IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 26, 2019, accepted August 19, 2019, date of publication August 29, 2019, date of current version September 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2938389

Not All Lies Are Equal. A Study Into the Engineering of Political Misinformation in the 2016 US Presidential Election

AXEL OEHMICHEN[1], KEVIN HUA[1], JULIO AMADOR DÍAZ LÓPEZ[1], MIGUEL MOLINA-SOLANA[1], JUAN GÓMEZ-ROMERO[2], AND YI-KE GUO[1]

[1]Data Science Institute, Imperial College London, London SW7 2AZ, U.K.
[2]Department of Computer Science and AI, Universidad de Granada, 18010 Granada, Spain

Corresponding author: Miguel Molina-Solana (mmolinas@ic.ac.uk)

The work of M. Molina-Solana was supported by the European Commission under Grant 743623. The work of J. Amador Díaz López was supported by the Imperial College Research Fellowship. The work of J. Gómez-Romero was supported by the Universidad de Granada under Grant P9-2014-ING and in part by the Spanish Ministry of Education, Culture and Sport under the José Castillejo Research Stays Programme.

ABSTRACT We investigated whether and how political misinformation is engineered using a dataset of four months worth of tweets related to the 2016 presidential election in the United States. The data contained tweets that achieved a significant level of exposure and was manually labelled into misinformation and regular information. We found that misinformation was produced by accounts that exhibit different characteristics and behaviour from regular accounts. Moreover, the content of misinformation is more novel, polarised and appears to change through coordination. Our findings suggest that engineering of political misinformation seems to exploit human traits such as reciprocity and confirmation bias. We argue that investigating how misinformation is created is essential to understand human biases, diffusion and ultimately better produce public policy.

INDEX TERMS Misinformation, data science, US elections, politics.

I. INTRODUCTION

Even if the term *fake news* reached the mainstream in the 2016 electoral campaign in the United States, this phenomenon —and mainly, worries about how agents might strategically act to influence people's beliefs and perceptions— appears every time a new technological breakthrough in communications emerges [2], [33]. For instance, the introduction of cheap printing presses and advertising business models allowed newspapers to increase their reach dramatically [24]. Partisans, ideologues and some ill-intentioned 'entrepreneurs' were among the beneficiaries that, by adopting such innovations and strategically promoting sensational stories, managed to increase sales.

In later years, the appearance of innovations such as the radio, TV and the internet, have unintentionally incentivised similar behaviours. The unintended effects of these technologies have been so significant that each of them has brought regulation aimed at avoiding deception and minimising the influence on public opinion. For example, radio and TV brought with them regulations on deceptive advertising and political campaigning in the United States [45] and just recently, the so-called *fake news* campaigns in social media have sparked discussions about reining *fake news* in countries like France [16] and the United Kingdom [43].

The recent explosion in usage of social networking sites to obtain pecuniary gains from sensationalist stories [2] or strategically influence political campaigns [5], [20], [38] has highlighted the need to better comprehend how misinformation is created, how it diffuses in social media [46] and how it can be spotted [14], [31], [36], [48]. Most efforts to understand the phenomena [1], [2], [6]–[8], [15], [32]–[34], [42], and develop solutions [11], [14], [31], [48] suggest human biases may foster the diffusion of misinformation. However, and despite having identified perverse incentives in the diffusion of information in other forms of electronic media, little attention has been paid to how misinformation is generated and seeded into online social networks. In this paper, we aim at closing this bridge by providing evidence of the engineering process behind political misinformation

The associate editor coordinating the review of this manuscript and approving it for publication was Wenge Rong.

and suggesting explanations into why such engineering may work. We argue that understanding how misinformation is created is necessary to explain observed differences in the diffusion of misinformation,[1] better understanding human biases and generating public policy.

To perform this study, we used a dataset containing tweets collected during four months just after the 2016 presidential election in the United States [3]. This dataset includes manually labelled tweets[2] that got re-tweeted more than 1000 times. This dataset was chosen for different reasons. First, it is now widely accepted that foreign agents strategically acted to influence public opinion [44]. If deceivers took strategic actions, we would expect to be able to identify them within this dataset. Second, [46] points out that tweets containing true information usually never diffused to more than a 1000 people. By using tweets that got retweeted at least 1000 times, we can single-out characteristics of tweets containing misinformation, given the tweet achieved a substantial level of diffusion. Also, such tweets may be most relevant from the public policy perspective (i.e. it is highly unlikely that policy-makers are both interested and capable in regulating misinformation that does not alter the general public opinion). Finally, the sets of annotations used within the dataset encompass deceptive information, rumours, subjective information and false information within the label of *misinformation*. The variety of annotations allowed us to extend and corroborate findings of research done in the context of false information only.

In summary, we begin by assuming individuals participate in social media to achieve a specific goal.[3] In the case of political misinformation, such goals may include pecuniary gains or influencing public opinion [2] though, in this work, we abstract away from the goal deceivers may pursue, and look to confirm the following:

> *Deceivers strategically engineer their social media posts.*

Therefore, in the rest of the paper, we will put forward evidence confirming this claim by showing that tweets containing misinformation, and the accounts spreading them, have significantly different features compared to those not containing them.

We present three sets of results. The first set shows differences between features of accounts spreading misinformation and other types of accounts. We find that, on average, the former has fewer followers but follow more other accounts, have fewer status updates and were created more recently, and tend to favourite more the content of the others. In this way, we show that accounts sharing misinformation significantly differ from others and suggest this behaviour appears to be intentionally designed to elicit reciprocity. The second set of results presents differences within the textual field of the tweet. We show statistical differences in syntactic style; i.e., misinformation usually contains more exclamations, capitalisation and digits. Furthermore, we present sentiment analysis showing differences in the usage of sentiment both, through time and in general. We argue that such differences in sentiment are intentional and aimed at exploiting human psychological biases. Finally, we support this assertion (i.e., that textual features are engineered to exploit psychological biases) by showing that tweets containing misinformation were more favourited than those containing regular information.

