IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 13, 2019, accepted June 25, 2019, date of publication July 2, 2019, date of current version July 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2925965

Characteristics of a Highly Cited Article: A Machine Learning Perspective

MOHAMED ELGENDI, (Senior Member, IEEE)
Faculty of Medicine, The University of British Columbia, Vancouver, BC V6T 1Z3, Canada
BC Children's & Women's Hospital, Vancouver, BC V6H 3N1, Canada
School of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada
e-mail: moe.elgendi@gmail.com

ABSTRACT Machine learning (ML) is a fast-growing topic that enables the extraction of patterns from varying types of datasets, ranging from medical data to financial data. However, the application of the ML methodology to understand the key characteristics of highly cited research articles has not been thoroughly investigated, despite the potential practical guidance that ML can provide for researchers during the publication process. To address this research gap, an ML algorithm known as principal component (PC) analysis is used to detect patterns in highly and lowly cited papers. In this paper, eight features (number of citations, number of views, number of characters with no spaces, number of figures, number of tables, number of equations, number of authors, and title length) are extracted from highly and lowly cited papers, leading to eight PCs (PC1–PC8). PC1 shows that the numbers of citations are positively correlated with the character count and negatively correlated with the title length. PC2 shows that the number of tables is positively correlated with the title length. PC3 shows that the number of figures is positively correlated with the number of tables. PC4–PC8 rank the importance of individual features in the descending order: number of equations, number of characters with no spaces, number of figures, number of views, and then the number of authors. The results of the ML analysis provide interesting and valuable tips for researchers, students, and all academic and non-academic writers who are seeking to improve their citation rates.

INDEX TERMS Natural language processing, text mining, artificial intelligence, scientific writing, citation analysis, bibliometrics.

I. INTRODUCTION

Thousands of academics are published at least once every five days [1]. While some of these publications attract attention and receive a high number of citations, others are scarcely cited, if at all. In 2003, Aksnes [2] investigated which publication features produce highly cited papers. However, the study focused only on authorship. In 2010, Habibzadeh and Yadollahie [3] investigated the impact of title length on citation rates and found that longer titles were associated with higher citation rates. However, their study examined only 22 articles and did not focus on highly cited papers, which may have limiting the conclusions that could be drawn. In 2011, Jamali and Nikzad [4] performed a similar analysis of 2,172 articles and reported that articles with longer titles were downloaded slightly less often than articles with shorter titles, contradicting the results of Habibzadeh and Yadollahie.

The associate editor coordinating the review of this manuscript and approving it for publication was Shirui Pan.

In 2014, Subotic and Mukherjee [5] found that title amusement level (i.e., the catchiness or playfulness of the title) was slightly correlated with more downloads and citations and that more amusing titles tended to be shorter. In 2015, Letchford *et al.* [6] found that papers with shorter titles may be easier to understand and hence attract more citations. Interestingly, none of the abovementioned articles quantified the number of words in a ''short'' or ''long'' title. However, the results do reveal that relatively comparable studies have achieved fairly inconsistent findings, making it somewhat difficult to extract any meaningful suggestions for naming articles.

To the best of the author's knowledge, no analysis has been conducted to determine the effectiveness of different publication features (number of views, number of characters with no spaces, number of figures, number of tables, number of equations, and number of authors, title length) simultaneously. With the advances in machine learning (ML), we can apply existing methodologies to uncover hidden relationships

between all publication features and investigate the importance of features and relevant intercorrelations.

II. MATERIAL AND METHODS

To ensure a generic and unbiased conclusion, articles were included in the analysis based only on citations, without consideration of the article type; the journal's impact factor, chronology, or discipline; or any other categorization.

A. DATASET FOR ARTICLE FEATURE ANALYSIS

The Multidisciplinary Digital Publishing Institute (MDPI) website provides access to highly and lowly cited publications in a given year. We used this website to identify 200 papers published in 2017 from all journals and in all subjects for inclusion in this study. Half of the article articles are highly cited, while the other half are lowly cited papers. The data can be downloaded from this link: http://dx.doi.org/10.21227/5493-9a35.

B. TITLE LENGTH DATASETS

In order to develop a solid conclusion with quantifiable values regarding title length, the top 100 citations from three indexes of citation databases were analyzed in addition to the MDPI dataset. The databases are described below:

- **Google Scholar**: This database [7] is publicly available. Approximately two-thirds of the database are books from all disciplines, and the database contains various cited papers. Only the 100 research publications with the most citations were used for this analysis. The number of citations for the 100 articles in this database ranges from 30,948 to 223,131.
- **Web of Science**: This database [7] is publicly available and contains various cited papers. Only the 100 research publications with the most citations were used for this analysis. The database covers all of Thomson Reuter's Web of Science and includes works on the social sciences and arts and humanities, conference proceedings, and some books published since 1900. The number of citations for the 100 articles in this database ranges from 12,119 to 305,148.
- **Altmetric**: This database is publicly available (https://www.altmetric.com/top100/2018/) and contains the 2018 Altmetric Top 100, which is an annual list of the research that most captured the public's attention last year. Only the 100 research publications with the most citations were used for this analysis. This list includes research published from 2013 to 2018. Individuals from various industries, publishers, and academic institutions are recently looking at the Altmetric score more with interest, which is why it is included in this study.

