Transformer-based Prediction of Emotional Reactions to Online Social Network Posts

Irene Benedetto[1,2] **Moreno La Quatra**[3]

Luca Cagliero[1] **Luca Vassio**[1] **Martino Trevisan**[4]

[1] Politecnico di Torino, {name.surname}@polito.it
[2] MAIZE SRL, {name.surname}@maize.io
[3] Kore University of Enna, {name.surname}@unikore.it
[4]University of Trieste, {name.surname}@dia.units.it

Abstract

Emotional reactions to Online Social Network posts have recently gained importance in the study of the online ecosystem. Prior to post publication, the number of received reactions can be predicted based on either the textual content of the post or the related metadata. However, existing approaches suffer from both the lack of semantic-aware language understanding models and the limited explainability of the prediction models. To overcome these issues, we present a new transformer-based method to predict the number of emotional reactions of different types to social posts. It leverages the attention mechanism to capture arbitrary semantic textual relations neglected by prior works. Furthermore, it also provides end-users with textual explanations of the predictions. The results achieved on a large collection of Facebook posts confirm the applicability of the presented methodology.

1 Introduction

Most Online Social Network (OSN) platforms allow users to annotate posts with personal reactions. Reactions to social posts not only indicate the user sentiment (e.g., *like, dislike*) but also reflect emotions or feelings (e.g., *sadness, love, care*). Moreover, the quantity of reactions is also a direct measure of the popularity of a post and, indirectly, can indicate the audience's enthusiasm for a specific topic. Such annotations are particularly relevant to marketers, advertisers, and policymakers because they can be exploited to profile OSN users' behaviors and personalize the offer of related services/products. At large, understanding how users react to different types of content on OSNs is of paramount importance to studying human behavior and the online ecosystem.

Predicting the emotional reactions triggered by an OSN post before its publication is particularly appealing as it enables ad hoc content revision and prioritization. The information available ante-publication encompasses the textual content of the post and a set of related metadata (e.g., publisher, publication date, presence of links or associated images). Previous approaches to emotional reaction prediction (Giachanou et al., 2018) rely on traditional occurrence-based text statistics on text, e.g., TF-IDF (Manning et al., 2008). Thus, they ignore the semantics behind the text. Furthermore, the prediction models are used as closed boxes and do not provide any explanations of the predicted reaction.

This paper proposes a new approach to emotional reaction predictions based on Transformers (Vaswani et al., 2017) and Shapley-based explanations. The proposed architecture encodes both the textual content of the post and the related metadata to obtain attention-based predictions of the number of reactions per type. On top of a multi-task regressor, we use an established explainable AI model, namely SHAP (Lundberg and Lee, 2017), to provide end-users with explanations on the most influential textual features.

The main contributions of this paper can be summarized as follows:

- The paper presents a new approach to predict emotional reactions to OSN posts ante-publication, i.e., disregarding post comments or replies.
- It formulates the emotional reaction prediction task as a multi-task regression problem, where the target variables are the number of reactions per type received by the post.
- It proposes a Transformer-based architecture combining both textual content and metadata. The adoption of Transformer-based approaches allows us to capture arbitrary semantic textual relations neglected by prior works.
- It provides end-users with textual explanations

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 354–364

of the generated predictions based on an established Shapley-based model.

- We extensively validate the proposed approach on a collection of real posts from the Facebook OSN. The results confirm the superior performance of transformer-based models and the prevailing role of textual content compared to metadata.

2 Related work

Despite the availability of various sources of data, accurately predicting content popularity on OSNs is still a challenging task due to their dynamic nature and the presence of various factors. Several techniques and models have been proposed to address this challenge. Typically the forecast objectives are popularity metrics, such as the number of likes a post will get. The features used for the prediction are obtained within the OSN, such as the content of the post, the author of the post, the previous posts, and antecedent metrics on the performance of the post, or are gathered from outside of the OSN, e.g., presence in newspapers and TV shows.

Popularity (likes, views, comments, etc.) is the main sign of success on OSNs, and, as a consequence, it is the focus of most pieces of research. Many works predict content popularity considering content intrinsic characteristics and social interaction features. In particular, Natural Language Processing (NLP) techniques are leveraged for helping predict future popularity. The factors that impact the popularity of posts on Facebook are identified in (Sabate et al., 2014), using an empirical analysis involving multiple linear regressions, with the most important factors being the number of followers and the presence of images. Similarly, (Ferrara et al., 2014) highlights the characteristics related to the dynamics of content consumption in Instagram, while (Gayberi and Oguducu, 2019) and (Carta et al., 2020) predict the popularity of a post by combining user and post features. The authors of (Rizos et al., 2016) predicts different Reddit news popularity indicators using the comment tree and the user graph, while (Li et al., 2013) considers the early views for prediction, focusing on features related to the intrinsic attractiveness of a video and the influence from a propagation structure. Fewer works in the literature deal with emotional reaction prediction. Among them, (Giachanou et al., 2018) simplifies the problem to a classification task. Conversely, (Krebs et al., 2017) predicts Facebook reaction quantity, using NLP on the text of the post and on comments and answers to the post after publication. Notice that content popularity can be also forecast from other information outside of the social network (exogenous), e.g., presence in newspapers, TV shows, etc. An example is the work by (Bertone et al., 2021) that forecasts Instagram and Facebook influencer popularity by extracting external data from Google Trends and then applying financial stock-market tools such as Bollinger Bands.

