Feedback Matters! Predicting the Appreciation of Online Articles A Data-Driven Approach

Catherine Sotirakou, Panagiotis Germanakos, Andreas Holzinger, Constantinos Mourlas

Feedback Matters! Predicting the Appreciation of Online Articles
A Data-Driven Approach

Catherine Sotirakou[1], Panagiotis Germanakos[2], Andreas Holzinger[3], Constantinos Mourlas[1]

[1]Faculty of Communication and Media Studies, National & Kapodistrian University of Athens, Sofokleous 1, 10559, Athens, Greece
{katerinasot,mourlas}@media.uoa.gr,
[2]Products & Innovation, SAP SE, Dietmar-Hopp-Allee 16, 69190, Walldorf, Germany,
panagiotis.germanakos@sap.com,
[3]Institute of Medical Informatics, Statistics and Documentation (IMI), Medical University Graz
Auenbruggerplatz 2, A-8036 Graz, Austria
andreas.holzinger@medunigraz.at

Abstract. The current era of advanced computational mobile systems, continuous connectivity and multi-variate data has led to the deployment of rich information settings that generate constant and close to real-time feedback. Journalists and authors of articles in the area of Data Journalism have only recently acknowledged the influence that the audience reactions and opinions can bring to effective writing, so to be widely appreciated. Such feedback may be obtained using specific metrics that describe the user behavior during the interaction process like shares, comments, likes, claps, recommendations, or even with the use of specialized mechanisms like mood meters that display certain emotions of readers they experience while reading a story. However, which characteristics can reveal an article's character or type in relation to the collected data and the audience reflection to the benefit of the author? In this paper, we investigate the relationships between the characteristics of an article like structure, style of speech, sentiment, author's popularity, and its success (number of claps) by employing natural language processing techniques. We highlight the emotions and polarity communicated by an article liable to increase the prediction regarding its acceptability by the audience.

Keywords: Data Journalism, Natural Language Processing, Sentiment, Emotions, News Articles, Computer-Assisted Content Analysis, Machine Learning

1 Introduction

In recent years, the advancement of computational systems and devices, along with the explosive growth and availability of open data have led to computational journalism' growth. Now, data journalists use in their news preparation and writing, software and

technologies which are found in the cross-borders of three different research disciplines: Computer Science, Social Sciences, and Media & Communications. According to [12] science and journalism were two fields that coexisted for the first time in Philipp Meyer's book [16] Precision Journalism, where he explained that the implementation of scientific methods used in social science experiments when used in reporting, produce very good results in both the news-gathering process and the final story. While computation has long been assisting journalists in different phases of the news production process, and what is known as computer-assisted reporting has been practiced since the 1960s [10], the scientific investigation of it has been more rigorously undertaken the last decade. According to an extensive literature review from Ausserhofer [4], a significant research output on data journalism and related fields has been produced only since 2010, with a prior small number of attempts to be regarded as isolated. Researchers and journalists usually study ways and explore methods to (a) locate and extract useful data (like information-rich textual data on the internet), (b) analyze and understand the meaning of massive datasets (i.e., trying to draw connections between the sender and the recipient of an email using email meta-characteristics), and (c) acknowledge more thoroughly the user feedback (Twitter data analysis) in their analysis and articles composition.

In our research, we mainly focus on the third research stream which refers to the *what* and *how* to incorporate the audience feedback in judging both the quality and the character of an article from authors' point of view facilitating their writing. The central motivational factor is the mind shift which has been observed lately regarding both the online news consumption and journalists' tendency to learn more about their audiences [1, 24]. Authors and editors used to ignore the audience feedback [5, 14]) relying mainly on personal predictions what they thought the audience desires might be or as Tandoc [33] suggests "they substituted their own preferences for those of their actual audience". According to the literature one of the reasons for this reluctant attitude towards the audience feedback was the fact that journalists should follow clear journalistic norms and remain autonomous without being affected by external preferences that might erode journalistic quality [5]. However, the explosion of big data and information technologies revolutionized the way that readers provide feedback to journalists (reviews, letters to the editor). Today, newsrooms have Web analytics tools that provide insights such as clicks, page views, viewing time, conversion funnels analysis and user flows that are only some of the audience metrics that help authors monitor how people interact with the online content on their newsroom's website. Accordingly, Natural Language Processing (NLP) techniques are widely applied in sentiment analysis so to determine which features users are particularly in favor of or to trace content that needs to be removed (in case of negative feedback).

