#Wikipedia on Twitter: Analyzing Tweets about Wikipedia

Eva Zangerle, Georg Schmidhammer, Günther Specht
Databases and Information Systems
Institute of Computer Science
University of Innsbruck, Austria
firstname.lastname@uibk.ac.at

ABSTRACT

Wikipedia has long become a standard source of information on the web and as such is widely referenced on the web and in social media. This paper analyzes the usage of Wikipedia on Twitter by looking into languages used on both platforms, content features of posted articles and recent edits of those articles. The analysis is based on a set of four million tweets and links these tweets to Wikipedia articles and their features to identify interesting relations. We find that within English and Japanese tweets containing a link to Wikipedia, 97% of the links lead to the English resp. Japanese Wikipedia, whereas for other languages 20% of the tweets contain a link to a Wikipedia of a different language. Our results also indicate that the number of tweets about a certain topic is not correlated to the number of recent edits on the particular page at the time of sending the tweet.

Categories and Subject Descriptors

H.1.4 [**User/Machine System**]: Human Factors; H.1.4 [**User/Machine System**]: Human Information Processing

General Terms

Experimentation, Human Factors, Measurement

Keywords

Wikipedia, Twitter, Quantitative Study

1. INTRODUCTION

Wikipedia is the primary and most extensive encyclopedia available online and is a central source of information for millions of users, making it the 6^{th} most visited site on the web [1], serving 450 million people every month [7]. The collaborative nature of Wikipedia has paved way many collaborative social media platforms enabling users to engage, contribute and communicate [19]. At the same time, online social networks (OSN) have evolved to a popular means of communication and enable millions of users to communicate, collaborate and share information online. Twitter is among the most successful OSNs in terms of the total number of users and interchanged messages every day. The Twitter platform currently serves 271 million active users who publish approximately 500 million tweets every day [3].

Wikipedia has been studied extensively in research over the last decade. In particular, researchers were interested in intrinsic factors contributing to the success of Wikipedia such as the community as the driving force behind Wikipedia and the quality of articles on Wikipedia. As for the community gathering around Wikipedia, researchers have analyzed the structure of the community in terms of e.g., interactions on discussion pages [21] or behavior and social aspects within groups [22, 31]. Another important factor that has been analyzed is the quality of articles on Wikipedia (e.g., in the form of duplicate detection [30] or coordination of contributions [20]). In contrast, extrinsic factors influencing the popularity and usage of Wikipedia have not been studied yet. One important extrinsic factor is the propagation of links to Wikipedia articles on OSNs. Recently, the Wikipedia Foundation has added a new feature to the Wikipedia Android app which allows users to easily share Wikipedia articles [32].

In this paper, we shed light on how Wikipedia is referenced by Twitter users by analyzing three different aspects: (i) the languages used within tweets and the Wikipedias referenced in these tweets, (ii) the Wikipedia articles and categories that are tweeted (content) and (iii) the relationship between the number of tweets about a certain article and the number of edits of this article. In particular, we aim to answer the following research questions:

- How popular are the various Wikipedias on Twitter and in which language contexts are these referenced?
- Which features do Wikipedia articles that are popular on Twitter exhibit/share?
- Does the number of tweets about a certain article correlate to a recent edit and hence, an update of the page?

Our analyses show that except for English and Japanese tweets, more than 20% of all tweets feature inter-language links (i.e., the language of the tweet differs from the language

OpenSym '15 August 19 - 21, 2015, San Francisco, CA, USA

ACM 978-1-4503-3666-6/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2788993.2789845.

of the Wikipedia linked to). Furthermore, we find that the distribution of tweeted Wikipedia articles features a long-tail distribution and that 64% of all articles of the English Wikipedia within our dataset are only tweeted about once. As for the correlation between the number of tweets about a certain article and the number of edits of the respective article, we find that these are not correlated for the general case. However, they may very well serve as an indicator for the occurrence of events.

The remainder of this paper is structured as follows. Section 2 characterizes approaches related to the analysis presented in this paper. Section 3 describes the dataset the analysis was performed upon and the cleaning steps we performed on this data. Section 4 presents the results and insights gained in the course of the presented analysis and Section 5 discusses these results. Section 6 concludes our paper in the light of our findings and provides an outlook upon possible future research directions.

2. RELATED WORK

Wikipedia is the most popular online encyclopedia [7] and this popularity naturally attracts researchers who investigated various aspects related to Wikipedia which also have been covered by survey papers [17, 28]. Research around Wikipedia includes e.g., examining and facilitating the data provided by Wikipedia and also analyzing the community behind Wikipedia and its effects.

