A SHAP-based controversy analysis through communities on Twitter

Samy Benslimane

benslimanesamy@outlook.fr

LIRMM, Univ Montpellier, CNRS

Thomas Papastergiou

LIRMM, Univ Montpellier, CNRS

Jérôme Azé

LIRMM, Univ Montpellier, CNRS

Sandra Bringay

AMIS, Paul-Valéry University

Maximilien Servajean

AMIS, Paul-Valéry University

Caroline Mollevi

Institut du Cancer Montpellier (ICM)

Research Article

Keywords: Machine learning, SHAP, Analysis, Explainability, Controversy, Communities

Posted Date: February 7th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3908863/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at World Wide Web on September 14th, 2024.
See the published version at https://doi.org/10.1007/s11280-024-01278-z.

A SHAP-based controversy analysis through communities on Twitter

Samy Benslimane[1*], Thomas Papastergiou[1], Jérôme Azé[1], Sandra Bringay[1,2], Maximilien Servajean[1,2], Caroline Mollevi[3,4]

[1*]LIRMM, Univ Montpellier, CNRS, Montpellier, France.
[2]AMIS, Paul-Valéry University, Montpellier, France.
[3]Institut du Cancer Montpellier (ICM), Montpellier, France.
[4]IDESP, UMR Inserm - Univ Montpellier, Montpellier, France.

*Corresponding author(s). E-mail(s): benslimanesamy@outlook.fr; samy.benslimane@lirmm.fr;
Contributing authors: thomas.papastergiou@lirmm.fr; jerome.aze@lirmm.fr; sandra.bringay@lirmm.fr; maximilien.servajean@lirmm.fr; caroline.mollevi@chu-montpellier.fr;

Abstract

Controversy encompasses content that draws diverse perspectives, along with positive and negative feedback on a specific event, resulting in the formation of distinct user communities. Research on controversy can be broadly categorized into two domains: controversy detection/quantification, which aims to measure controversy on a topic, and controversy explainability, which seeks to comprehend the reasons behind a topic's controversial nature. This paper primarily contributes to the realm of controversy explainability. We conduct an analysis of topic discussions on Twitter from a community perspective, investigating the role of text in accurately classifying tweets into their respective communities. To achieve this, we introduce a SHAP-based pipeline designed to quantify the influence of impactful text features on the predictions of three tweet classifiers. Our approach involves leveraging various text features, including $\boldsymbol{BERT}$, $\boldsymbol{TF-IDF}$, and $\boldsymbol{LIWC}$. The results, derived from both SHAP plots and statistical analyses, distinctly reveal the substantial impact of certain text features in tweet classification. Furthermore, our findings underscore the significance of this study and underscore the potential advantages of combining text and user interactions for a comprehensive understanding of controversy quantification.

Keywords: Machine learning, SHAP, Analysis, Explainability, Controversy, Communities

1 Introduction

Social media platforms, exemplified by Twitter, provide individuals with a significant avenue to articulate, share, and engage in discussions about their opinions and ideas on a wide array of subjects. Certain topics, marked by their ability to attract diverse and opposing viewpoints, often give rise to what is commonly referred to as controversy. This phenomenon is frequently instigated by impactful events, particularly within the realms of political discourse, climate change, gun legislation, and similar domains.

While controversy can be defined in various ways, for the purposes of this research, we adopt the perspective that controversial topics encompass those eliciting disparate viewpoints and feedback regarding a specific event, thereby polarizing users into two primary conflicting communities [1]. The examination of controversy on social media has gained increasing significance in numerous applications. This includes the identification of opinion divergences, mitigation of the spread of fake news, bridging gaps between communities, and addressing the impact of the "filter bubble" phenomenon, where algorithmic personalization limits information diversity and alters perception. The task of automatically detecting controversy poses a substantial challenge and has been extensively explored. Recent approaches, particularly in the context of social media, predominantly leverage structural information derived from user interactions, represented as graphs [1]. This methodology operates under the assumption that polarized attention aggregates into distinct communities centered around influential users. Figure 1 visually illustrates this community division on a controversial topic using a user retweet graph.

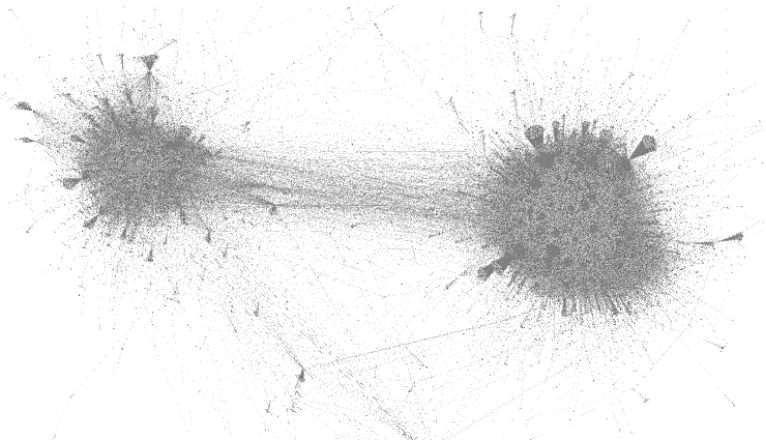

Fig. 1 User Retweet graph on the controversial topic PELOSI. The graph is represented using ForceAtlas2 [2] algorithm for spatial visualization.

Furthermore, researchers have explored the incorporation of textual content as a promising avenue for quantifying controversy. Recent studies have integrated textual information with structural data to enhance controversy classification tasks [3]. Notably, natural language processing (NLP) and deep learning techniques have been employed to enrich structural graph information and quantify controversy [4]. Understanding user behaviors in the context of impactful content has emerged as a significant challenge. Text analysis has gained attention in various applications, including fake news detection, email classification [5], and claims detection [6]. To unravel the relationship between word usage in a text and the cognitive and mental states of the author, psycho-linguists have developed the Linguistic Inquiry and Word Count (LIWC) tool [7]. This tool, examining thousands of dimensions, has been widely applied in research, such as identifying and analyzing claims [8] and complaints [9].

Remarkably, text analysis for controversy detection remains a relatively unexplored research direction. Notably, [10] stands out as the only work that delves into text analysis for controversy detection. This study investigates discussion features, including word usage and writing style, to gauge their predictive power for controversy and language sensitivity in Reddit posts. Word usage features aim to identify words more closely associated with controversy, considering their contextual relevance, while writing style features explore attributes such as text length, readability, part of speech (POS) tags, and sentiment analysis to differentiate between controversial and non-controversial text.

In this research, we explore the interpretability of controversy through the lens of SHAP (SHapley Additive exPlanations) [11], aiming to provide a fair assessment of the individual contributions of different text features in tweets to controversy detection. Notably, our work represents the inaugural endeavor to leverage SHAP to explicate the nuances of controversy.

SHAP, recognized as a fundamental advancement in explainable artificial intelligence, serves the purpose of comprehending the decision-making process of a given model. As a model-agnostic technique, SHAP is versatile and applicable to elucidate predictions generated by any existing machine-learning model. The foundation of SHAP lies in the Shapley value concept derived from cooperative game theory. This theoretical underpinning enables the fair distribution of the credit for a model's prediction among its contributing features. The term "fairly" is precisely defined mathematically, ensuring that the redistribution function of credit adheres to four key properties: efficiency (ensuring a complete distribution of the outcome among features), symmetry (guaranteeing identical rewards for features contributing equally), dummy (ensuring zero rewards for features that do not contribute to the outcome), and additivity (considering the additive rewarding of a feature in the presence of multiple game outcomes).

In summary, our study represents a pioneering effort in utilizing SHAP for the explainability requirements specific to controversies, shedding light on the intricate interplay of individual text features in the context of controversy detection.

Contributions. We are interested in studying the controversy of Twitter discussions. The subjectivity of such a concept is problematic, so we take a bigger point of view by analyzing it from the community perspective. Our contribution is then threefold.

1. We first **quantify controversy** on topics using both structural and textual properties, showing that textual information contains interesting features to help quantify controversy.
2. We propose a solution to **explain controversial topics**, through their communities, by investigating the contribution of features on different community-task classification models using SHAP. We investigate this solution by applying it to two relevant topics and show that the analysis generates interesting and promising results.
3. We finally investigate the **community evolution through time**, by looking at the contributing features of community prediction models in different timeframes of one controversial topic. Preliminary analysis shows interesting results and community behavior

The rest of this paper is organized as follows. Section 2 provides an overview of related work. Sections 3 presents our approach to analyse and explain controversy. Section 4 describes dataset we used in our approach evaluation. Section 5 presents and discusses the obtained results. Section 6 concludes the paper and highlights some future work.

2 Related Work

2.1 Explainability methods and techniques

Explainability of black-box machine learning systems is strongly needed to build user trust as its absence can negatively affect its adoption, mainly in some critical domains (Medical, etc.). Different terms are used to refer to explainability such as interpretability [12], transparency [13], understandability [14] and causability [15]. In this paper, we use explainability and interpretability terms interchangeably. Explainability methods are often categorized into model-agnostic and model-specific methods. The former can be used on any learning model, which is seen as a black box, when the latter is specific to a given ML model due to some assumptions made on it. Explainability methods are also categorized into global and local mehods. Global explainability methods explain the model as a whole when local explainability methods only explains a given prediction.

