Online Monitoring of Stance from Tweets: The case of Green Pass in Italy

Alessandro Bondielli[†], Giuseppe Cancello Tortora [*], Pietro Ducange[*], Armando Macri[*],
Francesco Marcelloni[*], Alessandro Renda[*]
[*] Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino 1, 56122 Pisa, Italy
[†] Department of Computer Science, University of Pisa, Largo B. Pontecorvo, 3, 56127, Pisa Italy
Email: {alessandro.bondielli, pietro.ducange, francesco.marcelloni, alessandro.renda}@unipi.it
{g.cancellotortora, a.macri1}@studenti.unipi.it

***Abstract*—Stance detection on social media has attracted a lot of attention in the last few years, as opinionated posts are an invaluable source of information which can possibly be exploited in dedicated systems. This is especially true in the case of particularly polarizing topics for which there is no clear consensus among population. In this paper, we focus on one of these topics, namely the EU digital COVID certificate (also known as Green Pass), with the objective of uncovering the stance towards it in a specific time period for the Italian Twitter community. To this aim, we first tested some classifiers for determining the most suitable one in terms of performance and complexity for the stance detection problem under consideration. Then, we compared several approaches aimed at counteracting the occurrence of concept drift, i.e., that phenomenon for which the characteristics of the dataset vary over time, possibly resulting in a degradation of classification accuracy. Our experimental analysis suggests that updating the classifier during the stance monitoring campaign is crucial for maintaining a satisfactory level of performance. Finally, we deployed our system to monitor the stance on the topic of Green Pass expressed in tweets published from July to December 2021 and to obtain insights about its evolution.**

***Index Terms*—stance detection, concept drift, text classification, Green Pass, Covid-19**

I. Introduction

Social media have become one of the most widespread tools to express opinions and ideas. Their ease of access and use provides a good platform for discussing current events, both at the global and local scales. Thus, they also provide an invaluable source of information to enable the understanding of how specific topics are perceived by large audiences by means of techniques in the fields of opinion mining and stance detection.

A growing amount of work is dedicated especially to the stance detection task. This task can be formulated as follows: given a piece of text (e.g., a social media post) concerning a pre-determined target (e.g., a topic or a concept), classify the text with one of three possible labels in the set {In Favor, Against, Neither} [1]. In other words, decide whether the piece of text expresses a "positive", "negative" or "neutral" stance towards the target, even if the target is not explicitly mentioned in the text [2]. For example, given the topic of mandatory vaccination, the goal is to decide whether the piece of text is in favour of, against, neither in favour of nor against mandatory vaccines (i.e., it has a positive stance towards mandatory vaccination). Stance detection has proven to be both an interesting task by itself and a helpful resource for other social media-related problems such as fake news and rumour detection.

An interesting aspect to take into account is that stance towards specific topics may often be influenced by real world events concerning that topic [3], [4]. This is especially true in the case of particularly controversial topics or concepts for which it is hard to identify a general consensus in opinion among the population. It is clear that, in a setting where (i) there is a continuous production of new data and (ii) real world events have the potential to shift the opinion and change the discourse around the topic, the distribution of data may be subject to *concept drift* [5]. In such a setting, training a classifier with an initial set of data and subsequently applying it to new data may yield sub-optimal performances (e.g., the classifier performance may deteriorate over time). Thus, appropriate strategies suited to deal with concept drift should be adopted for maintaining a good level of classification performance along the time.

In this paper, we propose a pipeline for retrieving tweets for a specific target and classifying them for stance. We focus specifically on the adoption of EU digital COVID certificate (also known as Green Pass). The certificate attests one of the following conditions: having had the COVID-19 vaccination, having tested negative for COVID-19 in the last 48 hours, or having had COVID-19 in the last six months. The introduction of the certificate has sparked various reactions in the Italian population, including strikes. A debate has also arisen on social media, with Twitter being the forum of conversation related to COVID-19, where opinions supporting or opposing the government measures can be found.

The goal of our work was twofold. On the one hand, we intended to uncover the opinion, expressed by users on Twitter, about the adoption of the digital certificate, and its evolution through the different stages of implementation during the pandemic. To this aim, we designed a stance classification system, exploring the performance of different

This work was supported in part by the Italian Ministry of University and Research (MUR) in the framework of the CrossLab Project (Departments of Excellence).

classification algorithms for the task, namely *Support Vector Machine* (SVM), *Logistic Regression* (LR) and *Complement Naive Bayes* (ComplementNB). We compared these models both in terms of classification performance (i.e. using accuracy and F1-score as metrics) and in terms of their complexity (i.e. analyzing the number of features and the training computational cost), in order to address the potentially high volume of data to be processed over time. On the other hand, we aimed to evaluate different methods for stance detection in an evolving setting, i.e., in the case of potential concept drift. In particular, we compared three different training and model updating approaches versus the classical approach in which a classification model is trained with an initial set of labelled data and never updated along the time. We show that the latter approach is the one which achieves the worst classification performances, even when an ensemble model is adopted.