C. METHODS

Principal component analysis (PCA) is one of the most popular techniques in artificial intelligence for dimensionality and reduction and for finding correlations between features without any prior knowledge. Understanding the relationship between features presents several challenges, but PCA mitigates them. PCA is important because it does not involve training or labeling, but automatically finds relationships. The overall PCA analysis was carried out as follows:

1) Combine highly and lowly cited papers into one matrix, X, with dimension (200 articles $\times$ 8 features).
2) Normalize using the Min-Max scaling, $Z = (X - X_{\min})/(X_{\max} - X_{\min})$, and ensure that Z has zero mean.
3) Obtain the covariance matrix $C = \frac{1}{8}\sum_{f=1}^{8} ZZ^T$, where f is the number of article features.
4) Perform eigen decomposition of C and compute the e_f eigenvalues and their corresponding eigenvectors, v_f, to satisfy the equation $Cv = ev$.
5) Sort the eigenvalues in descending order, $e_1 \geq e_2 \geq \ldots \geq e_8$, and match them to corresponding eigenvectors.
6) The eigenvector with the largest eigenvalue will be called the first principle component (PC1) while the last PC associated with the lowest eigenvalue will be called PC8.

Eight features are extracted from the MDPI database: number of citations, number of views, number of characters with no spaces, number of figures, number of tables, number of equations, number of authors, and title length. The features extracted from highly and lowly cited papers were compared using the t-test (p_t) and the Wilcoxon–Mann–Whitney test (p_w) for two independent groups. A p value of < 0.05 was considered significant. Pearson's correlation coefficient was used to calculate the correlation between features and between the features and PCs. We used Matlab 2018b software and Python 3.6.5 (Anaconda, Inc., default March 29, 2018, 13:32:41) [MSC v.1900 64 bit (AMD64)] on win32 software to analyze the data and Wordle to generate word clouds.

III. RESULTS

The MDPI database, one of the leading open-access databases of multidisciplinary journals, was used to download highly and lowly cited papers. The MDPI website allows one to view articles and rank them based on their citation rates. After ranking all articles published in 2017, the top 100 highly cited and 100 lowly cited papers were manually downloaded. Table 1 statistically compares the highly and lowly cited papers. The overall visual representation of Table 1 is shown in Figure 1.

As expected, the number of citations for highly and lowly cited papers is significantly different because the data was already classified into two categories based on the number of citations. The number of views and number of characters (no spaces) were also both statistically significant ($p_t < 0.001$ and $p_w < 0.001$). The number of figures slightly differs between highly and lowly cited papers ($p_w = 0.058$, which is not significant but on the boundary). The number of tables is statistically significant ($p_t = 0.002$ and $p_w = 0.006$). The number of equations is not different, but the title length and

TABLE 1. Statistical results regarding eight features of highly and lowly cited papers.

	Highly Cited Papers			Lowly Cited papers			w-test	t-test
	μ	σ	$\tilde{X}$	μ	σ	$\tilde{X}$	p_t-value	p_w-value
Citations	50.150	14.429	46.000	0.000	0.000	0.000	<0.0001	<0.0001
Views	4627.740	5712.518	3836.000	1348.230	800.880	1164.000	<0.0001	<0.0001
Characters	73,108	41,120	58,000	34,804	17.556	33,600	<0.0001	<0.0001
Figures	7.200	7.900	6.000	5.390	4.559	5.000	0.106	0.058
Tables	2.780	3.469	2.000	1.620	2.286	1.000	0.002	0.006
Equations	4.850	15.800	0.000	5.630	15.589	0.000	0.556	0.734
Authors	6.110	11.704	4.000	3.690	2.525	3.000	0.006	0.049
Title Length	10.990	4.140	10.000	13.020	4.853	13.000	0.001	0.003

[a]Here, μ = average, σ = standard deviation, $\tilde{X}$ = median, w-test = Wilcoxon rank-sum test, t-test = Student's t test, p_t value is the statistical hypothesis testing for the t-test while the p_w for the Wilcoxon rank-sum test.