Global agnostic-models aims to show the impact of a given feature on the prediction. It includes two main categories of approaches. The first category aims to visualize the effects of a feature on a prediction model. It is mainly based on Partial dependence function which is used when features are not correlated to capture the marginal effect of a feature on the prediction. It allows to show whether the relationship between the feature is simple (linear) or complex. Individual Conditional Expectation (ICE) is another technique used to detect any eventual heterogeneous relationship between the feature and the prediction. The second category of approaches aims to measure the contribution of a feature to the prediction performance to show to what extent a feature is contributing to the prediction performance. Permutation Feature Importance (PFI) [16] and Leave-One-Covariate-Out (LOCO) [17] are the most used techniques n thsi second category. PFI technique calculates the feature importance when a feature

is randomly shuffled a number of times and replaced by a dummy feature in the training data, testing data or validation data. LOCO technique is similar to the PFI, it only differs by leaving the feature out instead of replacing its real values by a dummy ones.

Local agnostic-models methods are not concerned with explaining the whole learning model. They only target to explain individual predictions. Different mechanisms are used including individual conditional expectation (ICE) function to compute the dependence of a given feature on each prediction separately, influence function that estimates the role of a feature by perturbing the training samples, anchor explanation that provides if-then rules to specify conditions, when satisfied, will give to the same prediction. It is worth to note that SHapley Additive ExPlanations (SHAP) provides both global and local model-agnostic explanations. It uses the shapley value from game theory to compute to the contribution of a feature.

2.2 Explainability of Controversy

Controversy interpretation aims to provide users with arguments that can explain why a controversial topic is controversial. There has been little work on controversial interpretation. An unsupervised stance-aware summarization approach is proposed in [18]. It focuses on the Twitter platform and considers controversy interpretation as an optimization problem and proposes a ranking model that generates the top k tweets that best summarize the conflicting stances of the controversial topic. Each tweet is mainly ranked regarding its stance indication which measures to what extent the tweet represents the stance of the community it belongs to. Controversial topics are limited to those generating two conflicting communities. A community is characterized by a stance community statistically identified by a set of (stance) hashtags. A regression model is also used to predict the rank of tweets regarding their articulation (to what extent the tweet is well written) and topic relevance (to what extent a tweet is related to the controversial topic). Relying on stance hashtags makes the approach topic dependent. It also does not take advantage of some characteristics that could be common to different controversial topics. Considering that an unsupervised task is not sufficient for controversial text summarization as its arguments space is complicated, an unsupervised expert-guided contrastive opinion summarization is proposed in [19]. It mainly relies on aligning ordinary opinions present in tweets with expert prior opinions. Expert opinions could be provided either by external resources or by users' annotations. , and tackles the summarization by an heuristic approach instead of an optimization problem. This approach has at least two main limits. It first necessitates an important involvement of the user to extract prior knowledge and secondly could not easily support the emergence of new ideas and arguments in tweets. The predictive power of individual features for controversy on Reddit social media is studied in [10]. Word usage, writing style, sentiment, and user involvement were considered. The obtained results show involvement features (numbers of preceding/succeeding comments) carry the most predictive strengths.

Arguing that controversy detection should be language and topic independent to offer better performance, the graph analysis technique is also exploited in [20] to look for some local motifs that could characterize controversial graph interactions. Dyadic and triadic network motifs (local pattern of the user interaction) along with their

frequencies are extracted from retweet and reply graphs and used as a feature to predict controversy. User texts are unfortunately not analyzed and the approach outputs are used more as features to predict controversy than to explain the controversy.

2.3 Controversy detection and quantification

A substantial amount of work has been done on controversy detection on social media. Most of them exploit user graph interactions and partitioning algorithms to identify the two main conflicting communities. User graph interaction can be a simple graph [1] or an attributed graph [21] to take advantage of user attributes (number of tweets per user, number of followers, etc.). To limit the impact of the echo-chamber phenomena, the user graph is augmented by adding new edges that materialize connections between users with opposite views [22]. Although these approaches are language and domain-independent and can then be applied easily to any topic discussion, it nevertheless presents the drawback of not taking advantage of extra information. Some works attempted to overcome these limits by exploiting for instance named entities to infer the tendency nature (positive, negative, neutral) of users towards some given named entities [23], and user's vocabulary to cluster users with more similarities in their vocabularies [24]. Some recent works consider controversy detection as a graph classification problem [3]. Graph embedding techniques (GNN) and NLP techniques are used to combine the structure of users' interactions and text content of discussions by encoding the whole discussion graph (structure and texts) into low-dimensional and dense vector spaces. All these approaches aim to quantify/detect controversy on a topic, but they don't help to understand why a topic is controversial.

3 Approach description

As said above, we explore controversy from the text and community perspective. To explain controversial topics, we propose a pipeline composed of four components, as depicted in figure 2. Section 3.1 will describe how the graph is processed and communities constructed, while section 3.2 will show how we process text for each user. Section 3.3 will detail how we quantify controversy based on structural and textual inputs. Finally, section 3.4 will describe how we explain controversy through community analysis with SHAP.

3.1 Graph building and partitioning

A topic T is represented by a set of tweets (including retweets) $T = \{t_1, t_2, \ldots, t_r\}$. t_i denotes the i^{th} tweet of the topic T_j. Each topic T_j is represented as an undirected user retweet graph, where two users (nodes) are connected if one has retweeted the other. To ensure the reliability of the partitioning, only the biggest connected component is kept in the final graph, as small groups of users can be unconnected to others.

To label users by their respective communities for each topic, we rely on the work in [1], and use the partitioning algorithm metis [25] to partition each graph into two communities. We consider that we only have the pros and cons of communities, and

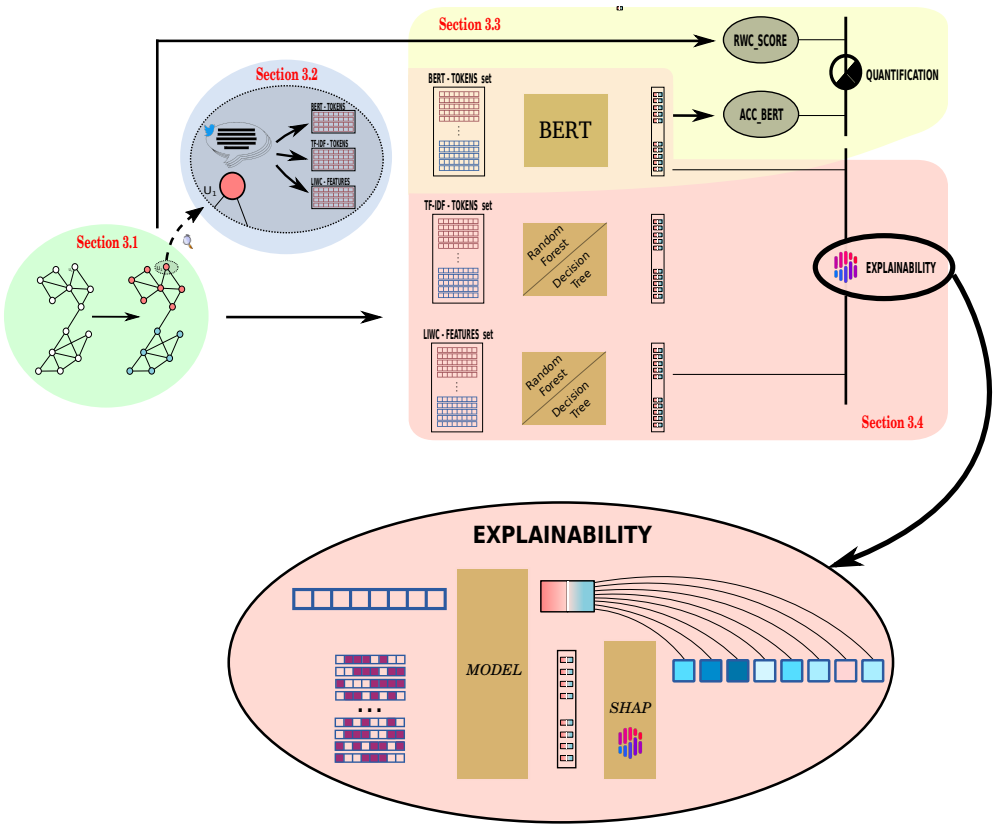

Fig. 2 Pipeline used for both quantifying and explaining controversy through our community analysis. The functioning of SHAP analysis in the Explainability section is detailed. The darker the blue (respectively red), the more the characteristic leads the model to predict the blue (respectively red) community, and conversely.

thus we do not take into consideration sub-communities. Each user is labeled by the community label (C_0 or C_1) it belongs to.

3.2 Text processing

Users gathered in the graph can be authors of one or multiple tweets, as well as none if they only retweet. Each tweet is labeled with the label of its original author (C_0 or C_1). Tweets from users that are not connected in the final connected graph are discarded from our analysis.

For each original tweet, different types of features can be created. In our approach, we considered three types of features generated from BERT, TF-IDF, and LIWC methods, but any other type of feature can be added. Finally, three sets of features are generated, BERT-TOKENS set, TF-IDF-TOKENS set, and LIWC-FEATURES set, according to their respective type of feature. The types of features that are used are presented below.

Table 1 Description of the first 2 levels of LIWC features.

