uc3m | Universidad Carlos III de Madrid

Master in Computational Social Science
2022-2023 Academic Year

Final Master's Thesis

"Fair play? Detecting and monitoring online hate speech during the 2022 Qatar World Cup"

Álvaro Sanz Castañeda

Tutor

Iñaki Úcar Marqués

Madrid, 20/06/2023

ABSTRACT

Social Media Listening techniques have gained relevance during the last decade due to their advantages for brands and institutions in terms of gathering information and strategic decision making. In this work, we combine Data Mining, Machine Learning and Data Visualization techniques to answer our central question: does the celebration of football matches correlate with an uptick in the proportion of online hate speech?

By taking the Spanish National Team's participation in the 2022 Qatar World Cup as our case study, we retrieve more than 4 million tweets sent during more than 2 weeks and build a Supervised Machine Learning binary classifier to categorize each one as hateful or not. Our findings indicate that football games are not enough to trigger upticks in hate speech by themselves and that cultural and historical precedents are more powerful factors.

Key words: Hate speech, Machine Learning, social media, football, Spanish National Team

RESUMEN

Las técnicas de Escucha en las Redes Sociales han ganado relevancia durante las últimas décadas dadas sus ventajas para empresas e instituciones en términos de recopilación de información y toma de decisiones estratégicas. En el presente trabajo, combinaremos técnicas de Minería de Datos, Machine Learning y Visualización para responder a nuestra pregunta central: ¿Correlaciona la celebración de partidos de fútbol con un aumento en la proporción del discurso de odio en Internet?

Tomando como caso de estudio la participación de la Selección Española en el Mundial de Qatar 2022, recuperamos más de 4 millones de tweets enviados durante más de 2 semanas y construimos un clasificador binario a través de Supervised Machine Learning para categorizar cada uno de ellos como relativo al discurso de odio o no. Nuestros hallazgos indican que los encuentros futbolísticos no son suficientes por sí mismos para inducir a incrementos en la discriminación online y que los factores culturales e históricos son más relevantes.

Palabras clave: Discurso de odio, Machine Learning, Redes Sociales, fútbol, Selección Española

DEDICATORIA

A mi madre y a mi padre, que son la razón de que pueda estar escribiendo este trabajo.

CONTENTS

1. INTRODUCTION

1.1. Context

Communication has undergone massive changes during the last decades thanks to the Internet and its successive advances, which have greatly extended the use of social media among people from all countries and modified the way in which we convey and receive information (Khanday et al., 2022). Online platforms serve as an interface from which to interact with other known and unknown users and debates are continuously established around many different topics, offering the possibility to easily interchange opinions.

Twitter has set itself as one of the most popular micro-blogging services (Monti et al., 2013) from which to do this, also resulting in a huge amount of available data due to the 'digital print' users leave with each interaction: this offers enormous possibilities in terms of compiling and analyzing tweets systematically, as we will do.

However, it is widely known that Facebook, Instagram, Twitter, and many other platforms also involve risks and dangers, being the production and diffusion of online hate speech one of them (Pereira-Kohatsu et al., 2019). Hate speech is a complex concept, as there are many possible perspectives and no consensus on what exactly constitutes it, but some authors have established that it's a communicative act (Burnap and Williams, 2015) targeted towards a person or a group based on some characteristics, like "race, ethnicity, sexual orientation, gender identity, disability, religion, political affiliation, or views" (Pereira-Kohatsu et al., 2019, p.2).

As in any other social 'terrain', social media reflects preexistent inequalities and is not aseptic, as it's also been noted its tendency to foster polarization and discriminatory rhetoric (Jaki and De Smedt, 2018). We must acknowledge that hate prevention and detection is not an easy task, as many times messages lie under the idea of 'free speech' and plurality of opinions, and legislation surrounding this matter changes across different countries, but these companies have been sometimes criticized for not doing enough on limiting online hate (Khanday et al., 2022).

Online hate speech is obviously not isolated from its 'offline' counterpart: both are heavily related, as they emerge from structural factors and specific situations which foster them. The 2015 European refugee crisis (Jaki and De Smedt, 2018), the "Black Lives

Matter" movement (Waseem and Hovy, 2016) or terrorist attacks like Charlie Hebdo or the Woolwich murder (Burnap and Williams, 2015) are just some examples of "triggers" which lead to hate speech spreading. We must then acknowledge that hate has two main components: the communicative 'side' and this physical counterpart. The second one tends to cluster in time after specific events (King & Sutton, 2013) and is considered the violent expression of hateful messages. This is one of the evident reasons why we should be concerned about this topic, in order to avoid harassments, aggressions, and even murders.

The case we're analyzing in this work, football, has a long history of documented patterns of racism (Back et al., 1998), sexism and masculinity attached to it (Dunn, 2014), as well as homophobia (Cleland, 2018). Although there are reasons to think they are progressively changing (not only in football, we should say), there's also a certain debate on whether it's effectively happening or not and at what pace. Indeed, the 21st of May of 2023 there was a huge international reaction to racist scandals regarding Vinicius Junior, a Real Madrid C.F. player, that took place during matches played at the highest levels of competition in Spain, which could be interpreted as a sign of hate and sports having a connection that is far from over.

In this contribution, our main objective is to address and quantify the relationship between football events and online hate speech through Social Media Listening techniques, by trying to answer if football matches result in upticks in the proportion of online messages characterized as hateful. To do so, we gather more than 4 million tweets from Spanish football fans during the 2022 Qatar World Cup and construct a Supervised Machine Learning model that elucidates whether a tweet is hateful or not, to later perform statistical operations and construct visualizations of the evolution of hate speech throughout the event.

Another work that also combines football and hate speech detection is from Alrababa'h et al. (2021), who set up a model to identify hateful tweets, although they restrict their scope to anti-Muslim ones, to test the 'parasocial contact hypothesis'. This theory states that contact with members from minorities tend to reduce prejudice towards that minority, but always under some conditions: exposition should be repeated, the experience should be positive, and the identities should be noted (Allport, 1954; Schiappa, Gregg and Hewes, 2005).