Our main concern is not simply to employ data-driven techniques predicting the probability for an article to gain more likes, claps, or recommendations but rather to go a step further and embrace a more human-centered approach, to understand the characteristics that make a great article based on the motivation and the scope that the journalist wants to communicate. This information will help on enriching an *a priori* knowledge that could be utilized as one more dimension and guideline, contributing to the success of the potential next article. For the scope of this paper, we highlight the

emotions and polarity that an article communicates in relation to other well-recognized characteristics of article's style and author's reputation. We tackle the problem of finding the interesting relations between the characteristics of an article namely the (i) article's structure, (ii) style of speech, (iii) sentiment (iv) author's popularity, and its success (number of claps), assuming that for online articles posted on a blogging platform it is possible to establish a predictive link between several content-based facets and success. We use several NLP and machine learning methods for cleaning, filtering, and processing text data like measuring term frequency and lexicon based approaches, as well as for checking the importance of each factor, and also to evaluate the predictive power of our model. Our preliminary results (from a sample of 3000 articles extracted from the Medium online publishing platform in 2017) has shown that indeed characteristics of the user, emotions expressed in the text, personal tone of speech and the use of uncommon words influence the prediction of the results regarding the level of acceptability by the audience.

2 Related Work

In this section, we describe the distinctive attributes of the text as determinants of articles' quality and acceptability by the broader audience. We emphasize on previous researches that consider sentiment, structure and writing style since those factors have been used to verify our experiments. Journalists write stories hoping that their writing will arouse emotions in audiences, sometimes positive like hope, joy or trust and others negative such as fear and anger depending on the given coverage. The emotional state of the online reader is a key factor for authors to understand the audiences' reactions to the news stories [6], e.g. know how the reader would respond to their article, if they manage to express the desired emotion correctly on their writing, etc. The digital age offers a constantly increasing number of sentiment-rich resources like news articles available on the Web while technology nowadays allows readers to leave feedback and show their appreciation easily. Such an appreciation is usually demonstrated through various actions like sharing e.g., an article or post of interest with the community or with the targeted audience, ranking the content based on the quality or interest it might present or following the owner of the content.

According to literature, emotional online news content especially awe-inspiring content is more likely to become viral [6]. Different NLP methods that are trying to capture emotional categories, sentiment polarity, and emotional dimensions [25] have been proposed by researchers to extract the emotions expressed in the texts. Intelligent natural language processing tasks are required to be more sophisticated to improve their accuracy, thus a variety of sentiment and emotion lexica and corpora [11, 21, 32,] have been created that can identify a plethora of emotions instead of just suggesting whether they express positive, negative or neutral sentiment. Indicators of collective user behavior and opinion are increasingly common features of online news stories and may include information about how the story made the readers feel. Typical example of such features is Facebook's set of emoticons called "Reactions", that urges users to express

their experienced emotions about a post by using a button that includes five different emotional states: Love, Haha, Wow, Sad, Angry [8, 34].

Other approaches have focused on determining what are the characteristics of a good structure for a news article regarding text coherence. As Louis & Nenkova, [18] have previously categorized in their work, there are three ways of testing text coherence "by exploring systematic lexical patterns, entity coreference and discourse relations from large collections of texts". In terms of measuring the quality of a given article as a whole, recent work in the field has decomposed news articles into a set of simpler characteristics that reveal different linguistic and narrative aspects of online news [2]. Arapakis and his colleagues after discussing with several journalists, editors and computational linguists, proposed "a multidimensional representation of quality", taken from the editor's perspective that groups the attributes of a news story into five categories: Readability, informativeness, style, topic, and sentiment, with each category having several sub-categories such as fluency and conciseness for the readability and subjectivity, sentimentality, and polarity for sentiment. Their findings suggest that the journalists' perception of a well-written story correlates positively with fluency, richness (feature from the category style) and completeness (feature from the category informativeness), while aspects like subjectivity and polarity proved to be weakly correlated. Therefore, this particular work suggests that sentiment is not of great importance when it comes to article quality, whereas text comprehension and writing style seem to be determinants of quality.