The direct interplay between Wikipedia and Twitter has—to the best of our knowledge—not been touched by research yet. However, there is substantial work on how Wikipedia and its data can be facilitated for Twitter-related research as data extracted from Wikipedia has proven to be a valuable source of information (not only) for Twitter-related research. Wikipedia's disambiguation pages and its category system provide a profound basis for computing semantic relatedness [13]. Li et al. make use of page titles, disambiguation pages and redirects aiming to perform Named Entity Recognition (NER) for tweets [23]. Osborne et al. facilitate page views for articles extracted from Wikipedia as an indicator for events to enhance first story detection for tweets [29]. Furthermore, Wikipedia is also facilitated for the classification of tweets. Genc et al. perform such a classification based on a mapping from tweets to the most similar Wikipedia article and subsequently computing the distance of these articles as the distance of the categories of the respective articles [14]. A similar approach is followed by Parker et al. who map tweets to Wikipedia articles to subsequently use the introduction of the respective article to match it against a medical dictionary to indicate whether the given tweet reports some medical issue. This data is collected to infer public health trends based on Twitter data. Xu and Oard propose an approach for clustering tweets by leveraging Wikipedia's linking history to disambiguate topics [35]. Also, Wikipedia is used to create a user profile for user interest classification. Michelson and Macskassy firstly detect topics for a given tweet by performing NER on the tweet and subsequently resolve these entities against the Wikipedia category hierarchy to extract a user's topics of interest [27]. Lim and Datta aim to classify users based on the celebrities they follow by extracting information about the celebrity's occupation (and hence, an interest category) from Wikipedia [24]. Kapanipathi et al. exploit a hierarchy graph representing Wikipedia's categories to derive a user's interests [18]. Based on a similar user modeling approach, Lu et al. propose a tweet recommendation system which is based on the Wikipedia concept graph [25].

3. TWITTER DATASET

In this section, we present the dataset underlying our studies and the methods used for analyzing the data.

3.1 Data Collection

To gather a representative and sufficiently large raw dataset, we facilitate the following data collection method. We make use of the public Twitter Streaming API which allows for retrieving tweets containing given keywords and associated metadata as JSON-objects [6]. In particular, we filter for tweets containing the term "wikipedia". In total, we were able to gather 4,530,967 tweets fulfilling the filter criterion between 2014/10/20 and 2015/03/10.[1]

As the Twitter Filter API is subject to Twitter's rate limiting, the number of delivered tweets matching the given keywords is capped by a rate limiting equal to the rate limiting of the public Streaming API (approximately 1% of all tweets). However, this rate limiting did not affect our crawling process as the number of tweets matching our query constantly was below this limit (maximum number of tweets crawled per day: 60,910 on 2014/11/19). Hence, we were able to crawl all tweets matching the given filter keywords during the given time period.

3.2 Dataset

Based on the data collection method described in the previous section, we obtain a dataset featuring 4,530,967 tweets. A basic summary of the crawled raw dataset can be found in Table 1 (cf. column "Raw"; the "Cleaned" column refers to the reduced dataset after having cleaned the dataset as described in Section 3.3). A total of 1,440,122 tweets within the dataset are retweets, which accounts for 31.78% of the tweets. As for the URLs within tweets, 68.53% of all tweets within the dataset contain at least one URL, whereby 63.24% of all tweets contain exactly one URL and 5.29% contain more than one URL (maximum: 6 URLs). Generally, the average number of URLs per tweet is 0.75 (SD=0.58, M=1). A further analysis of URLs showed that 73.72% of these URLs eventually lead to a Wikipedia page, corresponding to 54.47% of all tweets. Examining the use of hashtags we found that 19.48% of the crawled tweets contain at least one hashtag, whereby a total of 159,231 distinct hashtags is featured within the dataset. On average, each tweet contains 0.34 hashtags (SD=0.92, M=0). A detailed analysis of the hashtags used can be found in Section 4.1.3. As for the number of tweets composed per Twitter user account, the maximum number of tweets for a user in our dataset is 64,521. Generally, the average number of tweets per user is 2.62 (SD=66.16, M=1). A detailed analysis on the users within the dataset can be found in Section 4.1.1.

Figure 1 depicts the distribution of crawled tweets per day. The average number of total tweets crawled per day is 33,314, the maximum number of tweets per day is 60,910, whereas the minimum number of tweets per day is 11,086 (SD=5,689, M=33,553). The low amount of tweets crawled

[1]We had to fix a certain time frame in order to precisely describe the dataset. However, we still are crawling and updating the dataset.

Characteristic	Raw	Cleaned
Tweets	4,530,967	2,468,055
Retweets	1,440,122	659,641
Distinct Users	1,730,984	844,975
Mentions	3,334,848	1,880,687
Distinct Hashtags	159,231	118,912
Hashtag Usages	1,528,458	778,737
Distinct URLs	1,447,124	1,121,825
URL Usages	3,393,846	2,793,900

Table 1: Dataset Overview (Raw Dataset, Subset of Cleaned Tweets

at the end of December 2014 has to be lead back to an erroneous crawler run. We also looked into the sudden spike on November 19th, 2014 and found that on that day Emmanuel Sanders, an American football player sent out the following tweet: "Wikipedia said I died after the game last week..... Well.... I must be in heaven " Apparently, the Wikipedia article about Emmanuel Sanders has been changed such that it included information about his alleged death and he corrected this information in a tweet [10]. In total this tweet reached 35,872 retweets in total, with 26,850 tweets on 2014/11/19 alone causing the spike in our distribution.