Together, our results seem to confirm that the engineering of tweets bolstered the diffusion of political misinformation within the 2016 presidential election in the United States and that such engineering was geared to exploit human biases. As such, we suggest such engineering might be one of the causes political misinformation spreads faster and more broadly than other types of misinformation. Nevertheless, even if our results coincide with the more extensive study done by [46], we acknowledge the limitations our data may impose on our claim. Therefore, our main contribution is to underscore the need to study how such misinformation is generated. Realising that the effects of misinformation cannot be explained by human biases alone allows us to understand such biases, differences in diffusion processes better, and, ultimately, generate public policy.

Particularly, our study is related to the literature on deception and misinformation in social science [1], [2], [6]–[8], [15], [32]–[34], [42] and automatic deception detection. To the former, we build from the findings suggesting deceivers may want to strategically act to convey their information [11], [14], [31], [48]. Our main contribution to this literature is to extend the finding of strategic action to convey misinformation at the individual level to the collective level. This finding complements research and evidence on influence campaigns using bots and foreign agents by suggesting their behaviour should be studied systematically [21].

The following section describes methods and data, and Section III presents the results of our experiments, with a discussion on them following. The work concludes by summarising the main findings and presenting future lines of work.

II. MATERIALS AND METHODS

Our goal is validating the commonly-assumed hypothesis that misinformation is carefully crafted and engineered by their authors. In particular, we are interested in identifying some of the means by which that engineering process took place, and as such, those features that are significantly different from usual patterns.

A. DATA

We rely on an annotated dataset that we published previously [3], indicating whether a tweet contains a piece of misinformation or not. To the best of our knowledge, this was still

[1] In a comprehensive study, [46] find that falsehood travels faster, deeper and more broadly than truth; and, in particular, false news related to politics, travel faster and deeper than those related to other topics.

[2] Tweets were labelled following the categories established by Conroy et al. [12]

[3] A survey detailing political participation online can be found in [35]

the only publicly available labelled dataset about misinformation on Twitter (in contrast to others containing stories or news sources) at the time of writing these lines. This dataset enabled us to perform analyses looking for statistically significant differences in the features between tweets containing misinformation and regular information that reached more than 1000 retweets during the collection window. Though we acknowledge that this binary labelling has arguably its limitations, for our particular piece of research, the benefits of having a simpler and straightforward categorisation outperforms the drawbacks. Netflix recently took a similar approach with regards to its recommender [18].

For completeness, the rest of this subsection will briefly describe the dataset and its creation process. The dataset was compiled by collecting publicly available tweets related to the presidential election in the United States in 2016 using Twitter's public streaming API. For each tweet collected, Twitter's API provided several features (besides the tweet text), including the number of retweets, favourites, media, and URLs. We collected the tweets following Twitter's API terms and conditions.

There is nowadays a hot debate on ethical issues around collection, analysis and publication of social media data, with no consensus on what constitute good practice regarding its two main ethical issues: informed consent and minimization of harm [47]. For this research, we have aligned ourselves with the common practice in academia of assuming that users provide informed consent (for their tweets to be collected and analysed) by their acceptance of Twitter's terms of service. We can also confirm that, in an effort to minimize harm on individual users, we have not enriched the dataset with external sensitive data on the users, nor have we applied any algorithm to derive sensitive data. We did not pre-process the dataset to highlight particular tweets or users. What is more, in this paper we only refer to summarized results (i.e. we are not analysing/publishing any individual tweets).

An essential feature within Twitter is the ability to share someone's tweet through 'retweets'. This functionality enables users to pass forward to their followers an exact copy of someone else's tweet. There are many reasons why users might decide to retweet; e.g. to spread information to new audiences, to show one's role as a listener, or to agree with or validate someone else's point of view [9].

The sample (57,379,672 tweets ranging from November 2016 until March 2017) was collected using the following search terms and user handles: #MyVote2016, #ElectionDay, #electionnight, @realDonaldTrump and @HillaryClinton. This number includes original tweets and retweets. From them, only the tweets that have more than 1000 retweets were extracted, resulting in a total of 9001 tweets. It is relevant to note that a portion of those tweets is no longer publicly available through Twitter API: several authors have deleted tweets and some accounts have been closed for infringing Twitter's policy, resulting in tweets no longer being available.

TABLE 1. Agreement between the labelling performed by the two teams in the used dataset.

		Second team		
		misinformation	**regular**	**unsure**
First team	**regular**	6482	1444	330
	misinformation	213	133	7
	unsure	250	98	44

Should any researcher be interested in replicating the results we describe here, the dataset we used for our experimentation (a snapshot at the time of collection) is publicly available [3]. However, we encourage them to re-query Twitter's API with the tweets' IDs listed in the dataset in order to obtain an updated snapshot of the dataset. Due to deleted tweets not being returned, results are likely to slightly vary from those reported here.

Two teams of individuals labelled the tweets in the dataset (by manually inspecting the text field) as *misinformation* if its text could be considered within any of the categories described in [37], and as a *regular tweet* (i.e. tweet not containing misinformation) otherwise. The first team was composed of 6 individuals that self-reported little knowledge on US politics, while the second manually crosschecked every tweet with factual data. All individuals were over 18 and received no compensation.

We compared the annotations of the two teams and found out that they disagree on roughly a quarter of the tweets (see Table 1). A close inspection to those by a third team consistently found that annotations by the second team were to be considered more accurate. To ensure the highest possible standard, for the present study, we focus solely on the labelling of the second team. This data contains 1675 tweets (that were retweeted more than 1000 times) labelled as misinformation —18.6% of the 9001 in the dataset— and left aside their 381 unknowns (4.23%). After reviewing the study, Imperial College London considered it was exempt from further ethical review.

B. METHODOLOGY

We looked for statistically significant differences in the distributions of features through a Kolmogorov-Smirnov test [27] between the set of tweets that achieved 1000 retweets or more and contain misinformation, and those that achieved 1000 retweets or more and contain regular information. The null hypothesis was *h0: The two samples come from the same distribution* against the alternative hypothesis *h1: The two samples come from different distributions*. Kolmogorov-Smirnov test is particularly suited in this situation as it allows us to compare whether or not two samples come from the same arbitrary (i.e. not assuming normality) probability distribution.

