Multilingual Bot Accusations: How Different Linguistic Contexts Shape Perceptions of Social Bots

Leon Fröhling
GESIS - Leibniz Institute
for the Social Sciences
leon.froehling@
gesis.org

Xiaofei Li
RWTH Aachen
xiaofei.li1@
rwth-aachen.de

Dennis Assenmacher
GESIS - Leibniz Institute
for the Social Sciences
dennis.assenmacher@
gesis.org

Abstract

Recent research indicates that the online use of the term "bot" has evolved over time. In the past, people used the term to accuse others of displaying automated behavior. However, it has gradually transformed into a linguistic tool to dehumanize the conversation partner, particularly on polarizing topics. Although this trend has been observed in English-speaking contexts, it is still unclear whether it holds true in other socio-linguistic environments. In this work we extend existing work on bot accusations and explore the phenomenon in a multilingual setting. We identify three distinct accusation patterns that characterize the different languages.

1 Introduction

Social bots are described in academia as automated accounts in online media that have the ability to manipulate public opinion on large scale. While a plethora of current work on bots focusses on their detection, researchers are increasingly interested in how people perceive social bots and how they talk about them in the online sphere. In this work we extend a recent study by Assenmacher et al. (2024) who look at accusation situations, i.e. those instances where a user on Twitter (𝕏) accuses another user of being a bot. While earlier work only focused on the English speaking landscape of the platform we want to investigate the phenomenon in a multi-lingual setting, examining the following research question:

> *What are the differences in how users accuse each other of being a bot across different languages, and can we identify a taxonomy of these accusation types?*

2 Related Work

Up until recently, the largest part of research on social bots has been concerned with developing methods for their detection, trying to expose and characterize their efforts to systematically influence the discourse on social media (Yaojun Yan and Yang, 2023). While most of these approaches focus on the platform Twitter and only look on the English parts of it by default (Orabi et al., 2020), few methods are available that are by design multilingual (Martin-Gutierrez et al., 2021; Lundberg et al., 2019), helping to grapple with a phenomenon that has been found to impact the online public discourse and events in countries around the world (Woolley, 2016), cascading across cultural and linguistic borders (Xu et al., 2024). While researchers developing these methods have to start out from the difficulty of operationalizing a phenomenon that has been found to be changing in meaning even within the academic discourse (Grimme et al., 2017), recent work by Assenmacher et al. (2024) has found an even more drastic shift in the understanding of the concept bot by social media users, away from the academic definition of *a program that automatically produces content and interacts with humans on social media* (Ferrara et al., 2016) towards one that locates bots in the context of polarizing debates and insulting or even dehumanizing comments, effectively denying a (human) social media user their ability to meaningfully participate in the public discourse. Their empirical finding, restricted to the English linguistic context, has been backed up by evidence from a survey conducted by Kats and Sharif (2022), who report that more than a third of participants define bots as "fake accounts", "posing as actual humans", thereby trying to "sway public opinion". The findings by Schmuck and Von Sikorski (2020) further indicate that users might be impacted by the threat they perceive from bot campaigns on social media by exposure to news coverage on them, opening up a channel for different perceptions across regional contexts. Given such conceptual difficulties, it might not come as a surprise that researchers

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 14–32

report for different platforms and linguistic and cultural contexts, including Russian (Kolomeets et al., 2024), Chinese (Tian and Fussell, 2024) and English (Kenny et al., 2024), how platform users struggle to distinguish human from bot accounts. On top of this general confusion, recent research on human perceptions of bots has found that users tend to perceive accounts with opposing viewpoints as non-human (Wischnewski et al., 2021), easily dismissing their opinions (Schweitzer et al., 2024).

3 Data

To extend the study conducted by Assenmacher et al. (2024) to a multi-lingual context, we combine different datasets collected and annotated for complimentary purposes.

3.1 Multilingual Data Collection

The first step was to replicate the data collection for different languages of interest. In the English-only study by Assenmacher et al. (2024), they first collected all tweets containing the keyword bot, before then developing their own bot accusation classifier to select those tweets that could actually be considered bot accusations. We used the findings of Pfeffer et al. (2023) to select some of the most popular languages on Twitter for inclusion in our analysis, as well as Korean as a language for which we knew of existing research and general media coverage on popular bot campaigns (Keller et al., 2020) and German as the language of our own linguistic background. For each included language, we conducted extensive checks on the relevant keyword for our purpose of first collecting all tweets containing the language-specific version of the term bot, used in the same sense as the term bot in the social media context in the English language. We did so by comparing different translation tools applied to the term bot used in different constellations and contexts, as well as by searching Twitter with candidate keywords, to see whether potentially relevant tweets would show up. For all languages, we additionally consulted with colleagues who are native speakers of the respective language and asked for confirmation that the keyword we selected for the language would actually be the most likely keyword to refer to a bot on Twitter. We collected the data directly from the Twitter v2 API full-archive endpoint via the academic access. The data collection covered the twelve-year period from January 2011 to January 2023. We constructed the API queries following the pattern

[keyword]*is:reply lang:*[language],

where [keyword] and [language] are replaced with the respective keyword and language code for any of the considered languages. By design of the keyword-matching of the API, this query returns only reply-tweets in the specified language that contain the keyword either as a freestanding word or preceded and/or followed by a punctuation mark. Table 1 gives an overview of the included languages, the keywords used in the API queries, as well as the number of tweets collected.

Language	Keyword	Tweets
Arabic	بوت [bot]	333,221
French	bot	579,004
German	bot	289,260
Japanese	ボット [bot]	607,188
Korean	봇 [bot]	2,976,879
Portuguese	bot	1,920,350
Russian	бот [bot]	267,614
Spanish	bot	2,395,066
Turkish	bot	385,631

Table 1: Languages considered for the data collection, the language-specific equivalents of the English term bot used in the API query construction and the number of collected tweets.

3.2 Training Datasets

For some of the bot accusation detection methods presented below, we make use of the training dataset used by Assenmacher et al. (2024). They manually annotated a subset of 2,000 English-language tweets potentially containing bot accusations, reporting an inter-annotator agreement of $\kappa = 0.83$. While they fine-tune and evaluate exclusively for the English-language context, we hope to transfer some of their classifier's strong performance in detecting English bot accusations to the languages we are studying.

To supplement the use of the English-language dataset in developing the language-specific classifiers, we also sample training datasets directly from the tweets collected for each of the considered languages. These datasets are random samples of 3,200 tweets per language and form the basis for two different versions of training datasets. First, the tweets in their original languages are annotated using OpenAI's GPT-3.5,[1] leveraging the

[1] https://platform.openai.com/docs/models/

large language model's (LLM) zero-shot capabilities (prompt details in Appendix A). Second, the tweets are translated into English using Google Translate.[2] We acknowledge the potential for (systematic) errors when relying on LLMs and machine learning models for these tasks in the section on Limitations below, but concede that our work on such a broad range of languages on such a large scale would otherwise be infeasible.

3.3 Validation Datasets

To ensure that the bot accusation classifiers we introduce below still produce sufficiently valid results, we ask human crowdworkers - picked for their proficiency in the corresponding language and their familiarity with Twitter - to annotate subsets of 200 randomly sampled tweets per language, thereby generating high quality, groundtruth datasets. We collect three annotations per tweet, and label tweets as an *accusation* if at least two annotators considered it as such, and label them as *no accusation* otherwise. Appendix A provides details on the crowd annotations.