Another body of literature focuses on the prediction of the temporal dynamics of post popularity. The authors of (Vassio et al., 2021, 2022) study how influencer posts attract likes and reactions and the factors for content popularity evolution. Similarly, (Ahmed et al., 2013) identifies temporal evolution patterns and uses those to predict the future popularity of the content using data from Youtube, Digg and Vimeo, applying K-means clustering and simple linear forecasting technique. Finally, (Ramachandran et al., 2018) propose a model that reproduces the popularity attraction on Twitter, observing that hourly interactions decrease geometrically with time.

Other works focus on predicting other dimensions of popularity, and not the intensity itself. The authors of (Hu et al., 2017) predict popularity of posts by subdividing their evolution into three key moments ("burst", "peak", and "fade") using a Support Vector Regression technique. In a similar direction, (Yu et al., 2020) predicts when the popularity of Twitter hashtags reaches its peak, using an LSTM Deep Learning model with topological network information, social information, and Hashtag strings.

Differently from previous works, in this paper we focus on predicting not only a single metric, but multiple reaction types associated with specific sentiments. We keep the regression objective to make precise and fine-grained predictions, without simplifying it to a classification problem. Moreover, we only use data available at publication time, hence we do not need any other information like early popularity or comment content. Indeed, our approach is only based on the content and metadata of the post itself, and on the features of the creator's profile. We leverage modern Transformers to capture more subtle, semantic patterns in the text and perform better predictions.

3 Methodology

3.1 Problem formulation

Let u be a user of an Online Social Network (OSN) creating a post p at time t_p. p is characterized by a textual content c_p and a set of metadata (e.g., number of followers of the post creator at time t_p, the post publication date and time, and post author, and the presence of images, videos, or links associated with the post). The other OSN users can annotate p with 9 types of reactions in $\mathbf{R}$ (i.e., *like, love, wow, care, haha, wow, angry, comment, share*). More details on the annotations and metadata information considered in this study are given in Section 4.1.

Let $r \in \mathbf{R}$ be an arbitrary reaction type. We keep track of the number of reactions of that particular type produced by OSN users over time. Specifically, let $r(p,t)$ be the cumulative function of the number of reactions of type r received by post p at time t.

Given a post p we model the task of emotional reaction prediction as a *multi-task regression problem*, where the targets are the numbers $r(p,t)$ of reactions at time $t > t_p$ for every reaction type $r \in \mathbf{R}$.

Notice that:

- The prediction model *exclusively* considers ante-publication information. Hence, post comments and replies are ignored.
- Whenever not otherwise specified, we set the *prediction horizon* to *one day ahead*, i.e., the time elapsed after the post publication time t_p.
- The *popularity of the post creator* (e.g., the number of friends/followers) likely influences the absolute number of post reactions and it is considered in the post metadata. Furthermore, to properly handle imbalances in the target variable distributions, we address the prediction of the logarithmic function $\tilde{r}_{p,t} = \log_{10} \frac{r_{p,t}}{n_{o_p,t}}$ instead of $r_{p,t}$, where $n_{p,t}$ is the number of followers of the post owner o_p at time t. The use of the logarithm allows us to both aid in achieving training convergence and to compress the variability, while not losing information when the input values are too small.
- The presence of specific terms/expressions in the content of the post may trigger specific reactions (e.g., *Sounds great!* likely triggers a *wow* reaction). We consider this aspect by attending specific pieces of text by means of a transformer architecture (Conneau et al., 2019).

3.2 Regression Model

To address this multi-task regression problem, we propose the architecture depicted in Figure 1, hereafter denoted by Transformer+Metadata. It leverages both text and metadata data to attain precise emotional reaction forecasts. It consists of:

- A Transformer model, namely XLM-RoBERTa (Conneau et al., 2019), to represent each token in the input text.
- A subsequent average pooling operation, to derive the representation of the textual content of the post.
- A fusion stage between the text encoding and the metadata vector[1].
- A fully connected multi-task regression layer with logits to generate the output predictions.