To examine the potential effects of emotions, writing style and readability on article quality prediction in the journalism domain, researchers Louis & Nenkova [19], investigated science articles from the New York Times, by separating them into two categories "very good", in which articles from the authors whose writing appeared in "The Best American Science Writing" anthology series were included, and the "typical" category that included all the remaining articles. Their experiments showed that "excellent authors associated with greater degree of sentiment, and deeper study of the research problem" as well as usage of what is called beautiful language, meaning the unusual phrasing. This approach is similar to ours, however, this study explored only science articles from the New York Times, which we already know that are examples of good quality journalism. In our approach, we consider articles from a broad spectrum, written not only by professional journalists but mostly by amateur writers.

The scientific community also has been experimenting with predictive analytics [7, 20, 31]. In their work McKeown, et al [20] present a system that predicts the future impact of a scientific concept, based on the information available from recently published research articles, while Sawyer et al [31] suggest that award-winning academic papers use simple phrasing. However, a successful news story would be described by alternative characteristics in contrast to good academic writing. In another work, Ashok, Feng, & Choi [3] used statistical models to predict the success of a novel based on its stylistic elements. Their work examined a collection of books and movie scripts of different genres, providing insights into the writing style, such as lexical and syntactic rules, sentiment, connotation, and distribution of word categories and constituents commonly shared among the best sellers. The results suggest that successful writing includes more complex styling features like prepositions, pronouns, and adjectives

while unsuccessful novels have more verbs, adverbs, and foreign words, thus having higher readability. Moreover, this work found that the writing style of successful novels is similar to news articles.

Still, very little research is available on preprocessing noisy texts, which is usually done, particularly by large companies (e.g. Facebook) manually to identify and correct spelling errors, or other noise (whitespaces, boundaries, punctuation) [36]. This field is a topic of research for quite a long time, but the effects of cleaning text passages is still rarely described [37]. The problem of noise is underestimated within the machine learning community, but has serious consequences for language identification, tokenization, POS tagging and named entity recognition.

Finally, there are many works today that have applied in different cases for textual analysis factors like animate and personal pronouns, use of visual words, people-oriented content, use of beautiful language, sub-genres, sentiment, and the depth of research description [19] lexical choices such as thinking and action verbs [3]. Readability is also a significant factor, in their work [30] explored features like word and sentence length, cohesion scores and syntactic estimates of complexity. Their results showed that increased use of verb phrases provides better readability. In our current work we use those factors in combination since we believe can determine the quality and acceptance of an article regarding structure, style, author's popularity and emotionality.

3 Method and Dataset

We employ a data-driven approach aiming to respond to two typical related research questions: (a) To what extent the character of an article can reveal the reaction of the readers, producing more or fewer claps? and (b) would it be of importance to predict the character of an article to the benefit of the journalist? We investigate textual data from online articles on the Web so to help the authors to adjust their writing style, gaining more acceptance from their audience. We extract content-based features from the articles on the online publishing platform, Medium[1], based on relevant metrics concerning those proposed in the reviewed literature, such as the use of beautiful language, the tone of speech, polarity and sentiment, and genre.

Before we begin to delve into the effect of the different characteristics of an article to its success, we first need to extract features using text analysis techniques related to the four different aspects of an article that we suggest, namely the (i) article's structure, (ii) style of speech, (iii) sentiment (iv) author's popularity. We use several NLP methods for cleaning, filtering, and processing text data like measuring term frequency and lexicon based approaches. After defining the above categories we need to study their importance on the article's success, applying machine learning algorithms. A random forest classifier is used to evaluate the significance of the features above in the success of an article as well as to create a decision tree able to predict whether the claps count of a given article will be high, medium or low. We run experiments on a mixed-topic

[1] https://medium.com

dataset of over 3 thousand articles published at 2017 and downloaded at the beginning of 2018. The articles had a large distribution of claps ranging from 84 to 157K claps.

