Mining and Analysis of twitter users' stance in regards to 2020 US presidential candidates: a retrospective analysis

Giulio Fischietti
Department of Computer Engineering, Università di Pisa
Email: g.fischietti@studenti.unipi.it
N: 628043

1 Introduction

Repository location Indicate where the dataset repository is located.

Context The 2020 American Presidential Election was characterized by an intense competition between the two candidates: Donald Trump, for the republican party, and Joe Biden, for the democratic party.

During the period between September 2020 and January 2021 many events such as debates, impeachments and accusations for electoral fraud were reflected on social media discussions, with an increase of tweets posted by politicians and users' replies, which mostly expressed their opinion in a positive (*In Favor*) or negative (*Not in Favor*) manner.

Through analysis of these replies the main tendencies and preference of the electorate were found, and generated an algorithm that can distinguish tweets' and users' favouritism towards a specific politician and analyzed the change of their consensus along with the succession of events.

2 Method

This project can be broken down to following steps

Steps

1. Data Fetching
2. Data Cleaning
3. Data filtering
4. Data Labeling and Events Evaluation
5. Models Evaluation and comparision (static, incremental and sliding window)
6. Tweets' stance classification and analysis
7. Users' stance calculation and analysis

3 Data fetching, cleaning and description

The Dataset consists in multiple csv files containing:

- Joe Biden's tweets
- Donald Trump's tweets
- Replies to Joe Biden's tweets
- Replies to Donald Trump's tweets

In the period between September 1st and January 8th.

3.1 Data Fetch

The raw data has been fetched with 24 parallel scripts (to speed up the process), each one of them getting tweets from different and mutually exclusive intervals of dates, filtering through query parameters only the tweets that contained text and a mention to the respective candidate, excluding those that contained links/images.

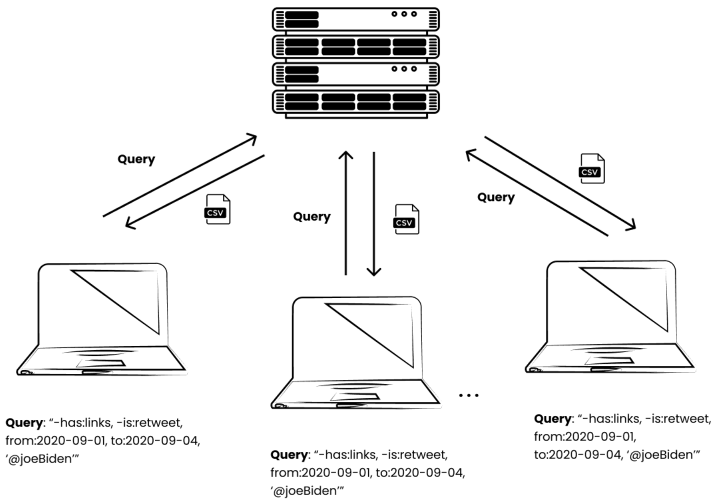

Figure 1: Workers that fetched the data through the Twint API

In the particular case of Donald Trump, which was banned since January 8th and his tweets not visible, the data was downloaded from this website.

The result of the queries is a list of csv files characterized by the following structure:

```
{"tweet": {
    "user_id": "id of the user",
    "id": "id of the tweet",
    "conv_id": "id of the conversation",
    "tweet": "content of the reply/tweet",
    "date": "date of the tweet in the format YYYY-MM-DD",
    }
}
```

3.2 Data Cleaning

After the fetch, a cleaning phase was done for every csv file, to unescape the texts, decode them correctly and remove non-text elements such as emojis.

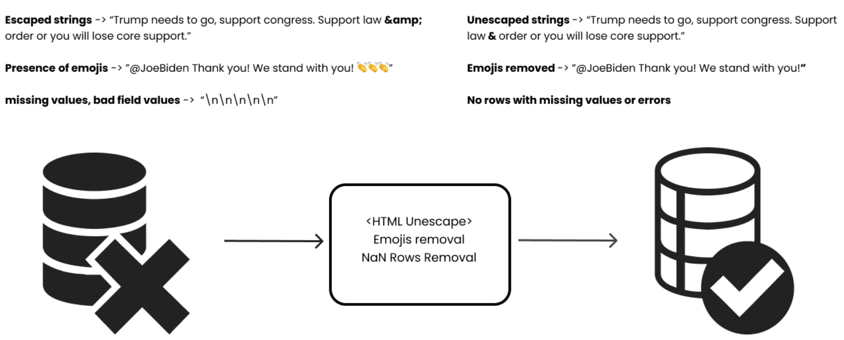

Figure 2: Data cleaning process

3.3 Data filtering

According to Twitter's documentation [2], a reply to an original post will have as conversation id the same id as the original tweet, but most importantly also the replies to other users' replies will share the same conversation id: this is why two extra filtering phases were necessary in order to obtain consistent results.

With this being said, all the tweets whose conversation id did not match any of the ids of the candidates' dataset, were removed: in particular, the remaining tweets that contained more than one mention were also removed from the dataset as it meant that the tweet was a reply to another user reply and not a direct reply to the candidate.