We compare Transformer+Metadata with the following baseline methods:

- Metadata-Only: A classical regression model relying on Linear regression, Adaboost regressor, Random forest regressor, and Multi-Layer perceptron. The models rely only on post metadata, i.e., they disregard the textual content of the post.
- TF-IDF+Metadata: An extended version of Metadata-Only considering also the content of the post encoded using an established occurrence-based text representation, i.e., the term frequency-inverse document frequency (TF-IDF) (Manning et al., 2008). Textual and metadata features are concatenated to feed the regression model. Unlike Transformer+Metadata, TF-IDF+Metadata does not rely on Transformers.
- TF-IDF: a simplified version of TF-IDF+Metadata based on textual features solely.

[1] On the training dataset the vector is normalized based on mean and variance.

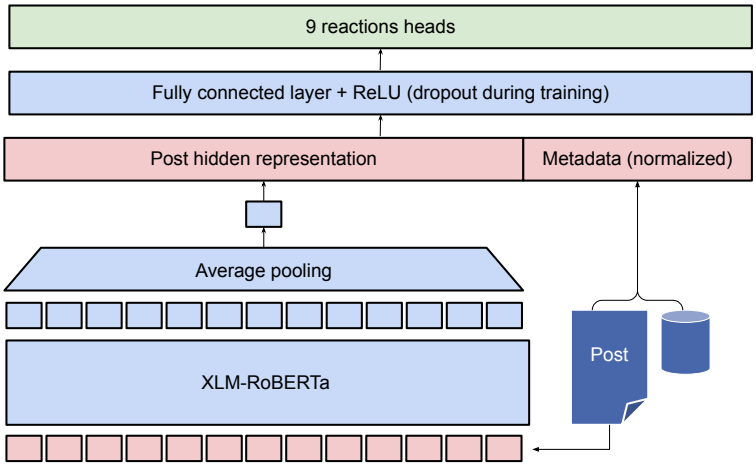


Figure 1: The proposed architecture.

- Transformer-Only: An architecture based on the RoBERTa transformer for text encoding. It ignores all metadata information. It corresponds to the left-hand side of the architecture proposed in Figure 1.

- Moving Average: an auto-regressive model based on moving average proposed by Facebook Inc. This naive approach forecasts a post's reactions using a moving-average over the last 10 posts of the same author. These forecasts are provided directly by CrowdTangle within the posts' metadata.

Our Transformer-based regressors (Transformer+Metadata and Transformer-Only) rely on XLM-RoBERTa (Conneau et al., 2019) for the following reasons:

- It achieved state-of-the-art performance on various natural language processing tasks including sentiment analysis, named entity recognition, and machine translation.

- It has been trained on a large-scale dataset including 100 different languages. Thus, it allows us to apply the proposed approach on multilingual OSN posts (e.g., foreign expressions words included in English posts).

3.3 Model Explanation

To provide explanations of the forecasts, we study the influence of textual features on the predicted reaction values. With the goal of explaining the influence of textual tokens on the prediction model performance, we apply the explainer on top of the Transformer-Only model. Specifically, for each reaction type we highlight the tokens in the input text that mostly influence the prediction of the value of a particular reaction type using the text explainer available in the SHAP library (Lundberg and Lee, 2017). SHAP relies on the concept of Shapley Value, which is established for game theory. It quantifies the value to each player in a cooperative game based on the contribution to the total payoff of the group.

Let $\mathbf{T}_p$ be the set of tokens contained in the content of post p. Given a regression model $\mathcal{R}$ and a target number of reactions $r(p,t)$ for post p, the explainer computes the SHAP Value $\phi(t_p)$ associated with each token $t_k \in \mathbf{T}_p$. SHAP quantifies the influences of token t_k on the $r(p,t)$'s prediction.

4 Experimental setup

4.1 Dataset

To evaluate the proposed techniques, we use a dataset of posts from the Facebook social network. To obtain the posts, we rely on the CrowdTangle platform and its API[2]. CrowdTangle is a content discovery and social analytics tool owned by Meta, which is open to researchers and analysts worldwide upon having a partnership agreement.

We consider posts created by the Facebook profiles of UK newspapers and TV/radio stations of national and local importance. The list of profiles is created and maintained directly by the CrowdTangle team and, in total, includes 1 165 profiles. The list includes the most popular UK broadcast-

[2] https://github.com/CrowdTangle/API Latest access: April 2023

ers (e.g., the BBC) or newspapers (e.g., the Daily Mail), while others represent local media companies. Notice that the same company can be present with multiple accounts. For instance, the BBC owns many Facebook profiles dedicated to different topics (politics, economics, etc.) and regions (Wales, Scotland, etc.). This study considers posts created over more than three years, from the beginning of 2019 to mid-2022. Over that period, the profiles created a total of $\approx 6\,M$ posts, thus $\approx 40\,k$ per week, on average.

In this work, we consider only influencers whose number of followers is greater than $100\,k$, called Mega and Macro in (Zarei et al., 2020). This decision was based on our observation that influencers with a smaller audience tend to elicit a limited number of reactions. We processed a total of 709,142 posts, which were split into train, test, and development sets in an 80/10/10 ratio.

