Measuring Controversy in Social Networks Through NLP

Juan Manuel Ortiz de Zarate[1(✉)], Marco Di Giovanni[2], Esteban Zindel Feuerstein[1], and Marco Brambilla[2]

[1] Universidad de Buenos Aires, C1053 Buenos Aires, Argentina
{jmoz,efeuerst}@dc.uba.ar
[2] Politecnico di Milano, Milan 20133, Italy
{marco.digiovanni,marco.brambilla}@polimi.it

Abstract. Nowadays controversial topics on social media are often linked to hate speeches, fake news propagation, and biased or misinformation spreading. Detecting controversy in online discussions is a challenging task, but essential to stop these unhealthy behaviours.

In this work, we develop a general pipeline to quantify controversy on social media through content analysis, and we widely test it on Twitter.

Our approach can be outlined in four phases: an initial *graph building* phase, a *community identification* phase through graph partitioning, an *embedding* phase, using language models, and a final *controversy score computation* phase. We obtain an index that quantifies the intuitive notion of controversy.

To test that our method is general and not domain-, language-, geography- or size-dependent, we collect, clean and analyze 30 Twitter datasets about different topics, half controversial and half not, changing domains and magnitudes, in six different languages from all over the world.

The results confirm that our pipeline can quantify correctly the notion of controversy, reaching a ROC AUC score of 0.996 over controversial and non-controversial scores distributions. It outperforms the state-of-the-art approaches, both in terms of accuracy and computational speed.

Keywords: Controversy · Polarization · NLP · Social networks

1 Introduction

Controversy in social networks is a phenomenon with a high social and political impact. Interesting analysis have been performed about presidential elections [44], congress decisions [21], hate spread [9], and harassing [31]. This phenomenon has been broadly studied from the perspective of different disciplines, ranging from the seminal analysis of conflicts within the members of a karate club [54] to political issues in modern times [1,3,10,13,36].

J. M. O. de Zarate and M. Di Giovanni—Equal contribution.

C. Boucher and S. V. Thankachan (Eds.): SPIRE 2020, LNCS 12303, pp. 194–209, 2020.
https://doi.org/10.1007/978-3-030-59212-7_14

The irruption of digital social networks [17] gave raise to new ways of intentional intervention for taking advantages [9,44]. Moreover, highly contrasting points of view in groups tend to provoke conflicts that lead to attacks from one community to the other, such as harassing, "brigading", or "trolling" [31]. The existing literature reports a huge number of issues related to controversy, ranging from the splitting of communities and the biased information spread, to the increase of hate speeches and attacks between groups. For example, Kumar, Srijan, et al. [31] analyze many defense techniques from attacks on *Reddit*[1] while Stewart, et al. [44] insinuate that there was external interference in *Twitter* during the 2016 US presidential elections to benefit one candidate.

As shown in [30,33], detecting controversy also provides the basis to improve the *"news diet"* of readers, offering the possibility to connect users with different points of view by recommending them personalized content to read [37]. Other studies on "bridging echo chambers" [19] and the positive effects of inter group dialogue [4,38] suggest that direct engagement is effective for mitigating conflicts.

An accurate and automatic classifier of controversial topics, therefore, helps to develop quick strategies to prevent miss-information, fights and biases. Moreover, the identification of the main viewpoints and the detection of semantically closer users is also useful to lead people to healthier discussions. *Measuring* controversy is even more powerful, as it can be used to establish controversy levels. For this purpose, we propose a content-based pipeline to measure controversy on social networks, collecting posts' content about a fixed topic (an hashtag or a keyword) as root input.

Controversy quantification through vocabulary analysis also opens several research avenues, such as the analysis whether polarization is being created, maintained or augmented through community's way of talking.

Our main contribution can be summarized as the design of a controversy detection pipeline and the its application to 30 heterogeneous Twitter datasets. We outperform the state-of-the-art approaches, both in terms of accuracy and computational speed.

Our method is tested on datasets from Twitter. This microblogging platform has been widely used to analyze discussions and polarization [36,39,46,50,53]. It is a natural choice for this task, as it represents one of the main fora for public debate [50], it is a common destination for affiliative expressions [23] and it is often used to report and read news about current events [43]. An extra advantage is the availability of real-time data generated by millions of users. Other social media platforms offer similar data-sharing services, but few can match the amount of data and the documentation provided by Twitter. One last asset of Twitter for our work is given by *retweets* (sharing a tweet created by a different user), that typically indicate endorsement [6] and hence they help to model discussions as they can signal "who is with who".

Our paper is organized as follows: in Sect. 2 we list and summarize other works about controversy and polarization on social networks, in Sect. 3 we present the datasets collected for this study, while Sect. 4 contains the

[1] https://www.reddit.com/.

step-by-step description of our pipeline. In Sect. 5 we show the results and we conclude with Sect. 6.

2 Related Work

Due to its high social importance, many works focus on polarization measures in online social networks and social media [2,10,11,20,22]. The main characteristic that connects these works is that the measures proposed are based on the structural characteristics of the underlying social-graph. Among them, we highlight the work of Garimella et al. [20] that presents an extensive comparison of controversy measures, different graph-building approaches and data sources, achieving a state-of-the-art performance. We use this approach as a baseline to compare our results.