As we mentioned earlier, for this research we collected data from an online publishing platform called “Medium”, that hosts a variety of articles and publications produced by either amateur or professional writers, journalists, bloggers, companies, and range from short to long articles, with topics that cover a variety of topics such as science, education, politics, well-being etc. Medium provides to the reader an automatically calculated display of the reading time on every article so they can know how much is required of them to read through an entire story. Like votes on Digg[2] stories, “claps” (formerly called “Recommend”) on Medium represent whether readers liked the story or not and would recommend it to other users on the platform (a user can clap more than once). In the world of Medium, the success of an article is measured regarding claps count, which is the number of times the readers have been clapped.

In this paper, we suggest that a more comprehensive understanding of the content of an article should comply with four directions, as content specification elements. Below we describe in more detail what each one represents along with the method and resources used.

3.1 Content Structure

Good structure is a key to higher quality articles [2, 3]. In our model, we use five characteristics of structure: genre, title words, reading time (since it is predetermined by the Medium), the proportion of images in the text and proportion of bullet points in the article.

Genre: The topic of the article reflect certain characteristics of its nature and it has been greatly investigated in previous work [2, 3, 19]

Title: The title of an article is a significant aspect of every story and is also a clickbait strategy that is being used a lot in journalism [13, 35]

Reading time: every article posted on Medium has its length measured automatically, by the time it would take the user to read it.

Images: The use of a great number of beautiful images is one of the best practices used in social media marketing.

Bullet points: According to research the extensive use of the so-called listicle; a mixture of 'list' and 'article' is an interesting phenomenon and its power lies “not only in the power of the format in and by itself but also in 'shareable factors' that are related to the individual listicle” [27].

3.2 Style

There are multiple aspects that characterize the style of speech, and stylistic variation reflects largely the meaning that is communicated [28]. Many researchers have examined this dimensions of writing style by measuring features like formal language [17, 28, 29], attractiveness and richness [2], and use of beautiful language that refers to

[2] http://digg.com

words with lowest frequency or with highest perplexity under the phoneme and letter models [19]. Except for the above style-related dimensions we propose the personal tone of speech, typically examined in communication research and more specifically in political speeches.

Tone of speech: Communication researchers have studied the self-referential or self-reflexive messages that the media and advertising companies use to attract the audience's attention [26]. Politicians like Barack Obama also have recognized the persuasive power of personal pronouns and they use it strategically to their rhetoric [23]. Moreover, personal pronouns are an indicator of the existence of people in the story and prior works have experimented with human-related stories [19] as a factor of success. In our study, we investigate the use of personal pronouns (I, you, we etc), reflexive pronouns like myself, yourself, and possessive pronouns (mine, yours, etc).

Beautiful phrasing: Using beautiful phrases and creative words can amuse the audience and according to findings of Louis and Nenkova, [19] they are discovered in high-quality articles. We apply Term Frequency Inverse Document Frequency (TF-IDF) to determine what words in the corpus of articles might be rarer and thus favorable to use in our model.

3.3 Sentiment

Emotions expressed in an article can motivate people to leave feedback, share and rate the content [3, 6, 19]. Theories in the nature of emotion, like Plutchik's model, suggest the existence of basic emotions such as joy, sadness, anger, fear, trust, surprise, disgust, and anticipation [25]. For this study we used the resource EmoLex11, from the NRC suite of lexica, that is based on Plutchik theory, to extract the sentiment and polarity expressed in the articles. This word-emotion association lexicon [22] contains 14,182 words labeled according to emotions and also includes annotations for negative and positive sentiments. We investigate four affect-related features: the density of both positive and negative emotions, the density of each emotion (joy, sadness, anger, fear, trust, surprise, disgust, and anticipation), emotion expressed in the title, and polarity. We compute the counts of emotion words, each normalized by the total number of article words, and the total count of all the emotion words both negative and positive.

