Understanding Community Rivalry on Social Media: A Case Study of Two Footballing Giants

Sopan Khosla
skhosla@adobe.com
Adobe Research
Bangalore, India

Siddhant Arora
cs5150480@iitd.ac.in
IIT Delhi
New Delhi, India

Abhilash Nandy
nandyabhilash@gmail.com
IIT Kharagpur
Kharagpur, India

Ankita Saxena
ankitasonu24@gmail.com
IIT Roorkee
Roorkee, India

Anandhavelu N
anandvn@adobe.com
Adobe Research
Bangalore, India

ABSTRACT

Detection of hate speech in online user generated content has become of increasing importance in recent times. Hate speech can not only be against a particular user but also against a group. Rivalry between two communities with opposing ideologies has been observed to instigate a lot of hate content on social media during controversial events. Moreover, this online hate content has been observed to have power to shape exogenous elements like communal riots [4, 6]. In this paper, we aim to analyze community rivalry in the football domain (Real Madrid FC vs FC Barcelona) based on the hate content exchanged between their supporters and understand how events affect the relationship between these clubs. We further analyze the behavior of key instigators of hate speech in this domain and show how they differ from general users. We also perform a linguistic analysis of the hate content exchanged between rival communities. Overall, our work provides a data-driven analysis of the nuances of online hate speech in the football domain that not only allows a deepened understanding of its social implications, but also its detection.

CCS CONCEPTS

• **Information systems** → *Social networks*; • **Computing methodologies** → *Natural language processing*; *Neural networks*; • **Social and professional topics** → User characteristics.

KEYWORDS

social media; OSN; hate speech; online communities; NLP; time-series; causal inference

IUI Workshops'19, March 20, 2019, Los Angeles, USA

ACM Reference Format:
Sopan Khosla, Siddhant Arora, Abhilash Nandy, Ankita Saxena, and Anandhavelu N. 2019. Understanding Community Rivalry on Social Media: A Case Study of Two Footballing Giants. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019* , 8 pages.

1 INTRODUCTION

Social media is a powerful communication tool that has facilitated easy exchange of points of view. While it has enabled people to interact with like-minded people, share information and support during a crisis [25], it has also resulted in a rise of anti-social behavior including online harassment, cyber-bullying, and hate speech [14]. With more and more people sharing web content everyday, in particular on online social networks (OSNs), the amount of hate speech is also steadily increasing. In recent years, we have witnessed a growing interest in the area of online hate speech detection and particularly the automatization of this task. Social networking communities like Facebook and Twitter are putting forth hateful conduct policies [15, 36] to tackle this issue.

User-defined communities are an essential component of many web platforms, where users express their ideas, opinions, and share information. These communities also facilitate intercommunity interactions where members of one community engage with members of another. Studies of intercommunity dynamics in the offline setting have shown that intercommunity interactions can lead to the exchange of information and ideas [3, 17, 29] - or they can take a negative turn, leading to conflicts [31]. If the participating communities have opposing views, their interactions can lead to exchange of hate content that maybe directed towards the community's *ideology* (or interest, team) or its members. These online exchanges can also lead to on-the-ground communal violence [4, 6].

In this work, we present the first comparative study on the exchange of hate content during inter-community interactions in football fan communities. This hate maybe directed

Community	Users (Tweets)		
	General Mentions	HS(S:all)	HS(S:rival)
Real Madrid FC	564,770 (2,065,846)	10,105 (12,827)	3,502 (4,234)
FC Barcelona	307,931 (894,479)	6,855 (12,568)	1,774 (2,112)

Table 1: Data Statistics

towards the club in general or the members (e.g. players, supporters, manager etc.) of that club. Specifically, we analyze the rivalry between two Spanish football communities, Real Madrid FC and FC Barcelona, on Twitter. Real madrid FC and FC Barcelona are considered two of the biggest football clubs in the world and enjoy large support all over the world. Prior work [19] studies how members of one community attack the members of other community on Reddit, which provides an explicit platform for users to create and participate in interest-based communities (called subreddits in case of Reddit). However, for our study, we choose Twitter as it provides a larger cross-section of general public.

In order to characterize the dynamics of hate between these communities, We first try to understand how hate speech changes over time. Specifically, does hate speech increase over time and does it spike during external events. Next, we try to understand the characteristics of users who spread hate and then we try to understand the hateful tweets themselves. We also try to analyze how hate from the rival community is similar or dissimilar to the general hate against a target community and its members.

