Selfies of Twitter Data Stream through the Lens of Information Theory: A Comparative Case Study of Tweet-trails with Healthcare Hashtags

Yuan Zhang
University of North Texas
Yuanzhang2@my.unt.edu

Hsia-Ching Chang
University of North Texas
Hsia-Ching.Chang@unt.edu

Abstract

Little research in information system has been carried out on the subject of user's choice of different components when composing a tweet through the analytical lens of information theory. This study employs a comparative case study approach to examine the use of hashtags of medical-terminology versus lay-language in tweet-trails and (1) introduces a novel $H_{(x)}$ index to reveal the complexity in the statistical structure and the variety in the composition of a tweet-trail, (2) applies radar graph and scatter plot as intuitive data visualization aids, and (3) proposes a methodological framework for structural analysis of Twitter data stream as a supplemental tool for profile analysis of Twitter users and content analysis of tweets. This systematic framework is capable of unveiling patterns in the structure of tweet-trails and providing quick and preliminary snap shots (selfies) of Twitter data stream because it's an automatic and objective approach which requires no human intervention.

1. Introduction

Composing a tweet on the Twitter platform involves a choice of combining typical components, such as photos, video clips, and up to 140-character textual content which may include hashtags, hyperlinks, and the @username "mention" function. An orchestrated presentation of tweet content usually improves the usability, effectiveness, and perceived quality of a campaign message. Health communication studies suggest that a well-crafted balance of words, numbers, images and other illustrations can improve comprehension more than using text alone [13]. However, there is no widely-agreed rule of thumb regarding how diversified the content should be when combining text with other multi-media components (*i.e., image and video*).

As people increasingly seek health information online, healthcare campaigns on social media platforms are gaining more attention. Twitter, one of the most popular social media platforms, attracts and connects users (*people or business accounts who construct or/and read tweets*) across the world through their information seeking and sharing behaviors. On the other hand, along with opportunities, Twitter brings challenges to healthcare campaigners when it comes to making an effective and efficient message. Communicating healthcare messages on Twitter is not as easy as it seems to be because 140 characters are sometimes insufficient to make a point on healthcare related topics. The solution for such issue often involves two options: (1) use a hyperlink to direct the audience to target webpages where more space is available for campaigners to operate, or (2) use image(s) and/or video to enhance the tweet content. Either approach increases the complexity in the structure (*the way different components are organized*) of the Twitter messages (*i.e., tweets*). Therefore, the more components a tweet contains, the more complex its structure appears. According to information theory, messages have meanings and "these semantic aspects of communication are irrelevant to the engineering problem" (*the structure of message*) [20]. The concept of entropy, inherited from thermodynamic to information theory by Claude Shannon, provides researchers with a means to examine the variety of the combinations of typical content components that eventually compose a tweet.

This study is an attempt to introduce a measure of the structural complexity in data stream on social media. In particular, this study focuses on understanding healthcare communication on Twitter by contrasting the structure of messages in a sample of tweets associated with healthcare-related hashtags through the lens of information theory.

2. Related works

2.1. Entropy and information theory

Information theory was developed by Claude Shannon during World War II in his work of modeling the electronic signal transmission [20]. The idea of measuring information storage capacity in logarithmic terms dated back to the 1920s [10]. Information theory was originally used in studies of telecommunication systems and applications in data compression, and then Warren Weaver extended it to analyzing human communication [19].

In Claude Shannon's information theory [20], entropy was defined as the amount of information which was calculated by the logarithm of (1) the effective number of microstates of a closed system, or (2) the effective number of possible values of a random variable. For a sequence of symbols, the set of probabilities could be represented as P_1, … P_n, and the entropy of this sequence was calculated by the equation below where H refers to the measure of information and uncertainty [20], or average surprise [2].

URI: http://hdl.handle.net/10125/50294
ISBN: 978-0-9981331-1-9

$$H = -K \sum_{i=1}^{n} p_i log\, p_i$$

Primarily adopted in engineering and computer science, Shannon's entropic equation has been used to evaluate the level of predictability [17], redundancy [12], and degree of randomness/complexity [15] in a well-defined system. Besides its applications in the natural sciences, information theory has also been applied to linguistic studies. In 1992, Brown et al examined the upper bound for the entropy of the English language [6]. In 2004, Borgwaldt, Hellwig, and de Groot estimated the word-initial entropy per phoneme in English [5]. In 2009, Chong, Sankar, and Poor examined the entropy of American Sign Language [8]. Another similar study focused on phonotactics and phonotactic learning was conducted by Hayes and Wilson [11].

With the advent and prevalence of social media, research interests have shifted to linguistic studies on the Twitter platform using information theory. In 2013, Neubig and Duh examined "information content" per character in a tweet with a quantitative approach and found that although Chinese and Japanese language has more information per character, a Chinese/Japanese tweet doesn't necessary contain more information than the ones in other languages [18]. The application of information theory on Twitter also reaches another aspect of tweeting activities. Ghosh, Surachawala, and Lerman introduced an entropy-based activity classification method to characterizing the dynamics of retweeting activities in 2011 and suggested its applications in automatic spam-detection and trend identification [9].

Information system was the third major academic discipline (*after natural sciences and communication science*) that chose information theory as a general model of information exchange [4] and applied it to research topics such as database and business analytics, etc.

2.2. Twitter research in healthcare

The first study to examine what researchers had studied about Twitter found that the majority of studies was the content analysis of tweets across different domains, followed by the studies of Twitter users and the platform itself [22]. Using full-text content analysis of 382 academic articles published from 2007 to 2012, Zimmer and Proferes also concluded that tweet content was the most popular source of data collection and analysis; approximately 60% of studies employed content analysis to analyze tweets in various research areas. Computer science, information science, and communications were the top three disciplines contributing to Twitter research [23].

Healthcare professionals face challenges when communicating campaign messages to the general public on Twitter because Twitter is a real-time information sharing system and tweets usually have a short life-cycle. A hashtag, prefixed with a # symbol, is used to index keywords or topics on Twitter. Considered as an innovation, the hashtag convention was suggested by a Twitter user and initiated on Twitter to allow users to easily sift through and diffuse information that attracts their interest [7]. In 2017, Beguerisse-Díaz et al captured and analyzed 2.5 million tweets with hashtag #diabetes, from late March 2013 to late January 2014 and identified four themes that emerged from the tweets as health information, news, social interaction, and commercial messages [3].