During the 2022 Qatar World Cup, Spain played against Costa Rica, Germany, Japan and Morocco, winning only the first match, drawing the two next and losing the last, so the 'positivity' assumption is not fulfilled. We don't value 90 minutes games as repeated exposition either, so the only condition that would hold in this case is the salient group identity: these are some of the reasons why we think this theory isn't applicable to our object of study, and why our hypotheses (discussed in Section 3.1.) follow the opposite direction.

As we mentioned, hate speech has direct implications on the individuals and groups targeted by it and can even lead to physical manifestations, so building an online hate speech classifier which can successfully monitor the 'hate level' at any given moment could be a helpful tool to prevent these actions from happening. Additionally, it can help legislators, institutions and people who work with minorities understand the patterns that hate follows and at what moments it peaks, in order to implement satisfactory measures and to spread consciousness on this issue to the rest of the population.

1.2. State of the art – Text Classification

Any text classification problem can follow different directions, with gradual complexity and also different strengths and weaknesses:

1. **Lexicon-based approaches**: based on a set of predetermined words and their occurrence in texts.
2. **Machine-Learning approaches**: based on language models which classify them according to some features.
3. **Deep learning approaches**: also based on models, but with far more complicated features which also capture the semantic relations and word embeddings (Pereira-Kohatsu et al., 2019).

While the first one is helpful for Sentiment Analysis tasks, such as the one conducted by Lingiardi et al. in which they use this semantic content to later map hate speech in Italy (2019), and the third is more precise and accurate, we'll focus on the second.

There's an enormous set of possibilities when trying to build a predictor of any kind, such as what model to use (Random Forest, Naïve-Bayes, Logistic Regression, Support Vector Machine, Neural Networks…) or which features to employ (Bag-of-Words, TF, TF-IDF, unigrams, word n-grams, character n-grams…). This set of debates have been already

compiled by authors like Pereira-Kohatsu et al. (2019) or Ayo et al. (2020) in the context of hate speech, reaching a common consensus: that each combination involves some advantages and disadvantages. The former also includes a working 'module' which adequately gathers all the necessary steps in order to build a successful classifier: data acquisition, data labeling (if the ML problem is Supervised), preprocessing and representation and construction of the classifier itself.

However, there are certain differences with the output that the models give: while some investigations just train and test them to observe their performances and propose enhancements for future contributions, others employ them to later classify a much larger set of tweets to answer some research questions, such as the one already described in Section 1.1 regarding Mohammed Salah, through its Diff-in-Diff design (Alrababa'h et al. 2021), the contribution by Ristea et al. to then compare their patterns with 'real-life' crimes around stadiums (2019) or to build regressions in order to conclude which variables are more related to the hate spreaders (Burnap and Williams, 2015).

Additionally, some other researchers have built and applied successive classifiers, depending on their needs, to sort tweets by their values for each category. Monti et al. collected more than 35 million tweets to which they applied a triple categorization regarding if they were political, if they were negative and if they were general in order to model political disaffection (2013), while Basile et al. also performed multiple classifications to distinguish between hate speech or not, the 'range' of their target (general or specific) and if the tweet was aggressive (2019).

2. METHODOLOGY

2.1. Operationalization and hypotheses

Before describing our set of hypotheses, it is necessary to adequately define what is hate speech and which characteristics define it, as it will later be a central part of stages like Tweet annotation. From now on, we will be using the definition given by Waseem and Hovy, as it gives a clear set of possible orientations of the messages classified as hateful:

"A tweet is offensive if it

1. uses a sexist or racial slur.
2. attacks a minority.
3. seeks to silence a minority.

4. criticizes a minority (without a well founded argument).
5. promotes, but does not directly use, hate speech or violent crime.
6. criticizes a minority and uses a straw man argument.
7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.
8. shows support of problematic hash tags. E.g."#BanIslam", "#whoriental", "#whitegenocide"
9. negatively stereotypes a minority.
10. defends xenophobia or sexism.
11. contains a screen name that is offensive, asper the previous criteria, the tweet is ambiguous
(at best), and the tweet is on a topic that satisfies any of the above criteria." (2016, p.89).

As previously mentioned, we will be taking as case study the Spanish National Team participation in the World Cup: this includes the 3 matches played in the groups stage and the first knockout round. In order to have our data contextualized, the temporal frame will be set from 3 days before the first match until 3 days after the last one, which leaves 20th of November of 2022 and 9th of December of the same year as the limits. The objective with this approach is to have a 'basal' level of hate the days before with which to compare to the time after them and to be able to capture the potential lasting effect days after. In that period, Spain played 4 matches:

• 23/11 vs. *Costa Rica*

• 27/11 vs. *Germany*

• 01/12 vs. *Japan*

• 06/12 vs. *Morocco*

For each of them, we'll consider tweets 2 hours prior to its start, 2 hours for the match itself and 2 hours after the end (3 hours in the case of Morocco, as it went to overtime and penalties) as our objective of study. Only using observations sent during the game itself would be reductionist, as the diffusion of hateful messages may not only be limited to the time between the first and last whistles.

Our dependent variable will be the percentage of hateful tweets relative to all the tweets sent during that span of time, and our two main hypotheses are the following:

H1. Moments before, during and after the celebration of football matches, the proportion of tweets categorized as hateful increases significantly in comparison to time spans when no matches are being played.

H2. This effect increases significantly when the supporter's team does not get a win.

2.2. Data Collection & Annotation

The first logic step to start our work is to construct the dataset with which we will later try to extract conclusions from. To achieve this, we performed a Data Mining process interacting with the Twitter Academic API which resulted in about 4.2M tweets (each one with information regarding their author, the time when the tweet was sent, its language, the users mentioned…).