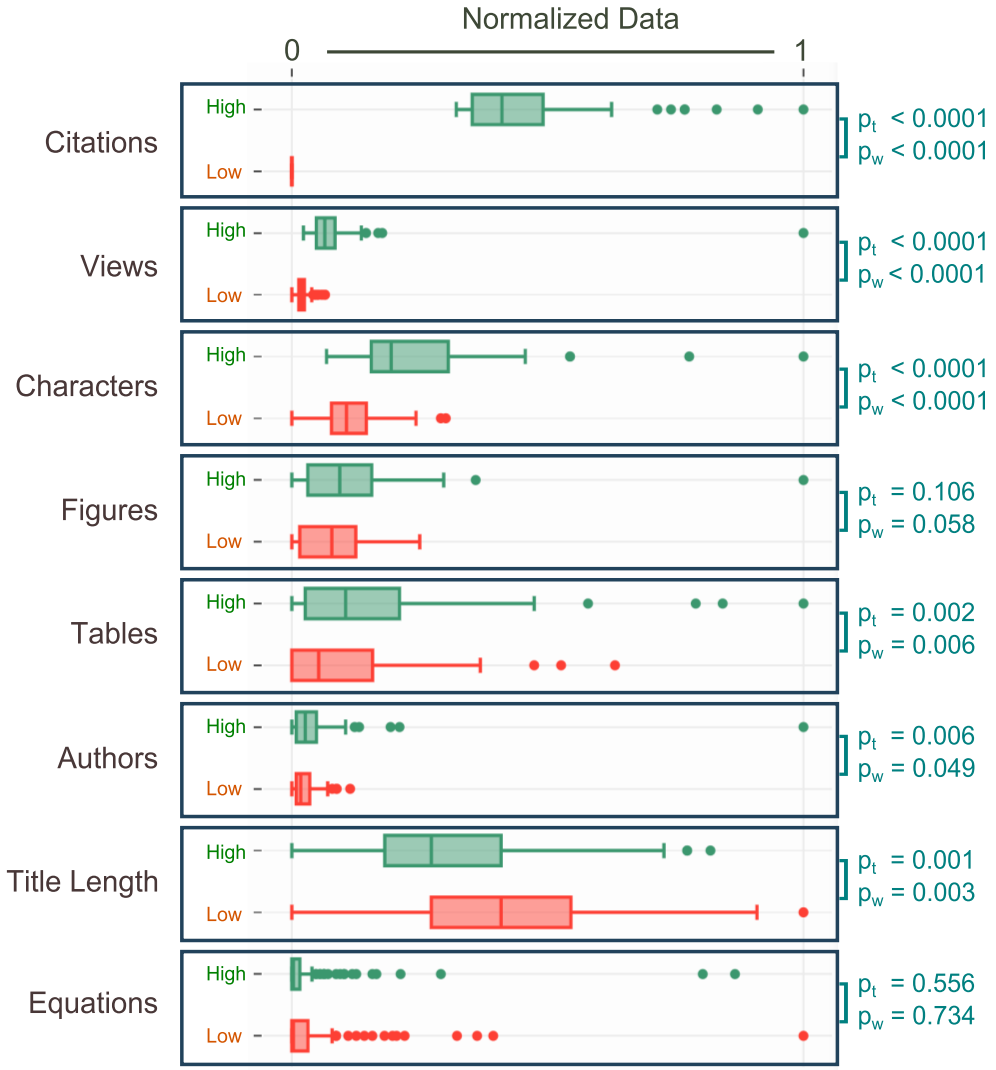

FIGURE 1. Statistical analysis of features of highly and lowly cited papers. Here, p_t value is the statistical hypothesis testing for the t-test while the p_w for the Wilcoxon rank-sum test.

number of authors are significantly different between highly and cited papers.

Figure 2 shows the eight PCs of the MDPI database. PC1 explains the most variance (40%), reflecting its relevance and importance. PC2 explains nearly 23%, while PC3 explains a little more than 14%. As expected, PC1 is the most volatile upon visual inspection, PC2 is the second most volatile, and PC8 is the least volatile.

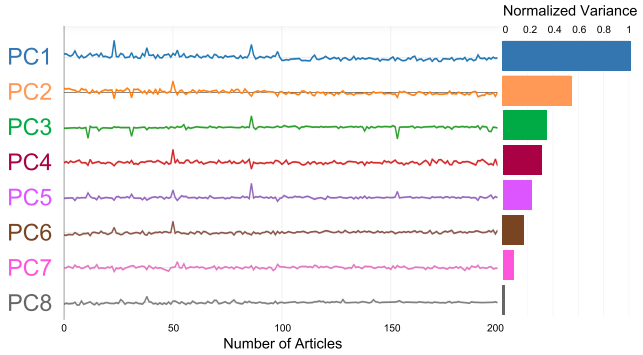

FIGURE 2. PCA of all article features extracted from the MDPI database. PC1 has the highest variance, while PC8 has the lowest variance.

Figure 3 shows a heat map of the correlation matrix between all PCs and features. The diagonal entries are all equal to 1. As can be seen, the heat map consists of four 8×8 blocks. The top right block is the correlation matrix for the principal components. We know that they principal components are orthogonal. As expected, this block contains zeros (numerically negligible values) in its off-diagonal entries, reassuring us that the principal components are mutually orthogonal (and therefore uncorrelated). The bottom left block is the heat map for the article features. We see that there is a slight correlation ($r = 0.51$) between characters and citations. However, the other features did not show any correlation. Also, the bottom left block is not informative.

The most interesting blocks of the heat map are the top left and the bottom right 8×8 blocks. Since the heat map is symmetrical, they are actually mirror images of each other. If we focus on the bottom right 8×8 block, which is demarcated with dashed black lines, within this block, the rows correspond to the article features, while the columns correspond to the principal components. The first column of the block is almost entirely dark red, signifying that PC1 is correlated