2 RELATED WORK

There have been notable contributions in the area of hate detection in Online Social Networks (OSN) and websites. [22, 24] use lexical features like word and character n-grams, average word embeddings, and paragraph embeddings. Other works [6, 11, 21, 30, 37] have leveraged profane words, part-of-speech tags, sentiment words and insulting syntactic constructs in pre-processing and as features for hate classification. Models used in the existing literature include supervised classification methods such as Naive-Bayes [20], Logistic Regression [38] [10], Rule-Based Classifiers [18], Random Forests [7] and Deep Neural Networks [1].

There have been fewer efforts towards characterizing hateful users online (people who post hate speech in OSNs). Chatzakou et al. [8] study the Twitter users in the context of #GamerGate controversy. In another work, Chatzakou et al. [9] use a supervised model to classify Twitter users into four classes: bully, aggressive, spam, and normal. Rudra el al. [27] characterize Twitter users, who post communal tweets during disaster events, based on their popularity, interests, and social interactions.

Silva et al. [30] use regex patterns like "I <intensity> hate <targeted group>." to identify hate target groups in terms of their class and ethnicity. Their system has very low recall as they only rely on very specific sentence structures. Another line of work [13] identifies individual targets using mentions in the hate tweets and uses Perspective API's *toxicity* and *attack_on_commenter* scores to detect if the hate speech is against the mentioned individual. In this work, we leverage the prior art in the area of stance detection for target-specific hate speech detection [32].

3 BACKGROUND

We define hate speech **(HS)** and hateful users **(HU)** according to the guidelines put forth by Twitter [36]. Any content that *promotes violence against or directly attacks or threatens other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease* is considered as "hate speech". On the other hand, "hateful user" is a user that perpetrates such type of content. Target Community **(T)** is defined as the ideology, team, interest, ethnicity etc. which the intended recipients of hate speech belong to. Whereas, source community **(S)** is the ideology, team, interest, ethnicity etc. that characterizes the group of users who post one or more hate tweets. For example, if a FC Barcelona supporter posts hate against Real Madrid or its members then FC Barcelona is considered the "source community" and Real Madrid is the "target community".

4 DATA COLLECTION

In this section, we provide details about the data collection and pre-processing pipeline.

Data Sources We collect data from three main sources. We collect tweets from *Twitter* using tweepy API, match fixtures and outcomes from *Foxsports.com* and other match statistics from *espn.com*.

Target specific tweets We collect relevant tweets in English language (via Twitter Search API) that mention the target community of interest, from June 2017 to May 2018. We create a list of entities (players, managers, owner, etc.) that belong to the target community of interest and use them as our search keywords for this.

Preliminary hate filtering We adopt a high recall data collection mechanism to represent a fair sense of hate speech in our domain. Similar to [13], we use a lexicon of abusive words adopted from [39] to retrieve English hate terms. After removing phrases that are context dependent, we use the resultant list of hate words to filter the extracted target specific tweets.

Source Community Identification We identify the community to which the Twitter user belongs by extracting their friends. We check if the user follows the official pages of

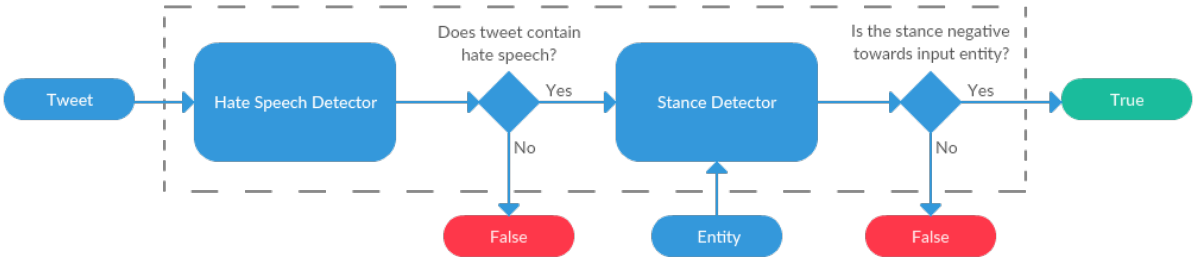


Figure 1: Schematic diagram of our target-specific hate speech detector

the community of interest (or its members) to categorize him as member of that community. Using this differentiation process, we categorize users into Barcelona supporters/ Real Madrid supporters/ neither.

