Social Media and Information Consumption Diversity

José Devezas Sérgio Nunes

INESC TEC and Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
{jld,ssn}@fe.up.pt

Abstract

Social media platforms are having a profound impact on the so-called information ecosystem, specifically on how information is produced, distributed and consumed. Social media in particular has contributed to the rise of user generated content and consequently to a greater diversity in online content. On the other hand, social media networks, such as Twitter or Facebook, have become information management tools that allow users to setup and configure information sources to their particular interests. A Twitter user can handpick the sources he wishes to follow, thus creating a custom information channel. However, this opportunity to create personalized information channels effectively results in different consumption profiles? Is the information consumed by users through social media networks distinct from the information consumed though traditional mainstream media? In this work, we set out to investigate this question using Twitter as a case study. We prepare two samples of users, one based on a uniform random selection of user IDs, and another one based on a selection of mainstream media followers. We analyze the home timelines of the users in each sample, focusing on characterizing information consumption habits. We find that information consumption volume is higher, while diversity is consistently lower, for mainstream media followers when compared to random users. When analyzing daily behavior, however, the samples slightly approximate, while clearly maintaining a lower diversity for mainstream media followers and a higher diversity for random users.

1 Introduction

Social media has become a part of our modern lives and a central service for information consumption, covering a wide range of topics, from personal events to worldwide news. Several studies [CHBG10, KWM11, MJA+11, LKSM14, CSR14] have focused on the study of social media through the characterization of users, usage patterns and content production. In this work, we take advantage of Twitter to study content consumption, giving particular attention to the characterization of the consumption patterns of news followers. As an information diffusion service, Twitter is frequently used for news broadcasting, either by citing a mainstream media news article, or even by directly serving as a communication channel to broadcast the news events themselves. Some studies have compared the content generated in Twitter with the content generated by mainstream media. These studies frequently focus on a collection of tweets, usually retrieved from the Stream API, and a collection of news articles from well known newspapers, for a common period of time. However, there are fewer studies that focus on analyzing the content consumed by each Twitter user on its own timeline and, to our knowledge, no study that distinguishes the content followed by Twitter users interested in mainstream media from the content followed by the majority of Twitter users.

In this work, we studied the home timelines of a collection of Twitter users, in order to understand the type of content that users follow on Twitter. Particularly, we were interested in comparing the general Twitter population with a specific group of mainstream media consumers. Our goal was to investigate to what degree the timeline of each Twitter user, i.e.

In: D. Albakour, D. Corney, J. Gonzalo, M. Martinez, B. Poblete, A. Vlachos (eds.): Proceedings of the NewsIR'18 Workshop at ECIR, Grenoble, France, 26-March-2018, published at http://ceur-ws.org

the information to which the user is exposed to, differs from the timelines' of other users. In other words, to understand if the experience of each user is unique or, on the contrary, if that experience is similar to that of other users. To achieve this goal, we characterized the anatomy of each individual timeline, presenting aggregated results per sample and studying the diversity of consumed information, overall as well as over time.

2 Reference Work

Bache et al. [BNS13] proposed a text-based framework for quantifying the diversity of documents based on their terms. Their approach was based on the application of Latent Dirichlet Allocation [BNJ03], to build a topic model for a given corpus, and the computation of the distance matrix between pairs of topics, using measurements such as topic co-occurrence and topic-word similarity. They estimated the diversity for each document, in relation to the corpus, by combining the distance matrix with the topic distribution for the document.

White and Jose [WJ04] evaluated several measurements of topic similarity, grouping them into association (Dice, Jaccard, Cosine, Overlap), correlation (Spearman, Kendall, Pearson), and distance (Euclidean, L_1 norm, Kullback-Leibler). For assessment, they used topics 101-150 from the TREC and the San Jose Mercury News 1991 collection. They pre-selected 10 topics, ensuring a variable number of overlap between the most relevant documents for each topic, and asked a group of 76 subjects to evaluate the similarity between each pair of topics using a 5-point scale (from highly dissimilar to highly similar). While the evaluation was done for only ten topics, according to their study, the most useful measurement group was the correlation, followed by the association group and, only then, the distance group.

