IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 24, 2021, accepted May 17, 2021, date of publication May 25, 2021, date of current version June 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3083578

Addressing Event-Driven Concept Drift in Twitter Stream: A Stance Detection Application

ALESSIO BECHINI, ALESSANDRO BONDIELLI, PIETRO DUCANGE, FRANCESCO MARCELLONI, (Member, IEEE), AND ALESSANDRO RENDA
Department of Information Engineering, University of Pisa, 56122 Pisa, Italy

Corresponding author: Alessandro Renda (alessandro.renda@ing.unipi.it)

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

ABSTRACT The content posted by users on Social Networks represents an important source of information for a myriad of applications in the wide field known as 'social sensing'. The Twitter platform in particular hosts the thoughts, opinions and comments of its users, expressed in the form of tweets: as a consequence, tweets are often analyzed with text mining and natural language processing techniques for relevant tasks, ranging from brand reputation and sentiment analysis to stance detection. In most cases the intelligent systems designed to accomplish these tasks are based on a classification model that, once trained, is deployed into the data flow for online monitoring. In this work we show how this approach turns out to be inadequate for the task of stance detection from tweets. In fact, the sequence of tweets that are collected everyday represents a data stream. As it is well known in the literature on data stream mining, classification models may suffer from concept drift, i.e. a change in the data distribution can potentially degrade the performance. We present a broad experimental campaign for the case study of the online monitoring of the stance expressed on Twitter about the vaccination topic in Italy. We compare different learning schemes and propose yet a novel one, aimed at addressing the event-driven concept drift.

INDEX TERMS Automatic stance detection, concept drift, social media analysis, text stream classification.

I. INTRODUCTION

Nowadays, millions of users mention and comment real world events by posting short messages, i.e. *tweets*, on the well-known Twitter platform. Its ease of use and widespread diffusion have rendered it a key source of information for a great variety of data mining applications. For instance, the analysis of tweets has been used for the early detection of real-time events [1], such as traffic congestion and incidents [2], earthquakes [3] and spread of epidemic [4].

Since political and social events typically fuel the online debates, another prominent field of application includes public opinion mining from tweets. A recent investigation [5] addressed the case of 2016 Brexit referendum, quantifying the average stance towards the topic and the influence exerted by Twitter users. More recently, the case of the vaccination topic in Italy has been investigated as well [6], [7]: an intelligent system has been devised with the aim of uncovering trends over time of the opinion expressed through tweets. The particular relevance of online monitoring services in Business Intelligence tools has been stressed for brand reputation [8] and for decision support systems [9].

The above mentioned applications demand for specific software modules, able to analyze and extract knowledge from the noisy and irregular textual content of tweets, relying on methods from data mining, machine learning, and NLP (Natural Language Processing) domains. Furthermore, Twitter can be seen as a particular form of a temporal data stream [10]. Indeed, the volume and features of the collected tweets may change over time, driven by real worlds events of social, cultural or political nature. This observation applies, for instance, to the online monitoring of public opinion: in opinion mining / stance detection tasks, a classification model is typically trained to discern between tweets that express a positive, negative, and possibly neutral opinion. The deployment of the learned model on the online tweet stream may be affected by *concept drift*, namely a change in the data distribution over time, and adequate countermeasures should be

The associate editor coordinating the review of this manuscript and approving it for publication was Yufeng Wang.

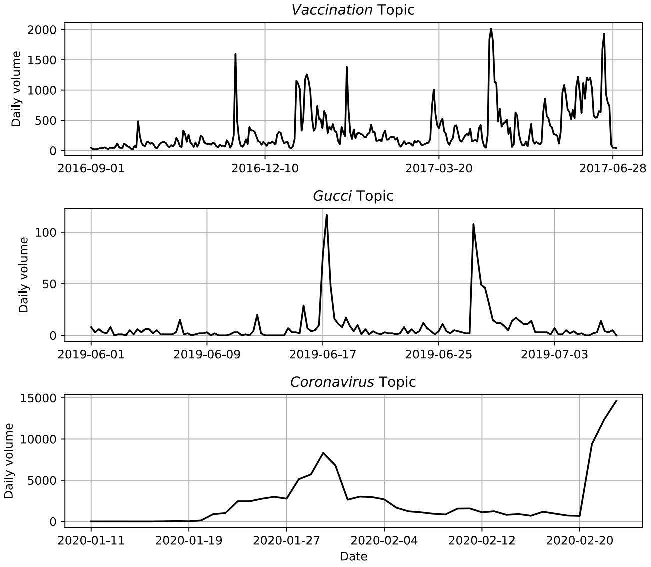


FIGURE 1. Daily number of italian tweets for different topics over different time windows. Top: tweets about vaccination topic. Center: tweets about Gucci brand. Bottom: tweets about coronavirus epidemic.

adopted for avoiding considerable performance loss. Figure 1 reports the daily number of tweets for three example topics (i.e *vaccination, Gucci, coronavirus*), analyzed in the Italian setting over three different periods. The three graphs share the presence of few spikes in the plot of the daily volume of tweets: we can easily verify that such peaks are placed immediately after events that triggered the online debate. As for *coronavirus* epidemic, the first peak relates to news about the severity of the infection in China, whereas the second one relates to the first confirmed cases in Italy. As for *Gucci* keyword, the first peak may be associated with the news of the leading role of the singer Harry Styles in the advertising of the new unisex fragrance of Gucci, and the second one may be associated with the broadcasting of the spot on Italian TV. The detailed analysis of the *vaccination* case study will be presented in Section V.

The present work stems from the observation that the way people express their opinion and thoughts may change over time according to the events-related landscape that characterizes the Twitter stream in online monitoring applications. The event-driven concept drift may alter the relation between the input data (text of a tweet) and the target variable (class of opinion/stance) in the supervised learning setting. In this work, we explore this phenomenon by analyzing the case of the classification of stance towards the vaccination topic in Italy. We extend the original vaccination dataset described in [6] with the aim of carrying out a long-term monitoring campaign. This enables the comparison of different general learning schemes in terms of classification performances. Furthermore, a novel learning scheme is proposed, based on the *semantic association* of the tweets extracted from different events, and specifically designed for fighting the event-driven concept drift. The term *semantics* denotes precise fields of a wide range of disciplines including linguistic and philosophy [11]. The term *semantic association* actually lacks a formal definition. According to Jabeen *et al.* [12], it can be considered as a semantic connection between textual units, such as words, sentences or entire documents [13], [14]. The semantic association between two units of text quantifies their degree of association based on one or more relations that occur between them. In the present work we consider tweets as units of text, and we consider them to be semantically associated if they relate to similar sub-topics within a broader topic. The appropriateness of vaccination is the main, broad topic under investigation; over time, the debate may evolve into several subtopics depending on the occurrence of real world events (e.g. news about the incidence of a disease, statements and political acts, discussion about the obligation for children). On the one hand we are interested in capturing the stance of users towards the vaccination topic in general, regardless of the specific subtopics. On the other hand, we also aim to capture how sub-topics are associated with, and thus related to, one another, to improve the performance of the classification system. This paper is organized as follows: Section II describes the background on stance detection and concept drift, and it discusses the related works. Section III frames the problem of tweets classification in presence of concept drift and introduces two baseline learning and evaluation schemes. Section IV describes the proposed, semantic-based, learning scheme. Section V presents the experimental setup: we describe the vaccination dataset adopted in this work, the basic classification pipeline, and the learning schemes involved in the empirical comparison. The results are reported in Section VI, while Section VII draws some conclusion.

II. BACKGROUND AND RELATED WORKS

In this section, we first recall the definition of the stance detection task and briefly review the most relevant works in the field of microblogging stance detection. Then, we formalize the notion of concept drift and report the most relevant adaptive solutions that have been proposed for text stream classification.

A. STANCE DETECTION

According to a recent definition [15], stance detection from text is a classification problem where the stance (or position) of the author of the text towards a target is expressed in the form of a category label in the set {Favor, Against, Neither}. In some works, the Neutral class is also added, or it replaces the Neither class. in a stance detection task, the target of interest is predetermined [16], and the stance towards it must be assessed even if it is not explicitly mentioned in the text.

