A Clone-based Analysis of the Content-Agnostic Factors Driving News Article Popularity on Twitter

Alireza Mohammadinodooshan, William Holmgren, Martin Christensson, Niklas Carlsson
Linköping University, Sweden
Emails: firstname.lastname@liu.se

***Abstract*—The significant impact of Twitter in news dissemination underscores the need to understand what drives tweet popularity. While the content of an article plays a role, several "content-agnostic" factors also influence tweet popularity. Previous studies have faced challenges in differentiating the effects of content-agnostic factors from content variations. To address this, the paper presents a comprehensive analysis of tweet popularity using a "clone-based" approach. The methodology involves identifying tweets linking the same or similar articles (clones) and studying the factors that make some tweets within clone sets more successful in attracting retweets. The analysis reveals insights into clone set characteristics, winners' success patterns, retweet dynamics over time, domain-based competition, and predictors of success. The findings shed light on the complex nature of popularity and success in social media, providing a deeper understanding of the content-agnostic factors that influence tweet popularity.**

I. Introduction

With all major news outlets actively promoting their news on Twitter and the majority of all Americans receiving their daily news from social media [1], Twitter has come to play an important role in the dissemination of news. Due to the significant influence of social media, understanding the factors that make a tweet popular is therefore increasingly important.

However, determining the factors that contribute to a news article's popularity on Twitter, and even more so determining the content-agnostic factors that impact the retweetability of a tweet linking to such an article, remains a complex task. While the article's content, such as its interest, relevance, and quality, is important [2], it is widely acknowledged that several "content-agnostic" factors also influence popularity. For example, in the case of news articles shared on Twitter, content-agnostic factors like the number of followers of the poster or the length of the tweet can impact how many times such a tweet is retweeted and therefore how frequently it is included in other users' personal feeds or search results.

Previous studies have explored tweet popularity, examining static and temporal properties of retweet counts. However, understanding how content-agnostic factors impact popularity has remained challenging. For example, news outlets with large social networks may appear more popular because they typically share links to more interesting content, not due to the direct influence of social network size on tweet popularity. Existing studies with datasets containing tweets with links to news articles of diverse content struggle to rigorously differentiate the effects of content-agnostic factors from those arising from content variations. This is a significant shortcoming since not all news articles are the same, and the factors that impact the successful promotion of a news article on Twitter may be heavily influenced by the interest in the article itself.

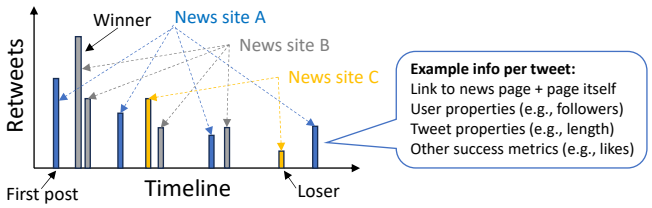


Fig. 1: Example clone set with posts linking news article clones posted by three news outlets.

To address the above shortcomings, in this paper, we present a comprehensive analysis of tweet popularity that accounts for the articles that are shared. In particular, we present a "clone-based" data collection and analysis (inspired by our prior work using clones to study YouTube popularity [3]), in which we first identify tweets linking the same article or a very similar article (which we call a "clone" of the original article), and then study what makes some of the tweets within such clone sets more successful in attracting retweets.

Fig. 1 illustrates the concept of a clone set. Here, distinct colors are used to denote tweets linking cloned versions of a news article published by different news outlets, while the order and height of the bars illustrate the relative timing and quantity of retweets for each such tweet, respectively. Using the clone concept, we can then control for the content and study which tweets are most successful and what content agnostic factors most influence a tweet's future success in generating retweets. For example, what factors most influence which tweet in a clone set will be the *winner*, and to what degree do we observe a pronounced *first-poster* advantage?

Our clone-based methodology offers important insights into content-agnostic factors affecting tweet popularity. As concrete examples, we next list ten example findings: (1) Clone set sizes and the number of website domains responsible for publishing

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11.
https://doi.org/10.1145/3625007.3627520

the articles show highly skewed distributions, with our results suggesting that a considerable portion of news stories are both replicated across numerous outlets and widely shared on Twitter. (2) The winners of big clone sets tend to receive more retweets, with the number of retweets following a power function. (3) Winners are predominantly posted early, but the first mover does not always obtain the most success. (4) Most clone sets link to a single domain, with the clone sets with most clones linking to NY Times, Forbes, and Bloomberg. (5) In clone sets with clones from different domains, Reuters was the most frequent winner, outperforming other domains. (6) The success of domains posting clones varies when competing against each other, with some domains frequently losing and others frequently winning. (7) The tweeter's characteristics play a significant role in the success of a clone, with winners and first posters typically having more followers and higher listing counts. (8) The length of the tweet text also influences success, with winners tending to use longer tweet texts. (9) Excluding public metrics such as likes, quotes, and replies (which also measure a tweet's popularity), the user follower count and the user verified status are the most important predictors of success, followed by tweet-related factors like the tweet count of the user and the tweet length. (10) Except for a smaller variation with Forbes, our domain-based analysis shows similar patterns across domains.

Overall, the study highlights the influence of various factors, such as clone set characteristics, winners' success patterns, retweet dynamics over time, domain-based competition, and predictors of success, shedding light on the complex nature of popularity and success in social media.

