Expert Systems With Applications 116 (2019) 209–226

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Monitoring the public opinion about the vaccination topic from tweets analysis

Eleonora D'Andrea[a], Pietro Ducange[b], Alessio Bechini[a], Alessandro Renda[a,c], Francesco Marcelloni[a,*]

[a] Dipartimento di Ingegneria dell'Informazione, University of Pisa, Largo Lucio Lazzarino 1, 56122 Pisa, Italy
[b] SMARTEST Research Center, eCampus University, Novedrate (CO), Italy
[c] University of Florence, Florence, Italy

ARTICLE INFO

Article history:
Received 22 January 2018
Revised 4 September 2018
Accepted 5 September 2018
Available online 6 September 2018

Keywords:
Opinion mining
Stance detection in tweets
Text mining
Tweet classification
Vaccines

ABSTRACT

The paper presents an intelligent system to automatically infer trends in the public opinion regarding the stance towards the vaccination topic: it enables the detection of significant opinion shifts, which can be possibly explained with the occurrence of specific social context-related events. The Italian setting has been taken as the reference use case. The source of information exploited by the system is represented by the collection of vaccine-related tweets, fetched from Twitter according to specific criteria; subsequently, tweets undergo a textual elaboration and a final classification to detect the expressed stance towards vaccination (i.e. in favor, not in favor, and neutral). In tuning the system, we tested multiple combinations of different text representations and classification approaches: the best accuracy was achieved by the scheme that adopts the bag-of-words, with stemmed *n*-grams as tokens, for text representation and the support vector machine model for the classification. By presenting the results of a monitoring campaign lasting 10 months, we show that the system may be used to track and monitor the public opinion about vaccination decision making, in a low-cost, real-time, and quick fashion. Finally, we also verified that the proposed scheme for continuous tweet classification does not seem to suffer particularly from concept drift, considering the time span of the monitoring campaign.

1. Introduction

Among existing social networks, over the past few years Twitter[1] has reached a widespread diffusion as a personal and handy information channel. Users typically broadcast information about personal or public real-life events, or simply express their thoughts, viewpoints or opinions on a given topic, product, service, event, etc., through a public Status Update Message called *tweet*. A tweet may contain also meta-information such as timestamp, location (in terms of GPS (Global Positioning System) coordinates, or user profile location), username, links, hashtags, emoticons, and mentions. Recently, it has been shown that tweets may represent a source of valuable information (Giachanou & Crestani, 2016): in fact, they are public and can be easily crawled with no privacy limitations, and their content can be analyzed with proper text/data mining techniques. Indeed, Twitter has been successfully used for the early detection of real-time events, such as traffic congestions and incidents (D'Andrea, Ducange, Lazzerini, & Marcelloni, 2015), earthquakes (Sakaki, Okazaki, & Matsuo, 2013), crime events (Gerber, 2014), or riots (Alsaedi, Burnap, & Rana, 2017).

However, analyzing tweets is more challenging than analyzing messages from other media like blogs, e-mails, etc. because of their limited length, which forces to operate at the sentence level rather than at the document level. Further, tweets are typically unstructured and irregular, may contain informal or abbreviated words (e.g., acronyms, hashtags), colloquial, idiomatic, or ironic expressions, misspellings or grammatical errors, making the conveyed information particularly noisy and fragmentary. This aspect is further worsened by the data sparsity phenomenon, i.e., a great amount of terms in a corpus occurs less than 10 times (Saif, Yulan, & Alani, 2012).

In the described setting, the extraction of meaningful information out of tweets resorts to *text mining* techniques, including methods from the fields of data mining, machine learning, statistics, and Natural Language Processing (NLP). Text mining refers to the process of automatic information mining also from unstructured natural language text. Text mining is hampered by the vagueness of natural language, due to the habit of people to make

* Corresponding author.
E-mail addresses: eleonora.dandrea@for.unipi.it (E. D'Andrea), pietro.ducange@uniecampus.it (P. Ducange), alessio.bechini@unipi.it (A. Bechini), alessandro.renda@unifi.it (A. Renda), francesco.marcelloni@unipi.it (F. Marcelloni).

[1] Twitter, www.twitter.com

https://doi.org/10.1016/j.eswa.2018.09.009
0957-4174/

frequent use of idioms, grammatical variations, slang expressions, or to assume an implicit context for a given word (Gupta, Gurpreet, & Lehal, 2009).

Sentiment analysis and opinion mining over the Web (especially in social networks, forums, microblogs, etc.) have recently and rapidly become emergent topics, mainly because of their potential in uncovering trends of the public opinion or social emotions (Rushdi Saleh, M., T., Montejo-Ráez, & Ureña-López, 2011; Mostafa, 2013; E. S. Tellez et al., 2017a, b). The terms "sentiment analysis" and "opinion mining" are currently used to refer to special subfields of the text mining research aimed at automatically determining, in natural language texts, the sentiment, or the opinion polarity (e.g., positive-biased or negative-biased) towards a certain *target* (Liu, 2010; Liu, 2015; Ribeiro, Araújo, Gonçalves, Gonçalves, & Benevenuto, 2016). They are considered challenging tasks, as even human experts may disagree about the sentiment associated with a text, e.g., because of the presence of ambiguity, sarcasm, or irony: such interpretation problems become even more difficult in the case of short and informal texts like tweets (Gokulakrishnan, Priyanthan, Ragavan, Prasath, & Perera, 2012; Valdivia, Luzíón, & Herrera, 2017).