The nature of the contents posted on Twitter, i.e. the popular status update messages dubbed *tweets*, makes the platform particularly suitable for stance detection studies. However, Twitter stance detection is regarded as a challenging task in the NLP panorama: Twitter users typically express their thoughts and opinions using unstructured and irregular sentences, with informal, abbreviated words, colloquial expressions, and often misspellings and grammatical errors. New words and hashtags continuously appear and become popular; irony, sarcasm or ambiguity are frequently present in messages, and Favor and Against stances can be expressed with both sentiment polarities. All these aspects, combined with the limited length of tweets, make Twitter a particularly harsh environment for automatic analysis of text messages.

Recent state of the art works on stance detection are in general far from achieving performances comparable to those of other NLP branches, e.g. ordinary sentiment analysis. Authors in [16] carried out an extensive experimental study that led to the following contributions: the introduction of a new stance dataset, the organization of a shared task competition on such dataset (SemEval 2016 - Task 6), and the development of a state-of-the-art stance detection system. The best performance has been obtained with a Linear SVM classifier; the exploitation of features drawn from training samples along with external resources led to an average F-score of 70.3. The outcomes of the above mentioned competition have been widely described [17], and the average values of F-scores, obtained by 19 different teams, ranged from 46.19 to 67.82. More recently, an approach based on a deep learning architecture to tackle the stance detection task on the same dataset has been proposed [18]: it consists of a two-phase LSTM model with attention, getting to a best-case average F-score of 68.84 and a best-case accuracy of 60.2%. The shared representation between stance and sentiment that has been proposed in [19] has not led to significant improvements, with an average F-score of 60.2. The rather poor performance figures are also observed with regard to other datasets, e.g. for the Chinese microblog stance detection task (NLPCC 2016), with a best case accuracy of 60.6% and average F-score of 62.2 [20].

B. CONCEPT DRIFT

A formal general definition of concept drift is given in [21]. Concept drift between two timestamps t_0 and t_1 can be defined as:

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

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

In a classification task the goal is to predict the categorical target variable y given the set of input variables $\mathbf{X}$. When dealing with continuous classification of data streams over time, a naive solution can be devised on training the classification model using data extracted in an initial time interval, and using the predictive model to classify new examples. In this setting, a change in the prior probability of classes $p(y)$ or in the class conditional probability $p(\mathbf{X}|y)$ may lead to concept drift, causing the performance of the classification system to deteriorate. Thus, such naive solution turns out to be inadequate for dealing with evolving data, and adaptive learning strategies, i.e. capable of reacting to concept drift, should be considered.

In practical applications the adaptation strategy depends on the particular type of concept drift: in [21] authors distinguish between different forms of changes in data distribution over time, referring to the one-dimensional toy example reported in Figure 2. *Sudden* or *abrupt* concept drift refers scenarios where the mean of the data distribution suddenly switches from one value (or concept) to another, without exploring intermediate values. This also happens in *gradual* concept drift, but in this case both concepts are maintained during a transitional period. Conversely, in *incremental* concept drift the data distribution switches from one concept to another exploring many intermediate concepts. Notably, data distribution concepts may exhibit redundancy or periodicity: when the drift restores already seen concepts it is referred to as *reoccurring* concept drift.

As the concept drift problem affects various domains and application areas, detection and adaptation techniques are extremely varied as well. A thorough description of such techniques is beyond the scope of the present work: in the following section we focus on the adaptation strategy in social network environment.

C. ADAPTIVE SOLUTIONS FOR TEXT STREAM CLASSIFICATION

Few works have addressed the issue of text stream classification in presence of concept drift.

Costa *et al.* [10] analyzed the issue of concept drift adaptation for a classification task over a tweet stream. The purpose of their classification problem was to predict the hashtag of each tweet on the basis of its text. They compared three different schemes: the time-window approach, the incremental approach, and the ensemble approach. The *time-window* approach learns a new model for each new chunk of tweets, thus implementing an abrupt forgetting mechanism; the *incremental* approach extends the learning set and retrain the model at each new chunk of tweets, preserving all previous examples; finally, the *ensemble* approach consists in combining the prediction of different classifiers trained on different time-windows. Their experimental setup was based on a dataset in which the order of tweets (and indeed the frequency of each hashtag) had been artificially altered in different ways to induce different types of concept drift; by using Bag of Words for numerical representation and SVM as classification algorithm, they observed that the incremental scheme outperformed the other approaches in almost all the drift settings. Nevertheless, they highlighted that such an approach is viable only whenever the storage capability is not a concern. On the other hand, the time-window model does not suffer from this problem, but it requires sophisticated policies for data forgetting, depending on the type of drift that occurs. In addition, it should be noted that the fictitious nature of the drift undermines the generalization of the results into real-world applications.

In a later work [22], the same authors deepened the analysis of the *ensemble* approach by comparing a novel scheme, named DARK (Drift Adaptive Retain Knowledge), with the

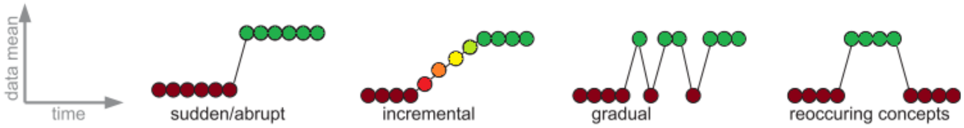


FIGURE 2. Different types of concept drift. Figure adapted from [21].

Learn++. NSE algorithm [23], an extension of Learn++ algorithm [24] for non-stationary environments. DARK consists in a dynamically weighted ensemble of classifiers. Differently from Learn++. NSE, it implements two forgetting mechanisms: the first one is regulated by the size of the training time window for each base classifier, whereas the second one is regulated by the number of classifiers in the ensemble. Corroborating the results presented in [25], they showed that an increased training window size positively contributes to the ensemble classification. At the same time, limiting the ensemble size makes DARK a lightweight solution, capable of achieving comparable or even better performance than the Learn++. NSE algorithm. The analysis presented in [26] shows how the choice of classifiers' performance metric for dynamic weighting and the use of a feedback strategy for exploiting misclassified examples allow boosting ensemble performance in presence of concept drift.

A completely different approach has been pursued in yet another research work [27]: the authors adopted an active learning scheme for sentiment analysis of tweet streams in the stock market domain, with the final goal of predicting the future value of stock prices on the basis of the public mood expressed on social network. The active learning scheme provides the algorithm with the ability to select new training data and query an expert for hand labeling. As a result, the classifier performance was improved.

However, none of the above mentioned works explicitly exploit the semantic information (intended as subtopic or scope of meaning) contained in the text messages. We argue that a semantic-aware approach could be helpful in text classification task, specifically when the text stream is affected by event-driven concept drift.

A recent work in this direction [28] proposes an *informative-adaptation-to-change* approach to deal with concept drift for polarity learning in opinionated data stream. A damped window aging mechanism gives lower importance to older objects and is embedded in the classification model, a Multinomial Naive Bayes classifier, with an aging factor continuously tuned according to the stream dynamics. Whenever a change between the *current window* vocabulary and *past window* vocabulary is detected, the parameter lambda is increased and the model tends to forget outdated data more rapidly. The experimental evaluation on the TwitterSentiment dataset [29] showed that the aging mechanism is beneficial for the classification task, but it also revealed that the informed vocabulary-based adaptation scheme is equivalent to a blind adaptation scheme, i.e. without change detection. This shortcoming may be due to several aspects, such as the extremely wide scope of the dataset, and the limited types of concept drift that the damped window model can cope with.

A novel method named Learn# [30] addresses the issues of long training time and catastrophic forgetting in incremental learning for a deep learning architecture. It shows remarkable results on several text classification tasks, yet without discussing the concept drift adaptation problem.

III. PROBLEM DEFINITION AND BASELINE LEARNING SCHEMES

The scenario of our investigation is a monitoring campaign, focused on a certain topic and based on a supervised learning task, carried out on the Twitter data stream. The case study concerning the stance towards the vaccination topic in Italy will be detailed in the experimental section.

