



Article

Profiling Astroturfers on Facebook: A Complete Framework for Labeling, Feature Extraction, and Classification

Jonathan Schler  and Elisheva Bonchek-Dokow * 

School of Computer Science, HIT (Holon Institute of Technology), Holon 5810201, Israel; schler@hit.ac.il

* Correspondence: elishevab@hit.ac.il

Abstract: The practice of online astroturfing has become increasingly pervasive in recent years, with the growth in popularity of social media. Astroturfing consists of promoting social, political, or other agendas in a non-transparent or deceitful way, where the promoters masquerade as normative users while acting behind a mask that conceals their true identity, and at times that they are not human. In politics, astroturfing is currently considered one of the most severe online threats to democracy. The ability to automatically identify astroturfers thus constitutes a first step in eradicating this threat. We present a complete framework for handling a dataset of profiles, from data collection and efficient labeling, through feature extraction, and finally, to the identification of astroturfers lurking in the dataset. The data were collected over a period of 15 months, during which three consecutive elections were held in Israel. These raw data are unique in scope and size, consisting of several million public comments and reactions to posts on political candidates' pages. For the manual labeling stage, we present a technique that can zoom in on a sufficiently large subset of astroturfer profiles, thus making the procedure highly efficient. The feature extraction stage consists of a temporal layer of features, which proves useful for identifying astroturfers. We then applied and compared several algorithms in the classification stage, and achieved improved results, with an F_1 score of 77% and accuracy of 92%.

Keywords: astroturfing; Facebook; machine learning; labeling; feature extraction



Citation: Schler, J.; Bonchek-Dokow, E. Profiling Astroturfers on Facebook: A Complete Framework for Labeling, Feature Extraction, and Classification. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2183–2200. <https://doi.org/10.3390/make6040108>

Academic Editor: Angelo Cangelosi

Received: 21 August 2024

Revised: 12 September 2024

Accepted: 26 September 2024

Published: 30 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

During the years 2019 and 2020, a political drama took place in the State of Israel: three consecutive elections were held, none of which resulted in the formation of a governing coalition. This frustrating political situation had an upside, from the point of view of our research: over this time period, a massive amount of data was generated on online social media (OSM) by politically minded users. One of the main interests in this type of dataset lies in identifying astroturfers. “Astroturfing” is defined by [1] as the practice of assisting politicians in promoting political or legal agendas by mobilizing people to sympathize and support their cause. This is achieved by creating an illusion of public consensus, which does not necessarily exist. This deceptive, strategic top-down activity [2] by political actors mimics the bottom-up activities of autonomous individuals.

As OSM platforms have become more widely used by the general population, they have also become more widely abused by actors interested in manipulating public opinion. Perhaps the most troubling is when astroturfing is organized by governments or political parties, who harness multitudes of users to undermine public opinion. These organizations are known as cyber troops [3]. The growth of astroturfing has gone hand in hand with the realization of its dangers. Today it is considered one of the most impactful threats online [4] and is prevalent across the globe [5]. The realization that the malicious use of OSM platforms vitiates trust and faith, the cornerstones of any democratic society, has prompted a surge of research to counter it, specifically research based on artificial intelligence [6] and machine learning.

This study contributes to this collective effort, by examining data from Facebook users during several rounds of recent Israeli elections to identify lurking astroturfers. Similar studies have been conducted on data from other countries [7–10]. Although countries have their own unique political system, the conclusions from each can often be useful to others. For instance, in the United States, the same two political parties are on the ballot in every election, whereas in Israel, the number of parties is not limited, and their identity and composition change from one election to another. Yet, insights gained by analyzing astroturfers operating in the Israeli system can be usefully applied to the American system. Moreover, due to its nature, the Israeli system can serve as a virtual Petri dish for the growth of astroturfers, at a rate much higher than in the American system. When elections are held repeatedly in this type of system, within a short time period, it provides a rare opportunity to collect and analyze a very large dataset containing a potentially high number of astroturfers.

Four main challenges present themselves when attempting to create a complete framework for differentiating between astroturfer and authentic user profiles on Facebook that are consistent with four key stages in the astroturfer identification process: collecting the data from Facebook, labeling enough data to create a dataset on which to run supervised learning algorithms, selecting and extracting relevant features, and choosing the algorithm best suited for the task. We have dealt with some of these challenges in previous studies. Here, we expand these solutions and incorporate them into one comprehensive framework.

This study innovates in several ways:

- In the data collection stage, we built a huge dataset with a multitude of interesting relevant information that benefited from the unique political circumstances at hand.
- In the labeling stage, we devised a novel approach for efficiently labeling the data that widens the bottleneck caused by the need to conduct expensive and time-consuming manual labeling.
- In the feature extraction stage, we presented an original set of features that adds an additional temporal layer to the data. These features emerged from the special election circumstances. The classification stage compared several learning algorithms and justified the use of the temporal layer of features.

This paper is organized as follows: in Section 2, we present related work for each of the four stages incorporated into our framework. Section 3 provides a general description of the framework. Separate sections are devoted to each stage in the framework. These provide details on the methods and results while highlighting the contributions: Section 4 describes how the data were collected, Section 5 describes how the data were labeled and presents our novel honeypot approach, Section 6 describes the feature set, along with the new temporal layer, and Section 7 presents the experimental results. We conclude in Section 8 with a discussion of the main contributions of this work, and some ideas for future research.

2. Related Work

Several comprehensive surveys of the astroturfing detection literature have been conducted [4,11]. One of the best-known approaches to identifying astroturfers is presented in [12], who developed a framework for analyzing the behavior of users and tracking the diffusion of ideas on Twitter in the run-up to the 2010 US midterm elections. Ref. [4] examined one of the first confirmed cases of electoral astroturfing during the South Korean presidential election of 2012, using a combination of graph-based techniques and content analysis. Ref. [13] used semantic analysis as well as behavioral analysis to discover paid posters on news websites. More recently, a form of authorship attribution using binary n-grams for content analysis was applied by [14] to social media data. In this section, we refer to those studies that are pertinent to our research, in that they attempt to tackle those parts of the astroturfer identification process for which we propose improvements. We begin with a clarification of the term “astroturfing”, and then, introduce studies dealing with data collection, labeling, feature extraction, and classification in the context of astroturfing.

One of the first in-depth studies of digital astroturfing appeared in [1]. Since then, the term has developed and taken various forms as astroturfing itself has mutated. Any attempt to automatically identify astroturfers must first define what exactly is being targeted. Different definitions are based on different aspects. One aspect is individual vs. organized since a user can either be acting independently and autonomously or can be part of an organized entity [15]. Another useful classification is the human vs. software agents: the user can be a real human, who, for whatever reasons, chooses to interact behind a fake profile, or the profile may be operated by what is known as a *bot* [16]. There are also intermediate states, sometimes referred to as *cyborgs*, that can be bot-assisted humans or human-assisted bots [17]. Terms such as *sybils* and *trolls* are also types of fake profiles [3,18]. Thus, the term *fake* can have different meanings in different contexts. Each meaning gives rise to an entity that can act as an astroturfer. Although some researchers make a distinction between the different types [19], we take a more inclusive approach. See, for example, [20], who uses the term *astroturfers* specifically for human entities acting in an orchestrated manner, as opposed to *bots*. In this study, we target profiles whose long-term behavior raises suspicions that they are fake, in the general sense that they do not belong to authentic human users, but rather to users—whether human or artificial, acting alone or in an orchestrated manner—with a clear-cut, political goal.