Outline: After describing our methodology for data collection and clone identification (§-II), we present a high-level characterization (§-III) of the winners and the relative success of different domains. We then study what factors most influence the success of a tweet (§-IV, §-V) before discussing related work (§-VI) and presenting our conclusions (§-VII).

II. Data Collection and Clone Identification

Clone Set Identification Framework: To collect clone sets, we developed a framework consisting of three main components: (1) a tweet retriever, (2) a text extractor, and (3) a clone finder. First, the tweet retriever retrieves tweets containing links to news articles using Twitter's API Academic Researcher product track. Here, we first obtained all tweets posted within an example timeline that contained an URL and then filtered the (resolved) URLs against the domains owned by a list of the most popular US news outlets. Second, the text extractor was used to extract the news article texts (but not figures, videos, etc.) from the URLs. Here, we used a combination of the open-source news-please Python module and a custom text extractor that we implemented for news websites with more complex structures (that news-please performed poorly on), as well as a per-domain specific crawler. Our custom-built crawler was built using the library Beautiful Soup. Finally, the clone finder module identifies potential clones by first grouping all tweets using the same URL and then applying a two-phase clone (or near-clone) identification approach on the extracted texts (when available). With the two-phase identification, we first use Simhash [4], a technique that generates 64-bit fingerprints for each text (64-bit has helped avoid collisions compared to 32-bit hashes), to check for similarity using a maximum hamming distance of 6 (ensuring a high recall), followed by calculating the pairwise cosine similarity on the vectorized texts (created using TF-IDF) of all pairs within a candidate cluster, so as to further refine the clone sets (and improve the precision). Our manual inspection showed that using a similarity threshold of 0.9 and combining these two phases (i.e., simhash for initial candidate clone identification followed by pairwise similarity tests within a cluster) reduces computational complexity (i.e., limits the required pairwise tests) and enhances accuracy (i.e., avoids unnecessary exclusion of potential clones, enabling more rigorous assessment in the subsequent cosine similarity).

TABLE I: Dataset overview

Age	Total tweets	Tweets in clone sets	Total clone sets	Largest clone set
1 year	1,398,359	1,219,244	75,902	10,165
1/2 year	1,128,696	988,331	70,773	4,057
1 month	1,021,421	883,816	65,151	6,673
1 week	928,587	811,558	62,684	9,146

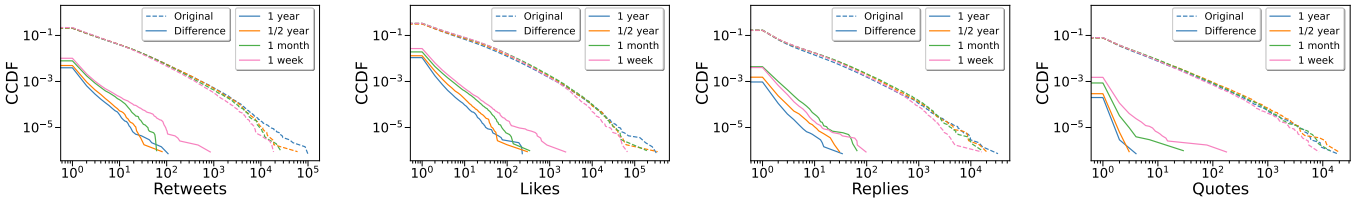
News Outlet Selection and Data Preparation: To select news outlets (for URL filtering and text extraction), we used the ranking lists of several independent rankings of US news outlets (e.g., *Allsides*, *opensources.co*, *pewresearch*, *statista*, *feedspot*, and *yougov*). The 69 selected news outlets represent a diverse range of topics, geographical locations, and audiences.

Datasets: Four datasets were collected based on the age of each post at the time of data collection (one-year old, half-a-year old, one-month old, and one-week old) and for each dataset we collected two snapshots: one when the posts are of the above listed ages and one that was collected one week later. In both cases, we collected all possible statistics about the tweet (including retweet statistics) and the tweeter of the tweet. The different aged datasets allowed us to analyze the effect of age differences on retweet behavior, while the retweet recollection one week later allows us to evaluate and reflect on the stability of the retweet counts over time.

Table I summarizes the size of the datasets. All datasets were collected over the week of March 2-8, 2021. Combined, the four datasets consist of 4.5M unique tweets including links to one of our predetermined URLs. Of these, most tweets (3.9M) are part of one of the 274,510 identified clone sets.

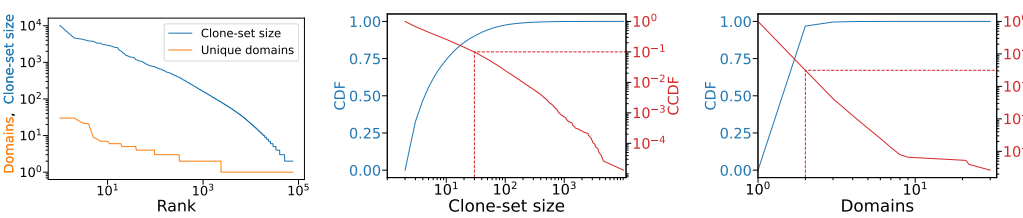
Success metric: To measure the successful spread of a tweet, we primarily use the number of retweets. This choice is motivated by the high importance of recommendations by friends and family (e.g., 83% believe more in such trust-earned advertisements than regular advertisements [5]) and world-of-mouth advertisement in general. We also show results for other public interaction metrics such as likes, quotes, and replies.