Regarding a text mining activity performed on tweets, it is particularly important to identify its goals, so to correctly use the proper term to refer to it. According to the definitions recently provided in Mohammad, Kiritchenko, Sobhani, Zhu, and Cherry, (2016), *sentiment analysis* tasks aim to generally determine whether a piece of text is positive, negative, or neutral, or alternatively to determine the speaker's opinion along with the relative target (the entity towards which the opinion is expressed). On the other hand, in a *stance detection* task the target of interest is pre-chosen, and the opinion towards it must be determined, no matter whether it is explicitly mentioned or not in the text. Notably, stance detection corresponds to a classification problem, with the additional difficulty that in a text a stance polarity can be expressed with any sort of sentiment polarities towards disparate targets (Mohammad et al., 2016; Mohammad, Sobhani, & Kiritchenko, 2017).

The exploitation of web opinion mining services is becoming prominent in several contexts like marketing, politics, recommendation systems, healthcare, etc. (Cambria, 2016; Pandey, Singh, Rajpoot, & Saraswat, 2017; Ducange, Pecori, & Mezzina, 2017). Different approaches have been proposed to study the reactions of social network's users to major events, so to uncover, explain, or predict the events themselves. Typical examples are the prediction of movements in stock markets (Bollen, Mao, & Zeng, 2011) and the outcome of political elections (Zhou, Tao, Yong, & Yang, 2013). Further, the monitoring of public health concerns (e.g. regarding vaccines or disease outbreaks) is attracting more and more interest: in Ji, Chun, Wei, and Geller, (2015), the concern level is quantified on the basis of the number of negative shared tweets, and it is correlated over time to the occurrence of news, with the purpose of identifying in real-time the effect of news on public concerns.

The vaccination topic has become controversial in recent years, also because of the news of the alleged connection (stated in a research article later retracted) between autism and MMR vaccine, against measles, mumps, and rubella. The influence of stances spread in social networks over individual behaviors and sentiments has been statistically detected (Salathé, Vu, Khandelwal, & Hunter, 2013). Indeed, online discussion groups have arisen, influencing the opinion of the population over vaccination decision-making in several countries (Bello-Orgaz, Hernandez-Castro, & Camacho, 2017). Hence, in some cases, a drop in vaccination rates has been noticed, increasing the risk of re-emergence of eradicated diseases. For example, the Italian Ministry of Health has detected, in March 2017, an increase of 230% of the number of measles cases, and an overall drop in vaccination coverage.[2] It is thus clear that the automatic monitoring of information over social networks is of primary importance for inferring the stance of the public opinion towards this topic: e.g., countermeasures can be taken in case of spreading of fake news, or the effect of social events can be detected. In this scenario, a *social event* is a topic that rapidly attracts the attention of social networks' users in a certain time interval, by increasing the number of user reactions (Nguyen & Jung, 2017).

This paper proposes an intelligent system for real-time monitoring and analysis of the public opinion about the vaccination topic on the Twitter stream. The reference case study relates to the Italian setting, currently with about 7 millions active Twitter users, where the vaccination topic has caused harsh debates. The system introduced in this work employs text mining and machine learning techniques, appropriately tuned, adapted, and integrated so to build the overall intelligent system. We present an experimental study aimed to determine the most effective solution out of some different state-of-the-art approaches for text representation and classification. The chosen approach was integrated into the actual final system and used for the on-the-field real-time monitoring of the public opinion. Thus, first, we employ text mining techniques to solve a multi-class classification problem by assigning the correct class label (*in favor of vaccination, not in favor of vaccination*, and *neutral*) to tweets. Then, we inspect the trend of polarity of the public opinion of Italian Twitter users, in particular in correspondence with local peaks of the daily number of tweets related to the vaccination topic. These peaks generally correspond to events concerning vaccination that may have raised the public opinion interest. An example of a social event of this type is the planned projection (subsequently canceled) in the Senate of the Italian Republic of the controversial anti-vaccine documentary film "*Vaxxed: from cover-up to catastrophe*".[3] Among other such events we can also recall news about the vaccination drop rates in Italy, and the discussion about the introduction of vaccination-related laws.

The paper has the following structure. Section 2 reviews the state of the art and the related work about sentiment analysis, opinion mining applications and stance detection, with reference to Twitter's messages. Section 3 describes the proposed system for stance detection about the vaccination topic from tweets, referring to the Italian scenario. Section 4 compares the results obtained by the proposed system with recent stance detection approaches. Section 5 shows the results of a 10-months monitoring campaign, presenting the relative trend of public opinion about vaccination, spotting out the influence of particular events and analyzing possible concept drift. Finally, Section 6 draws concluding remarks.

2. State of the art and related work

Since its early introduction in the research community in Web search and information retrieval (Dave, Lawrence, & Pennock, 2003), the term "Opinion Mining" has been used to emphasize the uncovering of judgments towards targets of interest in text analysis, but often it is assumed to cover a wider range of types of text analysis. Within this paper, according to Giachanou and Crestani, (2016), *sentiment analysis* is broadly intended to be the study of opinion, sentiment, mood, and emotion expressed in texts. Sentiment analysis is a broad research area having several sub-tasks. In its most general form, it can deal with detecting sentiment polarity, e.g., positive, neutral, and negative, or with identifying specific emotions, e.g., hate, anger, joy, and sadness. A specific task is called *subjectivity detection* and consists in discriminating between objective (neutral) and subjective (opinionated)

[2] Vaccination data in Italy, http://www.repubblica.it/salute/prevenzione/2017/05/12/news/i_vaccini_in_italia_i_dati-165262703/, (accessed 16 October 2017).

[3] Vaxxed: from cover-up to catastrophe, http://vaxxedthemovie.com/.

texts. This can be even more challenging than polarity classification, e.g., in the case of a news article citing people's opinions or vice versa (Liu, 2010; Cambria, 2016). Moreover, *stance detection* is the task aimed to determine the polarity of the opinion (in favor, against, or neutral) expressed in a text towards a given target entity, e.g., a topic, a product, a service, a person (Mohammad et al., 2016; Mohammad et al., 2017). The distinctive trait of stance detection, in the comparison with general sentiment analysis, is that the opinion polarity is detected towards a predefined target entity that may also be not explicitly mentioned in the text.

