Explaining controversy through community analysis on Twitter

Samy Benslimane, Thomas Papastergiou, Jérôme Azé, Sandra Bringay, Caroline Mollevi, Maximilien Servajean

Explaining controversy through community analysis on Twitter

Samy Benslimane
samy.benslimane@lirmm.fr
LIRMM, Univ Montpellier, CNRS
Montpellier, France

Thomas Papastergiou
thomas.papastergiou@lirmm.fr
LIRMM, Univ Montpellier, CNRS
Montpellier, France

Jérôme Azé
jerome.aze@lirmm.fr
LIRMM, Univ Montpellier, CNRS
Montpellier, France

Sandra Bringay
sandra.bringay@lirmm.fr
AMIS, Paul-Valéry University
Montpellier, France

Maximilien Servajean
maximilien.servajean@lirmm.fr
AMIS, Paul-Valéry University
Montpellier, France

Caroline Mollevi
caroline.mollevi@chu-montpellier.fr
Institut du Cancer Montpellier (ICM)
Montpellier, France

ABSTRACT

Controversy refers to content attracting different point-of-views, as well as positive and negative feedback on a specific event, gathering users into different communities. Research on controversy led to two main categories of works: controversy detection/quantification and controversy explainability. When the former aims to quantify controversy on a topic, the latter aims to understand why a topic is controversial or not. This paper mainly contributes to the controversy explainability. We analyze topic discussions on Twitter from the community perspective to investigate the power of text in classifying tweets into the right community. We propose a SHAP-based pipeline to quantify impactful text features on predictions of three tweet classifiers. We also rely on the use of different text features namely *BERT*, $TF-IDF$, and *LIWC*. The results we obtain from both SHAP plots and statistical analysis show clearly significant impacts of some text features in classifying tweets.It also highlights the relevance of the study as well as the potential benefits of combining text and user interactions to quantify controversy.

CCS CONCEPTS

• **Computing methodologies** → *Natural language processing: Machine learning; Discourse*; • **Information systems** → *Web mining*; • **Networks** → *online social networks.*

KEYWORDS

Data mining, NLP, Machine learning, Explainability, Statistical testing, Controversy, Social networks

1 INTRODUCTION

Social media, such as Twitter, constitutes a real opportunity for people to express, share and discuss their opinions and ideas on any topic. Some topics can attract diverse and opposite opinions and lead, in some cases, to what is known as controversy, often triggered by impactful events on political discussions, climate change, gun laws, etc. Many definitions of controversy exist, but we eventually retain and define controversial topics as topics attracting different point-of-views and feedback over a specific event, polarizing users into two main conflicting communities (usually agreeing and disagreeing with the event) [8]. Analyzing controversy on social media is getting increasingly important to several tasks, such as highlighting opinion divergences, reducing fake news spreading, or gaps between communities, and at the same time "filter bubble" [1] impact. Automatic controversy detection constitutes a real challenge and has been widely studied. Recently, in the context of social media, proposed approaches are mainly based on structural information extracted from user interactions represented as graphs [8], given the assumption that polarized attention is aggregated into different communities built around influence users. Figure 1 highlights this division between communities on a controversial topic.

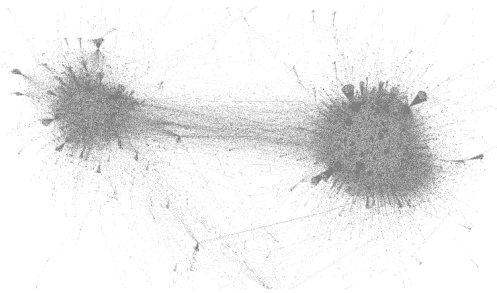

Figure 1: User Retweet graph on the controversial topic PELOSI. The graph is represented using ForceAtlas2 [10] algorithm for spatial visualization.

