Understanding the link between audience engagement metrics and the perceived quality of online news using machine learning

Sotirakou, C.; Trilling, D.; Germanakos, P.; Sinis, D.A.; Mourlas, C.

DOI
10.3233/web-210457
Publication date
2021
Document Version
Final published version
Published in
Web Intelligence
License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)
Link to publication

Citation for published version (APA):
Sotirakou, C., Trilling, D., Germanakos, P., Sinis, D. A., & Mourlas, C. (2021). Understanding the link between audience engagement metrics and the perceived quality of online news using machine learning. *Web Intelligence*, *19*(1-2), 63-86. https://doi.org/10.3233/web-210457

Web Intelligence 19 (2021) 63–86
DOI 10.3233/WEB-210457
IOS Press

Understanding the link between audience engagement metrics and the perceived quality of online news using machine learning

Catherine Sotirakou [a,*], Damian Trilling [b], Panagiotis Germanakos [c], Dimitrios Alexandros Sinis [a] and Constantinos Mourlas [a]
[a] *Faculty of Communication and Media Studies, National & Kapodistrian University of Athens, Sofokleous 1, Athens, Greece*
E-mails: katerinasot@media.uoa.gr, dsinis@media.uoa.gr, mourlas@media.uoa.gr
[b] *Amsterdam School of Communication Research, University of Amsterdam, Nieuwe Achtergracht 166, Amsterdam, The Netherlands*
E-mail: d.c.trilling@uva.nl
[c] *User Experience S/4HANA, Product Engineering, Intelligent Enterprise Group, SAP SE, Walldorf, 69190, Germany*
E-mail: panagiotis.germanakos@sap.com

Abstract. This article aims to explain the perceived quality of online news articles. Discovering which elements of a news story influence readers' perceptions could drive online popularity, which is the paramount factor of digital news readership. This work explores an approach to use tree-based machine learning algorithms to address this problem based on selected characteristics, which measure engagement, drawn from prior research mostly developed by communication scientists. A proposed extended model is used to examine the association between the engagement features and perceived quality concerning all the articles depending mainly on their genre. To demonstrate the capacity of using predictive analytics to facilitate journalistic news writing the proposed methodology is applied on a novel data set with 200K articles in total constructed from a blog site. The results of phase A, indicate interesting correlations between the features and the perceived quality of the articles. In stage B, the paper seeks to extract a set of rules that can be used as guidelines for authors in the writing of their next articles, indicating the probability of popularity that their articles may gain if these rules are taken into consideration.

Keywords: Natural language processing, news, computational journalism, computer-assisted content analysis, machine learning

1. Introduction

Publishers have always wanted to know more about what type of content people consume. For years they have been relying on focus groups, ratings, paid surveys, and circulation numbers to arrive at conclusions about audience attitudes and behavior related to their products. Decisions for future articles were made by editors based on vague assumptions of what the imaginative audience would have wanted to read and journalists used to write their stories having a certain persona in their minds usually created by the managers [10]. Before the web, estimating how many

*Corresponding author. E-mail: katerinasot@media.uoa.gr.

people read a given article out of a newspaper or a magazine or what the readers thought of it was hard to assess and only possible on a small scale. Technological developments and the plethora of available data have revolutionized the marketing and advertising industry, providing publishers with monitoring tools to track key performance indicators (KPIs) and sometimes even give access to micro-level data about individual user's behavior. Such tools have become a crucial element of modern newsrooms, which, for instance, display live ratings of the performance of individual stories on screens in their offices [15].

In particular, social media play an increasing role in news dissemination, as do search engines and news aggregators [21,42]. This means that more and more people rely on secret news feed algorithms and various other ranking programs for their daily news update. Therefore, positive feedback from online readers is coveted more than ever before, since social interactions such as likes, shares, and comments can determine how many people have a chance to come across the news content on social media [41,64]. Thus, for many digital organizations, it is an important goal to gain meaningful interactions from the users, so the filtering algorithms that decide the ordering of the posts will have their content appear higher in the news feeds of potential readers.

Yet, it may be short-sighted to just look at those articles that receive a lot of engagement, "guess" what makes them so popular, and base business decisions on it. To arrive at more sustainable guidelines to base decisions on, a more systematic approach is needed. In particular, one could ask: Can the analysis of these data yield meaningful answers to critical questions like what makes one article better than the others? Many communication scientists have tried to answer such questions by investigating traffic analytics (reading time, visits, etc.) and user engagement data from social media networks, such as Facebook and Twitter. A wide variety of measures of interest (such as attention, content popularity, different sub-dimensions of engagement, etc.) have been studied, and many features and various terms have been used to explain the phenomenon of a user's preference for one particular news article to others that leads to more likes, shares, clicks and so on [3,20,28,65,68]. Alongside engagement metrics, scientists have quite accurately predicted in the past the social popularity of news stories on Twitter before their publication, focusing only on article related features, and measuring social popularity as the total number of times an article's link appeared on Twitter [4].

Other scholars focus on the prediction of social popularity after an article is released, to harness the early measurements of an article's popularity to forecast its probability of success. For instance, data scientists who work for news outlets such as the Washington Post have managed to predict online popularity after publication to provide high-quality reading experiences to their audience [30]. The scientists defined popularity as the number of page views an article receives on the first day and created a system that used a variety of factors related to contemporariness, writing quality, social media networks and so on captured during the first half-hour after publication, that accurately predicts the article's popularity. Some have explicitly linked such features to the *quality* of texts. For instance, Park et al. [50] predicted the quality of online comments. For this, they used the selections that a newsrooms' editors made by picking 'top comments' as a ground truth. Similarly, in this work, the fact that the readers of a specific blogging platform, Medium.com, can "clap" one or multiple times to signal that they perceive an article to be of high quality is used.

This paper seeks to advance the existing literature by exploring which features that have been previously linked to engagement (also referred in this paper as engagement metrics, or engagement features, or features of engagement) with news article can be successfully used to understand the perceived quality of a text, as approximated by the claps it receives. It concentrates previous efforts that have successfully utilized different sets of features [20,65,68], and taxonomies [47], and proposes a framework with new features and extended feature dimensions. More specifically, the main contribution of this work is three-fold:

(a) to propose an enhanced model of engagement metrics based on previous research;
(b) to identify how different classifications of engagement metrics contribute to the perceived quality of articles depending on the various categories of news; and
(c) to recommend a set of rules, as guidelines to authors, based on the engagement metrics of the respective articles' categories, for increasing the probability to gain popularity.

A set of 200,000 articles from Medium.com, a well-known blog platform, were used in this study, and the so-called "claps" (a form of likes) an article receives served as a proxy for perceived quality. Also, a set of new features was created to measure how content characteristics, for instance, the length of a given article, affects whether the reader appreciates the story. Moreover, features measuring the author's tone of speech and the subjective manner of narrating a news story, along with their strong or not social network inside the blogging platform, were created to capture how a given author influences people's attitudes towards a news story.

A preliminary evaluation of this work produced some interesting outcomes. Indicatively, the perceived quality of the articles may take different meanings depending on the category that they belong to. Although we found some generic, rather expected results, like the significant contribution that the number of followers may have on the popularity of an article (irrespective of the category), a more close investigation revealed the almost equal importance of other features and the verification of our proposed extended model. Each category generates a different classification of features as the most important, which is able to influence their acceptance from the digital audience. As such, perceived quality should always be regarded taking into consideration the category that an article belongs to (or other contextual characteristics) since a possible generalization of the term could lead to misleading conclusions and interpretations concerning the fundamental elements that constitute it as a term and what it represents. Furthermore, this study was able to extract several important rules based on these engagement features for each category, which may guide the writing of the next article of the authors, providing them with insights into the probability of popularity that may gain, if followed.

The remainder of the paper is structured as follows: Section 2, presents an extensive literature review around the topic of investigation as well as a comparison with related work indicating the main contributions. In Section 3, The proposed-extended model is detailed, along with the main dimensions and employed features. Section 4 refers to the method of evaluation and the data-set used and Section 5 discusses the two phases of the data analysis, the results in perspective, as well as the proposed rules for each category of news that might increase their perceived quality. Section 6, concludes this paper and addresses future work.

2. Related work

Some of the popularity metrics used by the media organizations to distinguish valuable content except for the traditional methods (ratings, circulation) are based on user online behavior and include shares, favorites, likes, "most emailed", "trending", "top ranked" and so forth. These metrics along with certain exposure measurements such as page views and time spent on an article, can provide an overview of the perceived value of a news story. Even though clicks are a sign that a news article caught the user's attention, they may not provide too valuable actual feedback on the user's perceptions, but rather be a function of other reasons, such as the items' position in the design of the site [55].

However, when a user consciously decides to hit "like" or "share", that reflects a more personal attachment to and connection with the content and can serve as a valuable indication of the perceived quality of the article. After all, people also use these actions to construct a personal image, because what they share reflects on who they are or want to be [65].

To find if there are certain dimensions of online article's perceived quality, a trustworthy proxy is required, that signals that a news story is valuable to the audience. Thus, the blog platform Medium.com, which consists of a miscellaneous collection of articles, features an interesting combination of liking and sharing: It allows the user to "clap" for an article, which shows the author that their story was liked. Arguably, these claps can be a better proxy for perceived quality than the mere number of views or the number of "shared" on other platforms is, as reading or sharing does not necessarily imply a positive value judgment. One can, for instance, read an article because they feel bored, or share it with a disapproving comment or to mock it.

