# IDENTIFICATION OF FAKE AND AUTHENTIC SOCIAL MEDIA IDENTITIES

*Programming project 1-COSC2408*

**Warnakulasuriya Thamel - s3863814**

**Matthew Orwin – s3782972**

**Xinwei Li - s3816123**

**Yuxuan Bao – s3893703**

**Jiayi Shen – s3758915**

Contents

[IDENTIFICATION OF FAKE AND AUTHENTIC SOCIAL MEDIA IDENTITIES 1](#_Toc149505680)

[INTRODUCTION 3](#_Toc149505681)

[METHODOLOGY 3](#_Toc149505682)

[Data Curation 4](#_Toc149505683)

[Data pre-processing 4](#_Toc149505684)

[Data Splitting 4](#_Toc149505685)

[Models 4](#_Toc149505686)

[Random Forest: 4](#_Toc149505687)

[Decision Trees: 4](#_Toc149505688)

[Neural Networks: 5](#_Toc149505689)

[Gradient Boost Classifier: 5](#_Toc149505690)

[Multinomial NB: 5](#_Toc149505691)

[XGBoost: 5](#_Toc149505692)

[AdaBoost: 5](#_Toc149505693)

[Metrics Used 5](#_Toc149505694)

[DATA COLLECTION 6](#_Toc149505695)

[MEETINGS WITH THE CLIENT 7](#_Toc149505696)

[RESULTS 8](#_Toc149505697)

[DISCUSSION 9](#_Toc149505698)

[RECOMMENDATIONS 10](#_Toc149505699)

[CONCLUSION 11](#_Toc149505700)

# INTRODUCTION

In this vast and interconnected realm of social media, the issue of fake identities has become a thorn in the side of authentic online interactions. Picture this: a digital landscape where impostors lurk in the shadows, adopting false personas for a variety of unethical purposes. These fraudulent accounts can create a lot of havoc by spreading misinformation, inciting controversies, and even influencing public opinion and political events. On a more personal level, imagine the discomforting experience of stumbling upon a virtual doppelgänger using your identity to create a deceptive profile. It is not just a matter of distorting information; it also disrupts of trust in the online space. Social media platforms, to maintain the integrity of their ecosystems, constantly grapple with the challenge of identifying and purging fake accounts. Enter the scene: an algorithm designed to act as a vigilant guardian, capable of distinguishing between genuine users and digital impostors. This report delves into the development of such an algorithm aimed at unravelling the mysteries of fake identities and contributing to the restoration of authenticity in our increasingly digitalized social landscape.

The significance of our algorithm lies in its mission—to spot and deal with fake accounts on social media. The big goal here is to create an online space that is genuine and trustworthy. Imagine it as a digital detective, working behind the scenes to keep things real. The objective of our algorithm is straightforward: to sniff out those sneaky fake accounts. We want to distinguish between real users and impostors, making the online world a safer and more honest place. It's like having a virtual bouncer at the door, keeping out the party crashers and ensuring that the digital landscape is as authentic as it can be.

# METHODOLOGY

We took a deep dive into different algorithms which would eventually figure out which algorithm performed the best in identifying fake social media entities, testing a myriad of models to figure out which one is the superstar in helping us spot those pesky fake accounts. To rigorously assess their performance, we were able to test their performance against several open datasets and on some test data that was provided to us directly from the client. Our models included Random Forest, Decision Trees, Neural Networks, Gradient Boost Classifier, Multinomial NB, XGBoost, and AdaBoost. These models represent a spectrum of computational approaches, and our goal was to discern which one stands out as the most adept in distinguishing authentic users from fraudulent entities. The gravity of this undertaking is underscored by the need to fortify the digital landscape against deceptive accounts, promoting trustworthiness in online interactions.

## Data Curation

### Data pre-processing

The initial phase of our algorithm development involved meticulous data pre-processing to ensure the model's robustness and efficiency. We started off by cleaning and normalizing the dataset to establish consistency and to subsequently eliminate any potential noise. This step was very important in enhancing the quality of our input data, subsequently optimizing the model's ability to discern patterns.

