Detection of AI-Generated Social Bots on Twitter

Jesus Gonzalez, Shaunelle Riley, Valentina Silva

Abstract

The proliferation of AI-generated social bots on social media platforms poses a serious threat to public discourse and information integrity. This study evaluates and compares three classes of detection techniques—traditional machine learning (ML), deep learning (DL), and unsupervised clustering—on a recent Twitter dataset enriched with text and user-centric features. We preprocess and clean the raw JSON/CSV files, conduct exploratory analysis, and implement fourML classifiers (K-Nearest Neighbors, Logistic Regression, Support Vector Machine, Random Forest), a deep sequential model (hybrid LSTM), and PCA-driven K-Means clustering. Our results confirm that DL methods outperform classical ML across accuracy, precision, recall, and F_1-score, while clustering provides complementary insights without requiring labeled data. Key findings underscore the pivotal role of "is_verified" and sequential text patterns in bot identification but also highlight limitations due to small sample size and evolving platform policies.

1. Introduction

- **Problem statement:** Automated social bots increasingly influence online conversations, spreading misinformation at scale.
- **Objective:** To systematically compare ML, DL, and unsupervised methods on Twitter bot detection, to find the most reliable way to flag bots.

2. Dataset Description and Purpose

1. **Source & Format**
    - Raw data split into train.json, dev.json, and test.json (plus corresponding "_english" CSV exports). Each listing Twitter user accounts labelled as "bot" or "human."
    - Each record includes key features for each user:
    - **Numeric:** follower_count, friend_count, total tweets, days since account creation.
    - **Flags & categories**: is_verified (yes/no), language of tweets.
    - **Text:** the user's recent tweet text, which we later turn into sequences for deep learning.

2. **Size & Scope**
    - **Training set:** ~5,000 accounts (balanced bots vs. humans)

- **Validation and Test set:** ~1,000

3. **Project Purpose**
    - **Primary goal:** Maximize detection accuracy of AI-generated bot accounts on Twitter.

3. Methods and Why We Chose Them

Approach	Why
KNN, Logistic Regression, SVM	Simple and fast. Good starting points to see if basic features can separate bots from humans.
Random Forest	Tree-based methods handle mixed data types well and often give high accuracy with minimal tuning.
Hybrid LSTM	Reads tweets in order (like reading a sentence) to catch subtle text patterns that static features miss.
PCA + K-Means Clustering	An unsupervised approach: PCA reduces many features into a few summary dimensions, then K-Means groups similar accounts without labels.

4. Step by Step Workflow

Data Cleaning (data_cleaning.py & LSTM_data_cleaning.py)

- Filter accounts so that there are only accounts with at least 70% English tweets and limit the number of tweets per account to 20.
- Drop any columns that have null values or that do not contribute to our project goals. In this case, we drop more than 30 columns.
- Normalize tweets.
- We need to create new columns that can be useful for detection, so we need to count how many URLs, Hashtags, retweets, and links each user shares to see if there's any insights.
- Remove special characters (keep letters, numbers, and spaces) and stop words.
- Extract profile features like number of followers, friends, if they are verified, name, description, account age, and friends to follower ratio.
- Turned yes/no flags into 0/1, languages into one-hot columns.

- Scaled all numeric columns so they sit between 0 and 1.
- For text: tokenized tweets and padded sequences for the LSTM.

Original Dataset:

	ID	profile	tweet	neighbor	domain	label
0	17461978	{'id': '17461978 ', 'id_str': '17461978 ', 'na...	[RT @CarnivalCruise: 🎉 Are you ready to see wh...	None	[Politics, Business, Entertainment]	0
1	1297437077403885568	{'id': '1297437077403885568 ', 'id_str': '1297...	None	{'following': ['170861207', '23970102', '47293...	[Politics]	1
2	17685258	{'id': '17685258 ', 'id_str': '17685258 ', 'na...	[RT @realDonaldTrump: THANK YOU #RNC2020! http...	{'following': ['46464108', '21536398', '186434...	[Politics, Entertainment, Sports]	0
3	15750898	{'id': '15750898 ', 'id_str': '15750898 ', 'na...	[A family fears they may have been cheated out...	{'following': ['2324715174', '24030137', '2336...	[Politics]	0
4	1659167666	{'id': '1659167666 ', 'id_str': '1659167666 ',...	[RT @VonteThePlug: Yeah but he ain't got one h...	{'following': ['1628313708', '726405625', '130...	[Politics]	1
...	...	...	...	...	...	...
8273	1630890068	{'id': '1630890068 ', 'id_str': '1630890068 ',...	[@sethgoldberg17 @jaysonst Fan interference? I...	{'following': ['237453978', '462581299', '1706...	[Sports]	0
8274	713519580757536768	{'id': '713519580757536769 ', 'id_str': '71351...	[@C130Matt I think I heard a voice from out in...	{'following': ['36991422', '32567081', '133983...	[Sports]	1
8275	93345260	{'id': '93345260 ', 'id_str': '93345260 ', 'na...	[@savage_esquire That's unfuckingbelievable.\n...	{'following': ['714636670268792832', '23341114...	[Sports]	1
8276	1749309397	{'id': '1749309397 ', 'id_str': '1749309397 ',...	[@Jomboy_ Doesn't want to pull anymore Hammys\...	{'following': ['3124065581', '413364940', '211...	[Sports]	1
8277	50471224	{'id': '50471224 ', 'id_str': '50471224 ', 'na...	[The sports card market is unreal right now. P...	{'following': ['4202878276', '637216245', '129...	[Sports]	1

After preprocessing:

	label	num_links	num_hashtags	num_mentions	num_chars	processed_tweet	num_followers	num_friends	is_verified	account_age	tweet_length	statuses_count	favorites_count	listed_count	screen_name	description	friends_follower_ratio	sentiment_scores
0	0	0.055556	0.004636	0.016240	0.063536	['ready see newest ships name thanks partners ...	1.259957e-01	0.000159	1	0.857143	0.1	0.001906	0.000181	0.082824	SHAQ	VERY QUOTATIOUS, I PERFORM RANDOM ACTS OF SHAQ...	4.664524e-08	{'neg': 0.032, 'neu': 0.598, 'pos': 0.37, 'com...
1	0	0.025641	0.002140	0.008366	0.037059	['thank mc2020', 'great takeaways last nights...	6.261693e-03	0.000109	1	0.857143	0.1	0.001074	0.001216	0.005818	parscale	Owner @ Parscale Strategy. Senior Advisor Digi...	6.442566e-07	{'neg': 0.11, 'neu': 0.672, 'pos': 0.217, 'com...
2	0	0.042735	0.004636	0.006398	0.061808	['family fears may cheated 20000 donations rai...	2.688967e-03	0.001106	1	0.857143	0.1	0.037524	0.003760	0.003170	FOX13News	Bringing you the important stuff like breaking...	1.516363e-05	{'neg': 0.106, 'neu': 0.735, 'pos': 0.159, 'co...
3	1	0.025641	0.000000	0.006398	0.028440	['yeah but aint got happy song nigga always di...	1.093688e-04	0.000149	0	0.500000	0.1	0.000020	0.000930	0.000080	VonteThePlugNC	MOTIVATION 3 OUT NOW 🔥 Singles: 'Lil Shawdy' &...	5.024210e-05	{'neg': 0.08, 'neu': 0.671, 'pos': 0.249, 'com...
4	0	0.042735	0.000000	0.005413	0.036206	['aloha welcome back space', 'ms tree catches ...	1.034388e-01	0.000022	1	0.785714	0.1	0.000949	0.000226	0.048988	SpaceX	SpaceX designs, manufactures and launches the ...	7.882152e-09	{'neg': 0.013, 'neu': 0.895, 'pos': 0.092, 'co...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9920	1	0.034188	0.001070	0.015256	0.054554	['chicago cubs sign former texas tech reliever...	6.369723e-05	0.001867	0	0.571429	0.1	0.000453	0.000096	0.000193	AlanReifman	Texas Tech professor of human devt and family ...	1.080528e-03	{'neg': 0.006, 'neu': 0.734, 'pos': 0.26, 'com...
9921	1	0.023504	0.002140	0.008858	0.042350	['cardsfromattic gmas superior graded crumbled ...	6.932820e-05	0.000094	0	0.214286	0.1	0.009760	0.160603	0.000100	CardsFromAttic	Satirizing the sports card industry one tweet ...	4.998126e-05	{'neg': 0.081, 'neu': 0.735, 'pos': 0.185, 'co...
9922	1	0.014957	0.004993	0.014764	0.035630	['rt follow originalfunko chance win disney pa...	2.536398e-06	0.000452	0	0.500000	0.1	0.000574	0.002006	0.000005	bkgreen09		6.566248e-03	{'neg': 0.033, 'neu': 0.575, 'pos': 0.392, 'co...
9923	1	0.004274	0.000357	0.013780	0.042820	['run numbers buy bitcoin', 'imagine telling f...	1.264095e-06	0.000235	0	0.571429	0.1	0.000030	0.000800	0.000000	bilalko14		6.846232e-03	{'neg': 0.179, 'neu': 0.71, 'pos': 0.111, 'com...
9924	1	0.014957	0.003210	0.009843	0.032472	['catcher bruce maxwell kneeling national anth...	5.581716e-07	0.000214	0	0.357143	0.1	0.000040	0.000442	0.000000	IfYouBuildIt_	We're a father/son duo sharing the	1.410487e-02	{'neg': 0.06, 'neu': 0.8, 'pos': 0.141, 'compo...