As the hashtag convention has become popular on Twitter, it has provided more opportunities for and great convenience of information seeking and sharing. However, it is challenging for healthcare professionals to make the best use of the limited 140-character space and deliver an effective message. The reason is that health-related topics involve communicating sophisticated and sometimes confusing messages. Applying one or multiple hashtags in a tweet certainly extends its potential lifecycle by increasing the chances of being found and getting retweeted. However, the opportunity cost (*the loss of potential gain from other alternatives when one choice is made*) associated with this manner deserves further consideration because hashtags inevitably consume part of the 140-charater space.

Numerous healthcare hashtags have been used and shared on Twitter. In this study, these hashtags were defined and classified into two main categories: (1) medical-terminology hashtags whose prefixes and suffixes come from Latin and Ancient Greek, and (2) lay-language hashtags for medical/healthcare terms. For example, #glucose and #hypertension are categorized as medical-terminology hashtags, while #bloodsugar and #bloodpressure are categorized as lay-language hashtags.

Sometimes medical-terminology hashtags and lay-language hashtags have similar but not exactly the same meaning; other times these hashtags share the same semantic meaning. For example, glucose, a word in medical-terminology, is derived from the Latin word glucosium and its meaning is monosaccharide. In lay-language, glucose is called blood sugar. Although blood sugar does not refer to real cane sugar in human blood, it shares the same semantic meaning with glucose.

The difference in the usage of medical hashtags and lay-language hashtags is an important topic on the Twitter platform because it is wasteful to include them both given such limited space (*140 characters*). Healthcare professionals or agencies might be more likely to use #hypertension, however patients who are not familiar with medical-terminology and looking for tweets with #bloodpressure might not find these tweets.

3. Research method

Although the content of a healthcare message that carries the semantic meaning is highly constrained by the 140-character limit on Twitter, users can be creative about constructing their messages by combining typical components such as text, hashtags, hyperlink, image, video etc. Investigating the variety in such combinations, for

example, the ingredients of different components and the structure of a tweet-trail (*collection of tweets that typically share a common hashtag and sorted by the timestamp of each tweet*), provides insights in the tweeting activities in the context of healthcare communication, especially when these hashtags have similar or the same semantic meanings.

To tackle this issue, this study applied information theory to examine two pairs of tweet-trails with healthcare hashtags, namely #glucose versus #bloodsugar and #hypertension versus #bloodpressure, with a comparison of their statistical structures in terms of the choice of components to compose a tweet. The concept of entropy in this study, derived from Shannon's information theory, measures and compares the level of complexity in the structure of different tweet-trails.

3.1. Components of a tweet-trail

The first step to understanding the complexity in tweet-trails in terms of the structure is to define the level of granularity. In this study, the granular levels of a tweet are categorized as below:

Letter < Word < Component < Tweet

The left end of the spectrum (*i.e., letter*) represents smaller granularity whereas the right end of the spectrum (*i.e., tweet*) demonstrates greater granularity. To the best of our knowledge, no study has used the lens of information theory to investigate the tweet composition of typical components (*text, hashtag, hyperlink, image, etc.*) from which users can choose and construct their tweets.

This study employs a comparative case study method to show how medical-terminology hashtags and corresponding lay-language hashtags can be used to help in communication of healthcare messages. Using entropy as a measure, this study analyzes two pairs of healthcare tweet-trails with a specific focus on six typical components used in the tweets. As mentioned previously, there are many distinguishable components available to construct a tweet and the way of combining these components is unlimited, only depending on the choice of the tweet creator.

In this study, the granular components for composing a tweet are categorized as (1) image(s), (2) text with semantic meaning, (3) hashtag(s), (4) @username(s), (5) hyperlink, and (6) unused space. These six components serve as the fundamental "elements" or alphabet [15] for coding and calculating entropy based on Shannon's information theory.

The calculation of entropy in this study is based on the following premises: (1) All these components are independent of each other. Although the choice among different components to compose a single tweet is restrained by 140-character limit, this restriction does not affect the independence of entities in each alphabet. For example, in a sample of 100 tweets, there are 75 tweets with hyperlink and all these 75 hyperlinks are independent of each other; there is no restriction on choice of hyperlinks within the alphabet. (2) Each alphabet has a finite number of variables (*microstates*). In this empirical study, each tweet-trail contained a finite number of tweets, and in a given trail there was a finite number of different entities from typical components. (3) All the entities in each alphabet are discrete variables. (4) The empirical frequency of an entity in each alphabet serves as the probability of a variable in Shannon's equation. (5) The logarithm of the probability distribution is additive for independent sources.

3.2. Data collection and preparation

Two pairs of medical/healthcare hashtags versus their corresponding lay-language counterparts were retrieved using the hashtag-search function supported by NodeXL Pro software, version 1.0.1.378. Those two pairs of hashtags were #glucose versus #bloodsugar and #hypertension versus #bloodpressure. Regarding data filtering and cleaning, in order to calculate entropy value in a consistent way, the inclusion criteria in this study were: (1) All the tweets must be written in English. (2) All the tweets must contain at least one of the investigated paired-up hashtags. (3) The tweets must be unique, meaning no duplicate tweets in each sample dataset. (4) The tweets that contain video or gif image was excluded from this study because the entropy value of a video clip or a gif image file demand much more complex calculating technique and, therefore, will be included in future studies. (5) The tweets that contain emoji and/or special characters was excluded. The reason for this exclusion is that these symbols and emoji are dependent on display devices (*they do not look the same across different cellphone operation systems*) and they cannot fit into any of the six components which this study defines. The procedure of data collection and data cleaning of the two cases are summarized in Table 1.

Table 1. Summary of data collection process and data preparation

	Case 1		Case 2	
	#glucose	#bloodsugar	#hypertension	#bloodpressure
Data Collection Date	02-23-2017		02-12-2017	
Total Tweets Collected	190	165	250	250
Time Frame for Comparison	02-13-2017 to 02-22-2017		02-11-2017 to 02-12-2017	
Number of Tweets in Each Trail	96	95	61	96
Percentage of Tweets with Image(s)	42%	47%	15%	63%
Tweet(s) Contain Both Compared Hashtags	2		1	

During the data collection process, a variation of #bloodsugar was found: #bloodsuger. For the purpose of comparison, the tweets containing #bloodsuger were eventually excluded from this study. This phenomenon implies that #bloodsugar was used by users who occasionally spell incorrectly. On the other hand, no variation of hashtag spelling was identified in the data collecting process for the #glucose trail, indicating that people who use medical-terminology hashtags are less likely to make spelling errors.