Sentiment analysis and stance detection can resort to different classes of approaches: machine learning, lexicon-based, and hybrid approaches (Medhat, Hassan, & Korashy, 2014). As regards texts, they can be passed to classification systems in different formal representations that should be able to encode the contents with the required precision. Text can be represented as vector of numbers considering different schemes such as standard bag-of-words (BOW) (D'Andrea et al., 2015), word embeddings (Tang et al., 2014) and a combination of BOW and other features (for instance part-of-speech and word embeddings) (Mohammad et al., 2017). Machine learning employs supervised or unsupervised techniques to automatically extract knowledge directly from texts. Lexicon-based approaches rely instead on a predefined sentiment lexicon (e.g., WordNet (Miller, 1995), SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010), SenticNet (Cambria, Havasi, & Hussain, 2012)), a collection of words from the considered language, annotated with the relative sentiment polarities and strength values. The lexicon is then used along with statistical or semantic methods to perform sentiment analysis. Lexicon-based approaches are best suited for general boundless contexts (i.e., without topic), with well-formed and grammatically correct texts. On the contrary, they behave worse in bounded contexts (i.e., concerning a certain topic) or when an informal language is used, e.g., in social networks, due to the absence of context-related words in the lexicon. Further, in social networks like Twitter, the language undergoes continuous changes. In fact, a new invented hashtag or word can quickly gain popularity. It is thus clear that lexicon-based approaches, relying on predefined dictionaries, struggle to cope with such a dynamic setting. Supervised machine-learning approaches make use of sets of labelled texts, to be exploited for model training. Among the wide assortment of machine learning methods, deep learning approaches have recently become particularly popular also in the field of sentiment analysis (Zhang, Wang, & Liu, 2018). They exploit multiple layers of nonlinear processing units for feature extraction and transformation. Lower layers near to the inputs learn simple features, whereas higher layers learn more complex features thanks to the representation produced by lower layers.

In the following, we recall a few recent works for each class of approaches, focusing on the analysis of Twitter messages. As regards lexicon-based approaches, we mention two works Basile and Nissim, (2013) and Ortega Fonseca, and Montoyo, (2013). In the former, a tool based only on a polarity lexicon is applied on a topic-specific and a general dataset related to Italian tweets, both considering three classes. In the latter, the authors employ a three-step technique including text preprocessing, polarity detection, and rule-based classification based on WordNet and SentiWordnet lexicons. As it refers to machine learning solutions, supervised learning is the dominant approach in the literature. In Chien and Tseng, (2011) an SVM is used for the evaluation of the quality of information in product reviews. For the support to decision-making in marketing, it has been proposed a framework for summarization and SVM classification of opinions on Twitter (Li & Li, 2013). A Naive Bayes (NB) model on unigram features has been chosen in a system for real-time analysis of tweets related to 2012 U.S. elections (Wang, Can, Kazemzadeh, Bar, & Narayanan, 2012), with the aim of inferring the public sentiment toward candidates. A different text representation is proposed in Aisopos, Papadakis, and Varvarigou, (2011), employing *n*-gram graphs with distance-weighted edges, and making use of two classifiers (MNB and C4.5 decision tree) to perform both a two-way and a three-way classification of tweets posted in a time span of seven months. In Valdivia et al. (2017), the classification results, obtained starting from a baseline model comprising text elaboration and SVM classification, are improved by applying a majority vote to several methods for filtering out neutral reviews. In a recent work (E. S. Tellez et al., 2017a) it has been shown that the use of some traditional machine learning approaches (in the specific case, text elaboration, BOW vector representation, and an SVM classifier) can lead to good results in polarity classification of tweets.

Deep learning techniques have been intensely employed for text representation and classification. In particular, several techniques leverage *word embeddings*: they consist in a dense, continuous, representation of words in a low-dimensional space. The advantage of this vector representation is the possible encoding of general semantic and syntactic relationships between words, mapping similar words in close points of the representation space. Several unsupervised learning methods have been proposed to generate word vector representations from raw text. Well-known word embedding generation models like *Word2Vec* (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), *Glove* (Pennington, Socher, & Manning, 2014) and *Fast-Text* (Bojanowski, Grave, Joulin, & Mikolov, 2017) typically require text corpora containing billions of words to be trained. *Word2vec* learns representations by training a shallow neural network to reconstruct the linguistic contexts of words from target words or vice versa. *FastText* is a variant of *Word2vec* that breaks words into several *n*-grams (sub-words). In GloVe the training is performed on aggregated global word-word co-occurrence statistics from a corpus. Recently, several works either exploits pre-trained publicly available vectors, or train the models on specific corpora. Among deep learning architectures used in this field, Long Short-Term Memory (LSTM) Networks (Hochreiter & Schmidhuber, 1997) and Convolutional Neural Networks (CNNs) (LeCun, Bengio, & Hinton, 2015) represent state-of-the-art models. The authors in Cliché, (2017) pre-trained three well-known unsupervised learning models, i.e., *FastText, GloVe*, and *Word2vec*, using 100 million unlabeled tweets. To enrich the word representation with polarity information they fine-tuned word vectors using a distant supervision with a dataset of 5 million positive and 5 million negative tweets. Finally, they employed the fine-tuned word vectors to initialize a LSTM and a CNN model. An ensemble of such models achieved the best absolute performance in the *SemEval-2017 International Workshop on Semantic Evaluation, task 4 (Sentiment analysis in Twitter)* (Rosenthal, Noura, & Preslav, 2017). The authors in Xiong, Hailian, Weiting, and Donghong, (2018) learn sentiment-specific word embedding by exploiting both lexicon and distant supervised information. They fed several neural networks with a word representation combining word-level sentiment (i.e., lexicon information) and tweet-level sentiment (e.g. hashtag and emoticon) to obtain a multi-level sentiment-enriched word embeddings. Recently, in the framework of stance detection in Twitter, the authors of Dey, Ritvik, and Saroj, (2018), discussed a two-phases text classification scheme. In the first phase, a given tweet is classified as neutral or subjective with respect to the given topic. In the second phase, the stance of a subjective tweet is classified as in a favor or against towards the topic. In both phases LSTM networks are adopted as classification models. In general, it is important to notice that the accuracy of Deep Learning methods is typically achieved by resorting to massive training sets to support the learning phase.