3.4 Author's Popularity

Popularity indicates the total followers -potential readers- of an author on Medium and usually is used in social media network analysis, where both features like followers and following (users that a given user is following) play a crucial role on the user's position in the social network.

In preparation for the analysis, we further "cleaned" the dataset of 3030 articles by removing all the NaN values, html code, foreign languages and end up with 2990 useful articles. Furthermore, we stemmed the texts and dropped all the stop words and non-standard words and characters, such as punctuations, spaces, special characters and urls.

Note that our feature computation step is not tuned for the quality prediction task in any way. Rather we aim to represent each facet as accurately as possible. Ideally we

would require manual annotations for each facet (visual, sentiment nature etc.) to achieve this goal. At this time, we simply check some chosen features' values on a random collection of snippets from our corpus and check if they behave as intended without resorting to these annotations.

4 Analysis and Results

We started by converting the numerical variable that represents the number of claps into categorical one, having three values representing low, medium and high acceptance. Suitable python libraries were imported to automatically convert the numerical variables that represent claps in the three different groups, and thus get a quick acceptance segmentation by binning the numerical variable in groups.

Table 1. Ordered list of the importance of selected features affecting the acceptance of an article.

Feature	Importance
Following	0.094
Followers	0.071
Beautiful Phrasing (rare words)	0.060
Read Time (length)	0.052
Tone of Speech (self-reference)	0.052
Positivity	0.048
Surprise	0.047
Time passed (since publication date)	0.047
Sadness	0.046
Joy	0.046
Disgust	0.044
Anger	0.044
Anticipation	0.044
Density	0.042
Trust	0.042

Our next objective becomes now to develop a news classifier and study the effect of the selection of our predictor variables on the performance of the acceptance prediction model. The *randomForest* package from *sklearn* library in Python was used to create a classifier of our articles and additionally to measure the importance of the predictor variables. The importance of a variable is computed internally during the construction of the decision trees by checking the increase of prediction error when data for that variable is permuted while all others are left unchanged. In the case of Random Forest,

to determine the importance of the predictor variables, we calculated the Gini index for each of them. After that, we ordered from highest to lowest rate and kept the 15 most important variables out of 44.

The table above (see Table 1) presents the ordered list of the importance of the variables of the selected categories, such as authors' popularity, beautiful phrasing etc., starting from the most important variables that affect the acceptance of an article. A graphical representation is also depicted in Fig.1.

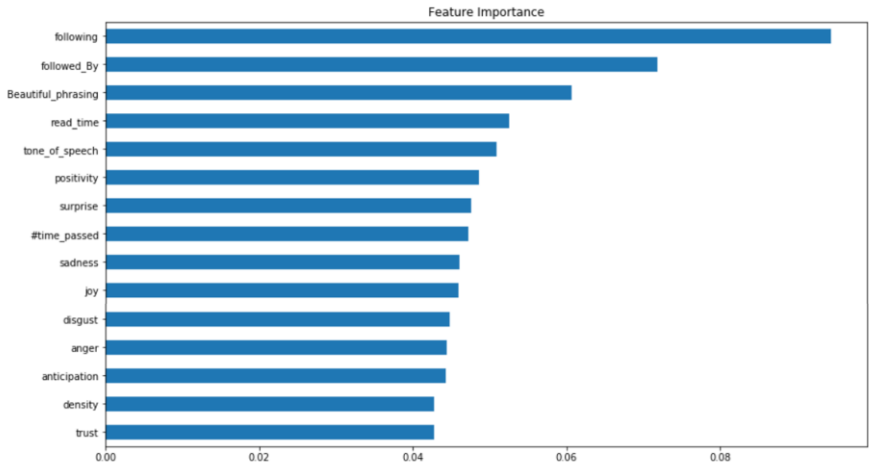

Fig. 1. Graphical representation of the importance of the features.