Dynamics of Tweet Interactions: Fig. 2 shows the Complementary Cumulative Distribution Function (CCDF) of the



(a) Retweets (b) Likes (c) Replies (d) Quotes

Fig. 2: CCDFs showing the per-tweet statistics for retweets, likes, replies, and quotes for each dataset.



(a) Rank plot (b) CDF/CCDF clone set sizes (c) CDF/CCDF unique domains

Fig. 3: Distribution of clone set sizes and domains for the one-year-old dataset.

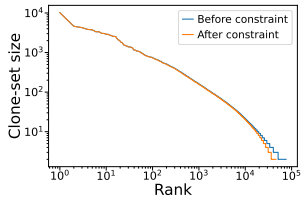


Fig. 4: Rank plot w/wo. constraint of unique winner.

number of retweets, likes, replies, and quotes for the original dataset ("Original") and the one-week gains between the two snapshots ("Difference"). Here, all curves are only slightly curved on log-log scale, suggesting heavy tailed distributions from a power-law-like family. Furthermore, the relative increases are small (especially for the datasets with older tweets), capturing the ephemeral nature of news and suggesting that most of the user interactions with these tweets already have taken place at the initial data collection. For example, except for likes (95% unchanged), 99% of the tweets see no change even for the 1-week old dataset, and for our primary metric (i.e., retweets), only 0.00001 of the tweets saw more than one hundred retweets during the second week.

Limitations: We acknowledge several limitations with our methodology. The findings may not generalize to other social media platforms. We do not consider the effect of Twitter's internal algorithms. The study is limited to linked news from the selected news outlets, which is based on public rankings but may not capture the full range of news sites. The text extraction process is not perfect and sometimes struggles with some pages with complex structures or that otherwise prohibit access. The choice of thresholds and parameters in the clone detection process (e.g., simhash hamming distance, cosine similarity threshold) is determined through manual evaluation and may not be optimal in all cases. Nevertheless, our manual sanity checking suggests that the methodology is able to achieve high accuracy and deliver clear clone sets.

III. High-level Characterization

Before analyzing the characteristics of a typical winner or determining the content-agnostic factors that most impact the success, we first provide a high-level characterization of the dataset and the relative success that different clones achieve.

Clone set sizes highly skewed: Fig. 3 shows a rank plot, the empirical cumulative distribution function (CDF), and the Complementary CDF (CCDF) of the clone set sizes and number of website domains responsible for publishing the articles. (While these stats are for the one-year old dataset, the relative shape of the distributions for all datasets are similar.) From the rank plot, we note that the biggest clone set consists of 10,165 clones (top ranked entry in Fig. 3(a)) and from the CCDFs we note that 10% of clone sets consists of more than 30 clones (red line in Fig. 3(b)) and 3% link to articles published on at least 2 domains (red line in Fig. 3(c)).

A. Winner-based Analysis

To be called "winner" of a clone set we required the tweet to have more retweets than the 2nd-most retweeted clone. While this resulted in a slight reduction of the number of clone sets it did not change our distribution statistics much (see Fig. 4).

Winner success follows power function of the clone-set size: As expected, the winners of big clone sets tend to be retweeted more. What is interesting is that the number of retweets that they obtain follows a power function of the clone-set size. This is seen by the straight-line characteristics of the green markers in Fig. 5, which shows the median number of retweets for the winners (1st rank) for different clone-set sizes. We also note that the majority of clone sets with less than 4 are not retweeted. These results show that there is a strong relationship between what is cloned and what is popular to retweet. We have also seen noticeable differences in the median number or retweets obtained by the winner (1st; green marker) and 2nd placed clone (blue marker). Here, it should be noted that both axes are on log scale.

Retweets and winners over time: To illustrate how tweets and retweets associated with specific clone classes within a clone set are distributed over time, we split the active time period of a clone set into five equal time periods. We call these time periods bins and count the number of tweets posted in each time bin that are of one of two tweet classes (tweets or retweets) and that are of one of the clone classes ("winners" or "losers"). Fig. 6 shows the percentage of each of these four tweets subsets that fell into each of the five bins. We note that 71% of all retweets (across all clone sets) occur during the

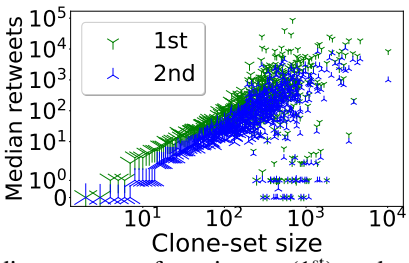


Fig. 5: Median retweets for winners (1st) and second places clones (2nd) for a one-year-old dataset. The size of each marker represents the number of clone sets of that size.

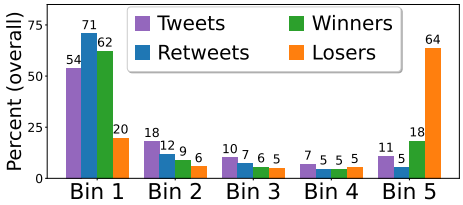


Fig. 6: Timing of all tweets and retweets, as well as "winners" and "losers". Here, the time bins are split equally between the first and last tweet observed within a clone set.

first time bin (20% of the lifetime of the clone sets) compared to only 54% of the origin tweets. It is therefore not surprising that most winners (62%) are posted during the first time bin. It is however also clear from the plot that it is not always the first mover (i.e., the clone that make the first tweet, and hence always in the first bin) that obtains the most success. The reason we see somewhat more winners (and losers) in the last bin (than in bins 2, 3, 4) is due to cases where we only had two clones in a clone set (each assigned to bins 1 and 5).