Among hybrid approaches, we can mention some recent works. In Ortigosa and Carro, (2014), the authors combine lexical-based

techniques and SVM classification to perform sentimental analysis on Spanish Facebook messages. In Agarwal, Xie, Vovsha, Rambow, and Passonneau, (2011), the authors employ a combination of unigrams and a selected set of features, on a manually annotated three-class dataset of English tweets. In Castellucci, Croce, Cao, and Basili, (2016), for a binary classification, texts are represented by using different categories of features, combining a BOW representation of the text, word-embedding semantic attributes, polarity information, along with other attributes. In Mohammad et al. (2017), the authors, by employing word embedding features in addition to *n*-grams, improve the accuracy of an SVM classifier for stance detection in tweets.

Taking into account vaccination, which is the target topic for our research, we can note that several works in the literature deal with the healthcare topic, including both texts and social network messages. In Botsis, Nguyen, Woo, Markatou, and Ball, (2011), the authors propose a text classification of reports collected from the U.S. Vaccine Adverse Event Reporting System related to the H1N1 vaccine, by employing SVMs. Chew, and Eysenbach, (2010) analyze the content of tweets related to the H1N1 outbreak to determine the kind of information exchanged by social media users. In Salathé et al. (2013), the authors employ a hybrid approach based on NB and maximum entropy classifiers to classify tweets as negative, positive and neutral with respect to the user's vaccination intent against H1N1. In Du, Xu, Song, and Tao, (2017), an SVM classifier is employed to assess the human papillomavirus (HPV) vaccination sentiment trend from tweets.

In this paper, we propose an intelligent system for monitoring public opinion regarding the stance towards the vaccination topic, with specific reference to the Italian case. Tweets are classified adopting the BOW representation of the texts, using *n*-grams as tokens, followed by an SVM model. Indeed, we experimentally showed that, for the specific context of stance detection on Twitter, the adopted text classification scheme outperforms recent state-of-the-art approaches, including text classification models based on deep-learning. In addition to the elaboration and classification of tweets, the system let us check in real-time increments of interest and stance changes of the public opinion, so that we can off-line associate them to context-related events of possible influence; this type of analysis is not present in works mentioned above. Moreover, we also verified that, over the time span of the real-time monitoring campaign, the system is characterized by a low classification concept drift.

The word embedding process deserves further discussion, as its proper use in sentiment analysis is not straightforward (Uysal & Yi Lu, 2017). Two recent works Mohammad et al. (2017) and Uysal and Yi Lu, (2017) have shown that training a word embedding model on a "background" domain-related corpus is beneficial for the task of stance or sentiment classification of tweets: both the background corpus and the classification dataset consisted of tweets collected in the same time window. However, in the present work, we adopted three publicly available word embeddings, pre-trained on the Wikipedia corpus. The rationale for this choice is twofold: (i) we do not have at our disposal a sufficiently large corpus of domain-related tweets in Italian, i.e. collected according to the same criteria used for the vaccination dataset, to emulate the training procedure presented in Mohammad et al. (2017); (ii) some recent works have successfully adopted pre-trained word embeddings for text classification (Wang et al., 2016; Uysal & Yi Lu, 2017).

3. The architecture of the proposed system for stance detection

In the following, we present the system to perform stance detection on Twitter, with reference to the vaccination topic in Italy. The system consists of three modules (see Fig. 1). The first module "Collection of tweets" (see Section 3.1) regards the collection of vaccine-related tweets from Twitter. The second module "Text representation" (see Section 3.2) applies a sequence of text elaboration steps to preprocessed tweets in order to transform them in numeric vectors. In the third module "Text classification and trend analysis" (see Section 3.3), the public stance towards the vaccination topic is studied. More in detail, first, an appropriate class label (namely, *in favor of vaccination, not in favor of vaccination*, and *neutral*, i.e., neither in favor nor not in favor of vaccination) is assigned to each tweet using a supervised learning model. Then, the classified tweets are analyzed to identify the trend over time of the public stance in Twitter, with particular reference to local peaks of the daily number of tweets concerning vaccination. We observed that local peaks of the daily number of tweets are related to particular events concerning vaccination (discussions in Parliament, approval of law establishing vaccination requirements, etc.). We have verified the presence of these events by analysing news on the vaccination topic. The following sub-sections describe in detail the steps performed in the three modules and the supervised learning stage (see Section 3.4). Although the focus of the paper is on Italian vaccine-related tweets, the proposed system is general and easily adaptable to any other topic or language.

3.1. Collection of tweets

The first module of the system consists of two main steps, i.e., *fetch and cleaning* of tweets, and *preprocessing* of tweets.