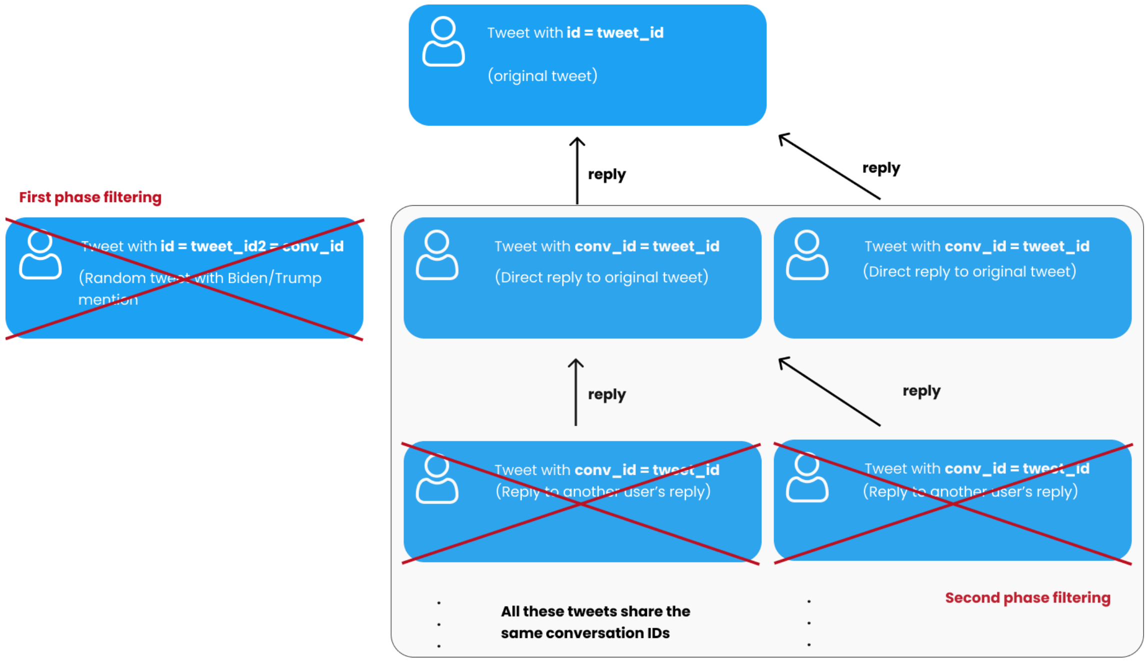

Figure 3: Data filtering process: all non-direct replies are removed

4 Data labeling and Events evaluation

The result of fetch clean and filtering of the replies are two datasets with the properties shown in Table 1, and the distribution of replies over time shown in figure 4 and figure 5: unsurprisingly, Trump and his followers have been particularly active during the same period of time counting almost ten times more replies and tweets than the other candidate.

	N of replies to Biden	N of replies to Trump
Total	2 171 689	19 099 744
Average (per day)	16 705	145 799
Standard Deviation	13 453	100 456
Max (in a Day)	70 959	757 679

Table 1: Dataset

To obtain the Train Sets, the period of time between September 1st and September 30th was chosen, extracting in that period a sample of 671 replies directed to each candidate, successively labeled as *In Favor* or *Not In Favor* (Table 2). Multiple queries were necessary in order to obtain a somewhat balanced dataset, as the opinion of the vast majority of repliers was polarized as *Not In Favor*.

Table 2: Train sets Class distribution obtained after the labelling phase

	N of replies to Trump	N of replies to Biden
Labeled as In favour	265	261
Labeled as Not in favour	406	410
Total	671	671

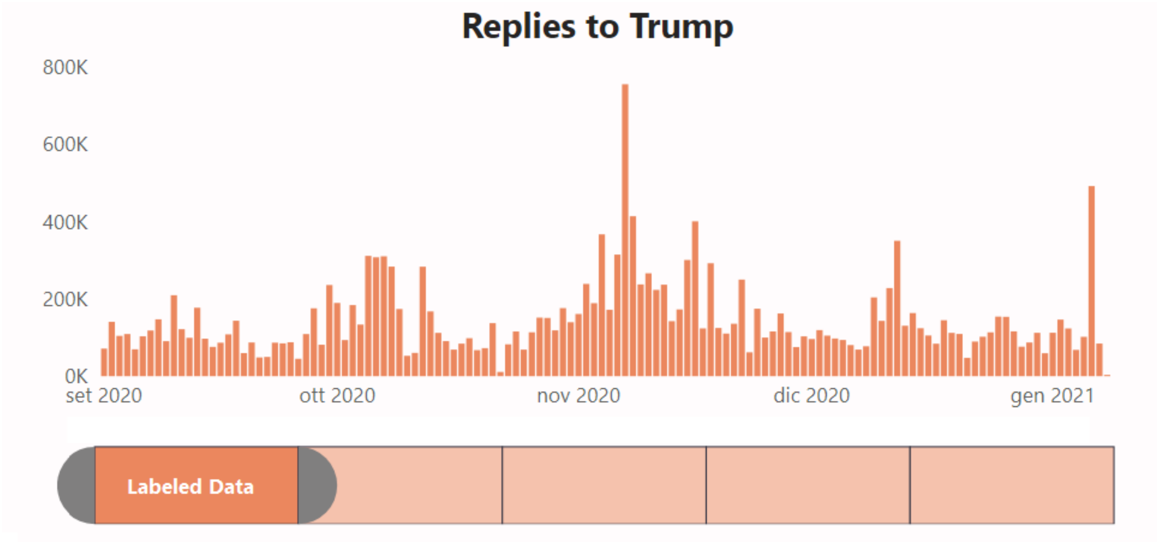

Figure 4: Replies to Donald Trump's tweets in the period between September 1st and January 8th

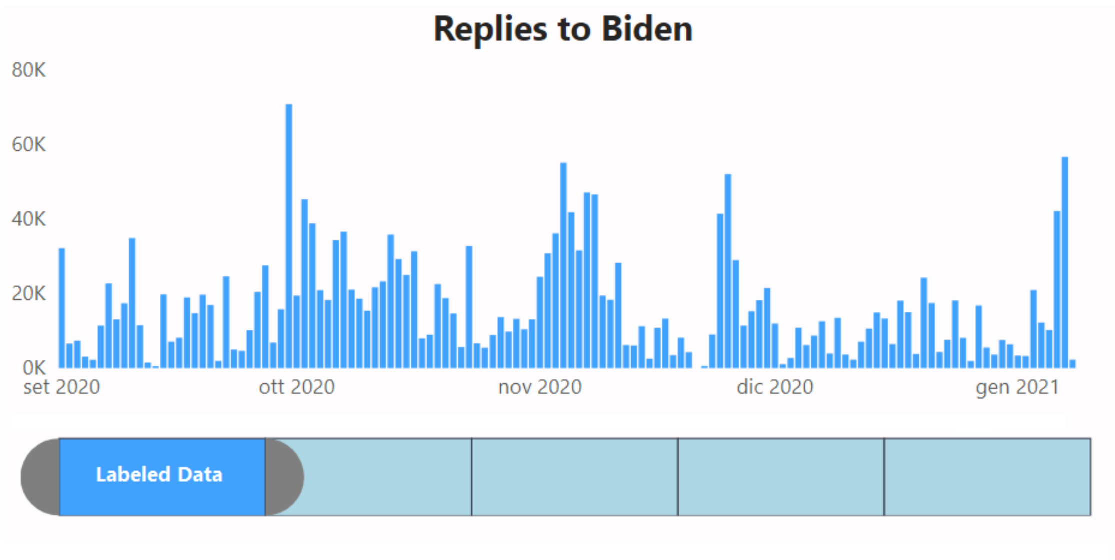

Figure 5: Replies to Joe Biden's tweets in the period between September 1st and January 8th