Zhao et al. [ZJW$^+$11] compared Twitter and mainstream media using topics models. They used a sample of the Edinburgh Twitter Corpus [POL10], originally collected from the Stream API and crawled news articles from the New York Times using their search function. Both datasets comprised documents for the timespan between November 11, 2009, and February 1, 2010. They used Latent Dirichlet Allocation to directly extract topics from the New York Times dataset, but, given the small size of tweets, they proposed a custom Twitter-LDA algorithm for topic detection in the Twitter dataset. In order to compare Twitter with mainstream media, they labeled detected topics using the categories provided by the New York Times, which they had to manually assign to their Twitter dataset. Moreover, they used three topic types to distinguish topics: event-oriented, entity-oriented, and long-standing. By looking at the distributions of topic categories and types, they discovered that Twitter provides more entity-oriented topics with low coverage on mainstream media, and that, although Twitter shows a low interest in world news, it helped spread news of important world events. The study we present here is similar in the sense that we also focus on understanding the position of mainstream media regarding Twitter, but it is also different in the sense that we keep our focus on Twitter, distinguishing between the home timelines of random users and the home timelines of mainstream media followers. Our study is centered around the individual (per user) consumption of content, for two distinct samples of users, as opposed to simply comparing the overall topics present in social media versus mainstream media. In particular, we are interested in studying the differences between the content that Twitter users are exposed to in their personal timelines.

There are multiple metrics that can serve as a diversity index [Jos06, Table 1], including True Diversity, Richness, Shannon Index, Simpson Index and Berger-Parker Index. Most diversity metrics are transformations of the effective number of types and have a particular interpretation dependent on the context of application. Our approach to studying diversity is based on the direct comparison of home timelines from individual users from two samples: one collected randomly and another one collected based on the preference to follow mainstream media accounts (i.e. users that share a common interest). We then analyze the cosine distances between all pairs of accounts within each sample, in order to quantify divergent behavior and thus estimate diversity.

3 Data Collection

In order to analyze the differences between the content that random users and mainstream media followers consume on their Twitter home timelines, we needed to indirectly obtain a sample of user home timelines. Given Twitter does not provide this feature directly through its API, our approach consisted on the following five steps:

1. Collect a sample of 20 users by generating random user IDs between 1 and the largest known user ID, from a recently created user.

2. Collect a sample of 20 users that follow at least 3 UK news accounts from the following list: @BBCNews; @guardian; @Telegraph; @Independent; @MailOnline; @DailyMirror; @TheSun; @daily_express; @metrouk; @daily_star; @standardnews;

3. For each collected user, fetch their followed accounts.

4. At the same time, for each followed account, fetch and store all their tweets for the past 14 days.

5. Locally, for each collected user, retrieve its stored followed account timelines, ordered by decreasing date, thus rebuilding the home timelines per user.

Each collected user, described in steps 1 and 2, was subject to a set of criteria to ensure a minimum level of expected activity and connectivity of the accounts (a basic check to discard inactive users):

- The user must have created at least one tweet in the last three months.
- The user must have at least three followers.
- The user must have created at least five tweets since the creation of the account.

The data was stored in an SQLite database. In order to define and describe each user sample, we used a "user_samples" table where we stored groups of user IDs, identified by a common sample ID. Each "user_sample" entry also contained a textual description detailing the data collection approach, as well as the user selection criteria (e.g., "Random users, generated by a random uniform sampling of Twitter user IDs between 1 and 3954358701, restricting language to 'en', last tweet date to 2015-07-15 16:45:43, follower count to 3 and status count to 5.").

In this paper, we characterize and compare the timelines for two user samples: "Sample Random 20", which represents the baseline as a collection of random Twitter users, and "Sample UK News Followers 20", which represents a particular group of users who have shown a general interest in mainstream media by following well-known UK news accounts.

4 Data Characterization

Overall, our collection contains 5,287,221 distinct tweets. However, as different accounts frequently have followed accounts in common, the timelines overlap, resulting in 7,758,779 analyzable tweets when looking at individual home timelines. "Sample Random 20" contains 947,068 distinct tweets, resulting in 1,080,789 (13.93%) of the overall analyzable tweets. "Sample UK News Followers 20" contains 4,685,800 distinct tweets, resulting in 6,677,990 (86.07%) of the overall analyzable tweets. Distinct tweets in "Sample Random 20" and "Sample UK News Followers 20" intersect, resulting in 345,647 common tweets. Users from "Sample Random 20" follow a total of 11,807 distinct users; on average, each user follows 621.42 users. Users from "Sample UK News Followers 20" follow a total of 22,082 distinct users; on average, each user follows 1,104.10 users.

The tweets for each user's followed account were collected for a period of 14 days, with slightly different start dates, resulting in an overall larger period of 55 days, from Jul 19 2016 to Sep 12 2016. The timespan for the home timelines of the 40 users in both samples only overlapped for a period of 13 consecutive days, from Jul 20, 2016, to Aug 2, 2016. We analyzed the average number of tweets over time, per day and per hour, respectively, for each sample. While "Sample Random 20" is moderately stable per day, with a coefficient of variation of 29.1%, "Sample UK News Followers 20" shows a more evident growth in the number of tweets, peaking at Jul 29 and having a coefficient of variation of 42.0%. Regarding the average number of tweets per hour, the maximum number of tweets for "Sample Random 20" was generated at 20:00 UTC, Jul 23, 2016 and at 16:00 UTC, Aug 1, 2016, worldwide, for "Sample UK News Followers 20", with coefficients of variation of 33.3% and 42.9%, respectively.