We employed the developed system to monitor the stance on the Green Pass topic expressed in tweets published from July to December 2021. We discuss the insights mined through this analysis.

The rest of the paper is organized as follows. In Section II we present related works in the areas of stance detection and techniques to address concept drift. Section III describes the collection of twitter data that we used in the experiments and the design of the stance classification system. In particular, the evaluation of three different classifiers for the task is presented. In Section IV we analyze the different approaches for dealing with concept drift, and provide experimental results on the collected Green Pass dataset. Section V provides an analysis of the evolution of the stance towards the EU digital COVID certificate in Italy, obtained by applying the best classification method to our dataset. Finally, Section VI draws some conclusions and proposes future works on the topic.

II. Related works

The problem of stance detection can be still considered as a challenge within the broader spectrum of NLP tasks. This is especially true if we take into account the fact that its most interesting applications are on social media such as Twitter. In social media, users typically express themselves and their opinions with a non-standard use of words, including abbreviations, colloquial expressions, grammatical errors, misspelling and neologisms. Moreover, it is often the case that irony and sarcasm are an integral part of many social media posts. All of these aspects render the detection of a precise stance for a piece of social media text even more of a challenge.

These issues have limited the research on stance detection with respects to other long-standing NLP tasks such as sentiment analysis. For example, authors of [6] describe the outcomes of a shared task for stance detection where F1 score values ranged from 46.19 to 67.82 among 19 different participant teams. In [2] a stance detection system is described, using a Linear SVM classifier incorporating both features obtained from the training data as well as external resources. The system obtained an overall F1-score of 70.3. In more recent works, the adoption of deep learning was proposed to tackle the problem of stance detection. In [7] a neural model based on LSTM and attention achieved a F1-score of 68.8 and an accuracy of 60.2. In [8] authors tried to tackle the problem by leveraging text representations learned for sentiment analysis tasks. However, the results does not show any significant improvement over other approaches. Moreover, it is worth to observe that no specific benchmark datasets are available and used. In fact, works such as [9] report similar best case performances with different datasets in Chinese.

When considering social media analysis, and especially in the case of continuous streams of data such as social media posts or news shared on social media, it is important to take into account also the problem of concept drift. Nevertheless, it interesting to point out different aspects of concept drift that may influence the implementation of automated strategies to face a specific task. Four different types of concept drift have been identified in the literature [5], namely *sudden*, *gradual*, *incremental* and *reoccurring*. As for the sudden concept drift, we refer to a scenario in which the distribution of values, and thus of concepts, suddenly shifts with a drastic difference between two subsequent time instants. A similar situation can be identified with incremental drift. In this case, however, the shift from a concept to another one is slower, and includes several different intermediate concepts. In gradual concept drift, it is possible to identify a transition period in which both concepts are active and valid at the same time. Finally, in reoccurring concept drift, we identify a shift between for example two concepts that is however reverted over time.

Several different solutions have been explored in the literature to face the problem of concept drift, that can broadly be distinguished into three main categories. First, we can consider *time-window* approaches. In this case, models are typically re-learned for new chunks of data (e.g., the tweets in a specific time window) and a forgetting mechanism of some kind is implemented to reduce or reset the weight of older data [10]–[12]. Second, *incremental* approaches typically leverage strategies for incremental training in order to cope with the generation of new data. Incremental training strategies include for example re-training the model from scratch with each chunk of new data [10], active learning [13], and methods based on deep learning models [14]. Third, several *ensemble* models have been proposed in the literature, providing good results [10]. In this case, an ensemble of different classifiers is used to provide the final label. It is interesting to point out that several works successfully attempted to merge ensemble and time-window strategies [15]–[17]. However, the most suitable approach often depends on the specific application and, to the best of our knowledge, the problem of detecting stance towards Green Pass from tweets has not been investigated so far in the literature.

III. Design of the stance classification system

In this section, we describe the intelligent system for stance detection on Twitter. We formulate the problem as a three class classification task, with the goal of labelling each tweet with *Positive*, *Negative* or *Neutral* stance towards the Green Pass.

The design of the system is inspired by one of our recent works [18], and encompasses the following steps:

1) data collection, cleaning and preprocessing;
2) selection of the best classification model.

In the following, we elaborate on these steps with reference to the case study of stance detection towards Green Pass.

A. Data collection, cleaning and preprocessing

Pursuing a good coverage in collecting relevant tweets, we used two variants of the target keyword, namely "greenpass" and "green pass". We focused on the Italian setting, filtering tweets by language, and on a specific time window, i.e. from July 1, 2021 to December 15, 2021. This period includes the day of enforcement of the green pass in Italy (August 6) and a number of other political and social events that have sparked the online debate.