2.1. Collecting Facebook Data

Astroturfing can be observed on various OSM platforms, each of which has its own characteristic modes of interaction and underlying data structure. For this reason, studies usually focus on one specific platform. We chose Facebook as our source of data, for both practical and theoretical reasons. Facebook is highly popular in Israel, much more so than Twitter, which seems to be more popular elsewhere [9,19,21–23]. Theoretically speaking, Facebook is a more challenging environment, and as such, less researched. The main challenge is posed by restrictions on automatic data extraction and analysis. By contrast, in the Twitter environment, data can be collected more easily. This can explain both the smaller number of studies on fake profile identification on Facebook as compared to Twitter, and also the smaller size of the datasets or shorter time spans in studies on Facebook, as compared to Twitter. See, for instance, the relatively small Facebook datasets in [24–26] and the short time span (three months) of [27]. There are two notable exceptions: one is the dataset collected by [28]. It spans 15 months and contains close to four million Facebook posts, of which several hundred were analyzed. However, they aimed to expose public Facebook pages promoting a malicious agenda, while falsely posing as authentic. It is important to note that on Facebook *pages* are not the same as *profiles*: pages generally represent organizations (political ones, in our context), and as such are usually managed by a team of people behind the scenes. Profiles, on the other hand, supposedly belong to individual users. Thus, while our work is similar to that of [28] in scope and size, we specifically attempted to uncover seemingly authentic and independent personal profiles that actually belonged to an organized entity attempting to manipulate public opinion.

The second recent exception is work by [29]. It analyzes several million comments of Facebook users which discuss posts on public pages of several verified news outlets to uncover the bots among these users. Since this study is similar to ours in terms of the domain and size of its dataset, as well as in its goal, we refer to it as state of the art and use it as a baseline for comparison with our results.

2.2. Labeling the Data: Obtaining Ground Truth

As with any supervised learning algorithm, the preliminary stage consists of collecting enough training examples that are already labeled. In the context of astroturfer identification, most such collections have been created manually by human experts, who assessed the authenticity of the profiles. While there are other approaches (such as enlisting crowdsourcing [28,30]), most works are not assisted by automatic techniques in the data labeling stages. The judges employed by [31], conducted Google Image searches to authenticate

user profile images. They also ran text searches by comparing personal information posted by a profile to information about the profile posted by its friends. A bot annotation protocol was devised by [32], and followed by [29], in which two human judges independently reviewed the user profiles until a conclusion was reached. Then, the two lists were merged and conflicting classifications were resolved.

The shortcoming of the manual approach is the fact that any random collection of profiles is expected to have a low percentage of astroturfers, so the amount of human labor invested is disproportionate to the number of astroturfers harvested. One method for overcoming this problem was devised by [26,33], using the notion of “social honeypots”; i.e., fake profiles created and configured specifically to attract astroturfers. The set of profiles attracted to the honeypot can then be handed over to human judges for manual labeling. In this way, a higher percentage of astroturfers are labeled than if a randomly collected set was used.

2.3. Selecting and Extracting Features

The features available for extraction from OSM profiles can be divided into several categories. A hybrid and diverse feature space, with features chosen from different categories, has higher value and contributes more to the robustness of the framework as a whole [4]. Below, we review the three main feature categories that have been used in previous works.

2.3.1. Profile-Based Attributes

The first category consists of attributes made publicly available by the profile owner, such as date of birth, gender, level of education, as well as the existence or absence of an avatar photo [31]. Profile attributes contain valuable information: one study showed how a profile can be classified as an impostor solely based on the photo [34]. It is interesting to note that what and how much information the user chooses to make public can in itself be used to gauge the authenticity of a profile since users setting up a fake profile might not be as eager to share personal data, however innocuous they may be. In addition, text analysis of the user name may also contain information pertaining to the task at hand. A number of other metrics have been proposed, such as how far back in time the profile came into existence. However, this information is not as easily accessible on Facebook as it is on Twitter [35].

2.3.2. Social Graph-Based Attributes

The second category takes the network of connections between the user and its friends and followers into account. This network can have different properties, as a function of the profile’s authenticity, that can be uncovered by graph-based algorithms [36]. The metrics include the number of friends, fluctuations in this number over time, and the structure of the graph induced by the profile [37]. The approval rate of friend requests has also been suggested in this context [38]. Ref. [39] also used graphs to describe the relationships between users and constructed a neural network that used this information, along with stylometry, to detect astroturfers in user-recommendation data.

2.3.3. Activity- and Content-Based Attributes

The third category is based on the text content created by the profile (e.g., posts and comments) as well as the modes of interaction the profile makes use of (e.g., “likes” and shares). The text content can be analyzed with text analysis algorithms. For example, Ref. [40] used text content alone to identify fake news on social media, regardless of who was producing it. By contrast, Ref. [41] ignored the text of the posts, and relied solely on modes of interaction with the post. Ref. [42] analyzed the underlying dynamics of the online interactions, and specifically, the pattern of the user’s reaction time, to differentiate between human users and bots. Ref. [43] also used interface interaction such as mouse clicks and keystroke data to distinguish bots from humans [43].

Different metrics can be calculated from the bare text, alongside the other modes of interaction. These metrics include the text length, the interaction frequency, and the ratio of

reactive comments to proactive comments (e.g., the number of comments made to existing posts versus the number of new posts made by the profile). For many other content-based features, see [35], who created BotOrNot for detecting fake Twitter accounts by generating several hundred content-based features (among others).

2.4. Machine Learning Algorithms

A large range of machine learning algorithms have been applied to the problem of astroturfer detection, including decision trees [31], naïve Bayes, logistic regression, support vector machines [33], Markov clustering [25], and neural networks [39]. Non-linear dynamics [42] have also been applied. We consider two works to be state of the art, both of which deal with a volume of Facebook data comparable to ours to identify fake profiles. Ref. [28] showed good results using logistic regression, and [29] achieved an F_1 score of 71% using an indicator capture-region model equivalent to that used by [32].

3. Framework

Below, we introduce our complete framework (see Figure 1). This framework allows for a comprehensive workflow, from the data collection stage, through the labeling stage, feature extraction, and finally, the machine learning of the classifier. While the framework in itself has features in common with other solutions, we present novel algorithmic contributions in each part, thus making this framework relevant and distinct from other works.

1. **Data collection:** This involved collecting a corpus of Facebook data consisting of public posts and comments from the public pages of political candidates over a time frame surrounding elections. Our dataset was collected over a time frame of 15 months, during which time three consecutive elections were held, which resulted in a huge, rich dataset. We extracted a collection of profiles from these posts and comments, which were then classified.
2. **Labeling:** This involved creating a subset of labeled profiles, which included enough astroturfers for a useful model to be built. This is time consuming and expensive since it is performed manually and the percentage of astroturfers within any random subset is very low. We expanded and perfected a honeypot approach, resulting in a sufficiently large labeled dataset, with a sufficiently high percentage of astroturfers, in a labor-efficient way.
3. **Feature extraction:** Given the raw statistics and textual features, this stage involved devising new and useful features for the task at hand. We introduced a temporal layer of features, which showed a proven advantage over other feature sets.
4. **Machine learning:** Given the labeled dataset with the extracted features, machine learning algorithms were applied. We report the results on a larger dataset while fine-tuning the parameters for optimal results.

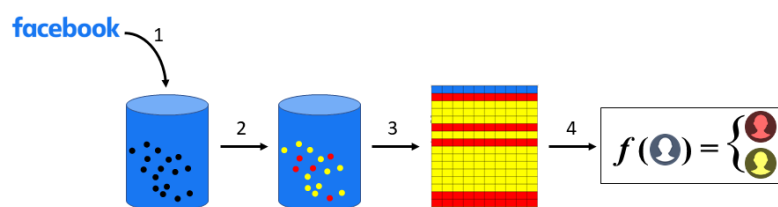


Figure 1. Framework: (1) data collection; (2) data labeling; (3) feature extraction; (4) model learning.

4. Data Collection

Between January 2019 and March 2020, we collected a large dataset of Facebook users who interacted with political candidates by commenting on their posts. Presumably, some proportion of those comments were generated by astroturfers, and we intended to use the dataset to develop a classifier to distinguish between authentic users and astroturfers.

The process of creating a dataset of astroturfers and authentic users began by selecting 12 leading Israeli politicians (from the 120 members of parliament) who represent different

parties and ideological wings (most of the 12 were party leaders) up for election. We hypothesized that, on the one hand, an analysis of 12 candidates would generate a sufficient amount of data to represent the parties, while, on the other, would be small enough to be feasible for the semi-manual labeling task to follow (see next section). The public Facebook pages of these politicians were the arena from which profiles were collected for the dataset: from these pages, we collected posts, and from these posts, we extracted comments and reactions (e.g., number of “likes”, number of shares).