(1) *Fetch and cleaning*. In this step, tweets are fetched according to some search criteria (e.g., keywords, time and date of posting, location of posting, hashtags). Although it is possible to resort to customizable tools designed for this purpose (Bechini, Gazzè, Marchetti, & Tesconi, 2016), our main requirement was to have a full coverage of the relevant tweets (and this is not guaranteed by using the plain Twitter APIs). We have reached our goal by employing the Java library GetOldTweets[4], which performs HTTP GET requests to directly collect tweets meeting the provided search criteria: In practice, it is able to carry out the same researches that may be performed using the Twitter Search web page.[5]

The downloaded set of raw tweets is reduced with the aim of discarding:

- duplicate tweets, i.e., tweets having same tweet id, possibly fetched in different searches;
- tweets written in other languages than the target one (Italian): this may occur because of the presence of keywords/hashtags with the same spelling in different languages (this has been accomplished using the Apache Tika[6] library for Java).

Regarding retweets (i.e., other users' tweets simply re-shared), we decided to maintain them in the dataset, as we think that the retweeting action, in this context, is a way of supporting/sharing the same opinion of another user.

(1) *Preprocessing*. In this second step, tweets are preprocessed by applying a Regular Expression (RE) filter, in order to extract only the text of each tweet, and remove all useless meta-information. In fact, each fetched raw tweet contains the tweet id, the user id, the timestamp, the location (if provided), a retweet flag, and the tweet's content. The tweet's content may include the user's text, hashtag(s), link(s), and mention(s). More

[4] GetOldTweets library available at https://github.com/Jefferson-Henrique/GetOldTweets-java/, (accessed 2017/06/30).

[5] https://twitter.com/search-home

[6] Apache, Tika https://tika.apache.org/, (accessed 2017/06/30).

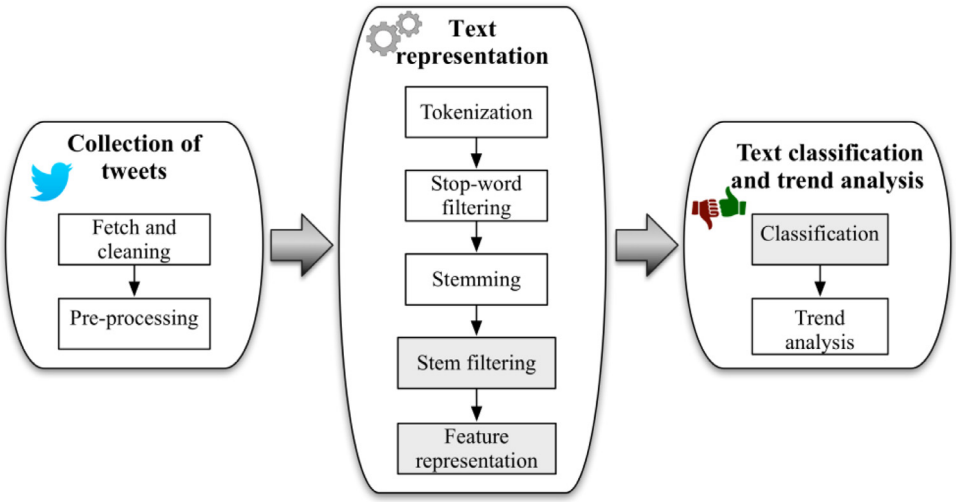

Fig. 1. Modules of the proposed system (grey blocks require information from the preliminary supervised learning stage).

in detail, using an RE filter, the tweet id, the user id, the location, and the retweet flag are discarded. The timestamp is temporarily discarded for the purposes of text mining elaboration, but it will be reconsidered for the analysis of the public opinion trend over time. From the tweet's content we discard: links, mentions, numbers and special characters (e.g., punctuation marks, brackets, slashes, quotes, etc.). Hashtags are not completely discarded, they are instead reduced to words (by eliminating the hash (#) symbol), so as not to lose relevant information. In fact, a common way of writing of Twitter's users is to use hashtags in sentences, in place of normal words. Finally, a case-folding operation is applied to the texts, in order to convert all characters to lower case form.

Hence, each tweet is represented as a sequence of characters. We denoted the j th tweet of the set as $tweet_j$, with $j = 1,\ldots, N$, where N is the total number of tweets considered in the subsequent steps.

3.2. Text representation

As discussed in Section 2, several methods have been proposed in the specialized literature for text representation and classification. To identify the most suitable scheme in our specific case, we experimented three categories of methods:

1. BOW text representation followed by classical machine learning algorithms for classification (D'Andrea et al., 2015).
2. A combination of BOW and word embeddings for text representation followed by classical machine learning algorithms for classification (Mohammad et al., 2017).
3. Deep learning-based approaches for text elaboration and classification (Cliché et al., 2017).

It is important to underline that the data preparation steps represent a crucial issue for the success of the overall system: this has been experimentally assessed also in the context of multilingual emotion classification (Balahur & Turchi, 2014; Becker, Moreira, & dos Santos, 2017). Thus, we cannot claim that the choices explored and selected through our experimentations are necessarily the optimal ones, but indeed they have found to deliver very good performances.

Similar to the work discussed in Balahur and Turchi, (2014) and Becker et al. (2017), we carried out an intensive experimental setup analysis for the identification of the parameters of the BOW text representation module. In particular, we considered different methods for the tokenization (word tokenizer, alphabetic tokenizer, N-gram with different values of N) and different strategies for the feature representation (binary approach, IDF and TF-IDF). For the sake of brevity, we show just the best combination that we have obtained.

In Section 4, we discuss in detail the results achieved by eleven methods selected from the three selected categories. For the specific text classification contest, the BOW text representation followed by an SVM classification model achieves the best results. Thus, we adopt this scheme for the text representation and classification in our system.

As regards the *text representation* module, its main steps are described in detail in D'Andrea et al. (2015). The main aim of the module is to transform the set of strings, representing the stream of tweets, into a set of numeric vectors, by eliminating noise and extracting useful information. In the following, we briefly recall the sequence of steps applied to the tweets, whereas Fig. 2 shows how a sample (vaccine-related) tweet is transformed as it undergoes the different text elaboration steps. The elaboration is carried out by employing the Java API for Weka (Waikato Environment for Knowledge Analysis) (Hall et al., 2009). The text elaboration steps, namely, *tokenization, stop-word filtering, stemming, stem filtering*, and *feature representation*, are described in detail in the following.