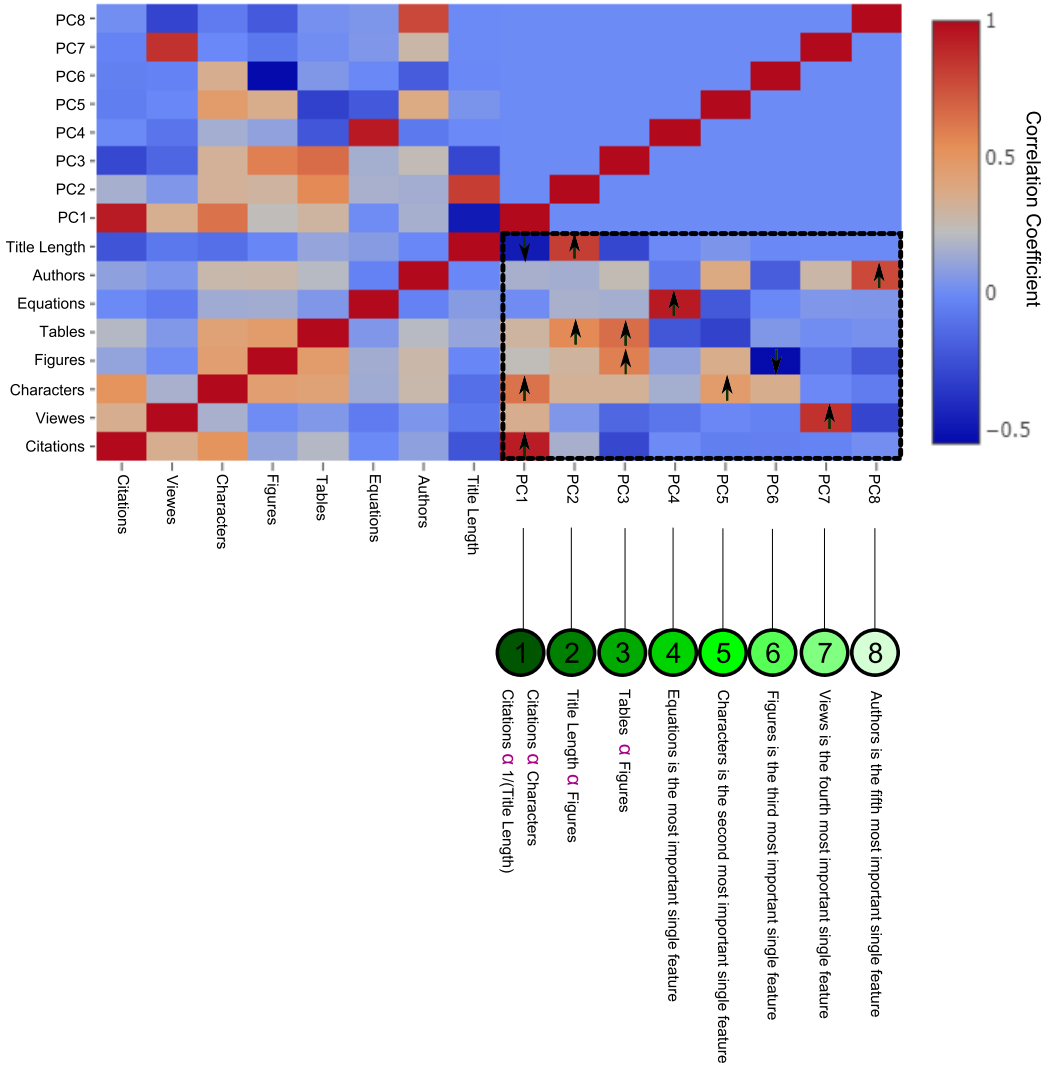

FIGURE 3. Heat map of the correlations between article features and PCs found within the MDPI database. The α refers to the hidden correlation. The black arrow indicates an absolute value of correlation coefficient greater than 0.5 is achieved.

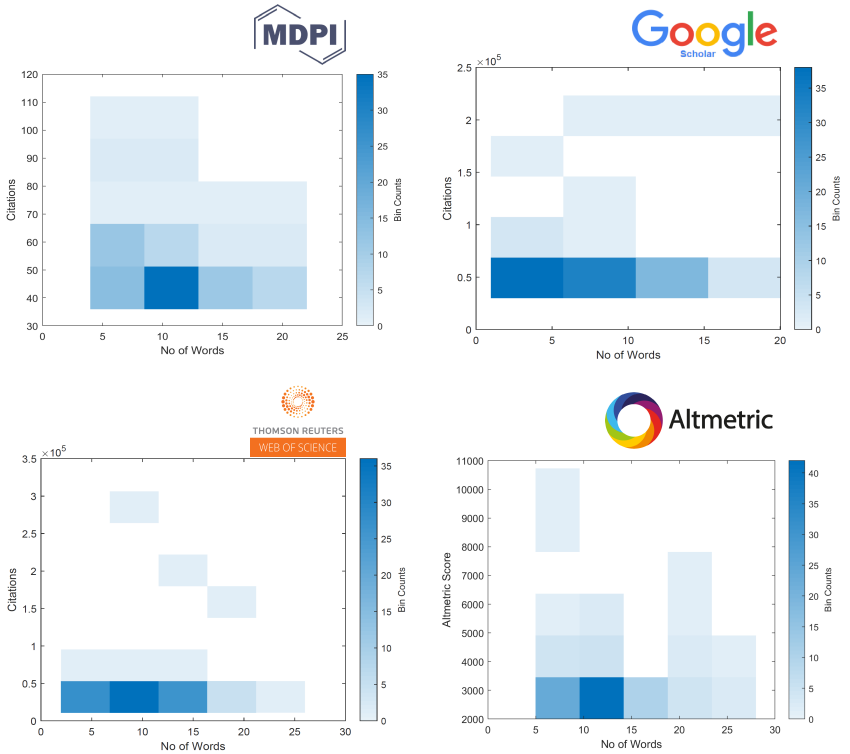

FIGURE 4. Binned scatterplot of the number of words in the titles of highly cited papers within four different datasets.

with the number of characters (no spaces), number of figures, and title length. Another interpretation of PC1 is that the number of citations is positively correlated with the number of characters (no spaces) and negatively correlated with the title length. Given that PC1 is the PC with highest eigenvalue, there is a hidden correlation between the number of citations and the number of characters with no spaces that moves in one direction, and title length moves in the opposite direction. The second column shows that the number of tables and title length are moving in the same direction, and both are strongly correlated with PC2. This suggests that there is a hidden correlation between the number of tables and title length. The third column, in the dashed box, shows that the number of tables and number of figures move in the same direction, and both are strongly correlated with PC3, suggesting that there is a hidden correlation between the number of tables and number of figures. PC4 to PC8 were unable to capture any correlation between article features, but they do rank the article features independently in descending order: number of equations, number of characters with no spaces, number of figures, number of views, and then number of authors.

Here, we investigate the number of words in the titles of scientific paper and the correlation of this feature with the number of citations for each paper, over all databases. For the MDPI database, the mean the number of words is $\mu = 11$, with a range from 4 to 22. For the Google Scholar database, the mean is $\mu = 7$, with a range of 1 to 20. For the Web of Science database, it is $\mu = 10$, with a range of 2 to 26. For the Altmetric database, it is $\mu = 12$, with a range of 5 to 28. Scattering the data may create confusion, and thus a binned scatterplot is used. The data space is partitioned into rectangular bins and displays, and the number of data points in each bin is indicated by different colors, as shown in Figure 4. By visually inspecting Figure 4, the optimal number of words in an article title is 10 ± 3 words according to the MDPI, Google Scholar, Web of Science, and Altmetric databases.