We know that for each peak of replies, most of the time, corrisponds an important event (in this case impeachments, debates, provocatory tweets etc...) that resonated in this social media: 6 peaks were identified for Trump and 8 peaks were identified for Biden, which some of the most relevant corrisponded to the following events, highlighted in fig.6 and fig.7:

- September 30th: first presidential debate
- October 5th: Trump, previously infected by Covid, is released from hospital
- October 12th: second presidential debate

- November 7th: Biden wins in Pennsylvania, winning the elections: Trump keeps calling for election fraud
- January 6th: Trump supporters storm Capitol

As different topics and events happened, they could possibly have caused a concept drift, a phenomena in which in the particular case of text mining and stance detection, means a change in the words used to express an opinion contrary or *In Favor* to a tweet leading to degradation of performance of a classification model over time: for this reason a certain number of peaks were selected from the replies of both candidates and for each peak a sample of 70 tweets was extracted, using them as a train set for a sliding window/incremental classifier model, then compared with a static model as shown in the following sections.

In the following figures is shown more in detaill the peaks and events considered in this analysis:

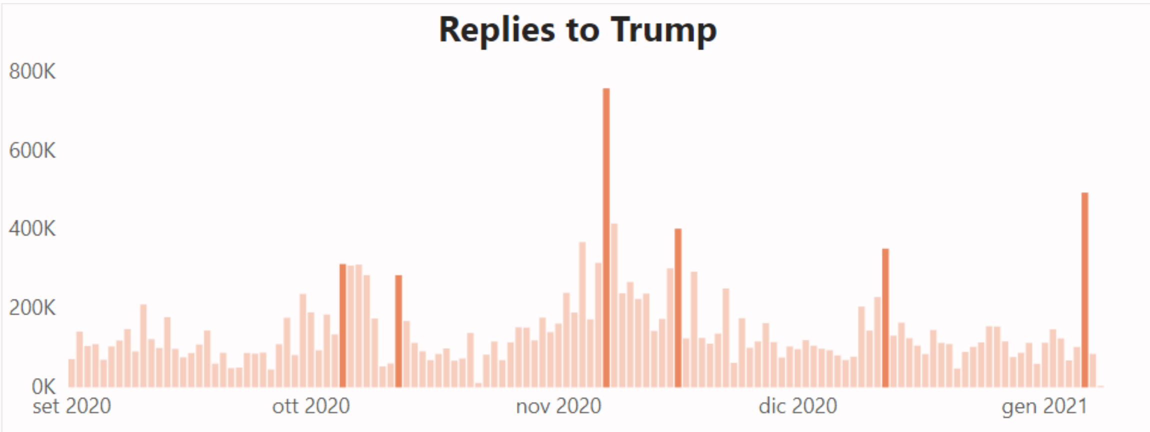

Figure 6: Peaks of Trump's twitter replies

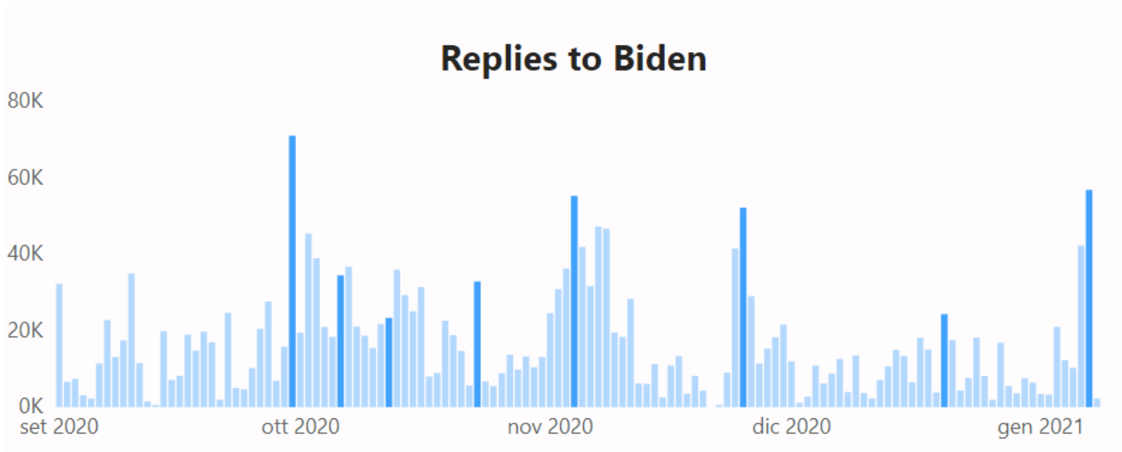

Figure 7: Peaks of Biden's twitter replies

5 Generating the models

In this analysis were trained and tested multiple classifiers and the one with the highest weighted f1-score was selected, and used to generate the static model: then its performance will be compared with the incremental and sliding window models.

The rapresentation of text used is the TF-IDF: after tokenizing and stemming the words, some tdidf-classifiers pipeline were tuned through a grid search, choosing the combination that lead to the best results.

In the following sections for each classifier will be shown the results for each candidate: while referring to the same algorithm, the model is actually different as it is generated from two different train sets obtained from each candidate replies.