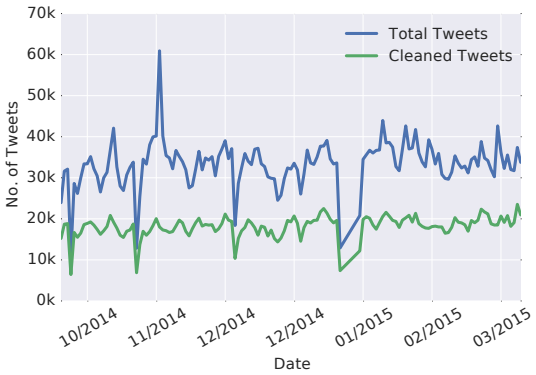

Figure 1: Tweets per Day Distribution

3.3 Data Cleaning

Initially, we performed an exploratory study of the data collected. We found that the dataset contains a myriad of tweets not pointing to Wikipedia. To get a deeper understanding of such behavior, we manually looked into these tweets. We found that many tweets simply mention Wikipedia in discussions without explicitly linking to it like e.g., in "No. #agile coding is not just simple copy and paste. Quality code isn't like writing a history essay from Wikipedia.". As we are mainly interested in how Wikipedia is facilitated in tweets and which articles, categories or features are mainly tweeted about, we choose to restrain our dataset. Therefore, we rely on the fraction of tweets within the dataset which actually contain a link to a Wikipedia page as (only) these tweets contain all the information necessary for our analyses. To be able to constrain the dataset such that it only comprises tweets containing a link to Wikipedia, we perform a filter operation based on the JSON representation of the tweets which also incorporates the fully expanded URL for possibly shortened URLs [4]. Based on this fully expanded URL, we remove all tweets not leading to any Wikipedia page or article. Furthermore, we normalize all URLs ("https://", "http://", "www.", ".mobile." and ".m." for mobile Wikipedia pages and anchors are removed).

An overview of the cleaned dataset and a comparison to the raw dataset can be found in Table 1 (cf. column "Cleaned"). Comparing the characteristics of these two datasets, the smaller, cleaned dataset exhibits similar characteristics (ratio of retweets, users, mentions, hashtag usages and distinct hashtags) except for the number of URL usages. Naturally, this is due to our filter criterion for the cleaned dataset. In the cleaned dataset, the average number of URLs per tweet is 1.10 (SD=0.37, M=1), compared to 0.75 (SD=0.58, M=1) in the raw dataset. This is also manifested in Figure 1 where the spike around November 19th is not reflected in the cleaned dataset as the trending tweet causing the spike did not contain a URL and hence, is not contained in the cleaned dataset (see discussion in Section 3.2). When looking at the number of distinct URLs within both the raw and the cleaned dataset, the number of distinct URLs is not reduced as much as the other characteristics of the two datasets. This can be lead back to the fact that within the raw dataset, 63.24% of the tweets generally contain a URL and 54.47% contain a Wikipedia-URL. Thus, this relatively small difference in the number of distinct URLs stems from the fact that 73.72% of all URLs point to Wikipedia and hence, are contained in the cleaned dataset as well. For our analysis, we choose to also include retweets in the cleaned dataset. We argue that retweets are often a sign of acknowledging the content of a tweet and to make it known to a wider audience [11] and hence, can be used as an indicator of interest of people for certain topics and thus, Wikipedia articles.

4. RESULTS

The following section provides the results of our analysis and discusses the findings. We firstly analyze the Twitter-related features of the tweets (URLs, hashtags, users) and subsequently shift our focus onto Wikipedia. In particular, we aim to relate the tweets within the dataset to the different Wikipedias, articles, categories and recent edits of the individual articles.

4.1 General Observations

In this section we share findings about general facts regarding the dataset—particularly related to Twitter-specific features.

4.1.1 Users

We find that the distribution of the number of tweets about Wikipedia per user is highly dominated by bots. When manually examining the top-20 user accounts with respect to the total number of tweets within the dataset, we find that 19 of these Twitter accounts are bots. The Twitter account which published the highest amount of tweets referring to Wikipedia is "Wikipedia Stub Bot" with 64,521 tweets within our crawling period. This account is a bot and sends out a tweet once a new article stub is created and asks for help in expanding it, like e.g., in "Someone created a Wikipedia article about "Norm Henderson". Help expand it! #Biography #Australia http://t.co/uaUYgD4hoF".

Hashtag	Count
#tanka	29,264
#biography	14,257
#wikipedia	11,392
#역정보	9,013
#gamergate	7,819

Table 2: Top-5 Hashtags

4.1.2 Mobile Wikipedia

As for the use of the mobile version of the Wikipedia website, we extract tweets linking to the mobile version of the Wikipedia website from the links (containing either ".m." or ".mobile.") and found that 22.04% of all tweets include a link to the mobile version of the respective article.