Overall, we collected 1,068,130 tweets: however, we opted for removing duplicates and messages with content expressed in a language different from Italian. At the end of this data cleaning procedure our dataset comprises 482,688 tweets. Their distribution over the considered time window is reported in Fig. 1. The trend suggests that the volume of tweets is remarkably high: starting from mid-July, there were on average more than 2000 tweets per day on the topic. Furthermore, the presence of several spikes in the plot denotes that the discussion on social networks is fueled by the occurrence of some socio-political events and considerably varies over time.

It is worth underlining that raw tweets scraped from Twitter generally result in a noisy dataset. Tweets, in fact, have certain peculiarities compared to other text corpora, including hashtags, emoticons, user mentions, to name but a few. Therefore, raw twitter data must be appropriately normalized in order to be effectively harnessed in an ML pipeline. We have applied the following preprocessing steps:

- converting tweets to lowercase;
- removing user mentions, URLs, emoticons and retweets;
- removing punctuation marks, brackets, quotes, special characters;
- replacing two or more spaces with a single whitespace.

B. Classification model selection

We opted for tackling the stance analysis as a supervised learning task: we reserved an initial portion of the tweets, sampled between July 1 and July 22 2021, to carry out an experimental comparison of three classification models and to select the best one. Once the best model was identified, we deployed it for uncovering user stance over the rest of the monitoring campaign. Such two phases are highlighted with different colors in Fig. 1.

The classification stage relies on a preliminary step of numerical representation of text: leveraging information from the related literature [2], [4], [18], a Bag of Words model has been adopted. Specifically, we performed tokenization with unigrams (n-grams with n=1), stop-word filtering, stemming, and vectorization with TF-IDF statistic as the weighting scheme.

As per the model selection, we manually labelled 2435 tweets to populate an initial labeled dataset; we have chosed the tweets in such a way to generate a balanced dataset (*Positive*: 828, *Negative*: 800, *Neutral*: 807) and to avoid a bias towards a specific class. Then, we assessed the performance of different classification models by using a 10-fold stratified cross-validation. In the empirical comparison we have included the following state-of-the-art classifiers: support vector machine (SVM), logistic regression (LogisticRegression) and complement Naive Bayes (ComplementNB). We resort to the Python implementation of the models available in scikit-learn[1].

Results of the 10-fold cross-validation are summarized in Table I and are evaluated in terms of Precision, Recall and F1-score by class. Furthermore, we report the overall accuracy and the macro-average F1 over the positive and the negative classes (F_1^{PN}). Such a metric is often used in the NLP field for problems where the identification of polarized classes is considered more relevant than the neutral one [19]. Hence the neutral class is excluded from calculation of the macro-average F1.

Performance figures in Table I are in line with those reported in the stance detection literature [1], [2] and mentioned in Section II. Indeed, an accuracy between 60% and 65% can be considered satisfactory. Notably, the recall on the neutral class is relatively low compared to the other classes, probably because it is easier to find characterizing patterns for polarized classes.

Overall, the three models achieve comparable performance; for the purpose of the downstream analysis, we selected the ComplementNB classifier. First, it ensures slightly higher performance than SVM and LogisticRegression; second, differently from other models, it supports online learning: whenever a new chunk of labelled tweets is available, it can be used for incrementally update the model parameters.

IV. Concept drift analysis

Concept drift can be formally defined as follows [5]. Given two timestamps t_0 and t_1:

$$\exists \mathbf{X} : p_{t_0}(\mathbf{X}, y) \neq p_{t_1}(\mathbf{X}, y) \qquad (1)$$

where p_{t_i} is the joint probability distribution at t_i between the set of input variables $\mathbf{X}$ and the target variable y.

As pointed out in Section I, the occurrence of socio-political events pertaining to the monitored topic may alter the volume and the characteristics of collected tweets. Therefore, as concept drift may affect the dataset, it is crucial to adopt proper countermeasures to prevent potential performance degradation that undermines the reliability of the classification system.

In this work, we assume that a collection of tweets can be also manually labelled by a user throughout the campaign. The availability of new labelled data allows monitoring the performance of the classification model and possibly updating it over time. However, since labeling each single tweet is impractical, we resort to a strategy for labeling a representative

