Identification of Influential Features for User Engagement with Wellness Content: Analysis of nyt_well Instagram Account

Seungjun Kim *

Posted Date: 12 April 2023

doi: 10.20944/preprints202304.0268.v1

Keywords: wellness; health; user engagement; social media; instagram; negative binomial regression

Article

Identification of Influential Features for User Engagement with Wellness Content: Analysis of nyt_well Instagram Account

Seungjun Kim

BA, University of California, Irvine, Department of Informatic; kims17@uci.edu

Abstract: Wellness is a multidimensional concept that touches upon the various physical, mental, emotional, spiritual, social and environmental facets of health. Interest towards and importance of wellness have been growing constantly for the past two decades and thus makes it crucial to understand which factors affect public engagement with wellness information for multiple stakeholders. The Instagram account of New York Times (NYT) specifically for sharing wellness content with the handle *nyt_well* was selected as the object of study. 773 posts from this account between March of 2019 and December of 2022 were collected and analyzed to answer the research question of which factors are most influential to public engagement with wellness content. Two negative binomial regressions were run on features including the type of post, length, word count, sentiment score and topic with number of likes and comments as the dependent variables for each of those regression models. Results indicated that the type of post and its sentiment score were the two most influential determinants of public engagement with p-values smaller than 0.05. While the effects of some of these factors aligned with findings from previous studies conducted on social media content not related to wellness (e.g., marketing), some others affected the two separate public engagement metrics in opposite directions, warranting future studies to investigate further on the cause of this phenomenon.

Keywords: wellness; health; user engagement; social media; instagram; negative binomial regression

Introduction

The Global Wellness Institute defines wellness as the active pursuit of activities, choices and lifestyles that lead to a state of holistic health. It is multidimensional in the sense that it touches upon the various physical, mental, emotional, spiritual, social and environmental facets of health [1]. More recent literature adds onto this definition of wellness by including the financial, vocational and relational aspects [2].

Interest towards and importance of wellness have been growing constantly for the past two decades. According to Google Trends, the search volume for the term wellness in January, 2023 was more than three times than that in January, 2004 [3]. In 2022, McKinsey conducted a survey on wellness suggesting that around 50% of US consumers now report wellness as a top priority in their day-to-day lives [4]. Various services and products that are intended to improve wellness, so called the wellness economy, represent 5.1% of global economic output as of 2020 [5].

Therefore, it is crucial to understand which factors affect public engagement with wellness information for multiple stakeholders. For media outlets including newspapers, it is of their interest to maximize engagement metrics such as the number of likes and comments which are often associated with advertisement revenue or their overall brand reputation and prestige. For non-profit organizations or government agencies that focus on promoting public health and wellness, understanding factors that affect public engagement with wellness content will guide them on how to design and better their campaigns and information sharing strategies for their target populations.

Among several places where wellness information can be accessed, social media is one of the most influential platforms where wellness information is posted and digested by people. As of 2022, approximately 302 million users were reported to be using social media in the U.S. [6] With such a massive user base, social media is increasingly receiving attention as an effective channel for conveying wellness related information to the public for awareness. A study indicates that social media is viewed as acceptable and accessible among multiple audiences and also promotes health equity among disadvantaged populations including senior citizens, people living in rural areas and with low income [7]. Social media show great promise as effective wellness communication channels for young adults (e.g., 18-30 years old) and also as tools to be utilized by wellness educators to present important risk management and disease prevention information to them [8]. In another study, nearly 85% of healthcare practitioners including nurses, nurse practitioners, pharmacists, physicians, and healthcare administrators agreed that social media can be an effective tool for the purpose of wellness education [9]. Instagram, in particular, has been a popular choice and was ranked the number five most downloaded application in the first quarter of 2018 [10]. On Instagram, users can share their life updates by posting image and video content. This nature of Instagram whose content is mainly based on visual cues can make it a better platform for engagement with users compared to other social media platforms. The enhanced intimacy it provides unlike text-based platforms (e.g., Twitter, Yik Yak) may offer another explanation for better engagement [11].