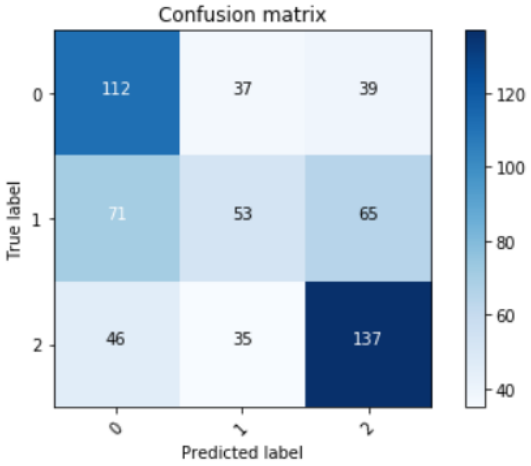

Fig. 2 Graphical representation of the Confusion Matrix.

For prediction purposes we split the dataset into two sets, one containing 80% of the total articles that was used as the training set and the other 20% was used to test the classifier so that the model can be trained and tested on different data. We run the model on the train data to construct the confusion matrix and compare the predictions with the observations in the validation data set, which is as we said different from the one used to build the model.

The labels above represent the low, medium and high numbers of claps according to the acceptance segmentation process we discussed earlier by binning the numerical variable into three groups. Thus, bin 0 represents the low acceptance of the users while bin 2 represents the group of articles that have a high number of claps. The bin 1 represents the intermediate state (see Table 2).

Table 2. The three groups (high, medium, low) of clap numbers.

Bucket	Start	End
0	87.0	1500.0
1	1500.0	5400.0
2	5400.0	55000.0

Studying the confusion matrix (see Fig. 2) we can see that the predictive power of our model is mainly focused on the:

- Label 0 representing articles having a very low number of claps (between 87 and 1500) and,
- Label 2 representing the famous articles (claps between 5400 and 55000).

The model seems to have a poor performance on the intermediate state for articles where the number of claps is between 1500 and 5400. The predictive power for the 0-label and 2-label is 63%.

Our experiments revealed several interesting insights. Our empirical results indicate that the following relationships between Medium users are the most important factors in predicting the audience's positive feedback, thus the author's readership depends highly on his online followers. This is a logical consequence because a greater number of followers results in a larger potential audience that is likely to leave positive feedback, especially if they already like the author's previous work.

We observed that our results are in line with the related literature, such as the power of self-reference in communicating a message and the fact that good writers are logophiles and prefer to use uncommon and long words in their appealing writing. Our findings suggest that making unusual word choices in writing might result in better feedback and accord with previous studies that proved the adults' preference in unfamiliar words to the everyday ones [9].

News articles are not only anonymously edited information but are essentially narrated, and according to Grunwald [15], they are "constructed by a personally involved, individual journalist performing a role as an engaged narrator using a variation

of communication acts" hoping to achieve a reliable and interesting deliverance of the message. Likewise, in our experiments, the use of personal tone of speech is highly correlated with success.

Another finding was that the only structure-based feature that correlates with success is article's length, while aspects such as genre, the number of title words, images, and bullet points present a poor correlation with positive feedback.

Moreover, our hypothesis that sentiment is of great importance is proven by our model, with aspects such as positivity and the emotion of surprise performing better than anger and disgust. Also, emotional density is an important feature for predicting the article's success in contrast to the aspect of polarity.

Finally, it is widely agreed that time factors like publishing date can change the audience appreciation depending on the platform's popularity over a certain period. We ran experiments where we incorporated the time factors into the model to accurately capture the user clapping over time, and the results were surprisingly accurate. It seems that platform's popularity in time is particularly influential factor in the accuracy of a model and will be further investigated in the future.

5 Conclusions

Nowadays, we are living in an era of rich information settings that generate constant and real-time feedback. In the area of Data Journalism, we observe a mind shift of journalists and authors towards extracting and learning more about their audiences. On the other hand, users-readers need to interact with a vast amount of contents residing in a variety of resources so to find what they are looking for. Especially, in the case of online articles' consumption such a reality makes them more selective and critical on which ones to read and stories to appreciate. This necessitates the consideration of the audience reflection and opinions in the composition of articles if we expect that they will meet their purpose and will be successful. Therefore, the main challenge is how can we figure out the character of the article based on the readers' feedback? Predicting the reaction of the audience toward a story has become of considerable importance not only for researchers and scientists but also for media organizations and digital managers that heavily invest in software to gain insights into readers behavior.