5.1 Static Model

The static model conists in a model trained with the data collected and labeled in an initial period of time (in this case the month of September), obtaining a model used to perform the detection/test in the whole period of time (in this case, until Januray 8th).
To do so, multiple classifiers such as Logistic Regression, Support Vector Machine, Passive Aggressive Classifier and Multinomial Naive-Bayes were considered, training and testing them in 5-fold cross validation.
For each one of them a Grid Search was performed, tuning the hyperparameters and choosing the best combination in terms of performance.

5.1.1 Logistic Regression

hyperparameter	Biden	Trump
C	1 000	1 000
penalty	l2	l2
max features	5 000	10 000
ngram range	1-2	1-3
tfidf norm	l1	l2

	Biden			Trump		
	P	R	f1	P	R	f1
Not in Favour	0.88	0.76	0.81	0.89	0.79	0.84
In favour	0.58	0.75	0.65	0.64	0.79	0.70
Weighted	0.79	0.76	0.77	0.81	0.79	0.79

Table 3: Best combination for Logistic Regression of Hyperparameters and table of CV results

5.1.2 Support Vector Machine

hyperparameter	Biden	Trump
C	100	1
gamma	0.001	0.001
max features	10 000	10 000
kernel	linear	linear
ngram range	1-3	1-3
tfidf norm	l2	l2

	Biden			Trump		
	P	R	f1	P	R	f1
Not in Favour	0.85	0.77	0.81	0.93	0.78	0.85
In favour	0.61	0.72	0.67	0.58	0.85	0.69
Weighted	0.77	0.76	0.76	0.84	0.80	0.81

Table 4: Best combination for SVM of Hyperparameters and table of CV results

5.1.3 Multinomial Naive Bayes

hyperparameter	Biden	Trump
alpha	0.5	0.9
max features	2 000	1 000
ngram range	1-3	1-2
tfidf norm	l2	l2

	Biden			Trump		
	P	R	f1	P	R	f1
Not in Favour	0.80	0.78	0.79	0.84	0.81	0.82
In favour	0.65	0.68	0.66	0.63	0.73	0.71
Weighted	0.74	0.74	0.74	0.78	0.78	0.78

Table 5: Best combination for Multinomial NB of Hyperparameters and table of CV results

5.1.4 Passive Aggressive Classifier

hyperparameter	Biden	Trump
C	100	1
loss	sq hinge	sq hinge
tolerance	1e-04	1e-05
max features	5 000	10 000
ngram range	1-3	1-3
tfidf norm	l1	l2

	Biden			Trump		
	P	R	f1	P	R	f1
Not in Favour	0.82	0.77	0.80	0.91	0.79	0.85
In favour	0.63	0.70	0.66	0.62	0.82	0.71
Weighted	0.75	0.75	0.75	0.80	0.80	0.80

Table 6: Best combination for Passive Aggressive of Hyperparameters and table of CV results

5.1.5 Final choise

The final choise was the support vector machines as in terms of performance (accuracy and weighted f1-score) reported significantly higher scores than the other classifiers: in particular, it has been chosen over the Passive Aggressive classifier as it had a lower training time, so this model will be use as static and then compared with its sliding window and incremental versions.

5.2 Incremental model

In the incremental approach the model was trained iteratively after the occurrence and detection/test of every event, adding in the dataset the sample extracted from that peak, consisting in 70 labeled tweets each.

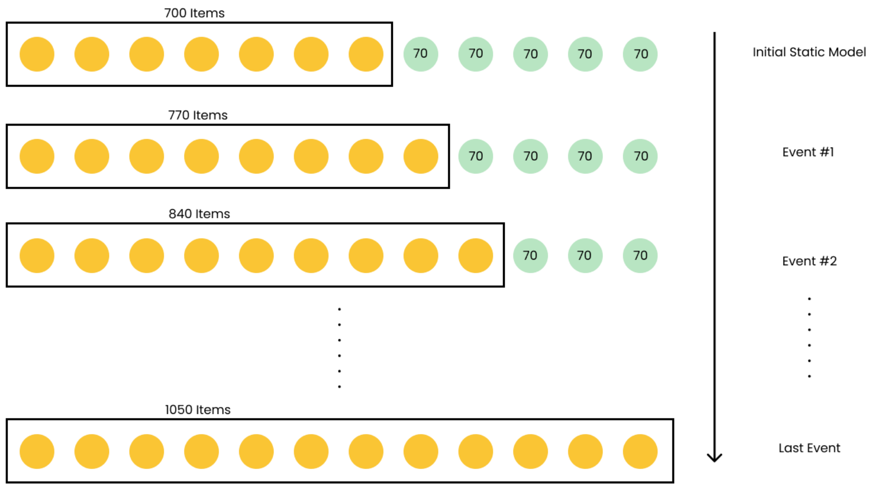

Figure 8: A simple scheme of the incremental model learning process

The result is a model whose dataset which uses to train increases over time, leading possibly to best performances but also to an increase of the training time, number of features and overall model size.

5.2.1 Static vs Incremental model

Comparing the performance in terms of f1-score of the static and incremental models (figure 8 and figure 10) we find some performance advantage in using an incremental model over the static one (only in the case of Biden) with about 4-5% (for Trump there was no significaant improvement) difference in score, while in both cases a loss of accuracy/fscore was experienced in both models in the 4th and 7th event which suggest a major concept drift after December: it would have been required a bigger sample for each peak in order to try to avoid these losses of accuracy/fscore.