The extraction process followed a two-way schema to ensure capturing as much football conversation as possible, which consisted of:

a) **Query search**: searching by the names, surnames, and Twitter handles of all the football players involved, by the countries played, their demonyms and national teams accounts and by some general football terms, applying both geographical and language filters.
b) **Semi-random users** from @SEFutbol: we retrieved information for the Spanish National Team 2.7M followers, to later apply some filters regarding number of tweets sent and creation date in order to avoid inactive users and bots. With the resulting users who met the criteria, we ran a random sampling process of 30.000 of them and we retrieved all their tweets between the time span delimited.

As our objective is to run a Supervised Machine Learning task, it is compulsory to have some tweets already tagged as hateful or not in order to build the training and testing splits for the model selection. The percentage of hateful tweets in regular conversational topics, such as the one we are working with, is not usually very high (this doesn't mean that it's not important), so we face an unbalanced classes problem. To address this challenge, we again combined randomization (2000 random tweets) with specific terms searching, such as common racial, homophobic or sexist slurs, sampling some of the tweets containing them (3000 in total) and binding them all in a single dataframe.

We generated a ‘hate’ variable and manually annotated them as hateful (1) or not hateful (0) according to the criteria expressed in Section 3.1. The final distribution of the 5000 tweets ended up like this:

Not hateful	Hateful
3932	1059

Fig. 1. *Distribution of manually annotated tweets*

With this process we tried to ensure that there were sufficient hateful tweets for further algorithms to be trained with, but not too many of them, as our objective of study is essentially an unbalanced distribution and it could bias these models in the next steps.

2.3. Data Preprocessing & Feature engineering

Once we already have our set of tweets with the desired variables, we must adapt the data to our specific needs for Text Classification. This includes transforming the texts from the tweets using Data Wrangling techniques and creating some variables when they meet some criteria.

In first place, we remove duplicated tweets (we may have captured the exact same tweet two or more times, when applying different queries, if they met two different conditions, or the same text repeated over time by an automatic account) and generate binary variables for Mentions, URLs and Hate terms presence.

We perform additional text transformations, starting from lower-casing all tweets, removing symbols and special characters like “\n”, “@” or “#” and transforming certain emojis to text, so that the algorithms then capture all of them depending on their emotions. Twitter users were left as text without its handle as we thought they could be a nice predictor for cases of famous football players when being massively attacked in social media for a determinate reason.

Emojis	Transformation
🤣😂	“tokenrisa”
🐌	“tokenrata”
👿🔪🤬👊🪓🤮💩	“tokenenfado”

👏💪👊💪❤️😻❤️♡	“tokenpositivo”

Fig. 2. *Tokenization of emojis process*.

Other emojis or non UTF-8 characters have been removed, as they offered little value to our analytic purposes. The next step in this process was to add to “tokenrisa” all possible Spanish ways of writing laughs, for which we employed an online consulted Regular Expression which captured a vast majority of the combinations, and even some English ones.

2.4. Document Features Matrix & NLP

Text Classification tasks differ substantially from other predictive approaches in the input given to the algorithm: while for traditional classification or regression predictions variables they tend to be numbers and factors, in this case the features are all the words used in the set of tweets. Manually building a dataframe with all words used in our 4,2M tweets isn’t feasible due to computational constraints maybe, but also because it’s not practical nor efficient.

Instead, we converted our set of tweets to a ‘quanteda’ corpus object (Benoit et al., 2018), taking as input each tweet as a document and the already generated features as document variables. This allows us not only to have the ‘bag-of-words’, but also those additional factors created in the previous Section.

However, the corpus object cannot be introduced as input to the algorithms, so we need to transform it to a ‘Document Features Matrix’ (‘DFM’, from now on), a kind of intermediate step between the dataframe and the ‘quanteda’ default object. When constructing it, we’ll split each tweet in tokens (minimal analysis units), while also removing punctuation signs, numbers and separators. In our case, we will be using unigrams (words) as our tokens’ unit: character n-grams were computationally expensive for such a large set of texts and word n-grams (two, three or more words altogether) gave worse results in terms of predictions.

After splitting all tweets in single words, we removed stopwords from the “ntlk” Python package (Bird et al., 2009) list for the Spanish language, to reduce the inputs given to the models which offer low or no information. We also performed an ‘stemming’ process to the words remaining: ‘stemming’ a word means reducing it to its lexical root, so that

words coming from the same base are computed as equal features. This will improve modeling performance and reduce unnecessary sparsity in the 'DFM'.

Once all these previous steps are done, we transform the corpus into 'DFM', computing the TF-IDF for each feature appearing: The TF-IDF goes beyond mere word frequencies and also measures the term's importance relative to the other documents. Even if we previously removed stopwords, we apply trimming to only leave words present in at least 5 tweets (2 in the training and testing splits, as the total amount is much lower), to avoid misspellings or other strange terms from biasing the analysis.

2.5. Model building

Once the previous operations have been applied to both the full dataset and the manual annotated subset, we search for the model that will later classify all tweets as hateful or not. To do so, we firstly use the 'set.seed()' function followed by a random split of 4500 tweets to train and 500 to test so that we obtain the same divisions in any execution of the code.

Giving to the 'caret' package (Kuhn, 2008) a 'quanteda' object as input has a particularity: we need to create a common 'DFM' with matched features between the train and test splits. The TF-IDF values will be 0 for those terms not present in the testing dataframe, but this is compulsory to give the model the same input (this step will also be necessary in Section 3.5.). Additionally, we create an object with the actual class of each document/tweet, so that we can later check the model's performance.

A 5-fold cross-validation process was performed with successive models, setting each one's specific parameters and setting "ROC" as the metric to tune the hyperparameters to. Algorithms like 'Random Forest', 'Logistic regression' or 'Neural Networks' were very computationally expensive in this case and were discarded because of the need to later classify millions of cases.

The algorithm with best performance, both in terms of Area Under the Curve and time, was 'XGBoost'. We predicted with the trained model setting the output as probabilities, not the class directly, so we could set the threshold that better fits for our purposes and that correctly balanced False Positives and False Negatives, with a final value of 0.3. This means that if the predictive task assigned a tweet a probability of more than 0.3 of being hateful, it would be classified as so.

