Content-based Stance Classification of Tweets about the 2020 Italian Constitutional Referendum

Marco Di Giovanni
Politecnico di Milano, Milan, Italy
Università di Bologna, Bologna, Italy
marco.digiovanni@polimi.it

Marco Brambilla
Politecnico di Milano, Milan, Italy
marco.brambilla@polimi.it

Abstract

On September 2020 a constitutional referendum was held in Italy. In this work we collect a dataset of $1.2M$ tweets related to this event, with particular interest to the textual content shared, and we design a hashtag-based semi-automatic approach to label them as Supporters or Against the referendum. We use the labelled dataset to train a classifier based on transformers, unsupervisedly pre-trained on Italian corpora. Our model generalizes well on tweets that cannot be labeled by the hashtag-based approach. We check that no length-, lexicon- and sentiment-biases are present to affect the performance of the classifier. Finally, we discuss the discrepancy between the magnitudes of tweets expressing a specific stance, obtained using both the hashtag-based approach and our trained classifier, and the real outcome of the referendum: the referendum was approved by 70% of the voters, while the number of tweets against the referendum is four times greater than the number of tweets supporting it. We conclude that the 2020 Italian constitutional referendum was an example of event where the minority was very loud on social media, highly influencing the perception of the event. Based on our findings, we suggest that drawing conclusion following only social media analysis should be performed carefully since it can lead to extremely wrong forecasts.

1 Introduction

On September 20 and 21, 2020, a constitutional referendum was held in Italy to reduce the number of parliamentarians (from 630 to 400). 69.96% of the voters approved it, with a voter turnout of about 51%[1]. Since the main Italian political parties supported the referendum, at first the outcome was obvious, but, through a huge activity on social media, opposers unsuccessfully tried to overturn the result. The referendum was a *confirmatory* referendum: voters were asked to approve a law. Thus, we refer to people that voted "yes", agreeing with the introduction of the new law that reduces the number of parliamentarians, as Supporters, and we refer to people that voted "no", against the introduction of the new law, as Opposers.

Since an always greater number of people share their thoughts online, social network analysis helps understanding the causes and forecasting the outcomes of political events, in parallel with already widely used approaches such as surveys and pools (Callegaro and Yang, 2018). Like surveys, *selection biases* are hard to remove. Social media users and citizens have different demographic distributions, resulting in under-represented categories of people (e.g., elderly people) (Mislove et al., 2011)[2]. Moreover, social media are also populated by bots, softwares that run accounts and automatically share content, introducing noise and bias in the collected data (Ferrara et al., 2016). These accounts are not run by real people and the data shared by them should not be included to perform analysis and statistics. However, a big advantage of the analysis of social media data is the higher magnitude of available data, easy to collect and process. It is often less expensive to collect content from social media than using classical approaches.

In this study we collect and analyze Twitter data about the Italian referendum in 2020. Our contributions can be summarized as follows:

- We collect and publicly share a corpus of $1.2M$ tweets about the Italian referendum in 2020. This is a rare and fundamental resource for NLP analysis, expecially stance detection, for non-English texts[3];

[1] https://en.wikipedia.org/wiki/2020_Italian_constitutional_referendum

[2] https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/

[3] The dataset is publicly available at https://github.com/marco-digio/italian-referendum-2020

Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, pages 14–23
Online Workshop, June 10, 2021.

- We design a content-based, semi-automatic, approach to label big magnitudes of textual data through hashtags. We obtain a set of $85k$ cleaned labeled texts with low human effort;
- We fine-tune an accurate text classifier to detect the stance of tweets (Support or Against the referendum). We also successfully apply it to classify tweets that the semi-automatic approach *cannot* label;
- We inspect three common text biases (length-bias, lexical-bias and sentiment-bias), observing that our dataset does not suffer from them;
- We discuss the discrepancy between the collected data from Twitter and the real outcome of the referendum, including possible further investigation essential to understand the phenomenon.

2 Related Works

Numerous published works correlate social media data with elections or referendums. The main and most studied recent event is the Brexit referendum, largely investigated from many different points of view (Howard and Kollanyi, 2016; Grčar et al., 2017; Del Vicario et al., 2017; Mora-Cantallops et al., 2019; Lopez et al., 2017; Llewellyn and Cram, 2016), but many other political events have been analyzed from a social media perspective (Tumasjan et al., 2010; Sobhani et al., 2017; Darwish et al., 2017; Pierri et al., 2020; Vicario et al., 2017).