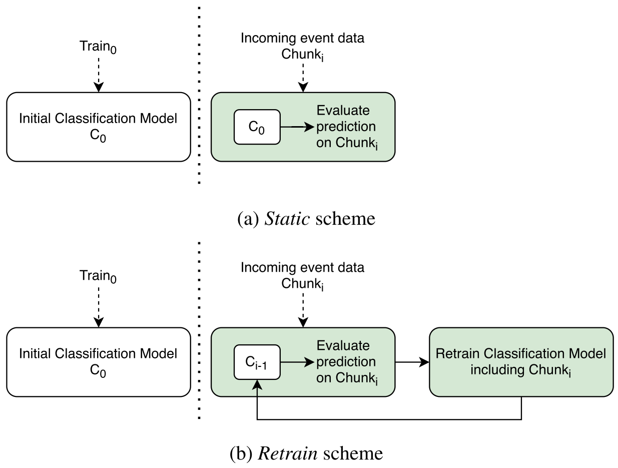
As we pointed out in Section I, the occurrence of a topic-related real-world event is often reflected in a spike in the daily volume of collected tweets, regardless of the topic under investigation (Fig. 1). In the online monitoring of stance based on Twitter stream it can be reasonably assumed that we are mostly interested in measuring the stance during such events. Thus, besides having a large number of messages available, we can assess how the event has affected people's stance about the topic.

In the following, we introduce the notation adopted along the paper.

- t_i indicates the time window of each event i;
- DS_i indicates the set of tweets collected within the time window t_i;
- $chunk_i$ indicates the labelled subset of DS_i.

In this work we do not discuss the problem of new event detection: looking at Fig. 1, we can hypothesize the exploitation of a peak detection algorithm on the daily volume of tweets, possibly combined with the analysis of the news stream on the topic. For the purpose of this work, we manually select the most relevant peaks and set a time window t_i of the size of 3 days from the date of the triggering event.

Furthermore, we assume that we can rely on an initial training set, $train_0$, prior to the online monitoring phase, that can be used to train an initial classification model C_0. This training set may be built with an initial portion of the data stream under investigation, or may belong to another dataset and be used in a transfer learning fashion. From an operational point of view, the cost of the online annotation of a set of tweets at each event ($chunk_i$) may be counterbalanced by the following advantages: on the one hand, the availability of labelled data enables the evaluation of the performance


Train$_0$
Incoming event data Chunk$_i$
Initial Classification Model C$_0$
C$_0$
Evaluate prediction on Chunk$_i$
Train$_0$
Incoming event data Chunk$_i$
Initial Classification Model C$_0$
C$_{i-1}$
Evaluate prediction on Chunk$_i$
Retrain Classification Model including Chunk$_i$

(a) *Static* scheme

(b) *Retrain* scheme

FIGURE 3. Schematic representation of two baseline learning schemes.

of any classification system throughout the monitoring campaign. On the other hand, such data can be used to update the classification model and to devise alternative learning schemes.

Indeed, our experimental analysis is designed to address the following research questions: (i) is the initial classifier adequate for the online classification of the tweets stream? and (ii) what is the most effective learning scheme to cope with the event-driven concept drift?

A. BASELINE LEARNING AND EVALUATION SCHEMES

Figure 3 illustrates two baseline schemes for learning and evaluation: *Static* (Figure 3a) and *Retrain* (Figure 3b).

Static scheme simply consists in using the initial model to classify the whole stream of tweets. The pseudocode of the approach is reported in Algorithm 1.

Algorithm 1 Static Learning and Evaluation Scheme

```
Require: stream: the stream of tweets
Require: $train_0$: initial training set
performance_metrics ← empty list
$C_0$ ← classification_model.train($train_0$)
for each new detected event $EV_i$ on stream do
    $DS_i$ ← collection of tweets related to $EV_i$
    $chunk_i$ ← labeled subset of $DS_i$
    metrics ← evaluate_prediction($C_0$, $chunk_i$)
    performance_metrics.add(metrics)
end for
return performance_metrics
```

Retrain scheme consists in retraining the classification model at each newly detected event. The pseudocode of the approach is reported in Algorithm 2. In data stream mining this approach is often referred to as *prequential evaluation* or *interleaved-test-then-train*. After collecting and labelling samples from an incoming event, we first evaluate the performance of the current classification model on the new labelled set. Then, such data are used to train the classification model. In particular, we incrementally extend the training set with the new labelled data and retrain the classification pipeline from scratch on the new extended training set. Notably, an alternative approach would consist in *incremental* learning: unlike the *Retrain* scheme, it does not replace an old model with a new one trained from scratch. Instead, it updates the existing model by just considering the new labelled data. An emblematic example is the optimization through stochastic gradient descent where the model can be updated with partial fitting on a minibatch of new data. However, as previously pointed out, the introduction of new words or hashtags is frequent in the Twitter Stream and the collection of new instances could alter the attribute space on which each sample is represented, thus giving rise to the so-called *feature drift* [31]: ignoring new words and not updating the attribute space may result in a significant loss of information. In this regard, the Hashing Vectorizer[1] can be used to incrementally training a text classification pipeline: the hashing trick [32] allows mapping a string token into a feature integer index, without the need to store a fixed vocabulary dictionary in memory. However, this approach has several drawbacks: (i) there is no way to evaluate the inverse mapping (i.e. from feature index to string representation), (ii) IDF weighting scheme cannot be applied (iii) hash collisions may occur. Furthermore, several recent works have highlighted that it delivers weaker performance compared to vectorization based on TF-IDF [33], [34]. We have therefore decided not to adopt a purely incremental approach, but rather an approach (i.e. *Retrain*) that re-evaluate the whole classification pipeline at every incoming event.

Algorithm 2 Retrain Learning and Evaluation Scheme

```
Require: stream: the stream of tweets
Require: $train_0$: initial training set
performance_metrics ← empty list
$train_i$ ← $train_0$
$C_i$ ← classification_model.train($train_i$)
for each new detected event $EV_i$ on stream do
    $DS_i$ ← collection of tweets related to $EV_i$
    $chunk_i$ ← labeled subset of $DS_i$
    metrics ← evaluate_prediction($C_i$, $chunk_i$)
    performance_metrics.add(metrics)
    $train_i$ ← $train_i \cup train_0$
    $C_i$ ← classification_model.train($train_i$)
end for
return performance_metrics
```

Static and *Retrain* schemes are diametrically opposed: the *Static* approach minimizes the computational cost, by just exploiting the information available in the initial training set *train*$_0$; the *Retrain* approach, conversely, leverages all the available data for model training, but it entails the cost of

[1] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.HashingVectorizer.html

periodically re-training the model. As a result, the *Retrain* scheme is designed to achieve adaptation to concept drift.

The inclusion of new instances in the training set is not the only path to adaptation: a popular approach consists in discarding outdated information from the model [28]. The most intuitive implementation of such a forgetting mechanism is through the adoption of a sliding window model or a damped window model. In the former case, least recent objects (outside the sliding window) are abruptly discarded; in the latter case, the weight of an instance decreases with its age. However, both approaches require the setting of a hyper-parameter, either the size of the sliding window or the decay factor, which makes them less attractive in real applications. Another major drawback of the traditional forgetting approaches is discussed in the next section, where the proposed learning scheme is introduced.

IV. THE PROPOSED *SEMANTIC* LEARNING SCHEME

It is widely recognized that, for concept drift adaptation, it may be useful to discard part of the information acquired in the past, that is no longer relevant for the current task. However, traditional forgetting schemes, namely sliding window model or damped window model, rely on the following hypothesis: the more recent instances are more important than the less recent ones. In other words this is equivalent to decide the type of concept drift and specifically to assume its *incremental* nature (see Fig. 2).

The dynamic of tweets stream from Twitter environment, though, cannot be traced back to a simple, monotonic, trend. The online discussion is often driven by real-world events and different types of concept drift, other than the incremental one, may arise, e.g. reoccurring drift. Therefore, we argue that it is not appropriate to give importance to the past instances merely on the basis of the remoteness in time from the current evaluation, but rather to conceive an alternative criterion. This is the rationale behind our proposed approach, namely the *Semantic* scheme: in the evaluation of the current event, the weight of the samples collected during a past event should be proportional to the semantic association between the current event and the past event itself. The pseudocode of the approach is reported in Algorithms 3 and 4 and detailed below.

A. THE OVERALL APPROACH

In the initialization stage (Algorithm 3, lines 2:3) we train the initial classification model with the available training set ($train_0$) as in the baseline approaches. Furthermore, we define a list of sets of labelled tweets, denoted as *train_pool*, which at the beginning only contains $train_0$. Whenever a new event i is detected we select the tweets related to the event from the stream (DS_i). The procedure Eval_Similarity (line 6) evaluates the similarity between the data of the new event and each subset of *train_pool*, traceable to a past event or to the initial training set, and it returns a list of weights of the same length as *train_pool*. The *train_pool* list is flattened (line 7) and the weight coefficient assigned to an event is transferred to each sample belonging to that event (line 8). Hence, a new classifier is trained by taking into account the weighted samples and is evaluated on the labelled set of the new event. Finally the *train_pool* is updated by including the recently labelled elements, i.e. $chunk_i$.