The performance the model is shown on Figure 3, as well as the Confusion Matrix just below, and they reflect the following metrics and results:

Metric	Value
Accuracy	0.8598
Balanced Accuracy	0.7727
Sensitivity	0.6168
Specificity	0.9286
Precision	0.7097
Area Under the Curve (threshold independent)	0.8865

Fig. 3.1 *Model's performance indicators*.

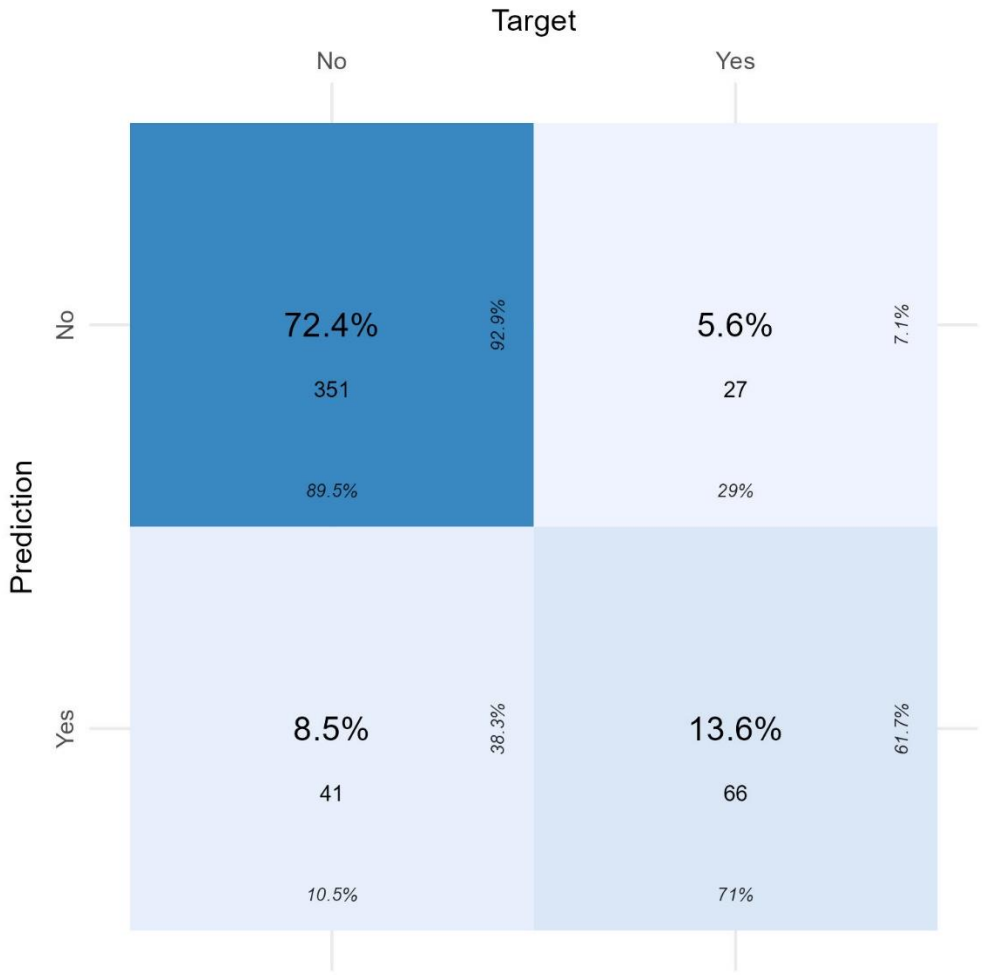

Fig. 3.2 *Model's Confusion Matrix*.

In addition, we constructed another corpus with two additional dataframes of 4196 and 4500 already annotated tweets, offered by Basile et al. (2019) and Pereira-Kohatsu et al. (2019), respectively, coming from previous works of online hate speech detection using Supervised Machine Learning techniques. However, the performances of the model didn't improve, and the measured metrics were all worse, which gives us a valuable clue: the annotation part is critic and very context-dependent, as hateful tweets captured from other temporal frames and from different conversations may bias the algorithms and make the performance decrease.

2.6. Full dataset classification

Once the model has been tested and its performance is proved, we advance to the next phase: applying it to our whole set of tweets, to give to each of them its hateful or not label. As specified in previous sections, the pre-processing phase was exactly the same except for the 'trimming' function, in which we only left features present in 5 or more tweets to avoid misspellings or uninformative words.

When trying to run the model for the more than 4 million tweets, we faced some obvious computational constraints regarding the available RAM memory of our domestic laptop: the size of vector that R Studio was trying to allocate was 161,2Gb. To avoid this problem, we constructed a loop in which successive iterations of 50000 tweets were classified using the model, with the following sequence:

1. Before setting the loop, keep open the training DFM, establish a batch size of 50000 tweets and create an empty dataframe in which we'll store all predictions, with two variables ("No" and "Yes").
2. We extract the 50000 corresponding tweets for each iteration from the full DFM, with all the tweets, into another DFM (from now on, 'dfmat_matched').
3. We match the features between all the tweets contained in 'dfmat_matched' and the training DFM, so that the model has that input too (if not, it would fail).
4. We predict using our pre-loaded model ('XGBoost') with the DFM resulting from step 3 as new data, specifying that the output should be probabilities too.
5. We store these predictions into the empty dataframe from step 1 and repeat successively binding all the rows. As each iteration of 50000 does not apply random sampling, but instead takes tweets in order from the beginning to the end, there's no problem identifying each prediction with its tweet.

Once this loop is over, we have an ‘all_predictions’ dataframe with the same number of rows as the initial dataframe. To associate each tweet with its final label, it’s necessary to generate a ‘hate’ factor variable with value “Yes” when the “Yes” probability is higher than the set threshold, 0.3, and we bind this column with the initial set of tweets.