Unlike the conventional statistical technique which compares two samples with same size, this comparison was based on different sized tweet-trails in the same time

period. In case 2, the sample size of #hypertension trail is much smaller than that of #bloodpressure trail due to the fact that tweets with #hypertension were much fewer published than the ones with #bloodpressure during that data collection period.

4. Data analysis with entropy calculations

The traditional entropy calculation is a straightforward process. However, it only involves one coding scheme at a time and generates only one entropy value for the scheme. Inspired by the work of Kearns and O'Connor, this study draws on their approach of calculating "form complexity" in moving image documents [14]. Furthermore, this study not only examines the complexity in the "statistical structure" [20] in a tweet-trail but also extends Shannon's original entropy equation to a multi-dimensional matrix by integrating six different content components with their own coding schemes.

Table 2 illustrates an example of the coding scheme and the matrix for calculating the entropy value of each component in a given tweet-trail (along with the *vertical direction*) and the synthetic value of $H'_{(tweet\text{-}x)}$ for each tweet in that trail (along with the *horizontal direction*). The operational definitions of the variables and their notations in this study were as follow: $H_{(x)}$ was the general notation of the matrix for entropy calculation. $H_{(trail)}$ was the final calculative result of the $H_{(x)}$ matrix. $H'_{(hashtag)}$, short for $H'_{(\#)}$, was the entropic value of component *Hashtag*. $H'_{(hyperlink)}$, short for $H'_{(HL)}$, was the entropic value of component *Hyperlink*. $H'_{(@username)}$, short for $H'_{(@)}$, was the entropic value of component *@username*. $H'_{(space)}$ was the entropic value of component *Unused Space*. $H'_{(text)}$, short for $H'_{(txt)}$, was the entropic value of component *Text with Semantic Meaning*. $H'_{(red)}$, $H'_{(green)}$, and $H'_{(blue)}$ were respectively the calculative results of entropy value of component *Image's RGB color*. For each tweet in a trail, $H'_{(tweet)}$ was the sum of each entity's $P(x_i) \times \log_2 P(x_i)$ value in that tweet.

Table 2: Example of the coding scheme and $H_{(x)}$ matrix

unit: bits	Text-based Content of a Tweet					Image in a Tweet			$H_{(x)}$
	Hashtag	Hyperlink	@username	Unused Space	Semantic Text	Red	Green	Blue	
Tweet 1	*Yes*	*No*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	$H'_{(tweet\text{-}1)}$
Tweet 2	*Yes*	*Yes*	*No*	*No*	*Yes*	*No*	*No*	*No*	$H'_{(tweet\text{-}2)}$
Tweet 3	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	$H'_{(tweet\text{-}3)}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Tweet n	*Yes*	*No*	*No*	*Yes*	*Yes*	*No*	*No*	*No*	$H'_{(tweet\text{-}n)}$
$SUM_{(tweets)}$	$H'_{(\#)}$	$H'_{(HL)}$	$H'_{(@)}$	$H'_{(space)}$	$H'_{(txt)}$	$\frac{H'_{(red)}H'_{(green)}H'_{(blue)}}{H'_{(image)}}$			$H_{(trail)}$

The nomenclature in this study complied with the following rules: (1) The denotation of letter H as entropy was inherited from Claude Shannon's information theory [20]. $H_{(x)}$ and $H_{(trail)}$ were both derived from the original entropy concept regardless either in a thermodynamic-closed system or for a social media data stream. (2) Denotation of all the $H'_{(...)}$ means that these variables were not the same as Claude Shannon's original entropy concept. These $H'_{(...)}$ were actually sub categorical entropy-calculation results for the granular components in a tweet-trail. They were at lower levels of the hierarchy of a well-defined set of interrelated coding schemes.

4.1. Measuring the textual content

Aside from the component of image(s), there are five different textual components that can be used to construct the content text of a tweet, namely (1) text with semantic meaning, (2) hashtag(s), (3) @username(s) mentions, (4) hyperlink, and (5) unused space. To calculate entropy for each component, the NodeXL Pro Software automatically collected Twitter network information for component "*Hashtag*s in Tweet" and *Hyperlink* in "URLs in Tweet" column. The *@username* component was identified as vertexes for each edge in NodeXL dataset. The component *Unused Space* for each tweet was calculated by the formula: *unused space equals 140 characters minus the length of the tweet*. The component *Text with Semantic Meaning* was the textural content of a tweet excluding all the components of *Hashtag*, *@username*, and *Hyperlink*.

Although the relationship between the choice of six components and the characteristics associated with Twitter profiles (*personal/business account, followers, favorites, tweet counts, etc.*) is not the focus in this study, it is assumed that different choices among the various combinations of the six components have conspicuous impact on efficiency of communication on the Twitter platform. For instance, the main goal of text with semantic meaning is to convey an idea or make a point. The "mention" function, namely @username, is usually viewed as a string or as specifying the recipient of the message. A hyperlink does not have a semantic meaning at all but it could direct the audience from the Twitter platform to other web resources. Hashtag is a hybrid feature; sometimes its semantic meaning serves as a phrase with grammatical value in a sentence; other times it serves as a navigation aid (*keywords*) for information retrieval.

For each previously identified component, the collection of all its entities is called the coding alphabet [15]. For this study, each alphabet was generalized by summing up the total number of unique entities in each component. The next step was to calculate the frequency of occurrences for each entity of a specific component in each cell of Table 1. Regarding calculating the logarithm of empirical frequency, this study chose 2 as the base of logarithm and then multiplied the frequency of an entity in an alphabet with its corresponding logarithm. Choosing 2 as base of logarithm makes the unit of the results of base 2 logarithm "Bits", as recommended by J.W. Tukey to Claude Shannon [20]. According to information theory, the calculation of a logarithm should use probability of occurrence of each entity in the scheme. However, in a real-world scenario especially in a study of social media data stream like this one where the theoretical probability was unavailable, the empirical frequency was used instead.

