What sets Verified Users apart?

Insights, Analysis and Prediction of Verified Users on Twitter

Indraneil Paul (IIIT Hyderabad), Abhinav Khattar (IIIT Delhi), Shaan Chopra (IIIT Delhi), Ponnurangam Kumaraguru (IIIT Delhi), Manish Gupta (Microsoft India)

Outline

A: PROBLEM AND MOTIVATION

- Perceived influence of verification
- Understanding what sets verified users apart

B: DATASET DESCRIPTION

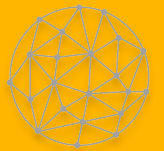
- Description of data collection
- Summary data statistics

C: METADATA/ACTIVITY ANALYSIS

- Study divergence of verified users from the rest for temporal activity and metadata signatures
- Deconstruct users into profiles

D: TOPIC ANALYSIS

- Study divergence between verified users and the rest for tweet topics
- Study divergence in topic diversity



Motivation

Reasons to care and intended outcomes

Ambiguity in Perception

Twitter, Facebook and Instagram have incorporated a **verification process** to authenticate handles they deem important enough to be worth impersonating.

However, despite repeated statements by Twitter about verification not being equivalent to **endorsement**, aspects of the process – the rarity of the status and its prominent visual signalling have led users to conflate the **authenticity** it is meant to convey with **credibility**.

Ambiguity in Perception

This perception of verification lending credence has led Twitter to receive a lot of flak in recent times, especially for harbouring **bias** against certain groups.

We try to demonstrate that the attainment of verified status by users can be explained away by less insidious factors based on user **activity trajectory**, **tweet contents**.

Visual Incentive

1. **Presence of authority and authenticity indicators:**
   Lends further credibility to the Tweets made by a user handle

2. **Presentation over relevance:**
   Psychological testing reveals that credibility evaluation of online content is influenced by its **presentation** rather than its relevance or apparent **credulity**

Attaining verified status might lead to a user's content being more frequently **liked** and **retweeted**.

Heuristic Models

The average user devotes only **three seconds** of attention per Tweet. This is symptomatic of users resorting to content evaluation heuristics.

One such relevant heuristic is the **Endorsement heuristic**, which is associated with credibility conferred to content by visual markers.

The presence of a marker such as a **verified badge** could hence, be the difference between a user reading a Tweet in a congested feed or completely ignoring it.

Heuristic Models

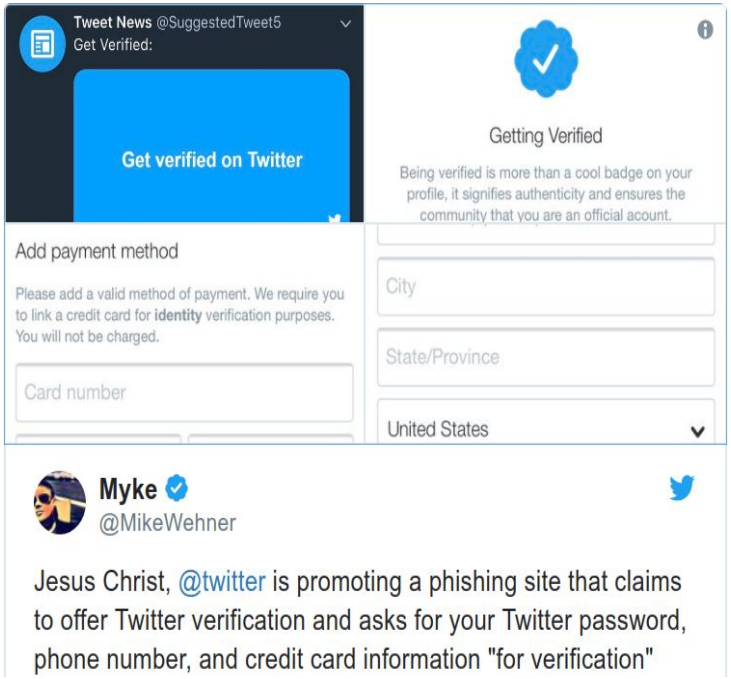
Another pertinent heuristic is the **Consistency heuristic**, which stems from endorsements by several authorities. This is important because a verified user on one social media platform is likelier to be verified on other platforms as well.