Figure 5 attempts to find similarity between different databases or common words that are used in highly cited papers. It shows that highly cited papers have some common words reflecting the topic of interest and focus of the papers. Below is a list the top five words in each database, ranked in descending order from left to right:

- **MDPI**:
  Review, cancer, monitoring, recent, and therapeutic.
- **Google Scholar**:
  Method, theory, analysis, applications, and learning.
- **Web of Science**:
  Method, protein, DNA, multiple, and new.
- **Altmetric**:
  Association, analysis, cancer, health, and study.

Table 2 shows that question marks were not used at all in highly cited papers and was used in lowly cited papers, but there was no significant difference ($p_t = 0.322$ and $p_w = 0.320$). Forward slashes were used sometimes in highly cited papers, and there was no significant difference with lowly cited papers ($p_t = 0.713$ and $p_w = 0.829$). Interestingly, dashes and dots were more commonly used in

FIGURE 5. Which words can be used to generate more impactful titles?

TABLE 2. Statistical results for symbol in the titles of highly and lowly cited papers.

	Highly Cited Papers			Lowly Cited papers			w-test	t-test
	μ	σ	$\tilde{X}$	μ	p_t-value	p_w-value	p_t-value	p_w-value
Question mark (?)	0.000	0.000	0.000	0.010	0.100	0.000	0.322	0.320
Forward slash (/)	0.060	0.343	0.000	0.050	0.297	0.000	0.713	0.829
Dash (-)	0.440	0.770	0.000	0.930	1.622	0.000	0.024	0.006
Colon (:)	0.430	0.498	0.000	0.270	0.468	0.000	0.014	0.038
Dot (.)	0.000	0.000	0.000	0.260	0.991	0.000	0.001	0.010

[a]Here, μ = average, σ = standard deviation, $\tilde{X}$ = median, w-test = Wilcoxon rank-sum test, t-test = Student's t test, p_t value is the statistical hypothesis testing for the t-test while the p_w for the Wilcoxon rank-sum test.

lowly cited papers, with a significant difference from highly cited papers ($p_t = 0.024$ and $p_w = 0.006$ for dashes; $p_t = 0.014$ and $p_w = 0.038$ for dots). Colons were used significantly more in highly cited papers ($p_t = 0.001$ and $p_w = 0.01$).

IV. DISCUSSION

A. DATABASE JUSTIFICATION

Google Scholar does not yet allow one to rank articles based on citation rate; perhaps this option will be available in the future. Moreover, the Web of Science does not provide access to articles based on number of views, number of characters, number of figures, number of tables, and title length. In addition to a search engine that allows one to search for articles based on citation rates, there is a need to identify open-access and non-open-access articles, as a recent study [8] showed that open-access articles receive more citations. To ensure fair and consistent analysis, there is a need for a dataset that contains highly cited and lowly cited papers published in the same year and provides access all of them, regardless of whether they are open-access or not. The MDPI offers a search engine that ranks all articles from 202 diverse, peer-reviewed, open-access journals. We downloaded and compared 200 articles (100 highly and 100 lowly cited articles) published in 2017. Note, MDPI provides the article ranking based on the access date. So, all downloaded articles are categorized highly cited and lowly cited based on the 17th of March 2019 access.

One of the important reasons to choose the MDPI database is the MDPI does not have any restriction over the title length, number of characters with no spaces, number of equations, and number of tables. This will make the comparison between papers published in 22 different disciplines more reliable and consistent, which is not the case when we compare articles published in journals with restrictions over format.

Altmetric is a relatively new metric that shows what both experts and non-experts are saying about published research output in mainstream media, policy documents, social networks, blogs, and other scholarly and non-scholarly forums. It includes citations on Wikipedia and in public policy documents, discussions on research blogs, mainstream media coverage, bookmarks on reference managers like Mendeley, and mentions on social networks such as Twitter and Facebook, which are not included in Google Scholar and the Web of Science. Recently, publishers, institutions, researchers, and funders have shown serious interest in Altmetric for its inclusion of social media. The Altmetric scores range from 2,001 to 10,724. Note that Altmetric does not allow all articles to be ranked based on their Altmetric score, although this option may be available in the future. However, Altmetric shows a unique way of including citations on social media, and therefore it will be used along with the Google Scholar and Web of Science databases to validate the title length results of the MDPI database analysis.

B. CHARACTER COUNT

Figure 1 shows that the number of characters, with no spaces, is significantly ($p_t < 0.0001$ and $p_w < 0.0001$) different between highly and lowly cited papers. Moreover, PC1, as shown in Figure 3, showed a strong correlation between the number of characters and number of citations. Interestingly, PC5 showed that the number of characters is the second most important feature. These results suggest that it is better to focus on writing longer papers, not shorter papers, to attract more citations.

Table 1 shows that the number of characters needs to be more than $33,600$, including references, which is approximately $5,600$ words. This result is in agreement with the number of words accepted in one of the highly impactful journals such as Nature (the 2017 impact factor = 41.5). According to Nature's most recent format (https://www.nature.com/nature/for-authors/formatting-guide), the maximum number of words accepted to publish a research article in Nature is $6,500$ words, including references. Note that Google's metrics (h5-index and h5-median) ranked Nature in 2018 as the most impactful journal in the world (https://scholar.google.com/citations?viewop=topvenueshl=en).