The results above are relatively consistent across the datasets and filtering methods used. One reason for this is the activity associated with most clone sets is short-lived. For example, in our one-year-old default dataset, the median and average activity interval (that we break into five bins) are 30 and 48 hours, respectively, whereas these times only reduce to 24 and 42 hours, respectively, for the one-week-old dataset. For the 1- and 6-month-old datasets the medians (26/25 hours) and the averages (41/41 hours) are relatively similar.

B. Domain-based analysis

We next compare the relative success of different domains as seen (1) across all clone sets and (2) across only the clone sets that contain competing clones associated with different domains. Throughout the paper, we call the second group of clone sets "mixed" clone sets. For each of these two cases, Fig. 7 shows the percentage of times each domain was responsible for the clone that was the "winner" or "first poster". We also consider the case when the combined set of clones of a particular domain collectively garnered more retweets than that from any other domain ("most total").

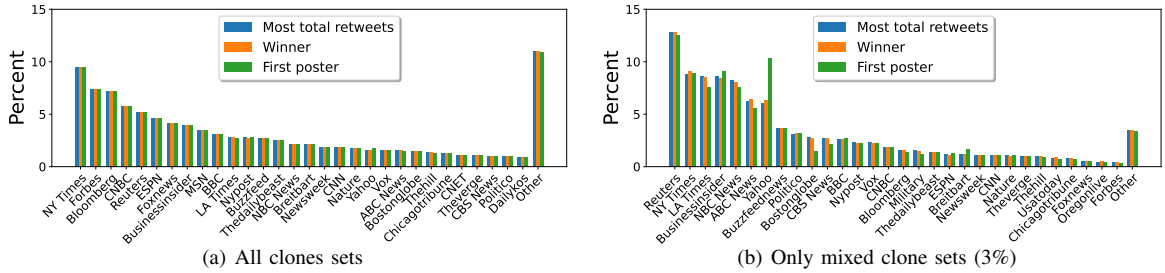
With 97% of the clone sets only linking to a single domain, it is not surprising that there are no major differences between the three metrics seen in the "all" case (i.e., Fig. 7(a)). The "all" figure instead captures the relative number of clone sets dominated by a domain, with the top-3 domains being NY Times, Forbes, and Bloomberg.

Winners when competing only against links to the same domain: Before looking at the mixed clone sets, let us first look closer at the "winners" and "first posts" of the top-7 domains in the non-mixed clone sets. For each of these domains (one domain per row), Fig. 8 shows CDFs for each of the following metrics: (1) the ratio of retweets of the "winner" and "first post" of each such clone set, (2) the ratio of retweets of the "winner" and the post that achieved the median number of retweets, (3) the ratio of retweets of the "winner" and the average number of retweets per post, (4) the ratio of retweets of the "winner" and the "loser", and (5) the relative fraction of retweets that the "winner" obtained. To aid comprehension, we provide a baseline (top row) that represents the results for all clone sets. The plots are color-coded to indicate their relative performance compared to this baseline. A green background signifies a larger median (y=0.5) value than the baseline, while blue values indicate smaller medians. Median values and their relative differences are depicted with orange dotted lines and percentage values, respectively. Reuters exhibits the largest relative improvements compared to the baseline across three metrics ("winner/first", "winner/median", and "winner/loser"). This suggests that winning posts from Reuters are often re-posts, surpassing the "median" and "loser" by a significantly greater margin compared to winners from other domains. Conversely, for Forbes, Buzzfeed, Bloomberg, and CNBC, the majority of winners are the "first posts", indicated by the vertical line in the first column of the plot, and their performance does not outshine the others to the same extent.

Mixed clone set analysis of competing domains: The relative rankings become more interesting when considering mixed clone sets (Fig. 7(b)) and how the ranking changes compared to the full dataset (Fig. 7(a)). Here, the most frequent "winner" (and "first poster") is Reuters, which goes from being ranked 6th to 1st and being responsible for almost 50% more winners than the 2nd ranked domain (NY Times). We also see some big drops in the rankings (e.g., Forbes went from ranked 2nd to last of the ranked domains) and a few cases with noticeable differences between the metrics.

Perhaps most noticeably, considering the mixed clone sets, Yahoo is the 2nd most frequent "first poster" but only achieves the 7th best "winning" percentage. This highlights that it does not always pay off being the first to post on a topic. One reason for this is that Yahoo clones tended to have relatively fewer retweets in general than the other top publishers. This is illustrated in Fig. 9, where we show the CDFs and CCDFs of the total number of retweets obtained across all clones linking the top-7 domains (blue curve) and their winning clone (red curve) together with the 75%-ile values. We note that domains with more "first poster" instances than "winner" instances (Yahoo and Business Insider) obtained the fewest retweets, whereas the opposite was true for the two domains that obtained the most retweets (NBC News and ABC News).

Head-to-head competition: The difference in the domains' relative success when their associated clones are posted also



(a) All clones sets (b) Only mixed clone sets (3%)

Fig. 7: Frequency that a domain had the most retweets in a clone set.