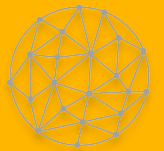
Hence, we posit that possessing a verified status can make a world of difference in the **outreach/influence** of a brand or individual in terms of the extent and quality.

Coveted Nature

Unsurprisingly, a verified status is highly **sought after** by preeminent entities and businesses, as evidenced by the prevalence of **get-verified-quick** schemes.

Instead of resorting to questionable schemes, accounts can follow our **insights** to increase their platform reach and improve their chances of verification.





Dataset

Collection sources, methods and summary

Collection Approach

We queried the Twitter REST API for the following:

1. The **@verified** handle on Twitter follows all accounts on the platform that are currently verified. We queried this handle on the **18th of July 2018** and extracted the user IDs.

2. We obtained the user objects for all verified users and subsetted for **English** speaking users obtaining 231,235 users.

3. Additionally, we leveraged Twitter's **Firehose API** – a near real-time stream of public tweets and accompanying author metadata.

Collection Approach

We used the Firehose to sample a set of 175,930 non-verified users by controlling for number of followers - a **conventional metric** of public interest.

This was done by ensuring that the number of followers of every non-verified user was **within 2%** of that of a unique verified user we had previously acquired.

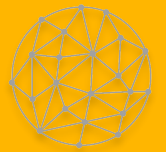
For each of the aforementioned user, data and metadata including **friends**, **tweet content** and **sentiment**, **activity time series**, and **profile reach trajectories** was gathered.

Collected Features

User Metadata	Temporal Features
Number of followers	Average number of followers last year
Number of friends	Average number of friends last year
Number of statuses	Average number of statuses last year
Number of public list memberships	Proportion of followers gained in last 3 months
Account age	Proportion of friends gained in last 3 months
	Proportion of statuses generated in last 3 months
	Proportion of followers gained in last 1 month
	Proportion of friends gained in last 1 month
	Proportion of statuses generated in last 1 month
	Average duration between statuses

Collected Features

Content Features	Number of POS tags[1] Frequency of POS tags[1] Average number of words per sentence Average number of words per tweet Character level entropy Proportion of long words[2] Positive sentiment score[3] Negative sentiment score[3] Neutral sentiment score[3] Compound sentiment score[3] Frequency of hashtags Frequency of retweets Frequency of mentions Frequency of external links posted
Miscellaneous Features	LIWC analytic summary score LIWC authentic summary score LIWC clout summary score LIWC tone summary score Botometer complete automation probability Botometer network score Botometer content score Botometer temporal score Tweet topic distribution[4]



Verified User Network

231,235

English language Twitter verified users

175,930

English languahe Twitter non-verified users

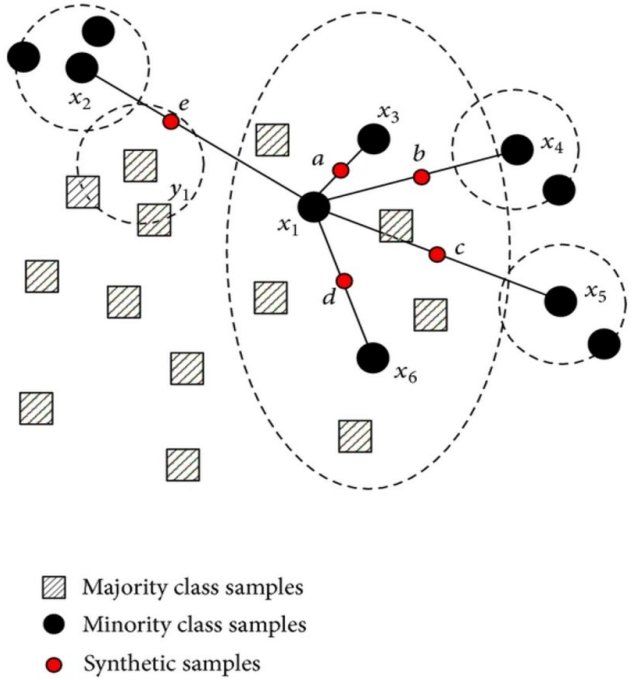
494 million

Tweets collected over a one year period

Class Imbalance

To prevent any effects of a **skewed class distribution** from affecting results, we applied two **class rebalancing** methods to rectify this.