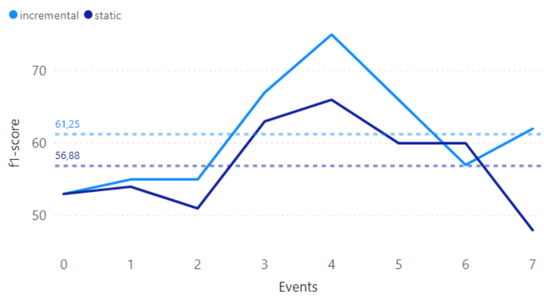

Figure 9: Biden model Incremental and static performance comparision

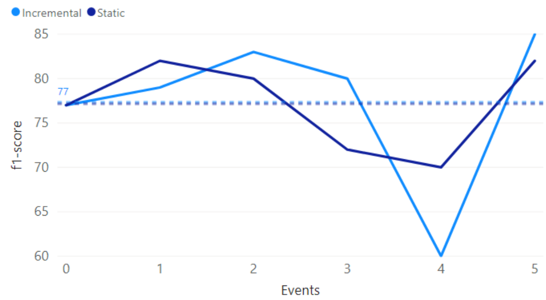

Figure 10: Trump model Incremental and static performance comparision

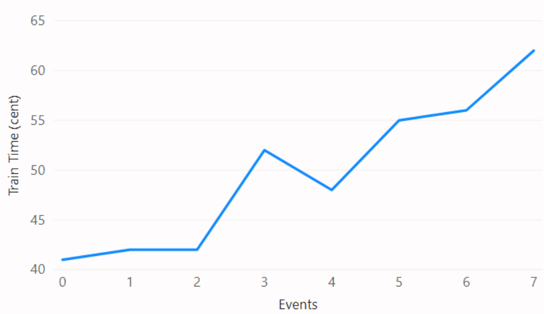

Figure 11: Training time of the incremental (Biden) model across the events

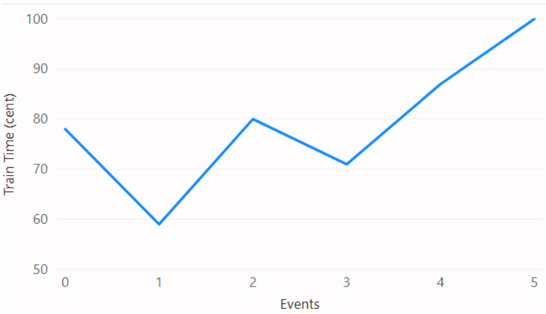

Figure 12: Training time of the incremental (Trump) model across the events

As expected, an increase in training time was experienced for both incremental models, in a linear manner.

5.3 Sliding window model

In the sliding window approach the model was trained iteratively after the occurrence and detection/test of every event, adding in the dataset the sample extracted from that peak, and removing the last sample, keeping constant the size of the training set.

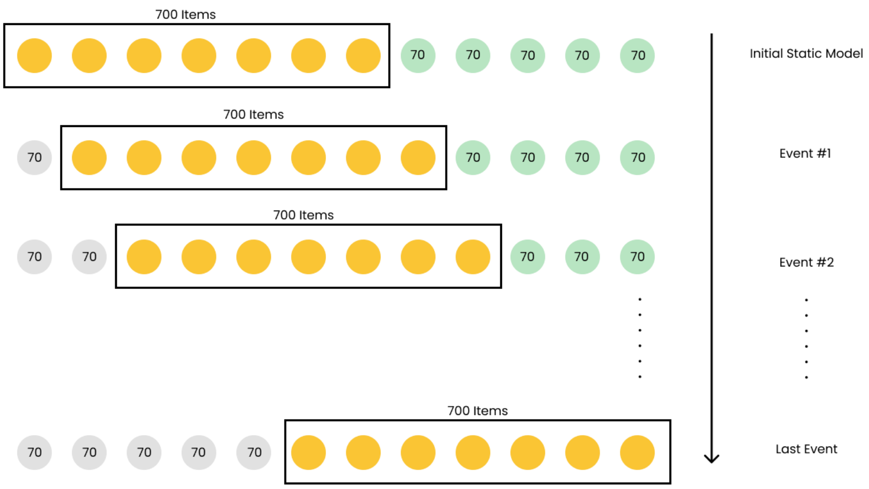

Figure 13: A simple scheme of the sliding window model learning process

The result is a model whose dataset which uses to train does not increase over time, keeping the model to an approximately constant size (as well as the training time required) while being able to adapt over the course of events.

5.3.1 Static vs Sliding window model

Also in this case choosing an adaptive model over the static one gave better results, only in the case of Biden (for Trump there was no significaant improvement) with 5-6% difference in score, while in both cases a loss of accuracy/fscore was experienced in both models in the 4th and 7th event which suggest a major concept drift after December and possibly it would have been required a bigger sample for each peak in order to try to avoid these losses of accuracy/fscore.

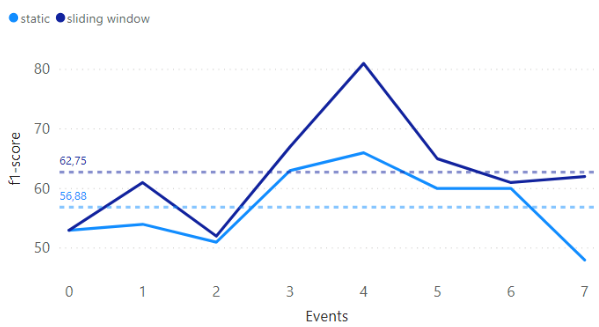

Figure 14: Biden model Sliding Window and static performance comparision

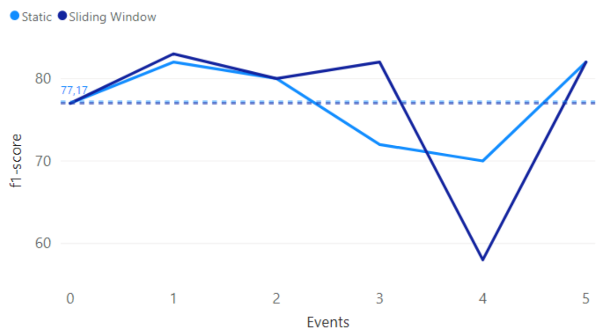

Figure 15: Trump model Sliding Window and static performance comparision

As expected the time required to perform each train was did not increase and remained almost constant over all the events