Although, there is not, to our knowledge, previous research that explicitly supports the correlation of claps with perceived quality, given the fact that the claps on Medium.com are public and other users can see the articles one clapped for, claps can result not only in feedback to the author but also as reader's recommendations to each another. Furthermore, the number of claps (it is possible to clap up to 50 times for the same story) is also used as an input for the personalization algorithm to determine which stories one's followers will see, and hence it can be seen as an indirect way of sharing. The current approach argues that claps on Medium.com can be used as a proxy to measure the perceived quality of an article and until now they have been under-researched.

Next follows a review of the research on the digital transformation of the modern newsroom in terms of the evaluation of the quality of their products and how online users on social platforms and recommendation algorithms have altered the traditional gatekeeping process. The choice of Medium.com as an optimal case for the research on the topic of perceived quality is explained and the concept of quality is defined. Then the literature review continues with a brief overview of how the digital journalist tends to think like a marketeer implicating in a way the digital audience's feedback in the decision making of the news agenda. After that, a section about previous attempts to use audience engagement to measure quality through popularity is presented and the actors involved in online readership that play a decisive role in what content ends up being evaluated are mentioned. The backbone of the literature review is then used to help build the framework and the model for the predictions.

2.1. *Measuring quality*

Quality is a fundamental concept in building a brand name, customer satisfaction, and serve as an advantage against competitors. The notion of objective quality refers to superiority on some predetermined ideal standard or standards [67], while perceived quality is highly dependent on the consumer's judgment [48]. Accordingly, Aaker defined perceived product quality as "the customer's perception of the overall quality or superiority of the product or service for its intended purpose, relative to alternatives" [1]. This work considers the standpoint of Aaker, building upon the assumption that online articles could be regarded as the content of the services that Medium offers to the digital audience, and investigates how their perceived quality is influenced by their features of engagement and the category (purpose) that they belong, in relation to the next important article close to it.

In the liquid modern era, journalism is constantly changing [13], and even though online news are extensions of print journalism, they move to novel directions and forms [12]. In this digital and interactive space, the role of the online audience is empowered and publishers judge the quality of their products with regard to what drives more traffic to their websites. According to recent studies, decisions within news organizations (for instance, about which articles to promote) are increasingly driven by engagement metrics [23,35,63,71].

Engagement data are crucial when it comes to news placement online, with evidence suggesting that in some cases they are even more important than editorial opinions [35]. Tandoc and Vos [17] after observing and interviewing journalists in three online newsrooms for 150 hours, concluded among other things that journalists balance their editorial decisions between traditional journalistic norms and audience influence expressed through social media engagement and traffic metrics. The same study stressed also the implications of the journalist promoting their stories on social media, thinking like a marketeer and giving in to the market demand can compromise editorial autonomy. Thus, it is of crucial importance to find out how to reconcile the high popularity of news content with the perceived quality as expressed by the user through certain engagement measures.

This study is focused in particular on the perceived quality of online news articles. A very broad definition of news is being used here that do not limit the focus on hard news (such as politics or economy) but follow Bogart's [7] definition, who defined news as the "useful current information that is not necessarily generally important, but is of considerable importance to some people". Therefore, in this study, the articles that appeared in the following categories on Medium.com: Lifestyle, Business, Technology, Sports, Health along the general News category are considered to belong in the broad spectrum of news articles.

2.2. *Digital audience*

Nowadays social media users' exposure to online news can happen to a great extent unintentionally as their peers use social media networks and messaging applications to share and discuss news, resulting in social media users being actively engaged with news stories they did not choose to read [45]. In conjunction with this, filtering algorithms used for personalization and marketing reasons are largely based on user's feedback making other users' sharing and liking of information extremely relevant to what actually ends up appearing in front of the online reader's screen. At the same time journalists work at commercialized media workplaces [14] where reader's engagement metrics and page views drive revenue and have the power to change the news agenda and the article's placement on the news site [64]. Thus, digital media outlets try to produce meaningful content for the audience so that their stories will gain more interactions and increase their chances of appearing in user's news feeds. Remarkably, many editors in an attempt to earn more page views and increase readership come up with witty titles, sensational writing, or even employ click-bait techniques [57].

A recent study analyzed 1.67 million Facebook posts made by 153 media organizations aiming to map the extent of clickbait practice, its impact, and whether "mainstream media" or "unreliable media" used clickbait more often [56]. The researchers created a model that distributed sub-word embeddings learned from a large corpus, with an accuracy of 98.3% and showed that mainstream media are increasingly using clickbait articles especially in stories about celebrity, entertainment or lifestyle news. But while using such strategies may work in the short run, structurally misleading people in their quality judgment may be harmful in the long run, as the brand image will suffer and they are unlikely to return. In particular, even if users click on such an article once, they will not endorse it (and hence, not help to spread it). Therefore, it is crucial to understand which features of an article can enhance its perceived quality.

2.3. Audience engagement

Although finding a news story interesting is a matter of personal taste, it is possible to create a taxonomy of so-called news values [19,25]. This approach suggests that certain attributes of news can influence journalists' and audience news selection [18,31] as well as engagement with digital news stories [65,70]. Yet, there is no full consensus on which features have the largest impact on the so-called "shareworthiness" [65]. On the one hand, Trilling et al. find geographical proximity, involving Western countries, conflict, human interest, negativity, positivity, and exclusiveness to increase engagement, with slight differences between Facebook and Twitter engagement [65]. Valenzuela et al. confirm the findings for exclusiveness and negativity and added – amongst other features – deviance and relevance to the list [68]. Garcia-Perdomo et al., in contrast, cannot confirm an influence of proximity, but do confirm an influence of conflict and human interest as well as deviance (here called unusualness) [20]. They also find impact and prominence to be relevant predictors. In contrast to [68], they also found some effect of usefulness. Kilgo et al. concluded that the influence of such news values was less important than emotional appeals; however, their conclusions are based on small samples [32]. Relatedly, other work investigated how news values differ between sensationalist and non-sensationalist content [9]. Yet, such content features are not the only features to consider to understand what explains engagement with online news and information. Following a more normative approach, Urban and Schweiger [66] investigated the readers' ability to judge the quality of news based on journalistic norms, namely diversity, relevance, ethics, impartiality, objectivity, and comprehensibility. They argued that the audience is able to recognize differences in quality only regarding relevance, impartiality, and diversity and that media brand images proved to be an important heuristic when recipients have to evaluate news quality.

The data availability and the increasing focus on audience engagement for measuring audience preferences have led many social scientists to try to measure quality through those consumer metrics. A significant attempt towards understanding the link between various online measurements and quality is the software "Metrics For News" [2], created by the American Press Institute. The way the platform works is news organizations provide their web analytics along with certain journalistic qualities they perceive important. The software then blends all the available data into one metric that is easy for publishers to comprehend and draw conclusions about their performance and design their business strategy. According to the information on the software's website, the platform helps news organizations to "grow audiences, deepen engagement and drive subscriptions".

2.4. Actors involved in online readership

Search engines, social media networks like Facebook, news aggregators, and publishing platforms such as Medium.com use algorithms that control the information a user sees [8]. Thus, algorithmic curation delivers alternative results to similar search terms according to the previously collected behavioral data on the given user who sought information [41,49]. For instance, every user that visits Medium.com has a different experience due to Medium's recommendation system that delivers a different mix of stories to every visitor based on factors like their reading history, the authors, and publications they follow, and so on. These "algorithmic gatekeepers" [40] are updated frequently so there is not an easy way for authors to figure out how the algorithm forms its decisions.

In addition, editorial decisions about the article position in a news website are crucial concerning what content users see and inevitably end up rating. Previous research has shown that article promotion on the homepage of a news outlet improves the likelihood of it being shared, read, emailed, and several case studies have shown that these decisions are heavily influenced by audience choices [17,63]. Similarly, to news organizations, managerial decisions, the editors of Medium.com with the help of specific algorithms, review, and select on a daily basis the best stories published on the website and distribute them across the platform, the app, the newsletter and so on. The selection is based on high editorial standards and the guidelines for writing an article eligible for curation are available on the website [39].

Furthermore, users themselves curate their personal information environment [64] to serve their individual goals, while the relationships between users that follow one another can shape readership as well. Similar to other social media networks Medium.com consists of a community of users that through their behavior can impact what stories other users see on their homepage [64]. As Singer [59] showed in her work about user-generated visibility, nowadays the traditional editorial decisions about which stories are important for the audience to read, are followed by a secondary gatekeeping process in which the users themselves filter out valuable information by liking, upvoting, downvoting, sharing, emailing, and so on. Since on Medium.com there is no negative feedback button like for instance in Reddit [54], the possibility of more balanced feedback that users can even out through time does not exist. Therefore, more followers imply more potential readers that can lead to more positive feedback.

The actors discussed above, unquestionably influence the number of people who see a given article on their screen, therefore it can also affect directly the perceived quality of this article.