We consider that a difference between sets is statistically significant if the p-value is lower than 1%. For the continuous variables with extreme values, we did the test on the (decimal) logarithm in order to have a more representative scale. For others (e.g. *num_hashtags*, *num_mentions*, *num_media*, *num_urls*), we compute the number of items per tweet. p-values smaller than 1e-16 are reported in all tables as 0.00.

Implementation-wise, we used R (v3.4) and Python 3 as programming languages, and the eTRIKS Analytical Environment (eAE) [30] as the analytical platform. The eAE is a flexible multi-tenant computational framework which enables the analysis of data at scale, enabling researchers to explore and analyse data concurrently and in the language of their choice.

III. RESULTS

In order to provide a comprehensive study on all relevant aspects related to crafting and engineering of misinformation, we studied several features of the tweets from three different angles: 1) information about the account, 2) analysis of the text of the tweet and 3) information about the tweet.

The primary source of data features was Twitter's API. From it, we obtained some of them and derived many others. The rest of this section focuses on checking whether or not the value distributions of the features of the tweets differ when referred to misinformation or not. We assume that those differences are due to a conscious engineering process. To do so, we compute statistically significant differences in their value distributions. Results of this analysis are shown both numerically and visually.

A. FEATURES OF THE ACCOUNT

Literature [39] has shown that particular accounts are more prone to generate misinformation. The US government has recently disclosed a list of Twitter accounts that were known to spread misinformation consistently [19].

Under this umbrella, we have studied features about the account (such as its number of followers, whether or not it is verified, whether it has an image in its profile, etc.) and features about the syntactical analysis of the name and description of the account. Results are shown respectively in Tables 2 and 3. Following ethical guidelines, note that we did not derive any sensitive data.

When contrasting accounts spreading misinformation to others, results in Table 2 show that the former has fewer followers but follow more other accounts, have more infrequent status updates and were created more recently, and tend to favourite more the content of others. Results reported in Table 3 lead us to conclude that individuals with higher numbers of special characters and capitals in their user name are more likely to tweet misinformation. These results are consistent with the observation that these accounts are short-lived and created programmatically, and therefore aimed to avoid name collision with existing ones.

TABLE 2. Analysis of features related to the account generating the tweet. The results (p-value and t-stat) come from the Kolmogorov-Smirnov test [27] on the distributions between the misinformation and the other tweets. Rows are ordered by p-value. Variables above the line are those whose differences are considered statistically significant (p-value smaller than 0.01).

	p-value	t-stat
user.followers_count	0.00	0.129
user.listed_count	0.00	0.121
user.verified	0.00	0.209
user.favourites_count	1.65e-12	0.103
user.friends_count	5.07e-11	0.094
user.statuses_count	2.32e-08	0.081
user.default_profile	4.96e-01	0.022
user.default_profile_image	1.00e+00	0.001
user.profile_use_bg_image	1.00e+00	0.009

TABLE 3. Features about the account generating the tweet (related to text analysis). Again, rows are ordered by statistical significance; significant variables are above the line. It is interesting to see that those are mostly the ones associated with spelling used by bots (randomly generated to avoid collisions).

	p-value	t-stat
user.description.caps	1.50e-04	0.059
user.name.caps	8.27e-04	0.053
user.name.weird_char	2.95e-03	0.049
user.screen_name.caps	3.64e-03	0.048
user.screen_name.weird_char	2.18e-01	0.028
user.description.exclam	2.53e-01	0.027
user.screen_name.digits	4.44e-01	0.023
user.description.digits	6.52e-01	0.020
user.screen_name.underscres	8.02e-01	0.017
user.description.nonstandard	8.14e-01	0.017
user.name.digits	1.00e+00	0.006
user.name.underscores	1.00e+00	0.000

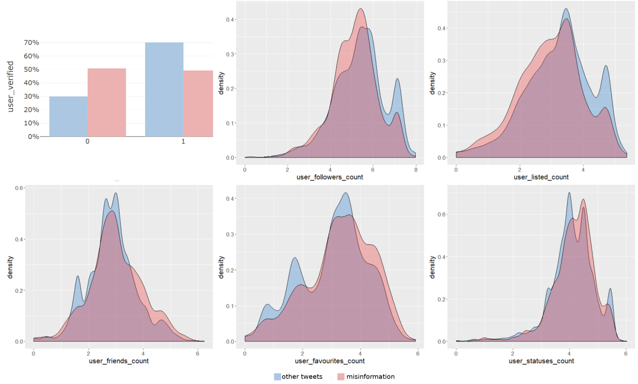


FIGURE 1. Density distribution of the variables from Table 2 that are statistically significant. The test for the proportion of verified account confirms an expected fact: the proportion of verified account is much weaker for tweets containing misinformation than for other tweets, suggesting that misinformation tends to be created by more 'anonymous' people.

However, it should be noted that a statistically significant difference does not mean that such a difference is substantial or ultimately meaningful. So, in order to better assess meaningful differences, we are also plotting the density distributions. Fig 1 visually displays the density distributions of the features that do have statistically significant differences in their distributions.

B. FEATURES OF THE TEXT OF THE TWEET

A second angle to examine misinformation is by looking at the text within the tweet and analysing, both, its formal components (syntax), and its meaning (sentiment).

TABLE 4. Features about the text of tweets. Again, rows are ordered by statistical significance; significant variables are above the line.

	p-value	t-stat
tweet.text.digits	8.02e-09	0.084
tweet.text.caps	1.41e-08	0.082
tweet.text.exclamations	4.40e-03	0.047
tweet.text.nonstandard	1.00e+00	0.008

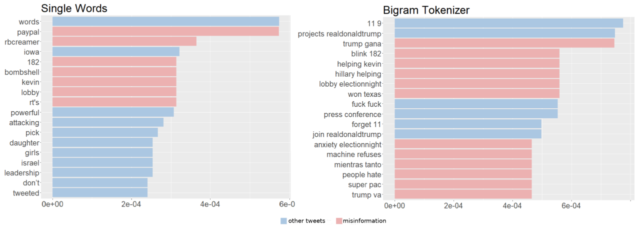


FIGURE 2. Most recurrent words in the tweets (single and bigram).