A general approach to quantify controversy in social media has been proposed by Garimella et al. (2018), designing a graph-based approach using solely on the underneath social graphs. This approach is language independent, relying solely on the social structure of communities of users, but computational expensive. Another approach has been proposed, that includes the content of texts to make more precise and fast computations (de Zarate et al., 2020).

We investigate this event from a content-based *stance detection* perspective (Küçük and Can, 2020), analyzing only user-generated content to detect the inclination about the referendum in Italy. There are few works about stance detection with non-English tweets (Vamvas and Sennrich, 2020). Lai et al. (2018) collect a similar dataset for the Italian referendum in 2016. They tackle the stance detection task by adding to simple NLP approaches, such as bag of hashtags, bag of mentions or bag of replies, network based features obtained by clustering the retweet/quote/reply networks with Louvain Modularity algorithm. They also analyze the datasets from a diachronic perspective by splitting the time window into four sections based on the dates of referendum-related events. Other works focus on the Italian political situation of Twitter users with content-based approaches (Ramponi et al., 2019, 2020; Di Giovanni et al., 2018). They collect tweets shared by politicians and their followers, and train accurate classifiers that predict the political inclination of users, without considering the social interactions: the content shared contains enough information to successfully perform classification of political inclination.

Similar tasks have been proposed at SemEval 2016 (Mohammad et al., 2016b), IberEval 2017 (Taulé et al., 2017), IberEval 2018 (Taulé et al., 2018) and finally at EVALITA 2020 (Cignarella et al., 2020), where teams were challenged to detect stances of manually labeled Italian Tweets about the Sardine Movement. We remark the difficulty of such tasks by looking at the performance of the best team (Giorgioni et al., 2020), that fine-tuned an Italian pre-trained BERT model (Devlin et al., 2019) and augmented the data with results from three auxiliary tasks.

A comparative study (Ghosh et al., 2019) shows that for stance-detection datasets of English texts from Web and Social Media, BERT model achieves the best performance, but there is still much room for improvements.

iovoto*	parlamentari	iovoto*taglioparlamentari
voto*	vota_efaivotare*	tagliodeiparlamentari
vota*	referendum	referendum2020_iovoto*
votare*	referendum2020	iovoto*_referendum2020
unitiperil*	maratonaperil*	cittadiniperil*

Table 1: List of keywords used to filter relevant tweets. They refer to *vote*, *parliamentarians*, *cuts* and *referendum*. We substitute * with no, si and sì (*yes* in Italian).

3 Data Collection, Description and Labeling

The dataset is collected from **Twitter**[4], a microblogging platform widely used to discuss trending topics, whose official API allows a fast and comprehensive implementation. On Twitter, users share *tweets*, small texts (up to 280 characters) that can

[4]https://twitter.com

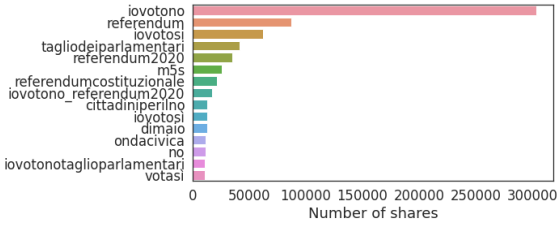

Figure 1: Mostly shared hashtags in the dataset.

be enriched with images, videos or URLs. Other users can *quote* (or *retweet*) another tweet by sharing it with (or without) a personal comment. A user can also *follow* other users to get a notification when they tweet (retweet or quote), and can be followed by other users.

We query data about the referendum held in Italy in September 2020 by searching Italian tweets, containing at least one of the keywords reported in Table 1, usually used as hashtags, but not always. In total we collected $1.2M$ Italian tweets posted between 01/08/2020 and 01/10/2020 by about $111k$ users.

The keywords are refined and validated iteratively. Starting from three keywords (referendum, iovotosì - IVoteYes, iovotono - IVoteNo), we inspect the most frequent hashtags and, if related to the topic, we add them to the query. In Figure 1 we show the most used hashtags in our complete dataset. Many frequent hashtags have no clear and safe connection with the referendum, thus we do not select them as keywords during the collection step, such surnames of politicians ("dimaio") and political parties ("m5s").

3.1 Hashtag-based Semi-automatic Labeling

Manually labeling big data sets is an expensive and not-scalable approach. Usually more than one annotator, fluent in the selected language, is required to produce a reliable label, and the time and cost to obtain a data set large enough to train an accurate classifier is usually high.