Each individual tweet in the samples has a unique $H'_{(tweet\text{-}x)}$ value. However, this $H'_{(tweet\text{-}x)}$ was not entropy value because entropy is a measure of the overall property for a closed system, therefore the concept of entropy could not be applied on single-tweet level. The entropy of each component in the tweet-trail was denoted as $H'_{(component)}$ and calculated using the following equation:

$$H'_{(component)} = P(x_{tweet\text{-}1}) \times \log_2 P(x_{tweet\text{-}1}) + \ldots + P(x_{tweet\text{-}n}) \times \log_2 P(x_{tweet\text{-}n})$$

The entropy of the textual content of the tweet-trail was denoted as $H'_{(content)}$, and was the integrated value that calculated by summing up all the values of entropy for each of the five different components as follows:

$$H'_{(content)} = H'_{(\#)} + H'_{(HL)} + H'_{(@)} + H'_{(space)} + H'_{(txt)}$$

4.2. Measuring the image component

An image can be numerically represented in many ways. According to Marr, representation is used to clarify certain characteristics of an entity in a system and to provide a scheme for coding [16]. Knowledge about patterns of the characteristics is crucial for determining a functional and appropriate representation for coding scheme in a system. In 2009, Anderson and O'Connor used RGB data to map color distribution of each frame in the Bodega Bay scene for structural analysis of the sequence of Hitchcock's movie "The Birds" [1]. Likewise, in this study, a set of three numbers, namely the average RGB values (*from 0 to 255*), was used to represent each image in a single tweet.

In each image there is a possibility of 256 shades of red, green, and blue color. In total over 16 million (256^3) combinations are available to represent a single image file. For those cases where a single tweet contained more than one image, the set of weighted average RGB values of all images in that tweet served as the numerical representation. This approach provided an objective way to token an image file without human intervention. In a repetitive test with over 700 images, this approach appeared to be effective and adequate. No identical set was assigned to different images. Sometime there are textual tweets contain the same image but with different contents; while other times tweets share both content and image, but those image files are in different resolutions. As a result of this method, the numerical set was identical for the same images across different tweets regardless of file size.

All the red values in each RGB set constructed the alphabet of red color for that tweet-trail, and so did the green and blue color. As shown in the following equation, the frequency of each value of red, green, and blue color was calculated and then multiplied by its own logarithm than adding up together to get the entropy of each color:

$$H'_{(red/green/blue)} = P(color_{image\text{-}1}) \times \log_2 P(color_{image\text{-}1}) + \ldots + P(color_{image\text{-}n}) \times \log_2 P(color_{image\text{-}n})$$

The synthetic value of entropy of the image component of the tweet-trail was denoted as $H'_{(image)}$ and was calculated by summing up all the values of entropy for each of the three colors as expressed by the following equation:

$$H'_{(image)} = H'_{(red)} + H'_{(green)} + H'_{(blue)}$$

The reason for such a configuration with the image component being composed of three different entropy values is that an image in a tweet takes up a certain amount of space in any display devices. The Twitter default size of the image (*440 X 220*) is always larger than the textual content (*140 characters*) of the tweet.

In a study of evaluating the effect of pictures on health communication, investigators found that "pictures closely linked to written or spoken text can, when compared to text alone, markedly increase attention to and recall of health education information" [13]. As a multi-media supplement for textual communication messages, image plays a crucial role not only in visualizing the main idea of the content but also in attracting users' attention in order to increase the probability of being retweeted. Therefore, it is arguable that the image compoent accounts for more proportions in the $H_{(x)}$ matrix than any of the other components alone.

4.3. $H_{(x)}$ as a variety index

The final product of the calculation matrix is $H_{(x)}$ and is calculated by the following formula:

$$H_{(x)} = H_{(trail)} = H'_{(content)} + H'_{(image)}$$

The calculated result of $H_{(trail)}$ was made up of eight entropy values from six different components in a tweet-trail (*the image component was composed of red, green, and blue three different color subsets*). These components were on a unique level of granularity of the tweet-trail to represent the diversity of the statistical structure in terms of choosing different components.

In this study, $H_{(x)}$ was used as an indicator of complexity in the structure of a tweet-trail. In addition, complexity in the structure is an indicator of the variety in tweeting behaviors in terms of choices for tweet composition. For example: individual users might involve more point to point communication using @username mention function while healthcare agencies might tend to embed hyperlink into their tweets to drive network traffic to the target webpages. For this reason, the structure of medical-terminology tweet-trail could be different from the one of lay-language under the assumption that users with different profiles have preference towards one hashtag of this pair over the other. Therefore, $H_{(x)}$ served as a variety index or an indicator of the complexity in the structure of a tweet-trail.

5. Data visualizations

This study employs radar graphs and scatter graphs as data visualization aids to get an intuitive demonstration. These graphs are viewed as selfies of the hashtags trails because they visualize the complexity in the structure and reveal the pattern of the characteristics of each individual tweet in the trail. The word "selfie" was originated from

social media platforms and refers to a photograph of oneself. In this study, the word "selfie" was introduced to represent the snap shot of a tweet-trail on Twitter because it provides information about the structure and composition of that trail and is unique for each individual tweet-trail.

5.1. Radar graphs

For the purpose of comparing and contrasting each pair of tweet-trails, values of the cells in the last row of Table 1, namely $H'_{(hashtag)}$, $H'_{(@username)}$, $H'_{(hyperlink)}$, $H'_{(space)}$, $H'_{(text)}$, and the sum of $H'_{(red)}$, $H'_{(green)}$, and $H'_{(blue)}$, were harvested and organized with six vectors on a radar graph by their weighted average proportion in the tweet-trail. Then, the radar graphs for each tweet-trail in a pair were placed together to build a combined radar graph for this pair of tweet-trails. A radar graph shows the weight of each component in the tweet-trail. The more weight a component gains, the closer the shape of radar gets to the vertex of that component.

5.2. Scatter graphs

As shown in Table 2, the value of $H'_{(component)}$ was calculated separately and then aggregated into $H_{(x)}$. On the other hand, for each tweet in the matrix, its own $H'_{(tweet)}$ was calculated by summing up all the $P(x_i) \times \log_2 P(x_i)$ entities for each component in that tweet (if presents). The rationale behind the summation is that (1) according to information theory, the entropy of the joint event is "equal to the sum of the individual uncertainties" [20], and (2) all the cells in the matrix have the same unit, bits; because the values of these cells are the calculative results of the frequency of an entity in a tweet multiplied by the logarithm of its frequency.

6. Data analysis and visualizations

Table 3 summarizes the calculated results of the $H_{(x)}$ matrix for the two pairs. The value in each cell was the result of entropy calculation of each component in a given trail. $H'_{(image)}$ in this table equals the sum of $H'_{(red)}$, $H'_{(green)}$, and $H'_{(blue)}$. $H'_{(content)}$ equals the sum of $H'_{(hashtag)}$, $H'_{(hyperlink)}$, $H'_{(@username)}$, $H'_{(space)}$, and $H'_{(text)}$. $H_{(trail)}$ equals the sum of $H'_{(image)}$ and $H'_{(content)}$. A comprehensive list of all tweets in the #hypertension tweet-trail and the calculating process of $H_{(x)}$ matrix for this trail are provided as appendix 1 and appendix 2.