1) SYNTAX

In this instance, we extracted several textual features (percentage of capital letters, digits and special characters) using regular expressions. We aim to analyse if any of those have a different probability distribution and thus confirm that an engineering process has indeed taken place. Table 4 confirms that a tweet with a high proportion of capital letters, digits and exclamation marks has more chances to be misinformation.

Along with the textual features, we analysed word frequencies (as single words and bi-grams). The most recurrent ones are listed in Fig 2 together with their frequencies; discrepancies between distributions are visible. However, it is important to note that the actual list of words and values are closely related to the underlying data and context.

2) SENTIMENT ANALYSIS OF THE TWEET TEXT

Finally, we performed sentiment analysis on the content of the tweets. Sentiment analysis aims at identifying, extracting, quantifying and studying affective states and subjective information. Often, sentiment analysis is used to determine the attitude of a speaker [41]; in our case, the author of a tweet. When human readers approach a text, they use their understanding of the emotional intent of words to infer whether a section of text is positive, negative or neutral, or perhaps characterised by some more nuanced emotion like surprise or disgust. Text mining tools are available to extract the emotional content of text programmatically [40], and they usually boil down to two main approaches [25]: lexical approach and machine learning approach.

Our first step was to analyse the sentiments of the whole text field —including hashtags— of tweets in our dataset using the National Research Council Canada (NRC) lexicon [29]. The NRC lexicon provides a dictionary that scores every word within the lexicon according to their emotional traits. With the NRC lexicon, we counted the scores associated to each emotion for each word within each tweet and aggregated such scores to provide a proxy for the emotions reflected within each tweet. Finally, we aggregated the counts for all of the emotions overall and through time. Fig 3 points out that misinformation generally exhibits less joy, trust and positive emotions but more surprise. Looking at the temporal evolution of the emotions (see Fig 4), we noticed that fluctuation of emotions is higher tweets with misinformation, while trust and positive emotions dominated in tweets not containing misinformation.

A sentiment score can also be computed using Deep Learning techniques [49] and sometimes perform better than the lexicon approach. The reason for this is that Deep Learning Techniques can capture subtleties that the lexicon approach cannot. In our case, we trained a model to classify tweets as being positive or negative using the *sentiment140* dataset (firstly described in [17]). Specifically, we used two widely-used word embeddings: *word2vec* from Google [28] and *fasttext* from Facebook [23] which were fed into an LSTM with 128 hidden units followed by a dense layer using sigmoid activation to calculate a positive or

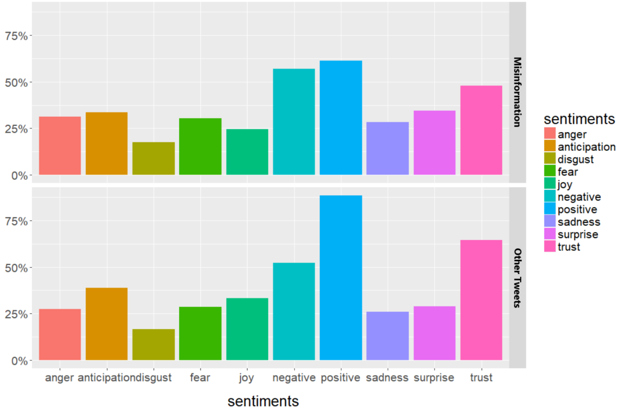


FIGURE 3. Comparison between the different core sentiments between tweets containing misinformation (top) and other tweets (bottom), following the Lexicon approach. Particularly relevant are the bars related with negative and positive sentiments, as they show that misinformation tends to have a less positive sentiment. It can also be observed that most of the overall contribution to positiveness comes from 'trust'.

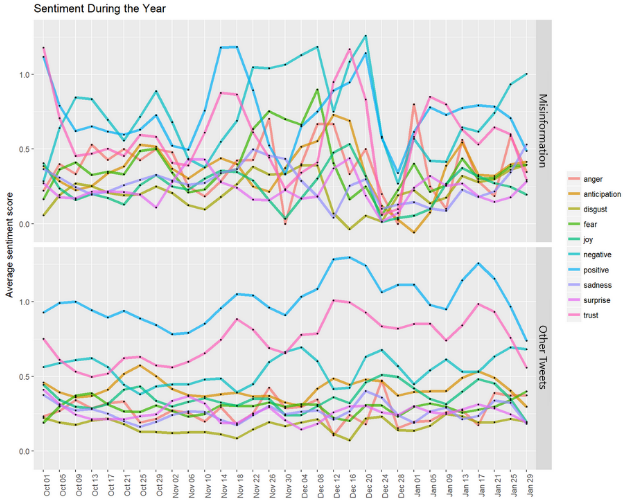


FIGURE 4. Evolution of the different core sentiments (with the Lexicon approach) over the course of the four months, between tweets containing misinformation (top) and other tweets (bottom).

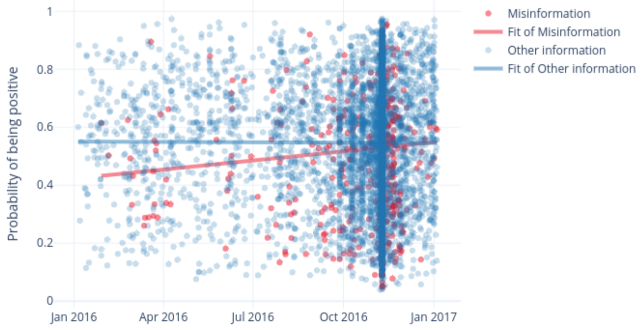


FIGURE 5. Difference of the evolution of the sentiment computed by a network using *fasttext* embeddings between tweets containing misinformation and other tweets. Each point represents a tweet in the timeline of our dataset and the probability of the tweet for being positive. The red line represents the mean of sentiment for misinformation whereas the purple line represents the mean for other types of information.