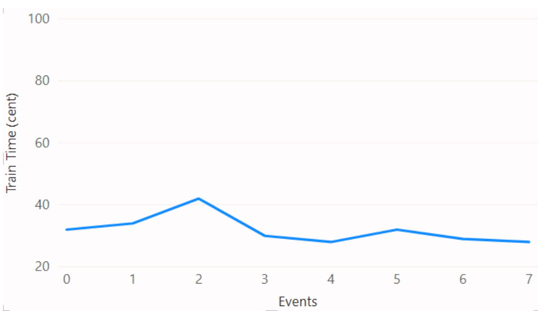

Figure 16: Training time of the Sliding Window (Biden) Model over the events

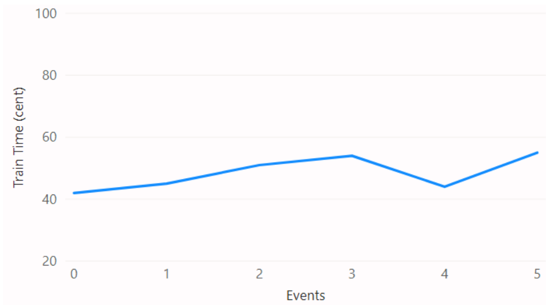

Figure 17: Training time of the Sliding Window (Trump) Model over the events

5.3.2 Final choises

After the previous considerations the models chosen to perform the classification of the tweets (replies) are:

- **Sliding Window** Model for replies to Joe Biden's tweets: as it showed significant improvements compared to the static and Incremental model, and also being lighter than the Incremental;
- **Static** Model for replies to Donald J. Trump's tweets: as the Sliding Window and Incremental models did not show significant improvements, the most simple and lighter solution was the Static Model;

6 Results

The results of the classification are analyzed in terms of *Tweet stance* and *Users' stance*: the analysis of tweet stance alone can not be enough to understand the public opinion, as there can be more active users who reply multiple times in a day (to the same post or in more tweets published by the candidate) and less active users (we also have to take into consideration that the results and the number of replies can be strongly dependent on how much the candidate is active, as it means greater confrontation and discussion with the followers.
In both analysis were taken in consideration both active and less active users.

6.1 Tweet Stance

The tweet stance was simply obtained classifying every tweet with the relative model and analyzing the results, without taking into account more active and less active users.

6.1.1 Joe Biden

For Biden, for the majority of time there was a prevalence of tweets classified as *Not In Favor* (with an average of 70 percent a day), but with the occurrence of some event the situation changed showing more In Favor tweets than the other classes.
(Results shown in Fig. 18)
Some of the peaks worth mentioning are:

- Over the whole month of September the tweets were mostly *Not In Favor* as the first debate approached (which would bring a significant increase of tweets in favor reaching a ratio of 1, almost), with a ratio of in favor and *Not In Favor* between 0.28 and 0.34
- November 4th, day after the end of elections, tweets probably expressed by supporters, with a ratio of In Favor and *Not In Favor* of 1.53
- November 7th, Biden officially wins in the state of of Pennsylvania, automatically winning the election: Trump refuses to leave the White House, ratio is 1.14

6.1.2 Donald J. Trump

For Trump, the biggest fraction of replies has been classified by the model as "*Not In Favor*" for the whole time, showing an overall in regards to Trump, exception made for some days before the 5th of October, in which we can almost see a tie between Favor and *Not in Favor*, as that day he was released from the hospital and off COVID-19. (Results shown in Fig. 19)
Some of the peaks worth mentioning are:

- Over the whole month of September the tweets were mostly *Not In Favor* as the first debate approached with a ratio of in favor and *Not In Favor* between 0.12 and 0.16
- October 3th and 4th he made some statements about his health as he was sick because of COVID-19, many tweets in his favor were posted to express support, with a ratio between 0.66 and 0.84
- December 23rd and 25th, Washington Times stated: "President Trump said he'll ask Congress to revise the just-passed 900 billion COVID relief package, saying it contains too much foreign aid and not enough help for struggling Americans." this news pssibly taken positively by his supporters bringing the ratio to 0.48