Furthermore, recognizing the significance of relevant features, we conducted feature extraction to identify attributes indicative of fake identities. This process involved a comprehensive analysis of the dataset, seeking patterns and characteristics that distinguish fraudulent from genuine identities. The success of our algorithm hinges significantly on the discriminative power of these features.

### Data Splitting

Once the data pre-processing was complete, we strategically divided the dataset into three subsets: training, validation, and test sets. The rationale behind this division was to facilitate effective model training, validation, and assessment of generalization performance.

To ensure a representative distribution of fake and real identities in each set, we employed stratified sampling. This approach ensured that each subset maintained a proportionate representation of both classes, preventing biases that might adversely impact model performance. The stratified split aimed to mirror the distribution of fake and real identities in the overall dataset, enabling the model to learn and generalize across diverse scenarios.

In addition to the standard train-test split, we leveraged cross-validation as a robust technique to assess the model's generalization capabilities. This involved iteratively training and validating the model on different subsets of the data, providing a more comprehensive evaluation of its performance across various scenarios.

## Models

### Random Forest:

After some research we found that it was ideal for identifying fake online identities, Random Forest's ensemble nature aggregates predictions from multiple decision trees. This handles complex datasets, ensuring robustness against overfitting. Its knack for discerning important features aligns with our goal of extracting relevant indicators of fraudulent identities.

### Decision Trees:

With transparency in decision-making, Decision Trees are valuable for interpreting the logic behind identifying fake identities. Their hierarchical structure captures intricate relationships, providing clear insights into features driving classification, especially adept at handling categorical data prevalent in online identity datasets.

### Neural Networks:

Neural Networks, with their capacity for learning intricate patterns and hierarchical representations, excel in complex classification tasks. For identifying fake online identities, their ability to capture non-linear relationships enhances accuracy and generalization.

### Gradient Boost Classifier:

Boosting weak learners, Gradient Boosting improves model accuracy by iteratively optimizing for misclassified instances. This adaptability suits the intricacies of identifying fake identities, gradually refining predictive capabilities.

### Multinomial NB:

Tailored for discrete features, Multinomial Naive Bayes is apt for scenarios where specific features distinctly characterize fake identities. Its simplicity and efficiency in handling high-dimensional data make it pragmatic for our classification problem.

### XGBoost:

Combining gradient boosting strengths with efficient computation, XGBoost is a go-to choose for complex tasks. Its regularization techniques mitigate overfitting, ensuring a balanced model adaptable to various data types in online identity datasets.

### AdaBoost:

AdaBoost's iterative refinement, focusing on misclassified instances, complements our goal of enhancing accuracy in identifying fake identities. Its ability to adapt and improve over successive iterations makes it a valuable addition to our ensemble of classifiers.

All in all, we were able to try out a diverse range of models which allowed us to strategically gain the unique strengths of each algorithm and subsequentially identify which model performed the best for the task at hand.

## Metrics Used

We used F1 score as our evaluation metrics for the models. The selection of the F1 score as our primary evaluation metric was deliberate and aligned with the requirements of identifying fake identities. The F1 score, a mean of precision and recall, emerged as the ideal metric for our task. Given the binary nature of our classification, it was imperative to strike a balance between minimizing false positives (precision) and capturing all actual instances of fake identities (recall).

False positives could potentially have severe consequences in misidentifying genuine users as fake, while false negatives might lead to overlooking actual instances of fraudulent activity. The F1 score, by harmonizing precision and recall, offered a comprehensive measure of our model's ability to discern both classes accurately.

In the context of identifying fake identities online, where the consequences of misclassification are substantial, optimizing for F1 score became paramount. This metric allowed us to fine-tune our model to achieve a balance that minimized both types of errors.

# DATA COLLECTION

After our initial meeting with the client, they expressed a desire for us to independently source data that could be used for testing and training our models. This data was to be gathered from various public social media platforms and open-source data resources. Following a meeting dedicated to data collection strategies, we decided to employ two distinct approaches.