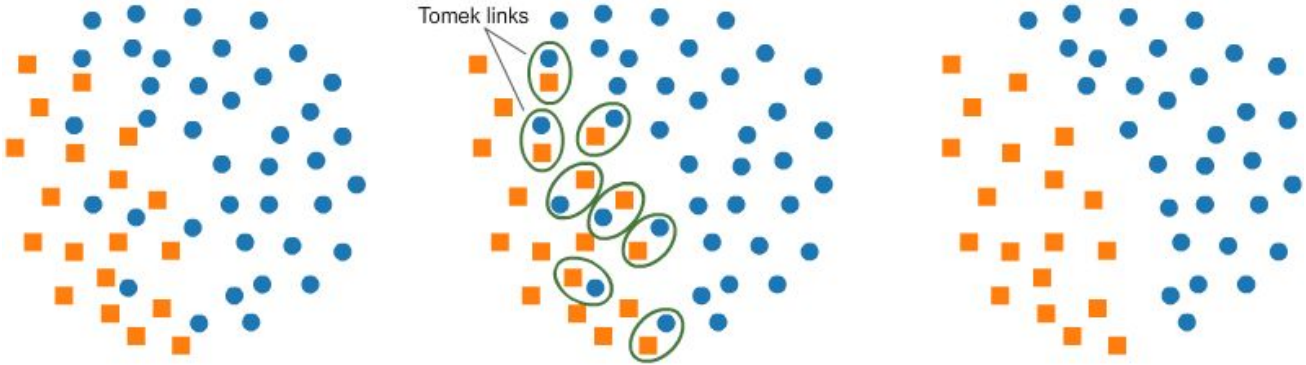
A minority oversampling technique called **ADASYN** was used. It creates **synthetic minority samples** based on **interpolation** between already existing samples.

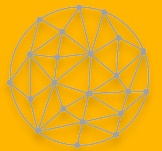


Class Imbalance

Additionally, we use a hybrid over and under sampling technique called **SMOTE Tomek** that also **eliminates samples** of the overrepresented class.

For a pair of **opposing class points** that are each other's **closest neighbours** (tomek link), the majority class point is eliminated.





Metadata and Activity Analysis

Investigating divergences in user features

User Data Classification

We commence our analysis by eliminating all features that could be deemed surplus to requirements. To this end, we employed an **all-relevant feature selection model** which classifies features into three categories: **confirmed**, **tentative** and **rejected**. We only retain features that the model is able to confirm over 100 iterations.

Using the rich set of features collected, we are able to attain a **near-perfect** classification accuracy of 99.1%. Our results suggest that a very competent classification of the Twitter user verification status is possible without resorting to complex deep-learning pipelines that sacrifice interpretability.