2.5. Main contribution in relation to related literature

Despite the advantages that “claps” offer compared to “shares” or “likes” for understanding the perceived quality of articles, they are rarely studied. To the best of our knowledge, no academic paper so far has studied how features that explain the perceived quality of articles on Facebook and Twitter can be used to explain “claps” on Medium.com. A reason for this may be that Twitter and Facebook as widely popular social networking sites draw most attention of both public and researchers. However, as discussed above, the “clap” as a signal that both implicates approval (as a “like”) *and* disseminates the story to one’s social network (like a “share”) is worth studying in detail for those who are interested in the perceived quality of an article. This study fills this gap. More specifically, the main contribution of this work is three-fold:

(a) It proposes a model of engagement metrics based on previous research that can quantify the perceived quality of different types of articles.
(b) It identifies how different classifications of engagement metrics contribute to the perceived quality of articles depending on the various categories of news. Also, the results show that an article’s genre affects the importance of features that reveal its perceived quality differently and should be treated as a situation-specific factor given the nature of each category. However, instead of merely focusing on identifying the role of different feature categories, this papers goes a step further and explores the nature of their effects.
(c) It recommends a set of rules, as guidelines to authors, based on the engagement metrics of the respective articles’ categories, for increasing the probability to gain popularity. While prior work on online popularity prediction on Twitter shows that there are ways to forecast a given article’s success before [4] and after publication [30], little work has examined how different attributes of an article before publication can influence its popularity. This study demonstrates for the first time to our knowledge, the importance of specific features that when combined, can shape online popularity and thus the perceived quality of a given article.

Henceforth, the current work concentrates previous efforts on the topic of identifying and predicting positive user interactions towards news content and test them using a new data set from a blog site. For this study various characteristics of an online news article have been used, such as its sentiment, novelty, and tone [20,65,68], along with the useful taxonomy of news tweets by Orellana-Rodriguez and Keane [47], that distinguished three dimensions of a tweet that influence its dissemination: user, content, and context features. The current study considers these three Twitter related feature dimensions and adjusts them to news articles, while also extends them by additionally proposing writing “style” as a category within content features.

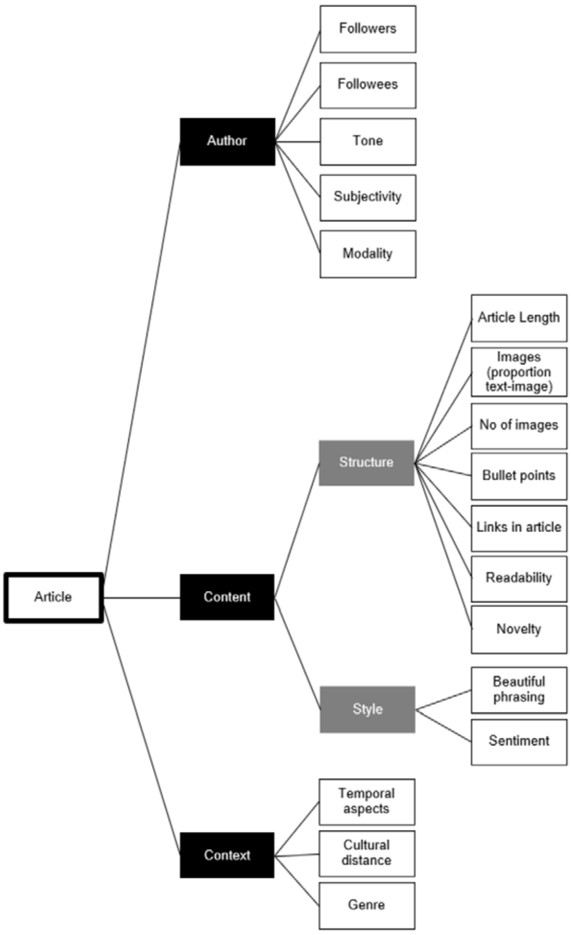

Fig. 1. Features that affect the perceived quality of the readers.

Then this taxonomy is used to an empirical test. By considering such a broad set of possible features, the current study sheds light on which ones can shape the perceived quality of news articles and which do not "work" depending on the genre, and proposes a specific set of guidelines to authors. In doing so, this paper will set a baseline and starting point for others who are interested in the prediction of "claps" or other online popularity metrics as well as for those who want to further refine the taxonomy developed in [47]. Also, earlier work that only focused on different subsets of the features considered in this research is extended and the exact nature of the relationship between these features and the perceptions of the audience is disentangled using a novel machine learning approach of rules extraction.

3. Model & feature extraction

The primary concern of this study is to create an inclusive model to measure and quantify the value of perceived quality across various article categories. The backbone of the model is structured using the overarching dimensions proposed by Orellana-Rodriguez and Keane [47], namely (i) author, (ii) content, (iii) context, as shown in Fig. 1.

As discussed above, perceived quality may be considered as a widely used validation factor and can be measured using various voting mechanisms for understanding the discrete actions of users on articles like acceptability, shareability, etc. Additionally, voting mechanisms are a way for communities such as Digg.com and Instagram.com to define content popularity and have been proven effective also for social Q&A communities like Quora.com to filter important answers [28]. Similarly, on Medium.com users' claps act as social signals to draw attention to influential people [51]. Besides following an author, every content creator can follow other users to expand their network.

Hereafter, the main features of our model are explained in detail along with the rationale for their selection:

3.1. Author features

Author's Popularity: This work could not have systematic data on the users themselves. However, the author's popularity which translates into the total number of followers on Medium can easily be calculated. Having a large number of followers may result in a great number of claps in a small period for a very small piece of text, while new authors have to work hard and produce long pieces of writing to achieve the same result [28].

Another way to determine an author's popularity is to look for the number of people they follow, as this can also contribute to build reputation and gain more votes, since users leverage networking communities by befriending other users and ask for votes in return [51].

Author's tone: Another characteristic of the author is the tone of speech he or she uses to communicate with the audience. According to marketing and communication specialists, the use of personal tone is a way for having a conversation with the audience, and expand one's brand, thus self-referential or self-reflexive messages are frequently used to attract the readers' attention [44]. Furthermore, previous research has shown that human-related stories are more likely to get positive feedback [37] and personal pronouns are an indicator of the existence of people in the story. This study investigates the use of personal pronouns (I, you, we, etc.), reflexive pronouns (myself, yourself, etc.), and possessive pronouns (mine, yours, etc).

Subjectivity: In principle, journalistic coverage is meant to be neutral, but recent research shows that social media drive journalists towards more subjective ways of presenting information [34]. To test whether more subjective news stories influence the crowd's appreciation the Python library TextBlob is used to measure subjectivity in textual data.

Modality: Grammatical modality is signaled by grammatical moods that express a speaker's general intentions and is implemented grammatically through three moods namely indicative, imperative, and subjunctive. To capture the degree of certainty of an author, we used pattern.en [11], which measures modality as a value between -1.0 and $+1.0$, where values $> +0.5$ represent certainty.

3.2. Content features

In the model, the focus is on content features that are expected to influence the reader. The model of Orellana-Rodriguez and Keane [47] is slightly extended by separating content to structure and style features.

3.2.1. Structure

More specifically the characteristics of the structure are article's length, number of images, image-to-text ratio, the existence of bullet points in the article, use of external links, readability metric, and novelty.

Article Length: The information depth is also depending on the average length of the articles. For instance, the New York Times prefers to focus on news depth rather than on news variation [53]. For that feature, the so-called "reading time" is used provided by Medium.com.

Images: The use of salient pictures is a way to attract audience attention and achieve link clicks. A recent study [38] revealed that stories with a photographic image of people in distress captured the attention of the participants who were more likely to donate money. Also, maintaining a good balance of text and images is an important factor for capturing attention and is used extensively for marketing campaigns. Consequently, to study images two different features are used: the total number of pictures in the article and the image-to-text ratio.

Bullet points: Listicles, meaning a mixture of 'list' and 'article' are very popular types of articles and according to research, they tend to be highly shareable [46]. Therefore, the existence of bullet points was calculated.

Links: According to literature [29] one of the signals of transparency in an article is the inclusion of external links to sources and documents, that could provide more verified sources, complementary information, or may present different angles. Furthermore, providing extra reading material can be a sign of extensive reportorial effort and the author's will to explain in more detail important matters to the general public.

Readability: Readability testing is a way to automatically check the clarity of writing. There are several metrics for measuring the article's complexity. In this paper, the Flesch-Kincaid Grade Level [33] is used, which is an advancement of the Flesch reading-ease test, and was developed for educational reasons, but has been widely applied to journalistic and political communication research [58,69]. Good readability means a text is accessible and easy to understand, therefore, when an article is too difficult (low reading ease), readers might be unable to comprehend.

Novelty: If an article covers new, exclusive information that is not present elsewhere, it may be more interesting for readers and enhance engagement [65]. A three-day sliding window was used and the normalized (to account for title length) number of nouns (which was determined using spacy [26]) that a headline shared with all concatenated headlines in the window to calculate a popularity score. A low popularity score can be interpreted as an indicator of high novelty [65].

3.2.2. Style

Beautiful phrasing: High-quality articles are often written using beautiful language [3,37], meaning more unusual phrasing and creative words that lead to more positive feedback [61]. This study uses Term Frequency-Inverse Document Frequency (tf·idf) as a proxy because the use of rarer words can be seen as an indicator for the use of beautiful language.

Sentiment: Emotional text can spark feelings in the audience and trigger different kinds of emotional responses. For instance, people are attracted to negative news because it satisfies the need for survival and security and helps to avoid threatening situations in their environment [5]. Both positivity and negativity have been shown to influence news sharing [61,65,68], with some psycho-physiological experiments showing that negative news evokes stronger and more sustained reactions than does positive news [60]. This study uses Vader [27], which calculates the overall polarity of the text along with the sentiment: positive, negative, or neutral. Other studies have focused on evidence that online content that causes awe or anger and anxiety emotions to the users increases the chances of becoming viral [6,22]. For the emotions extraction, we used the NRC suite of lexica, which is based on Plutchik's model, suggesting the existence of eight basic emotions: joy, sadness, anger, fear, trust, surprise, disgust, and anticipation [43]. For the measurement, the counts of emotion words are computed, each normalized by the total number of article words. Except for polarity and emotions, the overall emotional density expressed in the text is also calculated.

