Cognitive reinforcement for enhanced post construction aiming fact-check spread

Maria Araújo Barbosa[0000−0002−0464−0991], Francisco S. Marcondes[0000−0002−2221−2261], and Paulo Novais[0000−0002−3549−0754]

ALGORITMI/LASI, University of Minho, Braga, Portugal
pg42844@alunos.uminho.pt, francisco.marcondes@algoritmi.uminho.pt, pjon@di.uminho.pt

Abstract. Despite the success of fact-checking agencies in presenting timely fact-checking reports on the main topics, the same success is not achieved for the dissemination of these reports. This work presents the definition of a set of heuristics applicable to messages (posts) in the microblogging environment, with the aim of increasing their engagement and, consequently, their reach. The proposed heuristics focus on two main tasks: summarisation, emotion-personality reinforcement. The results were evaluated through an experiment conducted with twenty participants, comparing the engagement of actual and generated posts. From the results of the experiment, it can be concluded that the strategy used by the generator is at least better than the one used by the fact-checking journal Snopes in its Twitter posts.

Keywords: Post Generator · Social Networks · Twitter · Fake-news · Fact-checking · Emotions · Personality · Engagement.

1 Introduction

Social media platforms are very popular these days. Millions of users use *posts* to share their thoughts, opinions, news and personal information. In social media studies, the content of posts is often used as a basis for research as it provides insight into public opinion and what people are talking about [8]. Natural language processing (NLP) techniques transform the content of posts into data that can be interpreted by a computer. An example is the work of I. Singh *et al.* [5], which uses a GPT-2 and the plug-and-play language model to incorporate an emotion into the output of a generated text, ensuring grammatical correctness. L. Piccolo *et al.* [16], in turn, demonstrated a multi-agent approach to engaging Twitter users with fact-checkers; the idea is to encourage users to share verified content by educating Twitter users sharing URLs already flagged as fake.

A brief landscape of how fact-check reports are shared on Twitter is available in [9]. In short, considering 2020 data, the behaviour of `@snopes` (joined in 2008, 237.3K followers) is based on sending a fact-check link, but with a new and eventually witty headline for the tweet text. It uses a repackaging and retweeting strategy. `@PolitiFact` (joined in 2007, 673.6K followers), `@Poynter` (2007,

214.8K) and `@factcheckdotorg` (2009, 190.3K) all show similar behaviour, varying according to the style of the operator writing the post. `@Channel4News` (2008, 99.7K) also shows the same behaviour, but participates more in replies and some other "normal" Twitter interactions. Despite this, it has fewer followers than the others, but a higher growth rate. `@APFactCheck` (2011, 31.5K) is slightly more successful than the other agencies; its tweets are more instigative than journalistic, perhaps better suited to Twitter. It should be noted that it has fewer followers, although it is more successful in terms of engagement. Finally, `@MBFC_News` (2015, 4411) and `@TheDispatchFC` (2019, 1277) only tweet the title of the fact check and its link. For an illustration, refer to figure 1.

(a) `@snopes` traditional journalism

(b) `@Poynter` closer to "normal" tweets

(c) `@APFactCheck` provocative content

Fig. 1: Examples of slightly different strategies for creating tweet text

Considering the role of emotion in news [15], together with the empirical results presented in this landscape, it is possible to hypothesize that generating or tailoring microblogging posts to appeal to certain emotions and personality traits are able to improve engagement. Despite such a conception to be widespread, especially in supporting tools for digital marketing, there are few research papers testing analogous hypotheses, fewer considering mental processes [19].

In order to test this hypothesis, this paper builds a prototype post generator whose text is reinforced (or tailored) to specific emotions and mental processes. These prototype generations will be presented to a group of people in a laboratory environment along with real posts for evaluation. To reduce the scope, the prototype is focused on Twitter, and `@snopes` is chosen for comparison.

The paper is organized into two major sections. The first describing the prototype development and the second reporting the assessment and its results.

2 Tweet-Tailoring Heuristics & Prototype

For reference, a tweet is an online publication that typically consists of four different elements: *text-core*, *emoticons*, *hashtags* and *links*. The *text-core* contains the textual part of the post, comprising the relevant content/message. A *emoji* is a pictogram that can be embedded in the *text-core*. *hashtags* are search terms that can also be included in the post, and *links* are URLs to websites.

The source of the information is the Snopes fact-check reports. The information is extracted using a `beautifulsoup` scrapper, the fields searched for are: URL, date, newspaper name, title, claim, classification and the content of the report. The result is stored in a JSON file.