Hateful	Text
Yes	[1] "Esta arbitrando un moro co a mi q no me mientan"
Yes	[2] "Japoneses conchasdesumadre hijos de la gran puta radioactiva"
Yes	[3] "@alizariohi98 @TalebSahara Escoria analfabeta marroquí"

Fig. 4. *Random sample of hateful tweets*.

2.7. Visualization & Analysis

The final step from our Working Pipeline is the Visualization of the data obtained and its analysis through statistical methods to address our initial hypotheses. We’ve employed the ‘ggplot2’ R package to construct two graphics:

- A general one, in which the whole evolution relative to the whole temporal period is shown.
- A specific one showing the evolution in the Moroccan’s case, in order to distinguish if there are some types of events that triggered upsides in hate speech or if there’s no clear pattern.

To answer our hypotheses, we made use of Inferential Statistics through successive t-tests, which compare specific time spans against the rest of the World Cup and the analyzed matches between them. All code employed and described in the previous sections is available online at a GitHub repository for replicability and transparency purposes (Sanz, 2023).

3. RESULTS AND DISCUSSION

3.1. Visualizations

In Figure 5 we observe the overall level of hate speech between the 20th of November and the 10th of December, period in which Spain faced the four shown games. The first thing we observe is that there's a stable 'basal' level around the 1,7% mark, approximately, for the first 3 matches, with a normal degree of variance and no remarkable upticks or downfalls. In this period, we also find the lowest point of online hate when Spain was winning against Costa Rica by a large margin, which is reasonable: under favorable circumstances, aggressiveness towards other identities is not so prone.

Variance for the first 3 games	Variance for the full period
0.11	0.38

Fig. 5. *Variances for the time series*

The next two games, against Germany and Japan, don't show any peak of hate or specific situations in which there was a relevant increase either, and the maximum proportion of hateful tweets rounds about the 2% mark for some moments. These matches ended up in draws, and while the values are slightly higher in comparison with Costa Rica, we can't assume it's a consequence of the worse result.

Nonetheless, this stable trend below 2% starts changing during the days after the third result, when Spain has already classified to the eliminatory phase. For the first time, the 3% mark is surpassed, which may be due to the anti-Asian hate or due to the rivals for the first round being known, as Morocco was the opponent. In both cases, we observe how the trend changes to an increasing tendency for the first time, and how it's held in time (not a product of momentarily variations).

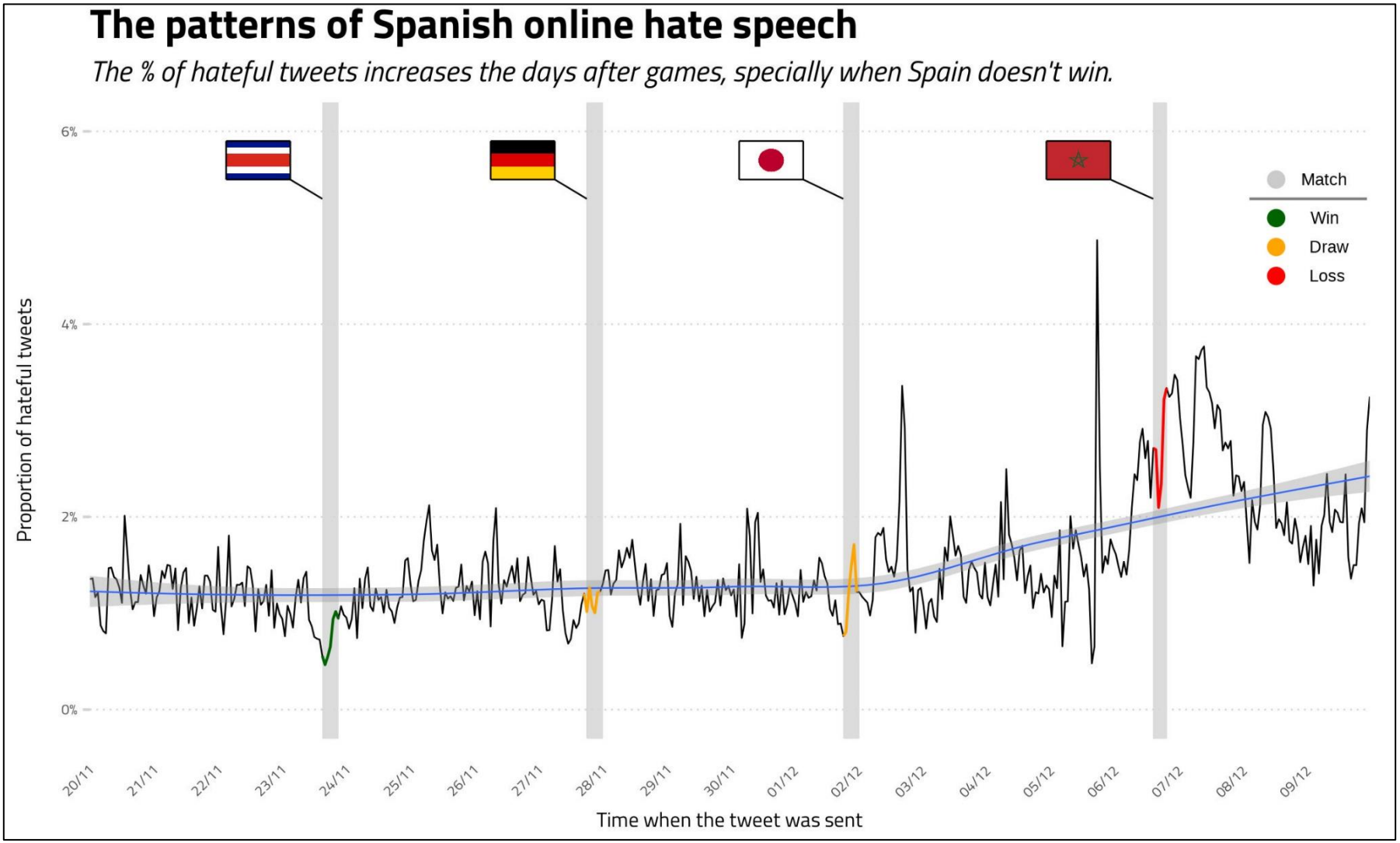

Fig. 6. *Hate speech levels visualization for the entire period (20/11/22 – 09/12/22)*

The days before the first round this upward tendency results in the maximum value of hate speech for an hour, almost in the 5% mark, and the proportion keeps growing until the match. When Spain loses, values round the 3,5% mark (more than the double compared to the previous 'basal' levels), and for a couple of days they stay near that zone. In the last moments we capture, we reach a sort of 'cooling' phase where the proportion slowly returns to 2%, with additional upticks at the end of our time series which we can't further analyze as we do not have the data (it could be motivated by some event or just product of variability).