Graph-based approaches have obtained impressive results when applied to detect stances in controversial debates (Garimella et al., 2018; Cossard et al., 2020). These approaches are mainly used to label user by looking at the nearest community in the social graph. They firstly define the graph structure, e.g. retweet graph, and then they apply community detection algorithms to partition the bigger connected component of the graph.

We design a content-based approach to semi-automatically label large sets of tweets. Different from the graph-based approaches, we label *single* tweets, while the graph approaches work at the user-level. The approach is based on *hashtags*, often used to express the inclination of users about a topic (Mohammad et al., 2016a). Trending hashtags attract audience and get the attention of other users in the social network[5].

We pick two main classes: in *Support* of the referendum and *Against* the referendum. We define as *Gold hashtags* the hashtags that clearly state a side in the vaccine debate. We plan to collect two sets of Gold hashtags, one for each side of the debate. If a tweet contains at least one of the Gold hashtags, we define its stance as the stance of the hashtag. Tweets containing at least one Gold hashtag from both sides are discarded. Firstly, we select two Gold hashtags, one for each side: #iovotosì (I Vote Yes) for the Support class and #iovotono (I Vote No) for the Against class. Note that in Italian the word *yes* is translated as *sì*, with the grave accent that is often omitted in informal texts, such as tweets. Thus, in the whole paper, every time we refer to the word *sì*, we include also the word *si*, without the accent. Two annotators manually validate this initial selection by inspecting 100 tweets for each class and finding only 4 tweets that clearly belongs to the opposite stance. They were used to attract the attention of the other side or to delegitimise a specific hashtag., e.g. “I cannot understand people that write #IVoteYes". However, our validation process confirms that these tweets are rare and introduce little noise to the data set.

We iteratively add new hashtags by inspecting the most frequent co-occurring ones and manually selecting the most pertinent ones, basing the selection on their meaning. An example of discarded hashtags is #conte (the surname of the Prime Minister of Italy at the time of the Referendum), highly co-occuring with #iovotono, since we cannot safely assume that it was used only by users Against the referendum. We also discard hashtags that co-occur with hashtags from both sides in similar percentages. An example is #referendum, obviously frequently used by both sides of the debate. Finally, after each iteration two annotators manually validate the selected hashtags, as previously described for the initial Gold hashtags. An hashtag passes the validation if the percentage of tweets that is

[5]Twitter has a specific section for trending hashtags and keywords https://twitter.com/explore/tabs/trending

	Tweets using both #IoVotoSì and #IoVotoNo
A	In a few days we will meet at the ballot boxes to express our preference about the #CutOfParliamentarians. While waiting, let's retrace the most famous referendums in the history of the Republic. #Referendum2020 #IVoteYes #IVoteNo
B	Let's dismantle some lies about #IVoteNO. The #CutOfParliamentarians is a reform that fixes the Italian distortion of having a very big number of elected people. Who talks about dictatorship is only using the usual fear strategy to keep a useless privilege. #IVoteYes

Table 2: Translated examples of tweets containing both the Gold hashtag #iovoto and #iovotosì. (A) shows a neutral tweet, (B) shows a Supporter attacking the point of view of people Against the referendum.

classified by at least one annotator as belonging to the opposite class is lower than 10%. We finally obtain two final sets of **Support Gold hashtags** and **Against Gold hashtags**, that allows us to get about $450k$ labeled tweets by manually labeling *few hundreds*. The selected Gold hashtags are the keywords reported in Table 1 that contains the * symbol. The symbol is substituted with the corresponding stance ("sì" or "no"). For example, #referendum2020_iovotono is a Gold hashtag for Against class, while #referendum2020_iovotosì (and #referendum2020_iovotosi) is a Gold hashtag for Support class. Since no other hashtag among the 50 most-frequent ones passes the full validation procedure, we end the labeling phase.

Note that we label tweets containing at least one hashtag from a single set in the corresponding class, while tweets with at least one hashtag from both sets as Both and tweets without any hashtag from both sets as Unknown. We remark that Both and Unknown tweets cannot be safely considered *neutral* since they can express a stance without explicitly using one of the selected hashtags, or using both of them (Table 2 reports an example of a neutral tweet labeled as *Both* (A) and a Support tweet labeled as *Both* (B). This is the main limitation of this semi-automatic labeling procedure: no neutral class can be safely defined, thus we can only train a binary-classifier, leaving for future works the design of a three-classes stance detector.

