



Beyond Words: Comparing Structure, Emoji Use, and Consistency Across Social Media Posts

Melanie Swartz¹ , Andrew Crooks² , and Arie Croitoru¹ 

¹ George Mason University, Fairfax, VA, USA
{mswartz2, acroitor}@gmu.edu

² University at Buffalo, Buffalo, NY, USA
atcrooks@buffalo.edu

Abstract. Social media content analysis often focuses on just the words used in documents or by users and often overlooks the structural components of document composition and linguistic style. We propose that document structure and emoji use are also important to consider as they are impacted by individual communication style preferences and social norms associated with user role and intent, topic domain, and dissemination platform. In this paper we introduce and demonstrate a novel methodology to conduct structural content analysis and measure user consistency of document structures and emoji use. Document structure is represented as the order of content types and number of features per document and emoji use is characterized by the attributes, position, order, and repetition of emojis within a document. With these structures we identified user signatures of behavior, clustered users based on consistency of structures utilized, and identified users with similar document structures and emoji use such as those associated with bots, news organizations, and other user types. This research compliments existing text mining and behavior modeling approaches by offering a language agnostic methodology with lower dimensionality than topic modeling, and focuses on three features often overlooked: document structure, emoji use, and consistency of behavior.

Keywords: Data mining · Social media · Emojis · User behavior modeling

1 Introduction

As social media users engage with online conversations and form virtual communities, social media analysis often focuses on the topics discussed [15], user activity patterns [9], and networks arising from interactions of users [10, 16]. Often overlooked is the communication style associated with a user's social media posts. For instance, analysis of the specific words used can provide a fingerprint of the individual posting the content [8] and reveal shared linguistic styles of the online community [5]. Users also adapt their language to address limitations and norms associated with technology [14] (e.g., character limits, availability of emojis). Within this paper, we propose to move beyond just words, as the structural components of a document's composition and the way in which emojis are used within a document, such as a tweet, also provide cues about the individual and social norms for online communication styles and preferences.

In this paper we introduce and demonstrate a language-agnostic methodology to characterize structures of content and emoji use within a document, measure consistency of structures across a set of documents, and cluster documents and users with similar patterns and behavior. By comparing these patterns and behaviors across users and user roles such as journalists, bots, and others, we can generate baselines and gain insights into the unique or shared structures of communication styles and emoji use.

Three main contributions of this research are: 1) a novel methodology for structural content analysis; 2) analysis of the structure of emoji use as the attributes, position, order, and repetition of emojis within a document; and 3) user behavior modeling with regards to consistency of structure of document and emoji use. Benefits of our approach include it is language-agnostic, requires less dimensions than traditional topic modeling, and yields additional measures that can be combined with other text and user metrics. Further, this paper addresses a gap of current social media analysis by focusing on the structural components of communication style, enables comparison of emoji use, and models consistency of user behavior based on social media content. In what follows, Sect. 2 provides an overview of current approaches to content analysis and analysis of emoji use, followed by our methodology in Sect. 3. We then present and discuss our results (Sect. 4), and conclude with areas for further work (Sect. 5).

2 Background

2.1 Content Analysis of Social Media

Content analysis of social media has mainly focused on the words contained in posts to identify discussion topics or associate groups of users based on their use of specific terms, hashtags, or group of words identified via topic modeling [15, 16]. Recently, content analysis combined with other metrics for user activity and network connections, has been applied in order to identify or categorize bots [10, 15, 16]. Analysis of the structure of social media is fairly nascent. [9] considered number of words in text in addition to user activity metrics to identify user intent in spreading misinformation. In addition to content length, [17] took into account content type such as presence of urls, and hashtags to describe activity associated with troll accounts. [1] examined the order of lexical properties of a tweet (such as place name, event date and time, event description) in order to improve the effectiveness of messaging during earthquakes. While [7] focused on the order of content within a tweet and impact on communication styles.