4.1.3 Hashtags

In a first basic analysis, we examine the set of hashtags used within tweets concerned with Wikipedia. Therefore, we extract the hashtags facilitated within the tweets and count the according total number of occurrences. Generally, each hashtag is facilitated in an average of 6.54 tweets (SD=88.33, M=1). Table 2 features the set of the most popular hashtags within the dataset. Within the cleaned dataset, the most popular hashtag is #tanka which refers to a genre of modern Japanese short poems [34]. 21,245 of these usages can be lead back to the retweet of one single trending tweet which contained this particular hashtag. Further frequently used hashtags include #biography, #wikipedia, #역정보 (Korean for disinformation) and #gamergate. The hashtag #gamergate is concerned with a conflict about sexism in video games which was very popular [12]. Interestingly, the distribution of hashtags within the raw dataset differs from the hashtag distribution of the cleaned dataset as #wikipedia is by far the predominant hashtag, followed by #gamergate, #tanka and #keepitfree.

4.2 Languages

In this section, we analyze the distribution and popularity of the different Wikipedias that users refer to within tweets. The aim of this analysis is to gain insights on which Wikipedias are popular within tweets and how this relates to the size and activity of the respective Wikipedia to answer our first research question. Therefore, we extract the language of the Wikipedia edition linked to from the URL mentioned within the tweet by applying regular expressions to the respective URL. Subsequently, we sum up all the references to the single Wikipedias. Table 3 gives a first overview on how links to the individual Wikipedias are distributed among the tweets within the dataset. As can be seen, the English Wikipedia is the predominant Wikipedia with a share of 52.81%, followed by the Japanese Wikipedia. These two Wikipedias are accountable for more 75% of all tweets, all other languages feature a considerably lower share.

To get an impression on to which extent these numbers relate to the characteristics of the different Wikipedias, we extract statistics about the different Wikipedias from the respective Wikipedia article [9]. These statistics include (i) the number of *articles*, (ii) the number of *edits*, (iii) the number of *total* articles and non-articles (including user pages, images, talk pages, project pages, categories and templates),

Language	Total	Share
English (en)	1,349,623	52.81%
Japanese (ja)	579,157	22.66%
Spanish (es)	140,396	5.49%
Turkish (tr)	78,235	3.06%
French (fr)	64,139	2.51%
German (de)	52,256	2.04%
Russian (ru)	44,347	1.74%
Arabian (ar)	38,757	1.52%
Korean (ko)	27,261	1.07%
Portuguese (pt)	26,442	1.03%

Table 3: Top-10 Wikipedias mentioned on Twitter

Measure	Spearman's ρ
Articles	.78*
Total	.76*
Edits	.65*
Users	.46*
Admins	.42*
Active users	.39*
Images	.39*
Depth	.35*

Table 4: Language Correlation (* indicates that correlation is significant at the 0.001 level (2-tailed))

(v) the number of *admins*, (vi) the number of registered *users*, (vii) the number of *users active* within the last thirty days, (viii) the number of *images* and finally, (iv) the *depth* of a Wikipedia[2]. In a second step, we analyze the correlations between the number of tweets about the individual Wikipedias and the popularity of these editions in regards to the extracted statistics about these Wikipedias (e.g., number of articles). As for the correlation between these attributes, we rely on the Speaman rank correlation coefficient. The results of the correlation analyses can be seen in Table 4. The number of tweets containing URLs pointing to Wikipedias and the number of articles within the respective Wikipedias are strongly positively correlated (ρ=.78, $p <$ .001 (2-tailed)). The same holds for the total number of articles and non-articles (total; ρ=.76, $p <$ .001 (2-tailed)). This implies that there is a correlation between highly populated Wikipedias in regards to both the number of articles and/or the number of non-articles and the number of tweets linking to these. On the other hand, we can only detect moderate correlation for the other measures (cf. Table 4). To deepen our understanding for these findings, we manually compare the top-10 Wikipedia editions (with respect to the number of tweets) to the statistics about the different Wikipedias. We find that the intersection between the 10 Wikipedias most frequently linked to in our dataset and the top-10 Wikipedias with respect to the number of articles comprises 5 Wikipedias (en, es, fr, de, ru)—despite the significant correlation for the full populations. When taking

[2]The Wikipedia article describes the depth of a Wikipedia edition as follows: "The "Depth" column (Edits/Articles x Non-Articles/Articles x [1-Stub-ratio]) is a rough indicator of a Wikipedia's quality, showing how frequently its articles are updated and edited. It does not refer to academic quality." [9].

Tweet Language	Wikipedia Language: en	ja	es	ar	fr	tr	de	id	ru	pt
en	**97.33%**	0.19%	0.42%	0.03%	0.33%	0.05%	0.35%	0.12%	0.10%	0.05%
ja	5.48%	**93.56%**	0.04%	0.01%	0.11%	0.03%	0.20%	0.01%	0.05%	0.01%
es	19.65%	0.28%	**77.48%**	0.01%	0.62%	0.03%	0.32%	0.07%	0.03%	0.51%
ar	26.58%	0.02%	0.12%	**72.79%**	0.17%	0.02%	0.02%	0.00%	0.00%	0.00%
fr	20.21%	0.19%	1.11%	1.92%	**74.73%**	0.03%	0.73%	0.02%	0.05%	0.17%
tr	20.78%	0.01%	0.17%	0.00%	0.18%	**77.62%**	0.83%	0.04%	0.10%	0.02%
de	21.15%	0.59%	1.41%	0.06%	0.44%	0.13%	**74.94%**	0.04%	0.04%	0.06%
id	**49.83%**	1.20%	1.77%	0.16%	0.60%	0.40%	0.91%	42.84%	0.06%	0.26%
ru	17.74%	0.10%	0.05%	0.00%	0.14%	0.03%	0.32%	0.00%	**78.38%**	0.01%
pt	28.90%	0.73%	6.91%	0.01%	0.75%	0.05%	0.46%	0.09%	0.03%	**60.87%**