We label *retweets* by looking at the hashtags in the original tweet, we label *quotes* by only looking at the hashtags in the quote itself, not at the quoted hashtags. In Table 3 we report the statistics of the obtained labeled dataset. Original tweets are tweets that are neither retweets nor quotes of other tweets, nor replies to other tweets.

Label	Tweets	Original	Retweets	Quotes	Replies
Support	93149	74086	2890	10572	5665
Against	364865	291185	15368	34559	24145
Both	4224	2796	145	246	1042
Unknown	353033	236743	16600	53119	47059
Total	815271	604810	35003	98496	77911

Table 3: Tweets Statistics.

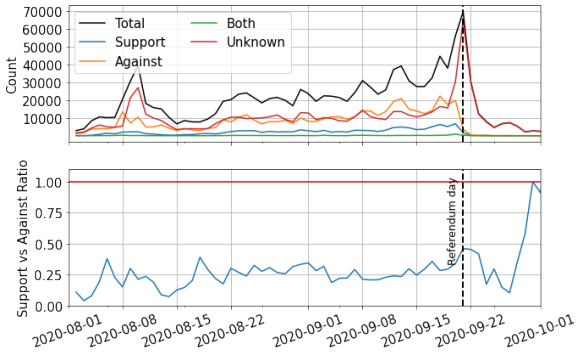

Figure 2: *Top*: Number of daily shared tweets, grouped by stance. *Bottom*: Daily Support vs Against Ratio. The higher the ratio, the greater the number of tweet Against the referendum. The red line (1) sets the value of equal number of Support and Against tweets.

3.2 Temporal Analysis

In Figure 2 (top) we show the distribution of tweets, grouped by their stance, during the time window selected, highlighting the referendum day. We notice a first peak around the August 8, due to an unrelated event about parliamentarians, that we accidentally included, since we used *parliamentarians* as a keyword to filter tweets. To remove noise and unrelated data, we discard all tweets posted before August 15 in the following analyses.

We also notice a huge peak of Unknown tweets during the referendum days, probably because users switched from the old hashtags *#IVoteYes* and *#IVoteNo* to their past tense versions (*#IVotedYes* and *#IVotedNo*). Thus, we discard tweets posted after September 19. Moreover, we do not want to influence our stance classification with tweets posted after the referendum.

In Figure 2 (bottom) we show how the ratio between Support and Against tweets evolves during the time window, observing constant values around 0.25 from August 15 to September 19. Thus, the daily number of tweets Against the referendum is four times bigger than the number of tweets Supporting it, further confirmed in Table 3, where the total number of Support tweets is four times smaller than the total number of tweets Against the referendum. We also notice big peaks and valleys outside

the selected time window, caused by the low number of daily posted tweets.

4 Data Analysis

In this section we describe the cleaning process, the stance classifiers and their results on the collected dataset.

4.1 Data Cleaning

Before training a stance classifier, we clean the text of tweets through the following procedure.

Texts are lowercased, URLs are removed and spaces are standardized. **We remove Gold hashtags** (see Table 1) since they were used to automatically label tweets and users, thus maintaining them will introduce a strong bias in the trained models. We keep the other hashtags since they could encode useful information and are not a clear source of bias. Tweets containing at least half of the characters as hashtags are also removed, since they are too noisy. They are usually used by bots to collect the daily trending hashtags. To prevent overfitting we remove duplicate texts, including retweets. We also remove texts shorter than 20 characters, that usually comment URLs or other tweets, being difficult to understand and contextualize. We keep emoji as they include useful information, e.g., the scissor emoji was mainly used by Supporters of the referendum since they want to *cut* the number of parliamentarians. We select only tweets shared after 15/08/2020 and before 20/09/2020, the first referendum day.

4.2 Stance classification

We analyze the dataset from a stance classification perspective.

Due to the impossibility to interpret the tweets labeled as Both or Unknown, we formulate the tweet stance classification task as a binary classification problem: the two classes represent tweets Supporting or Against the referendum. We obtain an unbalanced clean datasets: $85k$ tweets, of which 80% Against the referendum. To obtain a balanced dataset, over-sampling the *Support* class leads to slightly better results in the Validation dataset, but worse results on the Test set, probably due to overfitting, while under-sampling the *Against* class leads to worse results due to the removal of 60% of the original dataset.