2.2 Analysis of Emoji Use

There is still much more to be learned about the way visual content, including emojis, are used in social media [4]. However, most social media research pertaining to emojis has focused on the meaning of emojis or emojis as indicators of sentiment or sarcasm (e.g., [3]). Only recently has the emphasis shifted to the behavior and structure of emoji use. [12] revealed differences in the way emojis are used based on document types such as tweets, user names, and profile descriptions. [11] identified how emojis are used as structural markers based on where they are placed in text.

3 Measuring Document Structure, Emoji Use, and Consistency

In this section, we first describe how to represent structures of a document (3.1). Then we characterize the structure of emoji use as the attributes of emojis used (3.2) plus the position, order, and repetition of emojis within a document (3.3). Next, we explain how to measure consistency of structures across a set of documents (3.4). Finally, we describe how to cluster users based on structures (3.5) and consistency scores (3.6).

3.1 Document Structure, Content Structure, and Emoji Spans

In order to define the structures of a document, first identify the types of content associated with documents in the collection. For the purpose of this paper we use data from Twitter and view a single tweet as a document. For tweets, we identify content types: retweet indicator, text, emoji, punctuation, hashtag, mention, and url. Each document is divided into spans by content type, irrespective of spaces, and assigned a sequential number as span number. Span length is the number of features per span. Document structure is represented as a list with the content type and number of features for each span, in order of occurrence. Similarly, content structure is a list of span content types in order. Representing a document and content structure in this way enables comparison of documents based on the type or order of contents and enables grouping of documents with similar structural format and style.

For documents containing emojis, we identify which emojis are used as a sorted list of unique emojis. We use the emoji spans (i.e., the spans with content type of emoji) and document structure to describe the way that emojis are used in a document, which we refer to as the structure of emoji use. Figure 1 shows a sample tweet represented as document structure, content structure, emoji spans, and unique emojis. In the next two sub-sections, we demonstrate the structure of emoji use as the emoji attributes paired with the analysis of the position, order, and repetition of emojis within a document.



















Document contents	  @Nationals win  World Series  #Champs #Nats https://www.mlb.com							
Content type	Emoji	Mention	Text	Emoji	Text	Emoji	Hashtag	Url
Span number	1	2	3	4	5	6	7	8
Span length	3	1	1	1	2	2	2	1
Document structure	[(emoji,3), (mention,1), (text,1), (emoji,1), (text,2), (emoji,2), (hashtag,2), (url,1)]							
Content structure	[emoji, mention, text, emoji, text, emoji, hashtag, url]							
Emoji spans	[[ ,  , ], [], [ , ]]							
Unique emojis	[ ,  ,  ,  ,  ,  ,  , ]							

Fig. 1. Structures of a sample tweet.

3.2 Attributes of Emojis in a Document

For each emoji in the emoji spans and in the unique emojis list we describe the emoji along eight attributes noted below. These eight attributes were chosen because each can be used alone or in combination to enable comparison of emojis. Additional attributes

could be added such as sentiment or meaning. The first three attributes are from Unicode [2] and were chosen based on previous research showing the value of using emoji group and sub-group for comparison of emoji use [12]. The other attributes are based on heuristics used to sort emojis.

1. **Unicode Group:** Unicode assigns each emoji to one broad category (e.g., Smileys & Emotion, Animals & Nature, Food & Drink, Travel & Places, Objects, Symbols, Flags, and People & Body which also includes Activity).
2. **Unicode Sub-Group:** Unicode assigns emoji to sub-category (e.g., face-smiling).
3. **Unicode Name:** The Unicode emoji name (e.g., “face with tears of joy”).
4. **Type:** A label assigned by mapping sub-group to another descriptive property based on a research topic. For this paper, we use shape, anthropomorphic, and other.
5. **Anthro-type:** For anthropomorphic (human like) emojis, we map sub-groups to: face, face-gesture, hand-gesture, body-gesture, body-part, single person, multiple.
6. **Shape:** Indicated by emoji name: triangle, circle, square, star, heart.
7. **Color:** Indicated by emoji name: red, blue, yellow, pink, purple, orange, green, brown, white, black. We also include the five Fitzpatrick skin-tone colors used for emojis: light, light-medium, medium, medium-dark, and dark. We use the name because color appearance may vary across platforms.
8. **Direction:** Based on words in emoji name to indicate: up, down, left, or right.





