Measuring Controversy in Social Networks through NLP

Juan Manuel Ortiz de Zarate <jmoz@dc.uba.ar>
Marco Di Giovanni <marco.digiovanni@polimi.it>
Esteban Feuerstein <efeuerst@dc.ub.ar>
Marco Brambilla <marco.brambilla@polimi.it>

Controversy/Polarization

- Often in social media users discuss about controversial topics
- Existing literature shows different issues that controversy brings up such as: splitting of communities, biased information, hateful discussions and attacks between groups
- The detection of controversy is crucial

Controversy/Polarization

- Identifying controversy on a discussion allows to apply strategies to mitigate it, such as:
  - Improving the “news diet”
  - Bridging echo chambers
  - Defending from attacks

- In this work we propose a new vocabulary-based technique for identifying and quantifying controversy on a discussion

Related work

- Many previous works are dedicated to quantifying the polarization observed in online social networks

- The main characteristic of those works is that the proposed measures are graph-based

- Garimella et al.[1] present an extensive comparison of controversy measures, different graph-building approaches, and data sources, achieving the best performance of all

[1] GARIMELLA, Kiran, et al. Quantifying controversy on social media. ACM Transactions on Social Computing

Previous work

- Ortiz de Zarate et al.[2] presents a first approach for quantifying controversy using text

- In this work we developed a new technique less dependent on the graph, wider comparison of NLP models, higher heterogeneity of datasets and languages and better performance

- This allows to make new kind of analysis: “semantic frontier”, “semantic distance”, a new kind of technique to prevent controversy and more

[2] ORTIZ DE ZARATE, Juan Manuel; FEUERSTEIN, Esteban. Vocabulary-based Method for Quantifying Controversy in Social Media, ICCS2020

Datasets

- We use 30 different discussions that took place between 2015 and 2020, half of them with controversy and half without it

- They are in 6 languages: English, Portuguese, Spanish, French, Korean and Arabic

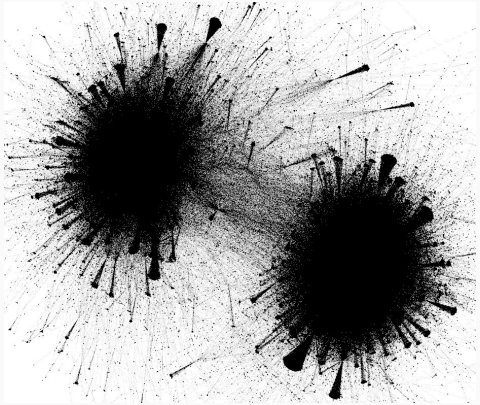
- Since our models require a large amount of text for training and since a tweet contains no more than 240 characters, we established a threshold of at least 100000 tweets

Datasets
Ground-truth definition

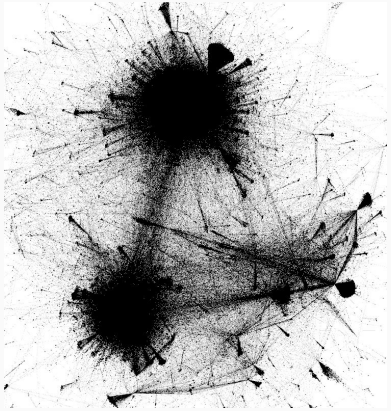
- To select new discussions and to determine if they are controversial or not we looked for:
  - Topics widely covered by mainstream media, and that have generated ample discussion, both online and offline
  - For non-controversy discussions we focused on “soft news", entertainment, impactful and/or dramatic
  - To validate that intuition, we manually checked a sample of tweets
  - For controversial debates we focused on political events such as elections, corruption cases or justice decisions.
- To furtherly establish the presence or absence of controversy in our datasets, we visualized the corresponding networks through ForceAtlas2