Table 3. Calculated results of $H_{(x)}$ matrix

Tweet-trail	H' (hashtag)	H' (hyperlink)	H' (@username)	H' (space)	H' (text)	H' (content)	H' (image)	H (trail)
#glucose	6.65	3.06	2.92	3.70	13.38	29.71	15.32	*45.02*
#bloodsugar	6.46	3.27	2.38	3.68	13.65	29.45	16.08	*45.53*
#hypertension	5.38	3.46	1.90	3.54	11.88	26.17	10.69	*36.87*
#bloodpressure	6.26	3.84	4.62	3.83	11.53	30.07	17.15	*47.22*

The values of $H_{(trail)}$ of the #glucose tweet-trail and #bloodsugar tweet-trail showed little difference (*45.02 versus 45.53*), suggesting these two tweet-trails had similar degree of complexity in their own structures. The value of $H_{(trail)}$ of the #hypertension trail is obviously lower than the one of #bloodpressure trail (*36.87 versus 47.22*) because there were only 61 tweets in #hypertension trail in contrast to the 94 tweets in #bloodpressure trail. This finding was consistent with the observations in the value of $H'_{(image)}$ of this pair of tweet-trails (*10.69 versus 17.15*) and $H'_{(content)}$ of this pair (*26.17 versus 30.07*), indicating that the total number of tweets in each trail was an influential factor on the final results of $H'_{(content)}$, $H'_{(image)}$, and $H_{(trail)}$ for that tweet-trail.

6.1. Radar graphs for #glucose tweet-trail versus #bloodsugar tweet-trail

Figure 1 illustrates the comparative radar graphs for case one, #glucose (*in blue*) versus #bloodsugar (*in orange*). The selfies of both tweet-trails almost overlapped because the #glucose trail and #bloodsugar trail had almost the same variation in the composition of the components in their respective structures (*shape and size*). This finding further indicates that these two medical hashtags are interchangeable in usage because the users made very similar choices in selecting components when composing their tweets.

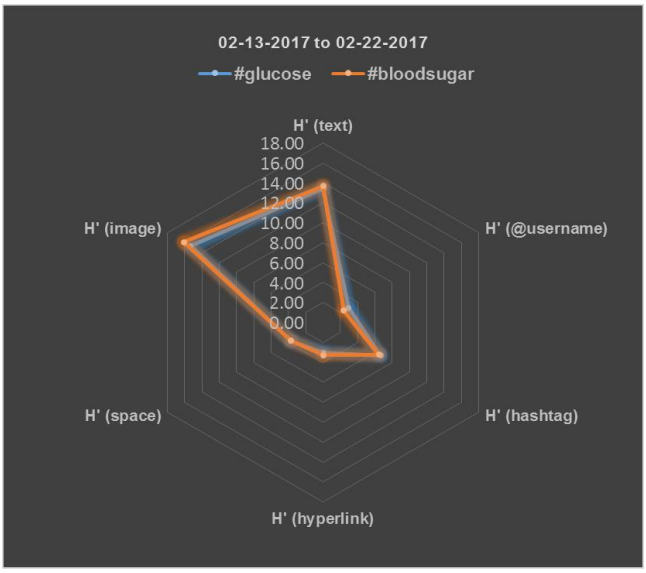

Figure 1. Combined radar graphs of #glucose versus #bloodsugar

A pilot study was conducted from January 26th to January 29th, 2017 to collect #glucose trail and from February 4th to February 9th, 2017 to collect #bloodsugar trail. The sampling and data cleaning process followed the same procedure as described in this study. Figure 2 shows the result of this pilot test. The #glucose trail and #bloodsugar trail have almost identical shape and size of radar graph even although they covered different time frame. When combining the finding of pilot test with the result of the formal study, it revealed consistency in the structures of this pair of tweet-trails, suggesting that the tweeting behaviors associated with these two hashtags were stable over time.

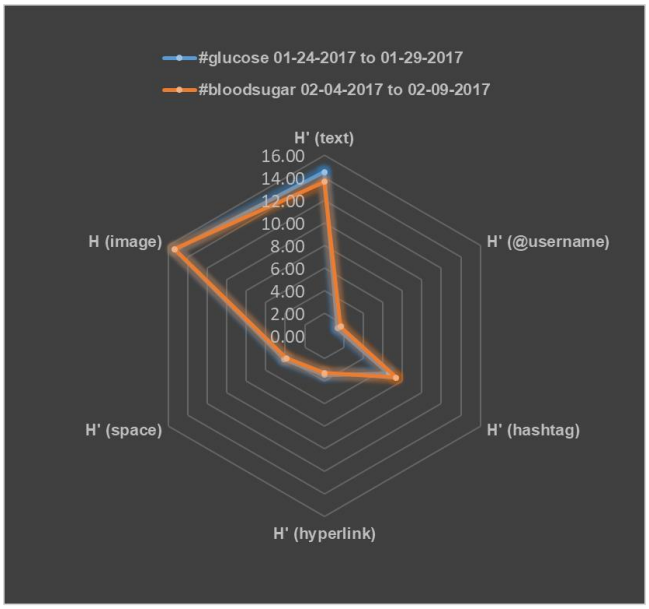

Figure 2. Results of the pilot study, Combined radar graphs of #glucose versus #bloodsugar

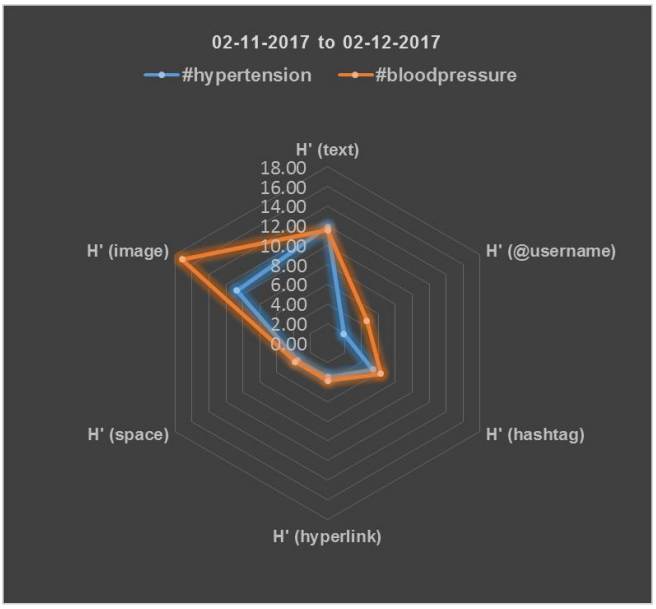

Figure 3. Combined radar graphs of #hypertension versus #bloodpressure

6.2. Radar graphs for #hypertension tweet-trail versus #bloodpressure tweet-trail

Figure 3 illustrates the paired-up radar graphs for case 2, #hypertension versus #bloodpressure. Unlike the results in case 1, the size of the selfie of the tweet-trail of #hypertension is much smaller than that of #bloodpressure. The reason for that is because the total number of tweets in the #hypertension trail was 61, 35% fewer than the 94 tweets in the #bloodpressure trail.