[1] https://scikit-learn.org/stable/

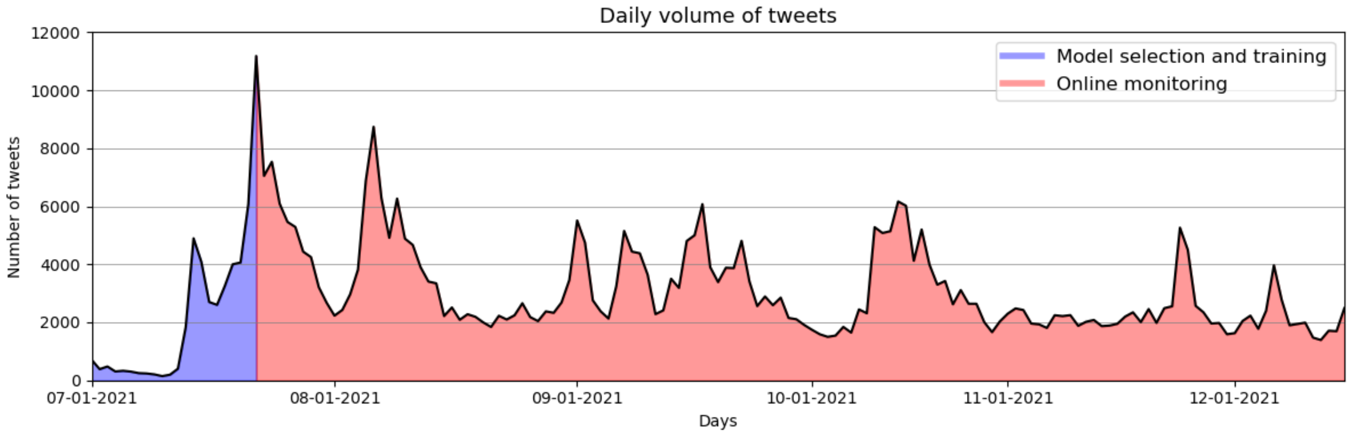


Fig. 1. Daily number of collected tweets from July 1, 2021 to December 15, 2021.

TABLE I
RESULTS OBTAINED BY THE THREE CLASSIFICATION MODELS EMPLOYED IN THE COMPARISON APPLYING A 10-FOLD STRATIFIED CROSS-VALIDATION.

Classifier	**Class**	**Precision**	**Recall**	**F_1-score**	**F_1^{PN}**	**Accuracy** (%)
ComplementNB	Neutral	0.6447	0.5328	0.5834	0.6745	64.76
	Positive	0.6035	0.7850	0.6824		
	Negative	0.7192	0.6212	0.6667		
LogisticRegression	Neutral	0.5928	0.5700	0.5812	0.6676	63.94
	Positive	0.6494	0.6800	0.6643		
	Negative	0.6742	0.6675	0.6709		
SVM	Neutral	0.5835	0.6059	0.5945	0.6626	63.94
	Positive	0.6542	0.6534	0.6538		
	Negative	0.6844	0.6587	0.6713		

set of tweets. Specifically, we considered a buffer of 200 elements: in a specific time window (around 2/3 weeks in our analysis) 200 tweets were randomly selected and inserted into the buffer along with the label assigned by the deployed classification model. It is worth to notice that the actual labeling process of the tweets in the buffer consists in checking the correctness of the estimated class associated with them. Whenever the buffer is full and re-labeled, the chunk of tweets is first used for testing the classification system performance and then, possibly, to update it.

Besides the 2435 tweets of the initial labelled dataset, we manually annotated an additional 1600 tweets during online monitoring from July 22, 2021 to December 15, 2021. Even though we tried to maintain a uniform distribution of the estimated classes in the buffer, the final proportion may vary upon the manual re-labeling. Table II shows the class distribution of each chunk of tweets and highlights that it is actually imbalanced in some chunks. Notably, the selected model (i.e., ComplementNB) has proved effective in dealing with imbalanced text classification [20]; on the other hand, textual data often is not actually generated according to a multinomial model (i.e., each word in a document is independent of every other), as assumed in the traditional Multinomial Naive Bayes model.

TABLE II
CLASS DISTRIBUTION OF CHUNKS OF TWEETS AFTER MANUAL LABELLING.

	Positive	**Negative**	**Neutral**
Chunk 1	101	79	20
Chunk 2	41	56	103
Chunk 3	41	64	95
Chunk 4	81	63	56
Chunk 5	75	68	57
Chunk 6	57	66	77
Chunk 7	65	69	66
Chunk 8	72	74	54

In this section, we first describe the considered approaches for updating the classification system in an online fashion based on chunks of tweets, and then we report the experimental results within the case study of the stance detection towards the Green Pass.