Facebook generates a unique ID for every user. For every comment or reaction to a post on a politician’s page, we extracted the publisher’s ID. Our dataset consisted of nearly half a million user IDs, which generated nearly four million comments and reactions on the pages of the selected politicians. For a small subset of the users that interacted with the chosen candidates’ pages (approximately 1500), statistical information was gathered from their public profile pages, such as the number of friends, presence of an image, presence of a sticker on the profile image, etc. Only information voluntarily made public by the user was collected in this way. No personal identifiable information was collected or stored.

Facebook’s strict privacy policies pose a challenge to data mining researchers. Nevertheless, we were able to create a large dataset over a time span of 15 months, similar to the size and time span of [28], while complying with Facebook’s terms and not violating users’ privacy, even though we did not restrict ourselves to public pages. Collecting such a large dataset of Facebook comments and users is an accomplishment in and of itself. Surprisingly, as the results show, we achieved high performance on the task of determining a profile’s authenticity using only public information, without relying on private profile information.

5. Labeling the Data: The Honeytrap Approach

The straightforward honeytrap approach, in our context, would require creating a fake political candidate profile. Obviously, this would be futile, in that it would not attract any astroturfers. Instead, we considered the existing profiles of the political candidates as natural honeytraps, in the sense that they attract astroturfer profiles by their very nature. Indeed, astroturfers are attracted to these profiles, as are innocent, normative users. One of the many features we generated emerged as useful in this context: namely, the ordinal number of a user’s comment within all comments on a post. Apparently, astroturfers are very often among the first commenters on posts. Of course, this makes sense, given their eagerness to make an impact with their online presence. But it turns out that this facet of their behavior is also a ‘tell’ that creates an astroturfer-rich pool of commenters in which they can more easily be found.

Based on this observation, instead of collecting all profiles commenting on the politicians’ posts, we limited ourselves to those profiles belonging to users who were among the first several commenters. Only these profiles were presented to the human judges for labeling. In other words, instead of selecting profiles from all comments, we selected only from the top K comments, with K set to 10. Choosing 50 posts randomly, and the 10 top comments from each post resulted in a set of 500 comments. By random sampling with returns from this set, we extracted a subset of profiles to be handed to the judges for manual labeling. It is interesting to note that while we had 500 comments, the size of the set of users who generated them was 300. This is due to another aspect of astroturfing behavior: not only are they quick to comment, but they also tend to comment more than once. This method indeed proved useful for zooming in on astroturfers: 20% of the profiles targeted in this way turned out to be astroturfers according to the human raters.

We hypothesized that the percentage of astroturfers labeled in this way would be significantly higher than for any random collection of profiles. In order to test this, we compared our method to three other data collection methods involving human judges:

1. Random baseline: Where 400 random profiles were selected from the complete pool of profiles who commented on the posts. Of these, only 9 (2.25%) profiles were labeled by the judges as astroturfers. This method makes the process highly inefficient.

2. Chosen posts: Where instead of sampling profiles from the pool of profiles who commented on the posts (as above), comments were sampled from the pool of all comments. This method is based on the insight that astroturfers generate comments at a much higher proportion than their own proportion in the population. This approach indeed yielded a higher proportion of labeled astroturfers: 46 of 400 profiles (11.5%).
3. Preliminary sifting: Where we applied preliminary manual sifting of profiles to create a set of suspicious-looking profiles, which were then presented to the judges for labeling, using the same guidelines. The preliminary sifting proved useful in that, of the 364 profiles, 76 (20.88%) were labeled as astroturfers.

As can be seen in Figure 2, our Top-10 method was on a par with the preliminary sifting approach in terms of the percentage of astroturfers produced. However, recall that the latter requires much more manual effort. To overcome this issue, our method employs human judges along with a variant of the honeypot technique that maximally capitalizes on human effort and time invested to produce a large-size, astroturfer-rich dataset.

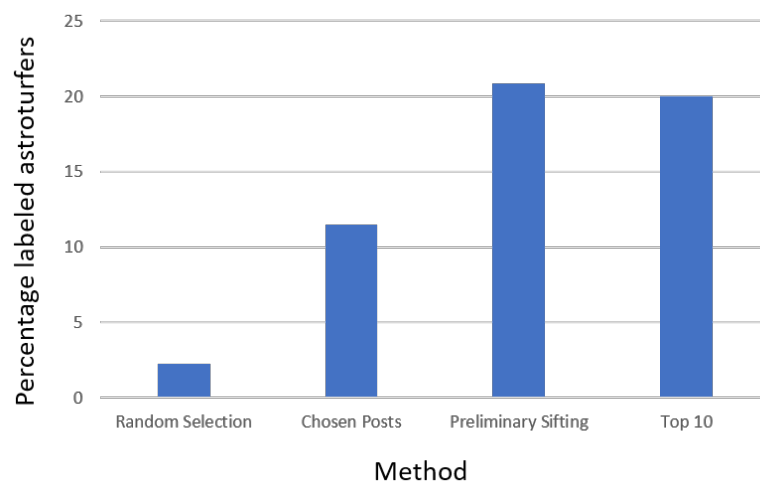


Figure 2. Percentage of labeled astroturfers by method.

Independent samples *t*-tests were conducted to compare the performance of the different methods. The statistical significance of these results is summarized in Table 1. Note that the Top-10 method was highly significant when compared to the random baseline, and significant when compared to the chosen posts method.

Table 1. Statistical significance of honeypot experiment I results: ** $p < 0.01$; * $p < 0.1$.

	Random Baseline	Chosen Posts	Preliminary Sifting
Chosen posts	**		
Preliminary sifting	**	**	
Top-10	**	*	

These results of **experiment I** of the honeypot method were presented in [44]. In the current work, we report on two further experiments. In **experiment II**, we tested our initial, intuitive, choice of $K = 10$ by testing and comparing different values of K and noting the percentage of astroturfers labeled for each value. The results are summarized in Figure 3 and confirmed that the choice of $K = 10$ produced the best results.

An independent samples *t*-test revealed that the choice of $K = 10$ was statistically significant ($p < 0.1$) as compared to the choice of $K = 1000$. These two values of K were the two extreme values tested. Although the intermediate values of K were not statistically significant, there was a dominant trend, which possibly would have been significant with a larger sample.

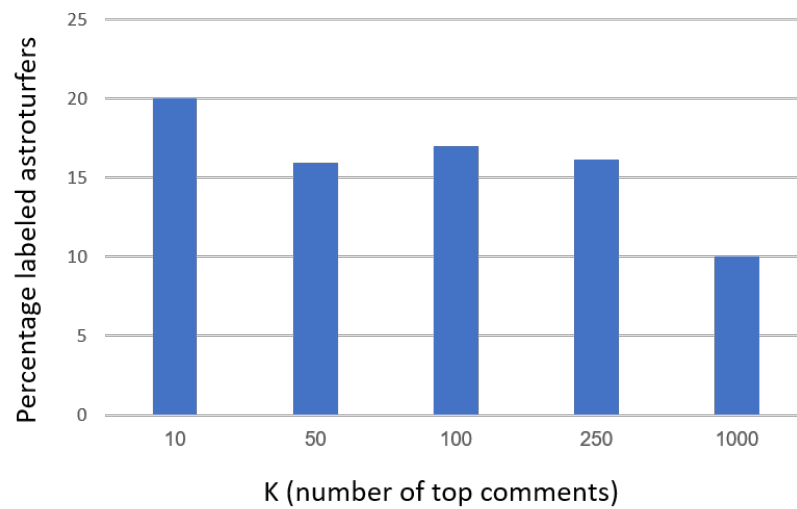


Figure 3. Percentage of labeled astroturfers for top K comments.

