Trust No One!
Using Machine Learning to Measure Distrust in Partisan WhatsApp Groups

Victor Soares Bursztyn, 3046613. Professor Doug Downey, EECS 349 (Fall 2018).

Introduction

On October 28th, amid widespread fake news and conspiracy theories that flooded social media[1], Brazilians elected far-right candidate Jair Bolsonaro their next president. With over 120M users in Brazil — the second largest market in the world —, the role that WhatsApp played in this electoral process has emerged as a major focus of attention and controversy because of its alleged effectivity in a large set of sucessful campaigns. Not only Bolsonaro has been loud about his use of the platform, but over 50 conservative congressmen who followed his steps got elected for the first time. However, due to WhatsApp's private nature and lack of transparency, it's hard to assess how exactly the platform was used. A recent study suggests that fake news that circulated massively through WhatsApp were more far-reaching than initially assumed as it revealed that 90% of Bolsonaro's constituency think they are truthful.[2]

On top of that, similarly to other elections worldwide, the expression "fake news" has been captured by partisan groups and used to promote distrust toward several institutions, ranging from mainstream media to traditional polling organizations. After collecting ~3M messages from hundreds of semi-public right-wing and left-wing WhatsApp groups during the Brazilian election, my prediction task is defined as follows: to predict when users from clearly partisan groups are calling something false (e.g., "a lie", "a fraud" or "fake news") as a function of all other terms recently used in the conversation. As illustrated in Fig 1, these neighboring terms can relate to polling organizations (marked in blue), to media outlets (green), to the opposition (purple), and even to traditional democratic institutions that rely on public trust, such as the voting booth (yellow).

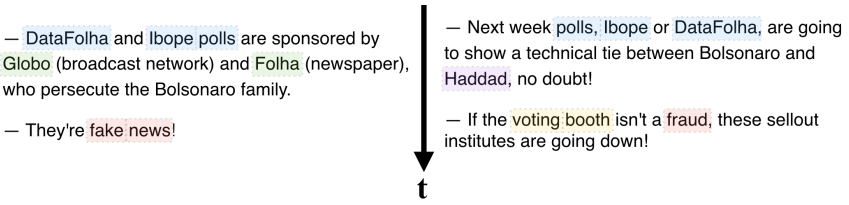
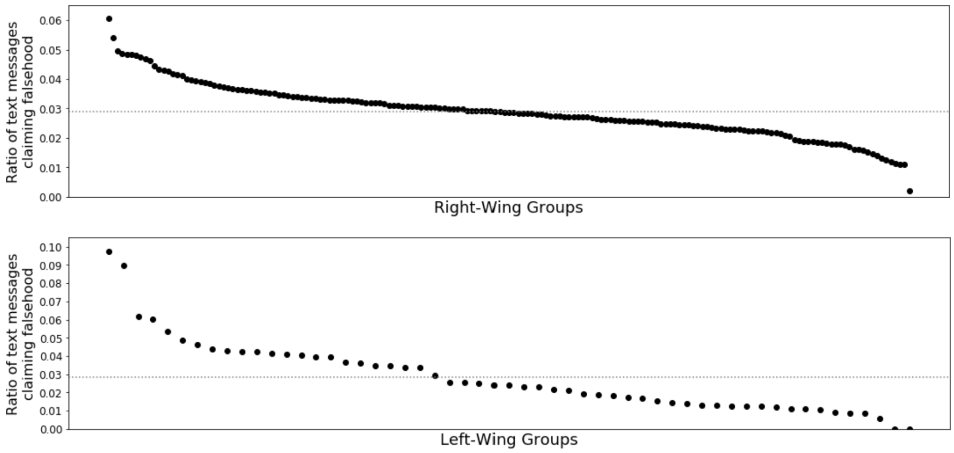


Fig. 1 - Two real examples of users calling something false in right-wing groups (freely translated).

This task is important because the behavior at hand took over the political debate during the Brazilian election, happening in a space (WhatsApp groups) and at a scale that are very challenging for political scientists to address. In this data set, almost 3% of all text messages contain terms claiming that something is false (Fig 2). Therefore, framing this behavior as a prediction task that depends on the neighboring terms can help unveiling the underpinning political discourse and measuring how it varied across the political spectrum.