Algorithm 3 Semantic Learning and Evaluation Scheme

```
Require: $stream$: the stream of tweets
Require: $train_0$: initial training set
Require: $k$: nearest neighbors parameter
Require: $N$: number of new event tweets used to assess similarity
 1: $performance\_metrics \leftarrow$ empty list
 2: $C_i \leftarrow$ classification_model.train($train_0$)
 3: $train\_pool \leftarrow [train_0]$
 4: for each new detected event $EV_i$ on $stream$ do
 5:     $DS_i \leftarrow$ collection of tweets related to $EV_i$
 6:     $events\_weights \leftarrow$ Eval_Similarity($train\_pool, DS_i, k, N$)
 7:     $train_i \leftarrow \bigcup train\_pool$
 8:     $w_i \leftarrow$ map_sample($train\_pool, events\_weights$)
 9:     $C_i \leftarrow$ classification_model.train($train_i$,$w_i$)
10:     $chunk_i \leftarrow$ labeled subset of $DS_i$
11:     $metrics \leftarrow$ evaluate_prediction($C_i$, $chunk_i$)
12:     $performance\_metrics$.add($metrics$)
13:     $train\_pool$.add($chunk_i$)
14: end for
15: return $performance\_metrics$
```

Algorithm 4 Evaluate Similarity

```
Require: $DS_i$: collection of tweets related to last event
Require: $train\_pool$: list of available labelled sets of tweets
Require: $k$: nearest neighbors parameter
Require: $N$: number of new event tweets used to assess similarity
 1: $subset_i \leftarrow$ random.sample($DS_i, N$)
 2: $kNN \leftarrow$ NearestNeighbors($k$).fit($train\_pool$)
 3: $indices, distances \leftarrow kNN$.get_neighbors($subset_i$)
 4: $events\_freq \leftarrow$ map_on_event($indices$)
 5: $events\_weights \leftarrow$ normalize($evenst\_freq$,$train\_pool$)
 6: $events\_weights \leftarrow$ scale($events\_weights$)
 7: return $event\_weights$
```

B. SEMANTIC SIMILARITY USING BERT

The Eval_Similarity procedure, described in Algorithm 4, assigns a weight to each subset of the *train_pool* based on the semantic association or similarity with the incoming event. We randomly select N tweets from DS_i (line 1): in this way we standardize the conditions of the various incoming events and reduce the computational load. For each of the N tweets in this subset, we evaluate its k nearest neighbors among the tweets of the *train_pool* (lines 2-3). Then, in *events_freq*, we store for each past event the cumulative number of times

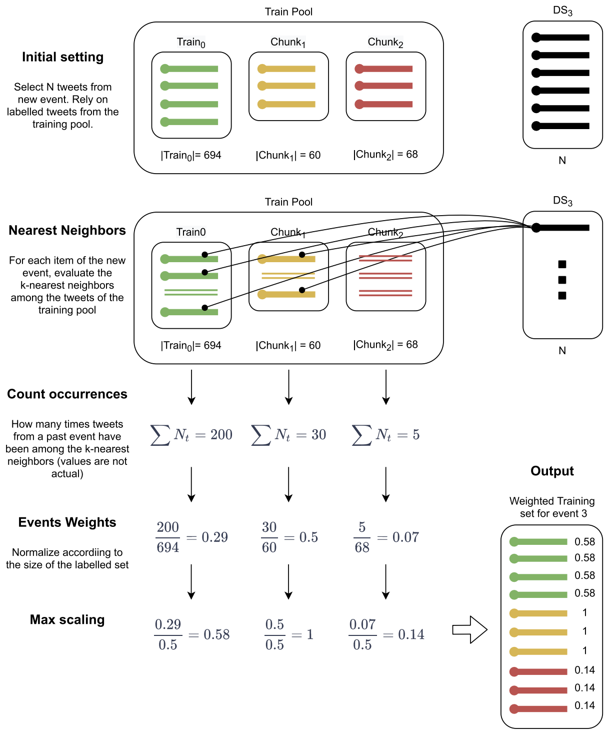


FIGURE 4. Example of operation of the evaluate similarity procedure for the incoming event 3: scaled events weights are used for weighting each sample of the training pool.

a tweet from the event itself has been among such k nearest neighbors. Since the subsets of *train_pool* may have different size, we normalize the *events_freq* array by dividing the value associated to an event by the size of the related labelled subset (line 5). In other words, given a subset J of labelled tweets from *train_pool* and the cumulative number N_t of times a labelled tweet t from J has been among the k nearest neighbors of the new event tweets, the weight w_J to be assigned to the samples of J is derived as follows:

$$w_J = \frac{\sum_{t \in J} N_t}{|J|} \tag{2}$$

Finally, we scale the weights by its maximum value. A schematic example of the *Evaluate Similarity* procedure is reported in Fig. 4.

The distance assessment underlying the Nearest Neighbors algorithm requires a numerical representation of the tweets that can encode semantically-driven information. To this aim, we resort to the recently proposed BERT language model [35]. Unlike classical word embedding, like Word2Vec [36], Glove [37] and FastText [38], which map the same token to the same vector regardless of its context, BERT is a language model based on the Transformer architecture [39] and implements a *masked language model* learning strategy to learn a contextualized representation of words and sequences. BERT models have been applied to a broad set of NLP tasks, including language understanding and question answering, by using them in a transfer learning setting where the pre-trained model is subsequently fine-tuned on the downstream task [35]. However, pre-trained models can be directly leveraged as feature extractor, in order to obtain contextualized representations of words and sequences. We opted to compute the final representation of each tweet as the dimension-wise average of the representations of individual words. We argue that the average vectors of semantically related tweets may be similar in the embedding space. As it is common practice for NLP related tasks, the similarity between embeddings used in the Nearest Neighbors algorithm is computed by means of the cosine distance. We must point out that other language models specifically focused on providing semantically-relevant sequence vectors, such as Sentence-BERT [40]. However, our choice of exploiting standard BERT models is motivated by the fact that no pre-trained Sentence-BERT models are specifically available for Italian.

We would also like to underline that the assessment of the semantic similarity relies on the availability of tweets for the newly detected event: we therefore assume that inference on the event will be performed with a slight latency. However, the similarity evaluation is an unsupervised procedure, and it does not require the costly and time consuming manual annotation of tweets to be performed immediately. For the purpose of our empirical comparison across learning schemes, and indeed in Algorithms 1,2 and 3, we supposed to have the labelled chunk of tweets immediately available for performance evaluation.

V. EXPERIMENTAL SETUP

In this section, we first describe the extended vaccination dataset, object of our experimental investigation, and then we provide the details of the basic text classification pipeline, shared among all the different learning schemes. Finally, we discuss the implementations of the learning schemes involved in the experimental analysis.

A. THE EXTENDED VACCINATION DATASET AND RELATED REAL-WORLD EVENTS

In one of our previous works [6] we introduced the ‘‘vaccination stance’’ dataset: we carried out a monitoring campaign of the vaccination topic in Italy over a period of intense public discussions ranging from September 1st, 2016 till June 30th, 2017. We queried Twitter API with a list of 38 vaccine-related keywords, such as #iovaccino (hashtag for ‘‘I vaccinate’’) or rischio vaccinale (vaccine risk). In the present work, we have extended the monitoring campaign until September 2019. The resulting daily volume of tweets is shown in Fig. 5; after filtering out non-Italian tweets and removing duplicate, we gathered 806,672 tweets.

Most of the peaks of the tweets volume distribution have been highlighted with a red circle in Fig. 5: by examining the news and press review of those days, we were able to match the peaks with specific real-world events, related to the vaccination topic. Table 1 collects the details of such events.