5 TARGET-SPECIFIC HATE SPEECH DETECTION

Tweets collected in the previous section merely mention the target community. This does not guarantee that they are actually talking about it. Also, despite the qualitative inspection of keyphrases, the filtered dataset still contained non-hate speech tweets. To mitigate the effects of obscure contexts in the filtering process, we propose a two-step classifier that would provide us with tweets that contain hate speech against the target entity.

Figure 1 shows the workflow for the proposed framework. A tweet is first passed through the hate-speech detection model. If the model classifies the text as positive for hate speech, then it is input to the stance detector along with the target entity of interest. If the stance detection model classifies it as negative towards the target, then we assert that tweet contains hate-speech against the target entity.

Does this tweet contain hate speech?

There has not been much prior work in modeling hate speech specific to sports domain. Therefore, we use an existing hate speech detection model trained on dataset from a different domain and see how it fares in our domain.

We use an LSTM model [1] trained on two popular datasets of general tweets, manually annotated for hate speech, to detect if hate speech is present in the tweet. Dataset introduced in [16] consists of 70K tweets manually annotated as abusive, hateful, normal and spam whereas the dataset proposed in [38] categorizes 20K tweets into sexist, racist and neither. We consider the tweets labeled as abusive, hateful, sexist or racist in the datasets as positive for hate speech. Spam samples are not used for training. We use an LSTM for modeling as it has been shown to capture the long-range dependencies in tweets, which may play a role in hate-speech detection.

Is this hate directed towards the target community?

Stance is used to define target-specific opinion (as against a general opinion) which can be favor, against or neutral. Stance helps to disambiguate between the generic sentiment or opinion of an individual with what the individual is referring to. In this work, we leverage stance detection algorithms as the second step in our target-specific hate detector.

To perform stance detection in tweets, we leverage a state-of-the-art TC-LSTM model introduced by [33] for target-dependent sentiment classification.

6 ANALYSIS

In this section, we present the analysis on the hate dynamics between two footballing giants - Real Madrid FC (also referred to as *madrid*) and FC Barcelona (*barca*). Tweets categorized by our model as hate speech against the target community t by users of source community s are referred to as HS(S:s, T:t) and the corresponding users who posted them as HU(S:s, T:t). Furthermore, S:all is used to represent all hate against a target community.

Is hate exchange a year round event?

We plot the time-series (Figure 2) of total number of hateful tweets exchanged between Real Madrid FC and FC Barcelona in the period ranging from June 2017 to May 2018 (football season 2017-18). We observe that the number of tweets with hate speech spike in isolation. A close inspection maps these spikes to football matches (also called *events* in rest of the paper) in which the target community was playing. In the absence of these events, we do not find a substantial amount of hate speech and hate speech in general does not seem to increase with time.

Studies [6, 28] have shown that online hate speech has increased over the years. However, it is difficult to address this question in retrospection as several offensive tweets are taken down by Twitter soon after they are posted.

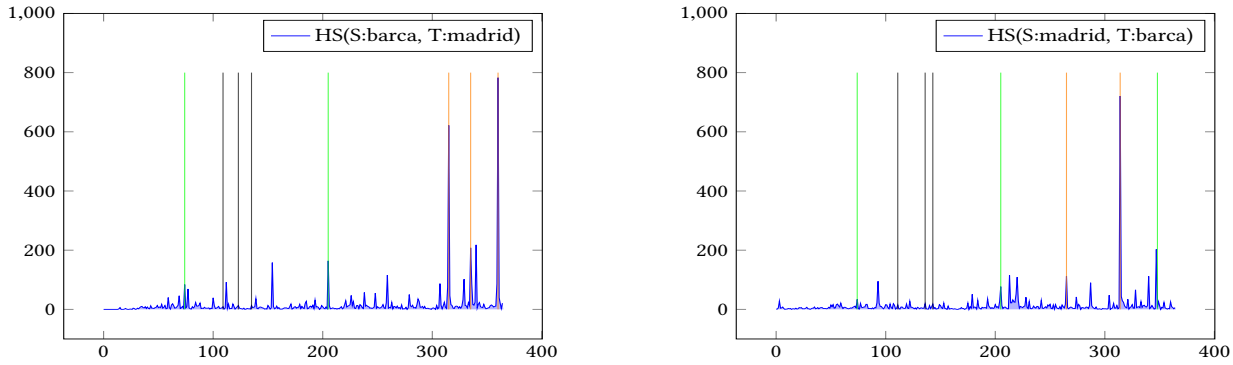


Figure 2: Number of hate tweets exchanged between Real Madrid FC and FC Barcelona in 2017-18 season. Vertical lines denote a sample of events during the season. Green corresponds to el-classico matches; black and orange lines represent matches which saw smallest and largest amount of hate respectively from the rival community.