In [20] the authors propose many metrics to measure polarization on Twitter. Their techniques, based on the structure of the endorsement graph, can successfully detect whether a discussion (represented by a set of tweets), is controversial or not, regardless of the context and, most importantly, without the need of domain expertise. They also include two methods to measure controversy based on the analysis of the posts' contents, both failing. The first of these methods starts with the embedding of tweets in vectors, the clustering of these vectors into two groups and a final computation of KL divergence[2] as a distance measure between clusters, and of I2 measure [27] to quantify the cluster heterogeneity. The second method is based on sentiment analysis. Their hypothesis is that controversial discussions have a higher variance than non-controversial ones. This approach is limited to the fact that it is dependent on language-specific tools that do not work reliably for languages other than English.

Matakos et al. [35] also develop a *polarization index* with a graph-based approach, not including text related features, modelling opinions as real numbers. Their measure successfully captures the tendency of opinions to concentrate in network communities, creating echo-chambers.

Other recent works [34,41,45] prove that communities may express themselves with different terms or ways of speaking, and use different jargon, which can be detected with the use of text-related techniques. Ramponi et al. [40,41] build very efficient classifiers and predictors of account membership within a given community by inspecting the vocabulary used in tweets, for many heterogeneous Twitter communities, such as chess players, fashion designers and members and supporters of political parties [15]. In [45] Tran et al. found that language style, characterized using a hybrid word and part-of-speech tag n-gram language model, is a better indicator of community identity than topic, even for communities organized around specific topics. Finally, Lahoti et al. [34] model the problem of learning the liberal-conservative ideology space of social media users and media sources as a constrained non-negative matrix-factorization problem. They validate their model and solution on a real-world Twitter dataset

[2] Kullback–Leibler divergence is a measure of how a probability distribution is different from a reference probability distribution.

consisting of controversial topics, and show that they are able to separate users by ideology with over 90% purity.

Other works for controversy detection through content have been made over Wikipedia [16,26] showing that text contents are good indicatives to estimate polarization. These works are heavily dependent on Wikipedia and can not be extrapolated to social networks.

In her thesis [25], Jang explains controversy via generating a summary of two conflicting stances that build the controversy. Her work shows that a specific sub-set of tweets is enough to represent the two opposite positions in a polarized debate.

A first approach to content-based controversy detection was made in [55]. The main difference between this work and [55] is that the techniques presented here are less dependent on the graph structure. Our new content-based pipeline introduces the possibility of defining and detecting concepts like the "semantic frontier" of a cluster. This opens new ways to activate interventions in the communities, such as the investigation of users lying near that frontier to facilitate a healtier interaction between the communities, or the analysis of users far away from the frontier to understand which aspects establish the real differences. Improvements on [55] (used as a second baseline in this work), include a wider comparison of NLP models and distance measures, a higher heterogeneity of datasets used, and results in better performances both in terms of AUC ROC scores and computational times.

3 Datasets

To test our approach, we collect 30 Twitter datasets in six languages. Each dataset corresponds to a manually selected topic among the trending ones. The collection is performed through the official Twitter API.

3.1 Topic Definition

In the literature, a topic is often defined by a single hashtag. We believe that this might be too restrictive since some discussions may not have a defined hashtag, but they are about a *keyword* that represents the main concept, i.e. a word or expression that is not specifically an hashtag but it is widely used in the discussion. For example during the Brazilian presidential elections in 2018, we collected tweets mentioning to the word *Bolsonaro*, the principal candidate's surname. Thus, in our approach, a topic is defined as a specific hashtag or keyword, depending on the discussion. For each topic we collect all the tweets that contain its hashtag or keyword, posted during a selected observation window. We also check that each topic is associated with a large enough activity volume.

3.2 Description of the Datasets

We collected 30 discussions (50% more than the baseline work [20]) that took place between 2015 and 2020, half of them controversial and half not. We selected

discussions in six languages: English, Portuguese, Spanish, French, Korean and Arabic, occurring in five regions over the world: South and North America, Western Europe, Central and Southern Asia. The details of each discussion are described in Table 2. We have chosen discussions clearly recognizable as controversial or not to have an evident groundtruth. Blurry discussions will be analyzed in future works. The encoded datasets are available on github[3].

Since our models require a large amount of text and since a tweet contains no more than 240 characters, we established a threshold of at least 100000 tweets per topic. Topics containing a lower number of tweets were discarded. To select discussions and to determine if they are controversial or not we looked for topics widely covered by mainstream media that have generated ample discussion, both online and offline. For non-controversial discussions we focused both on "soft news" and entertainment, and on events that, while being impactful and/or dramatic, did not generate large controversies. On the other side, for controversial debates we focused on political events such as elections, corruption cases or justice decisions. We validate our intuition by manually checking random samples of tweets.