Each post bears a textual caption with a maximum length of 500 characters and, optionally, can include a picture, a video or a link. In case the post includes a link to an external webpage, CrowdTangle reports the title of the linked webpage as extracted from the HTML and the first sentence of the webpage body – e.g., an article's content. In our experiments, we concatenate this extra text to the post caption for later processing. Moreover, CrowdTangle provides various metadata on posts, such as the creation time and, important to our analysis, a historical view of the reactions received. On Facebook, users can *comment* on a post – i.e., reply with a short text and interact with other commenters. Users also have the possibility to *react* to posts – i.e., expressing their feeling through a predefined set of 7 *emoji*. Namely, they can express *Like*, *Love*, *Care*, *Laughter*, *Surprise*, *Sadness* and *Anger*. The post's metadata include the temporal evolution of the number of comments and reactions (separately by type) received by a post. This information is provided at several time steps, whose granularity becomes coarser as time passes. During the first 24 hours after the post's creation, the metadata indicate the number of comments and reactions every fifteen minutes. After one day from post creation, numbers are provided with a daily granularity. For our analysis, we always consider the number of reactions and comments a post received after 1 day since its creation, i.e., $t = 24$ hours in $r(p, t)$.

Table 1 reports general statistics on the dataset, showing the median, mean and standard deviation of the number of reactions per post. We observe that the prevalent reaction is *Like*, as, in median, posts in our dataset receive 59 *likes*. The distribution presents a heavy tail, as already previously shown in the literature (Vassio et al., 2022), since the mean number of *p*er post is 333 and the standard deviation is rather large (1 768). Other reactions are rather frequent and in the second and third position we find *Comment* and *Share*, respectively, while others are certainly rarer, such as *Care*, *Angry* or *Sad*. Indeed, *Care* is the rarest reaction and is 0 for more than half of posts.

Table 1: Median, mean and standard deviation of target values, for each reaction.

	Median	Mean	Standard Deviation
Angry	1.00	41.04	257.05
Care	0.00	9.95	261.63
Haha	4.00	51.95	414.03
Like	59.00	333.80	1 768.86
Love	2.00	43.94	510.39
Sad	1.00	60.53	809.89
Wow	2.00	20.20	161.46
Comment	46.00	194.45	602.06
Share	13.00	117.56	993.57

4.2 Metrics

Given a post p by an influencer o_p, we have the predicted $\widehat{r}(p, t)$ and actual $r(p, t)$ reactions. To measure the performance of the different prediction models, we adopt the well-known Median Absolute Percentage Error (MedAPE). It is defined as follows:

$$\text{MedAPE} = median \frac{|r(p,t) - \widehat{r}(p,t)|}{r(p,t)} \times 100$$

The MedAPE measures, in percentage, to what extent the prediction deviates from the real value, which in our case is the number of reactions – here, we do not apply any normalization. Notice that the model's output and target values are still the logarithms of the reactions per follower and expected reactions per follower to reduce skewness. However, we opt to measure the performance on the original reaction number, as its prediction represents the objective of our work.

To understand how Transformer-based models operate, we leverage Shapley values. They represent the contribution of each token to the predicted value. The $\text{mean}(|\text{SHAPvalue}|)$ for a specific regression and reaction is defined as:

$$\text{mean}(|\text{SHAPvalue}|) = \frac{1}{N}\sum_{p=1}^{N}\phi_{i,p,t}$$

where $\phi_{i,p,t}$ is the function that computes the Shapley value of token i in post p for reaction t. We sum up the values for each token across all posts and divide them by the total number of posts to obtain the average token importance. The $\text{mean}(|\text{SHAPvalue}|)$ provides insight into the degree of importance of each token in the prediction.

4.3 Model evaluation and hyper-parameter setting

To evaluate Metadata-Only, TF-IDF+Metadata, and TF-IDF we perform hyper-parameters tuning using a 5-fold cross validation. To evaluate Transformer+Metadata and Transformer-Only we train the `xlm-roberta-base`[3] model version using a batch size of 16 and a learning rate of $5 \cdot 10^{-5}$. Models undergo a maximum of 5 epochs of training, with an early stopping criterion applied. Both models also have a weight decay of 0.01 and a warmup ratio of 0.06.

4.4 Hardware

Experiments were run on a machine equipped with Intel® Core™ i9-10980XE CPU, 2 × Nvidia® RTX A6000 GPU, 128 GB of RAM running Ubuntu 22.04 LTS. We provide detailed information about the models used for the evaluation and the fine-tuning procedure in the official project repository[4].

5 Results

In this section, we show and discuss the results of our experiments. We evaluate the proposed regression models on the dataset described in Section 4.1 in terms of Median Absolute Percentage Errors. Then, we delve into the Transformer+Metadata model and investigate the most relevant words to the regression task using the SHAP algorithm.