C. NUMBER OF AUTHORS

The results of this article show a correlation between the number of citations and the number of authors. Moreover, there is a significant difference ($p_t = 0.006$ and $p_w = 0.049$) in the number of authors between highly and lowly cited papers. PC8 showed that the number of author is an important feature, ranking it as the fifth aspect to think about during the writing process. It seems that multi-author papers gain greater exposure from the authors' institutions, labs, researchers, and students compared to single-author papers. In other words, each author has his own network, and bringing together all authors' networks will increase the number of readers who share the same research interests, which will in turn increase the likelihood of citations. Moreover, multi-author papers can benefit from self-citations [2]. The correlation between the number of citations and the number of authors may be affected by the quality of the research, as found by Lawani [9]. One may intuitively assume that when forces are joined and more than one person contributes to the work, the quality of the methodology, performance of the experiment, acquisition of funding, and quality of the paper will improve. So, it can be expected that multi-author papers will be cited more often than single-author papers.

D. NUMBER OF FIGURES

To the knowledge of the author, the number of figures has not been investigated in the literature. Figure 1 shows that the number of figures is slightly different ($p_w = 0.058$, which is not significant but indicates a trend) between highly and lowly cited papers. Moreover, PC3, as shown in Figure 3, showed a strong correlation between the number of figures and number of tables. Interestingly, PC6 showed that the number of figures is third most important feature. These results suggest that the more figures in a paper, the more information are conveyed to readers to help them understand the results. In open access journals there is no limitation on the number of figures; however, some other journals require a certain number of figures. In this case, combining multiple figures into one figure can be a good idea.

Table 1 shows that the median number of figures for highly cited papers is 6 while the average number of figures for highly cited papers is 7.2. This result suggests that at least 6 figures are needed to reflect relevance and impact, which is in agreement with the number of figures accepted by Nature. According to Nature's most recent format (https://www.nature.com/nature/for-authors/formatting-guide), the maximum number of display items (figures or tables) is 5 or 6.

E. NUMBER OF TABLES

To the knowledge of the author, the number of tables has not been investigated in the literature. Figure 1 shows that the number of tables is significantly different ($p_t = 0.002$ and $p_w = 0.006$) between highly and lowly cited papers. Two PCs (PC2 and PC3), as shown in Figure 3, reflected the importance of the number of tables. PC2 showed that there is a strong correlation between the number of tables and the title length. PC3 showed that the number of tables is correlated with the number of figures. In other words, the more tables exist in a paper, the more the information is conveyed as long as we avoid redundant representations of findings.

Table 1 shows that the median number of tables for highly cited papers is 2 while the average number of tables for highly cited papers is 3.4. This result suggests that at least 2 tables are needed to represent the analysis. Please note that the number of tables investigated here are independent from the number of figures. As mentioned, some journals such as Nature allows only a combination of 5 or 6 figures and tables.

F. NUMBER OF VIEWS

Figure 1 shows that the number of views is significantly different ($p_t < 0.001$ and $p_w < 0.001$) between highly and lowly cited papers. Moreover, PC2, as shown in Figure 3, showed a strong correlation between the number of views and the number of citations. Interestingly, PC6 showed that the number of views is the most important single feature. These results suggest that the larger the audience, the greater the likelihood of getting more citations. Therefore, it is our duty to let everyone–specialists and non–specialists–know about our recent publications. Any effort to raise awareness about recently published articles can contribute to an increased citation rate. In particular, use of all social media outlets, including Facebook, LinkedIn, Twitter, and ResearchGate, can be used to attract a large number of views on recently published articles.

G. NUMBER OF EQUATIONS

To the knowledge of the author, the number of equations has not been investigated in the literature. Figure 1 shows that the number of equations is not significantly different between highly and lowly cited papers ($p_t = 0.556$ and $p_w = 0.73$). Perhaps this is related to the fact that reviews are usually more commonly cited than articles that contain equations. PC4, as shown in Figure 3, showed that the number of equations is the most important feature. Moreover, Table 1 shows that the median number of equations for highly cited papers equals the median number of lowly cited papers. So including as many equation as needed is highly recommended, even though we did not find a significant difference between highly and lowly cited papers.

H. TITLE LENGTH

A study [3] including 22 journals found that studies with longer titles had more citations. However, there were two biases in their analysis: 1) all journals are from the medical discipline and 2) there was a statistically significant difference in the impact factors between the 22 journals used in their study (8 journals had an impact factor of > 10, and 14 had an impact factor of < 10, with a range of 0.35 to 50.02). Another study [10] found that title length was correlated to the number of citations. Again, the study was biased because 1) the sample size was small (25 highly and 25 lowly cited articles), 2) they considered only three medical journals, and 3) the impact factor of the three journals varied from 2.8 to 53. Interestingly, one study [11] reported that title length is not associated with the number of citations. However, there is a general consensus [4]–[6] that the title of articles is correlated with citation rate, and the majority of articles have recommended short titles. The optimal title is short, informative, and attractive, which is not an easy task, as Kane [12] mentioned. The title opens the door to readers, attracting their attention and giving them a brief overview of the content. Linguistically, the quality of the title plays a major role in attracting readers, as reported by Wang and Bai [13].

Choosing an article's title is a critical step in the writing process, and, arguably, it significantly impacts the likelihood that the presented research and results will be exposed to the intended audience. Put simply, if the title does not effectively capture the attention of readers and those searching databases with proper word choice, usage, and length, the likelihood that the paper will be opened and read is decreased, which consequently decreases the chance that the article will be highly cited. The title determines whether prospective readers readily identify with the topic area, discipline, and subject, and it helps with indexing and archiving the paper, which in turn impacts the search efforts of prospective readers and eventual citation. For example, Nadri *et al.* [14], found that articles in the statistics and biostatics category dominated the top 100 cited list in medicine. Recently, the importance of the topic is discussed thoroughly by Chen *et al.* [15] and Yi *et al.* [16].