[1] "Did WhatsApp help Bolsonaro win the Brazilian presidency?" - https://www.washingtonpost.com/news/theworldpost/wp/2018/11/01/whatsapp-2/

[2] "90% of Bolsonaro's constituents believe in fake news" - https://www1.folha.uol.com.br/poder/2018/11/90-dos-eleitores-de-bolsonaro-acreditaram-em-fake-news-diz-estudo.shtml

Fig. 2 - ~3% of all text messages in right-wing and left-wing groups contain terms claiming that something is false.

Furthermore, approaching this task with Machine Learning as opposed to descriptive analyses (e.g., creating tag clouds based on the most frequent neighboring terms) is important because it can provide better measurements on the particular triggers of the target behavior (i.e., promotion of distrust) in each partisan group. In this sense, a Machine Learning-based method can not only unveil the underpinning political discourse, but can also show how predictable this potentially complex behavior is across the political spectrum.

Data Set & Feature Engineering

As Fig 2 also shows, the data set comprises more right-wing groups than left-wing ones, which matches public perception and domain experts' knowledge on how WhatsApp was adopted by the Brazilian electorate. Interestingly, however, the behavior under study has some commonalities across partisan groups. First, the ratio of all text messages claiming falsehood is about 3%, which provides 19132 instances of the target behavior in right-wing groups and 3253 instances in left-wing groups. Second, as shown in Fig 3, the distributions of users calling something false have similar shapes. The view provided by the log-log CCDFs (Fig 3) shows that even their tails (i.e., the users who exhibit this behavior most frequently) are similarly shaped.

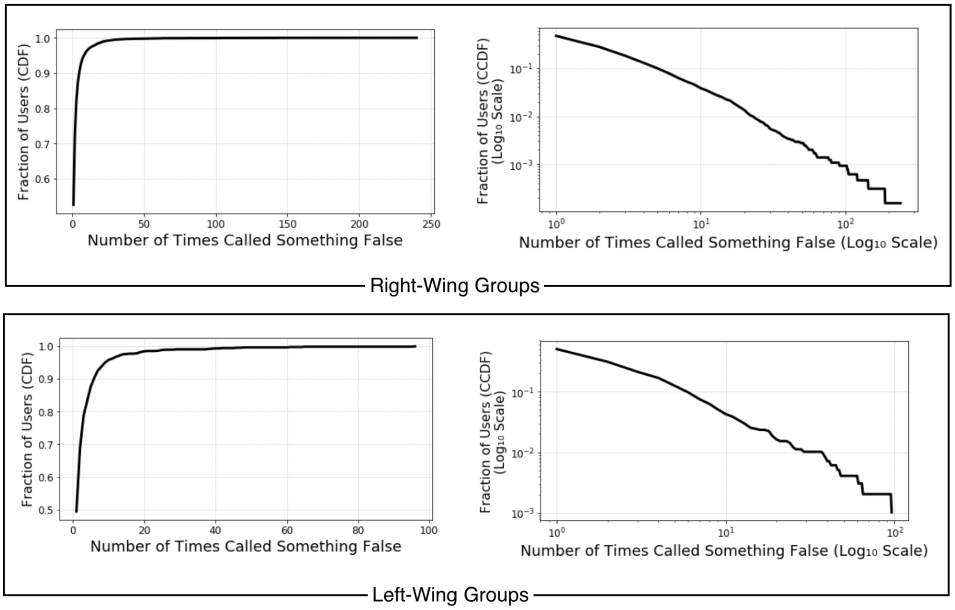


Fig. 3 - The distribution of users calling something false in right-wing and left-wing groups (CDFs and log-log CCDFs).

Knowing these distributions matters because it allows designing an experiment where the differential aspect is rigorously observed. This is done by focusing on the population that calls something false and gathering counter-factual evidence as precisely as possible — if a given user calls something false n times, gather n instances where the *same user* doesn't exhibit this behavior. This way, a statistical learner can find the terms that most reliably trigger the target behavior. Following this process for all users — *for the same set of users who calls something false and following the frequencies at which they exhibit this behavior, retrieving an equal amount of messages where nothing is called false* — leads to the creation of a well-balanced data set for the prediction task, with a total of 38264 instances from right-wing groups and 6506 from left-wing groups.