To furtherly establish the presence or absence of controversy of our datasets, we visualized the corresponding networks through ForceAtlas2 [24], a widely used force-directed layout. This algorithm has been recently found to be very useful at visualizing community interactions [49], as it represents closer users interacting among each other, and farther users interacting less. Figure 1 shows examples of how non-controversial and controversial discussions respectively look like with ForceAtlas2 layout. As we can see in these figures, in a controversial discussion the layout shows two well separated groups, while in a non-controversial one it generates one big cluster.

More information on the datasets is given in Table 2 in Appendix A.

4 Methodology

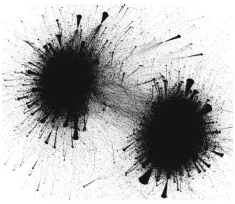
Our approach to measure controversy can be outlined into four phases, namely *graph building* phase, *community identification* phase, *embedding* phase and *controversy score computation* phase. The final output of the pipeline is a positive value that measures the controversy of a topic, with higher values corresponding to lower degrees of controversy.

Our hypothesis is that using the embeddings generated by an NLP model, we can distinguish different ways of speaking; the more controversial the discussion is, the better differentiation we obtain.

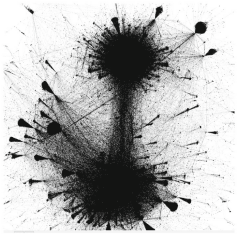
4.1 Graph Building Phase

Firstly, our purpose is to build a conversation graph that represents activities related to a single topic of discussion. For each topic, we build a retweet-graph

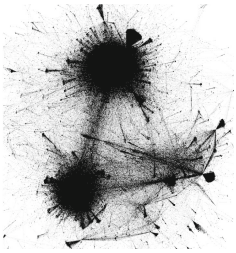
[3] Code and datasets used in this work are available here: https://github.com/jmanuoz/Measuring-controversy-in-Social-Networks-through-NLP.



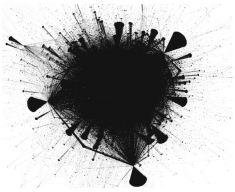
(a) Kavanaugh nomination



(b) Brazilian presidential election

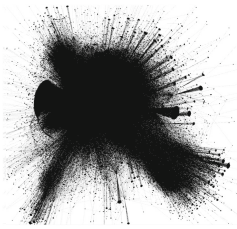


(c) Mentions to Argentinian ex-president



(d) Halsey concert

(e) Pop star birthday



(f) New album of EXO band

Fig. 1. ForceAtlas2 layout for different discussions. (a), (b) and (c) are controversial while (d), (e) and (f) are non-controversial.

G where each user is represented by a vertex, and a directed edge from node u to node v indicates that user u retweeted a tweet posted by user v.

Retweets typically indicate endorsement [6]: users who retweet signal endorsement of the opinion expressed in the original tweet by propagating it further. Retweets are not constrained to occur only between users who are connected in Twitter's social network, but users are allowed to retweet posts generated by every other user. As typically in the literature [7,9,18,32,36,44] we establish that one retweet among a pair of users is enough to define an edge between them. We do not use "quotes" to build the graph since, due to their nature, they can both signal endorsement and opposition, allowing users to comment the quoted tweet.

We remark that the "retweet information" is included in the tweets extracted, allowing us to build the graph without increasing the number of twitter API requests needed. This makes this stage faster than, for example, building a follower graph, another popular alternative.

4.2 Community Identification Phase

To identify the jargon of the community we need to be very accurate at defining its members. If we, in our will of finding two principal communities, force the

partition of the graph in that precise number of communities, we may be adding noise in the jargon of the principal communities that are fighting each other. Thus, we decide to cluster the graph using Louvain [8], one of the most popular graph-clustering algorithms. It is a greedy technique that can run over big networks without memory or running time problems, and does not detect a fixed number of clusters. Its output depends on the Modularity Q optimization, resulting in less "noisy" communities. In a polarized context there are two principal sides covering the whole discussion, thus we take the two biggest communities identified by Louvain and use them for the following steps. Since to have controversy in a discussion there must be "at least" two sides, if the principal sides are more than two, discarding the smallest ones will not impact the final result. In future work we will investigate these more complex situations. Up to here the approach we follow is the same as in [55].

4.3 Embedding Phase

In this phase, our purpose is to embed each user into a corresponding vector. These vectors encode syntactic and semantic proprieties of the posts of the corresponding accounts. They will be used in the next phase to compute the controversy score, since we need fixed dimension semantically significant vectors to perform the following computations.

Firstly, tweets belonging to the users of the two principal communities selected in the previous stage are grouped by user and sanitized. We remove duplicates and, from each tweet, we remove user names, links, punctuation, tabs, leading and lagging blanks, general spaces and the retweet keyword "RT", the string that points that a tweet is in fact a retweet. Many sentence embedding techniques have been developed in the literature, ranging from simple bag-of-words models to complex deep language models. To perform this step we selected two models among the most advanced ones, namely Fasttext and BERT, that embed texts into fixed dimension vectors encoding semantically significance and meaning.