The shapes of these selfies were also very different, indicating that this pair of tweet-trails had very distinct structures from the ones in case 1. One possible reason for the difference in the shapes of the radar graphs might be that the semantic meaning of "hypertension" is not exactly the same as that of "blood pressure". Hypertension in English means high blood pressure and its opposite word is hypotension, low blood pressure. The difference in the perception of semantic meaning caused users to make different choices among the six typical components when composing tweets. The reason why this study didn't include #hypotension tweet-trail was that #hypotension was not a popular hashtag and there were less than 10 tweets contained #hypotension collected during February 2017, causing insufficiency in data for generating visible radar graph (*the size of the radar graph was too small around the center to be an intuitive visualization aid*).

6.3. Scatter graphs for #glucose tweet-trail versus #bloodsugar tweet-trail

In Shannon's original entropy equation, the factor of time is absent. However, each tweet in this study in the tweet-trail has its own tweet timestamp. The timestamp of each tweet was combined with its own $H'_{(tweet)}$ harvested from the $H_{(x)}$ matrix and plotted on a separate scatter graph for each tweet-trail. Figure 4-1, Figure 4-2, Figure 5-1, and Figure 5-2 illustrate the distributions of each individual tweet in the given tweet-trails plotted with its $H'_{(tweet)}$ value along the time frame from February 13th to February 22nd, 2017. $H'_{(tweet)}$ was the synthetic value of a tweet because it was the sum of $P(x_i) \times \log_2 P(x_i)$ for all components in that tweet.

Although the combined radar graph showed high-level similarity in structures of these pair of tweet-trails, the scatter graph for each tweet-trail revealed very different pattern in terms of the density of tweeting/retweeting activities. The tweet-trail of #glucose (*Figure 4-1*) contained 96 tweets and the #bloodsugar tweet-trail (*Figure 4-2*) had 95 tweets. The size and shape of their radar graphs were the same, indicating they had identical data structures. However, the #bloodsugar trail had more intense tweeting/retweeting activities around February 15th, 2017. The tweets with #glucose in the ten-day timeframe were more evenly distributed. This finding suggests that although #glucose trail and #bloodsugar trail have similar structure in terms of their tweets data stream, the tweeting/retweeting activities that associated with each of these two hashtags thrived in different time frames. For example: #glucose trail was more active than #glucose trail during February 14th to February 15th, 2017, then #bloodsugar trail began to be dynamic from February 16th to February 19th, 2017 while #glucose trail was fading during that time. Then another uphill was observed in #glucose trail around February 22nd, 2017 while the activities of #bloodsugar trail started to decline.

6.4. Scatter graphs for #hypertension tweet-trail versus #bloodpressure tweet-trail

Figure 4-1. Scatter graph of #glucose tweet-trail from 02-13 to 02-22-2017

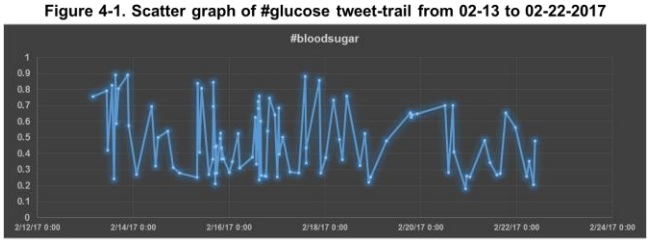

Figure 4-2. Scatter graph of #bloodsugar tweet-trail from 02-13 to 02-22-2017

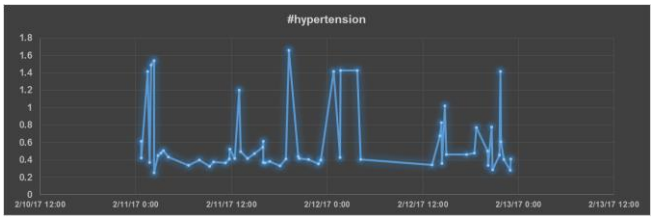

Figure 5-1. Scatter graph of #hypertension tweet-trail from 02-11 to 02-12-2017

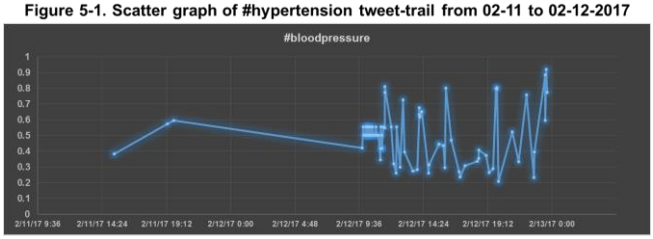

Figure 5-2. Scatter graph of #bloodpressure tweet-trail from 02-11 to 02-12-2017

The #hypertension trail (*Figure. 5-1*) had fewer tweets between February 11th and 12th, 2017; however, these tweets had relatively even distribution. In contrast, the #bloodpresure trail (*Figure. 5-2*) had many more tweets with unbalanced distribution. The comparison of this pair also shows approximately complementary feature in the density of distribution of their own tweets, the same pattern as what case 1 had revealed. However, the cause of this phenomenon cannot be explained solely by structural analysis so it will be further investigated in future studies.

For the purpose of gaining insights from a more intuitive demonstration, a small sample of the #bloodsugar trail was randomly extracted and marked at each timestamp with the capture of the tweet. After mapping the snapshots of each tweet with its own H' $_{(tweet)}$ value along the timeline on the scatter graph as shown in Figure 6, the pattern of characteristics of the distributed tweets emerged.

For any given tweet-trail, those tweets with higher H'$_{(tweet)}$ value had always been staying on the top area of the scatter plot, indicating relative higher complexity in terms of their statistical structures in contrast to the ones at the middle and bottom areas.

Those tweets with high H'$_{(tweet)}$ values were the ones mostly contained image(s) and almost every one of them was a retweet of some original tweet. A retweet means a reposted or forwarded message on Twitter. The tweets with lower synthetic value (*at the bottom area*) were those with low complexity in terms of the structure and low variety in terms of tweet composition. Those were mainly original textual tweets without any image attached.