In this paper, we propose a model to discover the characteristics of online articles that result in greater audience acceptance. Our model relates to four different dimensions of articles' characteristics namely, article's structure, style of speech, sentiment and author's popularity. We applied several NLP and machine learning algorithms to extract a consensus, in terms of features' importance and prediction on claps, about the data derived from the online publishing platform "Medium". From our experiments we can formulate a preliminary understanding that several attributes characterize online article's success. Our findings demonstrate that indeed characteristics of the user, emotions expressed in the text, personal tone of speech and the use of uncommon words are highly correlated with influence of acceptance by the audience.

In the future, we plan to focus on the development of more articles' characteristics and examine inherent correlations aiming at optimizing and further improving the prediction of our model. We also plan to adopt a human-centred viewpoint on the interpretation of the features, and their subsequent relationships, in an attempt to identify a weighted impact on the final result. Such a finding might trigger a deeper understanding on the requirements of a successful article based on the reactions of the audience so to be used as guidelines for the journalists and authors facilitating their success.

Acknowledgements. This work is supported by the General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (H.F.R.I.) in the context of the action '1st Proclamation of Scholarships from ELIDEK for PhD Candidates'.

References

1. Anderson, C. W.: Between creative and quantified audiences: Web metrics and changing patterns of newswork in local US newsrooms. Journalism, 12(5), 550-566 (2011).
2. Arapakis, I., Peleja, F., Berkant, B., & Magalhaes, J.: 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, pp. 1893-1902) (2016).
3. Ashok, V. G., Feng, S., & Choi, Y.: Success with style: Using writing style to predict the success of novels. In Proceedings of the 2013 conference on empirical methods in natural language processing, pp. 1753-1764 (2013).
4. Ausserhofer, J., Gutounig, R., Oppermann, M., Matiasek, S., & Goldgruber, E.: The datafication of data journalism scholarship: Focal points, methods, and research propositions for the investigation of data-intensive newswork. Journalism, 1464884917700667 (2017).
5. Beam, R. A.: How newspapers use readership research. Newspaper Research Journal, 16(2), 28-38 (1995).
6. Berger, J., & Milkman, K. L.: What makes online content viral?. Journal of marketing research, 49(2), 192-205 (2012).
7. Bergsma, S., Post, M., & Yarowsky, D.: Stylometric analysis of scientific articles. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 327-337). Association for Computational Linguistics (2012).
8. Chaykowski, K.: Facebook No Longer Just Has A 'Like' Button, Thanks To Global Launch Of Emoji 'Reactions', Forbes article: https://www.forbes.com/sites/kathleenchaykowski/2016/02/24/facebook-no-longer-just-has-a-like-button-thanks-to-global-launch-of-emoji-reactions/#29919a54692d.(2016) last accessed 2018/03/04.
9. Colman, A. M., Walley, M., & Sluckin, W.: Preferences for common words, uncommon words and non-words by children and young adults. British Journal of Psychology, 66(4), 481-486. (1975).
10. Cox, M.: The development of computer-assisted reporting. Informe presentado en Association for Education in Jornalism end Mass Comunication). Chapel Hill, EEUU: Universidad de Carolina del Norte (2000).