A. Approaches for concept drift adaptation

We compared the performance of five different approaches for reacting to concept drift in the classification of a stream of tweets. The approaches are described in the following:

- **static**: the classification system trained on the original set of 2435 labelled tweets is used to classify the whole stream of tweets; as the model is not updated, the static approach overlooks the concept drift adaptation problem and serves just as a baseline;
- **retrain**: at each available chunk of tweets, the training set is extended with new labelled data and the whole classification pipeline (i.e., TF-IDF and ComplementNB) is retrained from scratch;
- **sliding**: at each available chunk of tweets, the training set is updated adding the new chunk and removing an equivalent amount of less recent tweets. Then, the whole classification pipeline (i.e., TF-IDF and ComplementNB) is retrained from scratch as in the *retrain* approach. In other words, the sliding window model assumes the more recent instances are more important than the less recent ones, which in turn can be disregarded;
- **incremental**: at each available chunk of tweets, the classification model (ComplementNB) is updated based on the new labelled tweets with a *partial fitting*; unlike the *retrain* approach, the old model is not replaced with a new one trained from scratch, but its parameters, namely its internal statistics, are updated considering the new data distribution. Furthermore, only the classifier is updated whereas the attribute space, namely the vocabulary, generated based on the initial TF-IDF vectorization, is left unchanged throughout the online monitoring;
- **ensemble**: the prediction of the three *static* classifiers (SVM, LogisticRegression, ComplementNB) are combined based on a majority voting policy. This approach does not involve updating the model, but aims to assess whether using a composite classifier brings benefits in reacting to concept drift.

B. Experimental results

We evaluate the performance of each approach on the sequence of labelled chunks of tweets. The trends of F_1^{PN} and Accuracy are reported in Figs. 2a and 2b, respectively.

First of all, it is worth highlighting that plots in Figs. 2a and 2b show coherent patterns: given an approach, the measured F_1^{PN} and accuracy have similar trends; furthermore, the ranking of the approaches at every chunk generally does not depend on the considered metric. Results are also coherent with those reported in Table I within the cross-validation analysis, even if a strong variability across chunks can be observed: for instance, the fifth chunk of tweets is properly classified by all approaches, with an accuracy around 75%, whereas the eighth chunk shows a sharp degradation in performance, with an accuracy lower than 60%. Such a behaviour suggests that the stream of tweets is likely affected by the occurrence of concept drift; the increased complexity of some chunks, that hinders achieving high levels of accuracy, may also be induced by the tweet selection and annotation procedure. However, it should be noticed, similarly to previous comparative analysis [4], that we are primarily interested in investigating the relative performances among the different approaches, rather than the absolute ones.

Evidently, all approaches achieve exactly the same values of the metrics in the first chunk, as they were all trained on the same initial training set. The only exception is represented by the ensemble approach in which the performance of the same base model (ComplementNB) is combined with those of SVM and LogisticRegression. Interestingly, in the case of the first chunk, leveraging knowledge extracted from multiple models does not bring benefit in recognizing the stance expressed in tweets.

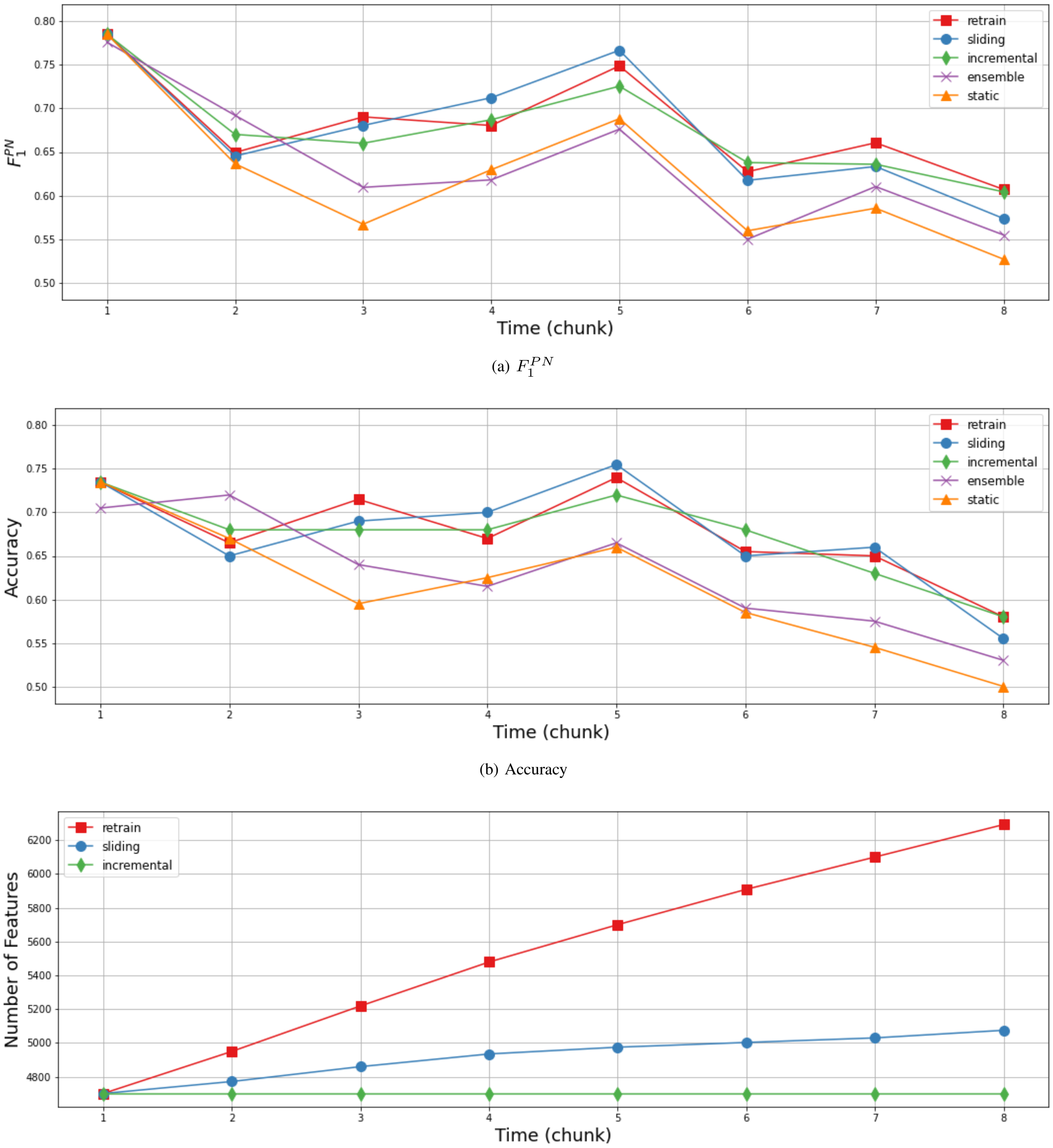
The *static* approach leads, in general, to the worst performance: F_1^{PN} and accuracy vary considerably along the monitoring campaign, with up to 25% discrepancy between the first and last chunks, and the approach is systematically below the others by about 5-10 percentage points (see Figs. 2a and 2b). A severe degradation of performance can be observed in the final part of the considered time window: this result confirms that relying on the initial classification model is inadequate for our long-term monitoring campaign and a proper strategy for counteracting concept drift needs to be adopted.