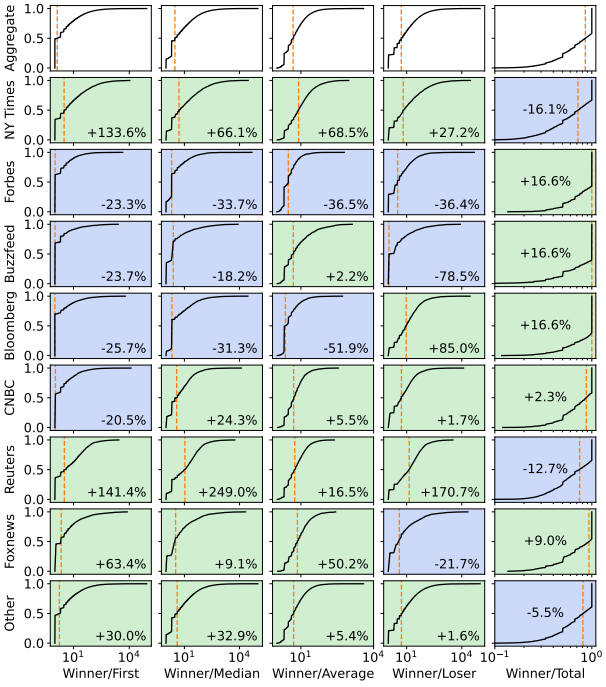


Fig. 8: CDFs of relative "winner" comparisons for the top-7 domains. Green/blue indicate higher/lower median than baseline (top row) and percentage values show the relative median differences compared to the median baseline.

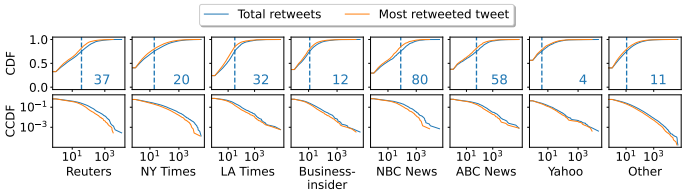
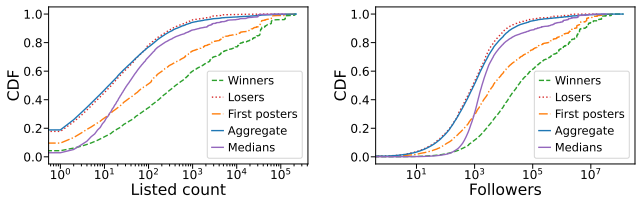


Fig. 9: CDF and CCDF of total retweets and most retweeted tweet in clone sets for each domain. The 75%-ile values are represented by the blue dotted line and a number.

became evident when comparing their winning percentages going head-to-head. Table II shows the fraction of times that a domain (row) won over another domain (column). We again use the top-7 domains in the mixed clone sets. Here, we use darker colors to indicate a bigger win fraction for the domain listed on the row over the domain listed on the column. Again, Yahoo and Business Insider are the most likely to lose (mostly yellow rows and red columns). Among the winners, Reuters



(a) Listed count (b) Followers

Fig. 10: CDFs of the two most impactful tweeter characteristics for different subsets of clones.

and LA Times stand out, with the biggest portions of pairwise wins (mostly red rows and yellow columns).

IV. What Makes a Winner?

A. Single-factor Analysis

We next discuss factors that were found to increase the success probability of a post. Here, Table III provides a summary of the percentage of clone sets where the "winner" had more or equal number/quantity of the variable of consideration than the "first post", "loser", and "median" clone. Furthermore, we split variables into three categories based on whether the variable captures the characteristics of the tweeter ("User"), the tweet itself ("Tweet"), and four measures of success ("Success"). For simplicity, we do not include less-good predictors, and in the following, we discuss the variables for which the "winner" had at least the same value for $\geq$80% of the samples and the metric was available for $\geq$10K comparisons.

Tweeter of a clone: The person/account responsible for a clone plays a big factor in its probability of success. While such correlations can be found in regular datasets, the role of the origin tweeter is perhaps more clearly and fairly captured when working with clone sets, as they neutralize the effects of the shared content.

To illustrate the effects of the tweeter characteristics, Fig. 10 shows CDFs of the number of listings and followers associated with the tweeters of the different clone categories. It is clear from the significant shift in the follower curves (note log-scale) that the accounts behind the "winners" (most retweeted clone), followed by the "first poster", compared to the other categories, that the "winning" accounts often already have built part of their success before the time of posting the tweets. By having attracted followers that will see their tweets, they are more likely to succeed also with future tweets and benefit from the rich-gets-richer effects (from the perspective

TABLE II: Head-to-head wins/competitions for one-year-old dataset. Darker cells indicate bigger win percentage for the domain listed for that row when competing with the domain listed for that column.

	Reuters	NY Times	LA Times	Businessinsider	NBC News	ABC News	Yahoo	Other
Reuters	–	11/13	0/1	15/15	5/9	5/7	187/194	64/84
NY Times	2/13	–	1/8	29/32	15/18	8/10	6/12	144/274
LA Times	1/1	7/8	–	5/5	4/8	17/27	140/147	38/56
Businessinsider	0/15	3/32	0/5	–	1/6	1/7	146/160	32/112
NBC News	4/9	3/18	4/8	5/6	–	11/12	136/141	32/60
ABC News	2/7	2/10	10/27	6/7	1/12	–	93/114	61/172
Yahoo	7/194	6/12	7/147	14/160	5/141	21/114	–	111/339
Other	20/84	130/274	18/56	80/112	28/60	111/172	228/339	–

TABLE III: Percentage of clone sets where the "winner" has higher or equal value for the considered variable relative "first post", "loser", and "median" clone. Except for cashtags (500 samples) and tweet video views (200 samples), both marked (*), we always had 10K+ sample comparisons (median 45K).