5 Information Consumption

When social media paved the way for pervasive communication, people became both producers and consumers. This introduced a shift in habits with potential implications to the quality and diversity of the consumed information. In order to better understand the impact of this change, we set to study how diverse timelines are, by focusing on what users consume, through their followed accounts. Our goal was to answer the following questions: Do random users and mainstream media followers have access to the same information through different channels? Or do the mainstream media still play a fundamental role in information diffusion that cannot be replaced by regular Twitter users and "word-of-mouth"?

5.1 Measuring Diversity

We aimed at characterizing and understanding the differences between the content consumed by random users and the content consumed by users with a particular interest in mainstream media. Our approach consisted of creating a user profile based on the tweets received in a user's timeline. Each tweet was preprocessed by removing emoji, links, mentions, 'RT' and punctuation, and by normalizing spacing, through the conversion of multiple spaces, tabs and new lines to a single space and the trimming of the text. We then created a document per user, containing a concatenation of all preprocessed tweets that appeared in the user's home timeline. Each document was converted

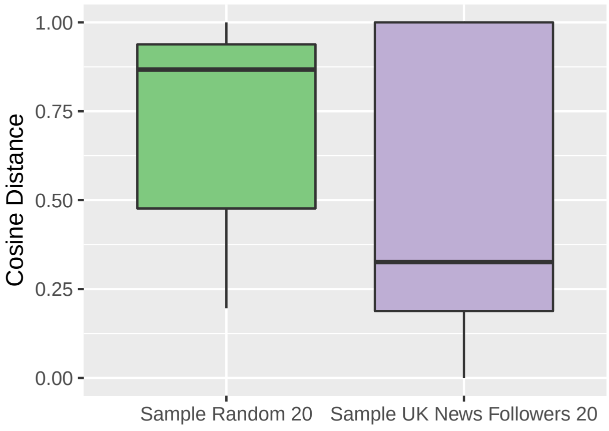

Figure 1: Cosine distances per sample, for all pairs of timelines.

to lower case and tokenized into unigrams, removing stopwords from several languages[1] and obtaining a document-term matrix, with the absolute term frequencies, per sample. Sparse terms were then pruned, ensuring a maximum sparsity of 0.996. This means that rare terms with more than 99.6% zeros, that were less useful in distinguishing user profiles, were simply discarded.

The resulting document-term matrix for “Sample Random 20” contained 19 documents and 228,165 terms — meaning that one of the users received no tweets during for the time span of the collection — and the document-term matrix for “Sample UK News Followers 20” contained 20 documents and 389,831 terms. In order to understand how diverse each timeline was, within either sample, we computed the cosine distance from each timeline to all others in the same sample. Timelines that are highly diverse will consistently have a high distance to most of the other timelines. Similarly, a sample will contain highly diverse timelines if the overall distances between all timelines are high, that is, timelines within a given sample considerably diverge in consumed content.

Figure 1 shows the box plot of the cosine distances between all pairs of timelines for each sample. As we can see, in particular through the median, “Sample Random 20” contains timelines that are more divergent among themselves (median cosine similarity is 0.87), while “Sample UK News Followers 20” contains timelines that are much less divergent among themselves (median cosine similarity is 0.33). We can say that mainstream media followers have less diverse information consumption habits when compared to a random sample of users.

5.1.1 Diversity over Time

We used a similar approach to study diversity over time, but instead of using a single user profile per timeline, we created a document per day for each user. This meant slicing the two original samples into 14 smaller parts, each part corresponding to one day, and repeating the study for each day.

Figure 2 depicts the dispersion of cosine distances between all pairs of timelines, per sample, over time. The daily behavior is consistent with the aggregated overall behavior, despite resulting in a slightly higher median cosine distance overall. This means that information consumption habits from random users are more diverse than mainstream media followers, but also that information consumption diversity for random users is lower per day than overall for the 14 days and, on the other hand, for mainstream media followers, it is higher per day than overall. This is quite expected, as the number of topics discussed in a single day are intuitively less than those discussed through the course of two weeks.

6 Conclusions

We have provided a consistent methodology to study the anatomy of a sample of Twitter timelines, focusing on content production and consumption, as well as on measuring overall and daily diversity. We studied the home timelines of two user samples: “Sample Random 20”, a random selection of users based on their numeric ID, and “Sample UK News Followers 20”, a selection of users that followed at least 3 out of 11 mainstream UK newspaper accounts.