The Instagram account of New York Times (NYT) specifically for sharing wellness content was chosen as the object of study with 274,000 followers as of January 11th, 2023. 773 posts from this account between March of 2019 and December of 2022 were collected and analyzed to address the following research question–which were the most influential factors that determined public engagement with wellness content?

Methods

Data collection

The Instagram account of New York Times (NYT) specifically for sharing wellness content with the handle *nyt_well* was selected as the object of study. NYT was considered a suitable choice for this study as it had 7.5 million subscribers as of 2020, the most number of paid subscribers in the world and may be able to offer insights that are more generalizable than other social media accounts that target very niche populations [12]. A total of 773 posts from when this account was first created in March of 2019 until December of 2022 were collected along with their metadata including their timestamp, type (e.g., image, sidecar and video), number of likes, numbers of comments and content. Posts that were completely irrelevant to the topic of wellness such as the celebration post for Halloween were excluded. The posts were scraped through a combination of the PhantomBuster Application Programming Interface (API) and manual collection as the API imposed limits on the number of minutes it would allow for collecting social media data for free. Metadata specific to the user were not collected due to potential privacy issues.

Identification of topic groups

Non-Negative Matrix Factorization (NMF) was used to identify various topic groups and keywords that represent each of them. The Term Frequency–Inverse Document Frequency (TF-IDF) algorithm was used to embed the text into vectors to which NMF were applied [13,14]. The appropriate number of topics (k = 7) to be modeled was determined via an iterative approach of trying all the values between 2 to 10 inclusive and manually checking for the quality of topics modeled.

NMF is an algorithm drawing from linear algebra concepts and theories that decomposes high-dimensional vectors into a lower-dimensional representation. It is known to enhance compression and interpretability compared to other dimensionality reduction techniques such as Principal Component Analysis (PCA) [15]. While other classic topic modeling algorithms including Latent Dirichlet Allocation (LDA) and BERTopic were also tested, they led to poor results with inconsistent

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 April 2023 doi:10.20944/preprints202304.0268.v1

keywords within topics or inability to identify more granular topic groups (e.g., lumping two separate topic groups into one). Furthermore, NMF has been previously applied to unstructured online comments of patients and clinical notes for topic modeling and showed promise especially for shorter texts such as tweets compared to other topic modeling algorithms [16–18]. The captions in Instagram posts were not lengthy in general which may have made the NMF a suitable choice.

Feature engineering

On top of the Instagram metadata collected, additional features including the length of the text, word count, and sentiment probability scores were created. The sentiment probability score, in particular, was calculated using the DistilBERT model made available by the HuggingFace interface [19]. For posts associated with negative sentiment, the probability scores were transformed into negative values. DistilBERT is a lighter version of the Bidirectional Encoder Representations from Transformer (BERT) model released by Google in 2018 which yielded state-of-the-art results on eleven Natural Language Processing (NLP) tasks and thus was selected as the model of choice for determining the sentiment of posts [20–21]. Based on this sentiment labeling, the percentages of positive and negative posts were calculated. Those percentages were recalculated within each topic group to see if proportions of positive and negative posts vary across topics.

Factors for post engagement

Once the explanatory variables for regression listed below in Table 1 were set up, several regression models were considered. The dependent variables were the number of likes and comments of Instagram posts, both of which are non-negative integers. Counts data are known to often display overdispersion manifested by highly right skewed distributions with high variability [22]. The data in our study shows potential overdispersion of likes and comments as illustrated by Figure 1. Due to these characteristics, the Ordinary Least Squares (OLS) regression or a typical linear regression with log-transformation of the data was deemed infeasible due to the inability to meet its assumptions [22]. While both Poisson and the negative binomial regression were popular selections for modeling count data, negative binomial regression became the final model of choice because of the less strict assumptions as opposed to Poisson regression which assumes the mean to be equal to the variance for accurate results. Model estimation was conducted separately for the number of likes and comments respectively. Lastly, software STATA (version 15) was used to run the negative binomial regressions.