In **experiment III**, we assessed whether time was a factor, as opposed to precedence per se. That is, we compared the percentage of astroturfers labeled in comments collected within the first M minutes of the post's appearance for different values of M. The results are summarized in Figure 4. The best values achieved (22%) were for the first 5 and first 10 min and were slightly better than the Top-10 method. Independent samples *t*-tests showed that the choice of $M = 2$ was statistically significant ($p < 0.1$) when compared to the choices of $M = 5$ and $M = 10$. So were the values of 5 and 10, as compared to 30 and 120.

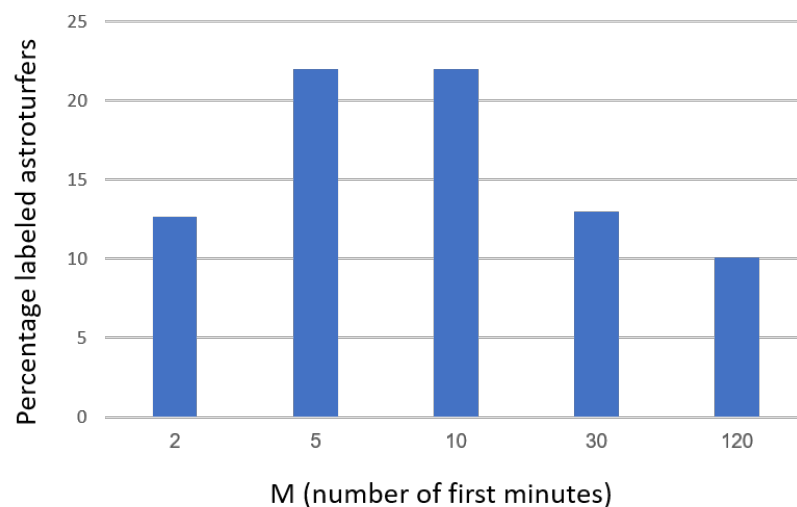


Figure 4. Percentage of labeled astroturfers for comments within the first M minutes.

This honeypot method made it possible to more efficiently label a dataset of 1234 profiles. The labeling was performed by three experienced judges, all familiar with the Israeli political scene as well as with OSM platforms in general and Facebook in particular. They received clear, detailed guidelines as to how to label profiles, analogous to the guidelines deployed by [32]. Each profile was assigned to two of the three judges. When the two assigned judges agreed on the labeling, the profile was labeled according to their joint decision. When they disagreed, the profile was not labeled and was not included in the dataset. Cohen's kappa score was used to assess the level of agreement between the pairs of judges. This procedure resulted in a value of 0.69 for this score.

6. Feature Selection and Extraction

By applying exploratory data analysis to these data, we extracted a total of some 80 features, which were divided into three subsets: profile information features, profile activity features, and temporal features. Initially, we intended to only use profile information and activity features, but as the political situation unfolded, and the data kept flowing in, with another and yet another round of elections, we realized that much information was to be found not only in the activity itself but also in how the activity changed within different temporal windows before, during, and after each round of elections. We next elaborate on these subsets of features.

6.1. Profile Information Features

Of the features associated with a user profile, some are provided by the user explicitly within the account data, such as name, age, gender, level of education, and marital status, and some can be generated from the account's activity. We did not make use of the explicit account data, as they are mostly private information, and as such require the user's permission to access them. In the following, we describe the account data features that are publicly available, either directly or indirectly.

- **Image:** A Boolean value indicating the profile picture's availability/authenticity. If it was both available and authentic, the value assigned was 1. Otherwise, the value was 0. The evaluation was performed manually.
- **Sticker:** A Boolean value indicating whether the user's profile image was a political-oriented sticker. If so, or if it had been so in the past, before being changed, the value was 1. Otherwise, it was 0. This feature was also manually evaluated.
- **Friends:** The number of friends of the user. If the user had no friends, or if the number of friends was not available, the value ascribed was 0. The rationale behind using zero for unavailable information was based on the fact that there were no profiles with zero friends in our dataset, so we could use the zero value as an indicator that the information was intentionally concealed.
- **Followers:** The number of followers of the user. If the user had no followers, or if the number of followers was not available, the value was 0. The rationale for using zero here was identical to the Friends feature.
- **Name's Charset:** The character set in which the user's name is presented (in our dataset it was most often Latin or Hebrew, but there were several others that occurred less frequently).
- **Multiple Aliases:** While every profile has a unique ID number, the name associated with the profile need not be unique. The same name can belong to more than one profile and, more relevant in this context, the same profile can have more than one name, since Facebook allows changing the profile's name. The existence of many names for one profile is a good indicator of a fake account. Whether a name change is significant, and therefore, suspicious or not, can be assessed by calculating the edit distance between the different names.
- **ID Account:** The URL of a Facebook profile can contain either the profile name or its ID number. This feature is a Boolean value indicating which of the two possibilities occurs in the profile's URL.
- **Account's Age Estimate:** Some profiles were created right before an election, and as such, were suspicious, and the closer the creation date to the election date, the higher the suspicion. Unfortunately for us, Facebook enables its users to hide their profile creation date. In order to overcome this problem, we made use of the fact that profile ID numbers on Facebook are ordinal, and are allocated sequentially. Thus, we were able to take those profiles for which both the date of creation and the profile ID were available and train a regressor, using a heuristic, to extrapolate an estimated date of creation for those profiles for which only the ID was available. Our assessment of the effectiveness of the method revealed that in 99% of the cases, the estimation was within one month (before or after) of the actual profile creation date.

6.2. Profile Activity Features

The set of profile activity features for each user profile was generated by tracking the account activity of the profile: information regarding all posts and comments by the user was registered. Both posts and comments on Facebook have a unique ID and a publication date, which we collected. The text itself was also stored. In addition, we also recorded several other pieces of information for each post: the name of the politician on whose page the post was published, whether it was a live post or not, and approximations of how many shares and how many comments the post received. For comments, we also recorded the ID of the post on which it was made, and a value indicating whether media content, such as an image or sticker, was included in the comment. By matching the IDs of the posts, comments, and users, all the posts and comments for a given user were retrieved and analyzed. This produced a rich feature set that encapsulated various aspects of the users' interactions on politicians' pages.

Various calculations were made on the raw numerical features, resulting in more features. The choice of these features was based on the ways fake profiles were expected to behave, as opposed to authentic profiles. For example, a high comment-per-time-unit ratio should reasonably raise suspicions regarding the authenticity of the profile generating those comments. The same reasoning applies to the feature of average time intervals between comments: a profile having a low value for this feature is less likely to belong to an innocent, normative user. Both of these features can be indicative of an active attempt to propagate ideas and opinions. Another example of suspicious behavior that can be captured as a feature is a peak in activity leading up to an election, with a sharp drop right after the election. Obviously, none of these features can be reliably used alone in order to reach a decision as to the authenticity of a profile. For example, some astroturfers could have a low comment-per-time-unit ratio, or there could be non-astroturfers with a high comment-per-time-unit ratio. This variability is demonstrated in Figure 5. It depicts box plots for several profile activity features: comments (a), URP (b), posts (c), WRS (d), NPC (e), and ATD (f) (see Table 2 for a description of these features). For the descriptive statistics of astroturfers and authentic users, see Table 3.

Table 2. Descriptions of the features appearing in Table 4.