Title polarity: The headline of an article is a significant aspect of every story as modern news readers and social media users scan the headlines before they decide to click [16,57]. To measure polarity expressed in the headline, again, Vader is used [27].

3.3. Context features

Context features explore the conditions in which an article was posted and include temporal and locational features.

Genre: The topic of the article reflects certain characteristics of its nature and it has been greatly investigated in previous work [3,37].

Temporal: The data set consists of articles from only one year, so only hour and type of day are used as temporal features.

Cultural distance: Previous research [65] found that geographical proximity and stories that mention Western countries increase social media engagement. This study, uses references to geolocations from the texts using the Mordecai system [24], which extracts toponyms and returns their coordinates.

4. Method & dataset

The analysis was conducted on a data set of a total of 200K articles taken from a single source, namely Medium.com, so that the extraction of the features would be consistent across all genres. The choice of the independent variables was made based on the characteristics that according to the studied literature seem to contribute to online content popularity. To quantify the perceived quality of the articles, this study opted for the official metric of this online community which is called claps, and serves as the dependent variable for our model. To determine if and to what extent the proposed characteristics of an article are related to its perceived quality a series of experiments took place. To this end, the model was trained using the aforementioned features and divided the prediction problem into two phases, the first was to predict the perceived quality of an article taking as input a mixed data set, while the input data for the second phase was each genre separately.

Table 1
Descriptive statistics

Category	No. articles	Mean (claps)	Median (claps)	75th percentile
News	17762	150	0	22
Technology	30883	236	2	59
Business	35768	65	0	14
Health	49072	39	0	3
Sports	12664	19	0	5
Lifestyle	54561	347	22	121

4.1. Data

More specifically, the data set was collected from the website Medium.com, which covers a range of topics from tech to politics to well-being, and for this study, articles from the following six categories News, Technology, Health, Business, Sports, and Lifestyle were retrieved. A Python program was written that scraped the website in January 2019 and collected articles published between September 2017 to September 2018. The total amount of articles before cleaning was 247.071 and had a large distribution of claps ranging from 0 to 292K. Before the analysis the data set had to be prepared, hence all the news articles were processed for stop words, nonstandard words and characters removal, stemming, and tokenization. In addition, articles not written in English were removed, NaN values, and HTML code from the text using regular expressions. After the preparation for the analysis 200,710 (Table 1) cases remained for inclusion in the experiments. One of the issues that appear studying the Table 1, is the fact that the vast majority of the articles have equal or near to zero claps. This causes the creation of imbalanced classes which is a common problem in machine learning classification since there is a disproportionate ratio of observations in each class. Standard classification algorithms, that do not take into account class distribution, are overwhelmed by the low-popularity class and they ignore and misclassify the minority of successful articles since there are not enough examples to recognize the patterns and the properties of the popular class. In this work, attention is taken using specific techniques and measures of quality to predict the rare but important class of successful articles.

For the extraction of the features, we used several Python libraries for text analysis, cleaning, filtering, counting words, and processing textual corpora. After determining the main categories of features for the model, the importance of each one was examined using a random forest classifier. More specifically, the number of claps was used as the dependent variable and created decision trees able to predict whether the claps count of a given article will be high or low. First, the experiments used a mixed-topic data set, (News, Technology, Health, Business, Sports, Lifestyle), and then the machine learning algorithm was tested on every genre separately to provide better insights into the different news categories. Each category had a different number of articles, and different statistical values in terms of claps distribution, while the median for the majority of the articles is 0 claps.

Articles with claps above the 99.8th percentile were considered outliers and were removed. Furthermore, the features were normalized and checked for feature correlation. Also, a Principal Component Analysis (PCA) was conducted to search for redundant features. However, the analysis showed that the large set of components could not meaningfully be reduced into a smaller one without losing a lot of variance.

5. Data analysis & findings – in two distinctive evaluation phases

The analysis was divided into two phases. In Phase A the whole data set of the articles was examined, to find the most important features and create a baseline for further analysis. After that, the study was focused on every different news genre trying to capture if there was any variation in the predictive power of the importance of each feature according to the genre the articles belonged to and in that case to what degree influences the perceived quality. In Phase B, a decision tree classifier was used to extract the rules of the best performing leaves that provided specific combinations of the engagement features. These features may increase the chances that a given article has

Table 2
Permutation importance for the top 10 features

Feature	Weight
Followed By	0.5810 ± 0.0240
Article length	0.0517 ± 0.0033
Following	0.0367 ± 0.0088
Flesch	0.0166 ± 0.0027
Polarity	0.0133 ± 0.0010
Tone of speech	0.0125 ± 0.0039
img/word	0.0087 ± 0.0026
No. Images	0.0081 ± 0.0020
Subjectivity	0.0062 ± 0.0044
Contains USA	0.0061 ± 0.0019

to gain greater popularity and thus higher perceived quality when they reach certain values according to the specific leaf.

This study separates articles with low and high claps by converting the numerical variable that represents the total number of claps into a categorical one. In both phases, a binary classification task was created aiming at constructing a model able to give us the importance of each feature as well as an interpretation of the prediction to enhance our understanding of a successful article. In our experiments, we used 80% of the data set for the training set and the other 20% for testing purposes. We used the scikit-learn [52] implementation of tree-based machine learning approaches (decision tree, random forest, xgboost) since tree models can be more interpretable than other complex models and provide more descriptive explanations of the effect of the features.

5.1. Phase A – evaluating the importance of the proposed model engagement metrics in relation to the perceived quality

In the first phase of experiments, the original data set was randomly sampled and a new one was created that consisted of 60K articles, 10K from each category (News, Technology, Health, Business, Sports, Lifestyle). The original data set can be seen in Table 1. To predict the appreciation of the articles it is crucial to get the perceived quality rank right, particularly at the high ranks, by correctly distinguish the most highly appreciated articles. Therefore, in the new data set, two buckets of articles were randomly sampled, the high and low with each consisting of 14K articles. More specifically, 14K articles that were placed in the top 25% of claps with the highest values were in the high-claps bucket, while 14K of 0 claps articles ended up to the low-claps bucket.

For prediction purposes, the randomForest package from the Python library scikit-learn [52] was used to create a classifier of the articles and to measure the importance of the predictor variables. The significance of a variable is calculated internally during the construction of the decision trees by checking the raise of prediction error when data for that variable is permuted while all others are left unaltered. The F-measure (F1) was adopted to evaluate the performance of the model. The best score of all between the three different classifiers was generated by the XGBoost algorithm with an F1-Score of 80%.

Afterward, for the interpretation of the random forest, the contribution of every feature on the prediction was measured. The variables were ordered from highest to lowest rated and the 10 most important out of 30 were kept. The ELI5 Python package for “Inspecting Black-Box Estimators” [36], was used for the calculation of permutation importance (Table 2).

To get better insights into which features contribute to the high or low-claps bucket, the importance score of every bucket was examined separately, starting from the most important variables that have more predictive power over the high-claps bucket.

Figure 2 presents the importance of the variables of the selected dimensions. The dimension of author is the first with Authors’ popularity (Followers & Followees, Subjectivity, Tone, Modality), second is the Content dimension and specifically Structure (Article Length, Images, Readability, Popularity score) and Style (Polarity).

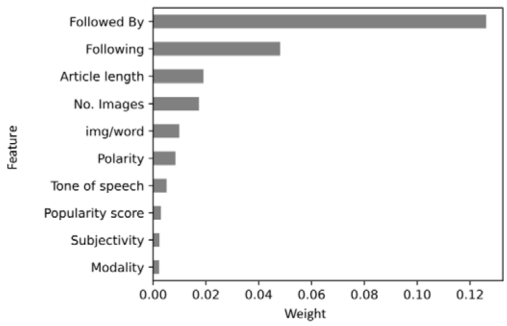

Fig. 2. Feature importance score for high claps bucket.

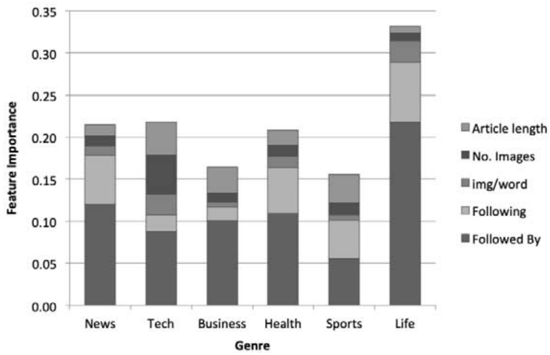

Fig. 3. The importance measures of top-5 features in six different article categories.

In general, as shown in Table 2 the social network of an author, along with length, pictures and polarity are the most important characteristics that influence the perceived quality of an article. Interestingly, the results reported in Fig. 2 indicate high accuracy and give an idea about the predictive power of the features contributing to the high-claps bucket, which will be used as the baseline for the analysis of the different news genres that will follow.

5.1.1. Evaluating the importance of each category's engagement metrics and their influence to the perceived quality

To check whether some features are more important depending on the genre, different data sets were created for every category and the same process was applied to split the articles into two classes of the same proportion of texts, one bucket with the 25% higher clapped texts and the low-clapped bucket with the ones that got 0 claps. Despite the fact that all articles were taken from a single source, differences exist on the distribution of the claps in every category as shown in Fig. 3.

The importance scores of the features revealed several interesting insights into the three dimensions of the model. The author dimension is the most influential regarding perceived quality, content follows with structure features being more important than style ones and the least significant is context. In more detail, the different data sets presented some slight differences, as detailed below.