We found that mainstream media followers consume a larger volume of information than random users. We analyzed the overall and the daily diversity over the course of two weeks, based on the cosine distances between all pairs of timelines, per sample. Both the overall and the daily diversity were consistently lower for the timelines of mainstream media followers, when compared to the timelines of random users. Interestingly, when analyzing the change from the overall two week aggregations to the daily aggregations, the samples diversities slightly approximate, but still result in a lower diversity within mainstream media followers and a higher diversity within random users.

Overall, we can say that, when compared to random users, mainstream media followers consume a narrower range of content, covering a smaller number of topics, with a higher production volume. This can be explained by the fact that users in this sample share a common interest (i.e. UK news), as opposed to the users in the random sample that have no common characteristic. As expected, mainstream media followers consume a less diverse variety of content. This

[1] We considered English, French, Spanish, Portuguese, Arabic, Russian, Greek and Hindi, but also typical expressions used in Twitter, like ‘via’ or ‘vs’.

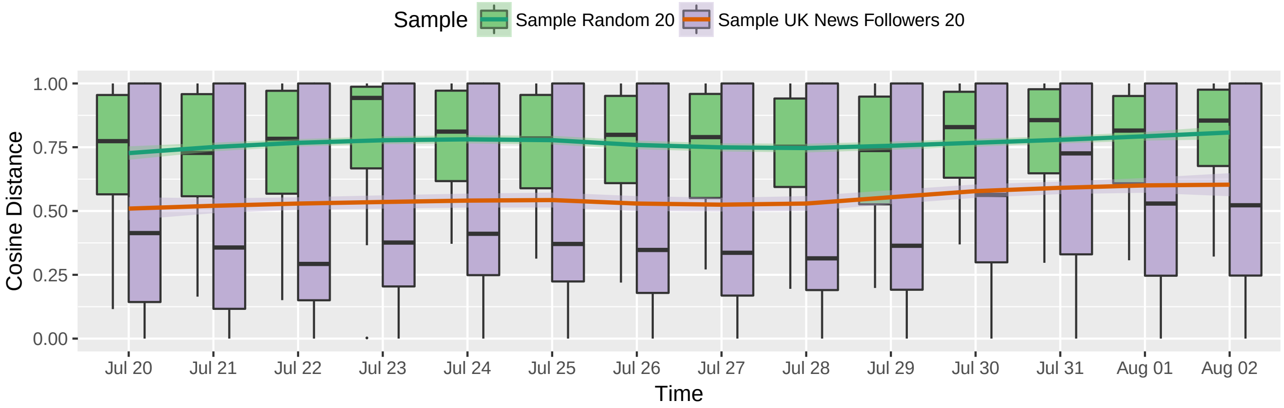

Figure 2: Cosine distances per sample, for all pairs of timelines, per day. The lines correspond to a locally weighted scatterplot smoothing (or LOESS, from LOcal regrESSion); they depict overall diversity per sample.

diversity is higher when we look at individual days, probably representing the coverage of multiple topics throughout a day, but lower when we look at the two week period, probably representing the convergence of topics.

In the future, we would like to analyze a larger sample of timelines, and also explore the diversity within topic-based communities, such as those focused on a given hashtag or those that share a geographical context.

7 Acknowledgments

José Devezas is supported by research grant PD/BD/128160/2016, provided by the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT). This work is partially funded by FourEyes, a Research Line within project “TEC4Growth – Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01-0145-FEDER-000020”, financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[BNS13] Kevin Bache, David Newman, and Padhraic Smyth. Text-based measures of document diversity. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’13*, page 23, 2013.

[CHBG10] Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and Krishna P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*, 2010.

[CSR14] Tiago Cunha, Carlos Soares, and Eduarda Mendes Rodrigues. Tweeprofiles: detection of spatio-temporal patterns on twitter. In *International Conference on Advanced Data Mining and Applications*, pages 123–136. Springer International Publishing, 2014.

[Jos06] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[KWM11] Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, pages 538–541, Barcelona, Catalonia, Spain, July 2011. AAAI Press.

[LKSM14] Yabing Liu, Chloe Kliman-Silver, and Alan Mislove. The tweets they are a-changin’: Evolution of Twitter users and behavior. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM 2014)*, Ann Arbor, MI, June 2014.

[MJA$^+$11] Alan Mislove, Sune Lehmann Jørgensen, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. Understanding the demographics of twitter users. In *Proceedings of the Fifth International AAAI*

Conference on Weblogs and Social Media (ICWSM 2011), pages 554–557. AAAI Press, 2011.

[POL10] Saša Petrović, Miles Osborne, and Victor Lavrenko. The Edinburgh Twitter Corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26, 2010.

[WJ04] Ryen W White and Joemon M Jose. A study of topic similarity measures. In *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, page 520, 2004.

[ZJW+11] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing Twitter and Traditional Media using Topic Models. In *Advances in Information Retrieval*, pages 338–349. Springer Berlin Heidelberg, 2011.