Fasttext [28]**.** This is a tool based on the skipgram model, where each word is represented as a bag of character n-grams. A vector representation is associated to each character n-gram; words being represented as the sum of these representations. This is a fast method that allows to quickly train models on large corpora and to compute word representations also for words that do not appear in the training data. We train this model with tagged data, accordingly to the output of Louvain (previous stage), representing the community of the user. To define the values of the hyper-parameters we use the findings of [52], where the authors investigate the best hyper-parameters to train word embedding models using Fasttext and Twitter data. We use the trained model to compute the text embedding.

BERT. Bidirectional Encoder Representations from Transformers (BERT) [14] is a deep state-of-the-art language representation model based on Transformers [48] pretrained in an unsupervised way on the entire Wikipedia dump for more than 100 languages. The model is designed for transfer learning, so it has to be finetuned for a few epochs for a specific tasks, inserting an additional fully-connected layer on the top, without any substantial task-specific architecture modifications. We use the BASE version of BERT (12 layer, 768 hidden dimension, 12 heads per layer, for a total of $110M$ parameters).

Given a dataset of tweets labeled accordingly to the output of Louvain (previous stage), we finetune BERT on a 2-classes classification task for 6 epochs (learning rate set to 10^{-5}). Since our goal is to obtain embeddings of tweets, after the training procedure we remove the fully-connected layer and we use the outputs of BERT as embeddings. In detail, BERT firstly split a sentence into tokens, adding the *[CLS]* token at the beginning. Then, it embeds each token into a 786-dimensional vector. Since we need a single vector of fixed length to compute our score, we select as aggregator the embedding of the *[CLS]* token. This is the same strategy selected during the fine-tuning step. We perform this stage using bert-as-service GitHub repository [51].

To train Fasttext and BERT in a supervised way, we need to create a training set with its labels. We label each user with its community, namely with tags C_1 and C_2, corresponding respectively to the biggest (Community 1) and second biggest (Community 2) groups. It is important to note that, to prevent bias in the model, we take the same number of users from each community, downsampling the first principal community to the number of users of the second one.

4.4 Controversy Score Computation Phase

To compute the controversy score, we select some users as the best representatives of each side's main point of view. We run the HITS algorithm [29] to estimate the authoritative and hub score of each user. We take the 30% of the users with the highest authoritative score and the 30% with the highest hub score and we call them *central users*.

Finally, we compute the controversy score r, using the embeddings of the central users $x_i \in \mathbb{R}^k$ and the labels $y_i \in \{1, 2\}$, imposing their belonging to cluster C_1 or C_2, computed during the community identification phase.

We compute the centroids of each cluster j with Eq. 1, where $|C_j|$ is the magnitude of cluster C_j, and a global centroid c_{glob} with Eq. 2.

$$c_j = \frac{1}{|C_j|} \sum_{i:y_i=j} x_i \tag{1}$$

$$c_{glob} = \frac{1}{|C_1| + |C_2|} \sum_{i} x_i \tag{2}$$

We define D_j as the sum of distances between the embeddings x_i and their centroids c_j using Eq. 3 for $j = 1, 2$, where *dist* is a generic distance function.

Similarly, D_{glob} is the sum of distances between all the embeddings and the global centroid.

$$D_j = \sum_{i:y_i=j} dist(x_i, c_j) \tag{3}$$

Because of the *curse of dimensionality* [5], measuring distances over big number of dimensions is not a trivial task and the usefulness of a distance measure depends on the sub-spaces that the problem belongs to [42]. For this reason, we select and test four distance measure: L_1 (Manhattan), L_2 (Euclidean), Cosine and Mahalanobis [12] distance (particularly useful when the embedding space is not interpretable and not homogeneous, since it takes into account also correlations of the dataset and reduces to Euclidean distance if the covariance matrix is the identity matrix).

The controversy score r is defined in Eq. 4.

$$r = \frac{D_1 + D_2}{D_{glob}} \tag{4}$$

Intuitively, it represents how much the clusters are separated. We expect that, if the dataset is a single cloud of points, this value should be near 1 since the two centroids c_1 and c_2 will be near each other and near the global centroid c_{glob}. On the contrary, if the embeddings successfully divide the dataset in two clearly separated clusters, their centroids will be far apart and near to the points that belong to their own clusters. Note that r is, by definition, positive, since D_1, D_2 and D_{glob} are positive too.

The datasets and the full code is available on github[4] and the results discussed in the following section are fully reproducible.

5 Results

In this section we collect the results obtained with the different techniques described above and we compare them to the state-of-the-art structured-based method "RW" [20] and our previous work "DMC" [55], a structure and text-based approach. In Fig. 2 we show the distributions of scores of Fasttext and BERT, using the four different distances described before, compared to the baselines "RW" and "DMC". We plot them as beanplots with scores of controversial datasets on the left side and non-controversial ones on the right side. Note that, since by definition "DMC" approach gives higher scores for controversial datasets and lower scores for non-controversial ones, the two distributions are reversed.