News: The News data set included almost 18K articles with the maximum number of claps to be 87K. The average number of claps for a News article on Medium.com (see Fig. 4) was 150 claps, and the 75th percentile was 22 claps that translate in 4434 articles. Thus the high-claps bucket for news was a random sample of 4K articles in the top 25% of claps and other 4K with 0 claps. The F1-score for the News data set was 82%.

Technology: The articles from Technology category were 30K and the maximum number of claps for a tech article was 111K, while the average number of claps was 236 (see Fig. 5), showing that readers on Medium appreciate tech articles a lot. Almost 8K articles were inside the 75th percentile which was 59 claps, so the total of 16K articles were used in this experiment. The F1-score for Technology data set was 83%.

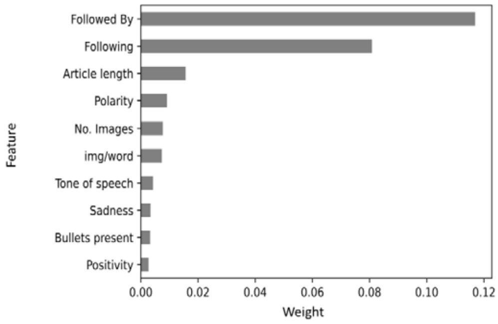

Fig. 4. Feature importance score for the high claps bucket in news data set.

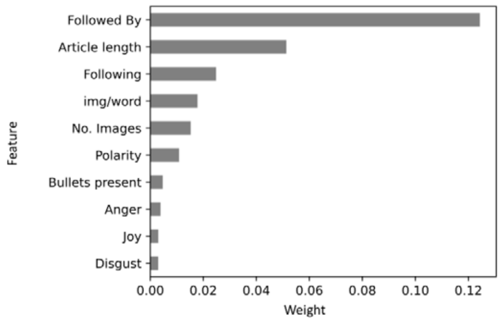

Fig. 5. Feature importance score for the high claps bucket in technology data set.

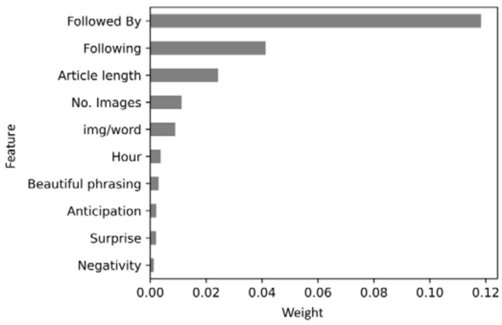

Fig. 6. Feature importance score for the high claps bucket in health data set.

Health: Almost 50K articles belonged to the Health data set where the more highly clapped article had gotten 36K claps. The average number of claps for a health-related article was 39 claps, and the 75th percentile was 3 claps (see Fig. 6), which corresponded to 11829 articles. Therefore the high-claps bucket was a random sample of 10K articles in the top 25% of claps and other 10K with 0 claps. The F1-score for Health data set was 78%.
Lifestyle: The Lifestyle data set had more than 54K articles and this topic seems to be the most favorable one, since it had the biggest average number of claps, namely, 347 (see Fig. 7) as well as the most highly appreciated article with 292K claps. More than 13K articles were in the 75th percentile which was 121 claps, thereafter 13K articles were used for the high-claps bucket and another 13K for 0-claps bucket. The F1-score for Lifestyle data set was 86%.
Business: The Business data set included almost 35K articles with the maximum number of claps to be 39K. The average number of claps for an article in the Business section of Medium.com was 65 claps (see Fig. 8), and the 75th percentile was 14 claps that translate in almost 9K articles. Therefore the high-claps bucket was

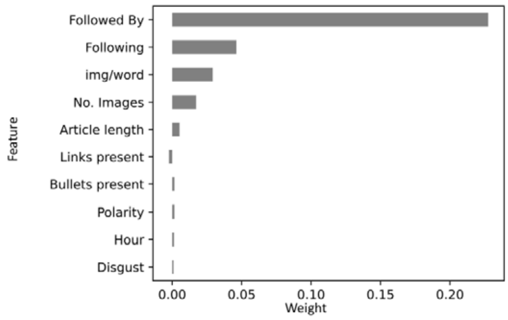

Fig. 7. Feature importance score for the high claps bucket in lifestyle data set.

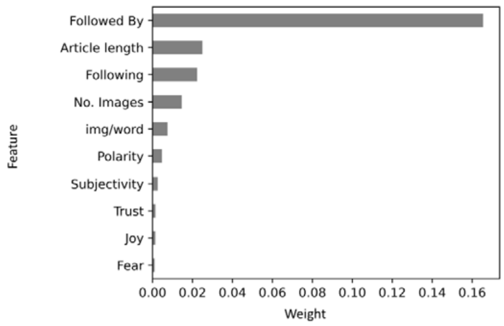

Fig. 8. Feature importance score for the high claps bucket in business data set.

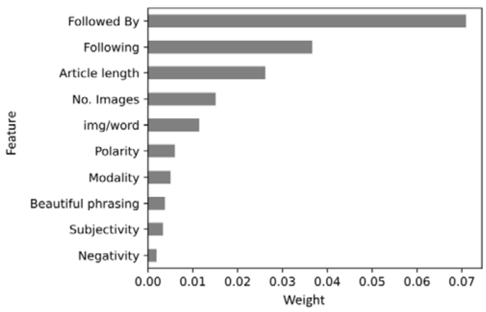

Fig. 9. Feature importance score for the high claps bucket in sports data set.

a random sample of 8K articles in the top 25% of claps and other 8K with 0 claps. The F1-score for Business data set was 76%.

Sports: The data set for Sports included more than 12,5K articles with the maximum number of claps that an article received being 7,5K. Furthermore, the average number of claps for sports articles was 19 (see Fig. 9) with the 75th percentile being 5 claps, meaning that 3K good cases and other 3K bad cases were used for inclusion to the model. The F1-score for Sports data set was 74%.

Table 3 shows the precision, recall and F1 scores of each bucket of the seven data sets.

5.1.2. Discussion of the results

First, the most important feature is the number of followers an author has, which is the only feature that remains stable in all data sets. This is somewhat expected because the more followers the author has, the more claps their articles will receive. Similarly, the number of people that the author follows influences the perceived quality of their article, meaning that the social network of the author has a positive effect. Second, the length of an article is

Table 3
Precision, recall, and F1 scores

	Precision	Recall	f1-score
Baseline	0.8	0.8	0.8
News	0.85	0.79	0.82
Tech	0.8	0.85	0.83
Health	0.75	0.83	0.78
Lifestyle	0.88	0.85	0.86
Business	0.77	0.75	0.76
Sports	0.74	0.74	0.74

associated with perceived quality. However, from the results of Phase A it cannot be interpreted if longer articles are correlated with better metrics or the opposite is true in this case.

Third, images seem to be of great significance in terms of positive feedback. Both the number of pictures and image-to-text ratio are among the top five important features in most data sets. This is a logical consequence since a picture communicates a great deal of information and can make the text more interesting and easy to understand. Fourth, sentiment has an impact on the crowd's endorsements, since the polarity of a text, which translates to the sum of positive and negative scores, appears to be strongly correlated to perceived quality. Moreover, positivity, negativity, sadness, neutrality, joy, trust, anger, fear, disgust, surprise, and anticipation also show up as less important features.

Interestingly, the findings suggest some differences depending on the category. Specifically, the existence of external links in the body of the article appears to be somewhat important in the case of Lifestyle, while beautiful phrasing is fairly relevant to Health and Sports categories. Meanwhile, readability (Flesch score) is associated with perceived quality when it comes to News articles. Notably, subjectivity is associated with high claps only in Business and Sports and tone only in the News data set. Likewise, the presence of bullet points in the text is quite significant in News, Technology, and Lifestyle, whereas grammatical modality is slightly correlated to the Sports genre. Surprisingly, the only temporal feature that ended up having a small degree of significance is the hour, which came up in Health and Lifestyle data sets.

According to the results, articles that belonged to different genres presented some slight differences regarding the importance of the predictive features. Additionally, one of the main objectives of this paper was to understand the relationship between the various features and their role in the identification of the perceived quality of the articles. Accordingly, the features of each article's category were sorted based on their importance and the top-10 and top-5 ones were compared with the baseline in Fig. 2. The main concern was to identify their similarity, and enrich the understanding regarding the influence that different genres have on the features that drive the value of the perceived quality of the articles. Calculating the cosine similarity and distance of each vector from the baseline (see Table 4) some interesting observations were drawn: (a) An article's genre affects the importance of the engagement metrics that reveal its perceived quality differently, and (b) the importance of the engagement metrics is positively correlated with the similarity that each genre has with the baseline vector (the average similarity of vectors for top-10 is 65% while for the top-5 raised to %80 – with the category Health to have even a correlation of 1 with the baseline). Such results dictate that perceived quality could not be quantified using a predetermined or fixed classification of the engagement metrics across articles' genres, but rather should be treated as a situation-specific factor considering always the contextual characteristics of an article. The latter might be able to influence the importance and the hierarchy of the features, assigning to perceived quality a different semantic understanding of what it represents.

Given the aforementioned understanding, in the next section, the engagement metrics of each genre are explored in more detail and a set of rules is generated for each one, expecting to estimate the perceived quality of articles in a specific category before their publication.