(1) *Tokenization* consists in transforming a stream of characters into a stream of processing units, called *tokens*, e.g., words, phrases. Thus, during this step, by choosing n-grams as tokens (with n up to 2) and after removing punctuation marks and special symbols (e.g., accents, hyphens), each tweet is converted into a set of tokens, according to the BOW representation. At the end of this step, the j th tokenized tweet, $tweet_j^T = \{t_{j1}^T, \ldots, t_{jh}^T, \ldots, t_{jH_j}^T\}$, is represented as the sequence of n-grams contained in it, where t_{jh}^T is the h th token, and H_j is the number of tokens in $tweet_j^T$.

(2) *Stop-word filtering* consists in removing *stop-words*, i.e., words providing little or no useful information to the text analysis: these words can hence be considered as noise. Common stop-words include articles, conjunctions, prepositions, pronouns, etc. Other stop-words are those typically appearing very often in sentences of the considered language (language-specific stop-words), or in the particular context analyzed (domain-specific stop-words). In this work, we employ a reduced version of the stop-word list for the

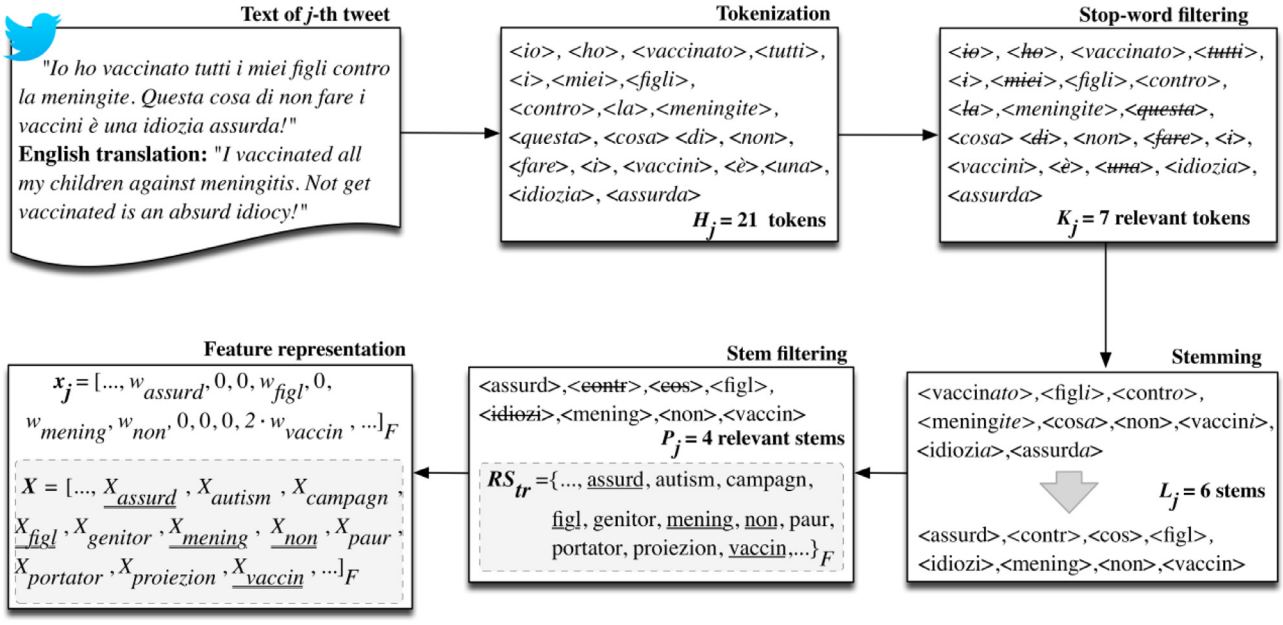

Fig. 2. Steps of the text elaboration (second module) applied to a sample tweet.

Italian language, available at the Snowball Tartarus website.[7] More precisely, we remove from the stop-word list: (i) all the verbal forms, and (ii) the words "*non*" (not) and "*contro*" (against), as we experimentally found that such words become important for opinion mining and sentiment analysis, thus they should be part of the text analysis process. At the end of this step, each tweet is cleaned from stop-words and thus reduced to a sequence of relevant tokens, $tweet_j^{SW} = \{t_{j1}^{SW}, \ldots, t_{jk}^{SW}, \ldots, t_{jK_j}^{SW}\}$, where t_{jk}^{SW} is the k th token and K_j, $K_j \leq H_j$, is the number of relevant tokens in $tweet_j^{SW}$.

(3) *Stemming* is a process typically required in dealing with fusional languages, like English and Italian (in our specific case). It consists of reducing each token (i.e., word) to its *stem* or root form, so to group words having closely related semantics. In this work, we exploit the Snowball Tartarus stemmer for the Italian language[8], based on the Porter's algorithm (Porter, 1980). Hence, at the end of this step each tweet is represented as a sequence of stems, $tweet_j^S = \{t_{j1}^S, \ldots, t_{jl}^S, \ldots, t_{jL_j}^S\}$, where t_{jl}^S is the l th stem and L_j, $L_j \leq H_j$, is the number of stems in $tweet_j^S$.