If we look further into detail of the game against the Moroccan National Team, available in Figure 6, we can also observe part of that 'preparation' phase, in which hateful users may be setting the terrain for targeting, in this case, Moroccan people: this may have an effect among other hate spreaders, who could join the conversation instigated by the ones who started the diffusion, also contributing to the hostile climate prior to the game.

When the match starts, at 16:00, there's about an hour and a half in which the level falls back to the 2% mark, and even lower; however, this is only the previous step to another significant uptick, probably because of the arrival to penalties and the 3-0 defeat. After that, the increase remains constant and scratches the 4% of tweets being hateful, which, considering the volume we handle (more than 4 millions), is not a negligible amount.

Looking at the 'general picture' of that day, the phases could be defined as:

1. *Preparation phase*: Some users start sending hateful tweets, progressively increasing the amount and involving other users, of which a certain amount will also join the conversation.
2. *U-Shaped phase*: Phase 1 ends approximately at the start of the game, even some moments prior to it, and the debate is not based on hate for some time. This is the time span when comments strictly related to the match itself may be sent. After this, the result starts to be adverse and hate increases again, up to some of its highest values.
3. *Sustained hate phase*: The final values of the U-shaped phase are held for many hours, with hateful users being very active and engaging others into the spiral. This is the phase when the targeted groups or persons are more vulnerable to attacks or aggressions.

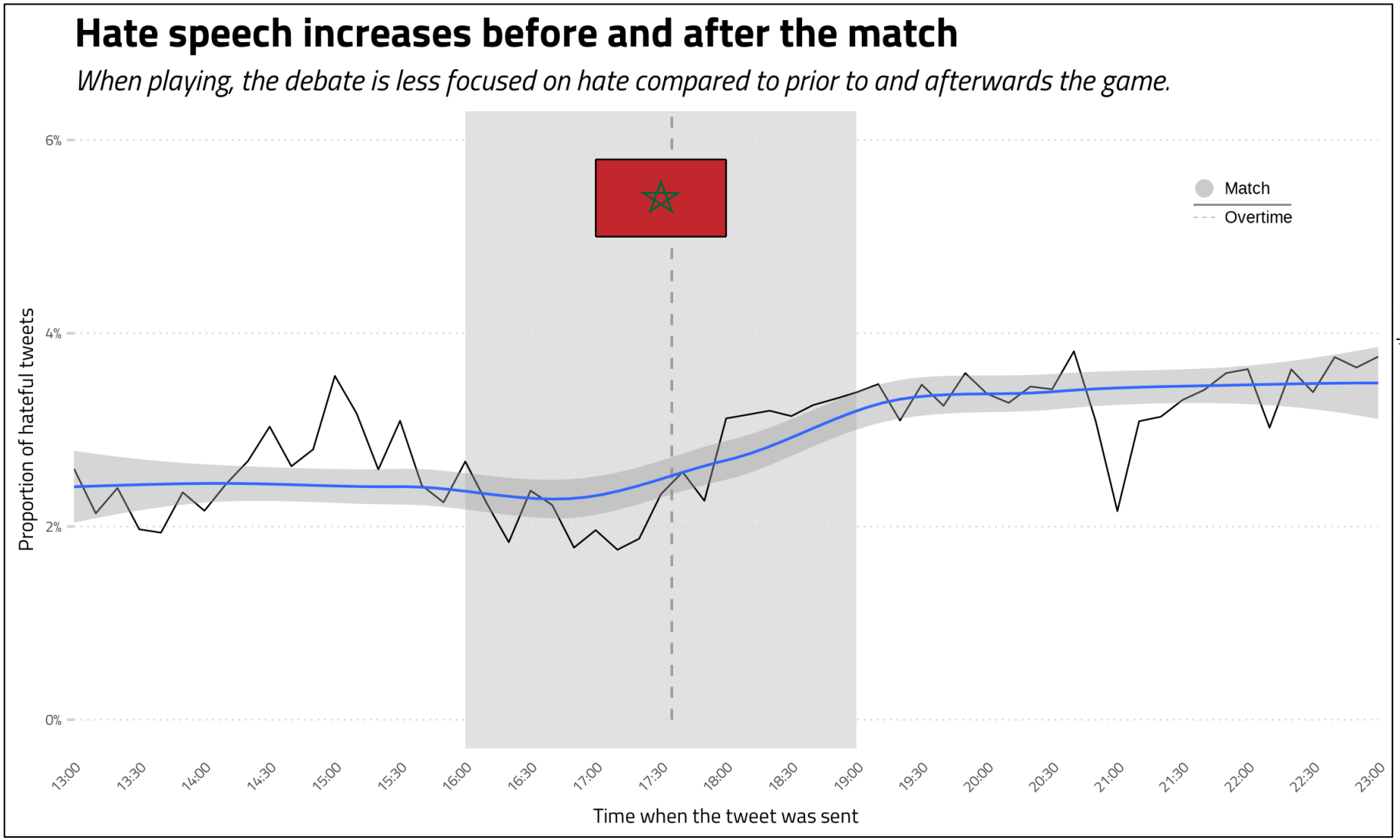

Fig. 7. *Hate speech levels visualization for the 6th of December (13:00 – 23:00)*

Nonetheless, all interpretations developed have been based on the plots, which is not enough. In Section 3.2., we'll aim to statistically confirm or reject the proposed hypotheses.