To demonstrate how the above set of attributes enables descriptive comparison of similarities and differences of individual emojis, consider these two emojis,  and . There are differences in appearance and type with one a red triangle (type of shape) and the other a hand holding up an index finger (anthropomorphic hand-gesture). Yet they are similar in direction of pointing “up”. Table 1 summarizes attributes for these emojis.

Table 1. Emoji attributes.

Emoji	Group	Sub-group	Name	Type	Anthro-type	Shape	Color	Direction
	Symbol	Geometric	Up-pointing red triangle	Shape	None	Triangle	Red	Up
	People & body	Hand-single-finger	Index pointing up: medium skin tone	Anthro	Hand-gesture	None	Medium	Up

3.3 Emoji Position, Order, and Repetition

Position of Emojis in a Document. We describe the general position of emojis in a document based on relative position of emoji as: first, beginning, middle, end, or last. We use document structure to derive relative position based on span number for the emoji content in relation to total number of spans in the document, divided into thirds, (e.g., the first third of spans is the beginning). Content in the very first and last spans are labeled as such. Documents with less than five spans are only first, middle, or last.

Emoji Order. The order of emojis and attributes are noted both within and across emoji spans. We take into account emoji order, as emoji color order within the same emoji span could result in a set of emojis taking on different meanings, based on context of text or user. For example, the set of heart emojis ❤️🤍💙 with color order red, white, blue could represent colors of a sports team (as in Fig. 1) or a country flag (e.g., Netherlands or United States). The order of emojis or attributes can also indicate a pattern. For example, “💙🌟text🌟💙”, represents a pattern we call emoji reversal which occurs when two consecutive emoji spans contain the same emojis or attributes, but the order in the second span is reversed.

Emoji Repetition. We categorize repetition of emojis or emoji attributes as three types: redundant, emphasis, and amplification. Redundant is the repetition of the same emoji or attribute within the same span, sometimes representing magnitude or quantity, (e.g., 😂😂😂). Emphasis is often used to draw attention and generally occurs when two emoji spans contain the same emojis regardless of order (e.g., “🚨 Alert 🚨”, “🌊 Blue-wave 🌊”). Amplification is repetition of an attribute across different emojis within the same span or across multiple spans (e.g., color red in: “🔴 Vote 🗳️ all Red 🔴”).

3.4 Measure of Consistency

With the structures of document contents and emoji use represented, we can then measure consistency of these structures across a set of documents, such as a user’s or group’s tweets. This measure makes it possible to highlight differences in behavior based on relative consistency in terms of document content, style, or emoji use.

To measure consistency, for a set of documents, iterate across the unique structures (U) (e.g., document structure, content structure, or emoji use). For each unique structure, divide the number of documents in the set with that structure (d_i) by the number of documents for the structure in the set with the most documents ($\max(Ud)$), then square the results. Calculate the measure of consistency for the set of documents, (C), as 1 divided by the sum of theses squares of normalized proportions of documents per unique structure. The resulting measure of consistency ranges between 0 and 1 with larger values indicating greater consistency and smaller values approaching 0 representing greater variation. The measure of consistency is represented by equation:

$$C = 1 / \sum_{i=1}^U \left(\frac{d_i}{\max(Ud)} \right)^2 \quad (1)$$

We chose this approach compared to other measures (e.g., Shannon or Simpson’s Index) to enable standardized comparison regardless of collection size and to support a variety of distributions for document counts per unique structures. In addition, we add weight for unique structures that comprise a greater proportion of a user’s documents.

3.5 Clustering by Structure, Content, and Emoji Use

Even though the text of individual documents varies greatly, users and documents can be clustered based on similarity of document structure, content style, or emoji use. We

also identify common structures used across of users, as well as identifying the users of specific structures via aggregation. These approaches support analysis of communication patterns to identify common or unique structures used, structures associated with specific types of users or groups, as well as identifying documents or users that may be related based on similar style defined by the structures used.