We select three models (one baseline and two commonly used architectures):

	Validation			Test		
Model	AUROC	$F1_w$	$F1_s$	AUROC	$F1_w$	$F1_s$
Baseline	0.50	0.78	0	0.50	0.52	0
FastText	0.74	**0.89**	0.56	0.65	0.59	0.18
BERT	**0.88**	0.86	**0.63**	**0.78**	**0.71**	**0.5**

Table 4: Area under ROC (AUROC), weighted F1 score ($F1_w$) and F1 score of the Supporters ($F1_s$) of the three models, as 5-fold Cross Validation on the training set (left) and on the Test Sets of 227 randomly selected and manually evaluated texts.

- Majority classifier (Baseline);
- FastText (Joulin et al., 2017), a fast approach widely used for text classification. Its architecture is similar to the CBOW model in Word2Vec (Mikolov et al., 2013): a look-up table of words is used to generate word representations, that are averaged and fed into a linear classifier. A softmax function is used to compute the probability distribution over the classes. To include the local order of words, n-grams are used as additional features, with the *hashing trick* to keep the approach fast and memory efficient. FastText is known to reach performances on par with some deep learning methods, while being much faster;
- BERT (Devlin et al., 2019), a Transformer-based model (Vaswani et al., 2017) that reaches state-of-the-art performances on many heterogeneous benchmark tasks. The model is pre-trained on large corpora of unsupervised texts using two self-supervised techniques: Masked Language Models (MLM) task and Next Sentence Prediction (NSP) task. Pre-trained weights are available on the Huggingface models repository (Wolf et al., 2020). We select a model pre-trained on a concatenation of Italian Wikipedia texts, OPUS corpora (Tiedemann, 2012) and OSCAR corpus (Ortiz Suárez et al., 2019), performed by MDZ Digital Library[6]. We fine-tune the model on our data[7].

4.3 Results

In Table 4 (left) we report the results of a 5-fold cross validation process. We select Area Under

[6] https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased

[7] Fine-tuning performed on a single NVIDIA Tesla P100, for 5 epochs. Best weights selected by minimizing the evaluation loss. Learning rate (10^{-5}) set through grid search.

the ROC curve (Fawcett, 2006), weighted F1-score (the F1 score for the classes are weighted by the support, i.e., the number of true instances for each class) and $F1_s$, the F1 score on the Support class (the under-represented class, that, by definition, a Majority classifier cannot detect).

Both FastText model and BERT outperform the Random Baseline approach, the latter obtaining higher AUROC and $F1_s$.

However, our goal is to predict the stance of tweets that do *not* share a Gold Hashtag. We use these models, trained on the big dataset labeled using Gold hashtags, to predict tweets that do not contain Gold Hashtags, thus tweets that, with the previously described automatic approach, were labeled as Unknown. Two human annotators manually labeled 500 randomly sampled tweets. After removing neutral and incomprehensible texts, we obtain a dataset of 227 tweets, of which 78 labeled as Supporters. We test our models on this dataset, the results are reported in Table 4 (right), confirming that even if there is a gap among the Validation performances and the Test performances, BERT did not strongly overfit the Training data.

Finally, we obtain an approximate statistic of the total number of tweets Supporting and Against the referendum by predicting the stance of every tweet previously labeled as Unknown ($110k$ tweets). It results in about 20% of Unknown tweets classified as Supporters, confirming the general number of tweets Against the referendum is four times bigger that the number of shared tweets Supporting it. However, we cannot validate this result since we do not have manually labeled the full dataset.

5 Biases analysis

In this section we inspect three common biases that often affect the accuracies of classifiers: Length of texts, Lexicon and Sentiment.

5.1 Length Analysis

The length of sentences, defined as the number of characters or tokens, often influences the prediction of a model, acting as a bias. In Figure 3 we plot the distribution of lengths of tweets calculated as the number of characters, after the cleaning procedure (there are no tweets shorter than 20 characters). There is no evident difference between the distribution of the number of characters in tweets labeled as Support or Against, suggesting that no length-bias is present in our dataset.

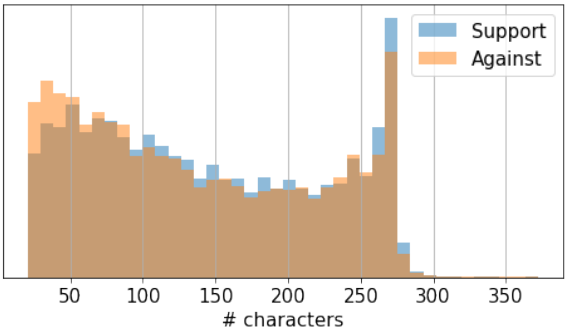

Figure 3: Length distribution of generated tweets grouped by stance. There is no significant difference in the normalized distributions.