4 Methods

The core challenge in multilingual research is the handling of texts that the researcher is not familiar with. In our particular endeavor of studying the linguistic phenomenon of bot accusations across different languages, the ability to reliably discern tweets in which other users are actually accused of being bots from mere discussions of the concept or unrelated posts on potential synonyms is essential to the subsequent analysis methods, which, nonetheless, also need to be adjustable to a multilingual setting.

4.1 Multilingual Accusation Detection

In the following, we develop different methods for the detection of bot accusations from tweets mentioning the keyword bot in different languages. All of these approaches do not require the researcher to be able to read or understand the language, building on either pretrained models that are inherently multilingual or specialized on a specific language, or on classifiers trained on the different versions of the training datasets introduced above. Apart from the difficulties of dealing with different languages, these methods are further constrained by compute and financial budgets. While the evaluation of the GPT-3.5 annotations on the training datasets described above is very promising, it would be prohibitively expensive to annotate the full datasets using the model, available only through a paid API.

4.1.1 NLI Classifier

The first method (subsequently referred to as $\textbf{Model}_{\textbf{NLI}}$) to detect bot accusations from tweets containing the (language-specific) keyword frames the classification task as a natural language inference (NLI) problem to leverage the zero-shot capabilities of pre-trained, language-specific NLI models. The biggest advantage of this approach is that it does neither require expensive annotated data nor access to expensive state-of-the-art models. While NLI models were originally developed to classify whether a hypothesis is either a contradiction, an entailment or neutral to a given premise, they can also be used for any classification tasks, by presenting the instance to be annotated as the premise and by phrasing the available labels as the hypotheses to be classified. Based on the entailment scores assigned to the different labels presented to the model in form of the hypotheses, a final label can then be constructed for any instance presented as the premise.

We directly use the texts of the tweets as premises, and construct hypotheses both for the *accusation* and the *no accusation* label. The templates used for hypothesis-creation were not selected for linguistic sophistication, but rather to be universally applicable, following the English example "This text is about [label]", where [label] for English would either be *accusing user of being bot* or *not accusing user of being bot*. Appendix B provides details on the pre-trained NLI models and the templates used to construct the hypotheses.

4.1.2 Multilingual BERT

The second method ($\textbf{Model}_{\textbf{Multi}}$) consists in fine-tuning a pre-trained, multilingual language model on the expert-annotated English language data used by Assenmacher et al. (2024). We use the *bert-base-multilingual-cased* model, [3] a BERT (Devlin et al., 2019) variant that has gained remarkable multilingual capabilities thanks to its pre-training on a corpus covering a total of 104 different languages. Most importantly, previous research has shown that fine-tuning the model to a task on data from one language also leads to improved performance on the task in other languages

[2] https://translate.google.com/

[3] https://huggingface.co/bert-base-multilingual-cased

(Pires et al., 2019). In contrast to the first method, the fine-tuning approach requires the existence of annotated training data. However, our hope in using this method is that by fine-tuning the model on the high quality English training data (some of) the strong performance reported by Assenmacher et al. (2024) for detecting bot accusations in English would transfer to the other languages included here. Appendix B details our fine-tuning setup.

4.1.3 Ensemble

The third method ($\textbf{Model}_{\textbf{Ensemble}}$) is designed to combine linguistic cues as best manifested in the data originally collected in the respective language with the expert annotations available only for the English language data. First, we fine-tune language-specific, pre-trained classification models on the training datasets in the original language, annotated using the GPT-3.5 model as described above. Second, we fine-tune an English-only, pre-trained BERTweet classifier (Nguyen et al., 2020) on the expert-annotated dataset provided by Assenmacher et al. (2024). To then annotate a tweet, we apply the language-specific classifier to the original version and the English-only classifier to the translated version. Only if both classifiers indicate that the tweet contains a bot accusation do we label it as such, otherwise it is considered to not be an accusation. We hope that the combination of these two crucial aspects improves the precision of the method, with the original-language classifier catching instances where important information is lost in translation and the English-only classifier catching instances where the annotation informed by zero-shot-GPT-3.5 deviates too much from the more precise expert annotations.

4.2 Multilingual Accusation Analysis

Once the tweets collected in the different languages have been classified using the accusation detection approaches presented above, the final step is to identify universal as well as language-specific patterns in the development of the phenomenon of social media users accusing each other as bots. Our choice of methods is inspired by Assenmacher et al. (2024), but we had to find ways to apply them across nine different languages and to make the results they produce comparable.

4.2.1 Word Embeddings Over Time

We use the proximity of the term bot to other terms in a language-specific word embedding space as an indication of the usage of the term and to detect shifts in its meaning. Word embeddings capture semantic relationships based on the co-occurrences of different terms by projecting words as vectors in a shared embedding space. To identify terms most closely associated with the term bot at different points in time, we calculate the cosine similarity between the vector for bot and those of all other vectors in the different embedding spaces that we trained using Word2Vec (Mikolov et al., 2013) for each language-year combination. The embeddings cover all years from 2011 to 2023 individually, aggregating only years with insufficient data into single embeddings. Since word embeddings are non-deterministic, we report those ten nearest neighbors of the term bot that show up consistently in five different runs of the embedding model, initialized with different random seeds.

4.2.2 Toxicity Measurement

To check whether tweets containing bot accusations are generally more toxic than their non-accusation counterparts, and to track the general development of the level of toxicity of accusation tweets over time, we measure the toxicity of accusations using the pre-trained Detoxify model (Hanu and Unitary team, 2020). While these models were optimized to measure toxicity across a number of languages, they do not cover Arabic, German, Japanese, and Korean. For these languages, we measure the toxicity of tweets translated into English using the English model variant.

4.2.3 Context Clustering

Shifting focus from the accusations themselves to the contexts in which they occur, we apply unsupervised clustering techniques to the original tweets preceding the bot accusations. We use multilingual sentence transformers (Reimers and Gurevych, 2020) to transform the original tweets into document embeddings, representing the tweets' semantic contents. By using cosine similarity to measure the distance between the embeddings of the different tweets, we are able to identify clusters of tweets that are supposedly concerned with similar topics and contexts, again per language-year combination as described above.

To help us interpret the resulting clusters, we first extract the most significant tokens of each cluster via cTFIDF scores. Based on these tokens that best summarize each cluster in contrast to the remaining clusters in the same embedding space, we

	$\text{Model}_{\text{Multi}}$			$\text{Model}_{\text{NLI}}$			$\text{Model}_{\text{Ensemble}}$			
	P	R	F1	P	R	F1	P	R	F1	S%
Arabic	0.250	0.011	0.022	0.600	**0.862**	**0.708**	**0.716**	0.667	0.690	3.31
French	0.794	0.476	0.595	0.684	**0.762**	**0.721**	**0.805**	0.629	0.706	7.45
German	0.877	0.475	0.616	0.789	**0.750**	**0.769**	**0.919**	0.658	0.767	6.35
Japanese	0.500	0.048	0.087	0.400	0.024	0.045	**0.650**	**0.619**	**0.634**	2.17
Korean	**0.538**	0.125	0.203	0.344	**0.750**	0.472	0.494	0.714	**0.584**	1.53
Portuguese	0.633	0.310	0.416	0.570	**0.610**	0.589	**0.720**	0.590	**0.648**	3.59
Russian	**0.925**	0.270	0.418	0.792	**0.891**	**0.838**	0.901	0.533	0.670	14.4
Spanish	0.833	0.429	0.566	0.740	**0.771**	**0.755**	**0.919**	0.564	0.699	9.76
Turkish	0.273	0.049	0.083	0.323	**1.000**	0.488	**0.575**	0.754	**0.652**	1.13
Overall	**0.769**	0.281	0.412	0.594	**0.720**	0.651	0.747	0.620	**0.678**	

Table 2: Performance of accusation detection models across nine different languages, as well as the percentage share (S%) of accusations detected by $\textbf{Model}_{\text{Ensemble}}$ for each language. The highest value for each metric and language is emphasized in bold. $\textbf{Model}_{\text{NLI}}$ tends to achieve higher recall, while $\textbf{Model}_{\text{Ensemble}}$ prioritizes precision over recall and has the highest F1-score overall.

prompt (details in Appendix B) GPT-3.5 to assign each cluster one of the provided labels for different topical contexts. The available labels were *automated behavior*, *polarizing debates*, *insults*, and *other*. The idea of this rather superficial approach is to still get a sense of the contexts in which bot accusations occur across different languages. The proportions of clusters are then tracked across time to observe shifts in the contexts that trigger bot accusations.