Notably, the *ensemble* approach falls also short of generalization capability in our evolving setting. Although performance sometimes improves compared to the *static* approach based on a single classification model, in general increasing the complexity of the system by combining predictions from different static models is still not enough to mitigate the impact of concept drift.

Retrain, *sliding*, and *incremental* approaches, which implement, even if in a different way, an updating of the classification system, outperform in general the *static* and *ensemble* approaches. They can be considered adequate to cope with the concept drift phenomenon. However, identifying the best approach is not trivial. Table III reports the average value of F_1^{PN} and Accuracy for the five approaches. Furthermore, for each metric, we also report the ranks scored by the approaches throughout the monitoring campaign, averaged over the eight values computed on as many chunks.

Results confirm that *retrain, sliding* and *incremental* approaches perform comparably and they consistently outperform *ensemble* and *static* approaches. On average, the performance of the three best approaches is comparable or better than that measured applying 10-fold cross-validation on the initial training set (Table I).

Finally, we also measure the number of numeric features used by the classification models at training time: Figure 2c highlights such trend just for the *evolving* approaches, namely *retrain*, *sliding* and *incremental*. Evidently, the number of features for the *static* and the *ensemble* approaches is constant over time, with the same value as reported for all approaches at the first chunk.



(a) F_1^{PN}

(b) Accuracy

(c) #Features

Fig. 2. Comparison of approaches along the online monitoring campaign in terms of (a) F_1^{PN}, (b) Accuracy, and (c) Number of features used at training time.

TABLE III

COMPARISON OF APPROACHES ALONG THE ONLINE MONITORING IN TERMS OF F_1^{PN} AND ACCURACY: AVERAGE VALUES AND AVERAGE RANKS.

	$\mathbf{F_1^{PN}}$		**Accuracy**	
	Average	Avg. Rank	Average	Avg. Rank
Retrain	**0.6812**	**1.94**	**0.6763**	**2.25**
Sliding	0.6768	2.44	0.6744	2.31
Incremental	0.6758	2.19	0.6731	**2.25**
Ensemble	0.6358	4.13	0.6300	3.88
Static	0.6223	4.31	0.6144	4.31

Figure 2c shows a substantial difference between the three approaches: the number of features considered by the classifier at each chunk grows significantly in the *retrain* approach, as new tweets are added to the training set, and moderately in the *sliding* approach, as besides adding tweets we also discard outdated ones. On the other hand it remains constant in the *incremental* case, since only the Bayesian classification model is updated. The fact that the incremental approach maintains good levels of performance despite not updating the attribute space suggests that our scenario is not particularly affected by *feature drift*: although new words or hashtags may appear over time, they are not essential for distinguishing stance classes, since no performance degradation is observed compared to approaches that consider them, namely *sliding* and *retrain*.