2.1 Summarization

The aim of this section is to obtain the post element *text-core*. To guide this procedure, the following requirements are established:

- Be consistent. Avoid contradictions and odd constructions.
- Be coherent. Have a meaningful flow of text.
- Be informative. Provide context with relevant information.
- Don't produce fake news.
- Have less than 280 characters (to meet Twitter's requirements);

In order to find the most appropriate approach, both extractive and abstractive summarization algorithms were investigated. Four extractive summarization algorithms were evaluated: 1) TextRank [12]; 2) Luhn's Heuristic Method [18]; 3) Selecting the sentence with the highest emotional value for sadness and surprise; and 4) Selecting the text in *allegation* (claim field) and *evaluation* (classification field) in the report. Five abstractive summarization models were also tested: 1) GTP-2 [1] (gpt2-medium); 2) BART (facebook/bart-base); 3) BERT (bert-large-uncased); 4) XLNet (xlnet-base-cased); and 5) T5 (t5-base).

To analyse the results obtained by these models, one hundred news items were randomly selected from the dataset and each of the previous approaches was applied. The results are accessed through automated and human evaluation. The former ensures that no fake news content is generated and that the maximum

[1] The GPT-3 is more robust version as it uses a large amount of data in the pre-training phase. However, it was not used as it is not available as open source.

length of the post is respected. The human evaluation was carried out by means of a questionnaire applied to a group of eight volunteers, who were asked to rate the coherence, consistency and informative quality of each post in binary terms, in order to analyse the subjective side of the results. The verification of fake news is carried out using the tool *Fake news classifier* (https://fake-news-detection-nlp.herokuapp.com). This classifier uses the BERT model, which achieved 98% accuracy and 99% recall and precision during validation. See table 1.

Table 1: Summary of evaluation results. The – means that the evaluation was not carried out because the approach had already been rejected.

Approach	**Size (N < 280)**	**True context**	**Acceptance**
T5	86%	96%	79%
BERT	60%	—	—
BART	100%	94%	25%
XLNet	60%	—	—
GTP-2	60%	—	—
Luhn's Heuristic method	23%	—	—
Text-Rank-NLTK	15%	—	—
Emotion-selection	63%	53%	—
Allegation sentence	100%	100%	83%

For extractive summarization, methods based on Luhn's heuristic method and Text-Rank-NLTK were discarded because the maximum post size is not respected in most examples. The emotion selection approach was also abandoned because it generated 37% of posts longer than 280 characters and almost half of them were classified as fake news (possibly because fake news is mostly associated with surprise and sadness). For abstractive summarization, 95% of the summaries from BERT, GPT-2 and XLNet were the same. GTP-2 was chosen to represent the results of this group. However, the GTP-2 model was then discarded as it leads to meaningless sentences (96% of the generated tweets are irrelevant, without knowledge or information related to the fact-checking news).

The results show a greater acceptance of the T5 model and the Allegation Sentence Witch as approaches chosen to obtain the *text-core* element of the post. Although the T5 model has a 14% chance of generating a post with a size greater than 280 characters, the content produced has only a 4% chance of creating a fake news. This model has a 79% acceptance rate among the study participants. The use of the allegation sentence, in addition to not producing fake content and the size always being within the expected range, has an acceptance rate of 83% in the human test carried out. On the other hand, the results of the BART model have a low acceptance among the volunteers in the study. The main reason is that in many cases the generated sentence is not complete and part of the message remains incomplete, as shown in Figure 2a. This means that the size is respected, but the consistency and informative character is lost.

Finally, figure 2b presents the actual tweet on this subject posted by Snopes in Twitter for comparison.

(a) BART model summarization. (b) Actual Snopes tweet.

Fig. 2: Tweet samples for reference.

2.2 Emotion reinforcement

The strategy of emotion reinforcement is to add *emojis* and *hashtags* to the *text-core* as described in the last section, and to emphasise keywords.

Positive and negative emoticons, classified according to [14], are added to the tweet in different numbers, up to four *cf.* [17], depending on the amount of space available (respecting the 280 character limit). The post will be updated with emoticons before hashtags because they are more successful in increasing engagement rates [11]. Hashtags allow people to find tweets, especially if they are trending and relevant. By applying Latent Dirichlet Allocation (LDA) with Dirichlet-distributed topic-word distributions to the text of the fact check report, relevant topics can be extracted (skip the inter-document step) [4] and those with higher emotional levels are included in the post as a hashtag.

In addition, to increase the emotional appeal of the tweet, the most important aspects were highlighted with capital letters. This allows the user to quickly identify the content of the post, which can lead to greater engagement. To do this, the *text-core* of the post is submitted to the *KeyBERT* [6] model to identify the keywords. This doesn't add any new information to the tweet.

For a reference, figure 3 depicts an instance of applying the suggested reinforcements to a plain allegation extracted as described in the last section. Roughly, the emotion assessment is done by a lexical approach *cf.* [7], based on the EMOLex dataset [13].