5 Results

In the following, we first evaluate the performance of the different accusation detection methods, before presenting the analysis results on the accusations detected via our method of choice.

5.1 Evaluation of Accusation Detection Methods

To select the most appropriate method for annotating the full datasets of tweets collect for the nine languages and reported upon in Table 1, we compare the performance of the different methods presented above on the validation datasets annotated by crowdworkers. In Table 2, we report the precision (P), recall (R) and F1-score (F1) of the different methods, with the F1-score being the harmonic mean of precision and recall and thus representing a trade-off between these two performance indicators.

For this specific task of classifying candidate tweets, that is, tweets containing the keyword bot in any of the considered languages, into those that contain bot accusations and those that do not, a high precision means that a high share of the

	Japanese	German	Russian
2011 - 2016	tweet, person, account, statement, laugh, block, follow, response, bot, thought	automatic, easy, tweet, word, programmed, account writes, reacts, think, probably	ban, pay, russia, dick, stupid, writes, really idiot, judging, people
2022 - 2023	account, block, person, tweet, fraud, thank you, probably, polite, think, bot	ukraine, easy, putin, twitter, russia, account, propaganda, profile, actually, russian	people, russia, stupid, really, idiot, putin, idiot, russian, judging, writes

Table 3: English translations of words closest to the term bot in the Japanese, German and Russian embedding spaces. Terms associated with automation highlighted in blue and those that are insulting or from a political context in red. We see that for Japanse, the term bot almost exclusively appears together with neutral, account-automation-related other terms, both for the first and the last years of data. In contrast, in Russian the term bot appears almost exclusively in company of insults or politics and patriotism related terms. For German, we see how the meaning of the term shifted over time - in the early years, it was associated with terms related to (account) automatization, while in the later years, it appears close to five different terms related to the highly politicized Russian war on Ukraine. Lists of nearest neighbors across all languages and years may be found in Appendix C Tables 9 to 17.

tweets labelled as accusations actually are accusations, while a high recall means that many of the actually existent accusations have been labelled as such. Similar to the argumentation by Assenmacher et al. (2024), we strive to balance precision and recall, but would consider precision to be the slightly more important measure, as we build our subsequent analyses on the assumption that we are working with tweets in which other users are accused of being bots. Since our initial data collection described above was designed to be as inclusive towards candidate tweets as possible, favoring recall over precision by requiring only the presence of the keyword bot, we now deliberately select an accusation detection method that does well on the precision metric across all considered languages. These criteria are fulfilled only by the $\textbf{Model}_{\textbf{Ensemble}}$ method, exhibiting F1-scores larger than 0.58 for each language and an overall F1-score of 0.678, as well as the highest precision for seven of the nine languages covered. In the last column of Table 2, we report the share of tweets that are classified as accusations when applying $\textbf{Model}_{\textbf{Ensemble}}$ to all tweets containing the keyword bot collected for each language.

5.2 Results of Accusation Analysis

We analyze the development of bot accusations on those tweets that have been classified as bot accusations by $\textbf{Model}_{\textbf{Ensemble}}$. We carefully tried to balance considerations regarding the precision and recall in the detection method as well as to validate the classifier using human annotations as groundtruth data, but still have to acknowledge that our final datasets used for analysis likely include tweets that do not actually contain a bot accusation (false positives), and that we likely missed tweets that actually are bot accusations (false negatives). However, we are confident that the datasets presented here still allow for a good enough approximation to the phenomenon of bot accusations, especially given the difficulties of conceptualizing and implementing any data collection and processing pipeline across nine different languages.

Based on the results of the different methods used for analysis, we assign the nine languages into three different groups, such that languages within the same group broadly exhibit the same development in the use of bot accusations over the years. We structure our presentation of the results along these groups.

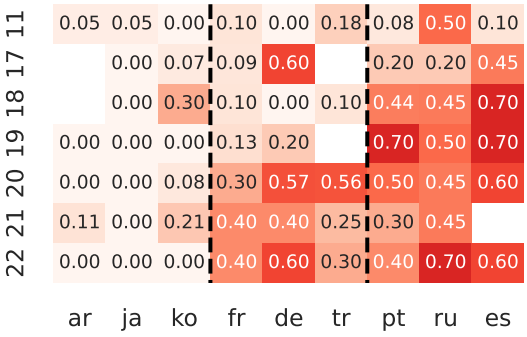

Figure 1: Share of original tweet clusters related to insults and political or politicized issues. Bot accusations in languages from Group 3 (pt, ru, es) consistently occur in contexts that are dominated by insults or discussions around political or politicized topics, which accusations in languages from Group 2 (fr, de, tr) only start doing in later years. Accusations in languages from Group 1 (ar, ja, ko) only rarely appear in these contexts.

5.2.1 Group 1 - Stable Automation

For accusations in Arabic, Japanese and Korean, we find that the term bot is consistently used in its original sense and in the context of terms related to automation and tweet technicalities (see Table 3). This continued use of the term in a non-derogatory, more neutral manner is also reflected in the toxicity level of the term in these languages, which - with the slight exception of Arabic during the years 2019 to 2022 - remains stable at a relatively low level, especially when directly compared with languages from the other groups (see Figure 3). Finally, when looking at the proportions of clusters characterized by insults and polarization (Figure 1) versus the proportions of clusters concerned with aspects of automation (Figure 2), we see that the first type of discourse only plays a minor or even negligible role in the original tweets leading up to the accusations, while automated behaviour is over the years consistently featured in the contexts of bot accusations.

We find that especially for Japanese and Korean the accusations were centered around gaming related content, for example:

> @USER 偶然ですよ。ボットだったから。私が勝たないと、負ける所だった (笑) [@USER It was a coincidence. Because it was a bot. I had to win or I would have lost lol.]

5.2.2 Group 2 - Shift in Meaning

For the languages in the second group - French, German and Turkish - we observe a pattern similar to what Assenmacher et al. (2024) report for bot accusations in English. While for these languages terms of automation are predominantly found in the vicinity of the term bot during the early years, this shifts in the years 2017 and 2018, with word embeddings in later years showing insults and dehumanizing language as well as political references much more closely associated with the concept bot (see Table 3). For the three languages from the second group, we also find a constant rise in toxicity in the accusing tweets starting around the year 2018 (see Figure 3), which is neither found in the toxicity of the languages in the first group, nor paralleled by a similarly pronounced rise in the non-accusing tweets (see Appendix C Figure 5). However, this reported shift from the term referring to technical aspects of automated behavior on the platform to an insult used in polarized and politicized contexts is best observed through the contents of the original tweets that precede the bot accusation. While Twitter users posting in French, German and Turkish discussed the concept bot predominantly in the context of automation up until the year 2019 (see Figure 2), this shifted drastically, with the years 2020 to 2023 showing a much higher prevalence of clusters related to insulting discussions and polarized debates (see Figure 1).