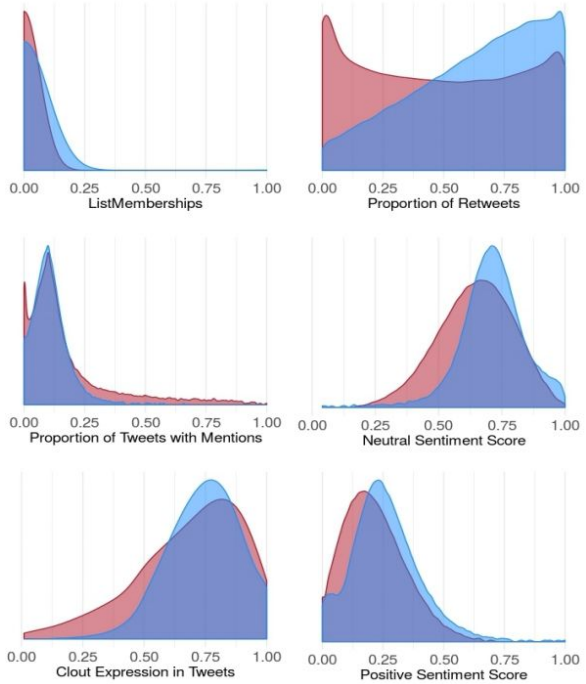
User Data Classification

Dataset	Classifier	Precision	Recall	F1-Score	Accuracy	ROC AUC Score
Original imbalanced data	Logistic Regression	0.86	0.86	0.86	0.859	0.854
	Support Vector Classifier	0.89	0.89	0.89	0.887	0.883
	Generalized Additive Model[1]	0.97	0.98	0.98	0.975	0.976
	3-Hidden layer NN (100,30,10) ReLU+Adam	0.98	0.98	0.98	0.983	0.977
	XGBoost Classifier	0.99	0.99	0.99	**0.989**	**0.990**
ADASYN class rebalancing	Logistic Regression	0.86	0.86	0.86	0.856	0.858
	Support Vector Classifier	0.89	0.89	0.89	0.891	0.891
	Generalized Additive Model[1]	0.97	0.97	0.97	0.974	0.973
	3-Hidden layer NN (100,30,10) ReLU+Adam	0.96	0.96	0.96	0.959	0.957
	XGBoost Classifier	0.99	0.99	0.99	**0.991**	**0.991**
SMOTETomek class rebalancing	Logistic Regression	0.86	0.86	0.86	0.860	0.856
	Support Vector Classifier	0.90	0.90	0.90	0.903	0.901
	Generalized Additive Model[1]	0.98	0.97	0.98	0.974	0.974
	3-Hidden layer NN (100,30,10) ReLU+Adam	0.97	0.97	0.97	0.966	0.968
	XGBoost Classifier	0.99	0.99	0.99	**0.990**	**0.991**

Feature Importance

To compare the usefulness of various categories of features, we trained **gradient boosting classifier**, our most competitive model, using each category of features alone.

Evaluated on randomized train-test splits of our dataset, **user metadata** and **content features** were both able to consistently surpass 0.88 AUC. Also, **temporal features** alone are able to consistently attain an AUC of over 0.79.

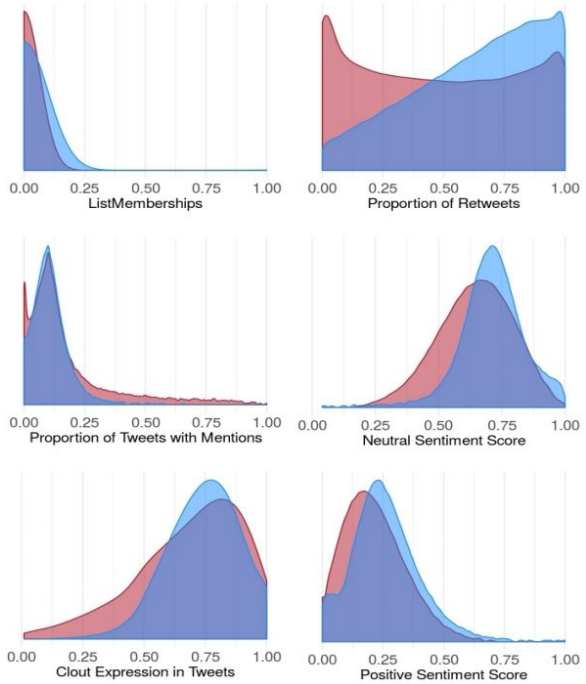


Feature Importance

The individual **feature importances** were determined using the **Gini impurity** reduction metric output by the gradient boosting model.

To rank the most important features reliably, the model was trained 100 times with varying combinations of hyperparameters.

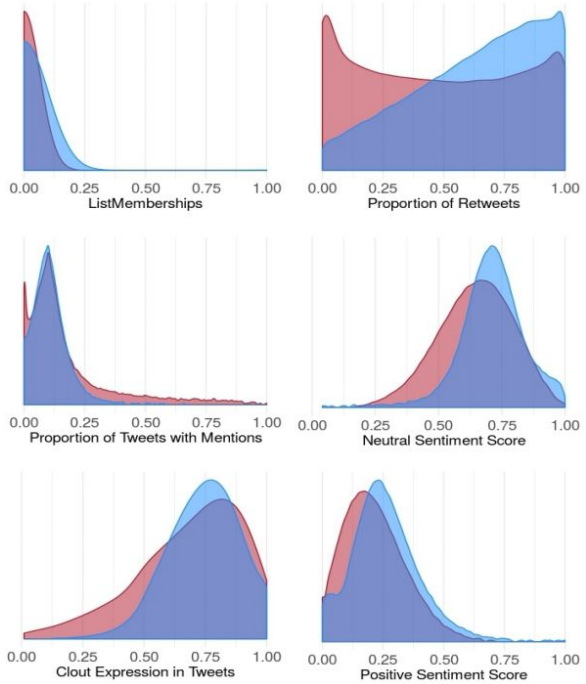
The most reliable discriminative features are shown.