One of the interesting factors that can impact citations is the interactivity between highly cited authors. Ding [17] investigated this area and found that highly cited authors do not coauthor with each other, but closely cite each other. This hidden dynamics between highly cited authors can influence certain research topic and increase citation to specific papers cited by them.

Journals, purpose-based libraries, and scientific search engines also partially rely on titles for search requests and proposing related articles [18]. The art of marketing and ''click bait'' (i.e., efforts on social media to cause audiences to click on the content) has been studied and applied with precision by various companies and organizations, which understand the significance of capturing a reader's/consumer's attention with effective wording. The same reasoning can be applied to writing research article titles, as there are many journals and publishers competing to push out similar work in large and competitive fields. The most highly cited paper according to each database is as follows:

- **MDPI**:
  'Liposomal Formulations in Clinical Use: An Updated Review' (word count = 8)
- **Google Scholar**:
  'Cleavage of Structural Proteins During the Assembly of the Head of Bacteriophage T4' (word count = 13)
- **Web of Science**:
  'Protein Measurement with the Folin Phenol Reagent' (word count = 7)
- **Altmetric**:
  'Mortality in Puerto Rico after Hurricane Maria' (word count = 7)

One observation regarding these highly cited papers is that the titles' word choice was not necessarily simple or accessible to a broad audience. This raises a question: Is the advice that researchers ought to create simplified titles to make papers more attractive credible? Based on this preliminary analysis, the advice is not necessarily true. In fact, it seems that there is no need to simplify the title to make the paper more attractive to the masses. However, the high

citation rates of papers within the identified word count range (10 ± 3 words) may be due to humans' short-term memory. According to Miller [19], a cognitive psychologist, the quantity of items that can be stored at any given time within a person's short-term memory is capped at a certain value. Miller's paper discussing short-term memory and the number of items that can be memorized, "The Magical Number Seven, Plus or Minus Two," remains one of the most highly cited papers in the history of psychology. The title of Miller's paper, which includes 8 words, aligns nicely with the results reported in this analysis.

Figure 4 shows other words that could be used in a title. One can see that it is relatively challenging to strike the appropriate balance between the number of words and proper word choice. In addition, research has found that it is difficult to generate a title that is impactful and attractive at the same time [12]. Researchers have spent their fair share of hours, days, or even weeks mulling over a title choice, often leading them to consult trusted colleagues and friends for their opinions. Alternatively, there are some academics who may not feel it is of the utmost importance to develop the perfect title, perhaps due to their excitement to publish results or pending timelines and competing workloads. Nonetheless, improperly representing research efforts through a poorly crated title may diminish the impact of any published results, and thus it is worth the extra time and effort, according to the analysis presented here.

Also, word choice may play an important role in titles' impact and effect. The words used in highly cited publications in the four discussed databases are shown in Figure 5. Using the following words in article titles may increase readership and, consequently, the number of citations: review, study, cancer, recent, new, association, analysis, method, theory, monitoring, therapeutic, applications, learning, protein, DNA, multiple, and health.

There are five main symbols used in titles: the question mark (?), forward slash (/), dash (-), colon (:), and dot (.). The slash, dash, and dot symbols have not been investigated in relation to citation rate in the literature. However, the colon has been investigated the most often, although the results were controversial. For example, one study [10] showed that the colon is associated with citations, and another study [4] showed the opposite. The latter study [4] also investigated the question mark in addition to the colon and found that the question mark is associated with fewer citations.

The results of this study showed that the question mark and forward slash did not have a significant impact on highly and lowly cited articles. However, the colon, dash, and dot showed significance. In particular, the colon is significantly used in highly cited papers, while the dot and dash are used significantly in lowly cited papers.

The ML findings shown in Figure 6 reveal six tips for writing a high-impact research article with an increased likelihood of receiving more citations:

1) choose a title that is 10 ± 3 words in length,
2) include 6 or more authors,
3) include 33, 600 characters (no spaces) minimum,
4) include 6 figures minimum, each figure can contain a number of smaller figures (e.g., Fig1a, Fig 1b, Fig 1c etc.), this is especially useful in cases where the targeted journal has placed restrictions on the number of figures,
5) include 2 tables minimum, and
6) use as many equations as necessary, as adding or reducing the number of equations will not have an impact on the readability or citation rate of the article.

FIGURE 6. An infographic summarizing how to write a paper that increases the likelihood of high citations. Applying these recommendations does not guarantee increased citation rates; rather, they are considerations when shaping a scientific paper. Of course, there are more essential features that improve citations and overall impact such as quality, importance, originality of the work, etc. Note that the number of recommendations are based on the MDPI paper submission format, which is a single column with no limit to the title length, number of characters with no spaces, number of equations, and number of tables.

Applying these recommendations does not guarantee increased citation rates. Of course, there are more important features that improve citation rates and overall impact, which can be called "non-formal features": publication in reputable journals such as Nature and Science, the reputation of the author(s), the originality and importance of the scientific content of the paper, the interest of the topic, the rigorousness and clarity of the results, the novelty of the results, the discipline of the journal, accessibility (i.e., open access vs. non-open access), the publication type (e.g., article, review, communication, etc), age of the references used, the quality of the reviews, and the quality of the editorial feedback.

During the literature review for this paper, no other related articles were found that could be included for comparison. To my knowledge there is no ML algorithm (or a mathematical model) applied to the same research question or

the same database. However, related controversial results found in the literature are discussed in the paper.

This paper tried to answer some of the questions raised on how to write and present research well [20]. Only ''formal features'', as defined in the paper as the eight features, were investigated using a multidisciplinary open access database. Interestingly, past related research in this area have combined open and non-open access article which can create biased results. Thus, findings in this paper are based on a fair analysis because all articles were open access, from various disciplines, and from the same publisher.