An important theme of politicization in this group was the alleged role of Russians in bot operations, for example:

> Brauchst nicht weiter mit dem Kerl zu diskutieren der ist ein Russen Bot.... [No need to argue with the guy, he's a Russian bot]

5.2.3 Group 3 - Stable Problematization

Similarly to languages included in Group 1, those in Group 3 - Portuguese, Russian and Spanish - do not show any significant shifts in the usage of the term bot. However, we find that the term has been constantly used with insulting and political connotations, right from the start of our data in 2011. Looking at the word embeddings for Russian in Table 3, we see that already in 2011 a number of insults are found close to the term bot, as well as references to foreign politicians or to Russia, potentially as an indicator of patriotic sentiments. This composition of terms associated with bot remains

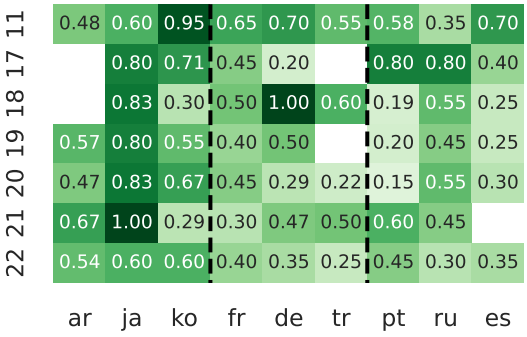

Figure 2: Share of original tweet clusters related to automated behavior. Bot accusations in languages from Group 1 (ar, ja, ko) appear over the years oftentimes in reaction to tweets that are discussing aspects of actual automation.

highly stable over the full period covered by our data. Looking at the toxicity measured in the accusations from these languages (see Figure 3), we observe relatively high levels from the beginning on, with slight increases over the full period, but no pronounced shifts as found for the languages in Group 2. Complimenting this impression, we see from the accusation contexts in Figure 1 that accusations in Portuguese, Russian and Spanish are already in the early years oftentimes found in the context of debates around political topics or in conversations that feature insulting and even dehumanizing language, much more so than languages from the other groups.

The following Russian tweet from 2012 is an early example of bot being associated both with insult as well as political motives:

> Нереально тупой бот @USER пытается пихнуть мне гламур Путинских вечеринок для гопоты. [The unrealistically stupid bot @USER is trying to shove the glamor of Putin's gopot parties at me.]

6 Discussion

In this study, we expand on existing research about social bot accusations by examining linguistic settings beyond English. We developed an ensemble of language-specific and translation-based models to detect bot accusations in nine different languages. Using this approach, we identified bot accusations on Twitter (𝕏) for each language from 2011 to 2023. Our findings reveal that the previously noted shift in bot accusations in English

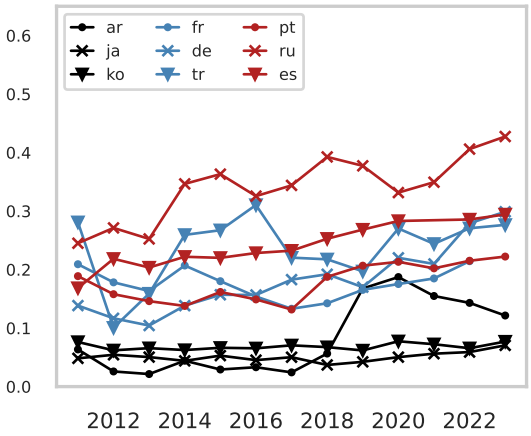

Figure 3: Development of toxicity scores for bot accusation tweets in different languages. Languages from Group 1 (ar, ja, ko) exhibit a relatively stable low level of toxicity in their bot accusations, while languages from Group 3 (pt, ru, es) exhibited (relatively) high levels of toxicity across the years. The bot accusations for the languages in Group 2 (fr, de, tr) started with low toxicity levels that and only started to permanently increase after 2018.

does not occur in every language. Specifically, discussions in East Asian languages, such as Korean and Japanese, show different patterns of bot accusations, with a stronger focus on automation-related topics, particularly in the context of gaming. In contrast, accusations related to polarizing political debates were seldom observed in these languages. On the other hand, we identified languages such as Russian, in which bot accusations were consistently associated with insults. Our findings have several implications. From a moderation perspective, it is important to understand that these accusations should not be treated equally. While it is true that we need to acknowledge that ”bot” is often systematically used as an insult, delegitimizing users' opinions and thus undermining constructive dialogue, the context and connotation of such accusations can vary significantly across different languages and cultures and, therefore, require different moderation strategies. We therefore highlight the risk of detection systems trained on English data only to fall short in generalizing to other context, emphasizing the need for diverse linguistic training to ensure accuracy and fairness.

7 Ethical Considerations

For our empirical study of the bot accusation phenomenon across a range of different languages, we are relying purely on publicly-accessible, user-generated Twitter posts. Using this type of data carries the usual privacy risks known from social media studies, which we, however, try to counteract by anonymizing all data immediately after its collection. We are not interested in studying individual-level bot accusations, but rather focus on societal-level patterns. We do not try to identify any of the individuals included in our datasets, and explicitly point out that the bot accusations we observe on the platforms are oftentimes directed towards actual human beings instead of automated accounts, as indicated both by our findings and those of Assenmacher et al. (2024). We therefore discourage anyone from trying to infer the status and degree of automation of an account from a bot accusation found in our datasets or elsewhere. In recruiting annotators and collecting their annotations, we followed the ethical considerations and best practices put forth by the platform provider Prolific,[4] including the guarantee that every annotator would receive an hourly pay equivalent (far) exceeding the required minimum pay as well as the informed consent and the possibility to withdraw from participation. To follow established practice in sharing research data collected from social media and still ensure full reproducibility and transparency of our results, we invite other researchers to contact us to mutually explore potentials for collaboration and the sharing of our collected research data.

8 Limitations

Most of this paper's limitations originate from its main conceptual and methodological challenge, the inherent multilinguality of the ambitious endeavor to study the same social media phenomenon across nine different languages. While we tried to handle this challenge as carefully as possible, we acknowledge a number of limitations that we were not able to overcome. First, and maybe most importantly, we rely on a number of pre-trained resources, particularly models. While the impressive zero-shot performances of advanced LLMs like GPT-3.5 has already been widely reported and used across a number of tasks and languages, the use of such general purpose, 'black-box' models should still be met with high attention and increased scrutiny. We tried to counterbalance our

[4] https://www.prolific.com/resources/ethical-considerations-in-research-best-practices-and-examples

reliance on the LLM's annotations in our different detection methods by ensuring that we validate the resulting classifier, by evaluating its performance on datasets that have been annotated by humans, without the mediating effects of any other methods or models. We decided to use the LLM-based method only after we observed satisfying performance relative to human annotations on the relevant task of identifying bot accusations across different languages. Regarding the use of the LLM for purposes of the (exploratory) data processing and analysis, particularly its role in labelling the different context clusters, we did not validate the model through crowd annotations, but rather relied on our own judgments and experiments. We manually checked many of the automatically annotated clusters, represented through their (translated) significant tokens, and found that the clustering decisions of the LLM were sufficiently reliable, particularly given the open-ended, exploratory nature of the task. A further limitation lies in the difficulty of reliably detecting bot accusations from tweets that are merely mentioning the keyword bot. This is already a challenging task for a single language, and even more so for nine different languages. As argued above, we tried to optimize for recall, i.e., the inclusion of as many accusations as possible, through a very broad initial data collection strategy, before then optimizing for precision, i.e., making sure that the tweets we identify as bot accusations actually are accusing other users of being a bot. We do so through our choice for the $\textbf{Model}_{\textbf{Ensemble}}$ method, the bot accusation method with the best F1-scores overall and - conceptually - the best setup to only label those tweets as accusations that are identified as such by different types of classifiers.