3.6 Clustering Users Based on Consistency

In addition, we cluster users based on their consistency scores for structures of document, content, and emoji use. We use the unsupervised clustering algorithm HDBSCAN [6] as it does not require defining the number of clusters, supports multiple dimension data, finds stable clusters within noisy data, and can handle clusters of varying density, size, and shape. For each cluster, we describe behavior traits as low, medium, or high consistency for each factor based on the greatest percent of users of that cluster falling within interquartile ranges for low (first quartile), medium (second and third quartiles), or high (fourth quartile). The composition of users in a cluster is then summarized based on additional information such as keywords in user profile descriptions or labeled data such as if account has previously been labeled as bot-like. Clustering users based on consistency enables comparison and grouping of users with similar behavior patterns associated with their communication style.

4 Experiment Results and Discussion

4.1 Experimental Setup

We apply our methodology to a corpus of 44 million tweets collected in October and November 2018 related to the 2018 U.S. midterm elections based on keywords, hashtags, and user accounts associated with candidates or political parties. For each of the 3.3 million unique users set of retweets and non-retweets we measure consistency of document structure and content structure. To improve consistency scores and reduce dimensionality, we modified the document structure by removing spans of punctuation and by not including the count of features for text spans. For the 30% of users of emojis in their tweets, we measure consistency of which unique emojis are used and also measure structure of emoji use represented as a vector of attributes, position, order, and repetition of emojis (eVAPOR). Using HDBSCAN we cluster users based on consistency scores for their retweets and non-retweets separately. We then describe the composition of each cluster and measure the percent of accounts labeled as bot-like based on Botometer scores. Next we present the results of our analysis.

4.2 Distributions of Consistency

We compared the distribution of consistency scores for users that sent more than two non-retweets or more than two retweets. Figure 2 shows the range of these scores for tweet text, document structure, content structure, unique emojis, and structure of emoji use. As expected, we found little user consistency in tweet text. In general, users were

more consistent in their non-retweets than retweets, especially for content structure and which specific emojis were used. This indicates users tend to use the same order, format, and often the same emojis for their own tweets, whether knowingly or not. Users had less consistency with retweets likely a result of retweeting multiple users. Analysis of user behavior for document structure, content structure, and emojis in tweets reduces dimensionality and yields new information compared to traditional text analysis.

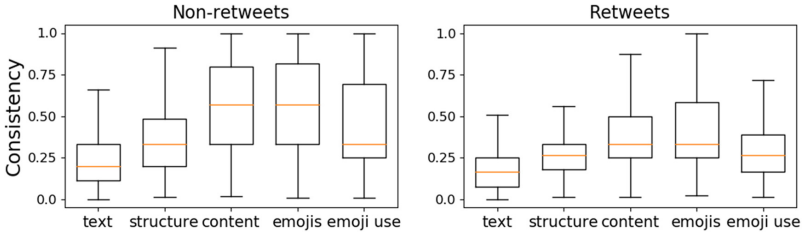


Fig. 2. Distribution of consistency scores for users sending non-retweets (left) and retweets (right) shown with interquartile ranges.

4.3 Analysis of Structures for Document, Content, and Emoji Use

Using the methodology presented in Sect. 3, we identified common structures of non-retweets and retweets used by a large percent of tweets or users. Table 2 shows the most common non-retweet content structures with emojis. Analysis of these structures used by bot accounts led to identification of additional accounts likely to be bots not yet labeled. Tweets of these accounts exhibited identical content structure and structure of emoji use, although the tweets had different text, urls, emojis, and document structures (e.g. same content type order with variation in number of urls, emojis, and mentions). Given the similarity of the user profile descriptions and names for these users, it would not be surprising if these accounts are related. This is just one of many examples we found demonstrating structural content analysis can identify specific styles of communication that may be a signature for an individual or group of users.

Table 2. Most common content structures with emojis for non-retweets.