5.2 Lexicon analysis

We check if tweets in different stances use similar lexicons. A big lexicon overlap in the dataset results in an accurate classifier that must learn the *meaning* of sentences, while a small lexicon overlap in the dataset allows the detection of specific words to be sufficient to make a prediction, neglecting the real meaning of the texts. We quantify the lexicon difference by computing the Pointwise Mutual Information (PMI) between words and classes (Gururangan et al., 2018).

A high PMI score of a word in a class is obtained when the word is used mainly in tweets belonging to that class. For this analysis, we discard Italian stop words collected from the NLTK library (Bird et al., 2009).

We report in Table 5 the first five words for each class, sorted by PMI score and the proportion of texts in each class containing each word. The frequency of words with higher PMI is low, thus we conclude that the two stances use mostly similar lexicons. A classifier cannot safely rely on the presence of specific words since the most indicative ones (higher PMI score) are not frequent enough. For example, the most frequent word among the top-5 is orgoglio5stelle, a keyword used by Supporters of the Referendum stating that they are proud of their party (5 stars) because the referendum was held by them. However, only 3% of the Supporter texts include this word.

5.3 Sentiment analysis

We distinguish between sentiment classification and stance classification by searching for a correlation between sentiment and stance in the datasets. Our goal is to have a stance classifier that does not

Support	%	Against	%
orgoglio5stelle	3.0	ondacivica	2.2
scissors emoji	0.3	30giorni_iovotono	0.5
laricchiapresidente	0.9	iostoconsalvini	0.5
pugliafutura	0.5	noino	0.4
rotolidistampaigienica	0.3	darevocealreferendum	0.4

Table 5: Top 5 tokens ranked by PMI (Pointwise Mutual Information) scores and the proportion of texts in each class containing each word.

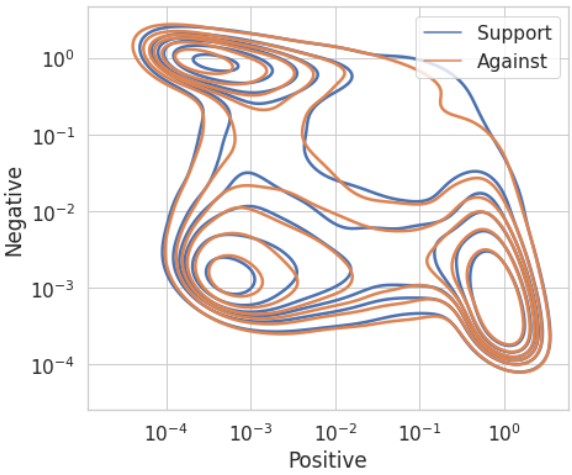

Figure 4: Sentiment distribution of generated tweets grouped by stance. There is no evident difference in the distributions. To improve the visualization, we use the same number of data points for both stances, down-sampling the texts Against the referendum.

rely on the sentiment of tweets to make a prediction. If Support and Against tweets are unbalanced in the Positive and Negative sentiment classes, the dataset contains a sentiment-bias.

We compute the sentiment scores of tweets and users using Neuraly's "Bert-italian-cased-sentiment" model[8] hosted by Huggingface (Wolf et al., 2019). It is a BERT base model trained from an instance of "bert-base-italian-cased"[9] and fine-tuned on an Italian dataset of $45k$ tweets on a 3-classes sentiment analysis task (negative, neutral and positive) from SENTIPOLC task at EVALITA 2016 (Barbieri et al., 2016), obtaining 82% test accuracy.

In Figure 4 we show the Kernel Density Estimation plot of positive and negative sentiment of tweets grouped by stance. The probability of being neutral is not shown as it can be obtained with $1 - p('positive') - p('negative')$. Since the distributions of the sentiments largely overlap, we conclude that there is no sentiment-bias in our datasets. It is further confirmed by looking at the actual predictions: for both Support and Against texts, 63% of them are classified as Negative, 25% as Neutral and 15% as Positive .

6 Discussion

6.1 Discrepancy between Twitter activity and the Referendum outcome

We notice a huge discrepancy between what users posted on Twitter and what citizens voted. The fraction of tweets and users that explicitly state their stance (and our prediction of tweets and users that do not) is very different from the final outcome of the referendum (69.96% of the voters approved it): the number of tweets with a Gold Hashtag Against the referendum is 4 times higher than the number of tweets with a Supporter Gold Hashtag, and the number of Unknown tweets that our best classifier predicts as Support or Against the referendum follows the same proportion. By looking only at what is shared online, we could have easily guessed that the Opposers won the referendum, while the real outcome is the opposite.