The first approach involved utilizing free APIs, such as the Twitter API and the MTEA GRAPHIC API, which are provided by popular social media platforms for obtaining user data. Upon exploring the Twitter API, we discovered that it had limitations in terms of the user information it could access. It could only provide publicly available data like user usernames and registration dates. To access tweets or more extensive data, it required paid access or the use of Python web scraping techniques. Following a second meeting, it became evident that this method was not feasible, and a decision was made to explore alternative avenues for data acquisition.

Subsequently, after discussing this situation with the client, they provided us with an email containing links to the main pages of numerous users, primarily on Instagram and Facebook. Our team then began investigating the MTEA GRAPHIC API, which functioned like a database query and returned user data in JSON format. However, it was found that while this API could provide some valuable data, it still did not meet our requirements for model training and testing.

In the subsequent meeting with the client, we collectively decided to utilize publicly available databases from sources like Kaggle and the UCI Machine Learning Repository for data collection and cleansing. This ensured that the data would be suitable for training and testing our models.

During the data collection process, several challenges arose that required extensive discussions and resolutions. Initially, after receiving the client's requirements, we delved into the discussions regarding what type of data was needed, where to obtain this data, and how to clean it effectively.

First and foremost, understanding the nature of the required data was crucial. We carefully assessed the data necessary for model testing and training based on the client's demands. However, the most intricate and complex aspect of the entire task revolved around how to collect this data.

We began by identifying the platforms from which we needed to collect data, namely Twitter, Instagram, and Facebook. Learning to use the APIs of these platforms presented significant challenges. A notable difficulty was that many of these APIs required developer authentication to gain access to usable data. After completing this step, we encountered further hurdles, especially with the Twitter API. It proved to be challenging to use, and the data it provided was highly limited. Access to more vital data required an upgrade, which meant paying for access.

As a result, the team shifted its focus to collecting the required data from Instagram and Facebook, specifically using the MTEA GRAPHIC API. We discovered that the free version of the MTEA GRAPHIC API could provide some usable data. However, this data often appeared corrupted after being saved, and the information obtained was still insufficient to support our model's testing and training needs.

Subsequently, the team turned to publicly available data resources like Kaggle and the UCI Machine Learning Repository to acquire the necessary data. Following this, a significant portion of our effort involved inspecting and cleaning the obtained data to ensure its suitability for our team's project.

# MEETINGS WITH THE CLIENT

In the first sprint, the team and the client primarily discussed project goals and requirements. The focus was on how to kickstart the project and what the ultimate objectives were. The client requested the team to undertake extensive research and preparatory work before diving into model development. During the second sprint, in meetings with the client, the team shifted their attention to model testing. They conducted tests with various models to determine the most suitable one and communicated the results with the client. Once the model to be used was finalized, the meetings between the client and the team shifted towards data acquisition and processing. The client provided the team with a data list and expected the team to obtain a dataset based on this list. In the third sprint, meetings between the team and the client focused on the completion of research and testing of the model, as well as the testing of the user interface, to ensure that the initial requirements were met, and the product could be delivered. The final meetings with the client were cantered around showcasing the team's work and achieving product delivery.

Regarding modifications to both the task at hand and the implementation of it, there were a few things that changed as a result of both our meetings and outside advice. After many weeks or research and model testing as described above, we set our sites on starting to implement both a machine learning algorithm to detect fraud based on variables taken from the online accounts themselves, as well as using natural language processing techniques to be able to assist and reinforce the model’s prediction. It was this second objective that was in the original scope of the project which changed as a mutual decision between our team and the client, based on the timeline of the project and the knowledge base that we as a team had. Another thing that was changed from our perspective was the addition of a user interface. Originally, it was our teams understanding that our deliverable was a machine learning algorithm that could detect fraudulent activity online, however, in discussions with the client, it was mentioned that a simple UI would be very nice if possible. Because of this addition, modifications to our project were made and we ended up implementing a simple UI using the model produced. These are just a few examples of things that were modified in our projects scope, we as a team were adaptable and able to work with our incredibly understand and patient client.