ForceAtlas2
Examples

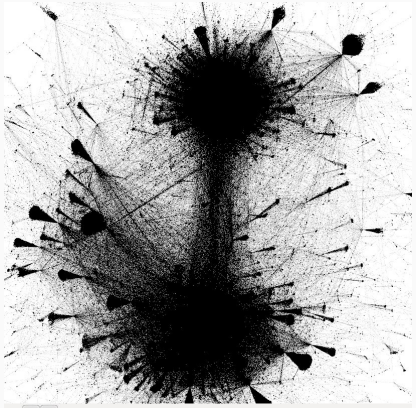
Kavanaugh discussion



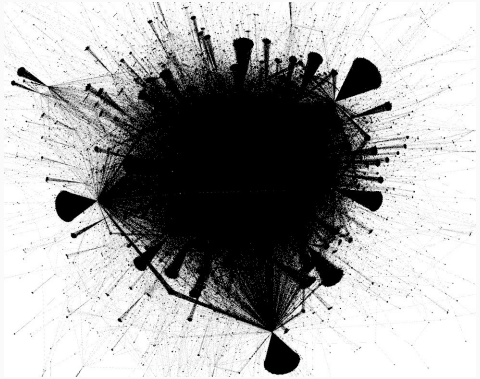
Macri discussion



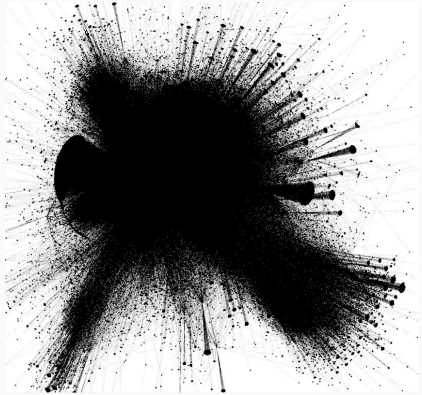
Bolsonaro discussion



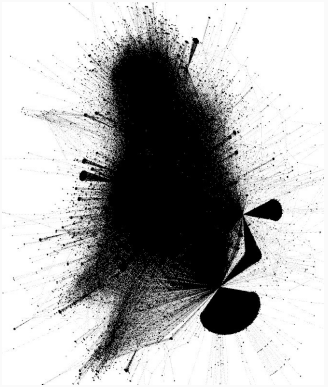
Halsey discussion



#EXODEUX discussion



#KingJacksonDay discussion



Datasets

Hashtag/Key wrod	#Tweets	Description and period
#LeadersDebate	250 000	Candidates debate, Nov 11-21,2019
pelosi	252 000	Trump Impeachment, Dec 06,2019
@mauriciomacri	309 603	Mentions to argentinian ex-president, Apr 05-11, 2018
Kavanaugh	260 000	Kavanaugh nomination, EEUU, Oct 05, 2018
Bolsonaro	170 764	Brazilian elections, Oct 27, 2018
#Al-HilalEntertainment	221 925	Al-Hilal champion, Dec 01,2019
#MiracleOfChristmasEve	251 974	Segun Woo singer birthday, 23-12-2019
Notredam	200 000	Notredam fire, Apr 16, 2019
#Wrestlemania	260 979	Wrestlemania event, Apr 08,2019
Messi	200 000	Lionel Messi Birthday, Jun 24, 2019

Methodology

- We propose a new technique for quantifying controversy by using a vocabulary approach

- Our method is a systematic technique for detecting controversy on any social digital network discussion

- It has a pipeline of 4 phases:
  - Graph Building phase
  - Community Identification phase
  - Embedding phase
  - Controversy Score Computation phase

Graph Building

- The objective is to build a network that represents the activity related to the discussion

- For each topic, we build a graph where we assign a vertex to each user who contributes to it and we add a directed edge from node *U* to node *V* whenever user *U* retweets a tweet posted by *V*

- Retweets typically indicate endorsement

Community Identification

- To identify a community's jargon we need to be very accurate at defining its members

- We cluster the graph using the popular algorithm Louvain
  - Structure-based algorithms
  - Not a fixed number of clusters
  - Very good performance

- We take the two biggest communities

Embedding

- In this phase, our purpose is to embed each user into a corresponding vector
- Tweets belonging to the users of the two principal communities selected in the previous stage are grouped by user and sanitized
- To estimate the embeddings we selected two models: Fasttext and BERT, and we trained them in a supervised way