For the purpose of performance evaluation and concept drift adaptation, we need labelled data for each event. To this aim, we have randomly sampled a subset of the related dataset

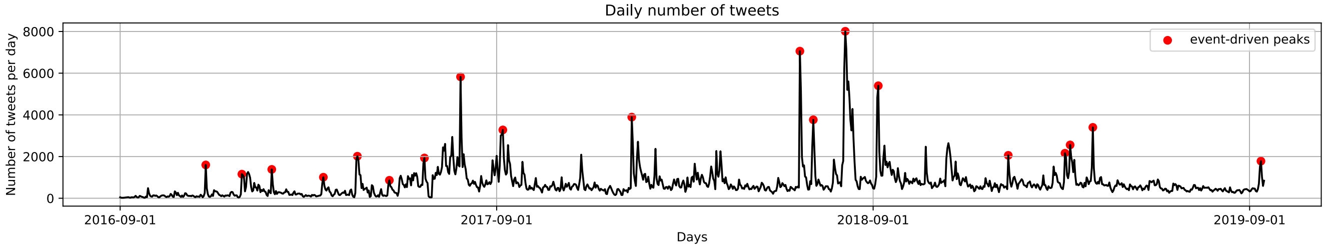


FIGURE 5. The daily number of tweets for the extended vaccination dataset. Peaks in the distribution are located at dates in which a real-world event has stirred up the online discussion about the vaccination topic. Red dots highlight the events that have been analyzed in our work.

TABLE 1. Real world context related events.

ID	Date	Event description
1	2016-11-22	Approval of the law establishing vaccination requirements for school children in Emilia Romagna Region;
2	2016-12-28	Death of a school teacher for meningitis in Rome;
3	2017-01-26	Agreement between Italian Health Minister and Italian Regions about vaccinations requirement;
4	2017-03-16	Increase of 230% cases of measles in Italy;
5	2017-04-19	Fake vaccinations in the Italian city of Treviso;
6	2017-05-19	Approval of the decree on vaccinations requirement (12 vaccines) in Italian kindergartens;
7	2017-06-22	Kid sick of leukemia died for measles in Monza;
8	2017-07-28	Decree about vaccines becomes law;
9	2017-09-07	President of Veneto region Zaia suspends the 2 years moratorium for children admission to school;
10	2018-01-10	Lega Nord party secretary Salvini declares about vaccination to delete former Health Minister law;
11	2018-06-22	Lega Nord party secretary Salvini declares about vaccination: 10 vaccines are too much;
12	2018-07-05	Health Minister Grillo modifies former minister law introducing self-certification;
13	2018-08-04	Doctor's order and italian regions against the extension about vaccines;
14	2018-09-05	Government's retreat about vaccination: vaccines remain mandatory;
15	2019-01-10	Grillo (M5S) and Renzi (former Prime Minister) sign the "pact for science" promoted by virologist Roberto Burioni;
16	2019-03-06	Lega Nord party secretary Salvini advocates a decree to force schools to keep unvaccinated children in class;
17	2019-03-09	The school principals: "On Monday those who do not have the certificate cannot enter the kindergarten.";
18	2019-04-01	A branch of M5S party declares in favour of the vaccines obligation;
19	2019-09-10	A woman goes on hunger strike following the exclusion of her daughters from kindergarten;

and manually labelled around 70 tweets, always pursuing a good balance among the three classes. We excluded from the sampling the tweets belonging to the $train_0$ set, built in the initial months of the monitoring campaign and possibly concomitant with the first few events. An overview on the number of tweets for each event is reported in Table 2.

B. THE STANCE CLASSIFICATION PIPELINE: PREPROCESSING, NUMERICAL REPRESENTATION, CLASSIFICATION MODEL

In our case study the stance detection problem is instantiated as a three-class classification problem intended as the assessment of whether the text of a tweet conveys an opinion in favor, not in favor, or neutral toward the target.

Whatever the learning scheme used for concept drift adaptation, the stance detection task requires the definition of a base classification pipeline. In the present work we rely on the algorithm described in [6], which has proven to be appropriate after an extensive model selection phase. It encompasses the following steps: the text of each tweet is pre-processed by removing links, mentions, numbers and special characters, and by converting it into lower case. A subset of stop words for Italian language are filtered out. Hence, to obtain a numerical representation, each tweet is first converted into a set of tokens according to the Bag of Words model: tokenization has been performed considering uni-grams (n-grams with n = 1) and bi-grams (n-grams with n = 2) and each word is reduced to its stem or root form with analogous semantics. Then, a vector of numeric features is computed by using TF-IDF index (Term Frequency - Inverse Document Frequency). A Support Vector Machine with linear kernel is used as classification algorithm. Such approach allows achieving an average accuracy of 64.84% using 10-fold stratified cross-validation on a well-balanced dataset of 693 labelled tweets, which is in line with the performance figures reported in Section II-A. The labelled dataset of 693 tweets collected between September 1st, 2016 and April 30th, 2017, is used as initial training set, namely $train_0$.

C. LEARNING SCHEMES, PARAMETER SETTING AND EVALUATION METRICS

Our empirical study is aimed at comparing the following learning schemes:

- **Static**: baseline scheme, as described in Algorithm 1;
- **Retrain**: baseline scheme, as described in Algorithm 2;
- **DARK**: state of the art approach proposed in [22] and discussed in Section II-C;
- **Semantic**: our proposed approach, as described in Algorithms 3 and 4;

Static and *Retrain* schemes do not require any configuration parameter. For the configuration of *DARK* scheme we take advantage of the main findings presented by authors in [22]: we test the approach varying the windows size (number of recent events considered for training a new classifier)

TABLE 2. Total number of tweets (DS_i) and labelled tweets ($chunk_i$) related to each event, along with their cumulative sum.

i	$Date$	$\|DS_i\|$	$\sum\|DS_i\|$	$\|chunk_i\|$	$\sum\|chunk_i\|$
1	2016/11/22	2535	2535	60	60
2	2016/12/28	3281	5816	68	128
3	2017/01/26	2352	8168	77	205
4	2017/03/16	2767	10935	85	290
5	2017/04/19	6082	17017	72	362
6	2017/05/19	3400	20417	73	435
7	2017/06/22	6094	26511	80	515
8	2017/07/28	8733	35244	69	584
9	2017/09/07	5209	40453	69	653
10	2018/01/10	7759	48212	69	722
11	2018/06/22	12198	60410	69	791
12	2018/07/05	7307	67717	69	860
13	2018/08/04	14210	81927	69	929
14	2018/09/05	12463	94390	69	998
15	2019/01/10	2058	96448	66	1064
16	2019/03/06	2161	98609	69	1133
17	2019/03/09	2559	101168	64	1197
18	2019/04/01	3401	104569	69	1266
19	2019/09/10	3288	107857	141	1407

TABLE 3. Comparison of learning schemes. Accuracy values obtained on the test set associated with each event and average values.

	Static	**Retrain**	**DARK**	**Semantic**
Ev_1	**0.600**	**0.600**	**0.600**	**0.600**
Ev_2	**0.647**	**0.647**	**0.647**	0.632
Ev_3	0.597	**0.610**	**0.610**	**0.610**
Ev_4	0.671	0.706	0.706	**0.718**
Ev_5	**0.500**	0.472	**0.500**	0.472
Ev_6	0.466	**0.562**	0.548	0.548
Ev_7	0.450	0.488	0.462	**0.500**
Ev_8	0.652	**0.696**	0.667	0.681
Ev_9	0.609	0.638	**0.667**	**0.667**
Ev_{10}	0.435	0.507	0.493	**0.522**
Ev_{11}	0.478	0.536	0.536	**0.580**
Ev_{12}	0.507	**0.594**	**0.594**	0.580
Ev_{13}	0.522	0.594	0.536	**0.638**
Ev_{14}	0.449	0.580	0.580	**0.609**
Ev_{15}	0.470	0.470	0.485	**0.515**
Ev_{16}	0.435	0.580	0.551	**0.623**
Ev_{17}	0.453	0.562	0.562	**0.625**
Ev_{18}	0.420	0.638	0.522	**0.667**
Ev_{19}	**0.546**	0.525	0.504	0.539
Avg	0.521	0.579	0.567	**0.596**

between 4 and 5, and the ensemble size (number of classifiers in the ensemble) considering both 4 and *all* available classifiers. Notably, the sliding window is just applied to the events data: $train_0$ is excluded from the forgetting mechanism, so that the training set is always large enough. Furthermore, we vary the criterion for dynamic weighting of base classifiers as in [26], by using simple majority voting or previous performance of each classifier, namely F-measure and accuracy. For the sake of clarity, in the following we just report the best configuration, which has been obtained with window size equal to 5, ensemble size equal to 4 and F-measure as the performance metric for the combination of classifiers.