Related to the limitation of being unable to achieve perfect accuracy in the bot accusation detection task, we acknowledge systematic biases that might occur because of linguistic particularities of and around the term bot. For instance, the German keyword we used, "bot", does not just refer to the concept we are interested in with this paper, but also translates to "offered". Similarly, the Arabic keyword "بوت" may also translate to "boot". We hope that, again, our choice of the ensemble model ameliorates such translation issues by combining a classifier that acts upon the original tweets with another classifier that works on a translated version of it. We hope that this study, as imperfect as it might be, still helps to advance our collective understanding of this interesting phenomenon beyond the much studied, English-only part of a social media platform like Twitter.

References

Dennis Assenmacher, Leon Fröhling, and Claudia Wagner. 2024. You Are a Bot! –Studying the Development of Bot Accusations on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:113–125.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Communications of the ACM*, 59(7):96–104.

Christian Grimme, Mike Preuss, Lena Adam, and Heike Trautmann. 2017. Social Bots: Human-Like by Means of Human Control? *Big Data*, 5(4):279–293.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Daniel Kats and Mahmood Sharif. 2022. "I Have No Idea What a Social Bot Is" : On Users' Perceptions of Social Bots and Ability to Detect Them. In *Proceedings of the 10th International Conference on Human-Agent Interaction*, pages 32–40, Christchurch New Zealand. ACM.

Franziska B. Keller, David Schoch, Sebastian Stier, and JungHwan Yang. 2020. Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. *Political Communication*, 37(2):256–280.

Ryan Kenny, Baruch Fischhoff, Alex Davis, Kathleen M. Carley, and Casey Canfield. 2024. Duped by Bots: Why Some are Better than Others at Detecting Fake Social Media Personas. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 66(1):88–102.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *arXiv preprint*. ArXiv:1412.6980 [cs].

Maxim Kolomeets, Olga Tushkanova, Vasily Desnitsky, Lidia Vitkova, and Andrey Chechulin. 2024. Experimental Evaluation: Can Humans Recognise Social Media Bots? *Big Data and Cognitive Computing*, 8(3):24.

Jonas Lundberg, Jonas Nordqvist, and Mikko Laitinen. 2019. Towards a language independent Twitter bot detector. In *Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries*, Copenhagen.

David Martin-Gutierrez, Gustavo Hernandez-Penaloza, Alberto Belmonte Hernandez, Alicia Lozano-Diez, and Federico Alvarez. 2021. A Deep Learning Approach for Robust Detection of Bots in Twitter Using Transformers. *IEEE Access*, 9:54591–54601.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint*. ArXiv:1301.3781 [cs].

Dat Quoc Nguyen, Tanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Mariam Orabi, Djedjiga Mouheb, Zaher Al Aghbari, and Ibrahim Kamel. 2020. Detection of Bots in Social Media: A Systematic Review. *Information Processing & Management*, 57(4):102250.

Jürgen Pfeffer, Daniel Matter, Kokil Jaidka, Onur Varol, Afra Mashhadi, Jana Lasser, Dennis Assenmacher, Siqi Wu, Diyi Yang, Cornelia Brantner, Daniel M. Romero, Jahna Otterbacher, Carsten Schwemmer, Kenneth Joseph, David Garcia, and Fred Morstatter. 2023. Just Another Day on Twitter: A Complete 24 Hours of Twitter Data. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:1073–1081.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Desirée Schmuck and Christian Von Sikorski. 2020. Perceived threats from social bots: The media's role in supporting literacy. *Computers in Human Behavior*, 113:106507.

Shane Schweitzer, Kyle S. H. Dobson, and Adam Waytz. 2024. Political Bot Bias in the Perception of Online Discourse. *Social Psychological and Personality Science*, 15(2):234–244.

Xinhe Tian and Susan R. Fussell. 2024. Could Chinese Users Recognize Social Bots? Exploratory Research Based on Twitter Data. In Constantine Stephanidis, Margherita Antona, Stavroula Ntoa, and Gavriel Salvendy, editors, *HCI International 2024 Posters*, volume 2119, pages 146–156. Springer Nature Switzerland, Cham. Series Title: Communications in Computer and Information Science.

Magdalena Wischnewski, Rebecca Bernemann, Thao Ngo, and Nicole Krämer. 2021. Disagree? You Must be a Bot! How Beliefs Shape Twitter Profile Perceptions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11, Yokohama Japan. ACM.

Samuel C. Woolley. 2016. Automating power: Social bot interference in global politics. *First Monday*.

Wentao Xu, Kazutoshi Sasahara, Jianxun Chu, Bin Wang, Wenlu Fan, and Zhiwen Hu. 2024. A multidisciplinary framework for deconstructing bots' pluripotency in dualistic antagonism. *arXiv preprint*. ArXiv:2402.15119 [cs].

Harry Yaojun Yan and Kai-Cheng Yang. 2023. The landscape of social bot research: a critical appraisal. In Simon Lindgren, editor, *Handbook of Critical Studies of Artificial Intelligence*, pages 716–725. Edward Elgar Publishing.

A Appendix Data

A.1 Crowd Annotations

Crowdworkers are recruited on the Prolific platform [5] and selected to be native-speakers of the relevant language, fluent in English, as well as regular users of Twitter. We required these criteria to ensure both sufficient linguistic capability to reliably comply with the annotation task and to be sufficiently accustomed to the jargon and customs of Twitter. We recruited twelve annotators per language and asked each annotator to label 50 candidate tweets, ensuring that each tweet would be annotated by three different annotators. We implemented two annotation checks and retrospectively checked the individual annotations for unusual patterns, but could not find any signs of low annotator attention or suspicious annotation behavior. Annotators were informed about the contents of the study before consenting to participate and were paid the equivalent of an hourly wage of 9 GBP, significantly exceeding the minimum wage requirement imposed by Prolific (6 GBP per hour).

In the following, we present our codebook used to instruct the annotators. We created slightly differing, language-specific versions, the example below is the German version. Figure 4 shows the annotation interface and Table 4 shows the number of accusations and non-accusations per language resulting from the crowd annotations.

[5] https://www.prolific.com/academic-researchers

Annotation Codebook for Bot Accusations Study

Available Labels

- **Yes** Choose 'Yes' to indicate that the tweet contains a bot accusation, i.e., that some specific user is said to be a bot.
- **No** Choose 'No' to indicate that the tweet does not contain a bot accusation, i.e., that no specific user is said to be a bot.
- **Not Sure** Choose 'Not Sure' to indicate that you cannot determine from the tweet alone whether a specific user is being accused of being a bot.
- **The text is not written in German / I do not understand the text** Choose this label if the tweet is written in a language other than German or if you just cannot make any sense of it.