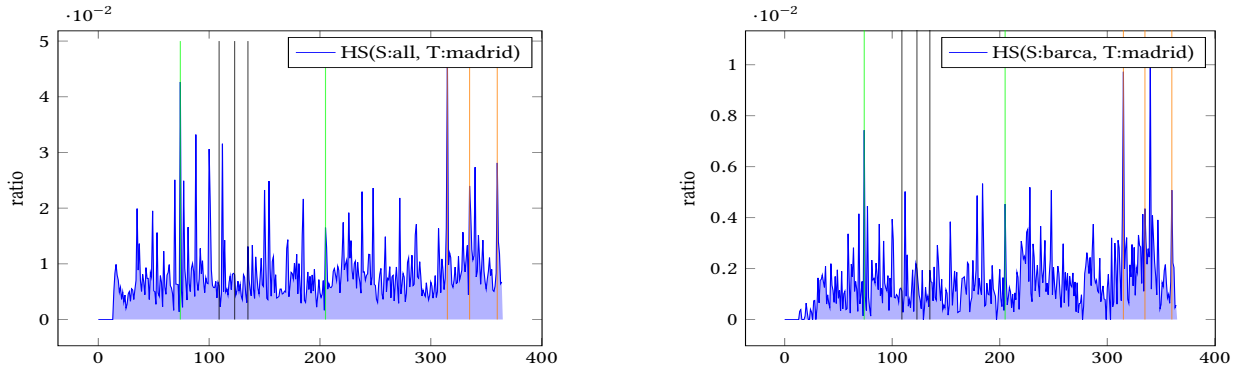


Figure 3: Ratio of 'hate speech towards Real Madrid FC' to 'General Mentions of Real Madrid FC' during the 2017-18 season.

Impact of offline events on hate speech online

Following the inference from the previous section, we now try to quantify the impact an event had on online hate speech. We posit that an event (a football match in our case) has high impact on hate speech if it results in an increased relative amount of hate speech against non-event days. To analyze the impact of events on online hate exchange between the two communities of interest, we plot a time-series of the ratio of HS(S:s,T:t) to general tweets mentioning target community t. (Figure 3 shows the time-series for T:madrid). We quantify their effect by treating them as interventions on observed time series. Following [25], we use Brodersen et al.'s technique [5] for causal inference on time-series to quantify the impact of football events on hate speech. The behavior of observed time-series (*treatment*) after a football match is compared with a counterfactual time series (*control*). Since, we do not observe the *control* time-series, we model it from observed time-series (in different time ranges) that correlate with the *treatment* series but were not affected by the event. Finally we use this setup to model the counterfactual of the treated time series using difference-in-differences approach.

A. Observed Time-Series: We define our treatment series with a timespan of two days before the match as *pre-treatment* period and two days after the match as *post-treatment* period.
B. Synthetic Control Group creation: We then identify possible *control* groups as time-series that occur in history, during same days of the week as the treatment series, with no event taking place during this time interval. We rank all *control* groups based on their similarity to the *treatment* group using Wilcoxon signed rank test and select the top 2 ranked time series for creating our synthetic control time series.
C. Impact Estimation: We finally use the difference between observed post treatment time series and the synthetic control time series to calculate the impact of event. The relative increase in online hate speech during an event is given by:

$$rel_{effect} = 100 * \frac{\sum t_k - c_k}{\sum c_k} \quad (1)$$

where t_k is the value of the *treatment* time series at time k, and c_k that of the *control* time series.