The less the two distributions overlap, the better the pipeline works. Thus, to quantify the performance of different approaches, we compute the ROC AUC. By definition, this value is between 0 and 1, where 0.5 means that the curves are perfectly overlapped (i.e. random scoring), while values of 0 and 1 correspond to

[4] https://github.com/jmanuoz/Measuring-controversy-in-Social-Networks-through-NLP.

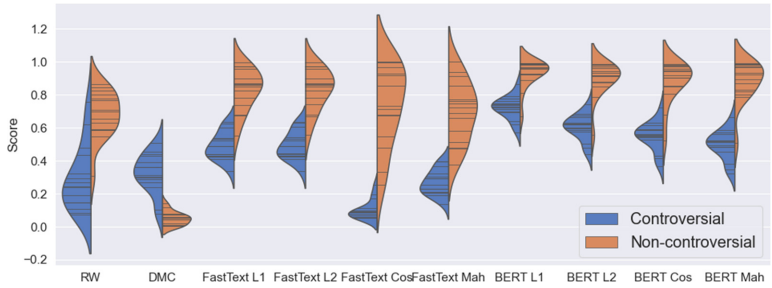


Fig. 2. Scores distributions comparison

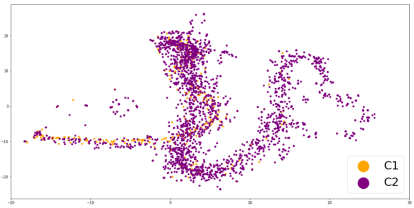
perfectly separated distributions. The comparison among the different distance measures is reported in Table 1. As we can see, the best score (the highest value) is obtained by Fasttext model with cosine distance, outperforming the state-of-the-art methods [20,55].

Table 1. ROC AUC scores comparison

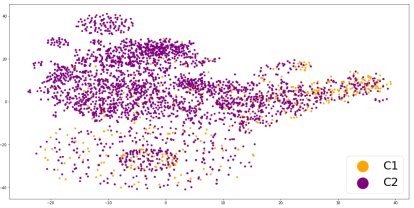
Method	L1	L2	Cosine	Mahalanobis	Baseline
FastText	0.987	0.987	**0.996**	0.991	–
BERT	0.942	0.947	0.942	**0.964**	–
DMC	–	–	–	–	**0.982**
RW	–	–	–	-	**0.924**

Even if BERT reached many state-of-the-art results in different NLP tasks [14], FastText suits better in our pipeline. Analyzing the wrongly scored cases we observe that BERT fails mainly with the non-controversial datasets, for example *Feliz Natal* dataset (0.51 controversy score). Our hypothesis is that, since BERT is a bigger and more complex model than FastText, sometimes it overfits the data. BERT is able to separate the two communities' ways of speaking even when they are very similar, not opposite sides of a controversy, exploiting differences that we are not able to perceive. To qualitatively check this behaviour we plot the embeddings produced by each technique by reducing their dimension to 2 with t-SNE algorithm [47] for visualization purposes.

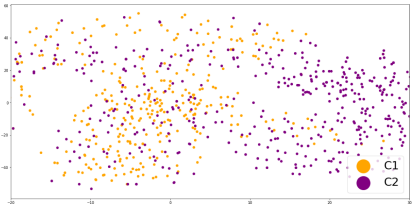
In Fig. 3 we show the reduced embeddings obtained by each method for two non-controversial datasets *Jackson's birthday* and *Feliz Natal*. The first dataset is correctly predicted as non-controversial by both methods and we can see that their embeddings are highly mixed, as expected. However *Feliz Natal* embbedings are mixed when Fasttext is used, while BERT is still able to split them in two separate clusters. This shows that, for the *Feliz Natal* case, BERT is still differentiating two ways of speaking.



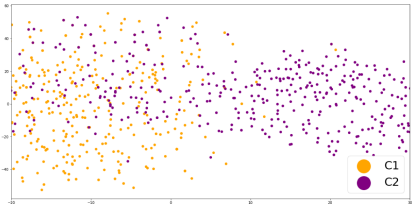
(a) Fasttext embeddings of *Kingjacksonday* dataset



(b) BERT embeddings of *Kingjacksonday* dataset



(c) Fasttext embeddings of *Feliz Natal* dataset



(d) BERT embeddings of *Feliz Natal* dataset

Fig. 3. t-SNE reduced embeddings produced by Fasttex and Bert

Computational Time. Figure 4 shows the boxplots over the 30 datasets of the total computational times (in seconds) of our two best algorithms, from the beginning (graph building stage) to the end (controversy score computation stage), compared to the baselines. Our approaches are faster than the baseline graph-based method (RW), while DMC approach is only faster than our BERT variant. Fastext approach outperforms both the baselines, allowing a quicker analysis when used in a real-time perspective, since intervention could be necessary for prevention of malicious behaviours, already described in Sect. 1.