	Variable	First poster	Loser	Median
User	**Followers**	86.5	87.6	89.6
	Listed count	83.6	83.7	86.5
	Following count	76.2	63.2	67.1
	User age days	74.8	67.5	71.3
	Number of tweets	73.5	68.3	72.2
Tweet	Tweet video views (*)	96.2	93.8	96.6
	Tweet text length	80.0	65.9	72.6
	Tweet age days	72.2	90.2	86.3
	Hashtags	71.2	59.6	86.5
	Mentions	69.7	54.2	76.4
	Cashtags (*)	65.4	54.8	94.1
Success	**Retweets**	100.0	100.0	100.0
	Likes	96.4	96.6	98.4
	Quotes	94.6	97.2	99.6
	Replies	93.3	93.4	98.7

of the tweeters). What is perhaps a bit more surprising is that the "first poster" often is a user that has relatively more listings and followers. The finding suggests that initial posters often have a substantial following, indicating that users who consistently share original content are more likely to gain followers over time. It is important to note that some "first poster" clones were shared by prominent news accounts promoting their own articles.

Tweet lengths: Consider next the tweet itself. Here, the use of clones for head-to-head comparisons becomes even more important (in part because the tweet sizes are more uniform). However, when comparing clones, we have found that the "winners" tend to use somewhat longer tweet texts than the other tweet categories. This is shown by the shift in the CDF of the "winner" category relative to the other CDFs (Fig. 11(a)) and the fact that this category has a significant larger fraction of tweets close to Twitter's 280-character limit (e.g., much sharper increase around this point). The success of longer tweets is also visible in the heat-map scatter plot (Fig. 11(b)), showing the number of tweets of different text lengths that obtained a certain number of retweets. Larger tweets exceeding the 280-character limit can be attributed to Twitter's inclusion of previous tweets in reply chains as mentions, which do not count towards the character cap [6]. Figure 11(c) demonstrates how this feature may enhance the popularity of longer tweets.

Public success metrics: Finally, we compare how much more success the "winners" achieved than the other categories using each of the four success metrics: retweet count (our default success metrics), likes, quotes, and replies. Fig. 12 shows both CDFs and CCDFs for each of these metrics. As expected, the "winners" outperformed the other categories noticeably with regards to all four metrics, and the "first post" always achieved the 2nd best of the five categories.

V. Modeling Success

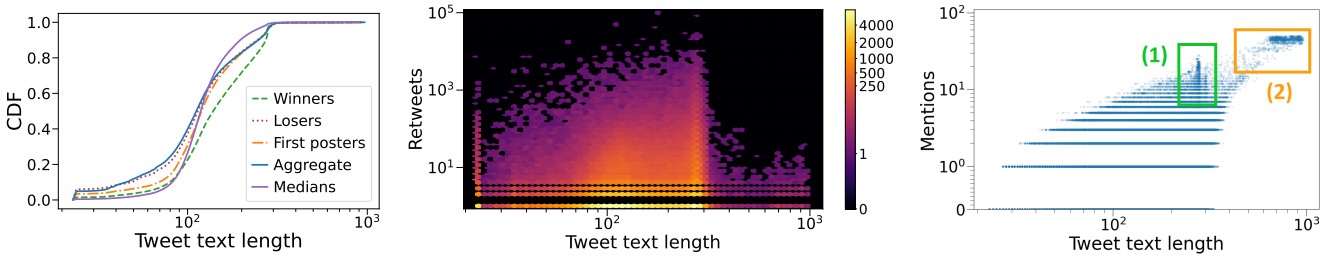
Building upon the empirical findings from the previous section, we next model tweet success using linear regression.

Pairwise correlations and multicollinearity: Before building regression models that take into account the clone sets, we first consider the correlations of the observed variables when ignoring clone set belonging. Fig. 13 shows the pairwise Pearson correlation coefficients as a matrix, with the matrix sorted so that the metrics are sorted from the metric with least-to-the-most correlation with the number of retweets. As expected, the best highest correlated metrics are the three other success metrics: likes, quotes, and replies.

Preliminary test models for the full dataset then indicated the presence of multicollinearity. To quantify collinearity, we calculated the Variance Inflation Factor (VIF) for each variable within clone sets. Although mentions and hashtags showed a slightly higher correlation with tweet text length, we decided to include them in the final model, considering their different variations. To mitigate multicollinearity, we used adjusted R^2 as selection criterion and applied all possible subset regression.

Best subset analysis: For each clone set, we conducted a best subset analysis, selecting models with the highest adjusted R^2 for each variable count. For the analysis presented here, we will exclude the three other success variables.[1] Fig. 14 shows the best subset analysis for clone sets without the three public success metrics. Here, the user follower count emerged as the most important predictor, included in 89% of models, followed by "user verified" (75% inclusion). Both predictors exhibited more significance, with a higher proportion of low p-values. User tweet count (70% inclusion) and user following count (66% inclusion) were the third and fourth most frequent predictors, with 28% significance each. Tweet text length

[1]In the case we applied it on all variables, the three most important variables were the other success metrics. In particular, the like count was the most significant predictor, present in 99% of the models and ranked as the most important variable in 98% of the models, the quote count was included in 85% of the models, and the reply count appeared in 65% of the models but only showed significance in 24%.



(a) CDFs lengths (b) Heat-map retweets (c) Scatter plot mentions

Fig. 11: Tweet text length plots for a one-year-old dataset.