Controversy Score Computation

- Take the 30% of the users with the highest authoritative and hub score, we call them *central users* and we use their embeddings
- Compute the two centroids of the clusters of the *central users* from C1 and C2 and the *global centroid*
- Compute D_1 and D_2, the two sums of distances of users of cluster C1 and C2 to their centroids, and the global distance D_{glob} as the sum of all users to the *global centroid*
- We try with Cosine, Euclidean, Mahalanobis and Manhattan distances

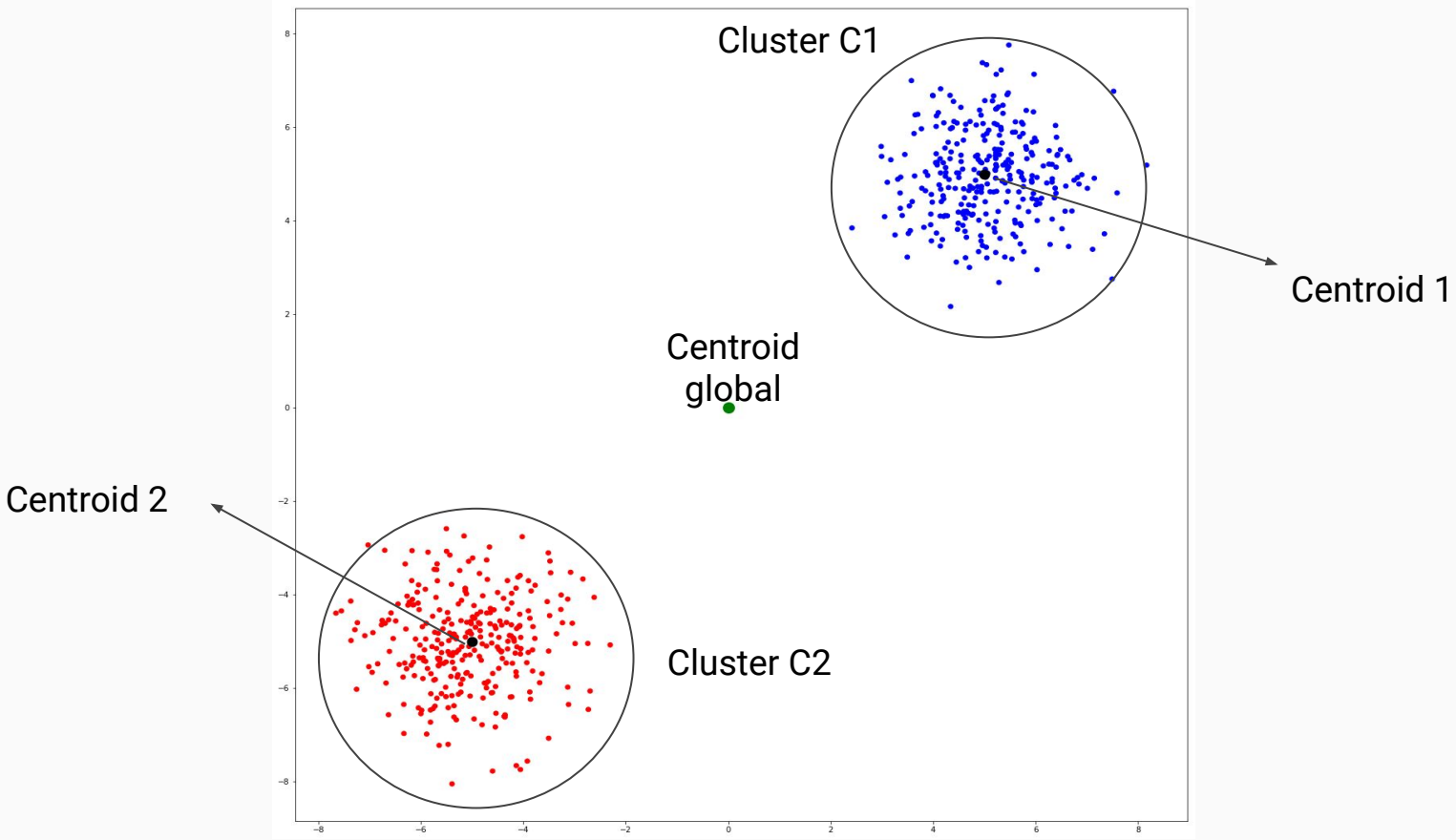
Controversy Score Computation

- Finally we compute the controversy score r with

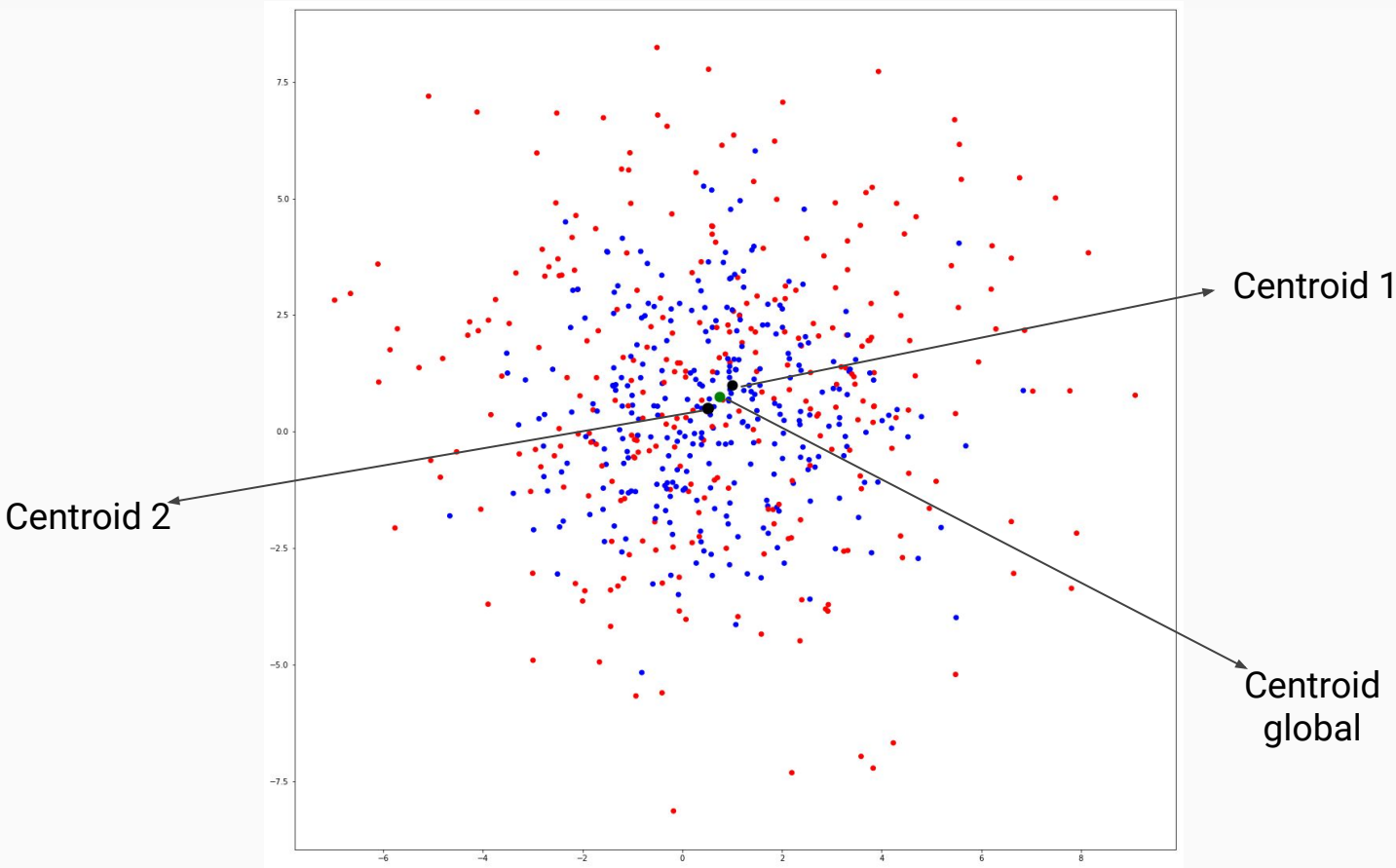
$$r = \frac{D_1 + D_2}{D_{glob}}$$

- *r* represents how much the clusters are separated
- If the dataset is a single cloud of points, this value should be near 1
- 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