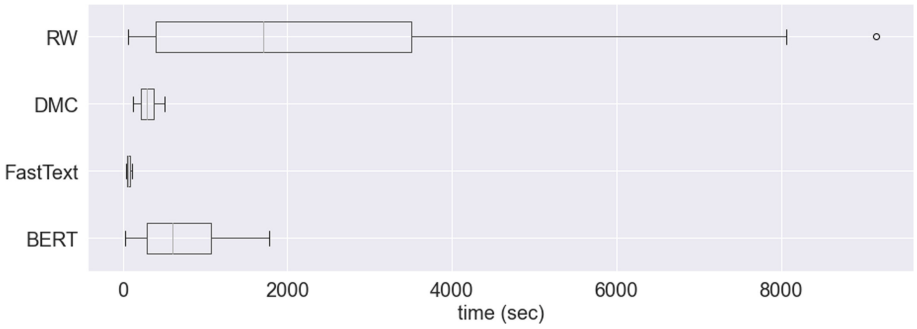


Fig. 4. Computational time comparison

6 Conclusions

In this work we designed an NLP-based pipeline to measure controversy. We test some variants, such as two embedding techniques (using Fasttext and BERT language models) and four distance measures. We applied these approaches on 30 heterogeneous Twitter datasets, and we compared the results. Our best approach, using FastText and cosine distance, outperforms not only the state-of-the-art graph-based method [20], where the authors state that content based techniques do not perform as well as structure based ones, but also our previous work [55], in terms of ROC AUC score and speed, due to the lower dependence on the graph structure and the insertion of a semantic contribute.

Our pipeline involves FastText, a fast model to encode sentences, or BERT, a more accurate language model, slower due to the complex finetuning process required. Fasttext obtains the best performance overall, reaching a ROC AUC score of 0.996. As we reported in the previous section, this is probably because BERT is so strong that it could differentiate ways of speaking even when they are not in controversy. Due to the nature of our pipeline, Fasttext performs better having also a much faster computing time. These results open to a whole new social network analysis to help people participate in healthier discussions, since these approaches allow us to detect faster and better the different points of view.

Since this approach on controversy detection shares similarities with previous works [20,55], we share some limitations too: *Evaluation*, difficulties to establish the ground-truth, *Multisided controversies*, controversy with more than two sides, *Choice of data*, manual collection of topics, and *Overfitting*, small set of experiments, although now we have 10 more discussions, it is still not big enough from a statistical point of view.

Our language-based approach has other limitations. Firstly, training an NLP model that can have a good performance requires significant amount of text, therefore our method works only for "big" enough discussions. However, most interesting controversies are those that have consequences at a society level, in general big enough for our method. Secondly, our findings are based on datasets coming from Twitter. While this is certainly a limitation, Twitter is one of the main venues for online public discussion, and one of the few for which data is easily available. Hence, Twitter is a natural choice. However, Twitter's characteristic limit of 280 characters per message (140 till short time ago) is an intrinsic limitation. We believe that our method, applied to other social networks like Facebook or Reddit, could perform even better, since having more text per user could redound on a more accurate computation of the controversy score.

Future work will involve also user-related analysis, such as the detection of users that are in the "semantic border", on controversial cases, and how they behave over time. This could be useful to find whether there are actors that may help to prevent polarization. We will also analyze which users lay on opposite semantic sides to quickly detect the main differences between two communities.

Finally, we will also detect and analyze the behaviours of users performing mixed interventions on a polarized debate, e.g. posting opinions of both sides of the controversy.

Appendix A Details on the discussions

Table 2. Datasets statistics, the top group represent controversial topics, while the bottom one represent non-controversial ones

Hashtag/Keywords	#Lang	#Tweets	Description and collection period
#LeadersDebate	EN	250 000	Candidates debate, Nov 11–21,2019
pelosi	EN	252 000	Trump Impeachment, Dec 06,2019
@mauriciomacri	ES	108 375	Macri's mentions, Jan 1–11,2018
@mauriciomacri	ES	120 000	Macri's mentions, Mar 11–18,2018
@mauriciomacri	ES	147 709	Macri's mentions, Mar 20–27,2018
@mauriciomacri	ES	309 603	Macri's mentions, Apr 05–11,2018
@mauriciomacri	ES	254 835	Macri's mentions, May 05–11,2018
Kavanaugh	EN	260 000	Kavanaugh's nomination, Oct 03,2018
Kavanaugh	EN	259 999	Kavanaugh's nomination, Oct 05,2018
Kavanaugh	EN	260 000	Kavanaugh's nomination, Oct 08,2018
Bolsonaro	PT	170 764	Brazilian elections, Oct 27,2018
Bolsonaro	PT	260 000	Brazilian elections, Oct 28,2018
Bolsonaro	PT	260 000	Brazilian elections, 30-10-2018
Lula	PT	250 000	Mentions to Lula the day of Moro chats news, Jun 11-10,2019
Dilma	PT	209 758	Roussef impeachment, 06-11-2015
EXODEUX	EN	179 908	EXO's new album, Nov 07,2019
Thanksgiving	EN	250 000	Thanksgiving day, Nov 28,2019
#Al-HilalEntertainment	AR	221 925	Al-Hilal champion, Dec 01,2019
#MiracleOfChristmasEve	KO	251 974	Segun Woo singer birthday, 23-12-2019
Feliz Natal	PT	305 879	Happy Christmas wishes, Dec 24,2019
#kingjacksonday	EN	186 263	popstar's birthday, Mar 24–27,2019
#Wrestlemania	EN	260 000	Wrestlemania event, Apr 08,2019
Notredam	FR	200 000	Notredam fire, Apr 16,2019
Nintendo	EN	203 992	Nintendo's release, May 19–28,2019
Halsey	EN	250 000	Halsey's concert, Jun 07–08,2019
Bigil	EN	250 000	Vijay's birthday, Jun 21–22,2019
#VanduMuruganAJITH	EN	250 000	Ajith's fans, Jun 23,2019
Messi	ES	200 000	Messi's birthday, Jun 24,2019
#Area51	EN	178 220	Jokes about Area51, Jul 13,2019
#OTDirecto20E	ES	148 061	Event of a Music TV program in Spain, Jan 20,2020