Feature Importance

Some features are intuitively separable, making an informed prediction possible. The **top 6 features** are sufficient to attain 0.9 AUC on their own right.

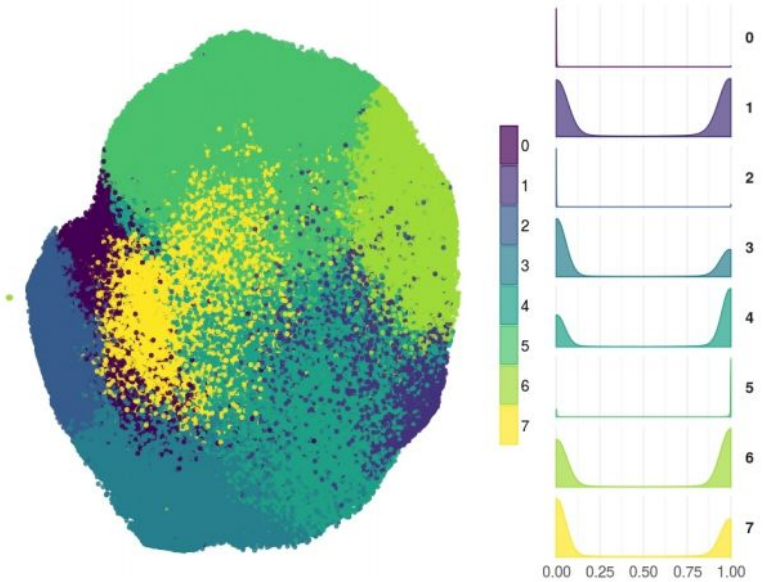
For instance, the very highest **public list membership** counts and prevalences positive sentiment in Tweets are populated exclusively by verified users while the very lowest propensities for authoritative speech as indicated by **LIWC Clout summary** scores are exclusively shown by non-verified users.



Profile Clustering

In order to characterize accounts with a **higher resolution**, we attempt to cluster them. We apply **K-Means++** on the normalized user vectors selecting the 30 most discriminative features indicated by the **XGBoost** model, eventually settling on 8 different clusters by tuning the perplexity metric.

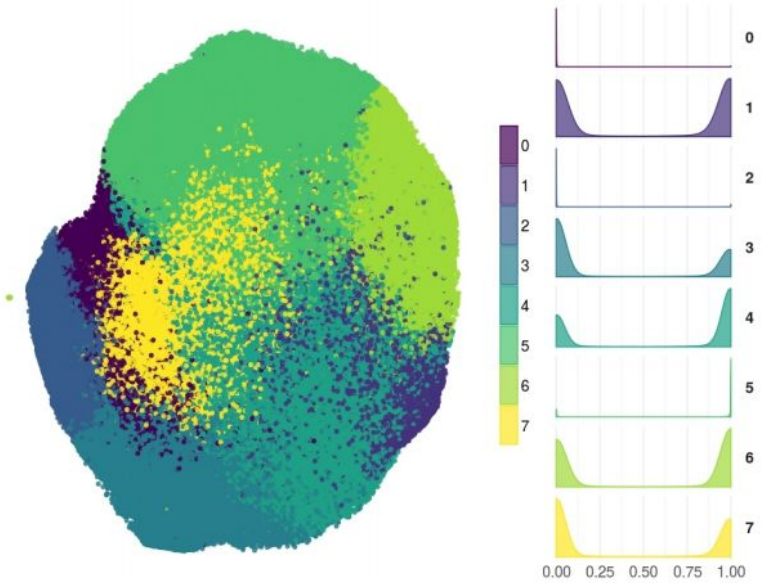
In the interest of intuitive visualization, **two dimensional embeddings** obtained via **t-SNE** are shown alongside.



Strongly Non-Verified

Cluster C0 can largely be characterized as the Twitter layman with a high proportion of experiential tweets. They have **short tweets**, high incidence of **verb usage** and score very high in the **LIWC Authenticity** summary.