Content Structure	Percent of tweets	Percent of users
[attention, text, emoji]	18%	31%
[text, emoji]	6%	15%
[attention, text, emoji, text]	4%	9%
[text, emoji, url]	2%	5%
[attention, text, emoji, hashtag]	2%	4%
[attention, emoji]	1%	4%

4.4 Clustering Users Based on Consistency

We compared consistency scores for user non-retweets and retweets across four dimensions: document structure, content structure, unique emojis used, and emoji use. Clustering users based on consistency scores in two dimensions reveals groups of users in the dataset with similar behaviors for document structure and emoji use, content structure and unique emojis used, Fig. 3. With the t-SNE algorithm we visualize the clusters of users with similar behavior across four dimensions, Fig. 4.

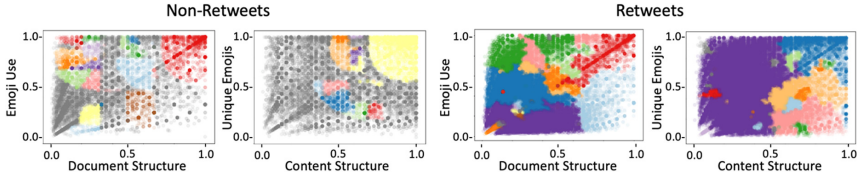


Fig. 3. Clusters of users with similar behavior across two factors in non-retweets (left) and retweets (right) Colors indicate cluster assignments.

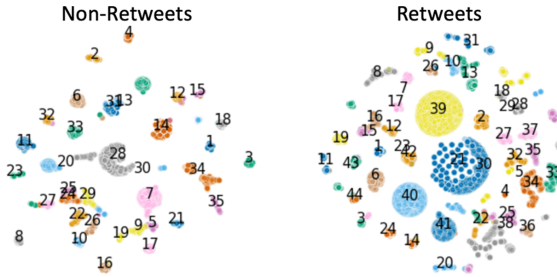


Fig. 4. Clusters of users with similar behavior across four factors for non-retweets (left) and retweets (right). Each cluster is labeled with a number.

4.5 Behavior Traits and Composition of Users in Clusters

For each cluster of users, we describe the behavior traits in terms of low, medium, or high consistency across each of the four dimensions. While most clusters had medium consistency for at least one dimension, 6.5% of non-retweet and 4.8% of retweet users were grouped into clusters that had high consistency across all four dimensions. We then calculated the percent of user accounts for each cluster that were likely bots based on Botometer scores [13]. One of the non-retweet clusters had 45% bots, compared to the average 19% for other clusters. While not all bot-like users had high consistency scores, this particular cluster did for each of the four factors. This could indicate that additional users within this cluster may be related or also bots but not yet identified by existing bot detection algorithms.

Next, we analyzed the composition of user roles per cluster and by behavior. We define users by role based on keywords in their user profile (e.g., journalists and news

organizations, marketers, businesses, celebrities, government, activists, veterans, students). Most users were clustered into groups with medium consistency scores across the four factors for non-retweets. However, the user group with verified user accounts and indicating the user is a journalist, reporter, or news organizations (which we label as ‘News’) had the greatest percent of users with high consistency across all four factors in non-retweets. Many of their tweets appear to be auto-generated using a template as tweets exhibited same structure but changing information such as top news and weather reports throughout the day. Similarly, ‘Marketer’ also had high percent of users with high consistency and tweets with similar structure and emoji use were indicating weekly or daily sales or specials. While most users had relatively low consistency in retweets, the user group of retired military veterans had the greatest percent of users with high consistency for which emojis were used and the way emojis were used in retweets. This could indicate that tweets with specific emojis and style of emoji use are more likely to be retweeted by this group. Table 3 summarizes the top user roles based on percent of users for categories indicating consistency of behavior associated with document structure and structure of emoji use in non-retweets.

Table 3. Top user roles and percent per consistency category for document structure and structure of emoji use in non-retweets.

		Consistency of structure of emoji use		
		Low	Medium	High
Consistency of document structure	Low	9% Celebrity 8% Activist	11% Bot 10% Activist	5% Coach 4% Government
	Medium	9% Bot 7% Coach	66% Marketer 45% Coach	12% Government 11% Veteran
	High	2% Business 2% Artist	12% Student 10% Celebrity	20% News 16% Marketer

While it is not easy to verify authenticity of a user account or role, we demonstrate how to identify unique and common patterns and traits among a group of users with the same attributes in their user profile description. Overall, our results reveal how new insight can be gained by identifying and analyzing communication style patterns of individuals and groups of users with similar roles or behaviors for consistency across structures or emoji use in their documents.