Semantic scheme requires the definition of a few parameters: k represents the number of neighbors of each tweet for the evaluation of the similarity between events. N represents the number of tweets of the new event used to assess the similarity with past events. In our experiments we set $k = 5$ and $N = 2000$ and we also investigate the sensitivity of the *Semantic* learning scheme w.r.t. the two parameters in Section VI-B. The choice of N obviously depends on the volume of available data: in our case study, we have chosen a value slightly lower than the minimum value of $|DS_i|$, as reported in Table 1. As for the pre-trained BERT model, we resort to AlBERTo [41], a BERT-based model pre-trained on Italian texts obtained from social media, and specifically tailored to Twitter. The dimensionality of the resulting vector space is 768.

TABLE 4. Comparison of learning schemes. F-measure values obtained on the test set associated with each event and average values.

	Static	**Retrain**	**DARK**	**Semantic**
Ev_1	**0.584**	**0.584**	**0.584**	**0.584**
Ev_2	0.633	**0.634**	**0.634**	0.614
Ev_3	0.594	0.608	0.607	**0.612**
Ev_4	0.661	0.702	0.703	**0.715**
Ev_5	**0.504**	0.473	0.502	0.473
Ev_6	0.469	**0.564**	0.551	0.551
Ev_7	0.451	0.485	0.459	**0.496**
Ev_8	0.652	**0.696**	0.668	0.682
Ev_9	0.612	0.646	0.672	**0.675**
Ev_{10}	0.403	0.496	0.479	**0.507**
Ev_{11}	0.412	0.479	0.464	**0.531**
Ev_{12}	0.498	0.572	**0.579**	0.556
Ev_{13}	0.483	0.573	0.504	**0.635**
Ev_{14}	0.445	0.580	0.588	**0.607**
Ev_{15}	0.461	0.459	0.483	**0.513**
Ev_{16}	0.382	0.569	0.539	**0.626**
Ev_{17}	0.449	0.560	0.561	**0.625**
Ev_{18}	0.371	0.633	0.514	**0.660**
Ev_{19}	0.518	0.498	0.475	**0.526**
Avg	0.504	0.569	0.556	**0.589**

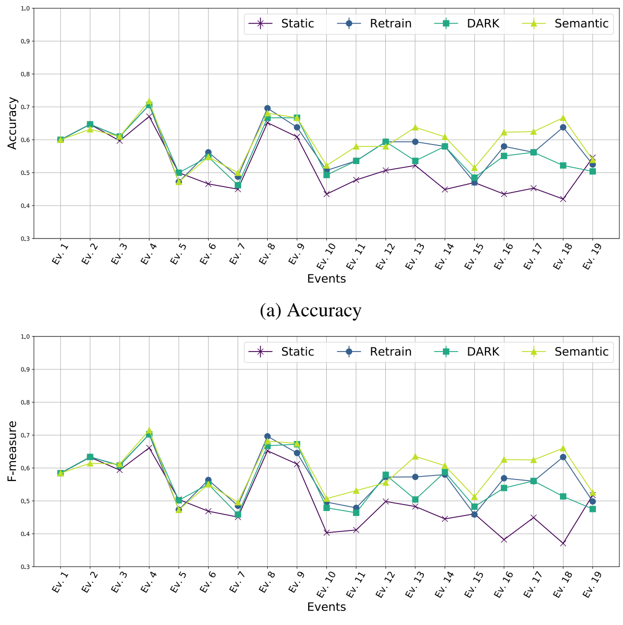
Performance are evaluated in terms of accuracy and macro-averaged F-measure.

VI. EXPERIMENTAL RESULTS

Results of the experimental campaign are reported in Tables 3 and 4 for accuracy and F-measure, respectively. The trends of the performance metrics over time are also shown in the graphs of Figures 6a and 6b.

First of all, we observe a considerable variability of the results among the various events. In many cases, performances are consistent with or better than expected (accuracy $> 60\%$) with figures roughly coherent with those reported in state-of-the-art works in the stance detection literature (see Section II-A). On the other hand, for few events, namely events 5, 7, 10, 15, performances are rather poor, regardless of the adopted learning scheme: the accuracy value settles around 50%, which can be considered inadequate even if we are dealing with a three-class classification problem. Possible explanations may lie in the occurrence of concept drift and the intrinsic complexity associated with certain events or deriving from the annotation procedure. These conditions may be quite common in real applications. Furthermore, it should be underlined that, in the framework of the empirical comparison, we are more interested in the *relative*, rather than *absolute*, performances of different approaches.

Fig. 7 reports the frequencies of the ranks scored by the four models throughout the whole monitoring campaign, both for accuracy and F-measure. The visual analysis of



(a) Accuracy

(b) F-measure

FIGURE 6. Comparison of learning schemes along the monitoring campaign in terms of (a) Accuracy and (b) F-measure.

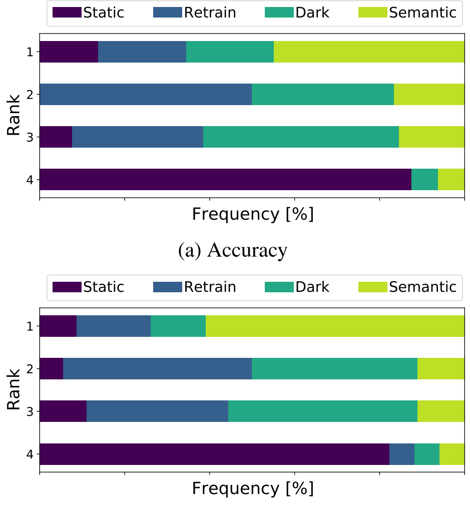
Figures 6a, 6b and 7 clearly reveals that the *Static* scheme achieves the worst results. The lack of a strategy for concept drift adaptation leads this scheme to achieve poor performance. Although such degradation does not seem to depend solely on the time distance from the initial training instant, it is more evident in the second half of the monitoring campaign.

Obviously, in the first event all the schemes perform identically since they are based on the same classifier, trained on $train_0$. In general the discrepancies between the approaches are more noticeable from event 10 onwards. The average values of accuracy and F-measure reported at the end of Tables 3 and 4 confirm that the highest gap occurs between *Static* learning scheme and the others. *Retrain* and *DARK* schemes show comparable performance: the slight difference in favour of *Retrain* approach mainly stems from events 13 and 18. We argue that the information embedded in the labelled sets of older events, those included in *Retrain* but excluded from the *DARK* model, seems to be essential to achieve a better classification.

Nevertheless, in our case study, exploiting and giving equal importance to *all* the past samples, as it happens in the *Retrain* scheme, is not the optimal strategy: our *Semantic* approach often outperforms other learning schemes both in terms of accuracy and F-measure and it achieves the highest average metrics values. The approach based on the semantic similarity of events allows to obtain a better adaptation to the evolving Twitter stream for the stance detection application.

A. QUALITATIVE ASSESSMENT OF SEMANTIC SIMILARITY

Explaining the effectiveness of the *Semantic* approach based only on the list of events previously reported is anything but trivial. In general, it can be noticed that almost all the events of the second half of the monitoring campaign are linked to political aspects, i.e. political declarations and statements of position on the vaccines obligations. This might motivate the slight superior performance of the *Semantic* approach after event 10.



(a) Accuracy

(b) F-measure

FIGURE 7. Frequencies of ranks scored by the four learning schemes on 19 events.

To have a better awareness of the outcome of the similarity assessment procedure based on BERT language model, however, we consider it appropriate to reduce the granularity of the analysis to single tweets. Table 5 shows the results of the k-Nearest Neighbors assessment procedure on two tweets extracted from event 14 and 17, respectively.

In the first case the tweet refers to the government majority decision about the obligation of vaccinations: all the five closest tweets belonging to the training pool make explicit reference to political entities and/or to the vaccines obligation. In the second example, the incoming tweet refers to the vaccines obligation for children to have access to school. The five most similar examples in terms of cosine distance date back to the most recent events and relate to the subjects of obligation, school and children.

B. PARAMETER SENSITIVITY

We investigate the sensitivity of the *Semantic* approach to its input parameters, namely k, the number of nearest neighbors, and N, the size of the subset of the new event dataset DS_i used to assess the semantic similarity with past events. The sampling procedure on DS_i makes the approach stochastic: indeed we should also evaluate the variability induced by using different seeds for the pseudo-random number generator. Specifically, we vary the parameters as follows:

- *seed*: three different integers denoted as A, B and C;
- k: in the set {3, 4, 5, 10};
- N: in the set {200, 500, 1000, 2000, 5000, 15000}.