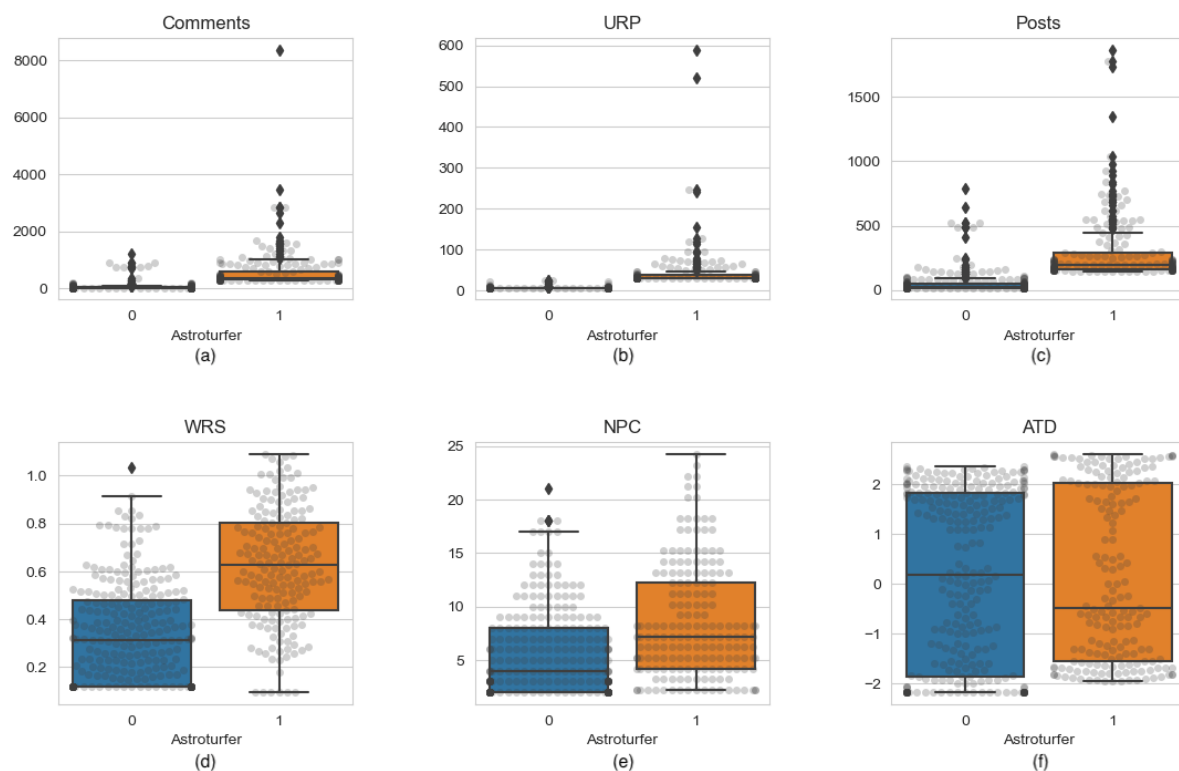
Feature Name	Feature Description
Comments	Total number of comments made by the profile
URP	Unique count of phrases repeated by the user
Posts	Total number of posts the profile engaged in
WRS	Word repetitiveness score
NPC	Number of different pages the profile commented on
ATD	Average time delta between comments in the same post
Followers	Number of followers
Sticker	Indicates if a profile image includes a political sticker
PPD ratio	Avg. post per day ratio: post-election/election month
Age	Estimated age of account
ADR ratio	Activity rate ratio: post-election/election month

Table 3. Descriptive statistics for different profile types.

	Authentic Users			Astroturfers		
	Min	Mean	Max	Min	Mean	Max
Comments	23.53	74.00	1217.53	276.45	617.00	8346.45
URP	7.47	8.00	25.47	30.98	48.00	587.98
Posts	17.97	57.00	784.97	140.87	307.00	2.60
WRS	0.12	0.34	1.04	0.10	0.62	1.09
NPC	2.00	5.50	21.00	2.20	8.70	24.20
ATD	−2.18	0.11	2.36	−1.95	0.13	2.60

Table 4. Mean values of top features for astroturfer versus authentic profiles for all feature categories.

Feature Name	Authentic Profile	Astroturfer Profile
Comments	74	617
URP	8	48
Posts	57	307
Followers	81	311
Sticker	0.17	0.51
WRS	0.34	0.62
NPC	5.5	8.7
PPD ratio	0.7	0.3
Age	60	79
ADR ratio	0.62	0.31
ATD	0.11	0.13

**Figure 5.** Boxplots for several features (in (a–f)). Astroturfer = 0 for authentic users, and Astroturfer = 1 for astroturfers.

The figure shows the distinctive behavior for each of the fake/authentic categories, but for some of the features, there was considerable overlap (most strikingly for the ATD feature (f), and somewhat less so, albeit still considerable, for the NPC (e) feature). However, using a large set of various features, each capturing a different aspect of the profile and its behavior, was sufficient for the statistical characteristics of fake versus authentic profiles to come through. In fact, a successful astroturfer detection system can only work when many diverse features are combined.

In the following, we describe the textual features, account activity metrics, and time delta features.

6.2.1. Textual Features

Astroturfers attempt to inculcate their ideas, and thus, tend to repeat themselves. With this in mind, we devised two repetition scores: one for words, and one for phrases. Both scores ranged from 0 to 1. For words, the minimal score of 0 indicated no repetition

by the user, and the maximal score of 1 indicated that the same words appeared in all comments by the user. For phrases, the word combinations that were repeated by the user were counted, and their ratio to the total number of the user's comments was calculated. We also calculated the ratio of non-textual comments (such as gifs or links) to the total number of comments by the user.

6.2.2. Account Activity Metrics

Different aspects of users' account activity were captured numerically. These included when the user's first and last comments appeared (within the dataset), the number of pages and posts on which the user commented, the number of comments authored by the user, the average number of posts per day, the ratio of days in which the user was active (within the dataset), and the ratio of "live" posts (among all posts) on which the user commented.

6.2.3. Time Delta Features

To gauge the frequency/intensity of the user's activity, we devised several time delta features. These included the average time deltas between adjacent comments by the user within the same day, within the same post, and, regardless of day or post, between the post's publication and the comment's publication. In addition, a time homogeneity score was formulated to estimate whether there was an underlying temporal pattern to the user's activity, or whether it was randomly distributed over time.

For each account activity feature, we calculated the mean value within each category (authentic and astroturfer). The ratio between the two means was used to rank the features. The features listed above in Table 3 are ordered according to this rank. The differences between the two categories are significant using the independent samples *t*-test (p -value ≤ 0.01).

6.3. Temporal Features

Observation of profile activities over time revealed that these were not static, but rather changed as election day approached. These changes show distinctive patterns for authentic users and astroturfers. For example, the values of one of the temporal features—PPW (posts per week)—over the course of the assessment period, are depicted in the graphs in Figures 6 and 7. Each week has one data point, and these data points are connected to form a continuous graph. The darker vertical shaded areas mark the month around each election date: the three weeks leading up to it, and one week following it. Within each election block, the dashed vertical line indicates the election date, and the solid vertical line pinpoints a week before the election date. Figure 6 shows the feature's average value for authentic users, and Figure 7 shows its average value for astroturfers. The feature's overall average (for both profile types taken together) is marked by a dashed horizontal line in both figures. These graphs show distinct behavior patterns for each profile type. For instance, while astroturfers' posts peaked on election day, authentic users' posts peaked beforehand, and then, dropped on election day. Also notable is the observation that astroturfers' activity was almost non-existent between elections, while authentic users continued to be active throughout the timeline.

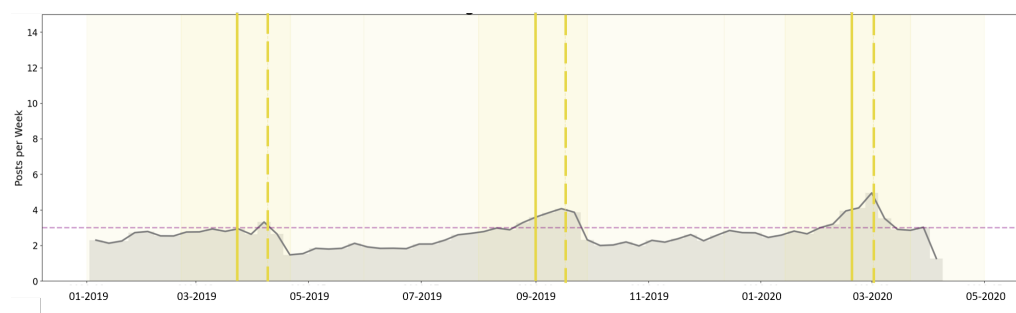


Figure 6. Posts per week of authentic users throughout the election period.