5.1 Analysis of Regression Performance

In this section, we show and discuss the performance of the different regression models. In Table 2, we report Median Absolute Percentage Error for all models, separately by reaction type. The table allows us to compare Transformer+Metadata with the other simpler models and quantify the benefits of this approach. For the sake of brevity, for TF-IDF+Metadata, Metadata-Only, and TF-IDF we report only the outcomes of the model that performs best on the validation set (i.e., MLP regressor).[5] Each cell in the Table indicates the MedAPE for a given combination of model and reaction. Notice that the value is reported in percentage. Since the reported error is relative, the number in a cell represents the percentage error with respect to the target magnitude.

Watching Table 2, we first observe the problem we address challenges all the proposed models, as in all cases the MedAPE is above 50%. *Care* is an exception, as the MedAPE assumes low values for Transformer-based models. This is expected, as this reaction often assumes value 0 as reported in Table 1. As such, the models tend to output 0 as the predicted value, leading to a correct prediction in most cases. In many cases, the simplest models provide very erroneous predictions, which we report in Table 2 with $> 100\%$ to indicate that the prediction model was not able to provide any kind of meaningful output.

Comparing the different regression models, we note that Transformer-based ones perform best thanks to the higher capability to capture semantic information from text. Overall, Transformer-Only and Transformer+Metadata have the best prediction accuracy for most reaction types. If neglect *Care* (which is most of the times 0), the best performance is achieved with *Wow*, where Transformer+Metadata achieves a MedAPE of 45.26%. Transformer+Metadata provides the best performance for all reactions except for *Share*, for which Transformer-Only performs best, with MedAPE of 66.55%. Overall, we observe that the impact of metadata in transformer-based models is relatively limited. Indeed, the improvements of Transformer+Metadata with respect to Transformer-Only are in most cases less than a percentage point. In this direction, notice the asterisks (*) in the table, that indicate whether the best-performing model offers a statistically significant improvement with respect to the given cell. The results show that Trasformer-Based models improve *significantly*

[3] https://huggingface.co/xlm-roberta-base Latest access: April 2023

[4] https://anonymous.4open.science/r/social-reactions-predictions-7CB0/README.md

[5] We selected the MLP regression after performing grid search on validation data with several configuration parameters that include support vector machines, *AdaBoost* regressor, multilayer perceptron regressor, and linear regression

Table 2: Comparison between regression performance on test data for each reaction type, in terms of MedAPE. The asterisk (*) indicates that the best-performing model (reported in bold) offers a statistically significant improvement (t-test with p-value=0.05) with respect to the given cell.

Model	Angry	Care	Comment	Haha	Like	Love	Sad	Share	Wow
Moving Average	>100% *	>100% *	97.20% *	>100% *	>100% *	>100% *	>100% *	86.20% *	>100% *
Metadata-Only	>100% *	>100% *	>100% *	99.98% *	98.03% *	97.12% *	>100% *	99.99% *	>100% *
TF-IDF	78.52% *	30.19% *	95.21% *	97.14% *	>100% *	89.66% *	>100% *	90.24% *	82.10% *
TF-IDF+Metadata	69.45% *	37.98% *	91.21% *	76.95% *	76.38% *	74.35% *	67.27% *	79.62% *	72.14% *
Transformer-Only	56.47%	0.10%	78.30%	62.42%	51.69% *	63.44% *	66.86% *	**66.55%**	50.22% *
Transformer+Metadata	**56.30%**	**0.05%**	**78.15%**	**62.47%**	**50.67%**	**62.07%**	**61.87%**	66.56%	**45.26%**

prediction performance, but in four cases Transformer+Metadata and Transformer-Only performance is not statistically different. Conversely, for traditional machine learning models, jointly considering text and metadata information significantly improves the performance – compare for example TF-IDF+Metadata with TF-IDF. This is likely due to the more limited informative content of the occurrence-based text representation.

5.2 Examples of Model Explanation

We now focus on Transformer-Only and consider the trained model that we obtain. The model is based on a Transformer architecture that bases its predictions on the post text. As such, we now investigate how Transformer-Only operates in making its decision and try to *explain* its choices with the help of the SHAP technique and library. The plots in Figure 2 respectively show the most influential tokens (at most 20) for three reactions, namely *Angry*, *Love*, and *Care*, as obtained using SHAP. Here, our main goal is to gain insights into the behavior of a non-interpretable model, such as the Transformer-Only, and identify the underlying factors driving OSN users' reactions. For each textual token, the plots report the $\mathrm{mean}(|\mathrm{SHAPvalue}|)$ (Lundberg and Lee, 2017) computed over a random sample of $N = 100$ posts in the test set. This value indicates the average feature influence on the output predictions. More specifically, the Shapley values are computed using the *auto* configuration, which automatically recommends the best Explainer for our model.