Do all events contribute to the hate speech equally? Our results show that outcome of matches seems to affect the hate

Hateful Users (HU)	General Hate (% of total tweets)		Event Overlap (% of total events)		Followers		
	Mean(%)	% Users who post > 10%	Mean(%)	% Users who post in > 20%	Mean	% Users with < 100	% Users with > 10,000
T: madrid							
S:all	9.87	43.73	1.176	0.08	2570.58	21.66	2.46
S:barca	10.09	45.92	1.145	0.01	2062.83	18.29	2.64
T: barca							
S:all	9.10	36.17	1.480	0.05	3000.74	37.00	3.70
S:madrid	9.36	39.40	1.384	0.00	1069.66	30.45	1.60

Table 2: Characterizing hateful users

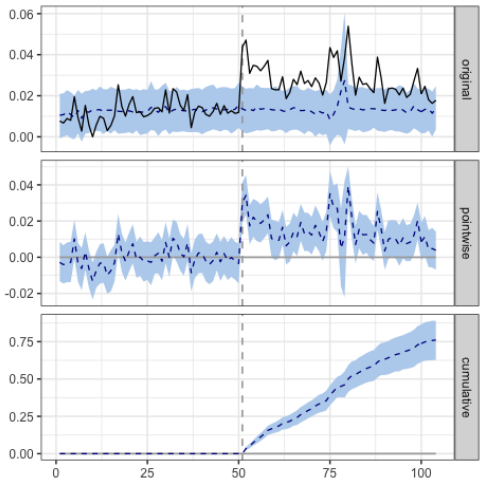


Figure 4: Example of impact estimation with counterfactual predictions, for the event "Champions League 2018 final" between Real Madrid FC and Liverpool FC. Top: Black is the observed series before/after the event, blue (dashed) is the counterfactual.

exchange. We find that losses and draws trigger greater hate speech than wins ($p < 0.001$). We also observe that Home Losses and Draws trigger greater hate in comparison to Away Losses and Draws ($p < 0.001$). In contrast, Away Wins trigger greater hate than Home Wins (HU(S:all, T:madrid): $p < 0.01$; HU(S:barca, T:madrid): $p < 0.001$;) particularly from the rival community. Moreover, group games trigger less hate in comparison to championship games (HU(S:all, T:madrid): $p = 0.05$; HU(S:barca, T:madrid): $p < 0.05$;). Matches which lead to elimination (HU(S:all, T:madrid): $p < 0.001$; HU(S:all, T:madrid): $p < 0.05$;) trigger greater hate. Matches played against the rival community tend to bring more hate from the rival community ($p < 0.001$). More generally, we find that rival community posts disproportionately higher hate for matches whose results directly impact them. A similar trend is observed for t = barca (omitted for brevity).

Characterizing hateful users

In this section we analyze HU(S:s, T:t). Users who post hateful content against the target community of interest.

Do they post hate speech in general? We investigate if the hateful users in our domain use hate speech in their general tweets as well. We extract their 3200 most recent tweets and classify them using our hate speech detection model (explained in the earlier sections) after excluding tweets that mentioned entities related to Real Madrid FC or FC Barcelona. We observe that users who post hate in our domain also propagate significant hate in general with around 10% of their general tweets being hateful (Table 2). Our results show that HU(S: rival) post more hate in general as against HU(S: all) with a higher percentage of HU(S: barca, T: madrid) crossing the 10% mark as compared to HU(S: all, T: madrid) (45.92% *vs* 43.73%). A similar trend is observed for T:barca.

Do they post hate in multiple events? We then analyze the user overlap across different events to see if there is a common set of users who post hate during multiple events. Any user who posts in a 24hr interval after the event is assumed to have posted because of that event. We find that around 90% HUs write hate tweets for only 1−2% of football matches throughout the year. Whereas, less than 0.1% HUs post hate in more than 20% of the total football matches in 2017-18 season (Table 2). Members of the rival community HU(S:rival), on average, show lower event overlap compared to HU(S:all). This is consistent with the findings in the *Analysis* section as the rival community is only interested in events which impact them directly.

Are they popular? Next, we check if the users who write hate tweets enjoy popularity on Twitter. We use *number of followers* as a metric to do the popularity analysis. As shown in Table 2, target-specific hate in our domain is posted by common masses (less than 100 followers) whereas, the popular users (more than 10,000 followers) seldom (< 4%) participate in this phenomenon. Popular members of Real

Madrid FC community seem to avoid hate speech against FC Barcelona (1.6%).