To further understand this discrepancy, we briefly inspect the differences in social characteristics of users. We label users as Support (Against) if they share only tweets previously labeled as Support (Against) the referendum. Figure 5 shows the normalized distribution of number of *followers* and number of *following* of users Supporting and Against the referendum. No difference in shape proves that the social audience of the two sides of users is quantitatively similar (the tails of the figures are cut for visualization purposes). Inspecting the most followed and following users (long tail of the distribution), we notice that among the top-10, exactly half of them are Supporters and half are Against the referendum, confirming our finding. Thus we conclude that Supporters won the referendum, not because they tweeted more than Opposers (they actually tweeted 4 times less than the people against the referendum), neither because they have more audience (the distributions of number of followers and following people is similar). We leave for future works the inspection of more detailed graph-related quantities, such as centrality of users in the network and topological measures to describe the graph structure.

[8] https://huggingface.co/neuraly/bert-base-italian-cased-sentiment

[9] https://huggingface.co/dbmdz/bert-base-italian-cased

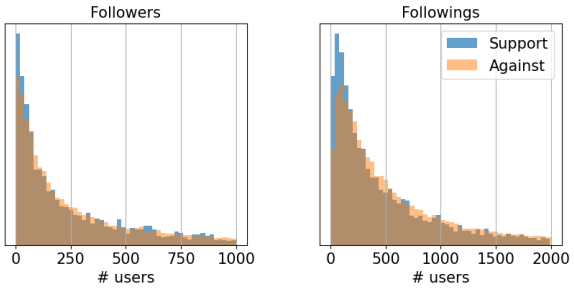

Figure 5: Distribution of followers (left) and following (right) users of users Supporting and Against the referendum.

We observed an event where the majority of voters were silent, or not even present on Social Media, while the minority was loud. This phenomenon implies not only that restricting the focus on social media to fully analyze an event could lead to extremely wrong forecasts, but also that the user perception of the general political situation can be influenced by an unrealistic image of the public opinion on social media that does not match the real sentiment towards the topic.

6.2 Ethical Considerations

Political inclinations of people is a sensitive topic. This work is meant to be a exploration on how to apply state-of-the-art NLP techniques to predict the stance of tweets about a political event, and whether they can help to perform more accurate forecasts of the outcome of a political event. Due to privacy issues, we do not share the trained model nor the obtained labels of tweets. However, we share the dehydrated collected tweets and the set of keywords to obtain the gold labels. These data allow researchers to reproduce the results but do not contain sensitive information, meeting the Twitter's Terms of Service[10]. In this study we prove that the political inclination of users can be detected by modern NLP approaches, *even if no evident hashtags of keywords are shared in a tweet.* Thus, we suggest a thoughtful and appropriate usage of social networks in order to keep private sensitive information.

7 Conclusion

Thanks to the last referendum in Italy, we collected a big Italian stance detection user-generated dataset. The dataset consists in $1.2M$ tweets, of which $85k$ are cleaned and labeled as Supporters or Against the referendum. The designed hashtag-based semi-automatic labeling approach allows us to train an accurate classifier that generalizes well also on tweets that do not contain Gold hashtags. We considered three common dataset biases (length-bias, lexicon-bias and sentiment-bias), confirming no significant dangers. Finally, we investigated the discrepancy between the fraction of collected tweets labeled by stance and the real outcome of the referendum, observing no clues that explain this difference. Based on our findings, we suggest that drawing conclusions following social media analysis should be performed carefully, and the results should be integrated with other other classical approaches such as surveys.

In future works, we aim to build a three-classes stance classifier, that can also predict *neutral* texts, since we observed big magnitudes of data that does not explicitly state a stance. We will also move the focus from tweets to users, detecting their inclination by looking at the history of shared tweets. We believe that the investigation of users that *changed* stance during the time window could help us understand how people opinions are influenced by social media. Finally, we observe that our classifier do not generalize well on other Italian stance-detection data sets, due to the high specificity of the task: the model learned the debate about the 2020 Italian constitutional referendum and its actors' inclination, but the knowledge obtained is not adequate to perform zero-shot transfer to other data sets. However, we plan to investigate if we can obtain boosts of performances in a multi-task and multi-source context, training a model on multiple similar tasks and data at the same time.