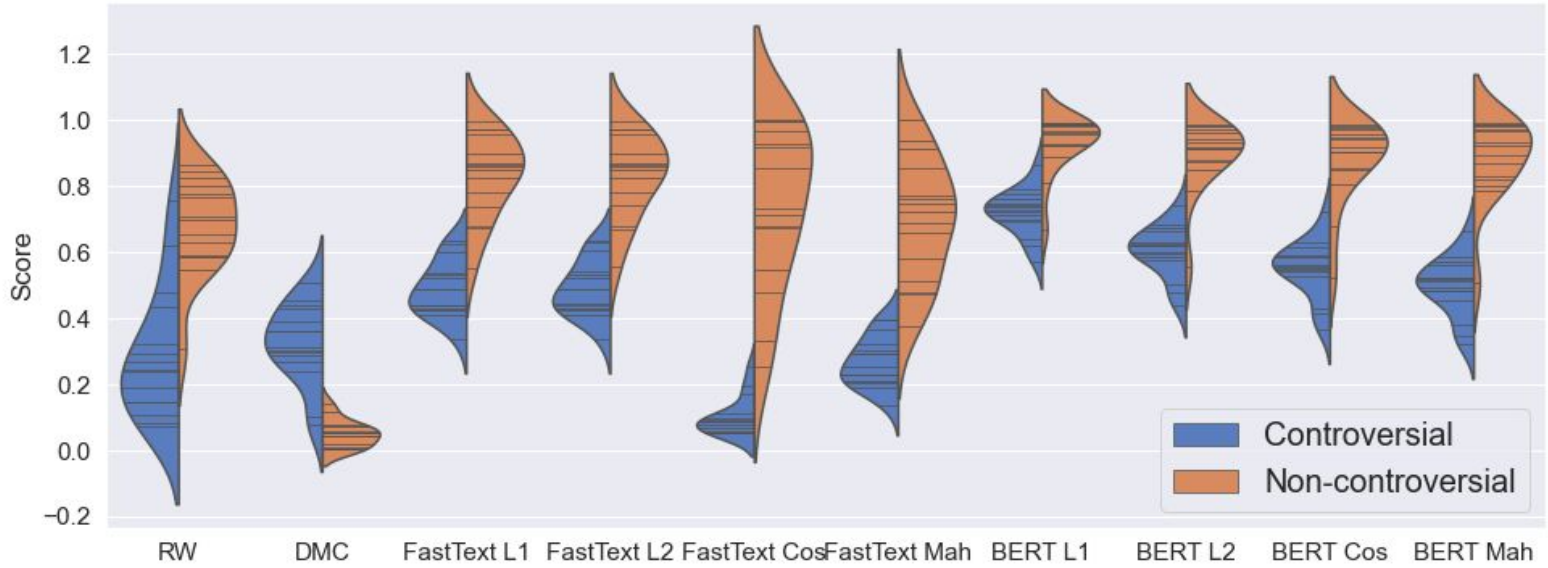
Controversial case



No controversial case



Results
Scores distributions comparison



Results
Scores distributions comparison

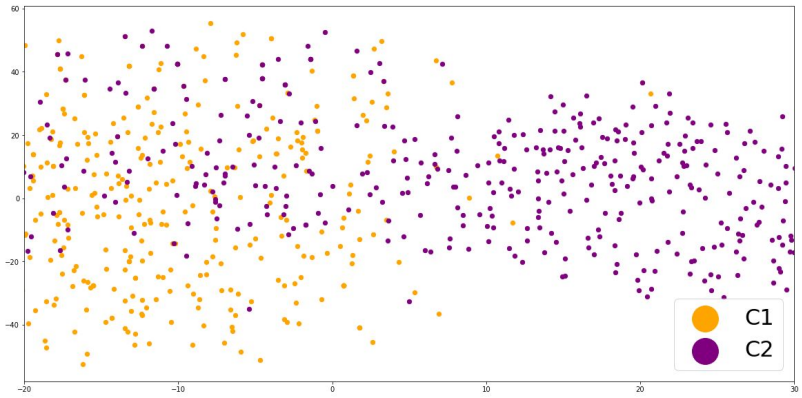
Method	Manhattan	Euclidean	Cosine	Mahalanobis	Baseline
Fasttext	0.987	0.987	**0.996**	0.991	
BERT	0.942	0.947	0.942	**0.964**	
DMC					**0.982**
RW					**0.924**

Discussion

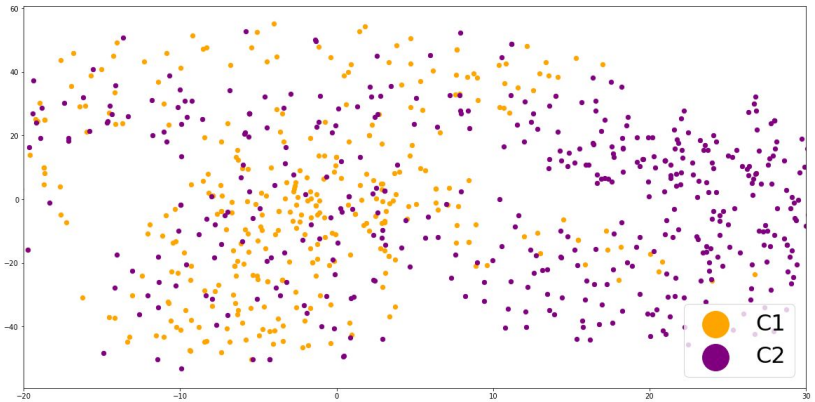
- Even if BERT paired many state-of-the-art results in different NLP tasks, FastText suits better in our pipeline
- We observe that BERT fails mainly with the non-controversial datasets
- Since BERT is a bigger and more complex model , it is able to separate the two communities' ways of speaking even when they are not opposite sides of a controversy, exploiting differences that we are not able to perceive
- To check this behaviour we plot the embeddings by t-SNE reductions