5.2. *Phase B – proposing a set of rules with respect to the engagement metrics of each category*

In this phase, the creation of a set of rules for each category of articles based on the engagement metrics discovered in the previous phase was investigated. One of the main aims of the study is to help authors during the writing

Table 4
Comparison of top-10 and top-5 articles categories' features similarity

Article types	Similarity (top-10)	Distance (top-10)	Similarity (top-5)	Distance (top-5)
Baseline	*1.000*	*0.000*	*1.000*	*0.000*
News	0.632	0.368	0.775	0.225
Technology	0.447	0.553	0.632	0.368
Business	0.707	0.293	0.775	0.225
Health	0.775	0.225	1.000	0.000
Sports	0.707	0.293	0.894	0.106
Lifestyle	0.632	0.368	0.775	0.225

process, by providing them with some initial insights regarding the possible popularity that their articles may have if written in a particular way. The proposed rules can point towards the direction of a high probability of gaining popularity depending on the category that an article belongs to.

5.2.1. Method & rule extraction

More specifically, for each genre, a threshold value was set, that results from the classification of articles based on the number of claps they have received and the definition of the 10% of them with the most claps as high popularity and the remaining 90% as low popularity articles. The articles then were given as an input to the decision tree classifier of scikit-learn library [52], the tree formation of which produced the rules for writing popular articles. These rules come from following the path from the root of the tree to every leaf node. Each node contains a numeric expression that involves an article feature, for instance, number of images $>= 2.5$, so the final rule for each path leading to a leaf node contains a set of arithmetic expressions associated with the logical "AND". At this point, it has to be mentioned, that binary decision trees introduce difficulties in rule extraction, due to the fact that they create many splits where attributes appear more than once in any path from the root to the leaf. Additionally, two more features were used in this phase, the (Followed By/Followers) ratio and the numbers of words used in the title.

5.2.2. Leaf node selection

From the above it can be concluded that each leaf node can lead to a rule, but the question is whether this rule is (a) important and (b) valid and robust. These two constraints can be satisfied, if the following conditions become true:

Importance: Since the purpose of this research is to find rules for popular articles, attention is focused on leaf nodes that improve the initial probability of high popularity, which is 10%. Thus, for a rule to be important, attention is turned to leaves with a probability of high popularity at least three times that of the original, i.e., greater than or equal to 30%.

Validity and Robustness: Machine learning algorithms, such as decision trees, use a large amount of input data for training and the rest for testing. So, there are two kinds of probability, depending on the data set. For the leaf node selection, what matters is the predicted probability calculated by applying the algorithm to the train set, that is, the number of observations of high popularity class that has been "captured" by that leaf over the entire number of observations captured by that leaf (during training). However, the probability for the same leaf node may differ when the decision tree algorithm is applied to the test data. If these two probabilities vary widely, it is obvious that we have an unstable node, which will not lead to a reliable conclusion. The difference between these two probabilities is called *misclassification error* for the specific leaf and it is a measure of validity and robustness of the extracted rule. Leaf nodes with misclassification errors lower than 0.1 are considered appropriate for rule extraction. It was noticeable, that those leaf nodes having a large number of articles, i.e. more than 1% of the training set, appeared to be more stable having low misclassification error. Based on the above preconditions for leaf selection, the quality of the extraction process cannot be jeopardized and the results can be generalized with certainty.

According to phase A, to answer the questions about the nature of the effects of specific important features to high perceived quality, the genre of the given article should be considered. Therefore, every genre was examined separately, with the rules produced from the decision tree being different. Also, for each genre, the mean misclassification error value was calculated for all the rules produced for the articles belonging to that genre.

5.2.3. Results – rules

In this section, the rules that have been produced for each category are presented:

News Articles

RULE #1 – Probability for high popularity: 88,59%

- Followers 3206–102679
- No. Images > 3,5
- Tone of speech > 8,587%
- Subjectivity > 0.452%
- F/F Ratio 6.406–155,913

RULE #2 – Probability for high popularity: 100,00%

- No. Followers > 102679
- No. Words > 421.5

misclassification error (mean): 0.0199

Tech Articles

RULE #1 – Probability for high popularity: 32,99%

- Followers 41.5–509.5
- No. Words > 812.5
- F/F > 1.159–5.905
- Tone of speech > 15.053 Contains USA $\leqslant$ 0.5

RULE #2 – Probability for high popularity: 31,90%

- Followers 509.5–5004.5
- No. Words 767.5–1193.5
- Bullets present > 0.5
- F/F $\leqslant$ 11.504
- Anticipation > 4.202

RULE #3 – Probability for high popularity: 54,77%

- Followers > 509.5
- No. Words 767.5–2914.5
- Bullets present > 0.5
- F/F > 11.504
- Density > 14.414
- Tone of speech > 13.563

misclassification error (mean): 0.0223

Business Articles

RULE #1 – Probability for high popularity: 34,34%

- Followers 112.5–802.5
- F/F > 5.164
- No. Images > 1.5
- Links present $\leqslant$ 0.5
- Negativity > 0.033

RULE #2 – Probability for high popularity: 32,69%

- Followers 802.5–22747.5
- No. Words $\leqslant$ 722.5
- F/F 20.341–347.402
- Flesch $\leqslant$ 16.25

RULE #3 – Probability for high popularity: 39,61%

- Followed By 802.5–3790.5
- F/F > 2.711
- No. Words > 722.5
- No. Images $\leqslant$ 2.5
- Contains USA $\leqslant$ 0.5

RULE #4 – Probability for high popularity: 57,65%

- Followed By 802.5–3610.5
- F/F 2.711–337.208
- No. Words > 722.5
- No. Images > 2.5

misclassification error (mean): 0.0278

Health Articles

RULE #1 – Probability for high popularity: 37,73%

- Followed By 162.5–804.0
- No. Words > 812.5
- F/F > 1.123
- Negativity > 0.065

RULE #2 – Probability for high popularity: 41,91%

- Followed By 1872.5–10073.5
- No. Words 351.5–1140.0
- Beautiful phrasing $\leqslant$ 3.104
- Tone of speech > 13.052

RULE #3 – Probability for high popularity: 52,10%

- Followers 804.0–10073.5
- No. Words > 1140.0
- F/F 1.489–18.16
- Negativity $\leqslant$ 0.106

misclassification error (mean): 0.0226

Sports Articles

RULE #1 – Probability for high popularity: 34,61%

- Followers > 1955.5
- No. Words 261.0–1305.0
- Modality $\leqslant$ 0.661
- No. TitleWords > 5.5
- Bullets present $\leqslant$ 0.5
- Beautiful phrasing $\leqslant$ 4.083

misclassification error (mean): 0.0379

Lifestyle Articles

RULE #1 – Probability for high popularity: 33,85%

- Followers 826.5–4098.5
- No. TitleWords > 5.5
- No. Words > 1476.0
- Positivity > 0.123

RULE #2 – Probability for high popularity: 40,49%

- Followers 4098.5–7976.5
- No. Words > 754.5
- F/F > 1.031–120.281

RULE #3 – Probability for high popularity: 95,06%

- Followed By > 9047.0–38530.5
- No. Words > 550.5
- F/F $\leqslant$ 230.884
- Surprise $\leqslant$ 4.936
- Positivity > 0.097

RULE #4 – Probability for high popularity: 99,79%

- Followers 9047.0–38530.5
- No. Words > 550.5
- F/F > 230.893
- Sadness $\leqslant$ 9.0
- img/word > 0.001
- Beautiful phrasing > 0.757

misclassification error (mean): 0.0188

5.2.4. Discussion of the results

As can be observed from the previous section, a sufficient number of rules (from one to four) were extracted for each article genre. Those rules may be used as guidelines for the authors in the writing of their next articles; indicating the probability of popularity that they may gain if they are taken into consideration. For increasing the quality, a specific number of leaf nodes was exploited for structuring the rules, rather than the whole data set, which might bring some variation in the engagement features' importance. More specifically, only one feature was involved in all the rules, the number of user's Followers. Each rule aims at authors of different popularity, as it is expressed through Followers, Followees, or the (Followers/Followees) ratio. Regarding the rest of the features taken into account, the Author and Content dimensions were primarily involved.

For the News articles, the results revealed two rules, with a very high probability of gaining popularity. Authors who have a large number of followers, from 3206 to 102679, can write articles that are very likely to become popular if they add more than three images in their articles. Furthermore, they should be more objective, but not absolutely, while expressing a personal opinion is desirable. They also have to use personal pronouns (more than 8,6% of the total words) and be followed by a lot more people than they follow. The second rule for the News articles, that guarantees 100% success relates to authors with over 102679 followers and articles of at least 421 words. Probably, these are extremely popular accounts, producing material that always leads to the public's satisfaction.

Tech articles' authors who have not managed to gain a reputation on the platform can achieve high popularity if they have many more Followers than Followees, write long enough stories, use personal pronouns, and don't refer to the United States. If they have a smaller number of Followers, writing shorter stories is acceptable, but they have to use lists with bullets, and words that indicate anticipation (more than 4,2% of the total words). If they want to increase the probability of high popularity they can write medium length or extensive stories in which use list-based articles, personal pronouns, and over 14% emotionally charged words. It is very important to have many more Followers than Followees.

The first rule for Business articles is for authors with a few Followers. To produce articles with an increased probability of popularity, they must be followed by at least five times the number of people following them. Also, they have to use at least two images, more than 3,3% negatively charged words, and not to use external links. If they have gained a little more Followers (more than 802.5), there are three ways to increase the popularity of their stories. They can write short and easy to read stories, otherwise, they have to write more extensive stories, including more (or less) than three images depending on the existence of US toponyms.