Table 1. Explanatory variables for negative binomial regression.

Explanatory Variables	Description
year	Categorical variable that indicates the year the post was uploaded to the account. (2019 to 2022)
month	Categorical variable that indicates the month the post was uploaded to the account. (1 to 12)
topic_num	Topic Number. (1 to 7)
length	Total length of the caption in each post.
wc	Word count of the caption in each post.
type	Type of post. (Image, Sidecar, Video)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 April 2023 doi:10.20944/preprints202304.0268.v1

sentiment_score	The higher the absolute value of the score is, the stronger the sentiment is. (-1 to 1)

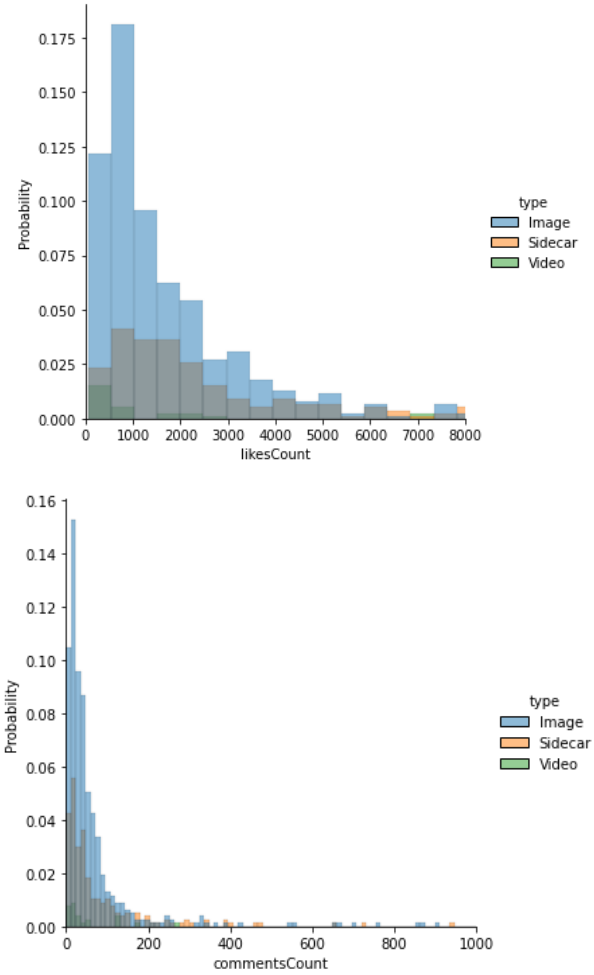

Figure 1. Distribution of number of likes and comments by post type

Results

Number of posts by type and year

In terms of the number of posts and its change over time, 36.5% and 46.2% of the posts were from 2019 and 2020 respectively, making up 82.7% of the posts in total. The number plateaued in 2021 and 2022. In terms of post types, 68.6% of the posts were images, 23.3% of the posts were sidecars that were made up of multiple images that users can swipe, and the remaining 3.1% of the posts were video content. Although there existed fluctuations in these proportions throughout different time periods, image posts were consistently the most prevalent type of post followed by sidecar and video posts.

Post themes

After applying the NMF algorithm on all the posts, the following seven topics were identified: (1) illustrations; (2) relationships; (3) professional tips; (4) kids; (5) victories; (6) parents; and (7) women. Topic 7 and 2 were the two major topics that had the most posts with proportions of 28.8% and 15.9%, respectively. Table 2 lists the counts and proportions of posts in each topic group.