# RESULTS

Results Table:

A white box with black text

Description automatically generated

As we can see in the results table presented above, our random forest and gradient boost classifier models were the ones that performed the best of all the models that were experimented with. As each model’s basic premise and underlying mechanisms were discussed earlier in the methodology section of this report, there is no need to discuss the reasons why are results are the way they are. However, we can go into detail regarding whether these results were expected or not and why that is the case.

As mentioned prior, it makes sense that our random forest model was one of the top performing algorithms due to it being an ensemble machine learning algorithm which combines numerous classifiers, meaning that we would expect an aggregate of different techniques to produce the most accurate result. When we compare this to some of the other techniques used that are not ensemble ML algorithms, the higher accuracy makes sense.

Like random forest algorithms, the gradient boost algorithm is also an ensemble classifier but using weak predictors which are usually in the form of decision trees. Whilst random forest and gradient boost are very similar, we can tell by the almost identical result, the key difference is that gradient boosting algorithms aggregates their decision trees during the process of creating them, adding trees together as they are created, compared to random forest that adds all created trees at the end of the entire process. However, whilst gradient boosting is often more effective than random forest, it is also more prone to overfitting and requires a lot more hyperparameter tuning. Which is why, due to a lack of experience with parameter tuning, we decided to use our random forest classifier for our final model as it was more likely to be a more comprehensive model with less chance of it having been over fitted.

One unexpected result that we had was that the neural network we trained was one of the worst performing models we tested, which was unexpected due the assumption that neural networks usually are more complex and therefore were thought to be better at predicting. However, after researching reasons for getting results like we have, we believe that one of the main reasons that our neural network performed so much worse than the other, more simple models, is simply the lack of data. Neural networks, whilst incredibly powerful, usually need massive amounts of data to be very effective, and since all our models were only using a dataset of approximately 750 pieces of data, we do not think that this was enough to allow for it to reach its peak efficacy. Other reasons for the disappointing performance include the potential need for more hyperparameter tuning or the need for more layers in the neural network. A possible answer for the lack of data is to artificially generate data using the data we already had, however we decided against this as we could not guarantee that the artificial data would have been accurate and thus, had the potential to ruin the model's ability to accurately predict on unseen data.

# DISCUSSION

The results show that both stochastic forest and gradient enhanced classifier models perform very well in identifying false identities. This is in line with expectations, as integration methods are generally good at handling complex data sets and minimizing overfitting. The unexpected underperformance of the neural network raises questions about its suitability for the given task of the data set at hand. One possible explanation could be that the data set size is relatively small, as neural networks typically benefit from large amounts of data for optimal performance. Further research and experiments may be needed to fine-tune neural network parameters or explore alternative architectures.

Given neural networks' reputation for handling complex patterns, their lower performance was unexpected. This result highlights the importance of not only relying on the complexity of the model, but also considering the appropriateness of the model for a particular task and data set. Challenges encountered during data collection, especially with APIs, can affect the performance of some models. For example, the limitations of the Twitter API may have led us to decide to turn to alternative data sources. Understanding and addressing these challenges is critical to perfecting future iterations of the algorithm.

The information presented lacks a clear reference to existing research on false identification on social media. To enhance the discussion, it would be useful to compare the results of the algorithm with those of relevant studies. This may involve examining the proposed algorithms for similarities, differences, and potential advancements. Although no direct comparison with existing research is provided, the algorithm's results can be seen as a contribution to the ongoing discussion about false identity detection. Future research may incorporate insights gained from this study, particularly in improving algorithms and addressing challenges encountered during data collection.

In conclusion, while the algorithm shows robust performance on some models, the unexpected results, especially those of neural networks, highlight the complexity of solving the problem of false identity detection in online Spaces. Further exploration and potential collaboration with existing research could enhance the overall understanding of effective strategies to combat deception on social media platforms.