Health articles' authors with a small number of Followers, or a ratio of Followers – Followees near to one, can simply increase the probability of high popularity by using negatively charged words to more than 6.5% of the total words. It seems like the negative news in this area can win the audience. If authors are more popular, they could write an article ranging from 352 to 1140 words in which they do not use rare vocabulary to a great extent (at most 3.1% of total words) but use enough personal pronouns as a percentage over than 13%. Alternatively, they should spend time writing a lengthy read of over 1140 words in which the negatively charged words do not exceed the 10.6% of the total words. Also, they should have 1.5 to 18 times more Followers than Followees.

To write stories with an increased probability of gaining popularity, Sports authors have to be already popular and have more than 2000 Followers. In this case, they should write a text of 261 to 1305 words in which they express certainty, but not in an absolute way. The use of rare vocabulary should be limited, the absence of a list format from the text is desirable, and the title should not be short, as it should consist of at least six words.

Finally, for the Lifestyle articles, moderately popular authors should write an extensive story, longer than 1476 words, of which over 12.3% should be positively charged. Finally, they should give their story a title of at least six words. For writers who have gained a bit of a reputation, their chances of success are easily increased if they have more Followers than Followees and write over 755 words. In the field of Lifestyle, if authors have acquired a large number of Followers already, they can almost certainly write articles that will become popular if the article's length is bigger than 550 words. If the Followers – Followees ratio is less than 230, the use of words showing surprise should be limited (not exceeding 5% of total words), however, positive words should exceed 12.3%. Otherwise, sad words have to be less than the 9%, rare words more than 0.75%, and at least one image per thousand words should be included.

6. Conclusion & future work

This work primarily discusses the prediction task of the perceived quality of articles published on the blogging platform Medium.com. It considerably extends a preliminary version of this work in [61] and [62] by introducing additional features (e.g., from [65]) and including a framework with three dimensions, namely, author, content and context (see [47]). The best predictor consistently is the number of followers, which is not surprising: if more people see it, more people will clap. Yet, other important factors have been revealed that could be interpreted as indicators for the popularity of an online article. Moreover, this research is one of the first in the area, to our knowledge, that places systematic emphasis on the concept of perceived quality of various articles in different categories and explores a number of characteristics (engagement features) so to be able to forecast the article's popularity. To quantify the perceived quality of the articles, the number of "claps" was the dependent variable of the proposed model. Furthermore, to define if and to what extent the features contribute to the future online success of an article, tree-based machine learning algorithms have been employed for the prediction tasks. The findings support what has been showcased already in previous research, that indeed factors related to the author's reputation, along with content-based characteristics of an article mainly length and images, and sentiment expressed in the text can be predictors of a reader's perceived quality.

In addition, results have also shown that an article's category affects the importance of features that reveal its perceived quality differently and should be treated as a situation-specific factor given the nature of each category. The latter, constitutes a main contribution of this work, making an important step toward the understanding of the perceived quality of online articles.

Also, on online publishing platforms like Medium, the main goal is to keep users interested for as long as possible by supplying them with the most relevant content. To that end, the underlying personalization scheme leverages data from previous interactions with the individual user to tailor their experience and provide them with the perfect fit between readers' preferences and the article's actual attributes. Articles posted on Medium.com are recommended to readers with similar preferences which means that articles are mainly judged by the right community of users. The only limitation in the proposed approach originates from the unknown level of dissemination of each article based on the personalization strategy used. Each article receives a different degree of exposure and, as expected, more famous authors often reach a greater audience. As a result, the "Followers" feature is proved to be the most important factor for a successful story.

The current research delves deeper into the exact relationship between the engagement features and the popularity by examining the production of rules that improve online articles' probability of gaining high popularity on the Medium platform, based on the prediction. This could be considered as one more significant contribution of this paper, where essentially, the proposed model's features have been further expanded and with the help of the machine learning algorithm, i.e., Decision Tree, a total of 17 rules have been formulated for the six categories of the articles. These rules can increase the probability of gaining high popularity from the initial 10% to a percent between 30% and 100%, by giving specific values to some of the features of the articles.

Of the 32 features that shaped the bottom-up model, 21 were those that participated in the rules that could affect the popularity of articles. Features concerning the popularity of the author, the writing style and the structure of the article were the ones that helped to gain high popularity. However, the feature that is involved in all the rules was again the number of followers, which due to its high importance, is a decisive feature of the rules. Observing the rules, it can be concluded that in most cases the author or the media outlet must have more followers than followees. Regarding the remaining of the engagement features, a universal conclusion cannot be drawn since the rules differ significantly, especially when it comes to different types of articles. This also adds an important nuance to previous work, based on which one may have assumed that stylistic features, for instance, may have played a greater role than we could empirically confirm.

This work can inspire more discussion and research towards different approaches to identify characteristics of news that resonate with perceived quality. Based on this paper, the future work could head to two directions. One is to try a top to bottom approach and use machine learning models to explore how and if features related to journalistic quality norms shape audience evaluations of an article's quality, or digital engagement metrics are only appropriate to capture user behaviors towards the quality of popular journalism. The second direction is to further substantiate the premise that claps are indeed a proxy for perceived quality. To that end, it would be worthwhile to design a study in which human annotators judge the quality of a large number of Medium.com articles without knowing the number of claps, and then correlating their judgment with the number of claps. This approach will be useful also in answering to what degree do unique characteristics of the individuals such as income, education level, gender, and age influence their preferences and perceptions of news quality.

7. Copyright

Acknowledgements

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under the HFRI PhD Fellowship grant (GA. 14540).

References

[1] D.A. Aaker, *Managing Brand Equity*, Simon and Schuster, The Free Press, 2009.

[2] American Press Institute, [n.d.], Metrics for News. https://www.metricsfornews.com, Accessed: 2020-02-03.

[3] I. Arapakis, F. Peleja, B. Berkant and J. Magalhaes, Linguistic benchmarks of online news article quality, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, 2016, pp. 1893–1902. doi:10.18653/v1/P16-1178.

[4] R. Bandari, S. Asur and B.A. Huberman, The pulse of news in social media: Forecasting popularity, in: *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[5] R.F. Baumeister, E. Bratslavsky, C. Finkenauer and K.D. Vohs, Bad is stronger than good, *Review of General Psychology* **5**(4) (2001), 323–370. doi:10.1037/1089-2680.5.4.323.

[6] J. Berger and K.L. Milkman, What makes online content viral?, *Journal of Marketing Research* **49**(2) (2012), 192–205. doi:10.1509/jmr.10.0353.

[7] L. Bogart, *Press and Public: Who Reads What, when, Where, and Why in American Newspapers*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1981.

[8] M. Carlson, Facebook in the news: Social media, journalism, and public responsibility following the 2016 trending topics controversy, *Digital Journalism* **6**(1) (2018), 4–20. doi:10.1080/21670811.2017.1298044.

[9] K. Danielle, Kilgo, S. Harlow, V. García-Perdomo and R. Salaverría, A new sensation? An international exploration of sensationalism and social media recommendations in online news publications, *Journalism: Theory, Practice & Criticism* **19**(11) (2018), 1497–1516. doi:10.1177/1464884916683549.

[10] R. Darnton, Writing news and telling stories, *Daedalus* **1975** (1975), 175–194.

[11] T. De Smedt W. Daelemans and T. De Smedt, Pattern for python, *The Journal of Machine Learning Research* **13** (2012), 2063–2067.

[12] M. Denis, *McQuail's Mass Communication Theory*, 6th edn, 2010.

[13] M. Deuze, Journalism in liquid modern times: An interview with Zygmunt Bauman, *Journalism Studies* **8**(4) (2007), 671–679. doi:10.1080/14616700701412233.

[14] M. Deuze and T. Witschge, *Beyond Journalism*, John Wiley & Sons, 2020.

[15] N. Diakopoulos, *Automating the News: How Algorithms Are Rewriting the Media*, Harvard University Press, Cambridge, MA, 2019.

[16] D. Dor, On newspaper headlines as relevance optimizers, *Journal of Pragmatics* **35**(5) (2003), 695–721. doi:10.1016/S0378-2166(02)00134-0.

[17] C. Edson, J. Tandoc and T.P. Vos, The journalist is marketing the news: Social media in the gatekeeping process, *Journalism Practice* **10**(8) (2016), 950–966. doi:10.1080/17512786.2015.1087811.

[18] C. Eilders, News factors and news decisions. Theoretical and methodological advances in Germany, *Communications* **31**(1) (2006), 5–24. doi:10.1515/COMMUN.2006.002.

[19] J. Galtung and M. Holmboe Ruge, The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers, *Journal of Peace Research* **2**(1) (1965), 64–90. doi:10.1177/002234336500200104.

[20] V. García-Perdomo, R. Salaverría, D.K. Kilgo and S. Harlow, To share or not to share: The influence of news values and popular social media content in the United States, Brazil, and Argentinia, *Journalism Studies* **19**(8) (2018), 1180–1201. doi:10.1080/1461670X.2016.1265896.

[21] E. Grieco and E. Shearer, [n.d.], Americans Are Wary of the Role Social Media Sites Play in Delivering the News, *Pew Research Center, Journalism & Media*, ([n.d.]), https://www.journalism.org/2019/10/02/americans-are-wary-of-the-role-social-media-sites-play-in-delivering-the-news, Accessed 13-March-2020.

[22] R.E. Guadagno, D.M. Rempala, S. Murphy and B.M. Okdie, What makes a video go viral? An analysis of emotional contagion and Internet memes, *Computers in Human Behavior* **29**(6) (2013), 2312–2319. doi:10.1016/j.chb.2013.04.016.