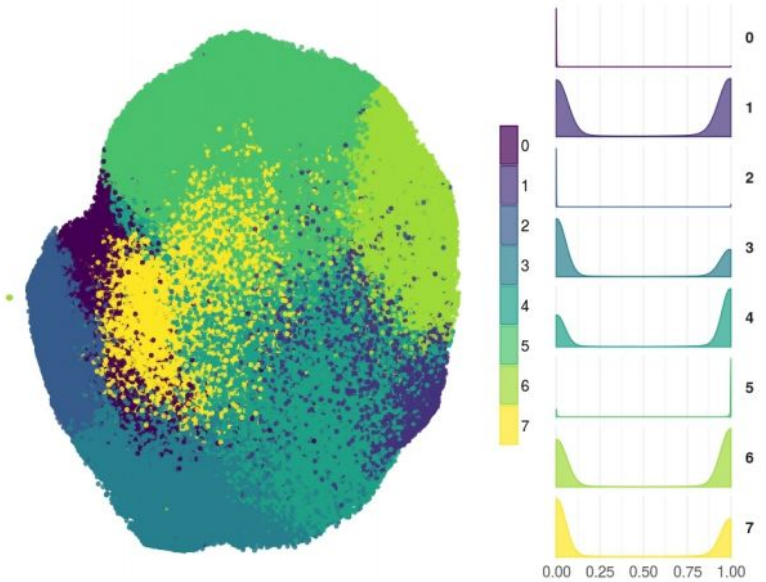
Cluster C2 can be characterized as an amalgamation of accounts exhibiting **bot-like behavior**. Members of this cluster scored highly on the **network and content automation** scores in our feature set. Extensive usage of **hashtags** and **outlinks** are observed.



Strongly Verified

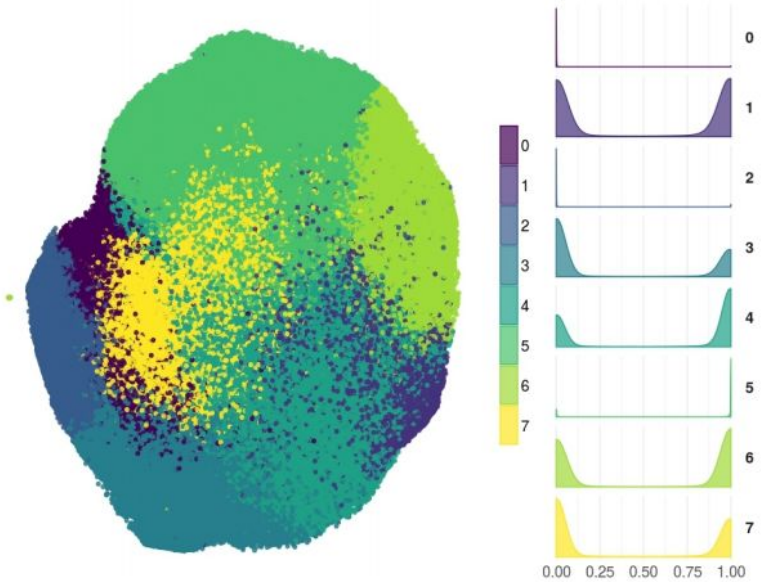
Cluster C4 having a tendency to post **longer tweets** and **retweet** more frequently than author content, while members of **Cluster C6** almost **exclusively retweet** on the platform.

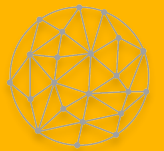
Cluster C5 is nearly entirely comprised of verified users and includes **elite Twitter users** that comprise the core of verified users on the platform. These users have by far the **highest list memberships** on average.



Mixed Clusters

Clusters C1, **C3** and **C7** are comprised of a mix of verified and non-verified users. Members of cluster **C1** are **ascendant** both in terms of reach and activity levels as evidenced by the proportion of their followers gained and statuses authored recently. Many users in **C1** have **obtained verification** in the data collection period. Members of **C3** and **C7** who are either **stagnant** or **declining** in their reach and activity levels and show very **low engagement** with the rest of the platform in terms of **retweets** and **mentions**.