1^{st} **level**	2^{nd} **level**
SUMMARY DIMENSION	WC (word count), WPS (word per sentence) BigWords, Dictionary word count, Analytic, Clout, Authentic, Tone
LINGUISTIC	Function (pronoun, determinant, adverb...), Verb, Adj, Quantity
PUNCTUATION MARKS	period, Comma, QMark, exclam, Apostro, OtherP
PSYCHOLOGICAL PROCESSES	Drives (affiliation, power), Cognition, Affect (emotion), Social (behavior & references)
EXPANDED DICTIONARY	Culture, Lifestyle, Physical, States, Motive, Perception, Time orientation, Conversation

3.2.1 Textual features

The first 2 sets of features are created from the pure textual contents of the tweet.
BERT-TOKENS. Based on a BERT tokenizer to pre-process data from text, we pulled the corresponding set of tokens. BERT tokenizer is a pre-processing step in BERT [26] models, which tokenizes input text by mapping each word to a unique index, adding special tokens to separate sentences, and encoding text using subwords for out-of-vocabulary words. It enables the input text to be passed into the BERT model presented in section 3.3.
TF-IDF-TOKENS. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical method to measure the importance of a word in a document compared to a corpus, by weighing the frequency of the term in the document against its rarity in the corpus. It is commonly used in information retrieval and text mining for feature extraction and text classification. We set the corpus dictionary from the same train set used for model classification, described in sections 3.3 and 3.4. Each tweet is tokenized independently.

3.2.2 Conceptual features

As introduced before, we aim to find meaningful explainable features that can help understand the controversy on social media.
LIWC-FEATURES. LIWC [7] analyzes textual content by helping to understand different psychological states such as thoughts, feelings, or personality, resulting in new insights, based on the statistical study of word use. These features are organized hierarchically and categorized in a hierarchy. The first 2 levels of features analyzed by LIWC are described in table 1 when a textual input has been given. Each feature has its own dictionary reflecting a psychological category of interest. The returned scores are computed from word count based on those dictionaries, and range between 0 and 100 (normalized by total word count), except for special features, such as word count (WC) or word per sentence (WPS). Each score is computed independently by tweet.

3.3 Controversy quantification

Based on community analysis, the controversy is quantified using structural and textual properties by looking into 2 different scores.

Structure-based controversy score. We first perform the Random Walk Controversy method (*rwc_score*) based on works from [1] on the graph described in section 3.1, focusing on user partitioning. This score has been chosen because it presents the best results among scores presented in [1]. The *rwc_score* is based only on structural information, generating multiple random walks from nodes of each community, and looking at the proportion of random-walk ending in the same community from where it started. A high *rwc_score* would correspond to better separate communities, thus a more controversial topic.

Textual-based controversy score. We secondly perform a community-based tweet classification only based on textual content from tweets. We base our work on a BERT-based model [26]. BERT is a machine learning model used for natural language processing. It is a transformer-based model, using multiple attention layers. BERT is pre-trained on a corpus of millions of text and is fine-tuned for our specific tasks. We split the set of tweets into 2 training and test sets, equally balanced between communities. The test set being equally balanced, we use the accuracy score of the test set *acc_bert* as the performance metric of the respective model A high *acc_bert* would correspond to a high capacity to predict communities using text, thus a more controversial topic.

Finally, we look at the complementarity of both properties, by multiplying both *rwc_score* and *acc_bert*.

3.4 Controversy explanation through communities

3.4.1 Statistical analysis of the generated textual features

A descriptive and statistical analysis of conceptual LIWC features on communities is presented and applied to topics labeled as controversial. The statistical analysis was performed using Matlab R0021b and the Statistics and Machine Learning Toolbox v12.2. Normality was tested using the Shapiro-Wilk parametric hypothesis test. For testing differences between groups one-way, Analysis of Variance (ANOVA) was employed when the assumptions of ANOVA were met. Otherwise, the Kruskal-Wallis non-parametric test was used. Linear correlation between variables was assessed using the Pearson product-moment correlation coefficient. Statistical significant correlations are considered as very strong if $|\rho| \geq 0.8$, as strong if $0.5 \leq |\rho| < 0.8$, and weak correlations otherwise.

For univariate analysis, Logistic Regression was performed. Significant influential outliers were removed, if the absolute value of the standardized residuals was above 3 and the Cook's distance above 4. The linear relationship between the continuous independent variables and the logit transformation of the dependent variable was assessed using a Box-Tindel test. Variables that failed to compile with the aforementioned assumption were not taken into consideration. Multivariate analysis was performed by

Multivariate Logistic Regression, using only the statistically significant variables from the Univariate analysis. Collinearity between variables was assessed by the Variance Inflation Factor (VIF). Variables having a VIF value above 5 were sequentially removed one by one, from the multivariate analysis, by considering variables in decreasing VIF order. Only variables having $VIF \leq 5$ were considered in the multivariate analysis. The statistical significance level was defined to 5% for all the tests.

3.4.2 SHAP-based analysis of classifier models

We now consider controversial topics presenting high *rwc_score* and *acc_bert* values computed by the controversy quantification component. This section assists us in analyzing, and explaining which features help tweet classification models towards one community rather than another. The more the topic will be seen as controversial, the better the community-based analysis of models will be. We analyze tweets from the test tweets set and originated from users of both communities, seeking to determine text features that can characterize communities. To analyze how much each text feature contributes to the tweet classification models, we rely on the SHAP method.

SHAP [11] draws its foundation from the collaborative game theory to explain a prediction/classification p(x) for a given instance x. Collaborative game theory can be viewed as a set of players who collaborate to achieve a common goal of the game and fairly divide the game reward. SHAP is a model-agnostic method. It can be used to explain any given prediction/classification model from its inputs and outputs. The explanation is given in terms of the marginal contribution of each feature value of the instance x to the p(x) output. In our case, the tweet classification model is the game, and tweet text features are the players. In this work, we consider three tweet classification models p: BERT, Random Forest (RF), and decision tree (DT) models. For each tweet classification model, we rely on its corresponding set of text features F as described in section 3.2. Given the same test tweet set of the topic T created in section 3.3, and the type of feature investigated F, we look at the marginal contribution of each feature.

From a coalition of features ($S \in F$), that does not contain the k^{th} feature f_k ($f_k \notin S$), the partial marginal contribution of the feature f_k for a given tweet t and a given classification model p is computed as

$$p\Big(t_{S\cup\{k\}}\Big) - p\Big(t_S\Big) \tag{1}$$

where $p(t_S)$ refers to the classification made by only using the features of the coalition S of the tweet t. All features that do not belong to the set S are discarded. Equation 1 represents the positive or negative benefit we obtain by adding the k^{th} feature to the features coalition S. Given a classification model p and its corresponding features F, the final marginal contribution of a k^{th} feature f_k of F for a given tweet t is denoted by $sv_k(p, F, t)$ and is computed by considering all possible features coalitions S as shown in equation 2. sv_k is called the shapley value of the k^{th} feature of F for a given instance tweet.

$$sv_k(p, F, t) = \sum_{S \subseteq Fn\{k\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[p(t_{S \cup \{k\}}) - p(t_S)\right] \tag{2}$$

The Shapley value [11] sv is computed for all single features, on all tweets of the test set. As shown in equation 3, the obtained result can be seen as a matrix $SV_{p,F}$ where $sv_{l,k}$ represents the contribution of the feature f_k for the tweet instance t_l. Each horizontal row of the matrix SV represents the contribution of the different features to the corresponding tweet classification model. Each vertical row represents the contributions of a given feature to the different tweet classification models. A mean of the vertical row values can be seen as the contribution of a given feature to the tweet classification model for the whole topic. Thus, each row of the matrix ensures the local explanation of a given tweet, where the whole matrix ensures the global explanation of the tweet classification model.

$$SV_{p,F} = \begin{pmatrix} sv_{11} & sv_{12} & ... & sv_{1m} \\ ... & ... & ... & ... \\ sv_{i1} & sv_{i2} & ... & sv_{im} \\ ... & ... & ... & ... \\ sv_{n1} & sv_{n2} & ... & sv_{nn} \end{pmatrix} \tag{3}$$

4 Dataset

Our work is based on Twitter and focuses on tweets related to several topics, controversial or not. We perform our analysis on 30 different datasets provided in [4], retrieved using the Twitter API. 15 topics have been manually labeled controversial and 15 non-controversial from multiple sources on mainstream media [4]. Non-controversial topics contain soft news such as entertainment or noticeable events with no controversy, while Controversial topics are mainly focused on political events (election, justice cases).

Each topic contains tweets retrieved from hashtags or keywords, corresponding to the respective event. Several pieces of information are retrieved by tweets, such as user-id, text, and retweet user information if recalled as a retweet. Only original tweets retweeted at least once are retained, as well as involved users. Notice that most users only retweet, and never publish original tweets. Tweets are cleaned up beforehand by replacing URLs and user tags with unique special tokens. From the data we got access to, some tweets might be missing in our datasets, depending on the topic, as tweets could have been deleted since the last time it was retrieved in [4]. The resulting dataset consists of 30 topics with their number of tweets ranging from 5 458 to 36 716, involving a number of users ranging from 3 696 to 161 612 per topic [24]. Table 2 resumes the frequencies of tweets, re-tweets, users, and users with at least one published tweet in the corresponding topic.