The data set is then extended in two ways:

1. *Retrieval of recent conversations* — For every instance, the three preceding messages are retrieved and included;
2. *Retrieval of external references* — For every message containing an external URL, the title of the URL is retrieved and embedded in that text message as this is how WhatsApp shows URLs in its interface;

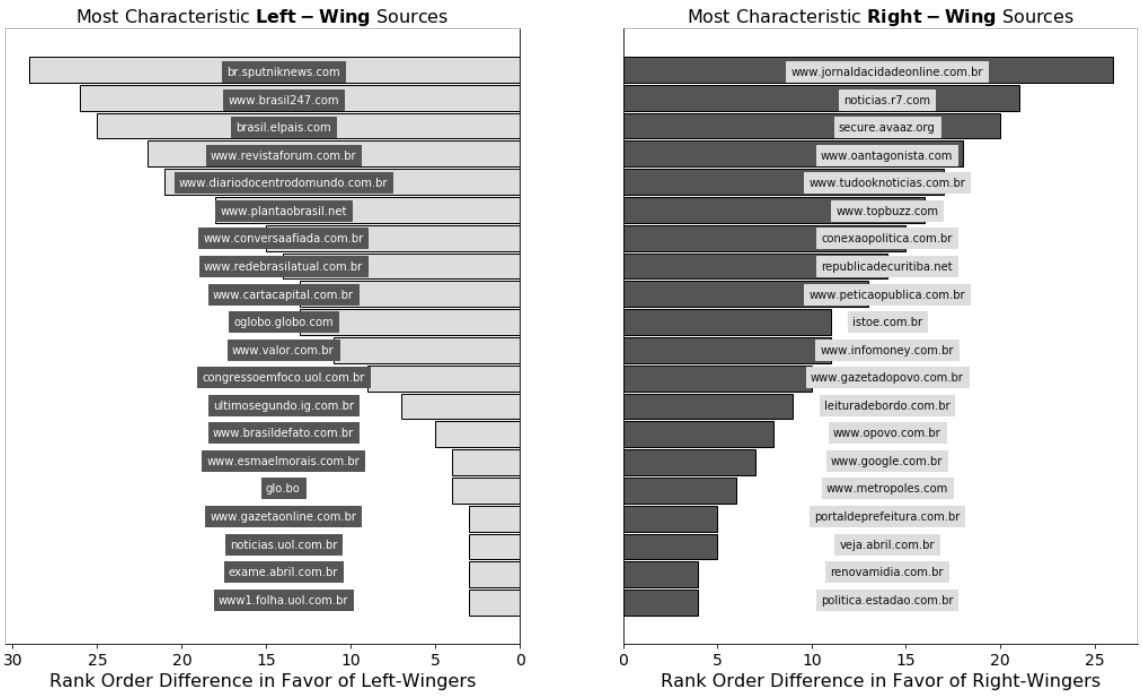


Fig. 4 - Largest rank order differences indicate news sources that are most characteristic of each partisan group.

Once the data set is designed and extended, it needs to be structured around content-defining features (mostly terms) that are both useful and adherent to Machine Learning algorithms. Fig 4 shows how certain features can help to characterize each subset of interest. In particular, it shows the most characteristic news sources based on *all* messages from right-wing and left-wing groups. This is done in two steps: first, I calculate the top 30 domains that were most frequently shared among right-wingers and left-wingers; then, I order these domains according to the largest discrepancies between the two internal rankings, thus maximizing the difference in rank orders. The rationale behind Fig 4, where the most characteristic features are pursued, can be applied within each partisan group in order to identify text features that (potentially) differentiate the target behavior from the counter-factual observations. With this in mind, the data set is structured in six steps:

Stop-words removal — For right-wingers and left-wingers, a set of stop-words is discarted including the original set of terms indicating that something is false (e.g., “lie”, “fraud” or “fake”);

Identification of most characteristic terms — Within each partisan group, for each half of the data set (i.e., with and without the target behavior), the 100 most characteristic terms are calculated analogously to Fig 4;

Selection of meaningful terms — Each term is evaluated as to whether it could refer to a meaningful and interpretable topic, and feature candidates that don’t signify much are discarted;

Extraction of URL domains — Also inspired by Fig 4, all URLs are extracted and included as features;

Temporal identification — Temporal *ids* are appended to indicate when these features are observed (in the actual message where the target behavior happens — t_0 — or in any of the three preceding messages);

One-hot encoding — Finally, the data set is represented using one-hot encoding w.r.t. the selected text features.