In light of the good performance and the reduced complexity, measured in terms of size of the attribute space, we selected the *incremental* approach for carrying out a retrospective analysis aimed at uncovering the stance of Italian Twitter users towards Green Pass.

V. MONITORING STANCE TOWARDS GREEN PASS

The second goal of this work is to provide some insight into the stance expressed towards the Green Pass by the Italian Twitter community. In order to understand how this aspect evolves over time we chose the *incremental* approach, which proved to be the most suitable for dealing with the evolving setting. We labelled all the tweets collected in the time window during our monitoring campaign. The overall volume of tweets retrospectively analyzed is around 450,000.

Figure 3 reports a summary of the stance monitoring.

For each day, we evaluate the percentage of positive, negative and neutral tweets and represent such values as stacked plots. Furthermore, the *overall stance* is also evaluated on a daily basis; on a given day i, the overall stance is computed as follows:

$$Stance_i = \frac{Positive_i}{Positive_i + Negative_i} \tag{2}$$

The line in the central part of the graph indicates the *overall stance* and is colored blue when it is biased toward positive stance (in favor of Green Pass) and red when it is biased toward negative stance (against Green Pass).

From the analysis of the plot we can derive some interesting insights. First and foremost, we see that most of the messages posted on Twitter are labelled as neutral. We can hypothesize that such tweets mostly include news reports concerning the implementation and enforcement of the Green Pass, or do not express a clear stance on the topic. We observe that the percentage of polarized tweets is on average between 20% and 30% for each of the two classes.

Moreover, if we consider the daily stance, we can identify three different trends across the period of analysis. We can observe in fact that for the first tweets collected, spanning July and August 2021, the daily stance tends to shift between positive and negative. In the period that spans between September and mid November 2021 we can instead observe a clear increase in negative-stance tweets. We can hypothesize that this may be due to the fact that, during this period, the Green Pass certificate became mandatory for additional activities, such as for working as a public employee (17 September 2021), for accessing public schools and universities (09 October 2021), and all work places (15 October 2021). Moreover, we must note that the measure sparked a series of protests starting from workers of the Trieste harbour and spanning across Italy. On this aspect, we can also hypothesize that this may be due to the so-called *negativty bias* [21]. There is evidence in the literature of a bias favouring the transmission of negative information on social media. In other words, it may be the case that certain groups of people with negative opinions may be more active on social media than people with opposing views. Conversely, in the latter period of our analysis, from late November to the end of December 2021, we see a shift in the stance of Twitter users towards more positive opinions. This also may be driven by several factors. On the one hand the strong increase in contagion across the population due to the emerging Omicron variant and thus a potentially clearer need of restrictive measures. On the other hand, a more active reaction of other parts of population against the protesters, and thus even more in favour of the Green Pass, following the same negativity bias pattern.

VI. CONCLUSION

In this paper, we proposed a pipeline for continuous stance detection in social media, and applied it to the case of the EU digital COVID certificate in Italy. We collected a dataset of tweets concerning the so-called Green Pass from Italian users. We performed a set of experiments aimed at identifying the best performing classifier for the task. We found out that, among the evaluated algorithms, Complement Naive Bayes performed best. Subsequently, we leveraged a labelled portion of our dataset in order to provide insights on the more reliable method for maintaining the classifier performances over time, thus effectively tackling the occurrence of concept drift. Our analysis suggests that an incremental strategy that leverages a partial fitting of the model provides the best trade-off between performances and computational cost of the model. Finally, we used the Complement Naive Bayes model in the incremental setting to perform a retrospective analysis of the stance of