Table 5: Distribution of Inter-Language Links (bold indicates highest value for given language)

into account the number of edits and hence, the activity on the particular Wikipedias, the overlap is 7 (en, de, fr, es, ru, ja, pt). Naturally, the distribution among languages on Twitter and hence, the origin of its users is not evenly distributed (e.g., Indonesia and Malaysia have a disproportionately high number of Twitter users [16]). I.e., the Japanese Wikipedia is the 13^{th} biggest Wikipedia, however, in our dataset it is the 2^{nd} most popular Wikipedia. This is also backed by the fact that Japanese Wikipedia is ranked $2nd$ in terms of total pageviews [2]. To reflect on this bias, we also incorporate the distribution of languages used on Twitter in our analysis. For this analysis, we detect the language of the tweets within our dataset by extracting the "lang"-field provided by the Twitter API, which contains information about the language of the respective tweet [5]. The distribution of languages used in the tweets contained in our dataset features English as the primary language (42.90%), followed by Japanese (21.92%), Spanish (5.77%), Arabian (2.56%), French (2.37%), Turkish (2.24%), German (1.75%), Indonesian (1.56%) and Russian (1.35%) and Portuguese (1.19%). All other language feature a share of less than 1%. For 5.51% of all tweets, Twitter's algorithms were not able to detect its language. Compared to the distribution shown in Table 3, apparently Indonesian and Arabic tweeters do not refer to the Indonesian resp. Arabic Wikipedia in the same proportion as other users of specific countries do.

Following up on these findings, we perform a comparison of the language used in a tweet and the language of the Wikipedia linked to. In particular, we are interested in inter-language links, i.e., tweets for which the language of the tweet differs from the language of the Wikipedia linked to. Table 5 presents the according results for our dataset. Due to the limited amount of space, we present the inter-language analysis for the top 10 languages in respect to the number of tweets in this particular language. English and Japanese tweets are referring to the Wikipedia of the same language in 97.33% resp. 93.56% of all cases. We will refer to such links as intra-language links in the following. On the contrary, within Indonesian tweets 49.83% of the Wikipedia links point at the English Wikipedia. Portuguese tweets link to the Portuguese Wikipedia in 60.87%, the English Wikipedia is linked to in 28.90% and the Spanish Wikipedia is featured in 6.91% of all tweets. Among the other languages, the share of intra-language links is between 72% and 78% and the share of links any non-English language is constantly below the 2% mark, mostly even lower.

4.3 Top Articles and Categories

The second research question is concerned with whether there are certain features which are shared by Wikipedia articles that are popular on Twitter. Therefore, we examine whether specific Wikipedia articles or categories were predominant within our dataset. For this analyses, we extract article titles and categories the articles were assigned to from the latest DBPedia dump. In order to cope with the considerable amounts of data when incorporating data from all Wikipedias, we choose to only analyze articles and categories of the English Wikipedia as it accounts for 52.81% of all Wikipedia links mentioned in the dataset and hence, is the most popular Wikipedia in regards to tweets.

Article	Tweets
diff	54,432
cod_wars	6,868
User:Giraffedata/comprised_of	4,541
matthew_ziff	2,100
kidz_bop	2,015
gamergate	1,703
old_revision	1,517
search	1,383
the_little_mermaid_(1989_film)	1,370

Table 6: Top-10 Articles

To get a first impression on this topic, we compute the most popular articles within the English Wikipedia in regards to how many tweets are dedicated to this particular article. This also includes resolving links containing the ID of an article, e.g., resolving the URL http://en.wikipedia.org/-?curid=7984881 to the English article about the University of Savoy). Additionally, we resolve all extended Wikipedia URLs which are URLs including a query string to e.g., point to a comparison page between two articles or requesting the history of a given page (diff). A full list of extended Wikipedia URLs can be found on the respective Wikipedia article [33]. In total, our dataset features 724,974 links to 336,605 distinct articles of the English Wikipedia. On average, each article is mentioned within 2.36 tweets (SD=12.09, M=1). Table 4 shows the top-10 articles. Interestingly, these include diff-pages as the most popular page mentioned on Twitter. Manually exploring the data, we find that these diff-pages are posted by bots which inform their followers

that a certain page has been changed and updated. E.g., in "Cessna 172 Wikipedia article edited anonymously from Pakistan http://t.co/8PSTwjBHTL" posted by the user "Pakistan Edits". Also among the top-10 mentioned Wikipedia articles are articles about the Cod Wars, a user article about how to (not) use the phrase "comprised of" and the Gamergate controversy (cf. also Section 4.1.3). Also, searches for certain terms on Wikipedia and the extended Wikipedia for retrieving old revisions of articles are among the most popularly tweeted Wikipedia URLs. However, it has to be noted that none of these single articles is mentioned within more than 1% of all tweets.