The first topic group consisted of posts with illustrations that mainly focused on personal stories of individuals facing and overcoming wellness challenges. The second topic group revolved around

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 April 2023 doi:10.20944/preprints202304.0268.v1

the theme of relationships with keywords including *family*, *mother*, *son*, *friends* and *daughter* being the top terms with the highest tf-idf values. The theme of the third topic group was professional tips with the word *newsletter* appearing recurrently in multiple forms (e.g., this week's newsletter, newsletter on X, newsletter with guides to Y). The word *subscribe* was another term that had a high tf-idf value among the posts in this group which reflected the campaign of the account to induce subscription to these guides and newsletters. Topics groups four and six were related to each other but focused on two different populations – kids and parents. Posts from the fourth theme featured content including children's book recommendations, home-schooling challenges and multiple aspects of life in school. Posts from the sixth theme were heavy on child care issues and expert advice was often quoted for addressing them. This was supported by the fact that terms *expert* and *spoke* were the two of the top ten terms with the highest tf-idf values. The seventh topic group on women mainly covered content about reproductive health of women which could be inferred from the keywords *baby*, *health*, *pandemic*, *mothers*, and *pregnant*. The fifth topic group was on victories people made against various wellness predicaments they experienced. In particular, the account set up the "small victories" series for parents who shared personal anecdotes on how they overcame problems they faced with child care.

Table 2. Breakdown of number of posts in each topic group modeled by NMF.

Topic Number	Number of Posts	Percentage
7	223	28.6%
2	123	15.9%
4	114	14.7%
6	112	14.5%
1	92	11.9%
3	72	9.3%
5	37	4.8%

65.8% of the posts were associated with negative sentiment. The proportions of posts labeled as negative within each group were higher than those labeled as positive across all topics except for the fifth topic group on victories. Table 3 displays the counts and percentage breakdowns of posts by sentiment in each topic group.

Table 3. Breakdown of posts by sentiment in each topic group.

Topic Number	Sentiment	Count	Percentage
1	Negative | Positive	67 | 25	72.8% | 27.2%
2	Negative | Positive	67 | 56	54.5% | 45.5%
3	Negative | Positive	54 | 18	75.0% | 25.0%
4	Negative | Positive	68 | 46	59.6% | 40.4%
5	Negative | Positive	17 | 20	45.9% | 54.1%
6	Negative | Positive	80 | 32	71.4% | 28.6%

 doi:10.20944/preprints202304.0268.v1

7	Negative | Positive	156 | 67	70.0% | 30.0%

Factors for post engagement

Sidecar posts, on average, received 1.16 times more likes than single image posts (P=.026). The type of post was statistically insignificant to the number of comments. Posts that fell under topic group 1 received both more likes and comments than those in other topic groups. The average number of likes and comments the posts received consistently increased over the years. Posts from the later half of the year (e.g., July - December), on average, got 1.96 times to 2.77 times more likes compared to January, the reference month. The month variable, however, was statistically insignificant to the number of comments.

Sentiment scores and word count were statistically significant to the number of comments but not to the number of likes. An additional point of negativity in sentiment was associated with an average increase 12% in the number of comments (P=.007). An addition of a word in a post was associated with an average increase of 3.3% in the number of comments (P=.001). However, the length of the post was negatively correlated with both the number of likes and comments. A unit increase in length of the post was associated with a 0.2% (P=.04) and 0.4% (P=.00) decrease in the expected number of likes and comments respectively.

Discussion

This study sought to understand which factors are the most influential in affecting public engagement with wellness information through the lens of an Instagram account run by a prominent newspaper source subscribed by a sizable population. It is worth noting that Large Language Models (LLMs) and topic modeling were used to create features including the sentiment score and topic of the posts in addition to the pre-existing metadata features. This approach allowed us to avoid manual annotations of each of the posts of their respective sentiment and topic labels, thereby saving cost and time.

In terms of the number of posts and its change over time, 36.5% and 46.2% of the posts were from 2019 and 2020 respectively, making up 82.7% of the posts in total. The number plateaued in 2021 and 2022. Figure 2 displays how the number of posts by post type fluctuates over time. This pattern can be potentially explained by the tendency of newly opened instagram accounts to post more content in the beginning to attract more followers but losing momentum once they reach a certain number of followers. In addition, the high number of postings in early 2020 may also be attributed to the COVID-19 pandemic which created a need for wellness and health information to be disseminated to the public in a timely manner.