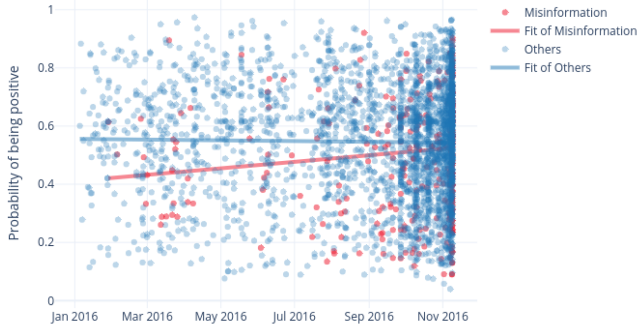


FIGURE 6. Similar to Fig 5 but adjusted only until the day of the election.

negative sentiment. We chose embeddings of size 300, the network was trained using RMSprop for 50 epochs and hyperparameters were tuned using the development set.

Fig 5 shows the overall probability of a tweet being positive using the *fasttext* embedding and the network described before. The average probability over the four whole months is more positive than negative, regardless of it being misinformation or not. However, the trend for misinformation is lower than for the other tweets, which is coherent with previous results stating that tweets containing misinformation tend to be more negative (Fig 3 and 4). The wider spread is explained by the scarcity of the misinformation at some periods (and also due to the dataset being imbalanced) and does not constitute a significant indicator.

C. FEATURES OF A TWEET

By feature of a tweet we refer specifically to the numbers of 1) hashtags, 2) mentions, 3) URLs, and 4) media elements in the text of the tweet; and to numbers of 1) retweets and 2) favourites achieved by the tweet the last time seen in the dataset. Table 5 lists all these features (a detailed description of each one can be found at Twitter's API website) together with the results from the Kolmogorov-Smirnov test. There are four variables in which the differences in distributions are different in terms of statistical significance (those above the horizontal line).

TABLE 5. Analysis of features related to the tweet. The results (p-value and t-stat) come from the Kolmogorov-Smirnov test [27] on the distributions between misinformation and the other tweets. Rows are ordered by p-value. Variables above the line are those whose differences are considered statistically significant (p-value smaller than 0.01).

	p-value	t-stat
tweet.fav_count	0.00	0.184
tweet.num_hashtags	0.00	0.169
tweet.num_mentions	0.00	0.267
tweet.num_media	2.07e-10	0.091
tweet.retweet_count	5.35e-01	0.022
tweet.num_urls	1.00e+00	0.005

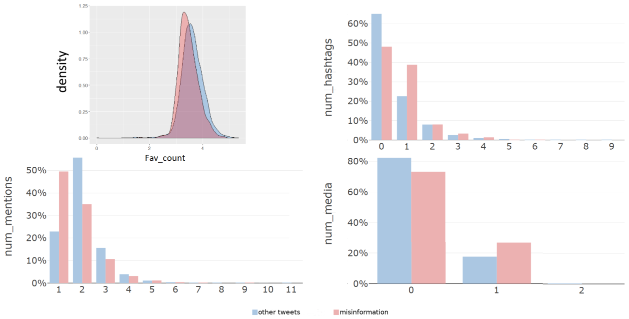


FIGURE 7. Distributions of the four significant variables related with the tweets (see Table 5). We can observe that tweets with misinformation have generally more hashtags and media but less mentions.

In addition to the features listed in Table 5, we looked at the distribution of tweet sources (iPhone, Android, web client, media studio, etc.) for both tweets containing misinformation and those not containing them, and both distributions were very similar: 40% vs 42% for iPhone, 33% vs 32% for the web client and 10% vs 9.5% for Android. Those marginal differences confirm there is no significant difference for this specific feature and do not offer any other meaningful insight.

Finally, we also analysed the most used hashtags in both subsets of tweets (leaving aside the ones used for collection), and we found that there is no statistical difference either between hashtags used in tweets containing misinformation and tweets not containing them. Fig 8 shows their frequency distribution. It is interesting to see that a couple of hashtags only appear in tweets labelled as misinformation. While this might be a product of the dataset, it is an issue that probably deserves further research. However, once again, the frequency of particular hashtags is hardly generalizable or usable on another dataset.

1) FEATURES ABOUT DIFFUSION OF TWEETS

As expected, we found out that the retweet count and the favourite count are correlated ($r = 0.665$); together with the number of followers of a user and the number of times she has been listed ($r = 0.911$). However, the evolution in the number of retweets (and favourites) dramatically varies depending on the tweets. Within our dataset, we can find examples for linear, exponential and polynomial over time.

In order to find out whether or not any of these is prevalent in misinformation, we computed the time (in hours) it takes

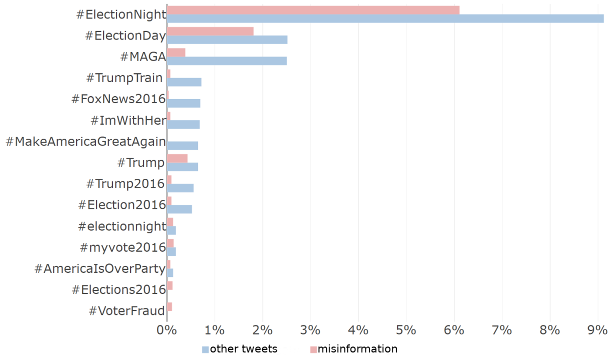


FIGURE 8. Frequency of appearance of most used hashtags in tweets containing misinformation (red) and not containing them (blue).

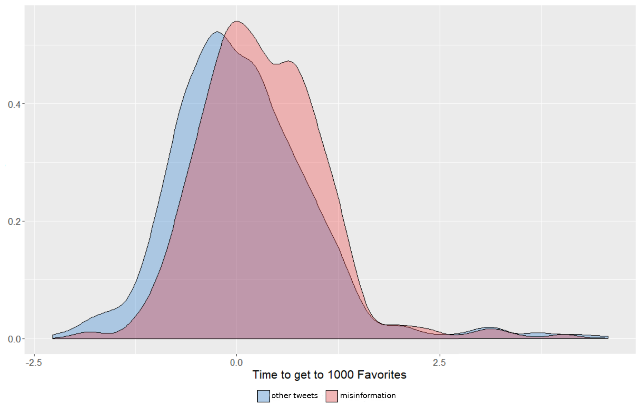


FIGURE 9. Distribution of the decimal logarithm of hours taken to get to 1000 favourites (*fav_timeto1000*) for both tweets containing misinformation and others. The associated p-value is 9.60e-11 which proves the significance of the diffusion pattern.

to get to 10, 20, 50, 100, 250, 500 and 1000 retweets (and to an equal number of favourites, respectively). Fig 9 shows the different distributions of the time to get to 1000 favourites (which is the most distinctive feature of this set) for tweets containing misinformation and other tweets. Tweets containing misinformation are generally slower to get a thousand favourites.