(4) *Stem filtering* consists in filtering out the stems, which are not considered relevant in the training dataset for the supervised learning stage (described in detail in Section 3.4). Thus, each tweet is cleaned from stems not belonging to the set of relevant stems RS_{tr}, and is represented as a sequence of relevant stems, $tweet_j^{SF} = \{t_{j1}^{SF}, \ldots, t_{jp}^{SF}, \ldots, t_{jP_j}^{SF}\}$, where t_{jp}^{SF} is the p th relevant stem in $tweet_j^{SF}$, and P_j, with $P_j \leq L_j$, is the total number of relevant stems in $tweet_j^{SF}$. Let F be the number of relevant stems identified in the training dataset.

(5) *Feature representation* consists in building, for each tweet, the corresponding vector of numeric features, i.e., $\boldsymbol{X} = \{X_1, \ldots, X_f, \ldots, X_F\}$, in order to represent all the tweets in the same F-dimensional feature space. The set of F features corresponds to the set $RS_{tr} = \{\hat{s}_1, \ldots, \hat{s}_f, \ldots, \hat{s}_F\}$ of relevant stems. Each tweet is thus associated with a vector of numeric features $tweet_j^{FR} = \boldsymbol{x_j} = \{x_{j1}, \ldots, x_{jf}, \ldots, x_{jF}\}$, where each element x_{jf} is set as follows:

$$x_{jf} = \begin{cases} TF_{jf} \cdot w_f & \text{if relevant stem } \hat{s}_f \text{ is in } tweet_j^{SF} \\ 0 & \text{otherwise} \end{cases}.$$

In the above equation, TF_{jf} is the term frequency of the relevant stem $\hat{s}_f$ in the j th tweet, whereas weight w_f expresses the importance in the training dataset of the f th feature, namely, the f th relevant stem $\hat{s}_f$, and is computed during the supervised learning stage (the computation of this weight is discussed in Section 3.4).

3.3. The text classification and trend analysis stage

As regards the text classification and trend analysis module, two steps are performed, i.e., *classification* and *trend analysis*.

(1) *Classification*. The fetched tweets are classified using a supervised classification model, namely an SVM classifier, previously trained during the supervised learning stage (see Section 3.4). The model assigns to each tweet, now represented with $\boldsymbol{x_j}$, a possible class label belonging to C, $C = \{C_1, \ldots, C_r, \ldots, C_R\}$, with R being the number of classes considered (in this work we have $R = 3$).

(2) *Trend analysis*. The classified tweets are analyzed over time, in order to infer changes (offline or even in real-time) in the public opinion about the vaccination topic. Such changes (e.g., spikes in the total number of tweets) may appear in correspondence with social known or unknown context-related events.

3.4. The supervised learning stage

As stated previously, a supervised learning stage is required before performing some of the steps of the second and third modules of the system, namely, *stem filtering*, *feature representation*, and *classification*.

To this aim, we need a collection of N_{tr} labelled tweets as training set. The training tweets were fetched using a set of context-related keywords as search criteria, and were preprocessed, as described in Section 3.1. Then, each tweet of the training set went

[7] Snowball stop-words list (Italian), http://snowball.tartarus.org/algorithms/italian/stop.txt, (last accessed 2018/07/16).

[8] Snowball Stemmer (Italian), http://snowball.tartarus.org/algorithms/italian/stemmer.html, (last accessed 2018/07/16).

through the following text mining steps: *tokenization, stop-word filtering*, and *stemming*. Finally, the complete set of stems CS_{tr} was extracted from the N_{tr} training tweets:

$$CS_{tr} = \{s, \ldots, s_q, \ldots, s_Q\} = \bigcup_{j=1}^{N_{tr}} tweet_j^S,$$

CS_{tr} is the union of Q stems extracted from the set of training tweets after the stemming step.

The importance of each stem s_q in CS_{tr} is represented by means of a weight w_q, computed as the Inverse Document Frequency (IDF) index (Salton & Buckley, 1988) as $IDF_q = \ln(N_{tr}/N_q)$, where N_q is the number of tweets containing stem s_q.

Then, each training tweet is represented as a vector of features in $\mathbb{R}^Q$, i.e, $\boldsymbol{x_t} = \{x_{t1}, \ldots, x_{tq}, \ldots, x_{tQ}\}$, where

$$x_{tq} = \begin{cases} TF_{tq} \cdot w_q \ if \ tweet_t^S \ contains \ stem \ s_q \\ 0 \ otherwise \end{cases},$$

with TF_{tq} being the term frequency (TF) of stem q th in the t th training tweet. Thus, we employ the well-known TF-IDF index.

Finally, in order to select the set of relevant stems $\hat{s}_f$ in RS_{tr}, a feature selection algorithm was applied as follows. First, the quality of each stem s_q was evaluated by means of the well-known Information Gain (IG) value (Patil & Atique, 2013) between feature S_q (corresponding to stem s_q) and the possible class labels in C. IG is computed as $IG(C|S_q) = H(C) - H(C|S_q)$, where $H(C)$ represents the entropy of C, and $H(C|S_q)$ represents the entropy of C, after the observation of S_q. Then, the stems are ranked in descending order and F stems, with $F \leq Q$, are selected among these. We experimented with different values for F. Consequently, each feature vector is reduced to the representation in $\mathbb{R}^F$ (discussed in step 5 of Section 3.2).

Lastly, the supervised classification models are trained by setting the values of their parameters. In our system, we adopted an SVM classification model. The SVM has been used successfully for text classification in the literature (D'Andrea et al., 2015; Mohammad et al., 2017; E. S. Tellez et al., 2017a, b). SVMs are discriminative classification algorithms based on a separating hyper-plane according to which new samples can be classified. The best hyperplane is the one with the largest minimum distance from the training samples and is computed based on the support vectors (i.e., samples of the training set). The SVM classifier employed in this work is the implementation described in Keerthi, Shevade, Bhattacharyya, and Murthy, (2001).