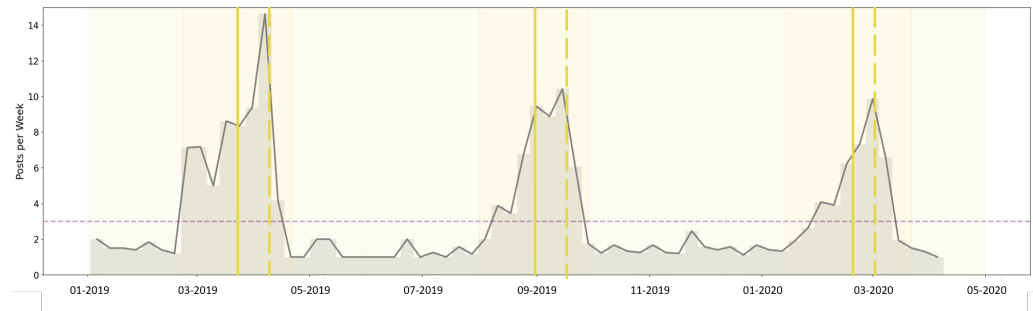


Figure 7. Posts per week of astroturfers throughout the election period.

Similar temporal dynamics were observed for other account activity features as well, leading to the hypothesis that splitting each of these features into time categories would result in a feature set that could capture the unique characteristics of the two profile types. We refer to this as the temporal layer that lies on the activity layer. After splitting the timeline into five periods, pre-election (PR), first half of the campaign (FH), election month (EM), election day (EL), and post-election (PO), statistical analyses for each period showed unique behavior for each of the profiles, as captured by many of the account activity features. Moreover, taking the ratio of an activity feature between two periods in which each of the profile types had its activity peak, namely, between PO and EM (PO/EM), resulted in a feature with high discriminatory power. For example, the PPD (posts per day) feature had a value of 0.7 for authentic profiles and 0.3 for astroturfer profiles. The ADR (active days ratio) had a value of 0.62 for authentic profiles, and 0.31 for astroturfer profiles. All the results were significant (t -test p -value ≤ 0.01).

With this added temporal layer, a total of some 80 features were extracted from the data. The features ranking the highest out of all the features (profile information, profile activity, and temporal features) are shown in Table 4, listed by rank. All account activity features from Table 3 appear in this table, as well as several of the newly devised temporal features. See Table 2 for a description of these features. In the following, we show how this enriched feature set considerably improved the classification results.

7. Applying Machine Learning Algorithms

The discovery of this novel temporal layer was made possible by the unique political circumstances at the time of the dataset creation. Only with consecutive elections held one after another did these features, and their importance, become known to us. As described above, these temporal features formed a third layer of features, atop the familiar categories of profile information and activity features. The results of previous attempts to learn a classification model from an earlier version of the dataset were described in [45]. Here, we show the results of machine learning algorithms on the extended dataset version, with additional fine-tuning of the different parameters.

Six different learning algorithms were applied to the dataset, with stratified cross-validation, using Python's scikit-learn package <https://scikit-learn.org/> (accessed on 25 September 2024): naïve Bayes, logistic regression, support vector machines, k-nearest neighbors, decision trees, and neural networks (using the Adam solver with four hidden layers).

Three different feature sets were used for each round of the learning algorithms. This was to assess the contribution of the temporal layer to the task at hand; namely, distinguishing between astroturfer profiles and authentic profiles relative to the contribution of the activity features per se. We also compared this to the contribution of the profile information features. The configurations tested were:

- I. Activity features alone (without profile information or temporal features);
- II. Activity features with profile information features (without temporal features);
- III. Activity features with temporal features (without profile information features).

We first discuss a comparison of the results for the basic feature set of activity features alone (configuration I), to those with the added profile information (configuration II). The accuracy, precision, recall, and F_1 percentages (for the astroturfer class) are presented in Table 5. For each performance metric column, two scores are provided: the percentage without the profile information, and with the profile information. Since we were dealing with an unbalanced dataset (a ratio of about 20/80 astroturfers to authentic users), we quote the F_1 values for the astroturfer class in the discussion. For most algorithms, these values did not lead to significant improvement using the added profile information (with the exception of the decision tree algorithm: F_1 values of 69% with the profile information versus 64% without it), and some even showed a possible deterioration (for example, logistic regression went down from 65% to 64% and k-nearest neighbors went down from 39% to 38%).

Table 5. Comparison of performance metrics (in percentages) for different algorithms, without (w/o) and with (w) profile information, and the percentage difference between them. Algorithms: NB (naïve Bayes), LR (logistic regression), SVM (support vector machine), KNN (k-nearest neighbors), DT (decision tree), and NN (neural network).

Algorithm	Accuracy	$\Delta\%$	Precision	$\Delta\%$	Recall	$\Delta\%$	F_1	$\Delta\%$
NB w/o	84		64		72		68	
NB w	85	+1.20	72	+12.50	65	−10.77	68	0.00
LR w/o	84		73		58		65	
LR w	88	+4.76	83	+13.70	53	−9.43	64	−1.56
SVM w/o	72		46		60		52	
SVM w	73	+1.39	49	+6.52	59	−1.69	54	+3.85
KNN w/o	76		58		29		39	
KNN w	77	+1.32	56	−3.57	28	−3.57	38	−2.63
DT w/o	86		68		60		64	
DT w	82	−4.88	76	+11.76	63	+5.00	69	+7.81
NN w/o	84		81		63		71	
NN w	90	+7.14	83	+2.47	64	+1.59	72	+1.41

Next, we present a comparison of the results for the set comprising the basic activity features alone (configuration I) to the results for the set with the added temporal layer (configuration III). The accuracy, precision, recall, and F_1 percentages (for the astroturfer class) are presented in Table 6, with the same layout as in Table 5.

Table 6. Comparison of performance metrics (in percentages) for different algorithms, without (w/o) temporal information and with (w), and the percentage of difference between them. Algorithms: NB (naïve Bayes), LR (logistic regression), SVM (support vector machine), KNN (k-nearest neighbors), DT (decision tree), and NN (neural network).

Algorithm	Accuracy	$\Delta\%$	Precision	$\Delta\%$	Recall	$\Delta\%$	F_1	$\Delta\%$
NB w/o	84		64		72		68	
NB w	87	+3.57	79	+23.44	73	+1.39	76	+11.76
LR w/o	84		73		58		65	
LR w	92	+9.52	84	+15.07	71	+22.41	77	+18.46
SVM w/o	72		46		60		52	
SVM w	72	0.00	57	+23.91	62	+3.33	60	+15.38
KNN w/o	76		58		29		39	
KNN w	74	−2.70	54	−7.40	39	+34.48	45	+15.38
DT w/o	86		68		60		64	
DT w	84	−2.38	72	+5.88	72	+20.00	72	+12.50
NN w/o	84		81		63		71	
NN w	91	+8.33	84	+3.70	68	+7.93	75	+5.63

To quantify the differences between the results of the two feature sets (with profile information versus without, in Table 5, and with temporal information versus without, in Table 6), we added a column of $\Delta\%$ for each metric. This value was calculated as the proportion of the difference between the two values, relative to the base value. For example, for the first and second accuracy values of Table 6, for the naive Bayes algorithm, we obtained a $\Delta\%$ value of $(87 - 84)/84 = +3.57$.

The results of this comparison show a clear performance improvement: the addition of temporal information was much more impactful than the addition of profile information for the task at hand. This improved performance was observed for each algorithm when comparing the performance of the algorithm to itself with the different feature sets. However, when comparing the performance of the algorithms to each other, the scores of the KNN and SVM algorithms, even with the enhanced feature set, i.e., using the temporal layer, were still significantly lower than the other algorithms, even when only using the activity-based features (e.g., F_1 values of 45% for KNN and 60% for SVM with the temporal layer, as opposed to 64–71% for the other algorithms with the basic feature set). Moreover, among the good performers, the complexity of the algorithm's model did not appear to be a factor: algorithms based on simple models had F_1 rates similar to those of algorithms based on more complex models (e.g., a neural network with four hidden layers). As mentioned above, two recent studies were conducted on Facebook datasets of comparable size and scope, with a similar aim of classifying fake profiles: Ref. [29], which we consider to be the state of the art, achieved an F_1 score of 71%; and [28], which, under optimal settings, reached the highest accuracy rate, 76.71%, using logistic regression. We also obtained the best performance with logistic regression (closely followed by the neural network). However, our F_1 score was 77%, and our accuracy rate was 92%, which is a significant improvement. While we cannot directly compare our results to theirs, since the datasets were not the same, our results on two different feature sets suggest that the right choice of features significantly boosts performance; specifically, the choice of features generated by temporally parsing the activity features.