V. CONCLUSION

In this paper, the impact of publication features (number of views, number of characters with no spaces, number of figures, number of tables, number of equations, number of authors, and title length) on the citation rate of articles was examined. Without prior training or knowledge, an ML method called PCA was able to detect a complex dynamic between the publication features. There is a significant positive correlation between the number of citations and the number of views, tables, and authors. Moreover, the number of citations is significantly negatively correlated with title length. One interesting result is that the number of equations is not correlated with the number of citations. These new findings are timely as, now more than ever, scientific papers are shared on a large scale across different media outlets. The results can be used as recommendations for producing article with a high chance of being cited, given quality content. Using this newfound knowledge with artificial intelligence and other significant elements of publishing, we can better arm ourselves as researchers and knowledge translators to ensure we are communicating our results and messages across to a broader audience while increasing our citation rates.

ACKNOWLEDGMENT

The author is grateful for the support from Mining for Miracles, BC Children's Hospital Foundation, Vancouver, British Columbia, Canada.

REFERENCES

[1] J. P. Ioannidis, R. Klavans, and K. W. Boyack, ''Thousands of scientists publish a paper every five days,'' *Nature*, vol. 561, pp. 167–169, Sep. 2018.
[2] D. W. Aksnes, ''Characteristics of highly cited papers,'' *Res. Eval.*, vol. 12, pp. 159–170, Dec. 2003.
[3] F. Habibzadeh and M. Yadollahie, ''Are shorter article titles more attractive for citations? Crosssectional study of 22 scientific journals,'' *Croatian Med. J.*, vol. 51, pp. 165–170, Apr. 2010.
[4] H. R. Jamali and M. Nikzad, ''Article title type and its relation with the number of downloads and citations,'' *Scientometrics*, vol. 88, no. 2, pp. 653–661, 2011.
[5] S. Subotic and B. Mukherjee, ''Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles,'' *J. Inf. Sci.*, vol. 40, pp. 115–124, Nov. 2013.
[6] A. Letchford, H. S. Moat, and T. Preis, ''The advantage of short paper titles,'' *Roy. Soc. Open Sci.*, vol. 2, Aug. 2015, Art. no. 150266.
[7] R. Van Noorden, B. Maher, and R. Nuzzo, ''The top 100 papers,'' *Nature*, vol. 514, p. 550, Oct. 2014.
[8] J. C. Clements, ''Open access articles receive more citations in hybrid marine ecology journals,'' *Facets*, vol. 2, pp. 1–14, Jan. 2017.
[9] S. M. Lawani, ''Some bibliometric correlates of quality in scientific research,'' *Scientometrics*, vol. 9, pp. 13–25, Jan. 1986.
[10] T. S. Jacques and N. J. Sebire, ''The impact of article titles on citation hits: An analysis of general and specialist medical journals,'' *JRSM Short Rep.*, vol. 1, pp. 1–5, Jun. 2010.
[11] F. Rostami, A. Mohammadpoorasl, and M. Hajizadeh, ''The effect of characteristics of title on citation rates of articles,'' *Scientometrics*, vol. 98, pp. 2007–2010, Mar. 2014.
[12] T. S. Kane, *The Oxford Essential Guide to Writing*. New York, NY, USA: Oxford Univ. Press, 2003.
[13] Y. Wang and Y. Bai, ''A corpus-based syntactic study of medical research article titles,'' *System*, vol. 35, pp. 388–399, Sep. 2007.
[14] H. Nadri, B. Rahimi, T. Timpka, and S. Sedghi, ''The top 100 articles in the medical informatics: A bibliometric analysis,'' *J. Med. Syst.*, vol. 41, no. 150, Oct. 2017.
[15] H. Chen, X. Wang, S. Pan, and F. Xiong, ''Identify topic relations in scientific literature using topic modeling,'' *IEEE Trans. Eng. Manag.*, to be published.
[16] Z. Yi, C. Hongshu, L. Jie, and Z. Guangquan, ''Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016,'' *Knowl.-Based Syst.*, vol. 133, pp. 255–268, Oct. 2017.
[17] Y. Ding, ''Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks,'' *J. Informetrics*, vol. 5, pp. 187–203, Jan. 2011.
[18] V. Soler, ''Writing titles in science: An exploratory study,'' *English Specific Purposes*, vol. 26, no. 1, pp. 90–102, 2007.
[19] G. A. Miller, ''The magical number seven, plus or minus two: Some limits on our capacity for processing information,'' *Psychol. Rev.*, vol. 63, no. 2, pp. 81–97, 1956.
[20] G. S. Patience, D. C. Boffito, and P. A. Patience, ''How do you write and present research well?'' *Can. J. Chem. Eng.*, vol. 93, pp. 1693–1696, Oct. 2015.

MOHAMED ELGENDI (M'02–SM'13) is currently a Senior Postdoctoral Fellow with the Department of Obstetrics and Gynecology and an Adjunct Professor with the Department of Electrical and Computer Engineering, The University of British Columbia (UBC). In addition to his over ten years of experience in the field of data analysis, he has received training on big data analysis and leadership in education from MIT. His experience in the areas of digital health, data analysis, and visualization includes his work at Global Health with the PRE-EMPT Initiative (funded by the Bill and Melinda Gates Foundation), the Institute for Media Innovation, Nanyang Technological University, Singapore, and the Alberta's Stollery Children's Hospital, Canada. He specializes in bridging the areas of engineering, computer science, psychology, and medicine for knowledge translation. He is also a Senior Fellow of the Howard Brain Sciences Foundation.