While several of the previous literature addressed the topic of engagement on Instagram and factors that affect it from a business, marketing and education perspective, not much of the same analysis has been conducted with a focus on public health and wellness [23,24]. This study sought to fill that gap by investigating how different factors of posts influence the level of engagement with wellness content in particular and also whether the way these factors affect engagement are similar or different from how they did for non wellness related posts. Posts with illustrations and sidecar posts were more likely to receive a higher number of likes and comments. This may be explained by the fact that sidecar formats and illustrations can offer richer, more narrative-based and emotional content which can incentivize users to engage more with the posts. This result aligns with findings from a previous study analyzing the posts from university Instagram accounts that carousel formats and emotional content increase both likes and comments. However, it also suggested that posts about sharing achievement can improve likes [23]. This lies contrary to the results of this study in which posts from topic group 5 on victories (e.g., achievement sharing from parents on child upbringing) did not receive, on average, more likes or comments in a statistically significant manner. It would be meaningful for future studies to make sense of why such contrasting conclusions arise and whether they are due to the different topic areas those Instagram posts cover or due to extraneous reasons.

Another interesting pattern was revealed from the ways word count and length of posts affected the number of comments a post received. Despite sharing a generally positive correlation with each other, these two variables affected the number of comments in opposite directions. This may reflect the preference of users for engaging with posts that include more information, narratives and stories and suggest that whether a post possesses such qualities or not is not just determined by its pure length but also by whether the post contains enough words to make it meaningful and this is something the word count variable can act as a proxy for.

The limitations of this study, however, should be noted. Although the NYT is the newspaper firm that retains the most number of subscribers in the United States, insights extracted from the posts from a social media account of a single newspaper venue may not be enough for generalization to a broader population. Therefore, further studies should be conducted on a bigger scale using data that encompass social media posts from multiple newspaper sources that preferably have disparate views and subscriber predispositions to ensure the diversity of the sample collected.

Moreover, this study revealed that several factors of posts affect distinct engagement metrics (e.g., number of likes and comments) differently. For instance, the word count and sentiment score were not statistically significant variables for number of likes but were statistically significant to the number of comments the posts received. Future studies may want to take a step further to understand why certain factors substantially affect the number of likes while they are statistically insignificant to the number of comments and vice versa.

Conclusion

The effects of factors including the type of post, length, word count, sentiment score and topic on mainly two metrics for public engagement with wellness content posted in the *nyt_well* Instagram account were explored in this study using binomial regression models. The type of post and its sentiment score were the two most influential determinants of public engagement with p-values smaller than 0.05. While the effects of some these factors aligned with findings from previous studies conducted on social media content not related to wellness (e.g., marketing), some others affected the two separate public engagement metrics in opposite directions, warranting future studies to investigate further on the cause of this phenomenon.

Conflict of Interest: The author has no other conflict of interest with any stakeholders.

References

1. What is Wellness? Global Wellness Institute. Accessed March 28, 2023. https://globalwellnessinstitute.org/what-is-wellness/
2. Stoewen DL. Dimensions of wellness: Change your habits, change your life. *Can Vet J*. 2017;58(8):861-862.
3. Google Trend. Google Trends. Accessed March 28, 2023. https://trends.google.com/trends/explore?date=all&geo=US&q=wellness&hl=ko
4. Still feeling good: The US wellness market continues to boom | McKinsey. Accessed March 28, 2023. https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/still-feeling-good-the-us-wellness-market-continues-to-boom
5. Statistics & Facts. Global Wellness Institute. Accessed March 28, 2023. https://globalwellnessinstitute.org/press-room/statistics-and-facts/
6. U.S.: social media users 2019-2028 | Statista. Accessed March 28, 2023. https://www.statista.com/statistics/278409/number-of-social-network-users-in-the-united-states/
7. Welch V, Petkovic J, Pardo Pardo J, Rader T, Tugwell P. Interactive social media interventions to promote health equity: an overview of reviews. *Health Promot Chronic Dis Prev Can Res Policy Pract*. 2016;36(4):63-75. doi:10.24095/hpcdp.36.4.01
8. Social Media: The Key to Health Information Access for 18- t... : CIN: Computers, Informatics, Nursing. Accessed March 28, 2023. https://journals.lww.com/cinjournal/Abstract/2015/04000/Social_Media__The_Key_to_Health_Information_Access.2.aspx
9. Pizzuti AG, Patel KH, McCreary EK, et al. Healthcare practitioners' views of social media as an educational resource. *PLOS ONE*. 2020;15(2):e0228372. doi:10.1371/journal.pone.0228372