TABLE 5. Qualitative assessment of tweets similarity: the five nearest neighbors of two tweets (sampled from event 14 and event 17 respectively). Original tweets and english translation.

Tweet	Ev. 14	**Vaccini: retromarcia della maggioranza sul rinvio dell'obbligo** ***Vaccines: reconsideration of the majority on the postponement of the obligation***
NN 1	Ev. 8	Approvata alla Camera la fiducia al governo su decreto legge che aumenta il numero dei vaccini obbligatori da 4... *The Chamber approved the confidence in the government about the decree law that increases the number of compulsory vaccines from 4...*
NN 2	Ev. 13	Vaccini, l'obbligo slitta: ok all'emendamento. Ma Fattori (M5S) vota contro - Politica - *Vaccines, the obligation slips: okay to the amendment. But Fattori (M5S) votes against - Politics -*
NN 3	Ev. 8	#Breaking #News Vaccini: la Camera conferma la fiducia al Governo, 305 sì. Alle 12 il voto finale sul provvedimento *#Breaking #News Vaccines: the Chamber confirms the confidence in the Government, 305 in favor. At 12 o'clock the final vote on the measure*
NN 4	Ev. 8	#IlFattoQuotidiano Vaccini, Camera approva la fiducia sul decreto: 305 sì, 147 contrari #Metapolitica *#IlFattoQuotidiano Vaccines, Chamber approves the confidence on the decree: 305 in favor, 147 against #Metapolitics*
NN 5	Ev. 8	Vaccini, la Camera approva la fiducia: oggi il voto finale sul decreto *Vaccines, The Chamber approves the confidence on the decree: today the final vote on the decree*
Tweet	**Ev. 17**	**Obbligo dei vaccini, 3 bimbi non possono entrare a scuola a Pagani** ***Obligation of vaccines, 3 children can not enter school in Pagani***
NN 1	Ev. 12	Vaccini, comincia demolizione decreto Lorenzin: tutti i bambini potranno andare a scuola anche senza certificato Asl "basterà autocertificazione" *Vaccines, the demolition of the Lorenzin decree begins: all children can go to school even without a certificate Asl "self-certification will suffice"*
NN 2	Ev. 13	Vaccini: nessun bambino sarà escluso da scuola a settembre *Vaccines: no child will be excluded from school in September*
NN 3	Ev. 11	Salvini, inutili 10 vaccini obbligatori, tutti i bimbi a scuola - Focus vaccini *Salvini, 10 compulsory vaccines are useless, all children at school - Focus vaccines*
NN 4	Ev. 16	VACCINI. DE LUCA, 'IN CAMPANIA BIMBI NON VACCINATI NON AMMESSI A SCUOLA' | Tv7 Benevento *Vaccines. De Luca, 'in campania unvaccinated children will not be admitted to school' | Tv7 Benevento*
NN 5	Ev. 16	Per fare andare a scuola i bambini non vaccinati non serve un decreto legge, basta vaccinarli... *To make unvaccinated children go to school you do not need a decree law, just vaccinate them...*

TABLE 6. Sensitivity of the *semantic* approach w.r.t. random *seed*, *k* and *N*. Average accuracy over all events.

	Seed A				Seed B				Seed C			
N \ k	3	4	5	10	3	4	5	10	3	4	5	10
200	0.591	0.593	0.593	0.594	0.588	0.587	0.593	0.594	0.590	0.588	0.590	0.594
500	0.596	0.595	0.596	0.594	0.595	0.593	0.595	0.593	0.594	0.592	0.593	0.592
1000	0.594	0.595	0.597	0.593	0.592	0.595	0.596	0.594	0.595	0.594	0.594	0.594
2000	0.593	0.593	0.593	0.591	0.593	0.596	0.596	0.593	0.595	0.594	0.594	0.592
5000	0.594	0.594	0.593	0.592	0.593	0.594	0.594	0.592	0.593	0.594	0.594	0.592
15000	0.594	0.594	0.593	0.592	0.594	0.594	0.593	0.592	0.594	0.594	0.593	0.592

Notably, when N exceeds the size of DS_i, we simply consider all samples from that dataset. Results are reported in terms of accuracy, averaged over all events, in Table 6.

Although such exploration of the three-dimensional parameter space is not exhaustive, it gives a preliminary insight on the parameter sensitivity. Results are substantially stable across different values of *seed*, N, and k: the average accuracy is almost always higher than 0.59 and indeed at least one percentage point higher than the second best learning scheme, i.e. *Retrain*. Intuitively, the higher the value of N, the lower the influence of the random seed. At the same time, the influence of sampling is greater if N is low, e.g. 200. Furthermore we can notice that a low value of k (e.g. 3, 4, 5) seems to be appropriate, not only from a computational point of view but also for slightly better performance.

VII. CONCLUSION

The analysis carried out in the present work focuses on an aspect that is often neglected in applications of knowledge discovery from Twitter: the volume and characteristics of the tweets stream vary according to real-world events that trigger the online debate. As a consequence, concept drift should be taken into account and proper adaptation techniques should be used.

With reference to the case study of stance detection towards the vaccination topic, we perform an experimental analysis to compare different learning schemes for the evolving setting. We have observed that just resorting to an initial classification model that does not implement any adaptation technique is not suitable, at least for our long-term monitoring campaign. Among the approaches designed for fighting concept drift, our novel learning scheme, based on the evaluation of the semantic similarity of events, shows promising results since it slightly outperforms both *retrain* and *ensemble-based* learning scheme. Future works will investigate the integration of the proposed approach in an end-to-end intelligent system for the online monitoring of the public opinion from Twitter: it might be interesting, for example, to find out how decisions made by government authorities during the Covid-19 pandemic were perceived by general public and affected the opinion towards policy makers.

REFERENCES

[1] A. Dhiman and D. Toshniwal, ''An approximate model for event detection from Twitter data,'' *IEEE Access*, vol. 8, pp. 122168–122184, 2020.

[2] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, ''Real-time detection of traffic from Twitter stream analysis,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2269–2283, Aug. 2015.

[3] T. Sakaki, M. Okazaki, and Y. Matsuo, ''Tweet analysis for real-time event detection and earthquake reporting system development,'' *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 919–931, Apr. 2013.

[4] L. Samaras, E. García-Barriocanal, and M.-A. Sicilia, ''Comparing social media and Google to detect and predict severe epidemics,'' *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, Dec. 2020.

[5] M. Grčar, D. Cherepnalkoski, I. Mozetič, and P. K. Novak, ''Stance and influence of Twitter users regarding the brexit referendum,'' *Comput. Social Netw.*, vol. 4, no. 1, pp. 1–25, Dec. 2017.

[6] E. D'Andrea, P. Ducange, A. Bechini, A. Renda, and F. Marcelloni, ''Monitoring the public opinion about the vaccination topic from tweets analysis,'' *Expert Syst. Appl.*, vol. 116, pp. 209–226, Feb. 2019.

[7] A. Bechini, P. Ducange, F. Marcelloni, and A. Renda, ''Stance analysis of Twitter users: The case of the vaccination topic in Italy,'' *IEEE Intell. Syst.*, early access, Dec. 15, 2020, doi: 10.1109/MIS.2020.3044968.

[8] A. Bechini, B. Lazzerini, F. Marcelloni, and A. Renda, ''Integration of Web-scraped data in CPM tools: The case of project Sibilla,'' in *Proc. 5th Int. Congr. Inf. Commun. Technol.* Singapore: Springer, 2021, pp. 279–287.

[9] P. Ducange, M. Fazzolari, M. Petrocchi, and M. Vecchio, ''An effective decision support system for social media listening based on cross-source sentiment analysis models,'' *Eng. Appl. Artif. Intell.*, vol. 78, pp. 71–85, Feb. 2019.

[10] J. Costa, C. Silva, M. Antunes, and B. Ribeiro, ''Concept drift awareness in Twitter streams,'' in *Proc. 13th Int. Conf. Mach. Learn. Appl.*, Dec. 2014, pp. 294–299.

[11] H. Zhuge, ''Interactive semantics,'' *Artif. Intell.*, vol. 174, no. 2, pp. 190–204, Feb. 2010.

