. (2021) focus on machine learning applications for fake news detection. Their study examines various supervised learning algorithms, including Random Forest, Support Vector Machines (SVM), and Naïve Bayes, to classify news articles as either fake or authentic. The researchers highlight the effectiveness of feature extraction techniques, such as TF-IDF vectorization, in improving classification accuracy. This work provides solid approaches for my project, as it demonstrates the effectiveness of these models in identifying deceptive content.

3. PROPOSAL DESIGN

The proposed tool aims to identify and flag misleading fitness content on social media. This tool could be developed by individuals familiar with machine learning and anyone willing to take the time to research previously-created detection tools such as email spam detectors. The majority of the project could be developed using Python in Google Colab, using existing machine learning libraries for model training and evaluation. The machine learning models would have to be trained and evaluated against fitness content posts on social media to assess their effectiveness in detecting misinformation.

3.1 Design Layout

The proposed detection tool will consist of three main components:

- Data Collection — Gathering fitness related posts from social media platforms, including captions, comments and hashtags.
- Feature Extraction — Identifying relevant features or trends, such as keywords and sentiment.
- Classification — Implementing machine learning models to classify posts as misleading or non-misleading based on extracted features.

Together each of these components will operate as a sequence, with data progressing through data collection, feature extraction and classification before producing a final prediction.

3.2 Data Collection

The first step in developing the misinformation detection tool is to collect and preprocess a dataset of fitness related social media posts from platforms such as Instagram, TikTok and YouTube. The extracted data should include captions, comments and hashtags related to fitness, bodybuilding, supplements and PEDs. Existing web scraping

tools and APIs can be used to help gather data. The dataset will be labeled manually to help classify posts as misleading or non-misleading with a focus on content promoting unrealistic body standard, unsafe practices or undisclosed PED use. To help prepare the data for analysis, text data should be cleaned and the TF-IDF technique should be used to convert text into numerical representation suitable for machine learning models. This dataset will serve as the training and testing foundation for the developed machine learning models.

3.3 Machine Learning

The proposed tool will implement supervised learning models for classification, starting with a baseline model and progressing to more complex models.

Baseline Model:

Logistic Regression — A simple model that establishes a baseline performance.

Advanced Models:

Random Forest — An ensemble method that builds multiple decision trees and combines their outputs to improve classification accuracy.

Neural Networks — Deep learning models used to capture complex relationships between words and overall context.

3.4 Evaluation

The tool's performance will be evaluated based on evaluation metrics, including accuracy, precision, recall, and F1-score.

- Accuracy — Measures overall correctness in classifying posts.
- Precision — Evaluates how many flagged posts are actually misleading.
- Recall — Measures how effectively the model identifies all misleading content.
- F1-Score — Provides a balanced measure between precision and recall.

The model with the best performance across these metrics will be used for the detection tool.

4. ANTICIPATED RESULTS

Once development is complete, the detection tool should be capable of running in Google Colab or a local Python environment. Further work will be needed to scale the detection tool into a web-based application for easier usage. The initial version of this tool will focus on text-based content, so future enhancements could expand to include the analysis of images and videos to detect misleading visual content. If the tool hits its expected outcome, it will be able to help gym-goers make more informed decisions while also laying the foundation for future integration into social media platforms to combat fitness-related misinformation.

While the anticipated results and potential of the tool are promising, potential challenges may arise during implementation. For instance, the dynamic nature of social media content may require continuous updates to the training dataset and model to maintain accuracy. Additionally, the tool's effectiveness may vary across different social media platforms due to variations in content formats. Nevertheless, the proposed solution would be a significant step forward in addressing fitness misinformation, with the potential to create a safer and more transparent digital environment for fitness enthusiasts.

5. CONCLUSION

My proposal explores and builds on existing machine learning techniques to combat the growing influence of social media on fitness culture and the potential harm caused by misinformation. While developing this detection tool is not a definitive solution to this problem, it would serve as a useful tool to promote more transparency and awareness over controversial health and fitness topics for gym-goers and social media users. By

analyzing captions, comments and hashtags, the tool will classify posts as “misleading” or “non-misleading,” empowering gym-goers to make safer and more informed decisions about their fitness habits. Beyond individual users, this tool could have the potential to assist social media platforms by offering a scalable and efficient way to filter deceptive and harmful fitness information.

6. FUTURE WORK

To develop this tool, software engineers would be needed to build and test the machine learning models for the detection tool. While this project provides a strong foundation for detecting fitness misinformation within text-based content, several areas of improvement and expansion remain. Future expansions should incorporate image and video analysis using computer vision techniques to detect misleading fitness content and deceiving transformations. By integrating multimodal machine learning, the tool could analyze both textual and visual misinformation, making it more comprehensive and applicable across diverse content formats.

Additionally, deploying this tool as a browser extension or a web application would make the application more accessible to anyone. Collaborations with social media platforms could lead to direct integration with content moderation systems, providing a scalable solution for misinformation detection at the platform level. With these enhancements, this project can continue evolving into a more robust, impactful tool that contributes to a safer and more informed fitness community.

REFERENCES

Cleveland Clinic Medical. (2023, February 7). *Anabolic Steroids*. Cleveland Clinic. https://my.clevelandclinic.org/health/treatments/5521-anabolic-steroids

Di Sotto, S., & Viviani, M. (2022, February 15). Health misinformation detection in the social web: An overview and a data science approach. MDPI. https://www.mdpi.com/1660-4601/19/4/2173

Khanam, Z., Alwasel, B., Sirafi, H., & Rashid, M. (2021, March 1). IOPscience. IOP Conference Series: Materials Science and Engineering. https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040/meta

Taherdoost, H. (2023). Enhancing social media platforms with machine learning algorithms and neural networks. https://doi.org/10.3390/a16060271

SportIntegrity.gov. (2023, May 26). The worrying trend of steroid use in young adults. Steroid Use Australia—Worrying Trend for Young Australians. https://www.sportintegrity.gov.au/news/integrity-blog/2023-05/worrying-trend-of-steroid-use-young-adults