7. Discussion

This study examined how healthcare communication messages on Twitter (*i.e. tweets*) were constructed by analyzing the complexity in structural components and variety of tweet composition in two pairs of tweet-trails with medical hashtags. Healthcare topics are sophisticated and healthcare communication messages usually resort to the aid of rich media such as image(s)/Video to visualize ideas and/or external hyperlink to direct audience to the destination webpage with further explanation. This phenomenon concurs with the results of the comparative case study in which 41.74% of the total tweets (*including all the samples of #glucose, #bloodsugar, #hypertension, and #bloodpressure trails*) incorporated image, and 69.36% of the total tweets contained a hyperlink.

In this study, the tweeting behavior was defined as the choices made among six typical components to construct a tweet. These observed tweeting behaviors were assumed to be associated with different types of Twitter users, meaning

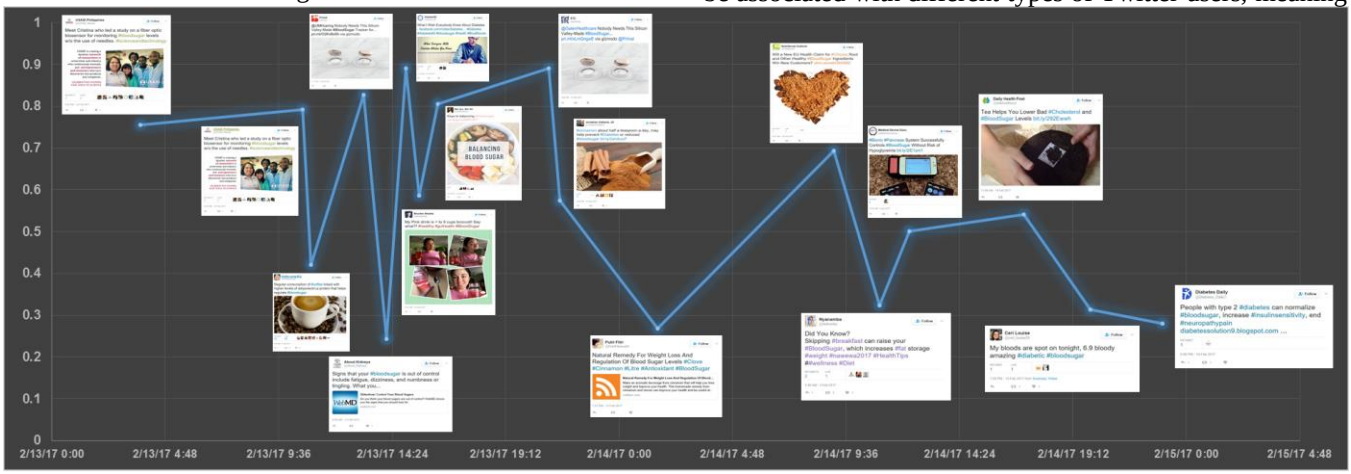

Figure 6. Captures of tweets in #bloodsugar trail distributed along with their timestamps from 02-13-2017 to 02-15-2017

that tweeting behaviors varied across Twitter accounts with diversified profiles. These components are independent of each other. Each tweet has limited space (i.e. 140 characters) to express its main idea. Therefore, a user's choice between the medical-terminology hashtags and lay-language hashtags requires consideration of the opportunity cost for the different options. Interestingly, the percentage of single tweet that contained both medical-terminology hashtag and lay-language hashtag was very low in both cases (*less than 2%*), indicating the fact that users tend to reduce the redundancy in hashtag usage by avoiding hashtags with similar or identical semantic meanings.

The findings from Figure 6 are summarized and organized in Figure 7 and Figure 8. As shown in Figure 7, the major factors that can differentiate tweets in the tweet-trail include (1) the complexity level of the tweet structure, and (2) the originality of tweet (*i.e., whether the tweets are original tweets or retweets*). In a given tweet-trail, a simple structure is defined in this study as a structure with a low level of variation in the combination of different components, while a complex-structured tweet means that the level of variation in the combination of different components in this tweet is high.

Complex Level in Structure	Originality: Original	Originality: Retweet
Simple	*Original tweets with Simple structure*	*Retweets of simple-structured tweet*
Complex	*Original tweets with Complex structure*	*Retweets of complex-structured tweet*

Figure 7. Classification of tweet types

Retweeting activities on the Twitter platform has a direct consequence on the structure of the original tweet: the increase in the complexity in its structure in contrast to that of the original tweet. Being retweeted leads to a higher synthetic value of $H'_{(tweet)}$ given all other conditions remain the same.

As presented in Figure 8, the results of this case study revealed the pattern that either being retweeted or applying a variety of components (*especially image*) when constructing a tweet contributes to relative medium to high synthetic value of $H'_{(tweet)}$. The potential application of this approach is to provide an alternative method of automatically detecting retweets with more information about the structure and composition of these retweets.

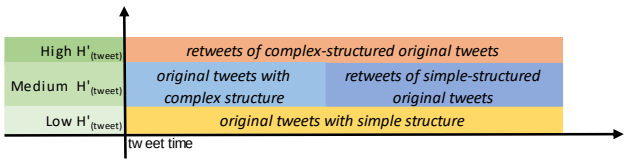

Figure 8. The distribution of individual tweet by the value of its $H'_{(tweet)}$

8. Limitations of the study

First, this case study only investigated two pairs of hashtags (*medical-terminology versus lay-language*) and the results may reflect only part of the story. More cases of medical hashtags with similar semantic meanings between medical-terminology and lay-language can be collected and compared in order to generalize the results.

Second, this study introduces $H_{(x)}$ as a variety index for analyzing the complexity in structure in a tweet-trail. However, it only reflects relative degree of complexity in statistical structures. According to information theory, the statistical structure of message is irrelevant to the semantic aspect of communication, which means complexity in structure doesn't necessarily lead to higher informativeness in its content. $H_{(x)}$ is not suitable as an indicator for evaluating the content value of these tweet-trails.

Third, this study assumes hashtags serve only as keywords for information retrieval. The investigated hashtags were not supposed to have grammatical value. However, in reality, hashtags sometimes serve as a phrase in a sentence. For situation like this, the data preparation involves more manual efforts or more sophisticated algorithm and the calculating process of $H_{(x)}$ matrix would be more complex due to the duality of hashtags.