10. The Top Mobile Apps, Games, and Publishers of Q1 2018: Sensor Tower's Data Digest. Accessed March 28, 2023. https://sensortower.com/blog/top-apps-games-publishers-q1-2018
11. Pittman M, Reich B. Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words. *Comput Hum Behav*. 2016;62:155-167. doi:10.1016/j.chb.2016.03.084
12. Ranked: The Most Popular Paid Subscription News Websites | Markets Insider. Accessed March 28, 2023. https://markets.businessinsider.com/news/stocks/ranked-the-most-popular-paid-subscription-news-websites-1030349711
13. Sammut C, Webb GI, eds. TF–IDF. In: *Encyclopedia of Machine Learning*. Springer US; 2010:986-987. doi:10.1007/978-0-387-30164-8_832
14. COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model | IEEE Journals & Magazine | IEEE Xplore. Accessed March 28, 2023. https://ieeexplore.ieee.org/abstract/document/9151169
15. Wang YX, Zhang YJ. Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Trans Knowl Data Eng*. 2013;25(6):1336-1353. doi:10.1109/TKDE.2012.51
16. Meaney C, Escobar M, Moineddin R, et al. Non-negative matrix factorization temporal topic models and clinical text data identify COVID-19 pandemic effects on primary healthcare and community health in Toronto, Canada. *J Biomed Inform*. 2022;128:104034. doi:10.1016/j.jbi.2022.104034
17. Shah AM, Yan X, Shah SJ, Khan S. Use of Sentiment Mining and Online NMF for Topic Modeling Through the Analysis of Patients Online Unstructured Comments. In: Chen H, Fang Q, Zeng D, Wu J, eds. *Smart Health*. Lecture Notes in Computer Science. Springer International Publishing; 2018:191-203. doi:10.1007/978-3-030-03649-2_19
18. Athukorala S, Mohotti W. An effective short-text topic modelling with neighbourhood assistance-driven NMF in Twitter. *Soc Netw Anal Min*. 2022;12(1):89. doi:10.1007/s13278-022-00898-5
19. DistilBERT. Accessed March 28, 2023. https://huggingface.co/docs/transformers/model_doc/distilbert
20. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. Published October 11, 2018. Accessed March 28, 2023. https://arxiv.org/abs/1810.04805v2
21. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv.org. Published October 2, 2019. Accessed March 28, 2023. https://arxiv.org/abs/1910.01108v4
22. Rietveld R, van Dolen W, Mazloom M, Worring M. What You Feel, Is What You Like Influence of Message Appeals on Customer Engagement on Instagram. *J Interact Mark*. 2020;49:20-53. doi:10.1016/j.intmar.2019.06.003
23. Full article: Factors Driving Social Media Engagement on Instagram: Evidence from an Emerging Market. Accessed March 28, 2023. https://www.tandfonline.com/doi/full/10.1080/08911762.2021.1956665?casa_token=EOQg4VsgY4EAAAAA%3AA8C3-Q618U5GTvlZHyLHokiE01jSvkZpPjBvkQPBkrbulf48YCjIJPavX_JVt0B0op36_E2h3UXi
24. Full article: Factors Influencing Engagement in Fashion Brands' Instagram Posts. Accessed March 28, 2023. https://www.tandfonline.com/doi/full/10.1080/17569370.2021.1938820?casa_token=goC0XAj93LAAAAAA%3AmvV3dbR9N-gUqbQfXzvji_Vz0UwFA5HURkHz3MN-66YkFvB0wFLnamPQ6GReVJafge3G70fpt6Or