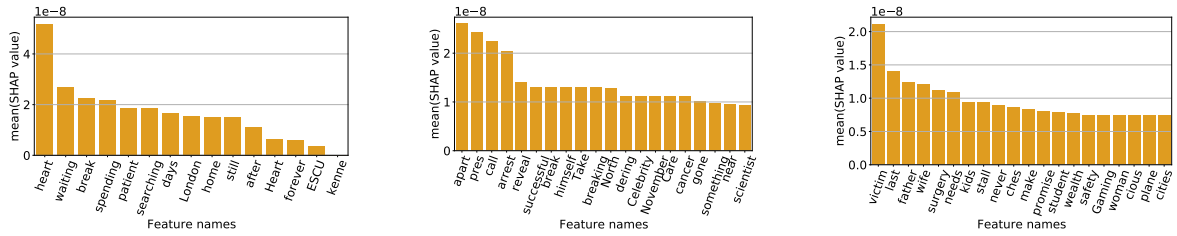
The Explainer is capable of detecting particular words and phrases having a significant impact on the types of reactions a post receives. This information could be useful for individuals and organizations who are looking to maximize engagement with their social media content. By incorporating more positive and uplifting language into their posts, authors may be able to increase the number of positive reactions (*Love* and *Ahah*) they receive, while reducing the number of negative ones (*Sad* and *Angry*).

Based on the SHAP values, we can gain insights into the types of language and content that are more likely to elicit certain reactions from OSNs users. The words and phrases associated with the *Care* reaction (i.e., Figure 2a) suggest that posts that convey a sense of patience and perseverance may be more likely to receive caring reactions. Similarly, the words and phrases associated with the *love* reaction (i.e., Figure 2b) suggest that posts that contain positive news, success stories, and celebrity updates may be more likely to receive love reactions. On the other hand, the words and phrases associated with the *angry* reaction (i.e., Figure 2c) suggest that posts that contain words related to victimhood, surgery, and family issues may be more likely to receive angry reactions. A more detailed analysis of the explanations can be found in Appendix A. It is important to note that these insights are based on the specific data and model used for the analysis, and may not necessarily apply to all contexts and users.

6 Limitations

Our methodology provides a novel approach to predicting emotional reactions to social posts. While our proposed methodology has shown promising results, further research is necessary to address some limitations and expand the scope of our findings.

Firstly, our model's reliance on only textual content and metadata may limit prediction accuracy, as other factors such as user demographics or multimedia content may also impact emotional reactions. Future work could explore the inclusion of additional features. Secondly, this study focuses on predicting emotional reactions to posts, rather than the reasons behind them. While our method provides textual explanations, they may not fully capture the complexity of user emotional responses.



(a) *Care* (b) *Love* (c) *Angry*

Figure 2: The feature importance of strongly correlated tokens computed as the mean of the absolute value of Shapley values for each reaction.

Lastly, our approach was evaluated on Meta posts, but may not generalize well to other social networks or domains. Further evaluations on different datasets are needed to confirm the applicability of our methodology across different contexts.

7 Ethical considerations

While our study minimizes privacy concerns by solely analyzing creator post content, we acknowledge the ethical implications of our model's predictions on individuals and communities.

Overall, the models proposed in the present study raise ethical concerns around the potential impact of our model's predictions on content creator behavior and the distortion of social media sentiment. Our model's predictions may incentivize content creators to prioritize emotional reactions over accuracy or factual content. To address this concern, responsible and ethical use of our methodology, alongside recognition of the limitations in capturing the complexity of user behavior, is necessary.

Finally, we observe that the dataset used in this work has been collected in a lawful and ethical environment. We used the CrowdTangle platform offered by Facebook, which contains only public posts of public profiles. The metadata of posts do not contain any sensitive information and only include the aggregate number of reactions they received.

8 Conclusions and Future Work

The paper explored the use of transformer-based models to predict the number of reactions antepublication to a social post. It focused on predicting the reaction counts for different types of positive and negative reactions by exploring the role of text and metadata information. Compared to prior approaches based on traditional text representations (e.g., TF-IDF), it achieved significant performance improvements thanks to the higher capability to capture the semantics behind the text. Transformers perform best compared to simpler methods even in the absence of metadata. The paper also leveraged a Shapley-based explainer to identify the tokens that mostly influence the prediction outcomes. The explanations meet the expectation, especially for very positive and very negative reaction types.

Acknowledgements

References

Mohamed Ahmed, Stella Spagna, Felipe Huici, and Saverio Niccolini. 2013. A peek into the future: Predicting the evolution of popularity in user generated content. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 607–616, New York, NY, USA. Association for Computing Machinery.

Fabio Bertone, Luca Vassio, and Martino Trevisan. 2021. The stock exchange of influencers: A financial approach for studying fanbase variation trends. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, page 431–435, New York, NY, USA. Association for Computing Machinery.