3.2. Statistical analysis

In Figure 7 we're able to see the results of successive t-tests comparing the mean proportions of hate speech for different time moments, to assess through inferential statistics if there are differences or not.

The first row compares moments when one of the matches is being played, including some hours before and some hours after it, with the rest of the available moments from the time series: the alternative hypothesis is that when playing, the hate speech level is higher than when not. However, we can't reject the null hypothesis, as the P-value is not lower than 0.05 or less, which means that we don't have any evidence of matches being moments of hate concentration.

Comparisons of the proportion of hate speech (T-Tests)

Test	Estimate	Mean X	Mean Y	P-value	H1	Conf. Level
All matches	-0.1041	1.402121	1.506230	0.751	greater	-
Morocco	1.2491	2.733601	1.484544	3.456133e-07	greater	***
After the matches	0.3368	1.766782	1.429993	5.028355e-07	greater	***

Fig. 7. *Successive T-tests for time spans shown on the "Test" column.*

We included a second test comparing difference in means for the game against Morocco with the rest of the time, and we do find support for the hypothesis that holds that there was a bigger level of hate in Twitter during that specific match, with a confidence level of 99%. Additionally, we wanted to check whether the 24 hours after sportive events also held a higher level of online hate compared to other moments, which also shows statistically significant results (99% of confidence) and a positive estimate.

Secondly, in Figure 8 the results for an ANOVA Test comparing all matches between themselves are shown: through this technique, we will compare not only the game in

which Spain was disqualified against the rest of the time, but also against the other matches.

ANOVA Test for all comparisons between matches						
Team 1	**Team 2**	**Estimate**	**CI Lower**	**CI Higher**	**P-value**	**Conf. Level**
Costa Rica	Morocco	2.0019	1.50937	2.4944	7.42e-10	***
Costa Rica	Germany	0.4113	-0.06186	0.8845	1.04e-01	-
Costa Rica	Japan	0.4699	-0.02259	0.9624	6.50e-02	-
Morocco	Germany	-1.5906	-2.08308	-1.0980	4.86e-08	***
Morocco	Japan	-1.5320	-2.04308	-1.0208	1.75e-07	***
Germany	Japan	0.0586	-0.43392	0.5511	9.87e-01	-

Fig. 8. *ANOVA Tests for all teams Spain played against.*

There are statistically significant results for all games in which Morocco is compared to, with a confidence level of 99%, and always coefficients showing that the Moroccan community was attacked in a higher proportion than the rest of the fanbases: this effect is of 2 percentage points against Costa Rica, and of about 1,5% against Germany and Japan.

All tests show a high statistical confidence for Morocco's match being the time unit when the Internet hate levels rise, as well as a shy but positive coefficient for the day after matches. This second finding could indicate that hostile comments and perceptions against nationalities, ethnic groups, sexual identities, and more, are slightly more prone to occur after games. Our first hypothesis focused more on the (approximately) 90 minutes when teams faced each other, as well as immediately close hours, but there's no evidence that suggests that. Tweets during the games could be more focused on the game itself and on the different events that happen within it, rather than in targeting someone or something to spread hate against, with posterior hours or days concentrating part of this speech.

4. CONCLUSIONS

4.1. Findings and interpretation

The goal of the present work was to assess the potential correlation between football matches and online hate speech, particularly in Twitter, taking as a case study the Spanish participation in the 2022 Qatar World Cup. Once the more than 4 million collected tweets were adequately cleaned and preprocessed, we constructed a machine learning classifier which allowed us to annotate each one of them as hateful or not. This also allowed us to later build graphical visualizations of the whole time series, observing peaks and trends, and to perform inferential statistical analysis to correctly compare the mean proportion of hate speech for some specific moments.

The results show that matches themselves are not enough to trigger the escalation of online hate speech, and that not winning the match isn't either: instead, we found that cultural and historical backgrounds are more relevant and the factor that later may convey a bad result as a precedent for targeting a specific identity or collective (although we were not able to discern what part of the variability is due to the loss and what part due to the rival). Moroccan immigrants, and the Moroccan identity itself, have been historically targeted by racist behaviors and attitudes by a part of the Spanish population, and it clearly reflects on the results obtained.

Non-favorable conditions might act as an incentive for racist, homophobic or sexist discourses to be shared, due to a perceived 'grievance' by the hate perpetrators. It's possible that the consecutive draws against Germany and Japan were not as meaningful as this loss, as Spain still scored one point in each one of them and historical patterns of discrimination against Germans are non-existent in Spain, while anti-Asian feelings are present but not as highly developed as anti-Moroccan and Islamophobic ones.

These two components, the cultural and the sports-related one, may interact in reality, being the second the one that conveys the first: as the outcome is not satisfactory for the hate spreader, and it comes from a perceived 'enemy' or 'opponent', it serves as the pretext from which to send a hateful tweet against it, or even committing crimes. One of the most controversial topics after Morocco's victory was the celebration from their fans in Spanish streets, seen as not legitimate and as provocative by hate-holders, which may also be the reason why hate speech held high levels up to the end of our temporal frame.

As 1) my team has lost and the other team's supporters are celebrating, 2) I make use of racist stereotypes and fallacies to justify my disagreement with it.

Additionally, we also discovered that the hours after each match are more prone to have increases in the relative frequency of hateful messages, although with a much more moderate effect than the former finding. This could be sign of a clustering effect of hate, as Burnap and Williams expressed (2015).

Our discoveries prove that Internet is not an aseptic terrain, and that it adequately reflects pre-existing social patterns and inequalities. There's a stable level of hate around the 1,6% mark, with the higher peaks reaching up to the 4% mark, which is concerning when viewing it in absolute terms: each hurtful message is belligerent towards (at least) someone and may be followed by violent events or crimes. Monitoring hate speech is a helpful task to develop public policies that consider at what moments it significantly increases and its escalation and de-escalation patterns, as well as for authorities, who with enough resources could visualize in real time peaks and orient their efforts in consequence.