Tweet Topic Analysis

Scrutinizing divergent Tweet topic choice and diversity

Topic Classification

To glean into Tweet topics we ran the **Gibbs Sampling** based LDA over 1000 iterations of sampling.

The number of topics was optimally fine tuned to 100 after trying out various values from 30 to 300 using **perplexity** values.

Instead of topic modelling on a **per-Tweet basis** and aggregating per user we apply the **author-topic model** collating all of a user's Tweets and topic modelling in one go. This is done to work around the fact that most Tweets are **too short** to meaningfully infer topics.

We use the default **document-topic densities** as well as **term-topic densities** as suggested in prior topic modelling studies.

Topic Classification

Our classification models demonstrate that it is eminently possible to **infer** the verification status of a user purely using the **distribution across topics** they tweet about, with a high accuracy.

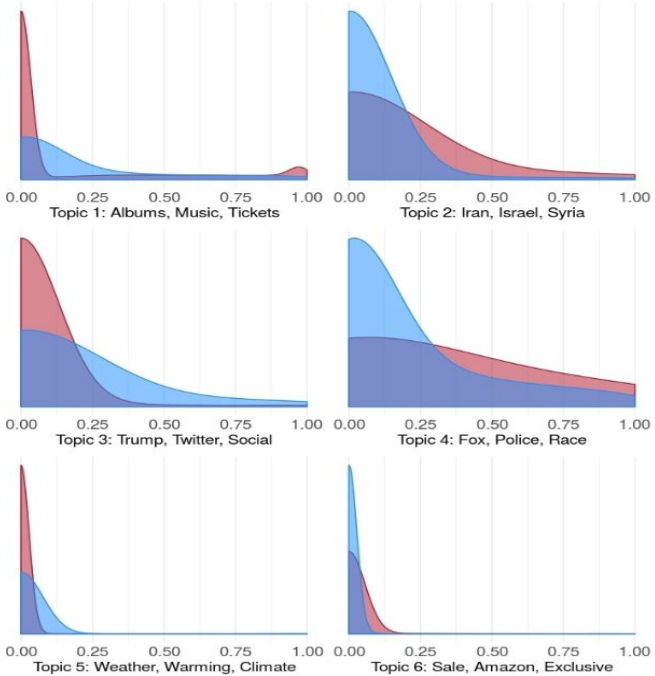
The most competitive classifier attained a classification **accuracy** of 88.2 %.

Classifier	Precision	Recall	F1-Score	Accuracy	ROC AUC Score
Generalized Additive Model (GAM)[1]	0.83	0.83	0.83	0.832	0.831
3-Hidden layer NN (100,30,10) ReLU+Adam	0.88	0.88	0.88	**0.882**	**0.880**
XGBoost Classifier	0.82	0.82	0.82	0.824	0.823

Topic Importance

In the interest of interpretability, we evaluate the **predictive power** of each topic with respect to verification status. We obtain individual topic importances using the **ANOVA F-Scores** output by GAM.

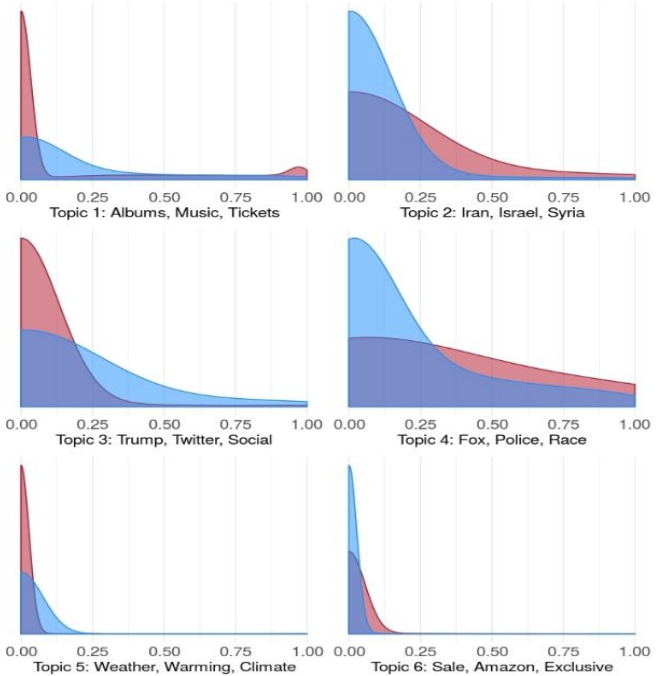
The procedure is run on **50 random train-test splits** of the dataset and the topics with the **lowest F-Scores** noted. Most discriminative topics with their **top 3 keywords** were noted.