Community analysis, performed in section 5.4, is topic-independent. We present the 2 controversial and non-controversial topics used in our analysis.

Textual features used to explain communities being topic dependent, we based the explainability section (section 3.4) on only 2 topics for simplification purposes, one being controversial (PELOSI) and one non-controversial (THANKSGIVING). These topics

Table 2 Frequencies of data retrieved from the dataset of each different topic. *Users_with_T* contains users publishing at least one tweet. The first 15 topics represent controversial topics, whereas the last 15 non-controversial topics.

Topic id	Tweets	RT	Users	*Users_with_T*	main keyword (Description) [LANG]
IMPEACHMENT-5-10	29 708	48 899	17 241	5026	Roussef impeachment [31oct–10nov, 2015]
MENCIONES-1-10ENERO	14 267	48 220	24 273	5587	Macri's mentions [1–11jan,2018]
MENCIONES-11-18MARZO	19 927	57 766	30 001	6747	Macri's mentions [11–18mar, 2018]
MENCIONES-20-27MARZO	13 509	67 468	32 562	5808	Macri's mentions [24–26mar, 2018]
MENCIONES-05-11ABRIL	31 741	142 285	59 799	11 011	Macri's mentions [5–10apr, 2018]
MENCIONES05-11MAYO	24 157	143 506	17 543	10 156	Macri's mentions [5–10may, 2018]
BOLSONARO27	10 972	86 719	43 118	5578	Brazilian elections [27oct, 2018]
BOLSONARO28	14 629	113 092	71 139	10 051	Brazilian elections [28oct, 2018]
BOLSONARO30	14 487	127 802	68 585	7461	Brazilian elections [30oct, 2018]
KAVANAUGH06-08	16 363	120 276	67 226	8410	Kavanaugh's nomination [8oct, 2018]
KAVANAUGH16	18 109	129 383	63 389	9519	Kavanaugh's nomination [3oct, 2018]
KAVANAUGH02-05	18 545	143 692	71 543	10 153	Kavanaugh's nomination [5oct, 2018]
LULA_MORO_CHATS	18 807	142 585	65 009	8921	Lula's mentions [1] [10–11jun, 2019]
LEADERSDEBATE	30 352	172 882	74 015	10 820	Candidates debate [11–21nov, 2019]
PELOSI	15 517	207 810	93 262	9122	Trump Impeachment [6dec, 2019]
AREA51	5458	153 285	102 426	3696	Jokes about Area51 [3–13jul, 2019]
OTDIRECTO20E	39 417	95 056	24 246	7561	Music TV program [13–20jan, 2020]
VANDUMURUGANAJITH	6088	113 092	8022	2026	Ajith's fans [23jun, 2019]
NINTENDO	20 669	102 431	88 928	6528	Nintendo's release [19–28may, 2019]
MESSICUMPLE	13 870	125 407	93 842	8136	Messi's birthday [23–24jun, 2019]
WRESTLEMANIA	36 716	104 496	57 861	10 515	Wrestlemania event [8apr, 2019]
KINGJACKSONDAY	20 077	107 063	38 529	11 457	popstar's birthday [24–27mar, 2019]
NOTREDAM	13 512	143 213	94 280	7147	Notredam fire [16apr, 2019]
THANKSGIVING	19 043	136 951	110 279	14 939	Thanksgiving day [28nov, 2019]
HALSEY	12 772	203 492	96 756	8203	Halsey's concert [7–8jun, 2019]
FELIZNATAL	23 451	193 499	161 612	18 940	Christmas wishes [25–26dec, 2019]
EXODEUX	18 355	135 303	36 174	6647	EXO's new album [7nov, 2019]
BIGIL	6655	171 092	25 217	4050	Vijay's birthday [21–22jun, 2019]
CHAMPIONSASIA	13 644	143 181	65 428	6582	Al-Hilal champion [24nov–1dec, 2019]
SEUNGWOOBIRTHDAY	20 507	192 865	17 543	5974	Segun Woo singer birthday [23dec, 2018]

Table 3 Descriptive statistics on the 2 communities retrieved for PELOSI and THANKSGIVING datasets.

-	PELOSI			THANKSGIVING		
	C_0	C_1	Total	C_0	C_1	Total
Tweets	10 430	5087	15 517	5531	13 512	19 043
Users	48 032	45 230	93 262	55 141	55 138	110 279
Users who tweet	5900	3222	9122	4781	10 158	14 939

have been chosen because they present high *rwc_score* and *acc_bert* scores. We present statistics of both datasets in 3.

PELOSI. Topic labeled as controversial regarding Nancy Pelosi's speech in Congress about former US president's Donald Trump first impeachment, on December 19, 2019 (Trump is blamed for abuse of power and obstruction of Congress). The speech, pushing for Trump's impeachment, is criticized for multiple reasons, by people defending the former president Donald Trump, but also the ones opposed Pelosi's positions, especially about being against abortion. Two major communities are represented, one "pro-Pelosi", where users support Nancy Pelosi, and one we called "against-Pelosi", where users are either against Pelosi or supporting Donald Trump. After performing user partitioning presented in section 3.1, and randomly checking tweets, we have

noticed that community C_0 (labeled 0) tends to represent people against the congresswoman Pelosi, anti-democrats, whereas community C_1 (labeled 1) comprises users either pro-Pelosi, against Trump, or pro-impeachment.

THANKSGIVING. Topic labeled as non-controversial gathering tweets referring to Thanksgiving 2019, a US annual national holiday to celebrate the harvest and other blessings of the past year.

5 results

We applied the first 3 steps of our method on the 30 topics presented in section 4 for controversy score quantification needs. However, as explaining how communities are represented is topic-dependent, the controversy explainability part of our method will be only performed on 2 topics that show high *rwc_score* and *acc_bert* scores. These two topics are labeled as controversial (PELOSI) and non-controversial (THANKSGIVING) respectively.

5.1 Graph processing

A fully connected graph is built from each of the 30 topics and partitioned into two distinct communities C_0 and C_1. User proportion ($CPROP$) between C_0 and C_1 is computed independently for each topic as per equation 4.

$$CPROP = \frac{min(|C_0|, |C_1|)}{max(|C_0|, |C_1|)} \tag{4}$$

The range of user proportion of our different graphs is large and varies from 0.05 to 0.99 with an average of 0.54. This shows clear structural differences between the different considered topics.

5.2 Text processing

We extracted for each topic three different text feature sets, namely the BERT-TOKENS set, TF-IDF-TOKENS set, and LIWC-FEATURES set. They will be used by our classification models. The LIWC features are retrieved using the LIWC app [2] on each tweet independently. Note that several topics are in different languages, we translate into English each tweet coming from other languages independently, using the deep translator python library [3], combined with the Google Translator algorithm.

5.3 Controversy quantification

We compare topic-related properties on our 30 topics independently. To better quantify the overlap between different scores of controversial and non-controversial topics, the sensitivity of model accuracies is measured using the area under the ROC curve (AUC ROC), 1 representing a perfect separation between topics, while 0.5 indicates indistinguishable communities. We intend to see if from a community perspective,

[2] https://www.liwc.app/
[3] https://pypi.org/project/deep-translator/

texts, in addition to structural information, can provide information about controversy, as well as find out if tweets of controversial topics from each community are easier to generalize and classify by our models.

Structure-based score. Based only on structural properties, the *rwc_score* is computed for each of the 30 topics presented in section 4. We retrieve a final AUC ROC score and obtain a high score of 0.88. This shows a good separation between topics, and from graph information, controversial topics show a similar behavior compared to non-controversial ones.

Textual-based score. For each topic t, the respective set of tweets X is split into two equally balanced train and test sets, using a ratio of 0.8.Based only on textual properties, *acc_bert* score represents the accuracy score on the test set for each topic. Concerning the BERT-based model used for classifying tweets, we extracted all 12 transformer layers and added an extra layer on top for classification. The model is trained until the training loss stops decreasing, with a learning rate of $2e^{-5}$. We optimized the model with Adam optimizer, using a decreasing learning parameter to avoid losing too much information from the first transformers-layers. We obtain an AUC ROC of 0.79 concerning the *acc_bert*, a high score which shows more generalizable tweets on communities on controversial topics, recalling better performance. We notice that some topics present significant user imbalances between communities, especially for the non-controversial ones. Looking only at topics having two strong communities with user proportion $CPROP$ higher than 0.2 (25 topics remaining), the AUC ROC score rises to 0.90 on *acc_bert* and reaches 0.91 on different *rwc_score*. Finally, when combining both *rwc_score* and *acc_bert* for each topic, we reach an AUC ROC of 0.91, which shows that both textual and structural information can be complementary in controversy quantification. Moreover, we notice that both *rwc_score* and *acc_bert* show similar behavior on ambiguous topics. They both struggle on the same non-controversial topic THANKSGIVING, having high scores, while the controversial topic LEADERSDEBATE presents 2 low values for both scores. That reinforces our conclusion that both text and user interactions contain useful information on the controversy.

5.4 Controversy explanation

5.4.1 Controversial statistical analysis

A descriptive statistical analysis (correlation and differences between groups), presented in section 3.4.1, was performed on the controversial topic (PELOSI dataset), for highlighting the insides of the dataset and better understanding the linguistic differences between the two communities. In this analysis, we used the LIWC features. The independence of the sample's observations was ensured by performing two preprocessing steps: (1) tweets are grouped by user, since a user can tweet multiple tweets, and the mean value of each LIWC feature is calculated, resulting in one observation per user and (2) all users participating into more than one topic have been discarded from the dataset (The amount of users discarded is less than 7%).