[23] N. Hagar and N. Diakopoulos, Optimizing content with A/B headline testing: Changing newsroom practices, *Media and Communication* **7**(1) (2019), 117. doi:10.17645/mac.v7i1.1801.

[24] A. Halterman, Mordecai: Full text geoparsing and event geocoding, *The Journal of Open Source Software* **2** (2017). doi:10.21105/joss.00091.

[25] T. Harcup and D. O'neill, What is news? News values revisited (again), *Journalism Studies* **18**(12) (2017), 1470–1488. doi:10.1080/1461670X.2016.1150193.

[26] M. Honnibal and I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, (2017), to appear.

[27] C.J. Hutto and E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[28] Y. Jin, J. Huang and X. Wang, What Influences Content Popularity? An Empirical Investigation of Voting in Social Q&A Communities, in: *PACIS*, 161, 2017.

[29] M. Karlsson, Rituals of transparency: Evaluating online news outlets' uses of transparency rituals in the United States, United Kingdom and Sweden, *Journalism Studies* **11**(4) (2010), 535–545. doi:10.1080/14616701003638400.

[30] Y. Keneshloo, S. Wang, E.-H. Han and N. Ramakrishnan, Predicting the popularity of news articles, in: *Proceedings of the 2016 SIAM International Conference on Data Mining*, SIAM, 2016, pp. 441–449. doi:10.1137/1.9781611974348.50.

[31] H.M. Kepplinger, *News Values. the International Encyclopedia of Communication*, 2008.

[32] D.K. Kilgo, K. Lough and M.J. Riedl, Emotional appeals and news values as factors of shareworthiness in Ice Bucket Challenge coverage, *Digital Journalism* (2017). doi:10.1080/21670811.2017.1387501.
[33] J.P. Kincaid, R.P. Fishburne Jr., R.L. Rogers and B.S. Chissom, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, 1975.
[34] D.L. Lasorsa, S.C. Lewis and A.E. Holton, Normalizing Twitter: Journalism practice in an emerging communication space, *Journalism Studies* **13**(1) (2012), 19–36. doi:10.1080/1461670X.2011.571825.
[35] A.M. Lee, S.C. Lewis and M. Powers, Audience clicks and news placement: A study of time-lagged influence in online journalism, *Communication Research* **41**(4) (2014), 505–530. doi:10.1177/0093650212467031.
[36] K. Lopuhin and M. Korobov, ELI5 is a Python package which helps to debug machine learning classifiers and explain their predictions. 2016–2017, to appear.
[37] A. Louis and A. Nenkova, What makes writing great? First experiments on article quality prediction in the science journalism domain, *Transactions of the Association for Computational Linguistics* **1** (2013), 341–352. doi:10.1162/tacl_a_00232.
[38] S.R. Maier, P. Slovic and M. Mayorga, Reader reaction to news of mass suffering: Assessing the influence of story form and emotional response, *Journalism* **18**(8) (2017), 1011–1029. doi:10.1177/1464884916663597.
[39] Medium.com, [n.d.], Medium's Curation Guidelines: Everything writers need to know. https://help.medium.com/hc/en-us/articles/360006362473-Medium-s-Curation-Guidelines-everything-writers-need-to-know, Accessed: 2020-04-09.
[40] P.M. Napoli, Social media and the public interest: Governance of news platforms in the realm of individual and algorithmic gatekeepers, *Telecommunications Policy* **39**(9) (2015), 751–760. doi:10.1016/j.telpol.2014.12.003.
[41] E. Nechushtai and S.C. Lewis, What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations, *Computers in Human Behavior* **90** (2019), 298–307. doi:10.1016/j.chb.2018.07.043.
[42] N. Newman, R. Fletcher, A. Kalogeropoulos and R. Nielsen, *Reuters Institute Digital News Report 2019*, Vol. 2019, 2019, Reuters Institute for the Study of Journalism.
[43] M. Nissim and V. Patti, Semantic aspects in sentiment analysis, in: *Sentiment Analysis in Social Networks*, Elsevier, 2017, pp. 31–48. doi:10.1016/B978-0-12-804412-4.00003-6.
[44] W. Nöth, Self-reference in the media: The semiotic framework, *Self-reference in the Media* **2007** (2007), 3–30.
[45] A. Oeldorf-Hirsch, The role of engagement in learning from active and incidental news exposure on social media, *Mass Communication and Society* **21**(2) (2018), 225–247. doi:10.1080/15205436.2017.1384022.
[46] A. Okrent, The listicle as literary form, *University of Chicago Magazine* **106**(3) (2014), 52–53.
[47] C. Orellana-Rodriguez and M.T. Keane, Attention to news and its dissemination on Twitter: A survey, *Computer Science Review* **29** (2018), 74–94. doi:10.1016/j.cosrev.2018.07.001.
[48] P.A.M. Oude Ophuis and H.C.M. Van Trijp, Perceived quality: A market driven and consumer oriented approach, *Food Quality and Preference* **6**(3) (1995), 177–183. doi:10.1016/0950-3293(94)00028-T.
[49] E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You*, Penguin, UK, 2011.
[50] D. Park, S. Sachar, N. Diakopoulos and N. Elmqvist, Supporting comment moderators in identifying high quality online news comments, in: *Conference on Human Factors in Computing Systems – Proceedings*, 2016, pp. 1114–1125. doi:10.1145/2858036.2858389.
[51] S.A. Paul, L. Hong and E.H. Chi, Who is authoritative? understanding reputation mechanisms in quora, 2012, arXiv preprint arXiv:1204.3724.
[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12** (2011), 2825–2830.
[53] T. Quandt, News on the World Wide Web? A comparative content analysis of online news in Europe and the United States, *Journalism Studies* **9**(5) (2008), 717–738. doi:10.1080/14616700802207664.
[54] Reddit, [n.d.], Reddit. https://www.reddit.com, Accessed: 2020-04-09.
[55] J.G. Robinson, [n.d.], Tow Center for Digital Journalism, https://www.cjr.org/tow_center_reports/how-journalists-imagine-their-readers.php, Accessed: 2020-01-07.
[56] M.M.U. Rony, N. Hassan and M. Yousuf, Diving deep into clickbaits: Who use them to what extents in which topics with what effects?, in: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ACM, 2017, pp. 232–239.
[57] J.M. Scacco and A. Muddiman, Investigating the influence of "clickbait" news headlines, *Engaging News Project Report*, 2016.
[58] M. Schoonvelde, A. Brosius, G. Schumacher and B.N. Bakker, Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches, *PloS One* **14**(2) (2019), e0208450. doi:10.1371/journal.pone.0208450.
[59] J.B. Singer, User-generated visibility: Secondary gatekeeping in a shared media space, *New Media & Society* **16**(1) (2014), 55–73. doi:10.1177/1461444813477833.
[60] S. Soroka and S. McAdams, News, politics, and negativity, *Political Communication* **32**(1) (2015), 1–22. doi:10.1080/10584609.2014.881942.
[61] C. Sotirakou, P. Germanakos, A. Holzinger and C. Mourlas, Feedback matters! Predicting the appreciation of online articles a data-driven approach, in: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Springer, 2018, pp. 147–159. doi:10.1007/978-3-319-99740-7_10.
[62] C. Sotirakou, D. Trilling, P. Germanakos and C. Mourlas, Opening the black box of perceived quality: Predicting endorsement on a blog site, in: *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, IEEE, 2019, pp. 388–392.

[63] E.C. Tandoc and R.J. Thomas, The ethics of web analytics: Implications of using audience metrics in news construction, *Digital Journalism* **3**(2) (2015), 243–258. doi:10.1080/21670811.2014.909122.

[64] K. Thorson and C. Wells, Curated flows: A framework for mapping media exposure in the digital age, *Communication Theory* **26**(3) (2016), 309–328. doi:10.1111/comt.12087.

[65] D. Trilling, P. Tolochko and B. Burscher, From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics, *Journalism & Mass Communication Quarterly* **94**(1) (2017), 38–60. doi:10.1177/1077699016654682.

[66] J. Urban and W. Schweiger, News quality from the recipients' perspective: Investigating recipients' ability to judge the normative quality of news, *Journalism Studies* **15**(6) (2014), 821–840. doi:10.1080/1461670X.2013.856670.

[67] Valarie and A. Zeithaml, Consumer perceptions of price, quality, and value: A means-end model and synthesis of evidence, *Journal of Marketing* **52**(3) (1988), 2–22. doi:10.1177/002224298805200302.

[68] S. Valenzuela, M. Piña and J. Ramírez, Behavioral effects of framing on social media users: How conflict, economic, human interest, and morality frames drive news sharing, *Journal of Communication* **67**(5) (2017), 803–826. doi:10.1111/jcom.12325.

[69] B. Wasike, Preaching to the choir? An analysis of newspaper readability vis-a-vis public literacy, *Journalism* **19**(11) (2018), 1570–1587. doi:10.1177/1464884916673387.

[70] P. Weber, Discussions in the comments section: Factors influencing participation and interactivity in online newspapers' reader comments, *New Media & Society* **16**(6) (2014), 941–957. doi:10.1177/1461444813495165.

[71] K. Welbers, W. van Atteveldt, J. Kleinnijenhuis, N. Ruigrok and J. Schaper, News selection criteria in the digital age: Professional norms versus online audience metrics, *Journalism: Theory, Practice & Criticism* **17**(8) (2016), 1037–1053. doi:10.1177/1464884915595474.