[12] S. Jabeen, X. Gao, and P. Andreae, ''Semantic association computation: A comprehensive survey,'' *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 3849–3899, Aug. 2020, doi: 10.1007/s10462-019-09781-w.

[13] M. Strube and S. P. Ponzetto, ''Wikirelate! Computing semantic relatedness using wikipedia,'' in *Proc. 21st Nat. Conf. Artif. Intell. (AAAI)*, vol. 2, 2006, pp. 1419–1424.

[14] S. Hassan and R. Mihalcea, ''Semantic relatedness using salient semantic analysis,'' in *Proc. 25th AAAI Conf. Artif. Intell. (AAAI)*, 2011, pp. 884–889.

[15] D. Küçük and F. Can, ''Stance detection: A survey,'' *ACM Comput. Surv.*, vol. 53, no. 1, pp. 1–37, 2020.

[16] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, ''Stance and sentiment in tweets,'' *ACM Trans. Internet Technol.*, vol. 17, no. 3, pp. 1–23, Jul. 2017.

[17] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, ''SemEval-2016 task 6: Detecting stance in tweets,'' in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 31–41.

[18] K. Dey, R. Shrivastava, and S. Kaushik, ''Topical stance detection for Twitter: A two-phase LSTM model using attention,'' in *Proc. Eur. Conf. Inf. Retr.* Cham, Switzerland: Springer, 2018, pp. 529–536.

[19] Q. Sun, Z. Wang, S. Li, Q. Zhu, and G. Zhou, ''Stance detection via sentiment information and neural network model,'' *Frontiers Comput. Sci.*, vol. 13, no. 1, pp. 127–138, Feb. 2019.

[20] 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. 145944–145952, 2019.

[21] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, ''A survey on concept drift adaptation,'' *ACM Comput. Surv.*, vol. 46, no. 4, pp. 1–37, Apr. 2014.

[22] J. Costa, C. Silva, M. Antunes, and B. Ribeiro, ''Adaptive learning for dynamic environments: A comparative approach,'' *Eng. Appl. Artif. Intell.*, vol. 65, pp. 336–345, Oct. 2017.

[23] R. Elwell and R. Polikar, ''Incremental learning of concept drift in nonstationary environments,'' *IEEE Trans. Neural Netw.*, vol. 22, no. 10, pp. 1517–1531, Oct. 2011.

[24] R. Polikar, L. Upda, S. S. Upda, and V. Honavar, ''Learn++: An incremental learning algorithm for supervised neural networks,'' *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 31, no. 4, pp. 497–508, Nov. 2001.

[25] J. Costa, C. Silva, M. Antunes, and B. Ribeiro, ''The impact of longstanding messages in micro-blogging classification,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–8.

[26] J. Costa, C. Silva, M. Antunes, and B. Ribeiro, ''Boosting dynamic ensemble's performance in Twitter,'' *Neural Comput. Appl.*, vol. 32, pp. 1–13, Nov. 2019.

[27] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, ''Stream-based active learning for sentiment analysis in the financial domain,'' *Inf. Sci.*, vol. 285, pp. 181–203, Nov. 2014.

[28] V. Iosifidis, A. Oelschlager, and E. Ntoutsi, ''Sentiment classification over opinionated data streams through informed model adaptation,'' in *Proc. Int. Conf. Theory Pract. Digit. Libraries*. Cham, Switzerland: Springer, 2017, pp. 369–381.

[29] A. Go, R. Bhayani, and L. Huang, ''Twitter sentiment classification using distant supervision,'' Stanford, CA, USA, Project Rep. CS224N, 2009, pp. 1–6.

[30] G. Shan, S. Xu, L. Yang, S. Jia, and Y. Xiang, ''Learn#: A novel incremental learning method for text classification,'' *Expert Syst. Appl.*, vol. 147, Jun. 2020, Art. no. 113198.

[31] A. Cano and B. Krawczyk, ''Kappa updated ensemble for drifting data stream mining,'' *Mach. Learn.*, vol. 109, no. 1, pp. 175–218, Jan. 2020.

[32] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg, ''Feature hashing for large scale multitask learning,'' in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 1113–1120.

[33] M. Di Giovanni, M. Brambilla, S. Ceri, F. Daniel, and G. Ramponi, ''Content-based classification of political inclinations of Twitter users,'' in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 4321–4327.

[34] I. Irawaty, R. Andreswari, and D. Pramesti, ''Vectorizer comparison for sentiment analysis on social media youtube: A case study,'' in *Proc. 3rd Int. Conf. Comput. Informat. Eng. (ICIE)*, Sep. 2020, pp. 69–74.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2019, pp. 4171–4186.

[36] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, ''Distributed representations of words and phrases and their compositionality,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[37] J. Pennington, R. Socher, and C. Manning, ''Glove: Global vectors for word representation,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[38] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, ''Enriching word vectors with subword information,'' *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, U. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.

[40] N. Reimers and I. Gurevych, ''Sentence-BERT: Sentence embeddings using siamese BERT-networks,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2019, pp. 1–11. [Online]. Available: https://arxiv.org/abs/1908.10084

[41] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile, ''AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets,'' in *Proc. 6th Italian Conf. Comput. Linguistics (CLiC-it)*, vol. 2481, 2019, pp. 1–6.

ALESSIO BECHINI received the Laurea degree in electronics engineering and the Ph.D. degree in information engineering from the University of Pisa, in 1996 and 2003, respectively. He is currently a Researcher with the Department of Information Engineering, University of Pisa. His research interests span the fields of concurrent and distributed systems, enterprise information systems, data management and integration, service management, with particular interest in bioinformatics problems. Currently, his work has been focused on issues in data mining for big data. He has served as the TPC co-chair, the general co-chair and the program co-chair of ACM international conferences.

ALESSANDRO BONDIELLI received the M.Sc. degree in digital humanities from the University of Pisa, Italy, in 2016. He is currently pursuing the Ph.D. degree with the Smart Computing Program, a joint program of the Universities of Florence, Pisa and Siena. He is currently working with the Department of Information Engineering, University of Pisa. His research interests include the problem of fact-checking and fake news detection, and more broadly the evaluation and application of machine learning and neural language models for NLP and text mining tasks.

PIETRO DUCANGE received the M.Sc. degree in computer engineering and the Ph.D. degree in information engineering from the University of Pisa, Italy, in 2005 and 2009, respectively. He is currently an Associate Professor of information systems and technologies with the Department of Information Engineering, University of Pisa. He has coauthored over 70 articles in international journals and conference proceedings. He has been involved in a number of research and development projects in which data mining and computation intelligence algorithms have been successfully employed. His main research interests include explainable artificial intelligence, big data mining, social sensing, and sentiment analysis. He is a member of the Editorial Board of *Soft Computing* journal (Springer). He served as the General Co-Chair of the First International Workshop on Higher Education Learning Methodologies and Technologies Online (HELMeTO 2019), in June 2019. He is currently the Program Co-Chair of HELMeTO 2021 and the Chair of the HELMeTO Task Force Steering Committee.

FRANCESCO MARCELLONI (Member, IEEE) is currently a Full Professor with the Department of Information Engineering, University of Pisa, Italy. He has co-edited three volumes and four journal special issues. He is the coauthor of a book and published more than 230 articles in international journals, books, and conference proceedings. His main research interests include data mining for big data, sentiment analysis and opinion mining, multi-objective evolutionary algorithms, genetic fuzzy systems, fuzzy clustering algorithms, and data compression and aggregation in wireless sensor networks. He has coordinated various research projects funded by both public and private entities. He has been selected to receive the 2021 IEEE TFS Outstanding Paper Award. He has been the TPC co-chair, the general co-chair, and the tutorial chair of some international conferences and has held invited talks in a number of events. He currently serves as an Associate Editor for IEEE Transactions on Fuzzy Systems (IEEE), *Information Sciences* (Elsevier), and *Soft Computing* (Springer), and the editorial board of a number of other international journals.

ALESSANDRO RENDA received the M.Sc. degree in biomedical engineering from the University of Pisa, Italy, in 2017. He is currently pursuing the Ph.D. degree with the Smart Computing Program, jointly awarded by the Universities of Florence, Pisa and Siena. He is currently working as a Research Fellow with the Department of Information Engineering, University of Pisa. His research interests include machine learning algorithms for data streams, applications of deep learning methodologies, and affective computing.