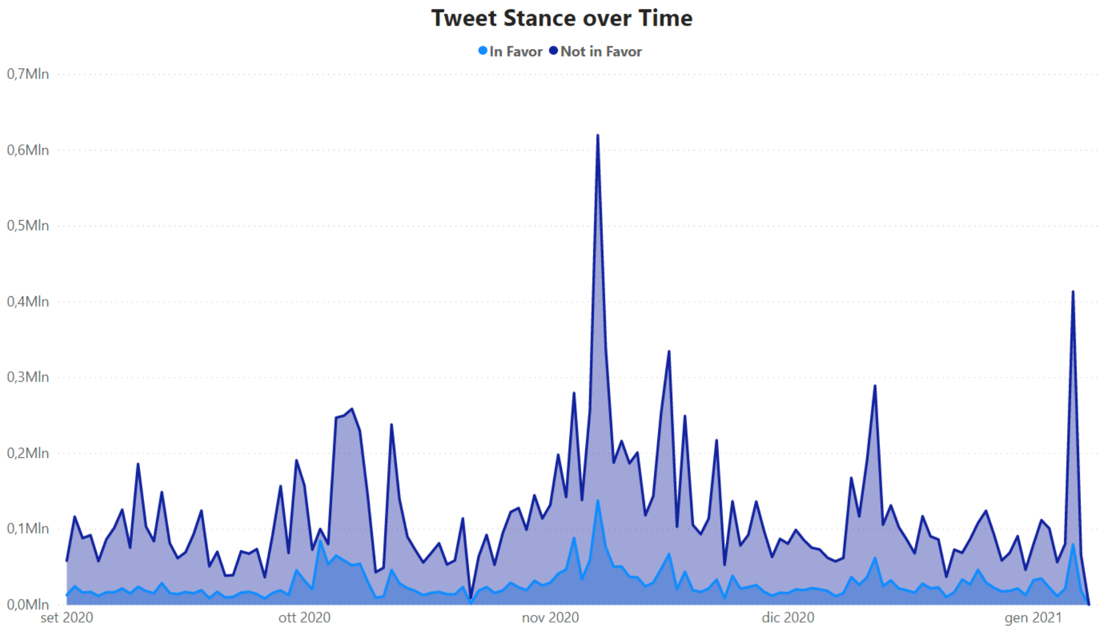

Figure 18: Tweet stance in regards to Trump

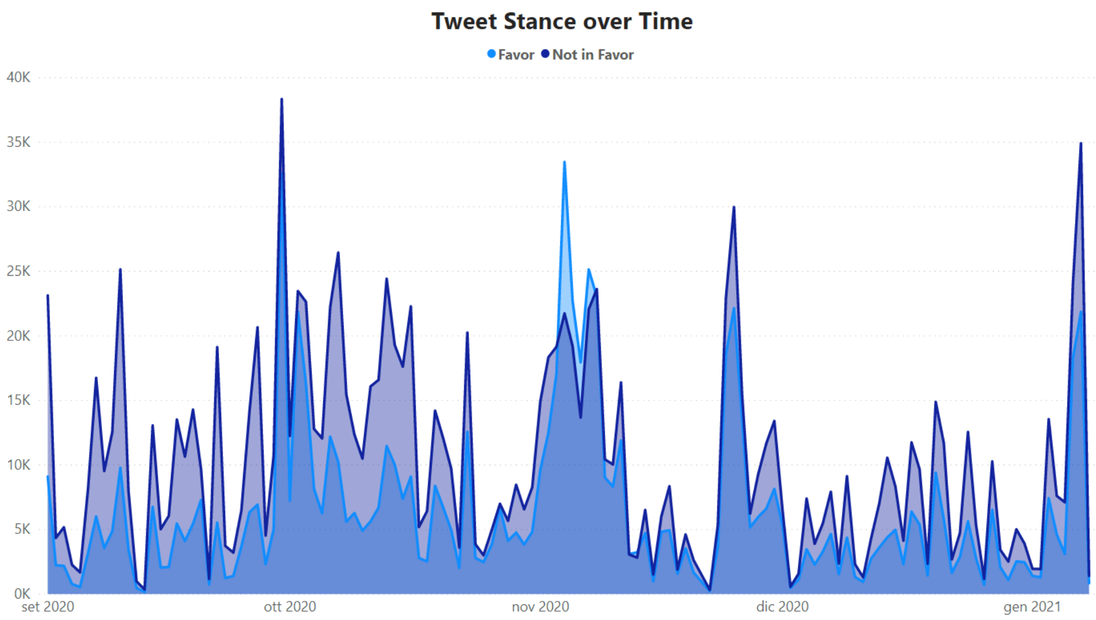

Figure 19: Tweet stance in regards to Biden

6.2 Users' Stance

As said before, there can be more active users that posts multiple replies making tweets stance a non rapresentative analysis of the public opinion so in this section is calculated and analyzed the stance of every user.
The Users' Stance was calculated using following parameter defined formulas, used in other [1]

$$UserScore = \frac{Positive + Negative}{Total}$$

And then discretized to get a Positive, Negative or Neutral Stance outcome:

$$UserStance = \begin{cases} UserScore > 0{,}5 & InFavor \\ UserScore < 0{,}5 & NotInFavor \\ |UserScore| \leq 0{,}5 & Neutral \end{cases}$$

This formula is applied for each time span between events, as the occurrence of an event can be crucial for the evolution of the opinion of every single user.

6.2.1 Joe Biden

In regards to Joe Biden the portion of users *In Favor* started from 30.5 percent, increasing significantly after the first event (first presidential debate) and kept increasing until reaching 41.6 percent, while the portion of neutral neutral decreased as well as the *Not In Favor*, even if in a slightly moderate manner.

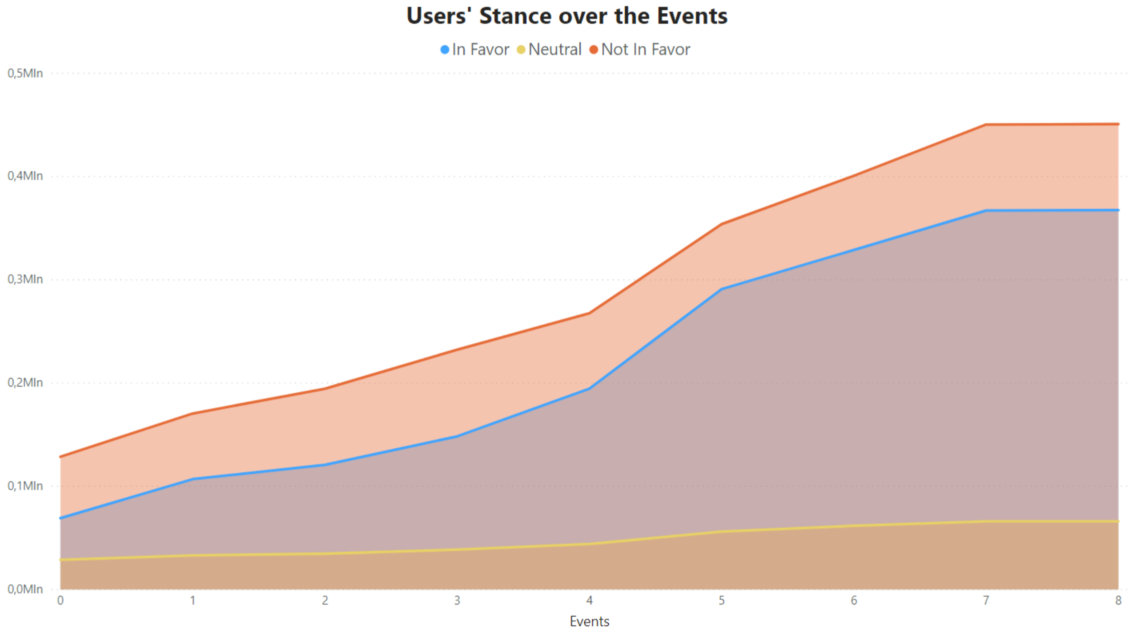

Figure 20: Users stance in regards to Biden over all the events

The greatest increase in consense was recorded at the 4th event, when he officially won the elections, while a slight decrease was noticed from the 5th and beyond, while Trump called for election fraud and his supporters ultimately stormed Capitol.