9. Conclusion

To the best of our knowledge, this study is the first one to apply information theory to evaluate tweet composition by accounting for the granularity of a tweet. It examines the use of hashtags of medical-terminology versus lay-language in Twitter data stream and introduces $H_{(x)}$ index as a measure to compare and contrast the statistical structures of the components in different tweet-trails. This index reveals the complexity in the structure and the variety of the components chosen in composing a tweet with a well-defined coding scheme.

Another contribution of this study is its novel data visualization tools to depict the measurement results. Both radar graph and scatter plot are intuitive demonstrations to illustrate the typical components of a tweet-trail, providing insights in tweet-composition styles in the context of health communication. The radar graph and scatter graphs work together to provide more insights when two tweet-trails have similar structures.

Third, this study proposes a systematic framework, the $H_{(x)}$ matrix which extends the classical entropy calculation to a multi-dimensional matrix for analyzing tweet-trails with complex structure. This methodological framework is designed for structural analysis of Twitter data stream as a supplemental tool for profile analysis of Twitter users and content analysis of tweets. Sometimes content analysis might be compromised by the celebrity effect, a tweet by a celebrity gets retweeted many times right after its birth, which causes high-dense burst in trail and distorts the trend in figure. When structural analysis is working together with content analysis and profile analysis, the whole picture of Twitter data stream would be much clearer than before.

This framework of $H_{(x)}$ index and matrix is unlikely to be a sole/major analyzing tool for studies of social media data stream. However, it is capable of unveiling patterns in

structure and provide quick and preliminary snap shots (*selfies*) of Twitter data stream because it's an automatic approach and requires no human intervention. The approach presented in this study could be argued the missing piece of a holistic analytic system and gives researchers an opportunity to observe the social media data stream from a whole new perspective and to examine what Claude Shannon called "the engineering aspect" of the events [20].

10. Future studies

First, video and emoji are important features that are commonly incorporated in a tweet. Therefore, future studies should consider including video and emoji as two extra components in the current coding scheme. Second, exploring alternative representations of the image(s) as a measurement in tweets could be another future research direction. The current solution of assigning a set of average RGB color to each image has a unique tendency. A dark image, in general, has relatively lower average RGB values than a bright one. Although the final effect is determined by the ratio of all six components and the image component only takes 3/8 of the total proportion, this difference in the values of RGB color still might result in minor difference in the values of $H'_{(tweet)}$ among different tweets and eventually bias the distribution of these tweets in scatter graph. Since $H'_{(tweet)}$ in the configuration of this study is a synthetic value made up of six different components, it is worth exploring whether adding new components to a tweet or using an alternative token for images would have a significant impact on the efficacy of the $H_{(x)}$ framework.

11. References

[1] Anderson, Richard L., and Brian C. O'Connor. "Reconstructing bellour: Automating the semiotic analysis of film." Bulletin of the American Society for Information Science and Technology 35, no. 5 (2009): 31-40

[2] Baldi, Pierre, and Laurent Itti. "Of bits and wows: a Bayesian theory of surprise with applications to attention." Neural Networks 23, no. 5 (2010): 649-666.

[3] Beguerisse-Díaz, Mariano, Amy K. McLennan, Guillermo Garduño-Hernández, Mauricio Barahona, and Stanley J. Ulijaszek. "The 'who'and 'what'of# diabetes on Twitter." DIGITAL HEALTH 3 (2017): 2055207616688841.

[4] Boell, Sebastian K. "Information: Fundamental positions and their implications for information systems research, education and practice." Information and Organization 27, no. 1 (2017): 1-16.

[5] Borgwaldt, Susanne R., Frauke M. Hellwig, and Annette MB de Groot. "Word-initial entropy in five languages: Letter to sound, and sound to letter." Written Language & Literacy 7, no. 2 (2004): 165-184.

[6] Brown, Peter F., Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. "An estimate of an upper bound for the entropy of English." Computational Linguistics 18, no. 1 (1992): 31-40.

[7] Chang, Hsia-Ching. "A new perspective on Twitter hashtag use: Diffusion of innovation theory." Proceedings of the American Society for Information Science and Technology 47, no. 1 (2010): 1-4.

[8] Chong, Andrew, Lalitha Sankar, and H. Vincent Poor. "Frequency of occurrence and information entropy of American sign language." (2009).

[9] Ghosh, Rumi, Tawan Surachawala, and Kristina Lerman. "Entropy-based classification of'retweeting'activity on twitter." (2011).

[10] Hartley, Ralph VL. "Transmission of information." Bell Labs Technical Journal 7, no. 3 (1928): 535-563.

[11] Hayes, Bruce, and Colin Wilson. "A maximum entropy model of phonotactics and phonotactic learning." Linguistic inquiry 39, no. 3 (2008): 379-440.

[12] Hayes, Robert M. "Measurement of information." Information Processing & Management 29, no. 1 (1993): 1-11.

[13] Houts, Peter S., Cecilia C. Doak, Leonard G. Doak, and Matthew J. Loscalzo. "The role of pictures in improving health communication: a review of research on attention, comprehension, recall, and adherence." Patient education and counseling 61, no. 2 (2006): 173-190.

[14] Kearns, Jodi, and Brian O'Connor. "Dancing with entropy: Form attributes, children, and representation." Journal of Documentation 60, no. 2 (2004): 144-163.

[15] Kinsner, Witold. "Is entropy suitable to characterize data and signals for cognitive informatics?" (2004): 6-21.

[16] Marr, David, and A. Vision. "A computational investigation into the human representation and processing of visual information." WH San Francisco: Freeman and Company 1, no. 2 (1982).

[17] Miller, George A. "What is information measurement?" American Psychologist 8, no. 1 (1953): 3.

[18] Neubig, Graham, and Kevin Duh. "How Much Is Said in a Tweet? A Multilingual, Information-theoretic Perspective." In AAAI Spring Symposium: Analyzing Microtext. 2013.

[19] Ritchie, David. "Shannon and Weaver: Unravelling the paradox of information." Communication research 13, no. 2 (1986): 278-298.

[20] Shannon, Claude E, A mathematical theory of communication. Bell System Technical Journal 27, no.3 (1948): 379-423.

[21] Watt, James. "Television form, content attributes, and viewer behavior." Progress in communication sciences 1 (1979): 51-89.

[22] Williams, Shirley A., Melissa M. Terras, and Claire Warwick. "What do people study when they study Twitter? Classifying Twitter related academic papers." Journal of Documentation 69, no. 3 (2013): 384-410.

[23] Zimmer, Michael, and Nicholas John Proferes. "A topology of Twitter research: Disciplines, methods, and ethics." Aslib Journal of Information Management 66, no. 3 (2014): 250-261.