Thus, overall, while the accuracy levels of all the tests were not remarkable, the comparison shows a significant improvement when applying the temporal layer. What remains is to choose the algorithm that is best suited to leverage this choice of feature set.

Now that we had a good classifier, we ran it on the complete dataset, i.e., half a million profiles, and found 3.01% classified as astroturfers. While this may seem to be a low value, our analysis showed that these astroturfers were responsible for 23% of all the comments. This huge volume underscores the potential influence of the activity of even a small group of astroturfers, and the danger it poses to democratic processes.

We conclude this section with a remark on the computational complexity of the classification process. Before applying any of the machine learning algorithms, the features must first be extracted, specifically, the temporal layer which we introduced here. This involves a basic statistical analysis for each profile to compute the averages for the relevant data. We do not view this as introducing significant overhead, but rather as being a step to ensure that the resulting method remains computationally comparable to other state-of-the-art techniques.

8. Conclusions and Future Research

The framework presented here forms a complete workflow for identifying astroturfers on Facebook. We make three novel contributions relating to the data repository, the labeling technique, and the temporal layer for the feature extraction stage. We briefly summarize each, and then, suggest some directions for future research.

Starting at the data collection stage, we described how a huge collection of profiles and comments was extracted from the public profiles of political candidates. Given the unique circumstances under which the data were collected, this corpus is unique not only in its size but also in its rich content. This valuable repository is now available for the use of the research community.

For the labeling stage, a novel approach was presented to increase the proportion of labeled astroturfers, without increasing the amount of manual labor involved. This technique is useful for enabling machine learning algorithms to be applied to a larger dataset that better represents the population from which it was sampled, thus improving the results of the algorithms.

Yet another innovation is our approach to the feature extraction stage. We devised a temporal layer of features that capture the changes in a profile's activity over the timeline of the elections. While these features came to light due to the contiguity of several rounds of elections, many are likely to be useful for stand-alone elections, which is the normative case in politics.

As our results show, the dataset we collected and the framework we introduced make valuable contributions to the task of distinguishing between authentic profiles and astroturfers. They can be used to gain insights into astroturfing and the characteristics of astroturfers. For example, we identified three main categories of behavior: the majority (some 72%) of the profiles identified as astroturfers are very active on public pages, writing dozens of comments a day, participating in discussions, and replying to other comments. A smaller set (some 18%) of astroturfers posts repetitive comments that promote specific ideas. In many cases (for 70% of astroturfers of this type) the comments are not directly related to the conversation but rather are repetitious slogans. An even smaller set (some 10%) of astroturfers just creates volume, without any actual content (such as "Bibi the king"). The activity of a large set of users consisted of only one or two very short comments of this type. While they do not deal directly with influencing public opinion, a significant volume of these comments may create this impact. In many cases, it is extremely difficult to decide whether these are authentic users or astroturfers (many of them have private profiles with a limited amount of information), which still leaves part of the problem open for future research.

While Facebook applies a proactive methodology of banning promoters of "fake news", in the case of astroturfers it is more difficult to apply these methods, since the content produced is similar to comments and reactions expressed by authentic users. However, revealing some of the user statistics to the public might empower people to make more enlightened decisions regarding how they relate to other users on social media, and in treating such users with appropriate caution.

There are many further research directions the dataset and the framework presented here point to. We present several suggestions which we consider to be the natural continuation of our work. First of all, it is important to note that our features did not include the content of the posts. Obviously, there is much information in the content which could also be leveraged for the task. From a wider perspective, devising a metric for gauging the contribution of astroturfers to the dissemination of hate speech on social media would be useful and is imperative when assessing any strategy for combating this phenomenon. Being able to quantify the influence that astroturfers have on public opinion is critical when devising a policy intended to reduce it. However, as great as our success might be in reducing the influence of astroturfers, we probably will not be able to eliminate astroturfing entirely. Therefore, being able to clean a dataset, for example, a dataset of responses to a survey, by identifying the authentic users' answers from the non-authentic ones, would be vital to drawing correct conclusions from the data. Within this type of dataset, not all the temporal features introduced in this study will be present. This is one of the limitations of our method. For such cases, the concept of a layer of novel features, atop the basic layer, can be extrapolated from the current research and applied appropriately. Hopefully, the work presented here, along with the original dataset on which it was performed, will lead to further insights in advancing ways to curb the threat of astroturfing.

Author Contributions: Framework, E.B.-D.; methodology, J.S.; software, J.S.; analysis, E.B.-D.; investigation, E.B.-D.; data curation, J.S.; writing, E.B.-D.; review, J.S. and E.B.-D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Only information voluntarily made public by the users was used in this research. No personal identifiable information was collected nor stored.

Data Availability Statement: The dataset described in this study, has been made publicly available at <https://github.com/jschler/ILElectionData> (accessed on 25 September 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Howard, P.N. *New Media Campaigns and the Managed Citizen*; Cambridge University Press: New York, NY, USA, 2005; pp. 93–144.
- Kovic, M.; Rauchfleisch, A.; Sele, M.; Caspar, C. Digital Astroturfing in Politics: Definition, Typology, and Countermeasures. *Stud. Commun. Sci.* **2018**, *18*, 69–85. [\[CrossRef\]](#)
- Bradshaw, S.; Howard, P. Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. In *Computational Propaganda Research Project*; Oxford Internet Institute: Oxford, UK, 2017.
- Mahbub, S.; Pardede, E.; Kayes, A.; Rahayu, W. Controlling astroturfing on the internet: A survey on detection techniques and research challenges. *Int. J. Web Grid Serv.* **2019**, *15*, 139–158. [\[CrossRef\]](#)
- Schoch, D.; Keller, F.B.; Stier, S.; Yang, J. Coordination patterns reveal online political astroturfing across the world. *Sci. Rep.* **2022**, *12*, 4572. [\[CrossRef\]](#)
- Shukla, R.; Sinha, A.; Chaudhary, A. TweepBot: An AI-driven online media bot identification algorithm for Twitter social networks. *Electronics* **2022**, *11*, 743. [\[CrossRef\]](#)
- Hegelich, S.; Janetzko, D. Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet. In Proceedings of the Tenth AAAI Conference on Web and Social Media, ICWSM, Cologne, Germany, 17–20 May 2016; pp. 579–582.
- Forelle, M.; Howard, P.N.; Monroy-Hernandez, A.; Savage, S. Political Bots and the Manipulation of Public Opinion in Venezuela. *SSRN Electron. J.* **2015**, 1–8. [\[CrossRef\]](#)
- Brachten, F.; Stieglitz, S.; Hofeditz, L.; Kloppenborg, K.; Reimann, A. Strategies and Influence of Social Bots in a 2017 German State Election—A Case Study on Twitter. *arXiv* **2017**, arXiv:1710.07562.
- Bessi, A.; Ferrara, E. Social Bots Distort the 2016 US Presidential Election Online Discussion. *First Monday* **2016**, *21*. [\[CrossRef\]](#)
- Chen, T.; Liu, J.; Wu, Y.; Tian, Y.; Tong, E.; Niu, W.; Li, Y.; Xiang, Y.; Wang, W. Survey on Astroturfing Detection and Analysis from an Information Technology Perspective. *Secur. Commun. Netw.* **2021**, *2021*, 3294610. [\[CrossRef\]](#)
- Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Flammini, A.; Menczer, F. Detecting and tracking political abuse in social media. In Proceedings of the International AAAI Conference on Web and Social Media, Barcelona, Spain, 17–21 July 2011; Volume 5.
- Chen, C.; Wu, K.; Srinivasan, V.; Zhang, X. Battling the internet water army: Detection of hidden paid posters. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), Niagara Falls, ON, Canada, 25–28 August 2013; pp. 116–120.
- Peng, J.; Detchon, S.; Choo, K.K.R.; Ashman, H. Astroturfing detection in social media: A binary n-gram-based approach. *Concurr. Comput. Pract. Exp.* **2017**, *29*, e4013. [\[CrossRef\]](#)
- Beğenilmiş, E.; Uskudarli, S. Supervised learning methods in classifying organized behavior in tweet collections. *Int. J. Artif. Intell. Tools* **2019**, *28*, 1960001. [\[CrossRef\]](#)
- Chavoshi, N.; Mueen, A. Model bots, not humans on social media. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018; pp. 178–185.
- Waite, C.; Bourke, L. Using the cyborg to re-think young people’s uses of Facebook. *J. Sociol.* **2015**, *51*, 537–552. [\[CrossRef\]](#)
- Yang, Z.; Wilson, C.; Wang, X.; Gao, T.; Zhao, B.Y.; Dai, Y. Uncovering social network sybils in the wild. *ACM Trans. Knowl. Discov. Data (TKDD)* **2014**, *8*, 1–29. [\[CrossRef\]](#)
- Chu, Z.; Gianvecchio, S.; Wang, H.; Jajodia, S. Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg? *IEEE Trans. Dependable Secur. Comput.* **2012**, *9*, 811–824. [\[CrossRef\]](#)
- Keller, F.B.; Schoch, D.; Stier, S.; Yang, J. Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Commun.* **2020**, *37*, 256–280. [\[CrossRef\]](#)
- Moon, B. Identifying Bots in the Australian Twittersphere. In Proceedings of the 8th International Conference on Social Media & Society, Toronto, ON, Canada, 28–30 July 2017; pp. 49:1–49:5.
- Patxi, G.G.; Puerta, J.G.d.l.; Gómez, C.L.; Santos, I.; Bringas, P.G. Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying. *Log. J. IGPL* **2016**, *24*, 42–53.
- Ellaky, Z.; Benabbou, F. Political social media bot detection: Unveiling cutting-edge feature selection and engineering strategies in machine learning model development. *Sci. Afr.* **2024**, *25*, e02269. [\[CrossRef\]](#)
- Rahman, M.S.; Huang, T.K.; Madhyastha, H.V.; Faloutsos, M. Efficient and Scalable Socware Detection in Online Social Networks. In Proceedings of the 21st {USENIX} Security Symposium ({USENIX} Security 12), Bellevue, WA, USA, 8–10 August 2012; pp. 663–678.