[10] https://twitter.com/en/privacy

References

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. *Overview of the Evalita 2016 SENTIment POLarity Classification Task*, pages 146–155.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

Mario Callegaro and Yongwei Yang. 2018. *The Role of Surveys in the Era of "Big Data"*, pages 175–192. Springer International Publishing, Cham.

Alessandra Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Sardistance @ evalita2020: Overview of the task on stance detection in italian tweets.

Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. 2020. Falling into the echo chamber: The italian vaccination debate on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):130–140.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Trump vs. hillary: What went viral during the 2016 us presidential election. In *Social Informatics*, pages 143–161, Cham. Springer International Publishing.

Juan Manuel Ortiz de Zarate, Marco Di Giovanni, Esteban Zindel Feuerstein, and Marco Brambilla. 2020. Measuring controversy in social networks through nlp. In *String Processing and Information Retrieval*, pages 194–209, Cham. Springer International Publishing.

Michela Del Vicario, Fabiana Zollo, Guido Caldarelli, Antonio Scala, and Walter Quattrociocchi. 2017. Mapping social dynamics on facebook: The brexit debate. *Social Networks*, 50:6–16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

M. Di Giovanni, M. Brambilla, S. Ceri, F. Daniel, and G. Ramponi. 2018. Content-based classification of political inclinations of twitter users. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4321–4327.

Tom Fawcett. 2006. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874.

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM*, 59(7):96–104.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *Trans. Soc. Comput.*, 1(1).

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: A comparative study. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 75–87, Cham. Springer International Publishing.

Simone Giorgioni, Marcello Politi, Samir Salman, R. Basili, and Danilo Croce. 2020. Unitor @ sardistance2020: Combining transformer-based architectures and transfer learning for robust stance detection. In *EVALITA*.

Miha Grčar, Darko Cherepnalkoski, Igor Mozetič, and Petra Kralj Novak. 2017. Stance and influence of twitter users regarding the brexit referendum. *Computational Social Networks*, 4.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Short Papers*, NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, pages 107–112. Association for Computational Linguistics (ACL).

Philip N. Howard and Bence Kollanyi. 2016. Bots, #strongerin, and #brexit: Computational propaganda during the uk-eu referendum.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. In *Natural Language Processing and Information Systems*, pages 15–27, Cham. Springer International Publishing.

C. Llewellyn and L. Cram. 2016. Brexit? analyzing opinion on the uk-eu referendum within twitter. In *ICWSM*.

Julio Lopez, Sofia Collignon-Delmar, Kenneth Benoit, and Akitaka Matsuo. 2017. Predicting the brexit vote by tracking and classifying public opinion using twitter data. *Statistics, Politics and Policy*, 8.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and (James) Rosenquist. 2011. Understanding the demographics of twitter users. volume 11.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Marçal Mora-Cantallops, Salvador Sánchez-Alonso, and Anna Visvizi. 2019. The influence of external political events on social networks: the case of the brexit twitter network. *Journal of Ambient Intelligence and Humanized Computing*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Francesco Pierri, Alessandro Artoni, and Stefano Ceri. 2020. Investigating italian disinformation spreading on twitter in the context of 2019 european elections. *PLOS ONE*, 15(1):1–23.

Giorgia Ramponi, Marco Brambilla, Stefano Ceri, Florian Daniel, and Marco Di Giovanni. 2019. Vocabulary-based community detection and characterization. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 1043–1050, New York, NY, USA. Association for Computing Machinery.

Giorgia Ramponi, Marco Brambilla, Stefano Ceri, Florian Daniel, and Marco Di Giovanni. 2020. Content-based characterization of online social communities. *Information Processing & Management*, 57(6):102133.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

M. Taulé, M. Martí, Francisco M. Rangel Pardo, P. Rosso, C. Bosco, and V. Patti. 2017. Overview of the task on stance and gender detection in tweets on catalan independence. In *IberEval@SEPLN*.

M. Taulé, Francisco M. Rangel Pardo, M. Martí, and P. Rosso. 2018. Overview of the task on multimodal stance detection in tweets on catalan #1oct referendum. In *IberEval@SEPLN*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. volume 10.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

M. D. Vicario, S. Gaito, W. Quattrociocchi, M. Zignani, and F. Zollo. 2017. News consumption during the italian referendum: A cross-platform analysis on facebook and twitter. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 648–657.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.