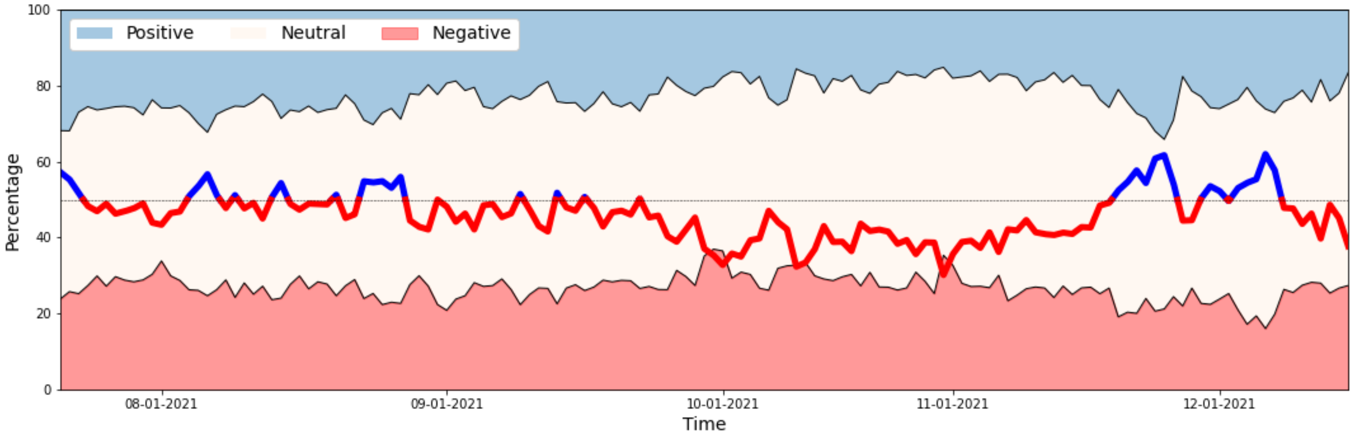


Fig. 3. Daily percentage of tweets by class as a stacked bar plot. The *overall stance* (Eq. 2) is represented as a line and is colored in blue when the positive stance prevails (average higher than 50%) and red otherwise. Best viewed in color.

users on the Green Pass. Our analysis suggests that we can clearly identify several trends in favour or against the pass, and link them to events unfolded in the time period we considered. In the future, we aim to evaluate a higher number of classification algorithms for stance detection, and different strategies to better cope with concept drift. We intend on the one hand to further the research on the complex problem of stance detection, and on the other hand to enable a clearer understanding of how people interact and react to complex problems on social media.

References

[1] D. Küçük and F. Can, "Stance detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, pp. 1–37, 2020.

[2] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *ACM Transactions on Internet Technology (TOIT)*, vol. 17, no. 3, pp. 1–23, 2017.

[3] A. Bechini, P. Ducange, F. Marcelloni, and A. Renda, "Stance analysis of twitter users: the case of the vaccination topic in italy," *IEEE Intelligent Systems*, vol. 36, no. 5, pp. 131–139, 2020.

[4] A. Bechini, A. Bondielli, P. Ducange, F. Marcelloni, and A. Renda, "Addressing event-driven concept drift in twitter stream: A stance detection application," *IEEE Access*, vol. 9, pp. 77 758–77 770, 2021.

[5] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR)*, vol. 46, no. 4, pp. 1–37, 2014.

[6] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 31–41.

[7] K. Dey, R. Shrivastava, and S. Kaushik, "Topical stance detection for twitter: A two-phase lstm model using attention," in *European Conference on Information Retrieval*. Springer, 2018, pp. 529–536.

[8] Q. Sun, Z. Wang, S. Li, Q. Zhu, and G. Zhou, "Stance detection via sentiment information and neural network model," *Frontiers of Computer Science*, vol. 13, no. 1, pp. 127–138, 2019.

[9] W. Li, Y. Xu, and G. Wang, "Stance detection of microblog text based on two-channel cnn-gru fusion network," *IEEE Access*, vol. 7, pp. 145 944–145 952, 2019.

[10] J. Costa, C. Silva, M. Antunes, and B. Ribeiro, "Concept drift awareness in twitter streams," in *2014 13th International Conference on Machine Learning and Applications*. IEEE, 2014, pp. 294–299.

[11] V. Iosifidis, A. Oelschlager, and E. Ntoutsi, "Sentiment classification over opinionated data streams through informed model adaptation," in *International conference on theory and practice of digital libraries*. Springer, 2017, pp. 369–381.

[12] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N project report, Stanford*, vol. 1, no. 12, p. 2009, 2009.

[13] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, "Stream-based active learning for sentiment analysis in the financial domain," *Information sciences*, vol. 285, pp. 181–203, 2014.

[14] G. Shan, S. Xu, L. Yang, S. Jia, and Y. Xiang, "Learn#: A novel incremental learning method for text classification," *Expert Systems with Applications*, vol. 147, p. 113198, 2020.

[15] J. Costa, C. Silva, M. Antunes, and B. Ribeiro, "The impact of longstanding messages in micro-blogging classification," in *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–8.

[16] Costa, Joana and Silva, Catarina and Antunes, Mário and Ribeiro, Bernardete, "Adaptive learning for dynamic environments: A comparative approach," *Engineering Applications of Artificial Intelligence*, vol. 65, pp. 336–345, 2017.

[17] ——, "Boosting dynamic ensemble's performance in twitter," *Neural Computing and Applications*, pp. 1–13, 2019.

[18] E. D'Andrea, P. Ducange, A. Bechini, A. Renda, and F. Marcelloni, "Monitoring the public opinion about the vaccination topic from tweets analysis," *Expert Systems with Applications*, vol. 116, pp. 209–226, 2019.

[19] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," in *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 2017, pp. 502–518.

[20] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the poor assumptions of naive bayes text classifiers," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 616–623.

[21] K. Bebbington, C. MacLeod, T. M. Ellison, and N. Fay, "The sky is falling: evidence of a negativity bias in the social transmission of information," *Evolution and Human Behavior*, vol. 38, no. 1, pp. 92–101, 2017.