Exploiting textual content constitutes certainly an interesting research direction for controversy quantification. Recent works also rely on adding textual content to augment structural information and perform controversy classification tasks [2]. NLP and deep learning techniques, enriching the structural graph information, are combined to quantify controversy [5].

[1] Algorithmic personalization limiting information diversity and perception.

Studying user behaviors [12] has also become a massive challenge, especially on tasks where content is impactful. Text analysis has received attention in text classification applications including fake news detection [1], or claims detection [15]. To capture the relationship between word usage in a text, and the cognitive and mental states of the author, a text analysis tool, called Linguistic Inquiry and Word Count (LIWC) [3], has been developed by psycholinguists. To the best of our knowledge, [14] is the only work that considers text analysis for controversy detection. Discussion features (word usage, writing style) have been studied to measure the predictive power of features for controversy and language sensitivity on Reddit posts.

In this paper, we look at controversy explainability from the SHAP perspective [16] to fairly measure how much each text feature of tweets is contributing to controversy detection. So far, our work is the first attempt to use SHAP for controversial explainability needs. SHAP, considered a core contribution to explainable artificial intelligence, is used to understand how a given model makes predictions. It is a model-agnostic method, which means that it can be used to explain predictions of any existing machine-learning model. SHAP is theoretically founded, and it exploits the Shapley value concept from the cooperative game theory. The Shapley value concept is a mean to fairly divide the reward of a game among its players contributing to the game outcome. The term "fairly" is mathematically defined, meaning that the reward redistribution function satisfies four properties: efficiency (guaranteeing complete distribution of the outcome among features), symmetry (guaranteeing that two features contributing equally have the same reward), dummy (ensuring zero rewards for features that do not contribute to the outcome), and additivity (considering additive rewarding of a feature in presence of several game outcomes.

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

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.

Paper organization. Section 2 provides a literature review of various works on the controversy. Section 4 describes the dataset we used to experiment. Sections 3 and 5 present our proposed methodology for controversy analysis and experimental evaluation results respectively. The last section concludes the paper.

2 RELATED WORK

This section presents an overview of controversial detection and quantification works in the context of social media.

2.1 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 [8] or an attributed graph [7] 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 [9]. 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 [17], and user's vocabulary to cluster users with more similarities in their vocabularies [18]. Some recent works consider controversy detection as a graph classification problem [2]. 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.

2.2 Controversy explainability

Controversy interpretation aims to explain why a controversial topic is controversial. Yet, despite its obvious importance, there has been unfortunately little work on it. Some works consider controversial explainability from the document summarization technique perspective. A ranking model is used to generate the top k tweets that best summarize the stances of each community of controversial topics by only exploiting the tweets [11]. Unfortunately, these approaches limit the controversial explainability to only providing arguments from each side of a conflicting debate. Graph analysis techniques are also exploited to analyze graph user interactions and look for local patterns that characterize controversial topics [4]. User texts are unfortunately not analyzed and the approach outputs are used more as features to predict controversy than to explain the controversy. [14] is the only work that considers text features analysis for controversy explanation needs. It aims to measure the predictive strengths of individual text features for controversy on the Reddit platform. The approach results clearly show that most features reflect controversy similarly across languages. The main drawback of this approach is that it lacks foundation as it is mainly based on a set of experiments. To the best of our knowledge, our work is the first attempt to analyze the predictive power of text features for controversy based on the well-founded SHAP method.

3 METHOD

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 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.