What are their key personality traits? To study the key characteristics of the personalities of HUs in our domain, we use the Twitter REST API to fetch the most recent 3200 tweets for each account. We exclude retweets as they might not reflect author's point of view and use IBM Watson Personality Insights API[1] for this analysis. It outputs a normalized percentile score for the characteristic. We study the results of the Big Five personality model, the most widely used model for generally describing how a person engages with the world. The model includes five primary dimensions: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness [2].

Figure 5 shows the distribution of scores of the Big Five personality traits for T:madrid. We find that HS(S:all) and HS(S:barca) have more similar personalities to each other than general mentions. Both HS(S:all) and HS(S:barca) exhibit lower Agreeableness than general mentions. Prior work [34] associates lower Agreeableness scores with suspicious and antagonistic behaviors. Our results indicate that HS(S:all) and HS(S:barca) are more self-focused, contrary, cautious of others, and lack empathy. For Conscientiousness, HS(S:all) and HS(S:barca) generally have lower scores than general mentions. Our results suggest that these users are laid back, less goal-oriented, and tend to disregard rules. Low Extraversion scores for both HS(S:all) and HS(S:barca) show that they are less sociable, less assertive, and more within themselves. HS(S:all) and HS(S:barca) have slightly higher, but statistically significant, scores for Neuroticism which indicates that they are more susceptible to stress and are more likely to experience anxiety, jealousy and anger. However, for Openness, the distributions for HS(S:all) and HS(S:barca) are close to general mentions (with median of approximately 0.19). We observe a similar trend for T:barca but omit here for brevity.

Characterizing hateful tweets

Are these hate tweets popular? We next investigate the popularity of target-specific hate tweets in our domain. We use the *retweet-count* of tweets to judge their popularity. We observe that for target community FC Barcelona (T:barca), hateful tweets from Real Madrid i.e. HS(S:madrid, T:barca), are retweeted less than general hateful tweets HS(S:all, T:barca) (with mean = 0.19 and 0.53 respectively) which in turn are substantially less popular than non-hateful tweets (mean = 5.47).

Content Characteristics. We use SAGE [12] to analyze salient words that characterize different types of tweets. SAGE attempts to find salient terms in a text (*child*) with respect to some base content (*base*). It creates clean topic models by taking into account the additive effects and combines multiple generative facets like topic and perspective distribution of words. In this analysis, we conduct three experiments (i) *child* = HS(S:all), *base* = tweets which mention the target community; (ii) *child* = HS(S:rival), *base* = tweets which mention the target community; and (iii) *child* = HS(S:rival), *base* = HS(S:all). We look at the top 10 salient words learned for the above-mentioned experiments (Table 3).

Exp.	Top 10 Salient Words
(i)	f**k, b**ch, f**ked, f**king, adore, c*nt, yer, kid, cheating, idiot
(ii)	f**k, f**ked, f**king, b**ch, c*nt, bulls**t, adore, sh**ty, sh*t, cheating
(iii)	granada, rampant, **messi**, bartomeudimiteya, , **valverde**, **forcabarca**, **viscabarca**, yer, madridiots, penaldo

Table 3: Top 10 salient words learned by SAGE for different experiments with T:madrid. Note the presence of barca specific keywords (bold) in (iii).

As a whole, both (i) and (ii) contain similar salient words. These words are mostly cuss words which is to be expected. The top salient words in (iii) contain mentions of the entities of the source community. On closer look, we find that these tweets try to demean the target community (e.g. players, managers or ideology etc.) in an attempt to glorify the source community. For example, this hate tweet by a Real Madrid FC supporter against FC Barcelona, *king dem ronaldo king dem left and right salute d king..i want to take this opportunity and say f**k all barcelona fans @fcbarcelona _es*, tries to glorify Ronaldo (an ex Real Madrid FC player) by calling him a King. Such a pattern of comparison is not visible in HS(S:all) which mostly focuses on the negatives of the target community. For example, this hate tweet against FC Barcelona from a user who is not a Real Madrid FC community member, *Dear @FCBarcelona , please take your sh*t (Bellerin) back. Please!.*

Psycholinguistic Analysis. We use LIWC2015 [26] for a full psycholinguistic analysis. We look at the following dimensions: summary scores, personal pronouns, and negative emotions.[2] In Figure 6 we can see that tweets with general (non-hate) mentions (NHM) of Real Madrid FC differ significantly from hateful tweets. Summary scores suggest that general tweets display higher values of tone than HS(T:madrid) suggesting that targeted hate speech is more hostile. HS(T:madrid) contains higher number of pronouns and is angrier than general tweets. Also, HS(T:madrid) is more informal and expectedly contain more swear words. It contains shorter sentences and uses more dictionary words on average as

[1] https://www.ibm.com/watson/services/personality-insights/

[2] LIWC2015 language manual [26] provides a detailed description for these dimensions.