Salvatore Carta, Alessandro Sebastian Podda, Diego Reforgiato Recupero, Roberto Saia, and Giovanni Usai. 2020. Popularity Prediction of Instagram Posts. *Information*, 11(9).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Emilio Ferrara, Roberto Interdonato, and Andrea Tagarelli. 2014. Online Popularity and Topical Interests through the Lens of Instagram. In *Proceedings*

of the 25th ACM Conference on Hypertext and Social Media, page 24–34.

Mehmetcan Gayberi and Sule Gunduz Oguducu. 2019. Popularity Prediction of Posts in Social Networks Based on User, Post and Image Features. In *Proceedings of the 11th ACM MEDES*, page 9–15.

Anastasia Giachanou, Paolo Rosso, Ida Mele, and Fabio Crestani. 2018. Emotional influence prediction of news posts. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Ying Hu, Changjun Hu, Shushen Fu, Mingzhe Fang, and Wenwen Xu. 2017. Predicting key events in the popularity evolution of online information. *PLOS ONE*, 12(1):1–21.

Florian Krebs, Bruno Lubascher, Tobias Moers, Pieter Schaap, and Gerasimos Spanakis. 2017. Social emotion mining techniques for facebook posts reaction prediction.

Haitao Li, Xiaoqiang Ma, Feng Wang, Jiangchuan Liu, and Ke Xu. 2013. On Popularity Prediction of Videos Shared in Online Social Networks. In *Proceedings of the 22nd ACM CIKM*, page 169–178.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Arthi Ramachandran, Lucy Wang, and Augustin Chaintreau. 2018. Dynamics and prediction of clicks on news from twitter. In *Proceedings of the 29th on Hypertext and Social Media*, HT '18, page 210–214, New York, NY, USA. Association for Computing Machinery.

Georgios Rizos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2016. Predicting News Popularity by Mining Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, page 737–742.

Ferran Sabate, Jasmina Berbegal-Mirabent, Antonio Cañabate, and Philipp R. Lebherz. 2014. Factors influencing popularity of branded content in Facebook fan pages. *European Management Journal*, 32(6):1001–1011.

Luca Vassio, Michele Garetto, Carla Chiasserini, and Emilio Leonardi. 2021. Temporal dynamics of posts and user engagement of influencers on facebook and instagram. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, page 129–133, New York, NY, USA. Association for Computing Machinery.

Luca Vassio, Michele Garetto, Emilio Leonardi, and Carla Fabiana Chiasserini. 2022. Mining and modelling temporal dynamics of followers' engagement on online social networks. *Social Network Analysis and Mining*, 31:012012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Hai Yu, Ying Hu, and Peng Shi. 2020. A prediction method of peak time popularity based on twitter hashtags. *IEEE Access*, 8:61453–61461.

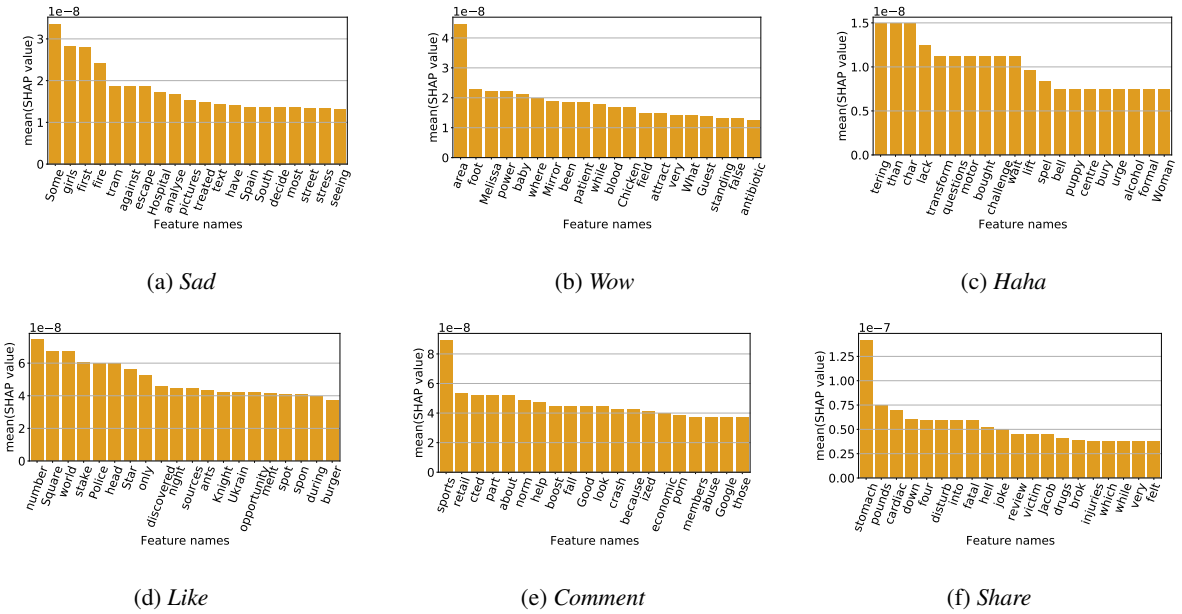
Koosha Zarei, Damilola Ibosiola, Reza Farahbakhsh, Zafar Gilani, Kiran Garimella, Noël Crespi, and Gareth Tyson. 2020. Characterising and detecting sponsored influencer posts on instagram. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 327–331.