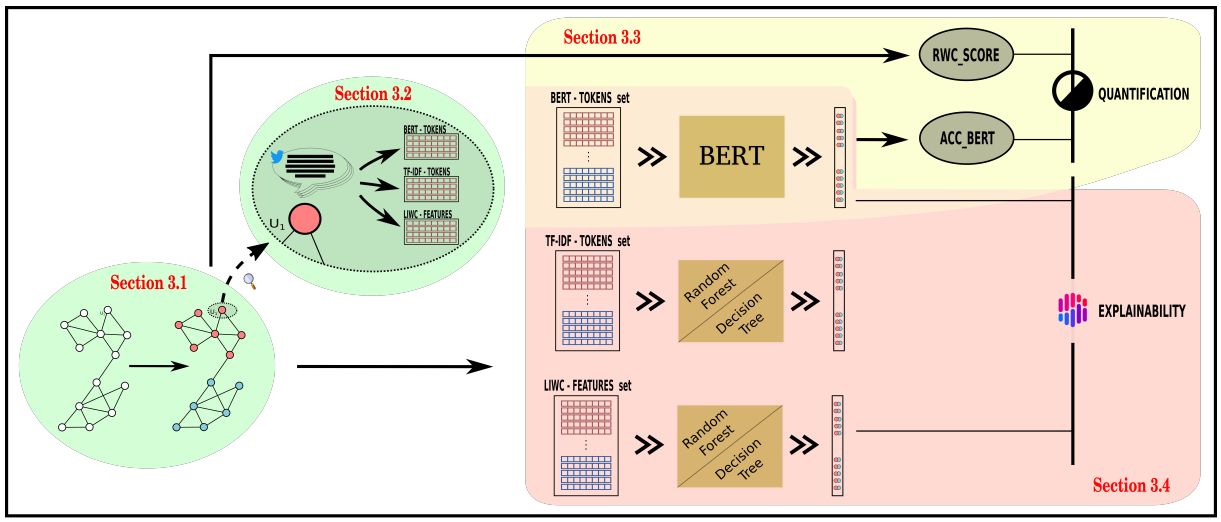

Figure 2: Pipeline used for both quantifying and explaining controversy through community analysis.

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 [8], and use the partitioning algorithm metis [13] to partition each graph into two communities. We consider that we only have the pros and cons of communities and 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.

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 [6] 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 [3] 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 [8] on the graph described in section 3.1, focusing on user partitioning. This score has been chosen because it presents the best results

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

among scores presented in [8]. 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 [6]. 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 set, 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 complementary 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 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.

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 for determining 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 [16] 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.

The Shapley value [16] sv is computed for all single features, on all tweets of the test set. 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 tweets 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.

$$SV_{p,F} = \begin{pmatrix} sv_{11} & sv_{12} & \dots & sv_{1m} \\ \dots & \dots & \dots & \dots \\ sv_{i1} & sv_{i2} & \dots & sv_{im} \\ \dots & \dots & \dots & \dots \\ sv_{n1} & sv_{n2} & \dots & sv_{nn} \end{pmatrix} \quad (1)$$

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 [5], retrieved using the Twitter API. 15 topics have been manually labeled controversial and 15 non-controversial from multiple sources on mainstream media [5]. 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

Table 2: 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

been deleted since the last time it was retrieved in [5]. 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. [18]. 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 have been chosen because they present high *rwc_score* and *acc_bert* scores. We present statistics of both datasets in 2.

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 2.

$$CPROP = \frac{min(|C_0|, |C_1|)}{max(|C_0|, |C_1|)} \quad (2)$$

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, 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.

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

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. 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.

Table 3: 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	$DT_{tfi+liwc}$	0.66
RANDOM-FOREST	TF-IDF	RF_{tfi}	0.69
	LIWC	RF_{liwc}	0.68
	TF-IDF + LIWC	$RF_{tfi+liwc}$	0.71
BERT	TEXT	$BERT_{text}$	**0.79**

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 explaining the high rwc_score. Table 3 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.