4.2. Limitations

During the development of this Thesis, we faced several challenges: firstly, computational constraints that prevented us from developing more features (word embeddings, through *word2vec*, for example) or employing more complex algorithms (such as Random Forest or Neural Networks), which may have given us better results at the testing phase and therefore more accuracy when predicting the whole dataset. Using unigrams as our minimal unit is a methodological decision that limits the sophistication of the model, as it doesn't capture well relationships between words and their usage inside tweets' sentences.

We should also be aware that in (almost) any extraction strategy regarding Twitter data there's a certain degree of 'noise' that we can't avoid including. When performing queries searches, for example, we may be capturing some messages which are not football related, as well as others that are not sent from Spain (just in the case of the language filtering); the best way to prevent this from happening excessively consists in setting strict filters and being careful with the accounts from which the information comes (to exclude bots, for example) and in being very specific when building the queries which will interact with

the Twitter API. Manually annotating the tweets to introduce inputs for the model may have also biased it towards the author's perception of what is hate and what is not, as well as using semi-filtered tweets searching for specific slurs might have oriented it towards some forms of hate rather than others.

In general terms, we should be aware that our case study is based on just 4 games in a very specific context, and that we could have even expanded more our temporal limits to capture further phenomena, or even have chosen more squads than just the Spanish National Team (although temporal and computational constraints influenced these decisions too).

Regarding the analytical part, we're aware that the successive T-Tests and ANOVA procedures belong to inferential statistics and that they offer limited possibilities. The scope of this Thesis was not based on constructing models from the gathered data, rather than developing visualizations and some statistical approaches to answer our hypotheses, but in a context where there's such a clear temporal dependency other methods could be more valid.

4.3. Future work

As we face the mentioned limitations in Section 4.2., there are some possibilities for further research regarding online hate speech analysis in order to overcome them. New approaches could focus more on modelling the classified tweets, by creating other features depending not only on internal aspects (time of the tweet, interactions…) but also on contextual factors, or on studying if there's a clustering effect on hate spreading or not. The temporal factor is another key aspect and advanced methods oriented towards modeling time series seem like a reasonable path to take too.

In order to reduce the noise problematic, one possibility is to build an initial classifier which successfully discerns between football-related tweets and those who aren't, so that further machine learning models only receive adequate data. As well as this 'pre' algorithm, it's also a possibility to build a 'post' model which classifies hateful tweets according to the game they belong to: this would be key in order to know to which specific rival or match each tweet is referring to.

Additionally, depending on more than one annotator is recommendable, as manually classifying tweets is essentially a subjective task and one tweet may have several possible

interpretations. Establishing a determined threshold of agreement among annotators to consider a message as hateful could be a simple way of overcoming this difficulty and preventing biases from being too prominent.

Regarding the case study chosen, possibilities are huge in terms of redirecting. We could have expanded the time span to also capture matches prior to the groups phase or broadened our scope so we could capture more games and hate in other languages and from other places of the world, in order to compare behaviors too. The World Cup is a relatively short competition, but domestic leagues offer many more games across a year, which may also be helpful in terms of unveiling patterns and successfully monitoring hate speech. At the end of the day, the better we perform this task, the safer citizens can be if adequate measures are taken by legislators and authorities.

5. BIBLIOGRAPHY

Allport, G. W., Clark, K., & Pettigrew, T. (1954). The nature of prejudice.

Alrababa'H, A., Marble, W., Mousa, S., & Siegel, A. A. (2021). Can Exposure to Celebrities Reduce Prejudice? The Effect of Mohamed Salah on Islamophobic Behaviors and Attitudes. American Political Science Review, 115(4), 1111-1128. https://doi.org/10.1017/S0003055421000423

Back, L., Crabbe, T., & Solomos, J. (1998). Racism in football: Patterns of continuity and change. Fanatics, 71-87.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., Rosso, P., & Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. Proceedings of the 13th international workshop on semantic evaluation, 54-63.

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). "quanteda: An R package for the quantitative analysis of textual data." _Journal of Open Source Software_, *3*(30), 774. doi:10.21105/joss.00774 <https://doi.org/10.21105/joss.00774>, <https://quanteda.io>.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & internet, 7(2), 223-242.

Cleland, J. (2018). Sexuality, masculinity and homophobia in association football: An empirical overview of a changing cultural context. International Review for the Sociology of Sport, 53(4), 411-423.

Dunn, C. (2014). Female football fans: Community, identity and sexism. Springer.

Jaki, S., & De Smedt, T. (2019). Right-wing German hate speech on Twitter: Analysis and automatic detection. arXiv preprint arXiv:1910.07518.

Kassimeris, C., Lawrence, S., & Pipini, M. (2022). Racism in football. Soccer & Society, 23(8), 824-833. https://doi.org/10.1080/14660970.2022.2109799

Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., & Malik, S. H. (2022). Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. International Journal of Information Management Data Insights, 2(2), 100120.

King, R. D., & Sutton, G. M. (2013). High times for hate crimes: Explaining the temporal clustering of hate-motivated offending. Criminology, 51(4), 871-894.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M., & Brena, S. (2020). Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. Behaviour & Information Technology, 39(7), 711-721.

Monti, C., Rozza, A., Zappella, G., Zignani, M., Arvidsson, A., & Colleoni, E. (2013). Modelling political disaffection from Twitter data. 1-9.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in Twitter. Sensors, 19(21), 4654.

Ristea, A., Langford, C., & Leitner, M. (2017). Relationships between crime and Twitter activity around stadiums. 1-5.

Sanz, Á. (20 de julio de 2023). Football and hate speech [GitHub Repository]. Github. https://github.com/alvarosc99/Football-and-hate-speech

Schiappa, E., Gregg, P. B., & Hewes, D. E. (2005). The parasocial contact hypothesis. Communication monographs, 72(1), 92-115.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. 88-93.