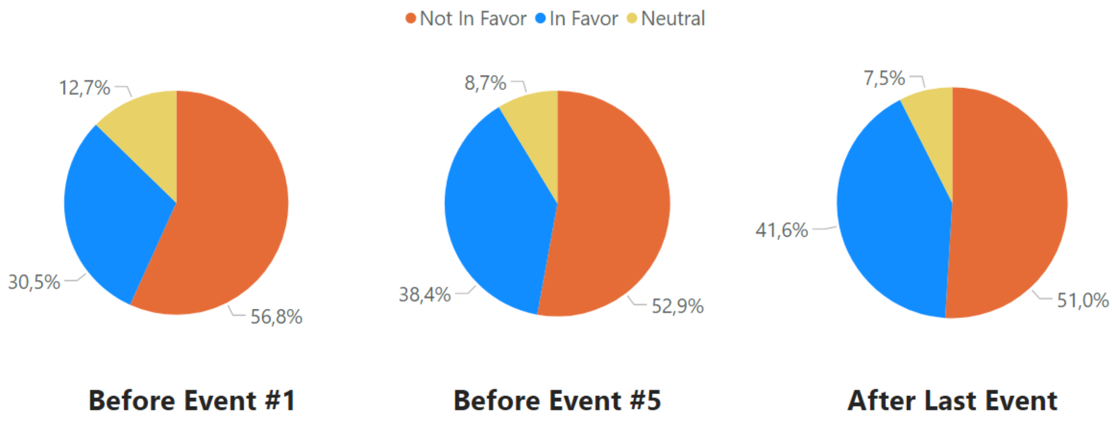

Figure 21: Users stance in regards to Biden at key events

6.2.2 Donald J. Trump

In regards to Donald Trump the portion of users in favor started from 16.3 percent, increasing significantly after the first event (Hospital dimission) reaching 22.8 percent and kept an almost constant value, while the portion of Not in Favor decreased after the event 3 and increased again after the last event, as Trump sopporters stormed Capitol.

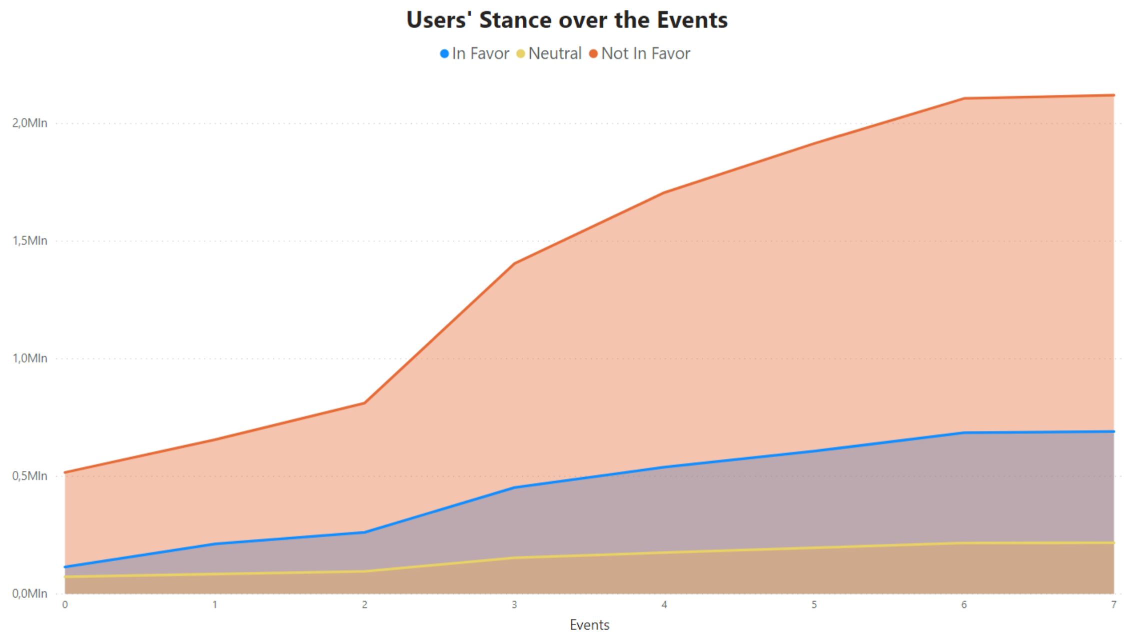

Figure 22: Users stance in regards to Trump over all the events

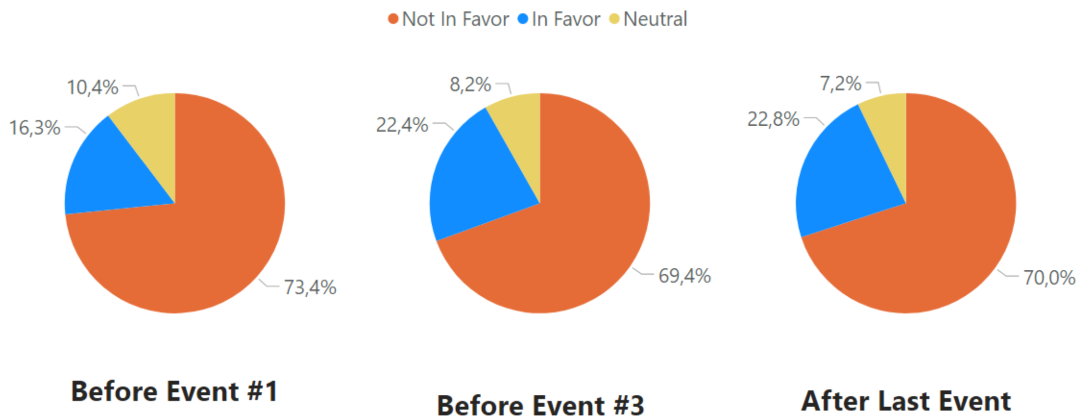

Figure 23: Users stance in regards to Trump at key events

References

[1] Francesco Marcelloni Alessio Bechini Pietro Ducange and Alessandro Renda. "Stance Analysis of Twitter Users: the Case of the Vaccination Topic in Italy". en. In: *IEEE* (2020). DOI: 10.1109/MIS.2020.3044968,.

[2] Twitter Developers. *Conversation threads using the Twitter API*. URL: https://developer.twitter.com/en/docs/twitter-api/conversation-id.