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

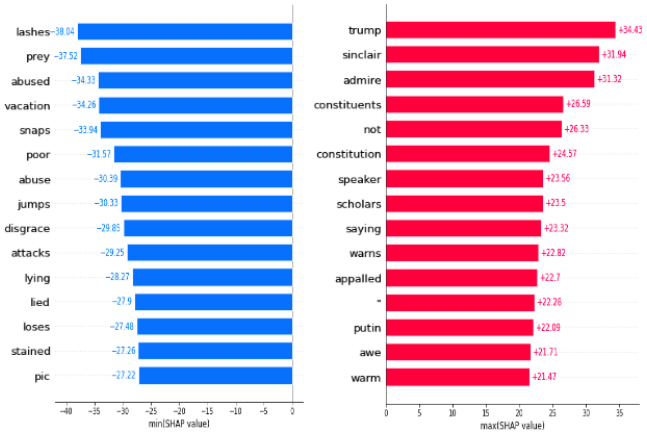

Figure 3: SHAP values are computed from tweets on $test_{pelosi}$, with BERT_{text}. Values correspond to tokens' impact on predicting one community. The figure shows the top-10 Tokens impacting prediction towards C_0 (left) and C_1 communities.

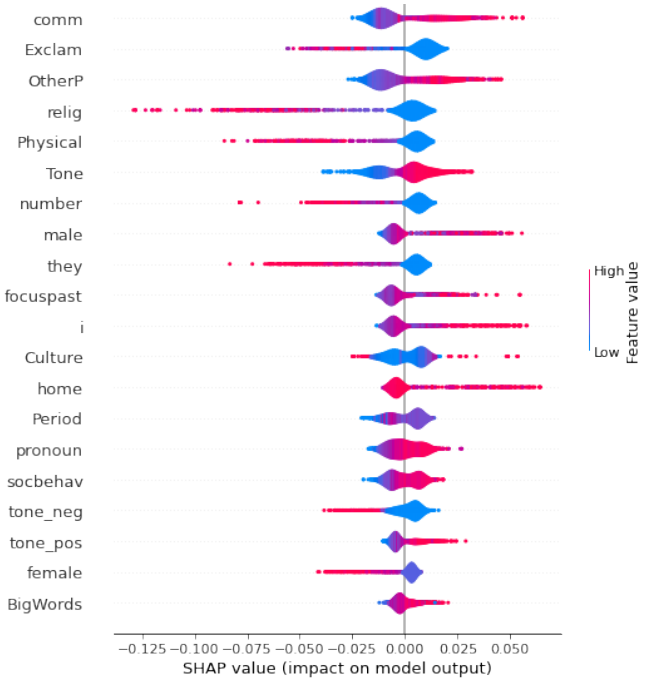

Figure 4: 20 most impacting LIWC features for model RF_{liwc} predictions on $test_{pelosi}$ set. The color scale, red (low feature value) to blue (high value) is represented for each sample. The larger the absolute SHAP value is, the more the feature pushes the model to predict the tweet to C_1 (inversely to C_0).

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. Impactful tokens of the train set are also analyzed to understand how the model learned to predict. We found that the lexical fields of tokens around the communities are very close to the analysis made previously. This shows a different way of communicating between these two communities, with distinct lexical fields.

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 4 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 Pelosi, usually use strong feelings or emphasis on their tweets. Figure 4 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 4), showing interesting behavior of our SHAP-based approach.

To conclude, regarding the proportion of users and tweets by communities in PELOSI, table 2 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.

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

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 5 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 $\text{RF}_{tfi+liwc}$.

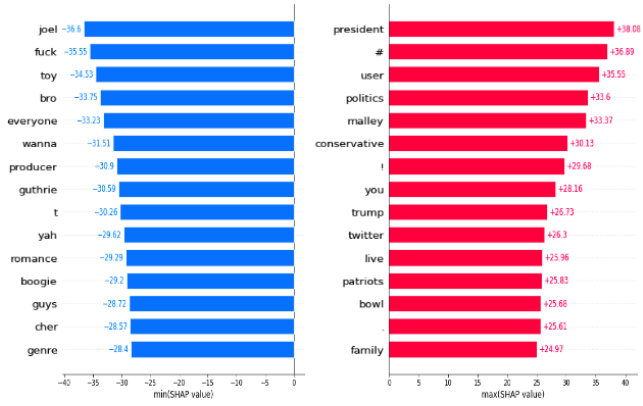

Figure 5: The top-10 tokens contribution of $test_{thanksgiving}$ set on BERT_{text}, based on SHAP values.