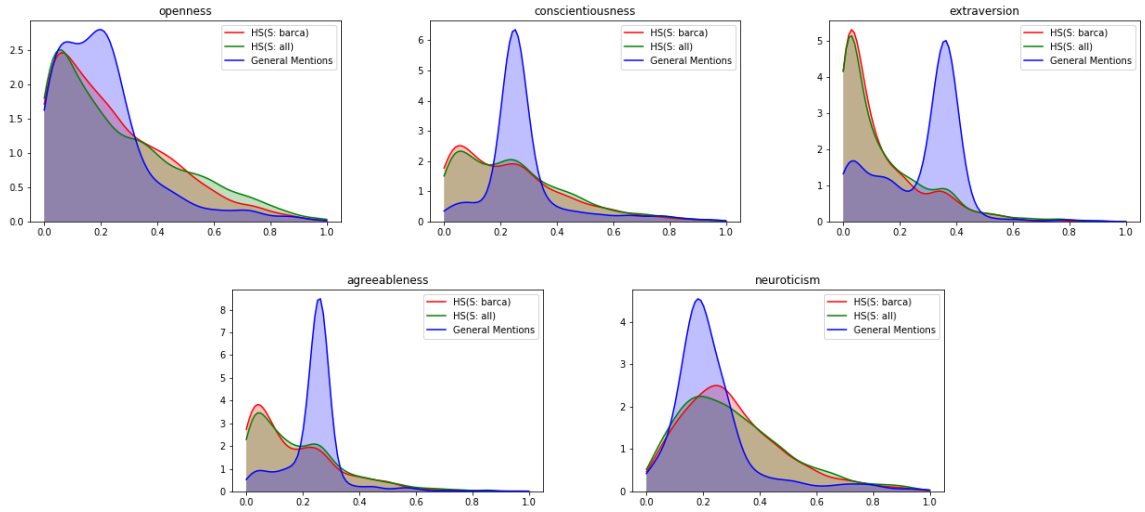
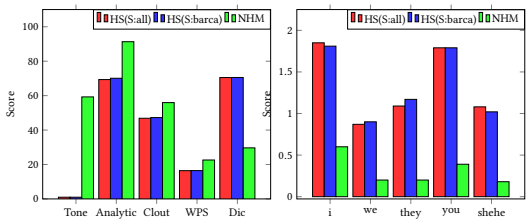
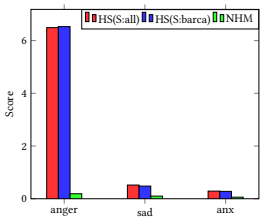


Figure 5: Distribution of scores for the Big Five personality traits for users who posted tweets that mention Real Madrid FC. *General Mentions* are tweets that mention the target community or its associated entities.



(a) Summary **(b) Personal Pronouns**



(c) Negative Emotions

Figure 6: Psycholinguistic Analysis of tweets which mention Real Madrid FC (T:madrid).

against the non-hate tweets mentioning Real Madrid FC. A similar trend is observed for T:barca (omitted for brevity).

7 DISCUSSION AND CONCLUSIONS

In this work, we provide a novel view of hate exchange between different communities in the football domain. We design a two-step model to detect target-specific hate speech. Using causal inference methodologies, we are able to measure the effect of external events on hate speech on social media. We show how rival communities post disproportionately high amount of hate during events which have a direct impact on their team's interests. We find that hateful users in our domain also post hate speech in general. They do not post hate in multiple events and do not enjoy generous popularity on social media. We show that their personality characteristics are significantly different from general users who post about the target community. Our analysis shows that hate tweets from rival community members differ in their theme from general hate tweets towards the target community. They try to glorify their team's players, playing style or ideology while demeaning the target community. However, the psycholinguistic analysis of the hateful tweets suggests that content from rival community does not differ from general tweets in terms of the emotional content, tone, pronoun usage or swear words.