A Detailed discussion of reaction explainations

In this section, we provide a more detailed explanation of the SHAP values associated with each type of reaction in the proposed Transformer-Only model. We discuss the tokens that are found to be strongly correlated with each reaction prediction and provide some insights into the types of content that may elicit these reactions from social media users. Figure 3 shows a detailed view of the explanation for the reactions that were not discussed in the main body of the paper. The histograms for the other reactions are reported in Figure 2. It is important to note that the $\mathrm{mean}(|\mathrm{SHAPvalue}|)$ presented in this section are not intended to identify the specific words or phrases that elicit each type of reaction. Rather, they are intended to provide insights into the tokens that are strongly correlated with the estimation of each reaction in our regression model.

Care

The *care* reaction (Figure 2a) is often associated with words and phrases that convey a sense of empathy and concern for others. Based on the scores provided by the explainability, the words that are most strongly associated with the care reaction include "heart", "waiting", "break", and "spending". These words suggest that users may react with *care* to posts that relate to personal struggles, offer support, or highlight the importance of patience and



(a) *Sad* (b) *Wow* (c) *Haha*

(d) *Like* (e) *Comment* (f) *Share*

Figure 3: $\mathrm{mean}(|\mathrm{SHAPvalue}|)$ values for most relevant tokens pertaining to different reaction types.

resilience. For example, posts that discuss a difficult experience or offer words of encouragement to those going through a tough time may be more likely to arouse this reaction.

Love

The frequency distribution of token relevance for the *love* reaction is illustrated in Figure 2b. The tokens that are strongly associated with positive emotions and affection tend to receive more *love* reactions from users.

The explainability model has identified certain tokens such as “successful”, “care”, “cancer”, and “scientist” that are useful to estimate the *love* reaction. This may suggest that users tend to react with love towards posts that feature success, care for others, good health, or scientific accomplishments.

Angry

The *angry* reaction (Figure 2c) is often characterized by the use of language that conveys frustration or anger. The explainability models indicate that the tokens most strongly associated with the angry reaction include “victim”, “last”, “father”, and “wife”. OSN users are more likely to react with anger to posts that pertain to unfair treatment or injustice, such as the victimization of vulnerable groups or the loss of a loved one.

Moreover, the words ”surgery”, ”needs”, and £kids” indicate that users may be more prone to reacting with anger to posts that deal with issues of health or well-being, such as inadequate access to medical care or basic necessities.

Sad

The *sad* reaction (shown in Figure 3a) is typically associated with phrases and words that convey negative emotions like grief, disappointment, or sadness. The explainability model’s relevance scores indicate that tokens such as ”fire”, ”hospital”, ”street”, and ”stress” are most strongly connected to the *sad* reaction. This suggests that posts related to traumatic events, medical emergencies, stress, and challenging experiences are more likely to elicit this reaction from users.

Although the words ”girls”, ”tram”, and ”escape” don’t seem directly linked to the *sad* reaction, they may also suggest that users react with sadness to posts related to personal struggles. For instance, stories of oppression or discrimination may evoke feelings of sadness in users.

Wow

The *Wow* reaction (Figure 3b) is triggered by content that suggests surprise or admiration. Tokens like ”power” and ”baby” indicate that users are more likely to react with amazement to displays of strength or cute and impressive babies, and stories that are out of the ordinary. Additionally, tokens like ”area”, ”foot”, and ”field” suggest that users may react with surprise to posts that relate to natural phenomena or impressive athletic feats.

Haha

The *haha* reaction (Figure 3c) is typically used to express surprise or amusement in response to certain words or phrases. Although some associated terms may be coincidental, other tokens show a strong semantic connection to the reaction. For example, words like ”transform”, ”motor”, ”challenge”, and ”puppy” appear to be highly correlated with the *haha* reaction. Such words suggest that users tend to react with humor or surprise to content that involves unexpected or funny instances related to technology, witty takes on everyday experiences or amusing content involving animals.

Like, Comment, and Share

The social media actions of *Like*, *Comment*, and *Share* are prevalent across various platforms. However, according to explainability models, the tokens linked with each of these actions do not necessarily indicate particular content that would elicit them (see Figures 3d, 3e, 3f).

For example, the tokens relevant to the *Like* reaction include ”number”, ”Square”, ”world”, and ”stake”, which do not have an apparent semantic connection with the positive emotions commonly associated with the *Like* reaction. Similarly, the tokens linked with *Comment* and *Share*, such as ”sports”, ”retail”, ”stomach”, and ”pounds”, do not necessarily suggest specific content that would encourage users to engage with a post by commenting or sharing.