Results
Embedding reductions

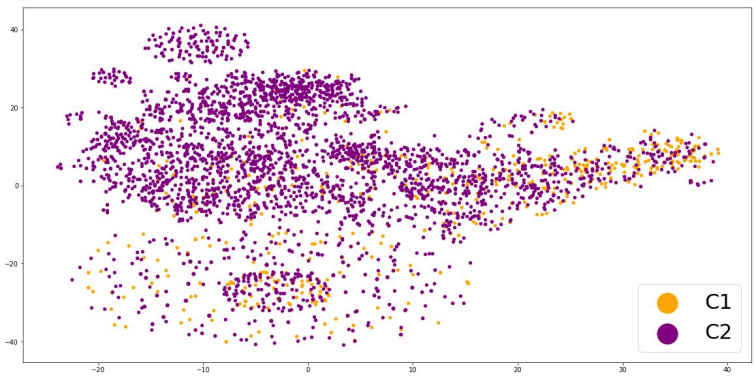
Feliz Natal BERT



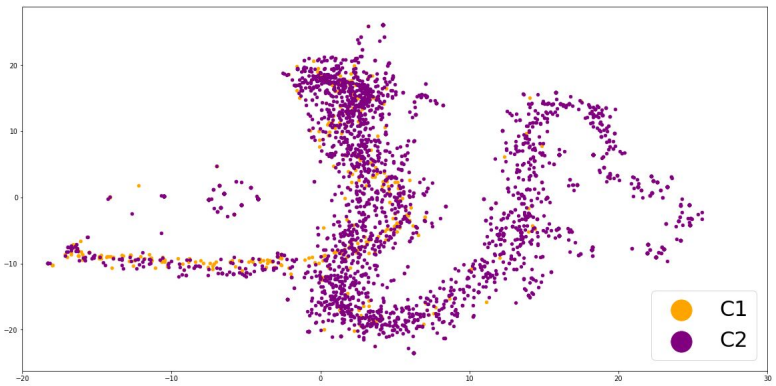
Feliz Natal Fasttext



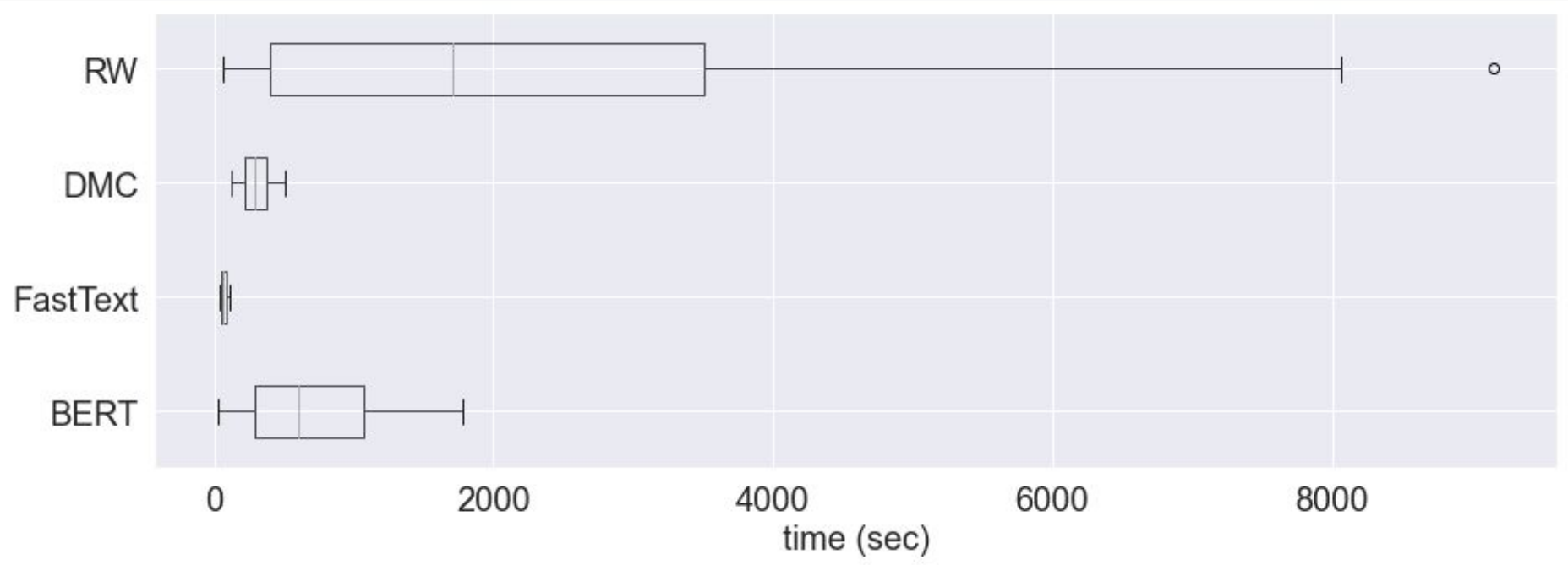
#KingJacksonDay BERT



#KingJacksonDay Fasttext



Results
Computing time by method



Conclusions

- We designed an NLP-based pipeline to measure controversy and we test some variants, such as two embedding techniques (using Fasttext and BERT language models) and four distance measures
- Our best approach, using FastText and cosine distance, outperforms the state-of-the-art graph-based method and also our previous work , in terms of ROC AUC score and speed
- 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

Discussion

- Limitations
  - Ground-truth, multi-sided and choice of data (same as Garimella et al.)
  - Data-size
  - Multi-language
  - Twitter Only
- Future works
  - User-related analysis, such as the detection of users that are in the “semantic border”
  - Analyze which users lay on opposite semantic sides to quickly detect the main differences between two communities
  - Detect and analyze the behaviours of users performing mixed interventions on a polarized debate

Thanks!

Questions?