Figure 2 presents the distribution of the number of tweets for a given article of the English Wikipedia. This distribution also includes the Wikipedia articles that were linked using a diff-URLs (see above). This figure depicts a longtail distribution, i.e., a few articles are tweeted about at a high frequency (relative to the total amount of tweets) whereas the majority of articles are only tweeted about rarely. This is supported by the fact that 15.44% of all articles account for 60% of all links whereas 216.177 articles are only mentioned in a single tweet. This accounts for 64% of all articles.

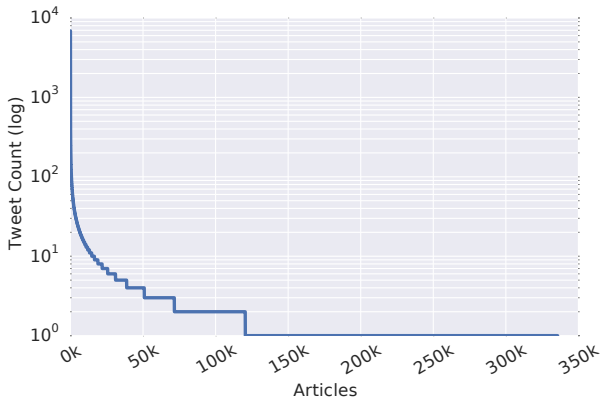

Figure 2: Tweets per Article

As for the top categories, we choose to analyze the leaf categories [15] which contain the articles referenced within our dataset. Our analysis shows that the distribution among the categories mostly tweeted about follows a longtail distribution similar to the distribution of articles. The 10 most frequently referenced leaf categories within our dataset can be found in Table 7. The most popular leaf category is "Living people" accounting for 14.6% of all tweets. Other categories among the most popular categories are dedicated with the film industry and wars. However, as can be seen from the table, the second most popular category accounts for only 2.5% of all tweets—reflecting the long tail of the distribution.

4.4 Correlation of Tweets with Edits

In our third research question we aim to investigate the relation between the number of tweets about articles and the number of edits of these articles around the time the tweet was published. We set this time frame to include all edits 24 hours before and 24 hours after the tweet has been sent. This time frame was chosen in coordination with the findings of Osborne et al., who found that Wikipedia lags

Category	Tweets
Living people	105,895
English-language films	18,331
American films	9,605
Wars involving the United Kingdom	7,487
American male television actors	7,255
20th-century conflicts	7,158
American male film actors	6,981
20th-century military history of the United Kingdom	6,968
Law of the sea	6,953
Wars involving Iceland	6,928

Table 7: Top-10 Leaf Categories

two hours behind Twitter in terms of how fast events are reflected on the respective platform [29]. Osborne et al. also performed experiments on how to choose the window-size for monitoring Wikipedia pages and found that choosing a window size greater than 48 hours did not yield significantly different results.

We rely our analysis on the public MediaWiki API which allows to retrieve all edits of an article within a given period [26]. The analysis is based on the cleaned dataset (see in Section 3.3) and we choose to only include articles of the English Wikipedia. These restrictions lead to a set of 715,977 tweets containing 724,974 URLs linking to 336,605 disinct articles. 81,630 articles were edited within 24 hours before or 24 hours after the tweet has been published which accounts for 24.25% of all articles. Among all articles, 543,788 edits on Wikipedia were performed within the given time frame and 91,577 edits were marked as minor edits (accounting for 16.84%). However, the decision on whether an edit is minor or not is left to the user on Wikipedia and hence, may not serve as a reliable indicator for how comprehensive the edit was. In total, 312,160 tweets link to an article which has been edited within the given time window of 48 hours (43.59% of all tweets). More precisely, 233,962 tweets link to an article which has been edited before the tweet and 215,192 link to an article which has been edited after the tweet. This implies that 136,994 tweets link to an article which was edited before and after the tweet.

To get an impression on the relation between the number of tweets about a certain article and the number of edits of a certain article within the given time window of 48 hours, we compute the correlation between these two distributions. Pearson's r shows a correlation factor of 0.04 (correlation is significant at the 0.001 level (2-tailed)). When excluding retweets from this analysis, r is 0.06. These findings imply that in the general case there is no relation between how actively edited a certain article is and its popularity on Twitter.

5. DISCUSSION

In the following we aim to get a closer look at the results and discuss the implications of our findings described in the previous section. Generally, this paper aims to analyze the interplay between Twitter and Wikipedia from three different perspectives: languages used on Wikipedia and Twitter, the topics and categories that are tweeted (content) and the

relationship between the number of tweets about a certain article and the number of edits of this article.