11. de Albornoz, J. C., Plaza, L., & Gervás, P.: SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In LREC, pp. 3562-3567 (2012).
12. Flew, T., Spurgeon, C., Daniel, A., & Swift, A.: The promise of computational journalism. Journalism Practice, 6(2), 157-171 (2012).
13. Frampton, B.: Clickbait: The changing face of online journalism. BBC News, 14 (2015).
14. Gans, H. J.: Deciding what's news: A study of CBS evening news, NBC nightly news, Newsweek, and Time. Northwestern University Press (1979).
15. Grunwald, E: Narrative norms in written news. Nordicom Review, 26(1), 63-79 (2005).
16. Journalism, Precision. "A Reporter's Introduction to Social Science Methods." (1973).
17. Lahiri, S., Mitra, P., & Lu, X.: Informality judgment at sentence level and experiments with formality score. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 446-457). Springer, Berlin, Heidelberg (2011).
18. Louis, A., & Nenkova, A.: A corpus of science journalism for analyzing writing quality. Dialogue & Discourse, 4(2), 87-117 (2013b).
19. Louis, A., & Nenkova, A.: What makes writing great? First experiments on article quality prediction in the science journalism domain. Transactions of the Association for Computational Linguistics, 1, 341-352 (2013a)
20. McKeown, K., Daume, H., Chaturvedi, S., Paparrizos, J., Thadani, K., Barrio, P., ... & Gravano, L.: Predicting the impact of scientific concepts using full-text features. Journal of the Association for Information Science and Technology, 67(11), 2684-2696 (2016).
21. Mohammad, S. M., & Turney, P. D.: Crowdsourcing a word–emotion association lexicon. Computational Intelligence, 29(3), 436-465 (2013).
22. Mohammad, S. M., & Turney, P. D.: Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, *29*(3), 436-465 (2013).
23. Nakaggwe, L.: The persuasive power of personal pronouns in Barack Obama's rhetoric (2012).
24. Napoli, P. M.: Audience evolution: New technologies and the transformation of media audiences. Columbia University Press (2011).
25. Nissim, M., & Patti, V.: Semantic aspects in sentiment analysis. In Sentiment analysis in social networks, pp. 31-48 (2017).
26. Noth, W.: Self-reference in the media: The semiotic framework. *Self-Reference in the Media*, *New York: Mouton de Gruyter*, 3-30 (2007).
27. Okrent, A.: The listicle as literary form. University of Chicago Magazine, 106(3), 52-53 (2014).
28. Pavlick, E., & Nenkova, A.: Inducing lexical style properties for paraphrase and genre differentiation. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 218-224) (2015).
29. Pavlick, E., & Tetreault, J.: An empirical analysis of formality in online communication. *Transactions of the Association of Computational Linguistics*, *4*(1), 61-74 (2016).
30. Pitler, E., & Nenkova, A.: Revisiting readability: A unified framework for predicting text quality. In Proceedings of the conference on empirical methods in natural language processing (pp. 186-195). Association for Computational Linguistics (2008).
31. Sawyer, A. G., Laran, J., & Xu, J.: The readability of marketing journals: Are award-winning articles better written?. Journal of Marketing, 72(1), 108-117 (2008).

32. Strapparava, C., & Mihalcea, R.: Semeval-2007 task 14: Affective text. In Proceedings of the 4th International Workshop on Semantic Evaluations (pp. 70-74). Association for Computational Linguistics (2007).
33. Tandoc Jr, E. C.: Journalism is twerking? How web analytics is changing the process of gatekeeping. New Media & Society, 16(4), 559-575 (2014).
34. Vaiciukynaite, E., Massara, F., & Gatautis, R.: An Investigation on Consumer Sociability Behaviour on Facebook. Engineering Economics, 28(4), 467-474 (2017).
35. Zheng, H. T., Yao, X., Jiang, Y., Xia, S. T., & Xiao, X.: Boost clickbait detection based on user behavior analysis. In Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data (pp. 73-80). Springer, Cham. (2017).
36. Petz, G., Karpowicz, M., Fuerschuss, H., Auinger, A., Stritesky, V. & Holzinger, A. Computational approaches for mining user's opinions on the Web 2.0. Information Processing & Management, 51, (4), 510-519, doi:10.1016/j.ipm.2014.07.011 (2015).
37. Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Winkler, S., Schaller, S. & Holzinger, A. On Text Preprocessing for Opinion Mining Outside of Laboratory Environments. In: Huang, Runhe, Ghorbani, Alia, Pasi, Gabriella, Yamaguchi, Takahira, Yen, Neily & Jin, Beijing (eds.) Active Media Technology, Lecture Notes in Computer Science, LNCS 7669. Berlin Heidelberg: Springer, pp. 618-629, doi:10.1007/978-3-642-35236-2_62 (2012).