Nonetheless, our analysis has limitations. Recent studies by Tufekci [35] and Morstatter et al. [23] have discussed the sample quality of the Twitter API. Since our analysis relies on keyword-based methods for retrieval of explicit hate speech, we cannot claim to have captured a complete representation of the hate exchange on Twitter. However, our main objective was to characterize hateful users and tweets in the sports domain with high precision and we believe that our careful filtering and classification models were able to do so.

REFERENCES

[1] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. *Proceedings of the 26th International Conference on World Wide Web Companion*, 759–760.

[2] Murray R Barrick and Michael K Mount. 1991. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology* 44, 1 (1991), 1–26.

[3] Václav Belák, Samantha Lam, and Conor Hayes. 2012. Cross-Community Influence in Discussion Fora. *ICWSM* 12 (2012), 34–41.

[4] Susan Benesch. 2014. Countering dangerous speech to prevent mass violence during KenyaâĂŹs 2013 elections. *Final Report* (2014), 1–26.

[5] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. 2015. Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics* 9 (2015), 247–274.

[6] Pete Burnap, Omer F Rana, Nick Avis, Matthew Williams, William Housley, Adam Edwards, Jeffrey Morgan, and Luke Sloan. 2015. Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change* 95 (2015), 96–108.

[7] Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science* 5, 1 (23 Mar 2016), 11. https://doi.org/10.1140/epjds/s13688-016-0072-6

[8] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Hate is not binary: Studying abusive behavior of# gamergate on twitter. *Proceedings of the 28th ACM conference on hypertext and social media*, 65–74.

[9] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. *Proceedings of the 2017 ACM on web science conference*, 13–22.

[10] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of ICWSM* (2017).

[11] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 3 (2012), 18.

[12] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 1041–1048.

[13] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. *ICWSM* (2018).

[14] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to Peer Hate: Hate Speech Instigators and Their Targets. *ICWSM* (2018).

[15] Facebook. 2016. Controversial, Harmful and Hateful Speech on Facebook. https://goo.gl/TWAHdr.

[16] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *International Conference on Social Web and Media Log* (2018).

[17] Howard Giles. 2005. *Intergroup communication: Multiple perspectives*. Vol. 2. Peter Lang.

[18] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering* 10, 4 (2015), 215–230.

[19] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 933–943.

[20] Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets against Blacks. *AAAI*.

[21] Altaf Mahmud, Kazi Zubair Ahmed, and Mumit Khan. 2008. Detecting flames and insults in text. (2008).

[22] Yashar Mehdad and Joel Tetreault. 2016. Do Characters Abuse More Than Words? *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 299–303.

[23] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *ICWSM*.

[24] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. *Proceedings of the 25th international conference on world wide web*, 145–153.

[25] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R Varshney. 2018. The Effect of Extremist Violence on Hateful Speech Online. *arXiv preprint arXiv:1804.05704* (2018).

[26] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

[27] Koustav Rudra, Siddhartha Banerjee, Niloy Ganguly, Pawan Goyal, Muhammad Imran, and Prasenjit Mitra. 2016. Summarizing situational tweets in crisis scenario. *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, 137–147.

[28] SafeHome.org. 2017. Hate on Social Media. https://www.safehome.org/resources/hate-on-social-media/.

[29] Muzafer Sherif. 2015. *Group conflict and co-operation: Their social psychology*. Psychology Press.

[30] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media. *ICWSM*, 687–690.

[31] Henri Tajfel and John C Turner. 1979. An integrative theory of intergroup conflict. *The social psychology of intergroup relations* 33, 47 (1979), 74.

[32] Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective LSTMs for target-dependent sentiment classification. *International Conference on Computational Linguistics* (2016), 3298–3307.

[33] Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Target-Dependent Sentiment Classification with Long Short Term Memory. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* abs/1512.01100 (2016), 3298âĂŞ3307.

[34] Ginka Toegel and Jean-Louis Barsoux. 2012. How to become a better leader. *MIT Sloan Management Review* 53, 3 (2012), 51–60.

[35] Zeynep Tufekci. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM* 14 (2014), 505–514.

[36] Twitter. 2016. Hateful Conduct Policy. https://support.twitter.com/articles/20175050.

[37] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. *International Conference Recent Advances in Natural Language Processing (RANLP)*, 672–680.

[38] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93. http://www.aclweb.org/anthology/N16-2013

[39] Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a Lexicon of Abusive Words–a Feature-Based Approach. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* 1, 1046–1056.