Correlation Analysis. Besides the obvious positive correlations that exist between variables belonging to connected hierarchical levels (i.e. having parent-child relations), we identified some interesting statistically significant correlations between variables. We need to note here that out of 101 parent-child feature relationships, only 17 were found to be statistically significantly correlated. A strong correlation exists between *Dic – Linguistics* variables ($\rho = 0.81232, p < 0.001$), which indicates the appropriateness of the dictionaries used in LIWC for capturing linguistic aspects. A more obvious correlation exists between prosocial behavior (Altruistic, helpful) and politeness: *prosocial – polite* ($\rho = 0.5336, p < 0.001$), although these two features do not belong to the same hierarchy. Another interesting negative correlation exists between *Clout* (the language of leadership, status) and *Authentic* (perceived honesty and genuineness) ($\rho = -0.3177, p < 0.001$) suggesting that users who speak about leadership and status are less polite. Finally, *negative tone* (including notions like bad, wrong, and too much hate) is correlated to *emotion* (including notions like good, love, happiness, and hope) suggesting that these opposite feelings coexist.

Differences between groups. The analysis of the means between the two different communities revealed some interesting facts. In the first place, out of the 117 features of LIWC-22, only 29 did not have statistically significant differences between the communities. For the *Summary* variable group, *Analytical thinking*, *Authentic* (perceived honesty), and *percentage of words having 7 letters or above* did not have statistically significant differences between the two groups. In terms of linguistic features, the use of 1st singular person (-0.591, p¡0.001), 3rd singular person (.2338, p=0.014) as well as 3rd person plural (0.2909, p¡0.001) have statistically significant differences between communities, while 1st plural or 2nd person mentions had no statistical differences between communities, where the differences are referring to C0-C1 means. The *psychological processes* group variables related to *Cognition* (0.3428, p=0.002), *positive ton* (-0.6473, p¡0.001), *negative tone* (0.7351, p¡0.001), *positive emotions* (-0.3333, p¡0.001), *anger* (0.1106, p=0.003), *female* (0.439, p¡0.001) or *male* (-0.3433, p¡0.001) have statistically significant different means between the two communities, while variables referring to *Insights*, *Differentiation*, *Emotion*, *Anxiety*, *Sadness*, *Prosocial behavior*, *Interpersonal Conflict*, or *Moralization* did not have statistically significant differences between communities. In the *Expanded Dictionary* category features, features referring to *Politics* (0.0179, p=0.002), *Ethnicity* (0.2971, p¡0.001), *Lifestyle* (0.2361, p=0.001), *Religion* (0.53895, p¡0.001), *Physical status* (e.g. medicament, food, health, illness, etc.) (0.62998, p¡0.001), *Sexual* mentions (0.1126, p¡0.001), or *Death* (0.073, p¡0.001) as well as features reflecting the focus of the user on the past (-0.5325, p¡0.001), the present (0.5382, p¡0.001) or the future (0.2594, p¡0.001) have statistically significant differences between the two communities. On the other hand, features that do not have statistically significant mean differences between the communities, include variables related to *Technology*, *Home*, *Acquire* (get, got, etc.), *Fatigue*, *Curiosity*, *Allure*, *Attention*, *Space*, *Feeling*, and *Non-fluencies*, giving us an indication that these features are not different among the two populations. Finally, punctuation features like the use of *Question* (0.30206. p¡0.001) or *Exclamation* (0.4118, p¡0.001) marks as well as the use of *Apostrophes* (-0.543, p¡0.001) have statistically significant differences between the two communities.

Univariate Analysis. Evaluating the predictive capability of the features, concerning the community prediction task, we performed a univariate Logistic Regression analysis. Out of 117 variables, 46 variables showed statistically significant contributions to the class prediction problem. From the variables referring to the Linguistic dimension, the 3rd person plural (*they*) variable impacted the prediction towards the C1 community (OR=1.1165 (1.0853, 1.1487), p¡0.001). Other statistically significant variables (e.g. *pronoun*, *you*, *number*, *prepositions*, *verb*, *adjective*, etc.) had an OR close to 1 in the range [0.9747, 1.0774]. The same applies to the Psychological processes features, where the 9 statistically significant features (e.g. *cognition*, *negative ton*, *family*, *female*, etc.) had an odds ratio near 1. The most interesting results were obtained for the Expanded dictionary variables where variables referring to *ethnicity* (OR=1.134(1.098, 1.1703), p¡0.001), *illness* (OR=1.222(1.096, 1.362), p¡0.001), *wellness* (OR=1.348 (1.041, 1.743), p=0.02), *mental state* (OR=1.814(1.407, 2.338), p¡0.001), *substances* (OR=1.14(1.256, 1.583), p¡0.001), *food* (OR=1.152(1.081, 1.228), p¡0.001) and *death* (OR=1.14 (1.07, 1.214), p¡0.001) showed a significant prediction ability towards the C1 community. Finally, the question and exclamation mark used were statistically significant predictors for the C1 community, having odds ratios of 1.05 and 1.01 respectively.

Multivariate Analysis. We, furthermore, evaluated the predictive capability of the statistically significant variables, as described in section 3.4.1, operating together in a linear regression model. Out of 46 statistically significant variables, 26 contribute to the community classification task. The majority of the variables, including variables from all categories, showed an odds ratio near 1. Again, the 3rd plural mentions (OR=1.176 (1.139,1.215), p¡0.001) belonging to the Linguistic Features and mentions on *wellness* (OR=1.43 (1.086, 1.883), p=0.01), *mental state* (OR=1.39 (1.144, 1.690), p=0.001) and *substances* (OR=1.24 (1.114, 1.398), p¡0.001), belonging to the Expanded Dictionary variables, had the most significant impact on the predictions of the user's community towards the C1 community. Finally, question and exclamation marks utilization, had a statistically significant impact on the community classification task, having odds ratios of 1.068 and 1.02 respectively.

We now investigate the community-based explainability step of our pipeline, reported in section 3.4, on the same PELOSI dataset, as well as a non-controversial one presenting high quantifying scores, THANKSGIVING, presented in section 4.

5.4.2 A controversial topic community-based analysis

PELOSI, labelled controversial, presents a $rwc\_score = 0.70$, $acc\_bert = 0.79$ and a combination score of 0.55. As shown in figure 1, we notice 2 separate communities, where users are strongly related to each other while being less related to the other community, which explains the high rwc_score. Table 4 shows the results of experiments applied to this topic. Our BERT-based model, considered as state-of-the-art in language modeling, can distinguish tweets coming from users in different communities with an 0.79 accuracy and exceed the performances of both *DT* and *RF* models using word features (TF-IDF). These results clearly show that the text contains impactful information on community analysis.

Table 4 Accuracy metric t on different combinations of model and features applied on $test_{pelosi}$, for the community-based classification task on topic PELOSI.

Model	Features	ID	**Accuracy**
DECISION-TREE	TF-IDF	DT_{tfi}	0.65
	LIWC	DT_{liwc}	0.62
	TF-IDF + LIWC	$\text{DT}_{tfi+liwc}$	0.66
RANDOM-FOREST	TF-IDF	RF_{tfi}	0.69
	LIWC	RF_{liwc}	0.68
	TF-IDF + LIWC	$\text{RF}_{tfi+liwc}$	0.71
BERT	TEXT	BERT_{text}	**0.79**

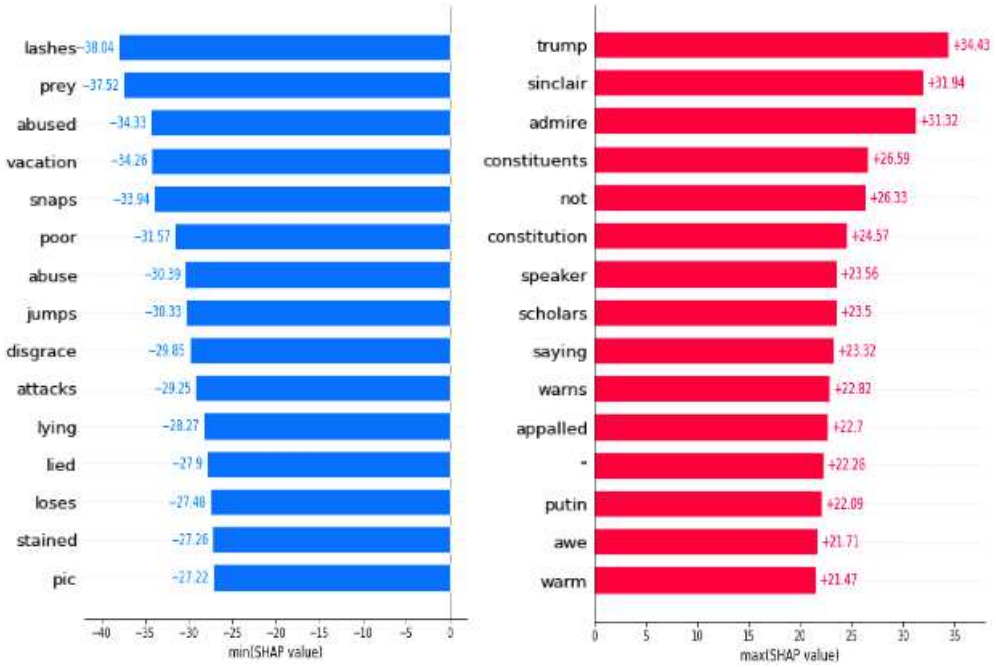

Fig. 3 SHAP values are computed from tweets on $test_{pelosi}$, with BERT_{text}. Values correspond to tokens' impact on predicting one community. The figure shows the top-10 Tokens impacting prediction towards C_0 (left) and C_1 communities.