5 Conclusions and Future Work

This paper introduces and demonstrates a new language-agnostic approach for structural content analysis and user behavior modeling by characterizing the structure and emoji use of a document, and then measuring and clustering by user consistency. With this methodology we described signatures of communication styles and behaviors for individuals, user groups, and clusters of users. We also identified users with document

structural properties and user consistency metrics similar to accounts already labeled as bot-like. Limitations of our study are that we focused on only one collection of tweets related to American politics and it is difficult to verify authenticity of user accounts. Areas for further research could compare tweet styles and author consistency for other topics and user roles such as sports, tourism, and health or message effectiveness. Structural content analysis and measuring consistency across documents, as presented in this paper, compliments existing text mining techniques and provides a new perspective for social media analysis by linking document style and user behavior.

References

1. Comunello, F., Mulargia, S., Polidoro, P., Casarotti, E., Lauciani, V.: No misunderstandings during earthquakes: elaborating and testing a standardized tweet structure for automatic earthquake detection information. In: ISCRAM (2015)
2. Davis, M.: Emoji Charts, v13.0. <https://unicode.org/emoji/charts/full-emoji-list.html> (2020)
3. Felbo, B., Mislove, A., Søgaaard, A., Rahwan, I., Lehmann, S.: Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In: EMNLP, pp. 1615–1625 (2017)
4. Highfield, T., Leaver, T.: Instagrammatics and digital methods: studying visual social media, from selfies and GIFs to memes and emoji. *Commun. Res. Pract.* **2**(1), 47–62 (2016)
5. Khalid, O., Srinivasan, P.: Style matters! Investigating linguistic style in online communities. In: ICWSM, pp. 360–369 (2020)
6. McInnes, L., Healy, J., Astels, S.: HDBSCAN: hierarchical density-based clustering. *J. Open Source Softw.* **2**(11), 205 (2017)
7. Pederson, J.A.: It's not what you tweet but how you tweet it: an experiment of orientation, interactivity, and valence in Twitter. Dissertation, Texas A&M Univ. (2016)
8. Pennebaker, J., Mehl, M., Niederhoffer, K.: Psychological aspects of natural language use: our words, our selves. *Annu. Rev. Psychol.* **54**(1), 547–577 (2003)
9. Rajabi, Z., Shehu, A., Purohit, H.: User behavior modelling for fake information mitigation on social web. In: Thomson, R., Bisgin, H., Dancy, C., Hyder, A. (eds.) SBP-BRiMS 2019. LNCS, vol. 11549, pp. 234–244. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21741-9_24
10. Schuchard, R., Crooks, A.T., Stefanidis, A., Croitoru, A.: Bot stamina: examining the influence and staying power of bots in online social networks. *Appl. Netw. Sci.* **4**(1), 55 (2019)
11. Spina, S.: Role of emoticons as structural markers in Twitter interactions. *Discourse Process.* **56**(4), 345–362 (2019)
12. Swartz, M., Crooks, A.: Comparison of emoji use in names, profiles, and tweets. In: ICSC, pp. 375–380 (2020)
13. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: detection, estimation, and characterization. In: ICWSM, pp. 280–289 (2017)
14. Walther, J.: Interaction through technological lenses: computer-mediated communication and language. *J. Lang. Soc. Psychol.* **31**(4), 397–414 (2012)
15. Wirth, K., Menchen-Trevino, E., Moore, R.T.: Bots by topic: exploring differences in bot activity by conversation topic. In: *Social Media and Society*, pp. 77–82 (2019)

16. Yuan, X., Schuchard, R., Crooks, A.: Examining emergent communities and social bots within the polarized online vaccination debate in Twitter. *Soc. Media Soc.*, **5**(3) (2019)
17. Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., Blackburn, J.: Disinformation warfare: understanding state-sponsored trolls on Twitter and their influence on the Web. In: WWW 2019, pp. 218–226 (2019)