Examples for tweets that are bot accusations

- A user is directly accused of being a bot
  - "@USER you're a bot!"
  - "I am sure that this is just another bot account..."
  - "This finally proves that Elon Musk is a bot - I knew it!" [*Accusations may also include people of public interest who are clearly not bots*]
  - "@USER Of course you are a bot, otherwise you wouldn't have these laser eyes." [*Accusations may also be meant sarcastically or ironically*]
- A user is addressed as a bot
  - "@USER a name with 8 numbers? bye, bot!"
  - "@USER ok bot"
- It is indicated that the previous user in a conversation (thread) is a bot
  - "@USER ^ bot" [*On XTwitter, the ^ is sometimes used as an upward pointing arrow, pointing towards the previous user in a conversation*]
  - "Default profile pic and joined 12/2023? #botalert"
- It is put into question whether a user is a bot or something else
  - "@USER Either you are incredibly stupid or just another bot?!"
  - "@USER So you admit you are a bot?"
- It is said that some part of an user's behavior is bot-like
  - "@USER stop it with your bot tweets"
  - "@USER why are you behaving like a bot then?"

Examples for tweets that are no bot accusations

- The word 'bot' is just being talked about, no accusation is being made
  - "there are too many bots on this platform..."
  - "@USER what am i supposed to do on bot lane???" [*The word 'bot' may sometimes be used in a different context, for example gaming*]
- A user is self-identifying as a bot
  - "I am a bot!"
  - "@USER how can you be sure that I am not just another bot?"
- A bot accusation is negated
  - "@USER At first I thought you were a bot, but now I am pretty sure you actually have a brain."
  - "@USER this does not seem to be a bot to me..."

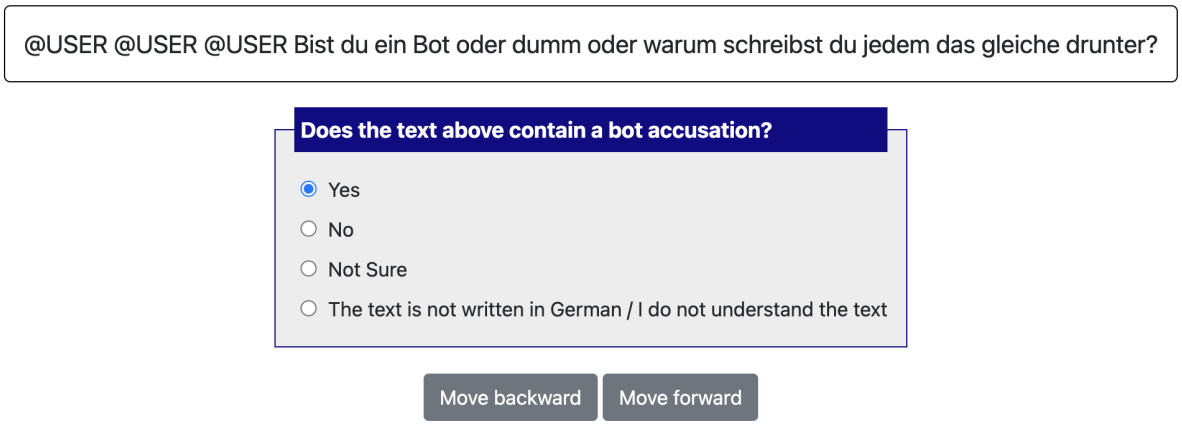

Figure 4: Annotation interface used to collect annotations from crowdworkers.

	Arabic	French	German	Japanese	Korean	Portuguese	Russian	Spanish	Turkish
accusation	87	105	120	84	56	100	137	140	61
non-accusation	113	95	80	116	144	100	63	60	139

Table 4: Number of instances labelled as accusations (**Acc.**) and non-accusations (**Non-acc.**) by annotators per language.

A.2 LLM Annotations

In the following, we show the prompt used for soliciting annotations from GPT-3.5 to annotate our training datasets. The placeholder [language] is replaced by the respective language, and [tweet] is replaced by the actual tweet to be annotated:

> Given the tweet below in [language], determine whether the user who wrote it is accusing other user(s) of being bot(s). Classify this text as "Yes" if the user is accusing other user(s) of being bot(s), "No" if there is no accusation of being a bot, or "Unclear" if it cannot be determined easily. Pay attention to the negation statement and think about it step by step. Tweet: [tweet], Classification:

B Appendix Methods

B.1 Zero-Shot Setup

Table 5 details the pretrained NLI models used for zero-shot classification in the different languages. Table 6 shows for each language the templates used as the hypothesis in the zero-shot setup, and Table 7 shows the candidate labels.

B.2 Multilingual BERT Setup

In the following, we make transparent the hyperparameters used for the fine-tuning of the pre-trained multilingual language model. This hyperparameter setup as well as the optimization algorithm, at the same time, serve as the default setup for the fine-tuning of every other model introduced above.

To fine-tune the pre-trained model to the accusation detection task, we add a dropout and a classification layer on top of the base architecture, using a 128 token input context, a dropout rate of 0.3, a learning rate of 2e-5, a batch size of 32, and an early stopping regime that interrupts training if the performance on a held-out evaluation set does not improve for five consecutive iterations. The Adam algorithm (Kingma and Ba, 2017) is used for optimization with $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$, individually adjusting the learning rates for each parameter to accommodate for low- and high-gradient parameters simultaneously. Due to resource constraints, we did not conduct any hyperparameter tuning but rather relied on a default constellation of parameters.

B.3 Language-Specific Classifiers

Table 8 details the pre-trained language models user for developing the language-specific classifiers.

B.4 Context Clustering

The prompt used to label the found clusters based on their most significant tokens is the following, with [language] and [keywords] being replaced by

the respective language as well as the most significant tokens of the cluster:

> I have clustered tweets in [language] and extracted the keywords of each cluster. Given one cluster below, if you are asked to classify it as one of the classes: automated behavior, polarizing debates, insults, and others. Which class would you assign it to and why? In addition, please translate all words into English. Return the results in json format: {{"class": "", "reason": "", "translations": ""}} List of keywords: [keywords], Class:

C Appendix Results

Figure 5 shows the toxicity over time of the original tweets, i.e., the tweets preceding the bot accusations, across different languages. Tables 9 to 17 show the ten nearest neighbors to the (language-specific) term bot over time and for each included language.

Language	Model
Arabic	MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7
French	MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7
German	MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7
Japanese	Formzu/bert-base-japanese-jsnli
Korean	muhammadravi251001/fine-tuned-KoreanNLI-KorNLI-with-xlm-roberta-large
Portuguese	MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7
Russian	MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7
Spanish	MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7
Turkish	MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

Table 5: Pretrained NLI models used for zero-shot classification in different languages.

Language	Template Hypothesis
Arabic	{} عن يتحدث النص هذا
French	Ce texte parle de {}.
German	Dieser Text handelt von {}.
Japanese	このテキストは {} についてのものです。
Korean	이 텍스트는 {} 에 관한 것입니다.
Portuguese	Este texto é sobre {}.
Russian	Этот текст о {}.
Spanish	Este texto trata sobre {}.
Turkish	Bu metin {} hakkında.

Table 6: Templates used for the hypothesis in zero-shot classification in different languages. All templates translate to the English *This text is about {}*.

Language	Template Accusation	Template Non-accusation
Arabic	روبوت بأنه المستخدم اتهام	روبوت بأنه المستخدم اتهام
French	ne pas accuser l'utilisateur d'être bot	accuser l'utilisateur d'être bot
German	den Benutzer nicht beschuldigen, Bot zu sein	den Benutzer beschuldigen, Bot zu sein
Japanese	ユーザーをボットとして非難しない	ユーザーをボットとして非難する
Korean	사용자를 봇으로 비난하지 않음	사용자를 봇으로 비난함
Portuguese	não acusar o usuário de ser robô	acusar o usuário de ser robô
Russian	не обвинять пользователя в том, что он бот	обвинять пользователя в том, что он бот
Spanish	no acusar al usuario de ser bot	acusar al usuario de ser bot
Turkish	kullanıcının bot olmadığını iddia etmek	kullanıcının bot olduğunu iddia etmek

Table 7: Templates used for the candidate labels in zero-shot classification in different languages. All templates translate to the English *accusing user of being bot* and *not accusing user of being bot*.