Finally, we also enriched the original dataset by computing the day and the hours at which the tweet was published (in the EST referential). The number of days which separates the creation of a tweet and the creation of its Twitter account is, on average, 1941 days for tweets containing misinformation and 2100 days for other tweets. This delta represents a computed p-value of 1.04e-08 which is indeed highly significant. This result suggests that accounts spreading misinformation were created more recently (already observed in [46]). This result is coherent with the fact that users who spread misinformation have their accounts eventually deleted and are forced to create new ones.

IV. DISCUSSION

Results presented in the sections above are intended to highlight differences in sources (accounts from which tweets are coming) and style (how textual elements of the tweets are crafted) between tweets containing misinformation and regular information. Our analysis shows that accounts sharing misinformation were created more recently, are less likely to be verified, and have fewer updates than those sharing other types of information. Also, these accounts tend to use weird characters in both their screen name and description. Moreover, accounts sharing misinformation have fewer followers but tend to follow others more often, and are more likely to favourite the content of others.

Such characteristics suggest that accounts sharing misinformation not only have a different profile from those sharing other types of information,[4] but also exhibit different behaviours such as following others and favouriting their content more often. Particularly interesting to our aim is to identify the before-mentioned behavioural differences. Research into reciprocity in online social communities [10], [13] has found that 'creating directed links to other nodes drive the latter to correspond by creating a link to the former'. This behaviour, known as *altruistic reciprocity*, has been widely confirmed across online social platforms [13]. We hypothesise accounts sharing misinformation may follow this behaviour as reciprocity may make it more likely to increase the visibility of their content.

Turning to stylistic differences, research in the field of automatic deception detection suggests syntactic differences could single-out deception. Reference [14] explains deceivers usually carefully craft their messages, but that syntactic leakages are, both, inevitable and detectable. Our results show such differences were present not only in the textual fields of tweets but also in screen name and descriptions of the accounts (Table 3). About the tone of the message, research into Bayesian decision theory [4], [22] has shown elements of novelty attract human attention because new information is central to update our understanding of the world. Fig 3 shows that our proxy for novelty (surprise) is overall more prevalent in misinformation than in regular information. Because this feature may be a product of the nature of misinformation and, hence, not engineered, we turned to analyse sentiment throughout the temporal dimension of our sample. From Fig 4, two facts stand out. First, if misinformation were novel by nature, it would be expected that the sentiment of surprise would be generally higher when compared to regular information. We found the mean sentiment of surprise for misinformation and other types of information to be 0.11 and 0.09 respectively. Second, there is a large amount of variation in sentiments for misinformation but not for regular information. Such a finding is puzzling as it lends itself to considerable speculation (e.g., are these drastic changes contradictory and, hence, the source of novelty? are these changes intentional? or are they part of the nature of misinformation?). Even if these findings are not helpful to settle the issue, it is easy to see that, both, the domain and variability

[4] Upon detailed inspection, it is, perhaps, no coincidence that profile is shared with that identified by [19], [26] as accounts dedicated to sharing misinformation during the 2016 election in the United States.

of the different emotions computed seem to be larger for misinformation. This variation may suggest, however, that misinformation may be dynamically changing to create polarisation.

As related by [42], political polarisation in the form of partisanship profoundly affects the way individuals process information in the presence of new evidence. Highly polarised information is more likely to be rationalised by individuals if they agree with such information. This effect is called *confirmation bias*. To test for the possibility such information may be engineered, we computed its sentiment. One would expect the slope of both curves in Fig 6 to be close to zero if the textual features of such tweets are not engineered. We would expect this because such a slope would be the product of a lack of coordination. However, what we see is that the slope for the fit of misinformation is positive. It could be argued that such coordination was the product of the election of Donald Trump. However, even after eliminating all the data just after November 9th 2016, such trend is still visible (see Fig 6).

Systematic differences in the source of misinformation, behaviours and features that are likely to increase the impact of misinformation, together with evidence pointing to continuous, seemingly coordinated, changes in tone of the textual elements of the tweets appear to point towards an effort to engineering political misinformation. The media have widely covered such an effort. However, to the best of our knowledge, our results are amongst the first efforts to shed light on the systematic way political misinformation was engineered. Furthermore, we have suggested such efforts are tailored to exploit known human biases such as confirmation bias or other effects such as that of *altruistic reciprocity*. Our analysis confirms a higher proportion of favourites for misinformation.

One might argue that bots drive this result. However [46] showed that the proportion of bots used by accounts spreading true information was the same to that of those spreading false information. Given that our dataset should be a random subset of theirs, we would expect such proportion to be maintained. The evidence presented so far raises the possibility that, at least within the context of our data, careful engineering of misinformation could be a driver for diffusion. Behaviours such as the above -average level of following and favouriting and the amounts of polarisation and novelty in the textual field of the message seem to exploit human biases. Together with the fact that the latter appears to be a product of coordination suggest these actions may help political misinformation diffuse faster and broader than other types of information. It is important to stress that most of our results match those of [46], which makes it possible that our results may translate into more general contexts.

We acknowledge that our study is limited by the data, both by its size and scope, and the hardness of having an accurate and unbiased labelling (as Table 1 illustrates). It seems clear that different cultural backgrounds, knowledge of the American culture, and English language proficiency induced vastly different perceptions on whether a piece of information is considered misinformation or not.

We are currently in the process of performing similar studies with datasets on different electoral processes that we have collected in the last year. Until then, the findings reported here should be taken more like potential leads rather than established truth.

ACKNOWLEDGMENT

Titan Xp GPUs used for this research were kindly donated by NVIDIA Corporation.

REFERENCES

[1] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: Divided they blog," in *Proc. 3rd Int. Workshop Link Discovery (LinkKDD)*, New York, NY, USA, 2005, pp. 36–43.

[2] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *J. Econ. Perspect.*, vol. 31, no. 2, pp. 211–236, 2017.