References

1. Adamic, L.A., Glance, N.: The political blogosphere and the 2004 US election: divided they blog. In: Proceedings of the 3rd International Workshop on Link Discovery, pp. 36–43. ACM (2005)
2. Akoglu, L.: Quantifying political polarity based on bipartite opinion networks. In: Eighth International AAAI Conference on Weblogs and Social Media (2014)
3. Al-Ayyoub, M., Rabab'ah, A., Jararweh, Y., Al-Kabi, M.N., Gupta, B.B.: Studying the controversy in online crowds' interactions. Appl. Soft Comput. **66**, 557–563 (2018)
4. Allport, G.W., Clark, K., Pettigrew, T.: The Nature of Prejudice. Addison-Wesley, Reading (1954)
5. Bellman, R.: Dynamic programming. Science **153**(3731), 34–37 (1966)
6. Bessi, A., Caldarelli, G., Del Vicario, M., Scala, A., Quattrociocchi, W.: Social determinants of content selection in the age of (mis)information. In: Aiello, L.M., McFarland, D. (eds.) SocInfo 2014. LNCS, vol. 8851, pp. 259–268. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13734-6_18
7. Bild, D.R., Liu, Y., Dick, R.P., Mao, Z.M., Wallach, D.S.: Aggregate characterization of user behavior in Twitter and analysis of the retweet graph. ACM Trans. Internet Technol. (TOIT) **15**(1), 1–24 (2015)
8. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. **2008**(10), P10008 (2008)
9. Calvo, E.: Anatomía política de Twitter en argentina. Tuiteando# Nisman. Capital Intelectual, Buenos Aires (2015)
10. Conover, M.D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., Flammini, A.: Political polarization on Twitter. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
11. Dandekar, P., Goel, A., Lee, D.T.: Biased assimilation, homophily, and the dynamics of polarization. Proc. Natl. Acad. Sci. **110**(15), 5791–5796 (2013)
12. De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L.: The mahalanobis distance. Chemometr. Intell. Lab. Syst. **50**(1), 1–18 (2000)
13. Del Vicario, M., Zollo, F., Caldarelli, G., Scala, A., Quattrociocchi, W.: Mapping social dynamics on Facebook: the Brexit debate. Soc. Netw. **50**, 6–16 (2017)
14. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805
15. Di Giovanni, M., Brambilla, M., Ceri, S., Daniel, F., Ramponi, G.: Content-based classification of political inclinations of Twitter users. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 4321–4327 (2018)
16. Dori-Hacohen, S., Allan, J.: Automated controversy detection on the web. In: Hanbury, A., Kazai, G., Rauber, A., Fuhr, N. (eds.) ECIR 2015. LNCS, vol. 9022, pp. 423–434. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16354-3_46
17. Easley, D., Kleinberg, J., et al.: Networks, Crowds, and Markets, vol. 8. Cambridge University Press, Cambridge (2010)
18. Feng, W., Wang, J.: Retweet or not?: personalized tweet re-ranking. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 577–586. ACM (2013)
19. Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.: Reducing controversy by connecting opposing views. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, pp. 81–90. ACM (2017)