Results

This work focuses on two learners whose representations provide clear insights on the most predictive features: Decision Trees (where terms are ordered by their *information gains*) and Logistic Regressions (where terms are ordered by their *coefficients*). In this context, predictive features (mostly terms) should show good precision and, hopefully, recall. After all, for either right-wingers or left-wingers, precision for the target class means finding the terms that most consistently identify users claiming something is false, while recall means finding the terms that identify all times when users are exhibiting such a behavior. However, since the entire range of topics causing users to promote distrust could be large and heterogeneous, it may be reasonable to prioritize precision over recall. In this sense, F0.5-score should reflect the appropriate

balance of precision and recall, and 10-fold cross-validation (CV) should be a good way of navigating the bias-variance tradeoff toward a stable model.
The learning task hence defined is performed using Weka 3.8.2 (Hall et al. 2009) over the generated data set as it allows rapid experimentations. Parameter sweeps are necessary to adjust two hyperparameters for Decision Tree learners — (i) the minimum number of instances per leaf; and (ii) the confidence factor used for pruning —, while 10-fold CV is still observed. Fig 5 shows learning performance w.r.t. the target class for both learners in each partisan group. It is worthy of note that Decision Trees (DTs) perform best in this prediction task for both right-wingers and left-wingers as they exceed 70% precision in a setting where 50% would be the naive baseline (i.e., the data set is perfectly balanced and this is a binary classification task). Regarding F0.5-score, DTs achieve 75% for right-wingers and approach 70% for left-wingers. Nevertheless, it is fair to note that all four learners shown in Fig 5 are at least partially successful when contrasted with the naive baseline, which suggests that the prediction task is correctly formulated.

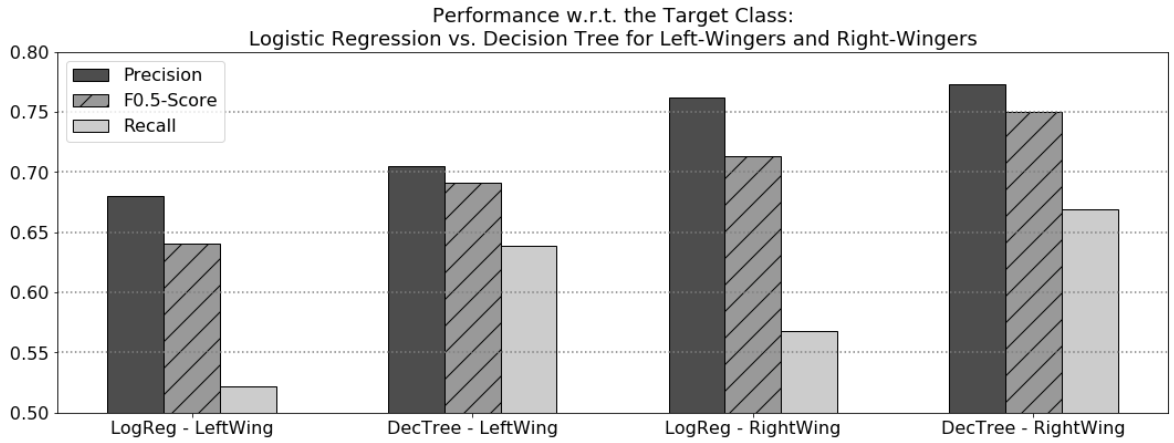


Fig. 5 - Logistic Regression and Decision Tree learners performing way above the naive baseline for right-wing and left-wing groups.

Furthermore, Table 1 shows the top 10 most predictive features in each learning setting, both in raw format and the corresponding translations (in parentheses). Cells in bold represent features that make the cut in both learners considering the same data set (either right-wingers or left-wingers).