Firstly, we discuss language features of tweets and particularly, the distribution of inter-language links. While we expected tweets to link to the Wikipedia of the same language, Indonesia does not conform to this hypothesis. To shed more light on this case, we explore Indonesian tweets and particularly look into links to English Wikipedia as this is the Wikipedia referenced most frequently within Indonesian tweets. One obvious hypothesis would be that those articles referenced within the English Wikipedia simply do not have an equivalent article on the Indonesian Wikipedia. To follow up on this hypothesis, we extract those articles and aim to match these against the set of all inter-language links gathered from DBPedia. This analysis shows that for 74.26% of all articles referenced, there is no Indonesian version available. However, for the remaining 25.74% there is an Indonesian version available. Our hypothesis is also supported by the fact that the Indonesian Wikipedia contains roughly 355,000 articles while the English Wikipedia contains 4,752,000 articles.

Intuitively, the number of edits of a certain article may serve as a good indicator for a recent event which is also reflected on Twitter. This has already been shown by Osborne et al. [29]. However, for the general case, we did not find any correlation between the number of edits of an article and the number of tweets about the given article. Osborne et al. showed that Wikipedia lags behind Twitter when it comes to events being reflected, Wikipedia may serve as a filter for spurious events [29]. When looking into events that took place during our crawling period, we can find the gamergate controversy within the set of the most frequently referenced articles and hashtags facilitated. Furthermore, the page with the most edits within our data is the gamergate talk page which also serves as an indicator for the extent of an event. Our findings suggest while there is a correlation between the number of tweets and the number of edits of the respective articles for events (i.e., circumstances evoking public response of any kind), this does not hold for the general case. We hypothesize that this is due to (i) the many small changes to a substantial number of articles by the community and (ii) the highly diversified set of articles and categories which are tweeted. When recomputing the correlation between the number of tweets and the number of edits of a given article without incorporating edits marked as minor edits, the correlation slightly rises from 0.04 to 0.06 (correlation is significant at the 0.001 level for all values (2-tailed)). As for the set of articles and categories that are tweeted, our findings in Section 4.3 suggest that while events such as the gamergate controversy are reflected in this distribution, the long tail of articles is only tweeted about once. Similarly, this also holds for the leaf categories of tweeted articles. This distribution features a longtail distribution which can also be lead back to the size of categories. I.e., the predominant category within our dataset is "Living people" which contains 703,532 articles for the English Wikipedia [8] (as of 2015/04/03) and hence, naturally many tweeted articles belong to this category. These distributions of articles and categories fact can also be related to our findings on the fact that bots are strongly represented within our dataset and as such, influence the distribution of articles being tweeted. In particular, bots send out a high number of tweets containing diverse and disparate articles strengthening the long tail distribution of tweeted articles. E.g., the Twitter account Wikipedia Stub Bot is the user with the highest number of tweets within the dataset (65,000 tweets within 4.5 months; cf. 4.1.1). This bot sends out tweets mentioning Wikipedia articles that have recently been created and asks for help in populating these articles. This indicates that these articles have not been edited before and hence, have not been tweeted before. To also reflect on how such tweets influence Wikipedia and to get an impression on how such tweets may influence the number of edits of a Wikipedia article, we analyze the number of edits performed of articles which have previously been tweeted by Wikipedia Stub Bot. Within the time window of 48 hours, we find that 68.21% of all articles posted experience a single edit—namely the creation of the respective article. This creation initially triggered the bot to publish a tweet about the respective article and hence, no edit followed within 24 hours of the publication of the tweet. 85.86% of all articles posted by this bot are edited less than five times within the given time window. Naturally, the population and editing of articles cannot solely be connected with tweets as there certainly are other channels to notice new articles or articles which require editing. However, we argue that the lack of edits suggests that the response of the community is limited.

6. CONCLUSION

In this work, we present an analysis on how Wikipedia is referenced within tweets based on a set of 4 million tweets. Our analysis shows that except for English and Japanese tweets, more than 20% of all tweets feature inter-language links. As for the topical analysis of tweeted articles, we find that these articles feature a longtail distribution and that 64% of all articles are only tweeted once. We find that the popularity of Wikipedia articles generally does not correlate with the number of edits of the respective article in a time window around the time the tweet was sent. However, we observe that events (e.g., the gamergate controversy) are reflected in the number of edits and the number of tweets about the event.

As for future work we want to further investigate inter-language links and look into why users make use of inter-language links. Furthermore, we are interested in analyzing to which extent certain events influence the popularity of Wikipedia articles on Twitter.

7. REFERENCES

[1] Alexa: The top 500 sites on the web `http://www.alexa.com/topsites/global`, accessed at 2015/05/27.

[2] Pageviews for All Wikimedia Projects (mobile + desktop) `https://dev.twitter.com/docs/api/1.1/post/statuses/filter`, accessed at 2015/05/26.

[3] Twitter: About Twitter `https://about.twitter.com/company`, accessed at 2015/05/26.

[4] Twitter API: Entities in Objects `https://dev.twitter.com/overview/api/entities-in-twitter-objects#urls`, accessed at 2015/05/26.