Exploratory Analysis (ExploratoryAnalysis.py)

- Plotted histograms and boxplots (e.g., followers distribution) to see differences.
- Checked correlations to spot overlapping features.
- Observed that bots often have fewer followers and fewer tweets, and are rarely verified.

2. **Train/Dev/Test Split** (LSTM_data_cleaning.py, SplittingDataset.py)

- Ensured each set kept the same bot/human ratio (70% train, 15% dev, 15% test) for the LSTM models, and (80% train, 20% test) for other classification algorithms.
- Saved these splits for reuse.

3. **Feature Preparation**
   - **For ML:** used the cleaned numeric and one-hot features directly.
   - **For DL:** combined padded tweet sequences with user features in a hybrid Keras model.
4. **Model Training & Tuning**
   - Ran grid searches for each ML model's key parameters (e.g., k in KNN, C in SVM).
   - For the Random Forest (RandomForrest.py): tuned number of trees and tree depth.
   - For the LSTM (LSTM.py): used an embedding layer, a bidirectional LSTM, and dense layers; monitored performance on the dev set and saved the best model.
5. **Unsupervised Clustering** (Kmeans_pca.py)
   - Applied PCA to get enough variance (80% and 90%) into a small number of components.
   - Ran K-Means (k=2, k=3) using those components.
   - Compared cluster membership to true labels to measure cluster purity.
6. **Evaluation** (EvaluateModel.py)
   - Measured accuracy, precision, recall, F_1-score, and ROC-AUC for each model on the test set.
   - Generated confusion matrices and ROC curves.
   - For clustering: reported purity score and silhouette coefficient.

5. Results & Explanation

Model	Accuracy	Key Takeaway (plain)
KNN	0.86	Looks at nearest neighbors; okay but misses complex patterns.
Logistic Regression	0.87	Uses a straight decision boundary; better than KNN but still limited.
SVM (RBF)	0.87	Finds more flexible boundaries; handles tricky cases better.
Random Forest	0.86	Builds many decision trees; strong at combining simple rules.
Bi-LSTM	87	Reads tweets as sequences; catches patterns other models miss—top performer.
PCA + K-Means	**0.88**	Groups similar accounts; good at rough separation but can't match labeled models.

Clustering Wins: While all the models performed very similarly regarding their accuracy, clustering came out on top with an accuracy of 88%.

Key features: "is_verified" surfaced as a top predictor in tree models; nonetheless, platform policy changes (e.g., paid verification) threaten its future reliability.

7. Conclusion

This project shows that clustering excels at spotting AI-generated bots on Twitter. This is very useful as it removes the need to have labelled data which is a big hurdle in data analysis. To stay ahead of smarter bots and policy changes, integrating LLMs, expanding data, and adding network features will be key.

Additionally, the reliability of the is_verified feature has diminished, because Twitter's paid verification system now allows bots to purchase verification, making it a less trustworthy indicator. Also, due to the size of the data, the algorithms were prone to overfitting.