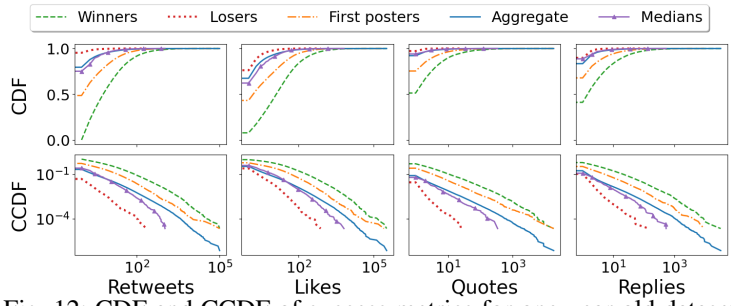


Fig. 12: CDF and CCDF of success metrics for one-year-old dataset.

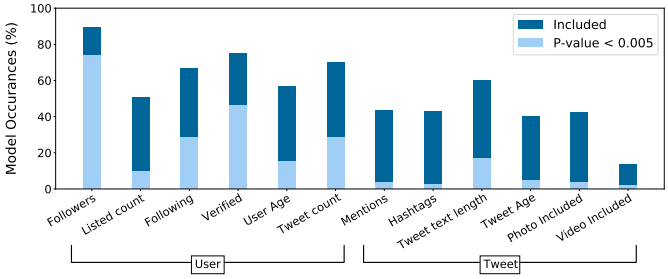


Fig. 14: Percentage of model occurrences derived from all clone sets through best subset selection (without public success metrics). Full bars show all occurrences, and the light colored parts indicate those where the predictor had p-value < 0.005.

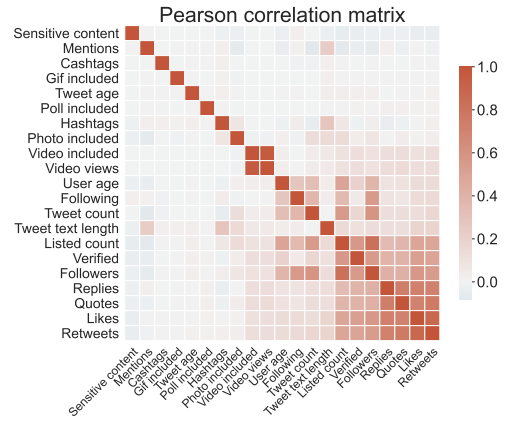


Fig. 13: Pearson correlation for different factors, sorted by correlation to retweet count.

appeared in more models but lacked significance in most, likely due to its correlation with mentions and hashtags. Photo included had limited significance, and video included was rarely selected. Models with 12 predictors performed worse on average compared to those with fewer predictors.

Domain-based models: Table IV presents the per-domain results for the clean clone sets, excluding public metrics. Once again, the most significant predictor is the user follower count, followed by user verified. The inclusion of user tweet count and user following count is closely balanced. However, in the clean subset, we observe that the variable "user verified" for the domain "Forbes" has a significantly lower inclusion rate, appearing in only about 40% of the best models. This contrasts with other domains, where it has a higher inclusion rate. Interestingly, in the mixed subset (results omitted due to space), "Forbes" exhibits a substantially higher value for "user verified", surpassing other domains. This suggests that the identity of the "Forbes" domain has a significant impact on the "user verified" variable, as the clean subset shows a notably lower inclusion rate compared to the mixed subset. Except for the user verified variable, the results do not significantly differ between the domains or between the mixed and clean analysis results, suggesting that there is no substantial difference in how tweet variables relate to the URL domain identity mentioned in the tweet. Finally, we note that "video included" appears to be the weakest indicator of success, consistently performing poorly. This finding aligns with the results shown in Fig. 14.

VI. Related Work

This work aligns with research exploring factors influencing post popularity on different social media platforms. Studies have investigated how content impacts engagement and diffusion. For example, [7] demonstrated the effect of emojis on user engagement while [3] revealed YouTube popularity factors when controlling for near-identical videos. A survey of this research line is available in [8].

Focusing on Twitter, studied here, various works have examined the impact of different factors. [2] highlighted content as an important factor for gaining retweets. [9] found that URLs and hashtags in the content have strong relationships with retweetability, and [10] showed that the subject of a tweet is the most informative content-based feature.

Contextual features also play a significant role. The number of followers and followees, as well as the account's age, seem to affect retweetability [10]. In a related study, [11] explored tweet features, including content and contextual factors, to

TABLE IV: Percentage of model occurrences derived from clones sets through best subset selection, with the exclusion of public metrics. Values with darker color indicate a higher percentage of model inclusion.

Domain	Bloomberg	Buzzfeed	CNBC	Forbes	Fox News	NY times	Reuters	Other
User followers count	0.910	0.839	0.862	0.939	0.901	0.891	0.897	0.898
User listed count	0.506	0.648	0.507	0.492	0.599	0.469	0.425	0.512
User following count	0.627	0.581	0.610	0.508	0.533	0.721	0.691	0.640
User verified	0.722	0.614	0.738	0.409	0.697	0.765	0.761	0.752
User age (days)	0.580	0.545	0.581	0.644	0.577	0.557	0.542	0.579
User tweet count	0.710	0.711	0.683	0.720	0.706	0.705	0.752	0.673
Mentions	0.423	0.518	0.479	0.614	0.424	0.431	0.432	0.443
Hashtags	0.432	0.366	0.467	0.515	0.505	0.416	0.430	0.439
Tweet text length	0.608	0.648	0.564	0.598	0.628	0.629	0.493	0.588
Tweet age (days)	0.355	0.555	0.419	0.515	0.400	0.392	0.399	0.407
Photo included	0.537	0.366	0.476	0.455	0.411	0.418	0.364	0.444
Video included	0.090	0.108	0.114	0.091	0.120	0.106	0.314	0.125

predict retweet counts and investigated retweet rates across diverse news domains. The timing of posts also impacts engagement [9], [12]. As an example, [9] noted a shift in engagement from virtue to vice content as the day progresses.