[3] J. A. D. Lopez, A. Oehmichen, and M. Molina-Solana, "Fakenews on 2016 US elections viral tweets (November 2016—March 2017)," Imperial College London, London, U.K., 2017. doi: 10.5281/zenodo.1048826.

[4] S. Aral and M. Van Alstyne, "The diversity-bandwidth trade-off," *Amer. J. Sociol.*, vol. 117, no. 1, pp. 90–171, Jul. 2011.

[5] D. Arnaudo, "Computational propaganda in Brazil: Social bots during elections," Oxford Internet Inst., Univ. Oxford, Oxford, U.K., Working Paper 2017.8, 2017.

[6] V. Bakir and A. McStay, "Fake news and the economy of emotions: Problems, causes, solutions," *Digit. J.*, vol. 6, no. 2, pp. 154–175, 2018.

[7] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.

[8] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, pp. 295–298, Sep. 2012.

[9] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," in *Proc. 43rd Hawaii Int. Conf. Syst. Sci.*, Jan. 2010, pp. 1–10.

[10] J. Cheng, D. M. Romero, B. Meeder, and J. Kleinberg, "Predicting reciprocity in social networks," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust 3rd Int Conf Social Comput.*, Oct. 2011, pp. 49–56.

[11] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks," *PLoS One*, vol. 10, no. 10, 2015, Art. no. e0141938.

[12] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," in *Proc. 78th ASIST Annu. Meeting, Inf. Sci. Impact, Res. Community*, vol. 52, no. 1, pp. 1–4, 2015.

[13] V. S. Dave, M. Al Hasan, B. Zhang, and C. K. Reddy, "Predicting interval time for reciprocal link creation using survival analysis," *Social Netw. Anal. Mining*, vol. 8, no. 1, p. 16, Dec. 2018.

[14] V. W. Feng and G. Hirst, "Detecting deceptive opinions with profile compatibility," in *Proc. 6th Int. Joint Conf. Natural Lang. Process.*, Nagoya, Japan, Oct. 2013, pp. 338–346.

[15] D. J. Flynn, B. Nyhan, and J. Reifler, "The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics," *Political Psychol.*, vol. 38, pp. 127–150, Feb. 2017.

[16] French Assemblée Nationale, "Loi n. 2018-1202 du 22 décembre 2018 relative à la lutte contre les fausses informations," *Journal Officiel de la République Française*, vol. 297, 2018, Art. no. 2. [Online]. Available: https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000037847559&categorieLien=id

[17] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford Univ., Stanford, CA, USA, Tech. Rep. CS224N, 2009. [Online]. Available: https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

[18] C. A. Gomez-Uribe and N. Hunt, "The Netflix recommender system: Algorithms, business value, and innovation," *ACM Trans. Manage. Inf. Syst.*, vol. 6, no. 4, 2016, Art. no. 13.

[19] House Intelligence Committee. (2018). *IRA Handles*. [Online]. Available: https://intelligence.house.gov/uploadedfiles/ira_handles_june_2018.pdf

[20] P. Howard, "How political campaigns weaponize social media bots," *IEEE Spectr.*, Oct. 2018. [Online]. Available: https://spectrum.ieee.org/computing/software/how-political-campaigns-weaponize-social-media-bots

[21] P. N. Howard, S. Woolley, and R. Calo, "Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration," *J. Inf. Technol. Politics*, vol. 15, no. 2, pp. 81–93, 2018.

[22] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vis. Res.*, vol. 49, no. 10, pp. 1295–1306, Jun. 2009.

[23] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 2, 2017, pp. 427–431.

[24] R. L. Kaplan, "Press, paper, and the public sphere," *Media Hist.*, vol. 21, no. 1, pp. 42–54, 2014.

[25] O. Kolchyna, T. T. P. Souza, P. Treleaven, and T. Aste, "Twitter sentiment analysis: Lexicon method, machine learning method and their combination," 2015, *arXiv:1507.00955*. [Online]. Available: https://arxiv.org/abs/1507.00955

[26] D. L. Linvill and P. L. Warren, "Troll factories: The Internet research agency and state-sponsored agenda building," Clemson Univ., Clemson, SC, USA, Tech. Rep., 2018. [Online]. Available: https://www.rcmediafreedom.eu/Publications/Academic-sources/Troll-Factories-The-Internet-Research-Agency-and-State-Sponsored-Agenda-Building

[27] F. J. Massey, Jr., "The Kolmogorov–Smirnov test for goodness of fit," *J. Amer. Stat. Assoc.*, vol. 46, no. 253, pp. 68–78, 1951.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. NIPS*, 2013, pp. 1–12.

[29] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, 2013.

[30] A. Oehmichen, F. Guitton, K. Sun, J. Grizet, T. Heinis, and Y. Guo, "eTRIKS analytical environment: A modular high performance framework for medical data analysis," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2017, pp. 353–360.

[31] S. Oraby, L. Reed, R. Compton, E. Riloff, M. Walker, and S. Whittaker, "And that's a fact: Distinguishing factual and emotional argumentation in online dialogue," in *Proc. 2nd Workshop Argumentation Mining*, 2015, pp. 116–126.

[32] G. Pennycook, T. D. Cannon, and D. G. Rand, "Prior exposure increases perceived accuracy of fake news," *J. Exp. Psychol., Gen.*, vol. 147, no. 12, pp. 1865–1880, 2017.

[33] G. Pennycook and D. G. Rand, "Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking," *J. Personality*, to be published.

[34] D. C. Polage, "Making up history: False memories of fake news stories," *Eur. J. Psychol.*, to be published.

[35] A. I. Pontes, M. Henn, M. D. Griffiths, and H. M. Pontes, "Validation of the online political engagement scale in a British population survey," *Aloma, Revista de Psicología, Ciències de l'Educació i de l'Esport*, vol. 35, no. 1, pp. 13–21, 2017.