Analysis of impactful tokens (words) based on BERT_{text} reinforces our conclusion, where BERT_{text} well-captured community-related features. Figure 3 shows the tokens with the most impact in predicting communities on $test_{pelosi}$ set. As expected, it highlights that words with negative connotations and pejorative tendencies ("abuse", "disgrace", "lying", "loses", "stained") strongly push the classifier to predict that the tweet belongs to the community C_0 of users attacking Pelosi. Some other tokens also emphasize conspiracies ("lashes", "snaps", "attacks"), probably against Trump. On the contrary, tokens representing positive qualifying adjectives ("admire", "warm", "awesome", "speaker") tend to impact the model strongly towards the community C_1,

containing users defending Pelosi. Since the purpose of C_1 is to promote Nancy Pelosi, it makes sense to have positive adjectives that describe her, unlike community C_0. It is worth remarking that tokens that are specific to the topic can also be representative of potential arguments of a community. Such tokens include 'constitution' (use of laws to request Trump's impeachment) or even 'Sinclair', a media from which a controversial question is drawn to embarrass Pelosi. Finally, we can also observe that users from C_1, at least compared to C_0, have more tendency to tweet or retweet by using the token "" to quote others.

The most impactful words in the training set, grouped in figure 4, are also analyzed, in order to understand how the model learns to correctly predict communities. We found that the lexical fields of the words around the communities are very similar to the analysis performed previously on the test set. This shows that a different way of communicating exists between these two communities, via distinct lexical fields.

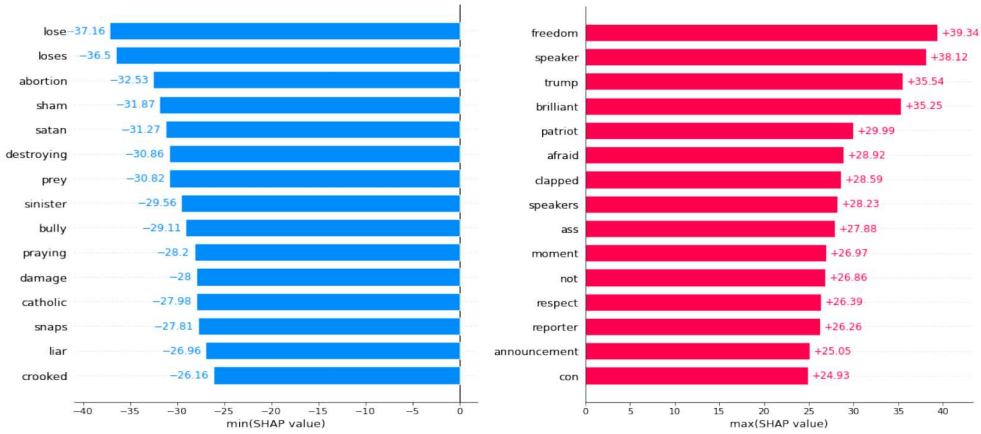

Fig. 4 SHAP values are computed from tweets of the training set $train_{pelosi}$, with BERT_{text}. Values correspond to tokens' impact on predicting one community. The figure shows the top-10 Tokens impacting prediction towards C_0 (left) and C_1 communities.

Concerning psychological states LIWC features, we reach a 0.68 accuracy on RF_{liwc}, which even goes up to 0.71 when combined with word TF-IDF features. Based on those results, we can assume that in this controversial case, LIWC can help characterize a tendency in a community in relation to another. Looking for psychological state tendencies in communities [4], figure 5 shows that top LIWC features impacting the prediction of RF_{liwc} on $test_{pelosi}$ are from the categories "Tone", "function", "Period", "Exclam", "OtherP", "Cognition", "Affect", "Social", "Lifestyle", "Physical" and "Time orientation", presented in table 1. We notice that punctuation plays an important role ("Exclam", "OtherP", "period"). The token "!" for instance shows a high impact on predicting community C_0, which is consistent, since users attacking

[4] a "perfect" feature would represent 2 well-separated clusters of colors, far away from the decision boundary.

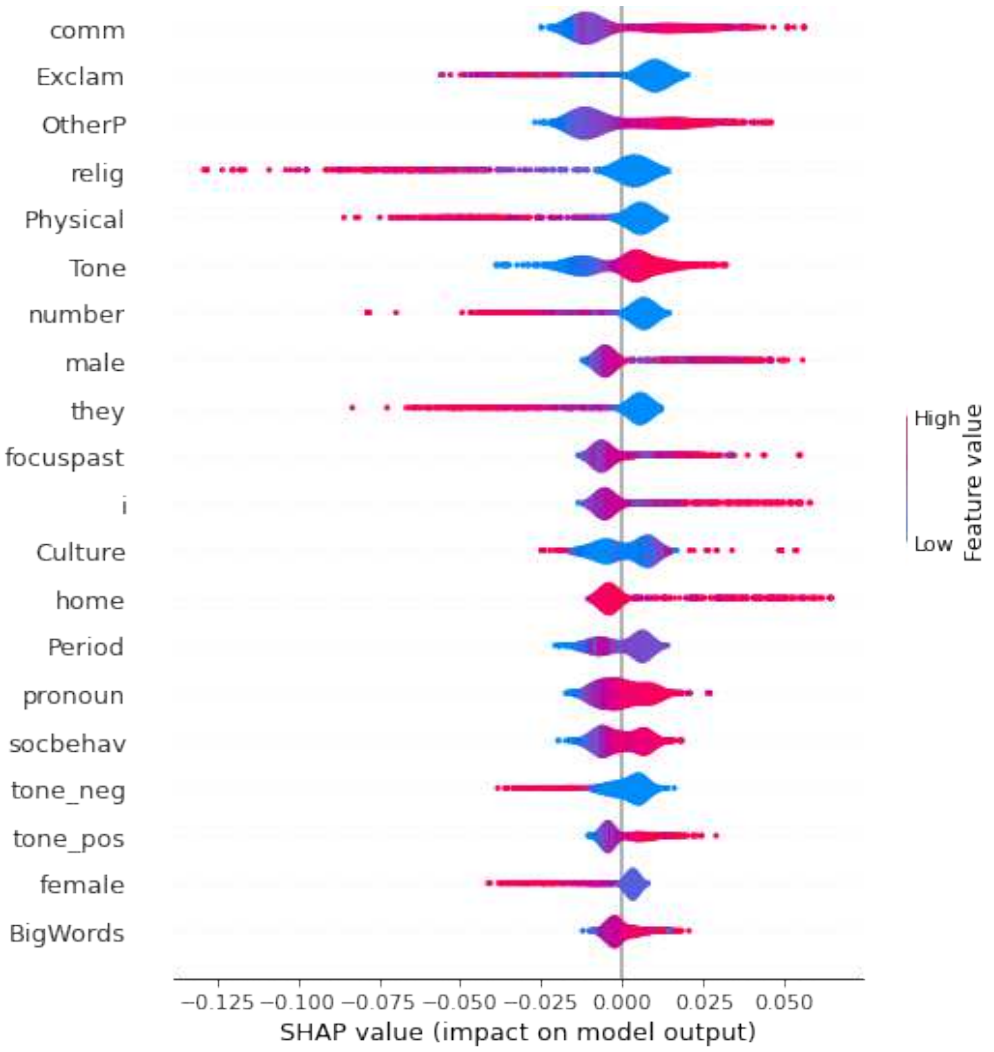

Fig. 5 20 most impacting LIWC features for model RF_{liwc} predictions on $test_{pelosi}$ set. The color scale, red (low feature value) to blue (high value) is represented for each sample. The larger the absolute SHAP value is, the more the feature pushes the model to predict the tweet to C_1 (inversely to C_0).

Pelosi, usually use strong feelings or emphasis on their tweets. Figure 5 also indicates functions like pronouns ("I", "They") or numbers impact model predictions. "They" tends to positively impact C_0 prediction compared to the 1st person singular ("I"). We can also pay particular attention to the tone and emotions felt in each community. We notice that tone is a very impacting and discriminating feature of the model. The rather inverted curves of the positive ("tone_pos") and negative ("tone_neg") tones reveal it, where C_1 users, who support Pelosi, are more likely to use a positive tone

than the C_0 community, which usually employs a more dramatic or polemic tone. This matches our conclusions regarding the analysis of BERT_{text} previously made. Finally, we notice that variables "home", "period" and "BigWords", have statistically no differences between communities, but are still identified as major contributors for our classifier (figure 5), showing interesting behavior of our SHAP-based approach.