25. Ahmed, F.; Abulaish, M. An MCL-Based Approach for Spam Profile Detection in Online Social Networks. In Proceedings of the 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications, Liverpool, UK, 25–27 June 2012; pp. 602–608.
26. Stringhini, G.; Kruegel, C.; Vigna, G. Detecting Spammers on Social Networks. In Proceedings of the 26th Annual Computer Security Applications Conference, Austin, TX, USA, 6–10 December 2010; pp. 1–9.
27. Gao, H.; Hu, J.; Wilson, C.; Li, Z.; Chen, Y.; Zhao, B.Y. Detecting and Characterizing Social Spam Campaigns. In Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, Melbourne, Australia, 1–3 November 2010; pp. 35–47.
28. Dewan, P.; Bagroy, S.; Kumaraguru, P. Hiding in Plain Sight: The Anatomy of Malicious Pages on Facebook. In *Social Network Based Big Data Analysis and Applications*; Kaya, M., Kawash, J., Khoury, S., Day, M.Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 21–54.
29. Santia, G.C.; Mujib, M.I.; Williams, J.R. Detecting Social Bots on Facebook in an Information Veracity Context. In Proceedings of the International AAAI Conference on Web and Social Media, Munich, Germany, 11–14 June 2019; Volume 13, pp. 463–472.
30. Feng, S.; Wan, H.; Wang, N.; Li, J.; Luo, M. Twibot-20: A comprehensive twitter bot detection benchmark. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Online, 1–5 November 2021; pp. 4485–4494.
31. Fong, S.; Zhuang, Y.; He, J. Not every Friend on a Social Network can be Trusted: Classifying Imposters Using Decision Trees. In Proceedings of the First International Conference on Future Generation Communication Technologies, London, UK, 12–14 December 2012; pp. 58–63.
32. Clark, E.M.; Williams, J.R.; Jones, C.A.; Galbraith, R.A.; Danforth, C.M.; Dodds, P.S. Sifting robotic from organic text: A natural language approach for detecting automation on Twitter. *J. Comput. Sci.* **2016**, *16*, 1–7. [\[CrossRef\]](#)
33. Kyumin, L.; Eoff, B.D.; Caverlee, J. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In Proceedings of the International AAAI Conference on Web and Social Media, Barcelona, Spain, 17–21 July 2011.
34. Alqudah, R.; Al-Qaisi, M.; Ammari, R.; Ta'a, Y.A. OSINT-Based Tool for Social Media User Impersonation Detection Through Machine Learning. In Proceedings of the 2023 International Conference on Information Technology (ICIT), Orlando, FL, USA, 4–6 April 2023; pp. 752–757. [\[CrossRef\]](#)
35. Davis, C.A.; Varol, O.; Ferrara, E.; Flammini, A.; Menczer, F. BotOrNot: A System to Evaluate Social Bots. In Proceedings of the 25th International Conference Companion on World Wide Web, Montreal, QC, Canada, 11–15 May 2016; pp. 273–274.
36. Nguyen, M.; Aktas, M.; Akbas, E. Bot detection on social networks using persistent homology. *Math. Comput. Appl.* **2020**, *25*, 58. [\[CrossRef\]](#)
37. Conti, M.; Poovendran, R.; Secchiero, M. Fakebook: Detecting Fake Profiles in On-Line Social Networks. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey, 26–29 August 2012; pp. 1071–1078.
38. Mezhuvev, V.; Sadat, S.N.; Rahman, M.A.; Refat, N.; Asyhari, A.T. Evaluation of the likelihood of friend request acceptance in online social networks. *IEEE Access* **2019**, *7*, 75318–75329. [\[CrossRef\]](#)
39. Chen, K.; Cheng, L.C.; Ye, M.Y.; Wang, J.H. A Graph Neural Network Approach to Detect Original Review Spammers of Astroturfing Campaigns. *Electron. Commer. Res. Appl.* **2023**, *62*, 101326. [\[CrossRef\]](#)
40. Raja, M.; Raj, L.A. Fake news detection on social networks using Machine learning techniques. *Mater. Today Proc.* **2022**, *62*, 4821–4827. [\[CrossRef\]](#)
41. Sharmin, S. Interaction Based Credibility Analysis of News on Facebook Using Machine Learning Methodologies. In Proceedings of the 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Dijon, France, 19–21 October 2022. [\[CrossRef\]](#)
42. Nikulchev, E.; Gusev, A.; Gazanova, N.; Magomedov, S.; Alexeenko, A.; Malykh, A.; Kolyasnikov, P.; Malykh, S. Engagement assessment for the educational web-service based on largest Lyapunov exponent calculation for user reaction time series. *Educ. Sci.* **2023**, *13*, 141. [\[CrossRef\]](#)
43. Chu, Z.; Gianvecchio, S.; Wang, H. Bot or Human? A Behavior-Based Online Bot Detection System. In *From Database to Cyber Security: Essays Dedicated to Sushil Jajodia on the Occasion of His 70th Birthday*; Samarati, P., Ray, I., Ray, I., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 432–449. [\[CrossRef\]](#)
44. Schler, J.; Bonchek-Dokow, E.; Vainstein, T.; Gotam, M.; Teplitsky, M. Utilizing Natural Honeypots for Efficiently Labeling Astroturfer Profiles. In Proceedings of the 22nd International Conference on Knowledge Engineering and Knowledge Management (Posters & Demos), 2020, Bozen-Bolzano, Italy, 16–18 September 2020; pp. 41–45.
45. Schler, J.; Bonchek-Dokow, E.; Vainstein, T.; Gotam, M.; Teplitsky, M. Profiling Astroturfing Facebook Users during Three Contiguous Israeli Election Periods. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 4331–4340.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.