20. Garimella, K., Morales, G.D.F., Gionis, A., Mathioudakis, M.: Quantifying controversy on social media. ACM Trans. Soc. Comput. **1**(1), 3 (2018)
21. Grčar, M., Cherepnalkoski, D., Mozetič, I., Kralj Novak, P.: Stance and influence of Twitter users regarding the Brexit referendum. Comput. Soc. Netw. **4**(1), 1–25 (2017). https://doi.org/10.1186/s40649-017-0042-6
22. Guerra, P.C., Meira Jr., W., Cardie, C., Kleinberg, R.: A measure of polarization on social media networks based on community boundaries. In: Seventh International AAAI Conference on Weblogs and Social Media (2013)
23. Hong, S.: Online news on Twitter: newspapers' social media adoption and their online readership. Inf. Econ. Policy **24**(1), 69–74 (2012)
24. 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 **9**(6), e98679 (2014)
25. Jang, M.: Probabilistic models for identifying and explaining controversy (2019)
26. Jang, M., Foley, J., Dori-Hacohen, S., Allan, J.: Probabilistic approaches to controversy detection. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 2069–2072 (2016)
27. Jeh, G., Widom, J.: SimRank: a measure of structural-context similarity. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543. ACM (2002)
28. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
29. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM (JACM) **46**(5), 604–632 (1999)
30. Kulshrestha, J., Zafar, M.B., Noboa, L.E., Gummadi, K.P., Ghosh, S.: Characterizing information diets of social media users. In: Ninth International AAAI Conference on Web and Social Media (2015)
31. Kumar, S., Hamilton, W.L., Leskovec, J., Jurafsky, D.: Community interaction and conflict on the web. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web, pp. 933–943. International World Wide Web Conferences Steering Committee (2018)
32. Kupavskii, A., et al.: Prediction of retweet cascade size over time. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 2335–2338. ACM (2012)
33. LaCour, M.: A balanced news diet, not selective exposure: evidence from a direct measure of media exposure. In: APSA 2012 Annual Meeting Paper (2015)
34. Lahoti, P., Garimella, K., Gionis, A.: Joint non-negative matrix factorization for learning ideological leaning on Twitter. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 351–359. ACM (2018)
35. Matakos, A., Terzi, E., Tsaparas, P.: Measuring and moderating opinion polarization in social networks. Data Min. Knowl. Disc. **31**(5), 1480–1505 (2017). https://doi.org/10.1007/s10618-017-0527-9
36. Morales, A., Borondo, J., Losada, J.C., Benito, R.M.: Measuring political polarization: Twitter shows the two sides of Venezuela. Chaos: Interdisc. J. Nonlinear Sci. **25**(3), 033114 (2015)
37. Munson, S.A., Lee, S.Y., Resnick, P.: Encouraging reading of diverse political viewpoints with a browser widget. In: Seventh International AAAI Conference on Weblogs and Social Media (2013)
38. Pettigrew, T.F., Tropp, L.R.: Does intergroup contact reduce prejudice? Recent meta-analytic findings. In: Reducing Prejudice and Discrimination, pp. 103–124. Psychology Press (2013)

39. Rajadesingan, A., Liu, H.: Identifying users with opposing opinions in Twitter debates. In: Kennedy, W.G., Agarwal, N., Yang, S.J. (eds.) SBP 2014. LNCS, vol. 8393, pp. 153–160. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05579-4_19
40. Ramponi, G., Brambilla, M., Ceri, S., Daniel, F., Di Giovanni, M.: Vocabulary-based community detection and characterization. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. SAC 2019, pp. 1043–1050. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3297280.3297384
41. Ramponi, G., Brambilla, M., Ceri, S., Daniel, F., Giovanni, M.D.: Content-based characterization of online social communities. Inf. Process. Manag., 102133 (2019). https://doi.org/10.1016/j.ipm.2019.102133, http://www.sciencedirect.com/science/article/pii/S0306457319303516
42. Sapienza, F., Groisman, P.: Distancia de fermat y geodesicas en percolacion euclidea:teoriaa y aplicaciones en machine learning. M.sc. thesis (2018). http://cms.dm.uba.ar/academico/carreras/licenciatura/tesis/2018/Sapienza.pdf
43. Shearer, E., Gottfried, J.: News use across social media platforms 2017. Pew Research Center 7 (2017)
44. Stewart, L.G., Arif, A., Starbird, K.: Examining trolls and polarization with a retweet network. In: Proceedings of the ACM WSDM, Workshop on Misinformation and Misbehavior Mining on the Web (2018)
45. Tran, T., Ostendorf, M.: Characterizing the language of online communities and its relation to community reception. arXiv preprint arXiv:1609.04779 (2016)
46. Trilling, D.: Two different debates? Investigating the relationship between a political debate on TV and simultaneous comments on Twitter. Soc. Sci. Comput. Rev. **33**(3), 259–276 (2015)
47. Van Der Maaten, L.: Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. **15**(1), 3221–3245 (2014)
48. Vaswani, A., et al.: Attention is all you need. CoRR abs/1706.03762 (2017). http://arxiv.org/abs/1706.03762
49. Venturini, T., Jacomy, M., Jensen, P.: What do we see when we look at networks. An introduction to visual network analysis and force-directed layouts. An introduction to visual network analysis and force-directed layouts, 26 April 2019 (2019)
50. Weller, K., Bruns, A., Burgess, J., Mahrt, M., Puschmann, C.: Twitter and Society, vol. 89. Peter Lang, Bern (2014)
51. Xiao, H.: Bert-as-service (2018). https://github.com/hanxiao/bert-as-service
52. Yang, X., Macdonald, C., Ounis, I.: Using word embeddings in Twitter election classification. Inf. Retrieval J. **21**(2–3), 183–207 (2017). https://doi.org/10.1007/s10791-017-9319-5
53. Yardi, S., Boyd, D.: Dynamic debates: an analysis of group polarization over time on Twitter. Bull. Sci. Technol. Soc. **30**(5), 316–327 (2010)
54. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**(4), 452–473 (1977)
55. de Zarate, J.M.O., Feuerstein, E.: Vocabulary-based method for quantifying controversy in social media. arXiv preprint arXiv:2001.09899 (2020)