Community structure and social influence are other key factors in popularity. For example, [13] underscored the influence of community structure on popularity while [14] tried to unify social influence and homophily in popularity prediction. [15] analyzed how sharer characteristics affect the recurrence of popularity evolution. Similarly, [16] highlighted the user's personality and role in tweet popularity.

Focusing on user behavior, [17] studied the survival timing of articles from 12 news sources. [18] categorized news sources on Twitter based on user sharing behavior and examined the engagement of different topics and user characteristics. [19] investigated the diffusion of true and false news stories on Twitter, concluding that false news diffused more and that humans were primarily responsible for spreading false news. [20] found that many tweets are retweeted more than the links are clicked, indicating differences in engagement.

While previous work indicates that contextual features are more effective than content features [21], an area still lacking sufficient focus is understanding popularity factors while controlling for identical content. This paper addresses this gap. As Borghol et al. [3] demonstrated for YouTube, controlling for similarity reveals new insights, making this analysis on Twitter data novel, valuable, and distinct from previous works.

VII. Conclusion

In conclusion, our clone-based analysis of tweet popularity on Twitter has uncovered important insights into the content-agnostic factors that influence the dissemination of news articles. We have identified patterns in clone set characteristics, retweet dynamics, domain-based competition, and predictors of success. The study highlights the significance of factors such as clone set size, winner characteristics, tweet text length, and user metrics like follower count and verified status. These findings provide valuable guidance for news organizations and social media users aiming to maximize the impact of their tweets. Overall, our research deepens our understanding of tweet popularity and contributes to the broader understanding of social media dynamics.

References

[1] A. Mitchell, E. Shearer, and G. Stocking, "News on Twitter: Consumed by Most Users and Trusted by Many," https://www.pewresearch.org/journalism/2021/11/15/news-on-twitter-consumed-by-most-users-and-trusted-by-many/, 2021, accessed: 2023.

[2] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *Proc.ICWSM*, 2010.

[3] Y. Borghol, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "The untold story of the clones: Content-agnostic factors that impact YouTube video popularity," in *Proc. ACM KDD*, 2012.

[4] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proc. ACM symposium on Theory of computing*, 2002.

[5] N. Nielsen, "Global trust in advertising," *The Nielsen Company*, 2015.

[6] "Counting characters," https://developer.twitter.com/en/docs/counting-characters, (Last accessed 2023).

[7] E. E. Ko, D. Kim, and G. Kim, "Influence of emojis on user engagement in brand-related user generated content," *Computers in Human Behavior*, vol. 136, p. 107387, 2022.

[8] D. Jatain, V. Singh, and N. Dahiya, "A multi-perspective micro-analysis of popularity trend dynamics for user-generated content," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 147, 2022.

[9] O. Zor, K. H. Kim, and A. Monga, "Tweets we like aren't alike: Time of day affects engagement with vice and virtue tweets," *Journal of Consumer Research*, vol. 49, pp. 473–495, 10 2022.

[10] M. Mahdavi, M. Asadpour, and S. Ghavami, "A comprehensive analysis of tweet content and its impact on popularity," in *Proc. IST*, 2016.

[11] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network," in *Proc. IEEE SocialCom*, 2010.

[12] S. Petrovic, M. Osborne, and V. Lavrenko, "Rt to win! predicting message propagation in Twitter," in *ICWSM proc.*, vol. 5, no. 1, 2011.

[13] S. Tsugawa, "Empirical analysis of the relation between community structure and cascading retweet diffusion," in *Proc. ICWSM*, 2019.

[14] Y. Shang, B. Zhou, X. Zeng, Y. Wang, H. Yu, and Z. Zhang, "Predicting the popularity of online content by modeling the social influence and homophily features," *Frontiers in Physics*, vol. 10, p. 915756, 2022.

[15] J. Cheng, L. A. Adamic, J. M. Kleinberg, and J. Leskovec, "Do cascades recur?" in *Proc. WWW*, 2016.

[16] S. Firdaus, C. Ding, and A. Sadeghian, "Retweet prediction considering user's difference as an author and retweeter," in *Proc. ASONAM*, 2016.

[17] D. Bhattacharya and S. Ram, "Sharing news articles using 140 characters: A diffusion analysis on Twitter," in *Proc. ASONAM*, 2012.

[18] M. Samory, V. Abnousi, and T. Mitra, "Characterizing the social media news sphere through user co-sharing practices," in *Proc. ICWSM*, 2020.

[19] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, pp. 1146–1151, 2018.

[20] J. Holmström *et al.*, "Do we read what we share? analyzing the click dynamic of news articles shared on Twitter," in *Proc. ASONAM*, 2019.

[21] M. G. Silva, M. A. Domínguez, and P. G. Celayes, "Analyzing the retweeting behavior of influencers to predict popular tweets, with and without considering their content," in *Proc. SIMBig*, 2018.