[5] Twitter: Introducing new metadata for Tweets `https://blog.twitter.com/2013/`

introducing-new-metadata-for-tweets, accessed at 2015/05/26.

[6] Twitter: Twitter Filter API https://dev.twitter.com/docs/api/1.1/post/statuses/filter, accessed at 2015/05/26.

[7] Wikimedia Foundation: Keep Wikipedia Free http://wikimediafoundation.org/wiki/Keep_Wikipedia_Free accessed at 2015/05/27.

[8] Wikipedia: Category:Living people http://en.wikipedia.org/wiki/Category:Living_people, accessed at 2015/05/28.

[9] Wikipedia: List of Wikipedias http://en.wikipedia.org/wiki/List_of_Wikipedias.

[10] Wikipedia says Emmanuel Sanders is dead, Sanders disagrees. *CBS*, 2014. http://www.cbssports.com/nfl/eye-on-football/24819294/wikipedia-says-emanuel-sanders\--is-dead-sanders-disagrees, accessed at 2015/05/26.

[11] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *43rd Hawaii International Conference on System Sciences (HICSS)*, pages 1–10, 2010.

[12] C. Dewey. The only guide to Gamergate you will ever need to read. *Washington Post*, 2014. http://goo.gl/Y0vr9L, accessed at 2015/05/26.

[13] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *IJCAI*, volume 7, pages 1606–1611, 2007.

[14] Y. Genc, Y. Sakamoto, and J. V. Nickerson. Discovering Context: Classifying Tweets through a Semantic Transform based on Wikipedia. In *Foundations of Augmented Cognition*, pages 484–492. Springer, 2011.

[15] X. Han, J. Liu, Z. Shen, and C. Miao. An Optimized k-nearest Neighbor Algorithm for Large Scale Hierarchical Text Classification. In *Joint ECML/PKDD PASCAL Workshop on Large-Scale Hierarchical Classification*, pages 2–12, 2011.

[16] L. Hong, G. Convertino, and E. H. Chi. Language Matters In Twitter: A Large Scale Study. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[17] N. Jullien. What we know about Wikipedia: A Review of the Literature analyzing the Project. *Available at SSRN*, 2053597, 2012.

[18] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth. User interests identification on twitter using a hierarchical knowledge base. In *The Semantic Web: Trends and Challenges*, volume 8465 of *Lecture Notes in Computer Science*, pages 99–113. Springer International Publishing, 2014.

[19] A. M. Kaplan and M. Haenlein. Users of the World, Unite! The Challenges and Opportunities of Social Media. *Business Horizons*, 53(1):59 – 68, 2010.

[20] A. Kittur and R. E. Kraut. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In *Proc. of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pages 37–46, New York, NY, USA, 2008. ACM.

[21] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages. In *International AAAI Conference on Weblogs and Social Media*, 2011.

[22] J. Leskovec, D. P. Huttenlocher, and J. M. Kleinberg. Governance in Social Media: A Case Study of the Wikipedia Promotion Process. In *International AAAI Conference on Weblogs and Social Media*, 2010.

[23] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. TwiNER: Named Entity Recognition in Targeted Twitter Stream. In *Proc. of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 721–730, New York, NY, USA, 2012. ACM.

[24] K. H. Lim and A. Datta. Interest Classification of Twitter Users using Wikipedia. In *Proc. of the 9th International Symposium on Open Collaboration*, pages 22:1–22:2. ACM, 2013.

[25] C. Lu, W. Lam, and Y. Zhang. Twitter User Modeling and Tweets Recommendation based on Wikipedia Concept Graph. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[26] MediaWiki API Main Page. http://www.mediawiki.org/wiki/API:Main_page, accessed at 2015/05/26.

[27] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: A first look. In *Proc. of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, pages 73–80, New York, NY, USA, 2010. ACM.

[28] F. Å. Nielsen. Wikipedia research and tools: Review and comments. *Available at SSRN 2129874*, 2012.

[29] M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis. Bieber no more: First Story Detection using Twitter and Wikipedia. In *Proc. of the Workshop on Time-aware Information Access. TAIA*, volume 12, 2012.

[30] S. Weissman, S. Ayhan, J. Bradley, and J. Lin. Identifying duplicate and contradictory information in wikipedia. *CoRR*, abs/1406.1143, 2014.

[31] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith. Finding Social Roles in Wikipedia. In *Proc. of the 2011 iConference*, pages 122–129, New York, NY, USA, 2011. ACM.

[32] Wikimedia Blog: Share a fact with friends on the Wikipedia Android app. http://blog.wikimedia.org/2015/04/02/share-a-fact-with-friends-on-android-app/, accessed at 2015/05/26.

[33] Wikipedia: Help:URL. http://en.wikipedia.org/wiki/Help:URL, accessed at 2015/05/26.

[34] Wikipedia: Tanka. http://en.wikipedia.org/wiki/Tanka, accessed at 2015/05/26.

[35] T. Xu and D. W. Oard. Wikipedia-based Topic Clustering for Microblogs. *Proc. of the American Society for Information Science and Technology*, 48(1):1–10, 2011.