Language	Pre-trained Model
Arabic	Davlan/xlm-roberta-base-finetuned-arabic
French	dbmdz/bert-base-french-europeana-cased
German	dbmdz/bert-base-german-cased
Japanese	cl-tohoku/bert-base-japanese-v3
Korean	KoichiYasuoka/roberta-base-korean-hanja
Portuguese	neuralmind/bert-base-portuguese-cased
Russian	bert-base-multilingual-cased
Spanish	dccuchile/bert-base-spanish-wwm-uncased
Turkish	burakaytan/roberta-base-turkish-uncased

Table 8: Pre-trained models for language-specific classifiers.

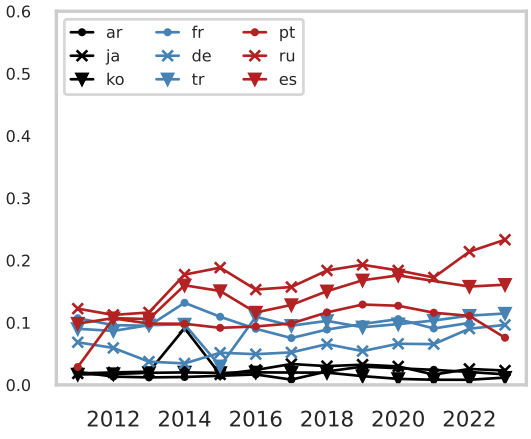

Figure 5: Development of toxicity scores for original tweets in different languages. The toxicity levels of the tweets preceding the bot accusations are generally lower and less volatile across the included languages. Strikingly, the increase in bot accusation toxicity after 2018 for languages from Group 2 is not paralleled by a similar increase in the original tweets of these languages.

Year	Terms Arabic (English Translations)
2011-2016	الله (God), تحديث (update), حسابك (account), الرسمي (official), ممكن (possible), اخوي (brother), البرنامج (program), تويت (tweet), الحساب (account), برنامج (program)
2017-2018	كه (electricity), وهمي (fake), البوت (bot), تويت (tweet), حساب (account), انسان (human), ياخي (brother), الله (God), راسك (head), تويتر (Twitter), ابيض (white), لازم (necessary)
2019	حساب (account), تلعب (play), البوت (bot), تويتر (Twitter), عادي (normal), شكلك (look), نوب (noob), ياخي (brother), طول (length), الله (God), واضح (clear), الناس (people)
2020	تلعب (play), الله (God), شكلك (look), اصلا (originally), اللعبه (game), رد (reply), حقير (despicable), شخص (person), عادي (normal), وهمي(fake), انسان (human)
2021	شخص (person), يرد (responds), حساب (account), بشري (human), وهمي (fake), لاعق (licker), الله (God), يجاوب (answers), حقيقي (real), بصراحة (frankly), اشك (doubt), شريحة (slice)
2022-2023	يرد (responds), بيس (base), كلمة (word), شخص (person), الله (God), حساب (account), انسان (human), الحساب (account), يكتب (writes), طبيعي (natural), العسكر (military), تويتر (Twitter)

Table 9: Nearest neighbors to the term بوت (bot) in the Arabic word embedding space.

Year	Terms French (English Translations)
2011-2016	compte (account), mots(words), spam (spam), répondre (respond), temps (time), monde (world), fake (fake), regarde (look), robot (robot) vraiment (really)
2017	compte (account), temps (time), phrase (sentence), probablement (probably), programmé (programmed), fake (fake), écrit (writes), réponse (response), troll (troll)
2018	propagande (propaganda), compte (account), russe (Russian), temps (time), photo (photo), merde (shit), répond(responds), mec (guy), croire (believe), voir (see)
2019	compte (account), petit (small), cas (case), troll (troll), temps (time), vraiment (really), humain (human), russe (Russian), bonne (good), jamais (never) répondre (respond)
2020	compte (account), troll (troll), temps (time), profil (profile), message (message), humain (human), vraiment (really), répondre (respond), vrai (true), fake (fake)
2021	troll (troll), compte (account), vie (life), france (France), cas (case), humain (human), propagande (propaganda), monde (world), répondre (respond), chose (thing)
2022-2023	compte (account), troll (troll), vraiment (really), créé (created), merde (shit), profil (profile), répondre (respond), fake (fake), bloquer(block), gros (big)

Table 10: Nearest neighbors to the term bot in the French word embedding space.

Year	Terms German (English Translations)
2011-2016	automatisch (automatic), einfach (easy), tweet, wort (word), programmiert (programmed), account, schreibt (writes), reagiert (reacts), denke (think), wahrscheinlich (probably)
2017	account, twitter, einfach (easy), schreiben (write), tweets, glaub (believe), follower, langsam (slow), fragen (ask), hashtag, dummer (stupid), offensichtlich (obvious), doof (dumb)
2018	hör (listen), aufstehen (get up), missbrauchen (abuse), einfach (easy), fake, account, profil (profile), antworten(answer), tweet, troll, völlig (completely), dummer (stupid), nazi
2019	antworten (answer), troll, einfach (easy), gesellschaft (society), aktiv (active), account, melden (report), beleidigen (insult), arbeitest (work), tweet, jederzeit (anytime) frage (question)
2020	troll, einfach (easy), account, leute (people), trump, fleisch (meat), blut (blood), tweets, profil (profile) automatischen (automatic) follower, russischer (Russian), dumm (stupid)
2021	troll, account, fake, tweets, leute (people), person (person), twitter, propaganda, schreibt (writes), follower, profil (profile) leben (life), antworten (answer), russischer (Russian)
2022-2023	ukraine, einfach (easy), putin, twitter, russland (Russia), account, propaganda, profil (profile), eigentlich (actually), russischer (Russian), antworten (answer), schreiben (write)

Table 11: Nearest neighbors to the term bot in the German word embedding space.

Year	Japanese Terms (English Translations)
2011-2016	ツイート (tweet), 人 (person), アカウント (account), 発言 (statement), 笑 (laugh), ブロック (block), フォロー (follow), 反応 (response), ボット (bot), 思っ (thought)
2017	アプリ (app), ウェブサイト (website), 突然 (suddenly), 体感 (body sensation), デジタル (digital), 力 (power), 証明 (proof), 問題 (problem), 基礎 (foundation)
2018	問題 (problem), 笑 (laugh), 損ない (harmless), 数学 (mathematics), 同じ (same), 人 (person), アカウント (account), 自動 (automatic), 返信 (reply), 思っ (thought), 意味 (meaning)
2019	だろう (probably), 同じ (same), 人 (person), 質問 (question), まし (better), 自動 (automatic), 名前 (name), 思う (think), ツイート (tweet), でしょう (probably), 変 (strange), ガチ (serious)
2020	ブロック (block), 無能 (incompetent), ジャンル (genre), 突っ (thrust), 推薦 (recommendation), 首 (neck), 込ん (crowded), ツイート (tweet), 迷惑 (nuisance), アカウント (account)
2021	自動 (automatic), ツイート (tweet), まし (better), たぶん (probably), 名前 (name), しれ (know), 垢 (account), 思っ (thought), 業者 (dealer) 思う (think), 人 (person), フォロー (follow)
2022-2023	アカウント (account), ブロック (block), 人 (person), ツイート (tweet), 詐欺 (fraud), ありがとう (thank you), だろう (probably), ござい (polite), 思う (think), ボット (bot)

Table 12: Nearest neighbors to the term ボット (bot) in the Japanese word embedding space.