Regarding the classification, we notice that all our models tend to slightly better predict and recognize tweets from the community against Pelosi (C_0) than the other (C_1), regardless of the features used, with for instance a 0.89 accuracy using our BERT-based model for C_0 against only 0.86 for C_1. It is also suggested by 5, where it seems that SHAP values have much higher absolute values, impacting in a stronger way the model towards class C_0.

To conclude, regarding the proportion of users and tweets by communities in PELOSI, table 3 shows that in this politically controversial topic, the community "attacking" the matter of the topic (C_0) is more prominent than the defending community (C_1). This being only a partial and simplified interpretation, further analysis could be developed from this impact analysis around this controversial topic, helping the overall understanding of the diverse communities.

5.4.3 A non-Controversial topic community-based analysis

The following non-controversial topic THANKSGIVING presents a $rwc\_score = 0.78$, $acc\_bert = 0.74$, and a combination score of 0.55. Moreover, this topic shows 2 strong communities (proportion $CPROP$ is higher than 0.2) while being labeled as non-controversial. This topic has been chosen for investigation, to understand what misleads the quantification of both controversy scores, especially the BERT-based model for predicting correct communities of tweets.

By plotting the graph, using the same force-layout algorithm used for PELOSI in figure 1, we notice that the community C_1 has users that are extremely related to one another, while the other has more distant users. This could explain the excessively high rwc_score. However, the 2 communities do not seem very distant, compared to the topicPELOSI. Secondly, from experiments made using the BERT-based model BERT_{text} on the test set, we recall a 0.74 accuracy (acc_bert score). By training a random-forest with LIWC features ($\text{RF}_{tfi+liwc}$) on the same test set, we obtain a 0.70 accuracy. Based on the same analysis presented in section 3.4, Figure 6 shows the most impactful features using the BERT-based model. We notice that if C_0 contains words/tokens that do not necessarily belong to a common category, C_1 contains 7 politic-related words (e.g. "president", "politics", "trump"). C_1 users seem to talk more about politics (while being strongly related to one another), suggesting that the topic might be related to some controversial sub-topic. C_0, on the opposite, seems to be more relaxed, without gathering users on a particular domain. This can explain the topic's high capacity for community-classification tasks, compared to non-controversial topics. We remark that "politic" belongs to the top-20 most impactful features, based on SHAP, in $\text{RF}_{tfi+liwc}$.

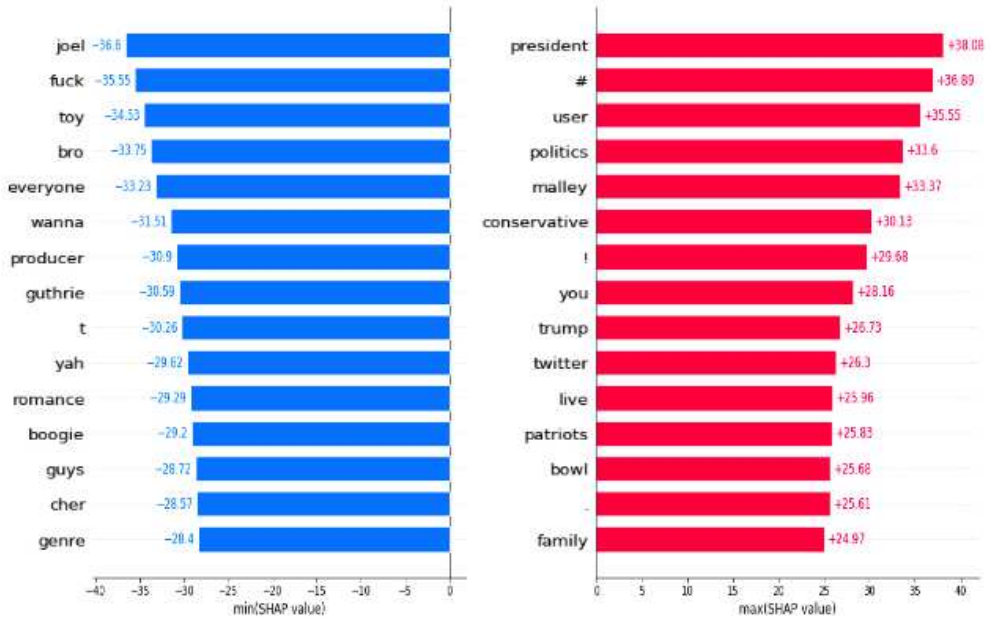

Fig. 6 The top-10 tokens contribution of $test_{thanksgiving}$ set on BERT$_{text}$, based on SHAP values.

5.5 Evolution of controversy through time

To analyze the temporality and potential evolution of communities, we apply the same approach to the Pelosi dataset, labeled as controversial, by splitting it into two distinct sets. The first dataset, *pelosi_before*, encompasses all tweets and retweets before the commencement of Pelosi's speech. The second dataset, *pelosi_after*, comprises all tweets and retweets after the commencement of the speech. The objective is to examine the evolution of communities before and after the speech to study the behavior of communities surrounding the controversial event.

Indeed, the evolution of contributing community characteristics allows for a better understanding of the controversy's development, as well as the study of the impact of an event on the subject. We will analyze the most contributing features of each community before and after the speech. To achieve this, we will precisely apply the same approach as demonstrated in Section 3. Each type of model will be trained for each of the two datasets, and the SHAP analysis will be applied to the test sets created for each dataset.

Figures 7 and 8 aggregate, for each community created from the graph partitioning, the evolution of textual features of the most contributing test sets for predicting communities of BERT models before and after Nancy Pelosi's speech. In general, it can be observed that, for both communities, the predictive strength of the most contributing words is significantly higher after the speech than before (the average score on the top 15 words/tokens for communities $c0$ and $c1$ being 7.88 and 4.89 before the speech, compared to 15.1 and 12.58 after). This indicates that the model

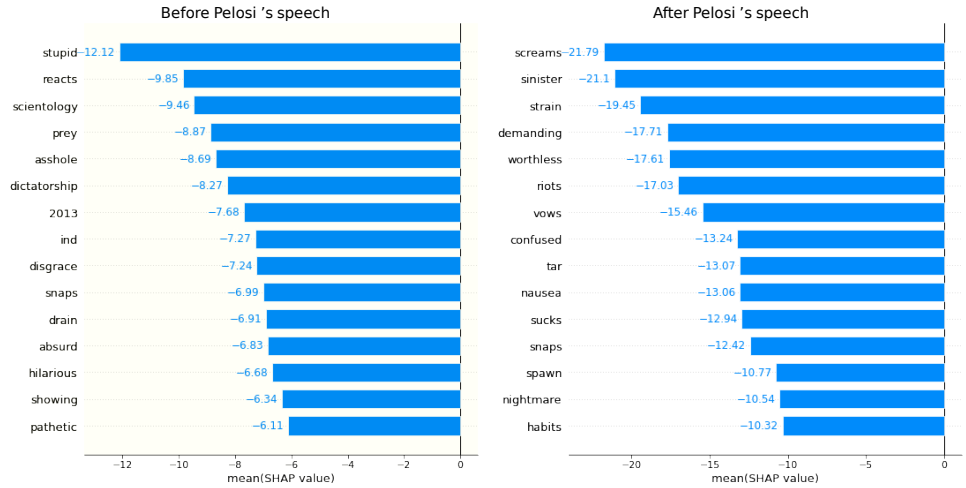

Fig. 7 The top-10 tokens contribution for the community $c0$ of the $pelosi\_before$ (left) and $pelosi\_after$ test sets on the BERT_{text}, based on SHAP values.

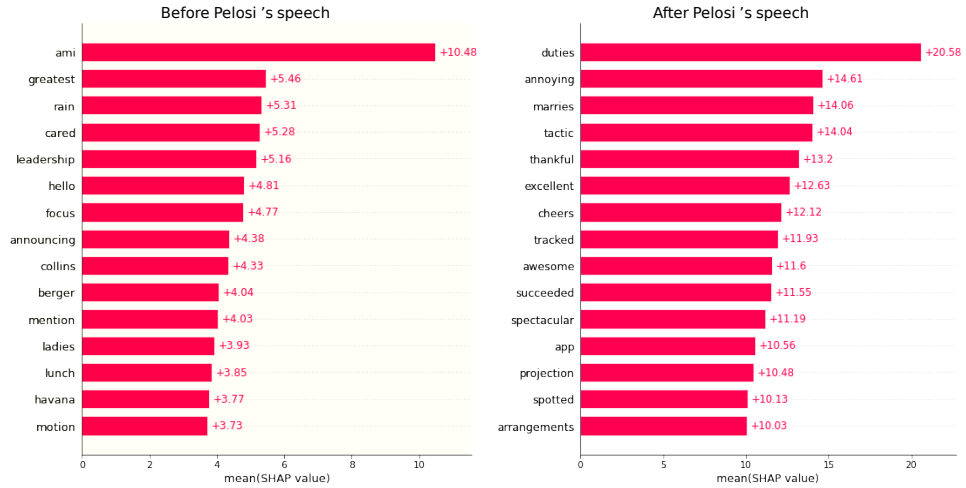

Fig. 8 The top-10 tokens contribution for the community $c1$ of the $pelosi\_before$ (left) and $pelosi\_after$ test sets on the BERT_{text}, based on SHAP values.

is better able to recognize features for classifying tweets, potentially representing a better portrayal of polarized communities on the subject.