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 complementary with user interactions on controversial topics. 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 [8]. This work has the potential to aid future research on enhancing basic quantification measures for controversies, by integrating significant textual data with structural information, as well as looking at users at the border of communities. Analysis of a specific topic getting widely controversial by understanding communities can help with many research tasks, such as studying and quantifying controversial topics over time. An interesting perspective could be to generalize this approach to different social media. Including sub-communities in the analysis also remains a challenging task.

REFERENCES

[1] B. Ahmed, G. Baloch, Arif Hussain, Abdul Baseer Buriro, and Junaid Ahmed. 2021. Analysis of Text Feature Extractors using Deep Learning on Fake News. *Engineering, Technology and Applied Science Research* 11 (04 2021), 7001–7005.
[2] Samy Benslimane, Jérôme Azé, Sandra Bringay, Maximilien Servajean, and Caroline Mollevi. 2021. Controversy Detection: A Text and Graph Neural Network Based Approach. In *22nd Conference on Web Information Systems Engineering*, Vol. 13080. 339–354.
[3] Ryan Boyd, Ashwini Ashokkumar, Sarah Seraj, and James Pennebaker. 2022. The Development and Psychometric Properties of LIWC-22. (02 2022).
[4] Mauro Coletto, Kiran Garimella, Aristides Gionis, and Claudio Lucchese. 2017. Automatic controversy detection in social media: A content-independent motif-based approach. *Online Social Networks and Media* 3-4 (2017), 22–31.
[5] Juan Manuel Ortiz de Zarate, Marco Di Giovanni, Esteban Zindel Feuerstein, and Marco Brambilla. 2020. 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. 194–209.
[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT Conference: Human Language Technologies, Volume 1*. 4171–4186.
[7] Hanif Emamgholizadeh, Milad Nourizade, Mir Saman Tajbakhsh, Mahdieh Hashminezhad, and Farzaneh Nasr Esfahani. 2020. A framework for quantifying controversy of social network debates using attributed networks: biased random walk (BRW). *Soc. Netw. Anal. Min.* 10, 1 (2020), 90.
[8] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying Controversy on Social Media. *ACM Trans. Soc. Comput.* 1, 1 (2018), 3:1–3:27.
[9] Pedro Henrique Calais Guerra, Wagner Meira Jr., Claire Cardie, and Robert Kleinberg. 2013. 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.
[10] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PloS one journal* 9 (06 2014).
[11] Myungha Jang and James Allan. 2018. 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*. ACM, 1221–1224.
[12] Myungha Jang, Shiri Dori-Hacohen, and James Allan. 2017. Modeling Controversy within Populations. In *Proceedings of the SIGIR International Conference on Theory of Information Retrieval, ICTIR*. ACM, 141–149.
[13] George Karypis and Vipin Kumar. 1995. METIS – Unstructured Graph Partitioning and Sparse Matrix Ordering System, Version 2.0. (01 1995).
[14] Philipp Koncar, Simon Walk, and Denis Helic. 2021. Analysis and Prediction of Multilingual Controversy on Reddit. In *Web Science Conference 2021*. 215–224.
[15] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. [n. d.]. Context Dependent Claim Detection. In *25th International Conference on Computational Linguistics: Technical Papers* (2014-08). 1489–1500.
[16] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874 [cs.AI]
[17] Marcelo Mendoza, Denis Parra, and Álvaro Soto. 2020. GENE: Graph generation conditioned on named entities for polarity and controversy detection in social media. *Inf. Process. Manag.* 57, 6 (2020), 102366.
[18] Juan Manuel Ortiz De Zarate and Esteban Feuerstein. 2020. Vocabulary-Based Method for Quantifying Controversy in Social Media. In *25th International Conference on Conceptual Structures, ICCS*, Vol. 12277. Springer, 161–176.