Year	Korean Terms (English Translations)
2011-2016	진짜 (real), 사진 (photo), 패러디 (parody), 아뇨 (no), 자동 (automatic), 저건 (that is)
2017	계정 (account), 진짜 (real), 생각 (thought), 사진 (photo), 다른 (different), 사람 (person), 블락 (block), 정보 (information), 세상 (world), 독촉 (urge), 트윗 (tweet), 사실 (fact)
2018-2018	계정 (account), 사람 (person), 트위터 (Twitter), 진짜 (real), 생각 (thought), 자동 (automatic), 사실 (fact), 팔로 (follow), 알티 (retweet), 정보 (information)
2019	진짜 (real), 사람 (person), 아마 (probably), 계정 (account), 생각 (thought), 정도 (degree), 트윗 (tweet), 사실 (fact), 되어 (become), 존나 (damn), 신음 (groan), 얘기 (talk)
2020-2020	사람 (person), 진짜 (real), 계정 (account), 트윗 (tweet), 생각 (thought), 마음 (mind), 알티 (retweet), 얘기 (talk), 자동 (automatic), 있는 (existing), 트위터 (Twitter)
2021	진짜 (real), 사람 (person), 계정 (account), 트윗 (tweet), 트위터 (Twitter), 아마 (probably), 생각 (thought), 마음 (mind), 가요 (song), 봇임 (a bot), 자동 (automatic)
2022-2023	진짜 (real), 사람 (person), 트윗 (tweet), 계정 (account), 정도 (degree), 아마 (probably), 사실 (fact), 알티 (retweet), 자동 (automatic), 생각 (thought), 있는 (existing)

Table 13: Nearest neighbors to the term 봇 (bot) in the Korean word embedding space.

Year	Portuguese Terms (English Translations)
2011-2016	manda (send), puta (whore), lixo (trash), block (block), tweets (tweets), merda (shit), boca (mouth), achar (find), responde (responds), frase (phrase)
2017	mundo (world), pessoa (person), fica (stay), milhões (millions), cara (guy), fake (fake), ruim (bad), merda (shit), começando (starting), safado (naughty)
2018	cara (guy), perfil (profile), fake (fake), pessoa (person), mulher (woman), news (news), conta (account), twitter (Twitter), merda (shit), falar (speak), seguidores (followers)
2019	perfil (profile), cara (guy), governo (government), conta (account), falar (speak), merda (shit), tweet (tweet), fake (fake), twitter (Twitter), foto (photo), fala (speech)
2020	merda (shit), fala (speak), caralho (fuck), lixo (trash), ninguém (nobody), fica (stay), foto (photo), gado (cattle), presidente (president), cara (guy), imbecil (imbecile)
2021	gado (cattle), pessoa (person), merda (shit), conta (account), imbecil (imbecile), cara (guy), tweet (tweet), fala (speech), foto (photo), falando (speaking)
2022-2023	caralho (fuck), desgraçado (wretched), lula (Lula), bozo (idiot), block (block), fica (stay), país (country), merda (shit), ninguém (nobody), humano (human), lixo (trash)

Table 14: Nearest neighbors to the term bot in the Portuguese word embedding space.

Year	Russian Terms (English Translations)
2011-2016	бан (ban), судя (judging), россии (Russia), тупой (stupid), хуй (dick), пишет (writes), реально (really), дебил (idiot), платят (pay), людей (people)
2017	вероятно (probably), истории (stories), скажите (tell), обама (Obama), понятия (concepts), запись (record), президентом (president), черт (damn), идиот (idiot)
2018	тупой (stupid), внимания (attention), идиот (idiot), страшная (terrible), бан (ban), ум (mind), орать (yell), нахуй (fuck off), сша (USA), терять (lose)
2019	заплатил (paid), россии (Russia), понятно (clear), тупой (stupid), судя (judging), путин (Putin), реально (really), followers (followers), тупая (stupid), бан (ban)
2020	россии (Russia), идиот (idiot), мнение (opinion), судя (judging), понятно (clear), тупой (stupid), знает (knows), пишет (writes), хуй (dick), типичный (typical)
2021	россии (Russia), людей (people), судя (judging), идиот (idiot), тупой (stupid), слова (words), смысла (sense), ответ (answer), пишет (writes), страны (countries)
2022-2023	людей (people), россии (Russia), тупой (stupid), судя (judging), идиот (idiot), путин (Putin), дебил (idiot), русские (Russians), реально (really), пишет (writes)

Table 15: Nearest neighbors to the term бот(bot) in the Russian word embedding space.

Year	Spanish Terms (English Translations)
2011-2016	troll (troll) , real (real), peña (crowd), seguidores (followers) , tweets (tweets) , hola (hello), foto (photo), alguien (someone), puto (fucking) , programado (programmed)
2017	gobierno (government) , alguien (someone), seguidores (followers) , gente (people), ignorante (ignorant) , foto (photo), troll (troll) , perfil (profile) , pobre (poor)
2018	pobre (poor) , mierda (shit) , troll (troll) , seguro (sure), tuits (tweets) , años (years), vida (life), pagado (paid) , ignorante (ignorant) , argumentos (arguments)
2019	publicaciones (publications), año (year), siguiendo (following) , denuncia (report), unió (joined), granja (farm), boca (mouth), socialista (socialist) , familia (family)
2020	procede (proceeds), morro (kid), inmediato (immediate), interese (interest), instante (instant), power (power), masivo (massive), metiche (nosy), orgánico (organic)
2022-2023	boludo (idiot) , míseros (miserable) , ladillas (crabs , pulgas (fleas) , servido (served), seguidos (followed) , masivo (massive), entrando (entering) , inmediato (immediate)

Table 16: Nearest neighbors to the term bot in the Spanish word embedding space.

Year	Turkish Terms (English Translations)
2011-2016	salak (idiot) , belli (obvious), hesapsın (account) , bak (look), adamsın (man) , muhtemelen (probably), düşünüyorum (thinking), piç (bastard) , galiba (apparently)
2017	hesabı (account) , robot (robot) , kardeşim (brother), belli (obvious), sağol (thanks), takipçi (follower) , sahte (fake) , gerçek (real) , botsun (bot) , otomatik (automatic)
2018	beyinsiz (brainless) , şaka (joke), sıralı (orderly), hesap (account) , güzel (beautiful), çocuğu (child), kardeş (brother), cevap (answer), fav (favourite), botsun (bot)
2019	hesaptır (account) , hemen (immediately), salak (stupid) , bak (look), insan (human), orospu (whore) , takipçi (follower) , botsun (bot) , güzel (beautiful), tweet (tweet)
2020	botsun (bot) , sıstem (system) , bayanısın (lady), hesap (account) , yorum (comment) , bak (look), sanalcı (virtualist), troll (troll) , adam (man) , muhtemelen (probably)
2021	botsun (bot) , hesap (account) , troll (troll) , aynı (same), büyük (big), boş (empty), insan (human), takip (follow) , adam (man) , vatan (homeland), belli (clear), salak (fool) ,
2022-2023	botsun (bot) , hesap (account) , sahte (fake) , troll (troll) , belli (obvious), yalan (lie), tane (piece), insan (human) cevap (reply), takip (follow) , gerçek (real), profil (profile)

Table 17: Nearest neighbors to the term bot in the Turkish word embedding space.