Regarding community C0, which predominantly comprises anti-Pelosi users, as shown in Section 5.4.2, it can be observed that the vocabulary used is different: only one common word ('snaps') is used, clearly indicating a different lexical field before and after the speech. An interesting analysis also involves the significance of the word

"absorption" in the model's prediction for this community after the speech, indicating that this topic was one of the controversial points surrounding her speech.

For community c1, it is notable that after the speech, the most important tokens/words revolve more around the quality of the speech ("excellent," "thankful," etc.), whereas before Pelosi's speech, these words were more focused on the context of the event. This preliminary analysis of communities allows for the extraction of behavioral trends within controversial subjects.

6 Conclusion

This paper presented a controversy analysis pipeline on Twitter to quantify controversial topics and explain controversy through a community perspective. We relied on the use of different sets of text features and on the well-founded SHAP method to better identify the contributions of text features in three distinct tweet classifiers. Experiments we conducted show that the community-based explanation works well on topics having high *rwc_score* and *acc_bert* scores, even if non-controversial topics can also have structured communities being easily identified, without being controversial. This confirms that apart from the fact that controversy is a subjective notion, controversy should be considered in a fuzzy and non-binary way, and quantifying it could help people understand to what extent a topic is controversial. Moreover, this analysis shows that text has also interesting features, and is complementary to user interactions on controversial topics.

Furthermore, experiences have shown that the analysis can be extended to explore user behaviors within communities, incorporating a temporal aspect. Indeed, studying the behavior stemming from specific events can provide insights into the evolution of these communities.

The study is based on 30 topics, and it is then not easy to generalize its results. Nevertheless, we proposed a general pipeline to analyze controversy from the community perspective and showed some tendency over controversial topics. Moreover, our interpretation is based on weak user labels, even if the partitioning method has recently shown good results [1].

Extending the temporal analysis of communities by delving into the analysis and prediction of controversy evolution remains an interesting perspective, by exploring the identification of precursor signals for controversies within discussions [27]. This subject could empower decision-makers in various fields to anticipate and mitigate potential controversies. The methodology could involve combining user interactions with the content of their messages and leveraging Graph Neural Network (GNN) techniques [3] to measure and quantify the evolution of the user graph over time. The integration of spatial and temporal dimensions is particularly compelling, as it allows for a more comprehensive tracking of controversy growth. By examining the propagation of information among users and extracting sequential patterns from temporal series, we aim to enhance our understanding of controversy dynamics.

Several sub-problems arise in this context, such as predicting potential actions and intentions of new users is a valuable aspect of our research [28], or anticipating the trajectory of controversies holds significant potential for improving the effectiveness

of involved entities [29]. Similar studies have been conducted in various social media domains. In the financial sector, machine learning methods applied to social media data are commonly employed to predict stock prices [30, 31]. Extending this forecasting capability to online debates, our research aligns with efforts to predict the outcomes of contentious discussions [32].

References

[1] Garimella, K., Morales, G.D.F., Gionis, A., Mathioudakis, M.: Quantifying controversy on social media. ACM Trans. Soc. Comput. **1**(1), 3–1327 (2018)

[2] Jacomy, M., Venturini, T., Heymann, S., Bastian, M.: Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PloS one journal **9** (2014)

[3] Benslimane, S., Azé, J., Bringay, S., Servajean, M., Mollevi, C.: Controversy detection: A text and graph neural network based approach. In: 22nd Conference on Web Information Systems Engineering, vol. 13080, pp. 339–354 (2021)

[4] Zarate, J.M.O., Giovanni, M.D., Feuerstein, E.Z., Brambilla, M.: Measuring controversy in social networks through NLP. In: 27th International Symposium on String Processing and Information Retrieval, SPIRE, Orlando, USA, October 13-15, 2020, vol. 12303, pp. 194–209 (2020)

[5] Iqbal, K., Khan, M.S.: Email classification analysis using machine learning techniques. Applied Computing and Informatics (2022) https://doi.org/10.1108/ACI-01-2022-0012

[6] Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., Slonim, N.: Context dependent claim detection. In: 25th International Conference on Computational Linguistics: Technical Papers, pp. 1489–1500

[7] Boyd, R., Ashokkumar, A., Seraj, S., Pennebaker, J.: The development and psychometric properties of liwc-22 (2022)

[8] Nakov, P., Barrón-Cedeño, A., Martino, G.D.S., Alam, F., Kutlu, M., Zaghouani, W., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Mubarak, H., Nikolov, A.: Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets (2022)

[9] Preoţiuc-Pietro, D., Gaman, M., Aletras, N.: Automatically identifying complaints in social media. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5008–5019. Association for Computational Linguistics, ??? (2019). https://doi.org/10.18653/v1/P19-1495

[10] Koncar, P., Walk, S., Helic, D.: Analysis and prediction of multilingual controversy on reddit. In: Web Science Conference 2021, pp. 215–224 (2021)

[11] Lundberg, S., Lee, S.-I.: A Unified Approach to Interpreting Model Predictions (2017)

[12] Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B.: General pitfalls of model-agnostic interpretation methods for machine learning models. In: xxAI - Beyond Explainable AI - International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers. Lecture Notes in Computer Science, vol. 13200, pp. 39–68 (2020)

[13] Barclay, I., Preece, A.D., Taylor, I.J., Verma, D.C.: Quantifying transparency of machine learning systems through analysis of contributions. CoRR **abs/1907.03483** (2019) 1907.03483

[14] Lipton, Z.C.: The Mythos of Model Interpretability (2017)

[15] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. WIREs Data Mining Knowl. Discov. **9**(4) (2019) https://doi.org/10.1002/WIDM.1312

[16] Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001) https://doi.org/10.1023/A:1010933404324

[17] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-Free Predictive Inference For Regression (2017)

[18] Jang, M., Allan, J.: Explaining controversy on social media via stance summarization. In: 41st International SIGIR Conference on R&D in Information Retrieval, Ann Arbor, MI, USA, July 08-12, 2018, pp. 1221–1224. ACM, ??? (2018)

[19] Guo, J., Lu, Y., Mori, T., Blake, C.: Expert-guided contrastive opinion summarization for controversial issues. In: Proceedings of the 24th ACM International Conference on World Wide Web. WWW '15 Companion, pp. 1105–1110 (2015)

[20] Coletto, M., Garimella, K., Gionis, A., Lucchese, C.: Automatic controversy detection in social media: A content-independent motif-based approach. Online Social Networks and Media **3-4**, 22–31 (2017)

[21] Emamgholizadeh, H., Nourizade, M., Tajbakhsh, M.S., Hashminezhad, M., Esfahani, F.N.: A framework for quantifying controversy of social network debates using attributed networks: biased random walk (BRW). Soc. Netw. Anal. Min. **10**(1), 90 (2020)

[22] Guerra, P.H.C., Jr., W.M., Cardie, C., Kleinberg, R.: A measure of polarization on social media networks based on community boundaries. In: Seventh International Conference on Weblogs and Social Media, ICWSM. The AAAI Press, ??? (2013)

[23] Mendoza, M., Parra, D., Soto, Á.: GENE: graph generation conditioned on named entities for polarity and controversy detection in social media. Inf. Process. Manag. **57**(6), 102366 (2020)

[24] Zarate, J.M.O.D., Feuerstein, E.: Vocabulary-based method for quantifying controversy in social media. In: 25th International Conference on Conceptual Structures, ICCS, vol. 12277, pp. 161–176. Springer, ??? (2020)

[25] Karypis, G., Kumar, V.: Metis – unstructured graph partitioning and sparse matrix ordering system, version 2.0 (1995)

[26] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT Conference: Human Language Technologies, Volume 1, pp. 4171–4186 (2019)

[27] Jamra, H.A., Savonnet, M., Leclercq, É.: Identification of weak signals in a temporal graph of social interactions. In: IDEAS'22: International Database Engineered Applications Symposium, Budapest, Hungary, August 22 - 24, 2022, pp. 34–42. ACM, ??? (2022)

[28] Almarzouqi, A., Aburayya, A., Salloum, S.A.: Prediction of user's intention to use metaverse system in medical education: A hybrid sem-ml learning approach. IEEE access **10**, 43421–43434 (2022)

[29] Mohapatra, A., Thota, N., Prakasam, P.: Fake news detection and classification using hybrid bilstm and self-attention model. Multimedia Tools and Applications **81**(13), 18503–18519 (2022)

[30] Swathi, T., Kasiviswanath, N., Rao, A.A.: An optimal deep learning-based lstm for stock price prediction using twitter sentiment analysis. Applied Intelligence **52**(12), 13675–13688 (2022)

[31] Akbiyik, M.E., Erkul, M., Kämpf, K., Vasiliauskaite, V., Antulov-Fantulin, N.: Ask" who", not" what": Bitcoin volatility forecasting with twitter data. In: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, pp. 688–696 (2023)

[32] Masud, R., Syamsurrijal, M., Baharuddin, T., Azizurrohman, M.: Forecasting political parties and candidates for indonesia's presidential election in 2024 using twitter (2023)

Declarations

Ethical Approval

This declaration is "not applicable".

Funding

This work was supported by grants from Janssen Horizon endowment fund.

Availability of data and materials

The dataset used has been retrieved from Zarate and al, 2020 [4].

Conflicts of interest/Competing interests

The authors have no relevant financial or non-financial interests to disclose. The authors have no conflicts of interest to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.