Topic Importance

Though there is some overlap between topics, there are **clear patterns** to be observed on some topics using which an **informed prediction** can be made.

For instance the users who tweet most frequently about consequential topics like **climate change** and **national politics** are all verified while controversial topics like **middle-east geopolitics** and mundane topics like **online sales** are something verified users devote limited attention to.

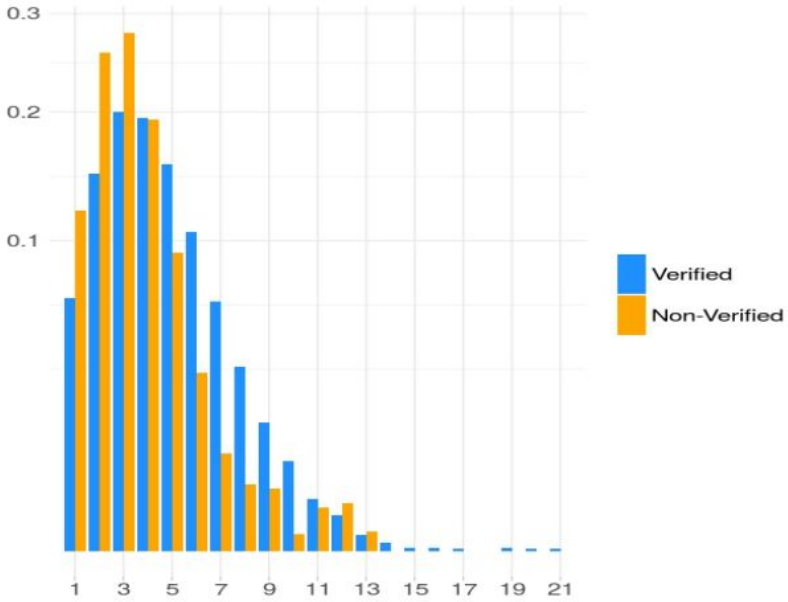


Topical Span

We next inquire about the **diversity** of Tweet topics.

In order to obtain an **optimal mix** of the number of topics per user in an **unsupervised** manner, we leveraged the use of an Hierarchical Dirichlet Process.

Inference is done using an **Online Variational Bayes** estimation using the previously stated hyperparameters.

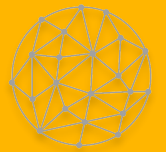


Topical Span

A trend is observed with non-verified users clearly being over-represented in the **lower reaches** of the distribution (1–4 topics), while a comparatively substantial portion of verified users are situated in the **middle** of the distribution (5–10 topics).

Also noteworthy is the fact that the very **upper echelons** of topical variety in tweets are occupied exclusively by verified users.

Shown are the **two most topically diverse** handles with 13 and 21 topics respectively.



Wrapping Up

Summary of contributions and possible future applications

Key Contributions

Full Featured Dataset

Released a fully featured dataset of 407k+ users, containing 79+ million edges and 494+ million time stamped Tweets.

Successful Classification

We are the first study to successfully attempt at discerning as well as classifying verification worthy users on Twitter.

We obtain a near perfect classifier in the process.

Actionable Findings

We unravel the aspects of a profile's activity and presence that have the greatest bearing on a user's verification status.

Future Applications

1. **Superior verification heuristic**
   Aforementioned deviations likely constitute a unique fingerprint for verified users which can be leveraged gauge the strength of a user's case for such status

2. **Actionable insights to improve online presence**
   Obtained insights can be used to significantly enhance the quality and reach of one's online presence before resorting to prohibitively priced social media management solutions

3. **Realistic synthetic influential profile generation**

Research Acknowledgements

IIIT Hyderabad

IIIT Delhi

Microsoft India

Thanks!

Any questions?

Find me at **ineil77.github.io**

Contact me at **indraneil.paul@research.iiit.ac.in**

For details refer to **paper preprint**