2.3 Mental process reinforcement

The strategy for strengthening mental processes is based on replacing words in the *text-core* with synonyms associated with a particular mental process when-

(a) Plain allegation sentence: `anger` 0.0; `anticipation` 0.0, `disgust` 0.0, `fear` 0.0, `joy` 0.0, `negative` 0.0, `positive` 0.0, `sadness` 0.0, `surprise` 0.0, `trust` 0.0.

(b) Emotional reinforcement: `anger` 0.0; `anticipation` 0.0, `disgust` 0.0, `fear` 0.0, `joy` 0.0, `negative` 1.0, `positive` 0.0, `sadness` 0.0, `surprise` 0.0, `trust` 0.0.

Fig. 3: Emotion Reinforcement Instance

(a) Emotion reinforced sentence: `paranoid` 0.2, `neuroticism` 0.5, `schizoid` 0.4.

(b) Neuroticism reinforcement: `paranoid` 0.0, `neuroticism` 0.9, `schizoid` 0.1.

Fig. 4: Mental Reinforcement Instance

ever possible. The approach is lexical and the reference data set is the MENTALex [10]. See figure 4 for an example.

3 Cognitive Reinforcement Assessment and Evaluation

In order to evaluate the results delivered by the prototype and, ultimately, to test the hypothesis of this paper, an experiment was conducted with 20 participants (the recommended number for a statically significant usability study [1]), between 12 and 18 September 2022. The anonymised volunteers were accustomed Twitter users recruited at the university, from whom the authors obtained informed consent. The experiment was conducted in a laboratory environment using the microblogging simulator [3]. The number of participants was chosen according to the guidelines suggested in [2].

Participants were asked to interact with the platform in the same way they do when scrolling through their Twitter feed. The posts received voluntarily refer to ten news items. A total of 30 posts: 10 generated by the T5 model, 10 generated by the allegation sentence, and 10 posts extracted from the `@snopes` Twitter page. All posts were randomly presented to the user on the platform, without

the user knowing which were generated by the prototype and which were from the `@snopes` Twitter page.

The plot in figure 5 shows the resulting engagement broken down by approach (T5 with cognitive reinforcement, allegation sentence with cognitive reinforcement, and actual Snopes tweets). In total, there were 434 interactions with the platform, an average of 21 interactions per participant and 14 interactions per post; in terms of interaction types, *like* was the most used by participants, followed by *follow* and *retweet*. This pattern is consistent with the one in [9].

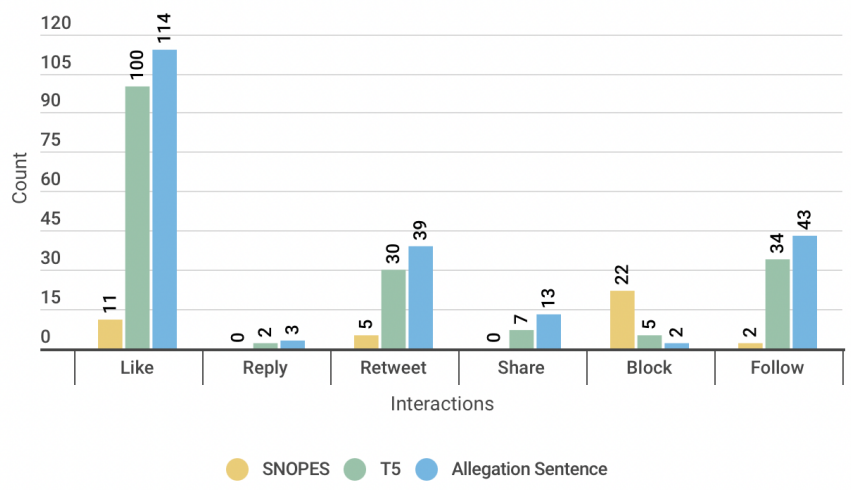

Fig. 5: Interactions count per approach

Note that in this study, unlike Twitter, the *block* interaction is associated with the post and not with the author's page. This means that when the participant *blocks* a post, it has a negative connotation towards the post. It takes an opposing viewpoint and its use shows a lack of interest in the post.

Compared to the other approaches, engagement with actual Snopes tweets is very low and includes most *block* interactions. See figure 2b for a comparison with figures 3 and 4. Broadly speaking, Snopes tweets use more formal text and fewer emoticons and hashtags compared to the proposed heuristic.