# RECOMMENDATIONS

The issue of identifying fake and authentic social media identities is paramount in today's digital landscape, where misinformation and online fraud are rampant. While significant strides have been made in the development of identity verification techniques, continued research and innovation are essential to keep pace with evolving tactics employed by malicious actors. This section provides recommendations for potential improvements and future research directions in the field of identifying fake and authentic social media identities.

First, Behavioral Analysis: Investigate behavioral patterns that distinguish fake from authentic identities. Identifying deviations in posting habits, language use, and interaction dynamics can be a valuable addition to current methods.

Second, User-Centered Approaches: Empower users with tools and features to actively manage their online identities. Research on user-centered identity management can lead to more effective identity verification processes and increased user awareness.

Third, Cross-Platform Identity Verification: Develop techniques for cross-platform identity verification. Coordinated fake identities often operate across multiple platforms; therefore, linking and tracking identities across these platforms is crucial.

Fourth, Collaboration and Data Sharing: Encourage collaboration between social media platforms, tech companies, and researchers to share data and insights. A collective effort is essential in combating the widespread issue of fake identities.

Fifth, User Education and Awareness: Develop educational initiatives to raise user awareness about the risks associated with fake identities and the importance of vigilance. Educated users are more likely to report and avoid engaging with fake accounts.

In conclusion, advancing the identification of fake and authentic social media identities requires a multidisciplinary approach, combining technological innovation with ethical considerations. The recommendations provided offer a roadmap for improving identity verification processes, thereby strengthening the online environment, and protecting users from malicious actors. Continued research, collaboration, and user empowerment are pivotal to achieving these objectives.

# CONCLUSION

Data Quality and Relevance: During the data management phase, we conducted a comprehensive evaluation of the quality and relevance of the specific data required for the project. This involved assessing data accuracy, completeness, and timeliness, ensuring that the information met the project's objectives. We recognized that data quality is paramount in ensuring the success of any data-driven task, as the integrity of our analyses and models depended on it.

Workflow Planning: Prior to commencing the project, we recognized the pivotal importance of meticulously planning the workflow. This step involved mapping out the entire data collection, processing, and analysis process, including setting clear objectives and milestones. A well-thought-out workflow streamlined our efforts, minimized potential bottlenecks, and ensured efficient progress throughout the project.

Exploration of Diverse Approaches: We also gained a profound understanding of the significance of exploring various data collection methods and analysis techniques. By experimenting with different modes of data acquisition and analysis, we were able to adapt and fine-tune our approach, allowing us to address evolving challenges effectively. This adaptability was crucial in achieving optimal results, given the dynamic nature of data-related tasks.

Client Collaboration and Expectations: Collaborating closely with the client and gaining a deep understanding of their needs and expectations was a cornerstone of our success. We ensured that we were aligned with the client's goals, which enabled us to tailor our efforts accordingly. Effective communication and a clear understanding of the client's objectives were instrumental in delivering a solution that met their specific requirements.

Skills Development and Team Collaboration: The project provided a valuable opportunity for our team to acquire an array of problem-solving skills and enhance our collaborative abilities. We encountered numerous challenges that demanded innovative solutions, and through effective teamwork and knowledge sharing, we were able to overcome these hurdles. This experience underscored the importance of continued skill development and team collaboration in tackling complex projects.

Practicality of Implemented Solutions: The ultimate measure of success was the practicality of the solutions we implemented. We ensured that the methods and tools used for data collection, analysis, and model training were not only effective but also feasible within the constraints of the project. This practical approach ensured that the solutions were sustainable and could be applied to similar projects in the future.

These key findings represent crucial insights gained throughout the project, underscoring the importance of planning, adaptability, collaboration, and practicality in successfully managing and utilizing data for our client's objectives.

accounts on social media. The overarching goal here is to foster a genuine and trustworthy online space. Picture it as a digital detective silently working behind the scenes to ensure the authenticity of the digital landscape. The objective of our algorithm is crystal clear: to sniff out those cunning counterfeit accounts. We aspire to differentiate between genuine users and impersonators, thereby rendering the online world safer and more truthful. It is akin to having a virtual security guard at the entrance, fully preventing unwelcome intruders and upholding authenticity in the digital realm possible.