[36] V. L. Rubin, "Deception detection and rumor debunking for social media," in *The SAGE Handbook of Social Media Research Methods*, L. Sloan and A. Quan-Haase, Eds. London, U.K.: Sage, 2017. [Online]. Available: https://uk.sagepub.com/en-gb/eur/the-sage-handbook-of-social-media-research-methods/book245370

[37] V. L. Rubin, Y. Chen, and N. J. Conroy, "Deception detection for news: Three types of fakes," in *Proc. 78th ASIS & T Annu. Meeting, Inf. Sci. Impact, Res. Community*, 2015, Art. no. 83.

[38] S. Sanovich, *Computational Propaganda in Russia: The Origins of Digital Disinformation*, S. Woolley and P. N. Howard, Eds. Oxford, U.K.: Project on Computational Propaganda, 2017, p. 24. [Online]. Available: http://comprop.oii.ox.ac.uk/

[39] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "The spread of low-credibility content by social bots," *Nature Commun.*, vol. 9, Nov. 2018, Art. no. 4787.

[40] J. Silge and D. Robinson, *Text Mining With R: A Tidy Approach*. Newton, MA, USA: O'Reilly Media, 2017.

[41] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie, *General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA, USA: MIT Press, 1966.

[42] B. Swire, A. J. Berinsky, S. Lewandowsky, and U. K. H. Ecker, "Processing political misinformation: Comprehending the Trump phenomenon," *Roy. Soc. Open Sci.*, vol. 4, no. 3, 2017, Art. no. 160802.

[43] The Digital, Culture, Media and Sport Committee, "Disinformation and 'fake news': Final report," House Commons—Digit., Culture, Media Sport Committee, London, U.K., Tech. Rep. HC 1791, Feb. 2019. [Online]. Available: https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf

[44] US Department of Justice. (2018). *Russia-Trump Inquiry: Full Text of Mueller's Indictment*. [Online]. Available: https://www.justice.gov/file/1035477/download

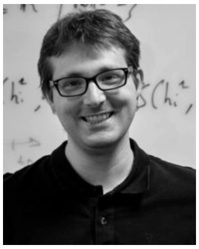
[45] US Federal Trade Commission. (1914). *Dissemination of False Advertisements*. [Online]. Available: https://www.law.cornell.edu/uscode/text/15/52

[46] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, pp. 1146–1151, May 2018.

[47] H. Webb, M. Jirotka, B. C. Stahl, W. Housley, A. Edwards, M. Williams, R. Procter, O. Rana, and P. Burnap, "The ethical challenges of publishing Twitter data for research dissemination," in *Proc. ACM Web Sci. Conf.*, New York, NY, USA, 2017, pp. 339–348.

[48] H. Zhang, Z. Fan, J. Zheng, and Q. Liu, "An improving deception detection method in computer-mediated communication," *J. Netw.*, vol. 7, no. 11, p. 1811, Nov. 2012.

[49] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 8, no. 4, p. e1253, 2018.



AXEL OEHMICHEN received the M.Eng. degree in computer science and applied mathematics from ENSEEIHT, in 2012, and the Ph.D. degree from the Imperial College London, in 2018.

After working for a year in the City of London for Societe Generale CIB, he joined the Data Science Institute, Imperial College London, as a Research Assistant, where he is currently a Postgraduate Research Associate.

His current research interests include the creation of flexible and scalable platforms for collecting, storing, and analyzing large scale data in a privacy preserving fashion.

KEVIN HUA received the M.Eng. degrees in general engineering and statistics from École Centrale de Lyon, and the M.Sc. degree in computer science from Université Claude Bernard de Lyon 1, in 2017.

He has three years of experience in data science applied in diversified fields (consultancy, bank, insurance, and social medias). He joined Imperial College London as a Research Associate in applied data science.

His current research interests include analyzing closely data, making suitable visualizations to provide intelligible knowledge and making precise machine learning models to optimize profitability.

JULIO AMADOR DÍAZ LÓPEZ received the Ph.D. degree in economics from the University of Essex.

He held different research positions in the U.K. and abroad. He is currently a Research Fellow with the Imperial College London. His current research interest is applied machine learning (ML), including big-data studies of online political participation and applying ML to categorize public opinion and automatically identifying fake news.

Dr. Amador is currently dedicated to the study of misinformation.

MIGUEL MOLINA-SOLANA received the degree in computer science and the Ph.D. degree from the Universidad de Granada, in 2007 and 2012, respectively.

From 2012 to 2015, he was a Research Associate with the Universidad de Granada in the FP7 Project Energy IN TIME. He was a Research Associate on visualization with the Data Science Institute, Imperial College. He is currently a Marie Curie Research Fellow with the Imperial College London, U.K. His current research interests include applied work in machine learning and knowledge representation in diverse domains, such as music, energy management, and business.

Dr. Molina-Solana is the Principal Investigator of the H2020 Project DATASOUND: Understanding Data with Sound.

JUAN GÓMEZ-ROMERO received the degree in computer science and the Ph.D. degree in intelligent systems from the Universidad de Granada, in 2004 and 2008, respectively.

He was a Lecturer with the Applied Artificial Intelligence Group, Universidad Carlos III de Madrid, from 2008 to 2013, and a Research Associate in the EU FP7 Project Energy IN TIME with the Universidad de Granada, from 2013 to 2017. He was also a Visiting Researcher with the Data Science Institute, Imperial College London, from 2016 to 2017. He has been a Senior Research Fellow with the Computer Science and Artificial Intelligence Department, Universidad de Granada, since 2016. He has participated in more than 20 projects in security, ambient intelligence, and energy efficiency. His current research interests include the use of semantic representation models and machine learning techniques to perform automatic reasoning toward higher-level information fusion.

Dr. Gómez-Romero is the Principal Investigator of the projects BIGFUSE: Semantics for Big Data Fusion and Analysis: Improving Energy Efficiency in Smart Grids and PROFICIENT: Deep Learning for Energy-Efficient Building Control.

YI-KE GUO received the degree (Hons.) in computing science from Tsinghua University, China, in 1985 and the Ph.D. degree in computational logic from Imperial College, in 1993, under the supervision of Prof. J. Darlington. He is currently a Professor of computing science with the Department of Computing, Imperial College London, where he is also the Founding Director of the Data Science Institute and leading the Discovery Science Group. He holds the position of CTO of the tranSMART Foundation, a global open source community using and developing data sharing and analytics technology for translational medicine.