As this study was conducted in a controlled environment, it is possible to use an appropriate measure of engagement through $\frac{interactions}{visualisations}$. The idea behind this metric is that the number of interactions with a post divided by the number of times that post is presented would reveal how engaging it is. This is perhaps the simplest metric of engagement, but it is not often calculated as most online social media do not provide the number of times a post has appeared. Assuming that each volunteer was exposed to the same number of tweets, the tweet

produced by the prototype has an average engagement rate of 28% (for the T5 model) and 35% (for the allegation sentence). Therefore, allegation sentence with cognitive reinforcement is the most appropriate heuristic considering the issues addressed in this paper.

4 Conclusion

This paper started from the hypothesis that enhancing tweets with emotions and personality traits would increase engagement with fact-checking tweets. In order to test this hypothesis, a set of heuristics was proposed, resulting in a prototype. The results of this prototype were then submitted for human evaluation in a laboratory environment.

Among the models studied, the T5 model and the allegation sentence produced the best results, which were then used to build the prototype. Following the literature guidelines, the use of emoticons and hashtags succeeded in reinforcing the emotional dimension. Furthermore, based on the Adaptive Personality Theory, the reinforcement of the neuroticism process helped to improve the overall result. Therefore, according to the data presented, the hypothesis is confirmed with an increase of 35% of the engagement score.

For future work, the presented heuristic needs to be divided in order to test each trait as an independent variable, with each one optimised to obtain the maximum effect. In addition, a psychological validation of the generated content will certainly contribute to the improvement of the study.

Acknowledgements

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020;

References

1. Alroobaea, R., Mayhew, P.J.: How many participants are really enough for usability studies? In: 2014 Science and Information Conference. pp. 48–56. IEEE (2014)
2. Alroobaea, R., Mayhew, P.J.: How many participants are really enough for usability studies? In: 2014 Science and Information Conference. pp. 48–56. IEEE (2014)
3. Barbosa, M.A., Marcondes, F.S., Durães, D.A., Novais, P.: Microblogging environment simulator: An ethical approach. In: Advances in Practical Applications of Agents, Multi-Agent Systems, and Complex Systems Simulation. The PAAMS Collection: 20th International Conference, PAAMS 2022, L'Aquila, Italy, July 13–15, 2022, Proceedings. pp. 461–466. Springer (2022)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
5. Goswamy, T., Singh, I., Barkati, A., Modi, A.: Adapting a language model for controlled affective text generation. In: Proceedings of the 28th international conference on computational linguistics. pp. 2787–2801 (2020)

6. Grootendorst, M.: Keybert: Minimal keyword extraction with bert. (2020). https://doi.org/10.5281/zenodo.4461265, https://doi.org/10.5281/zenodo.4461265
7. Jurafsky, D., Martin, J.H.: Speech and Language Processing. draft (https://web.stanford.edu/ jurafsky/slp3/), third edn. (2023)
8. Lal, S., Tiwari, L., Ranjan, R., Verma, A., Sardana, N., Mourya, R.: Analysis and classification of crime tweets. Procedia computer science **167**, 1911–1919 (2020)
9. Marcondes, F.S.: A fact-checking profile on twitter (2020), data can be found in ALGORITMI Centre, University of Minho, Braga, Portugal.
10. Marcondes, F.S., Barbosa, M.A., Queiroz, R., Brito, L., Gala, A., Durães, D.: Mentalex: A mental processes lexicon based on the essay dataset. In: Artificial Intelligence XXXIX: 42nd SGAI International Conference on Artificial Intelligence, AI 2022, Cambridge, UK, December 13–15, 2022, Proceedings. pp. 321–326. Springer (2022)
11. Mention: Twitter report (2018), https://mention.com/en/reports/twitter/emojis/
12. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing. pp. 404–411 (2004)
13. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Computational intelligence **29**(3), 436–465 (2013)
14. Novak Kralj, P., Smailovic, J., Sluban, B., Mozetic, I.: Sentiment of emojis. PloS one **10**(12), e0144296 (2015)
15. Peters, C.: Emotion aside or emotional side? crafting an 'experience of involvement'in the news. Journalism **12**(3), 297–316 (2011)
16. Piccolo, L., Blackwood, A.C., Farrell, T., Mensio, M.: Agents for fighting misinformation spread on twitter: design challenges. In: Proceedings of the 3rd Conference on Conversational User Interfaces. pp. 1–7 (2021)
17. Sims, S.: 7 tips for using emojis in social media marketing (2017), https://www.socialmediatoday.com/marketing/7-tips-using-emojis-social-media-marketing
18. Verma, P., Pal, S., Om, H.: A comparative analysis on hindi and english extractive text summarization. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) **18**(3), 1–39 (2019)
19. Zhang, H., Song, H., Li, S., Zhou, M., Song, D.: A survey of controllable text generation using transformer-based pre-trained language models. arXiv preprint arXiv:2201.05337 (2022)