LogReg - LeftWing	DecTree - LeftWing	LogReg - RightWing	DecTree - RightWing
NOTAS_t0 (bills)	**NOTAS_t0 (bills)**	ARTISTA_t-1 (artist)	ESPALHAR_t0 (spread)
TSE_t-1 (Superior Electoral Court)	**TSE_t0 (Superior Electoral Court)**	ARTISTA_t-2 (artist)	LEGISLAÇÃO_t0 (legislation)
EMPRESAS_t-2 (companies)	**TSE_t1 (Superior Electoral Court)**	**COMUNICAÇÃO_t0 (communication groups)**	ELEITORAIS_t0 (electoral)
WHATSAPP_t0	VERDADE_t0 (truth)	KIT_t0 (“gay kit” factoid)	mentiramparamimsobreojair.com_t0 (they-lied-to-me-about-Jair)
YOUTUBE_t-3	CAIXA_t0 (illegal campaign contributions)	**URNAS_t0 (voting booths)**	**URNAS_t0 (voting booths)**
ELENÃO_t-3 (colloquial reference to opposition)	plantaobrasil.net_t0 (Brasil-on-call)	PETRALHAS_t-1 (colloquial reference to opposition)	**COMUNICAÇÃO_t0 (communication groups)**
ELENÃO_t-1 (colloquial reference to opposition)	NOTÍCIAS_t0 (news)	FOLHA_t0 (mainstream newspaper)	VERMELHO_t0 (colloquial reference to opposition)
MÍDIA_t-1 (media)	chat.whatsapp.com_t0	IMPRENSA_t0 (press)	LENGTH_t0

LogReg - LeftWing	DecTree - LeftWing	LogReg - RightWing	DecTree - RightWing
TSE_t0 (Superior Electoral Court)	**WHATSAPP_t0**	MÍDIA_t0 (media)	ENTREVISTA_t0 (interview)
FOLHA_t-3 (mainstream newspaper)	TSE_t2 (Superior Electoral Court)	COMUNISMO_t0 (communism)	HADDAD_t-2 (Fernando Haddad)

Table 1 - Top 10 most predictive features in each learning setting from Fig 5.

Interestingly, from this angle, learners diverge more than they converge. This happens because of their different approaches in regard to how to represent (and learn) hypotheses that model the prediction task. However, the points of convergence can be strong indicators of terms that trigger the target behavior.
Among right-wingers, “URNAS” (as in voting booths) and “COMUNICAÇÃO” (as in communication groups) appear to be highly predictive of people calling something false. Several other terms related to these two subjects (“ELEITORAIS” referring to “URNAS” *vs.* “ENTREVISTA”, “ARTISTA”, “FOLHA”, “IMPRENSA” and “MÍDIA” referring to “COMUNICAÇÃO”) provide additional evidence supporting this claim.
Among left-wingers, “NOTAS” is a false positive since it refers to a spam message offering fake bills. Taking it off the table, “TSE” (as in Brazil’s Superior Electoral Court) and “WhatsApp” appear to be highly predictive of people calling something false. Indeed, they do relate to the same subject: while the Superior Electoral Court is the institution supposed to watch over the election and enforce its rules, WhatsApp has emerged as a powerful and opaque platform strongly appealing to the far-right. Other terms related to the subject (e.g., “CAIXA” and “FOLHA”, referring to allegations of illegal campaign financing involving WhatsApp and reported by the newspaper) can corroborate this interpretation.

Conclusion & Future Work

In general, contrasting with the naive baseline for this binary classification task, achieving 65%—75% of F0.5-score should be considered a success. However, the 5–10% superior performance among right-wingers indicate that the phenomenon under study is more easily predicted within right-wing groups, at least with the current set of features (terms) obtained in this work. This is an interesting measurement as it elicits some underlying uncertainties that could otherwise be burried in descriptive analyses. At the same time, measuring the most predictive terms using different hypothesis spaces (from different Machine Learning algorithms) may provide stronger evidence on the subjects most likely to trigger the target behavior across partisan groups. In this sense, the method described herein helps to address a relatively new research space (public WhatsApp groups), to answer a contemporary RQ (promotion of distrust), and it does so by (i) framing a domain-specific statistical problem as a Machine Learning problem and (ii) resorting to highly interpretable learners.
Finally, future work could include other views on the set of terms unveiled by this work — to verify, for instance, if the messages promoting distrust of the voting booth originated from particular regions within the federation (i.e., through state area codes).

References

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.