4. Comparing stance detection approaches

In this Section, we compare the results achieved by recent state-of-the art approaches for stance detection. Obviously, in the experimental comparison, we consider the dataset extracted for the specific context of stance detection regarding vaccination in Italy. First, we describe how we generate the adopted dataset. Then, we show the results achieved by the different methods selected for our experimental comparison campaign. We recall that this campaign was carried out for identifying the most suitable scheme, to embed in our stance detection system, for text representation and classification. It is intended that the results depend also on the data preparation steps chosen for our system, out of the wide range of possible ones, as underlined in Balahur and Turchi, (2014) and Becker et al. (2017).

4.1. Data set extraction

In order to compare the different text representation and classification approaches, we needed to collect and label a set of vaccine-related tweets. Thus, we collected tweets by using, as search criteria, the date of posting, and a set of vaccine-related keywords, chosen based on a preliminary analysis consisting of: i) the reading of newspaper articles about vaccines and vaccine-related events, and ii) interviews with medical experts. As date of posting, we considered a time span of five months, from September 1st, 2016, to January 31st, 2017. We chose this time span as the controversy about vaccines in Italy deeply increased in this period, according to Google Trend data.[9] The keywords employed refer to different sub-contexts: i) the vaccination topic itself; ii) diseases possibly caused by negative effects attributed to vaccines; and iii) vaccine-preventable diseases. Further, we also took into account three widely used hashtags, namely, *#libertadiscelta* (hashtag for "freedom of choice"), *#iovaccino* (hashtag for "I vaccinate"), and *#novaccino* (hashtag for "no vaccine"). Based on these criteria, we took into account 38 keywords (including synonyms, or singular/plural variations of the keywords). The set of keywords employed is listed in Table 1. We wish to point out that in a few cases the keywords were used in combination (i.e., logic and) with the keywords "vaccino", "vaccini", in order to fetch only tweets related to the vaccination context. We have read a large number of the tweets collected by using this procedure and we can confirm that almost the totality of the tweets are related to vaccination.

Next, we cleaned the set of tweets, by removing duplicated tweets and non-Italian tweets. In fact, some keywords, e.g., "autismo", are spelled in the same way also in Spanish, or some keywords, e.g., "big pharma", "vaxxed", may lead to fetch English tweets.

Finally, we randomly selected and manually labelled $N_{tr} = 693$ training tweets (about 3% of the fetched tweets) to employ in the learning stage. The training dataset consisted of 219 tweets of class *not in favor of vaccination,* i.e., tweets expressing a negative opinion about vaccination, 255 tweets of class *in favor of vaccination*, i.e., tweets expressing a positive opinion about vaccination, and 219 tweets of class *neutral*. The class *neutral* may include news tweets about people dead or fell ill due to vaccines or to missed vaccinations, neutral opinion tweets, and off-topic tweets containing the keywords selected (e.g., tweets related to the vaccination of pets). Fig. 3 shows an example of manually labelled training tweets. We chose to manually label tweets despite being an expensive and tedious activity. Recently, an emerging common practice to automatically label tweets exploits the kind of emoticon associated with the tweet (Gokulakrishnan et al., 2012; Aisopos et al., 2011; Agarwal et al., 2011). However, we did not take into account this approach as it presents a few problems: (i) the emoticon is often absent (especially in tweets concerning health topics); (ii) emoticons are rarely associated with tweets containing negative sentiments or contrary stances (Park, Barash, Fink, & Cha, 2013); (iii) some emoticons, e.g., those for "sad" or "happy", may actually help us to distinguish between positive-sentiment tweets and negative-sentiment tweets, while other emoticons, e.g., "surprise", may lead to a wrong labelling, as they are not clearly associated with a specific sentiment. In addition, stance detection is different from sentiment analysis. E.g., the tweet "Il film Vaxxed è molto interessante, felice che venga diffuso!:-) " ("The film Vaxxed is interesting, happy it is distributed!:-) ") refers to a *positive* sentiment (manifested also through the emoticon), but expresses a *not in favor* stance about the vaccination topic.

4.2. Experimental comparisons

The experiments were performed using a 10-fold stratified cross validation (CV) procedure. Being 693 the number of labelled tweets, at each iteration, the classification model is trained on

[9] Google Trends, https://trends.google.it/trends/.

Table 1
Set of keywords (with corresponding English translation) used to fetch tweets (please note that, in some cases, the keywords differ from each other only for the singular or plural form).

Context	Italian keyword (English translation)
Vaccination topic	"complotto vaccini" (vaccines conspiracy); "copertura vaccinale" (vaccination coverage); "vaccini", "vaccino" (vaccine(s)); "big pharma"; "rischio vaccinale", "rischi vaccinali" (vaccine risk(s)); "vaxxed"; "trivalente" (trivalent); "esavalente" (hexavalent); "vaccinati", "vaccinata", "vaccinato", "vaccinate" (vaxxed); "quadrivalente" (quadrivalent); "vaccinazione", "vaccinazioni" (vaccination(s)); "libertà vaccinale" (vaccination freedom); "obiezione vaccinale" (vaccination objection); "età vaccinale" (vaccination age); "cocktail vaccinale" (vaccination cocktail); "controindicazioni vaccinali" (vaccine contraindications)
Negative effects attributed to vaccines	"paralisi flaccida" (flaccid paralysis); "autismo" (autism); "malattie autoimmuni" (autoimmune diseases); "evento avverso", "eventi avversi" (adverse event(s));
Vaccine-preventable diseases	"meningite" (meningitis), "morbillo" (measles); "rosolia" (rubella); "parotite" (mumps); "pertosse" (whooping cough); "poliomelite" (polio); "varicella" (varicella); "MPR" (italian acronym for measles, mumps, rubella);
Hashtags	#novaccino (hashtag for "no vaccine"); #iovaccino (hashtag for "I vaccinate"); #libertadiscelta (hashtag for "freedom of choice")

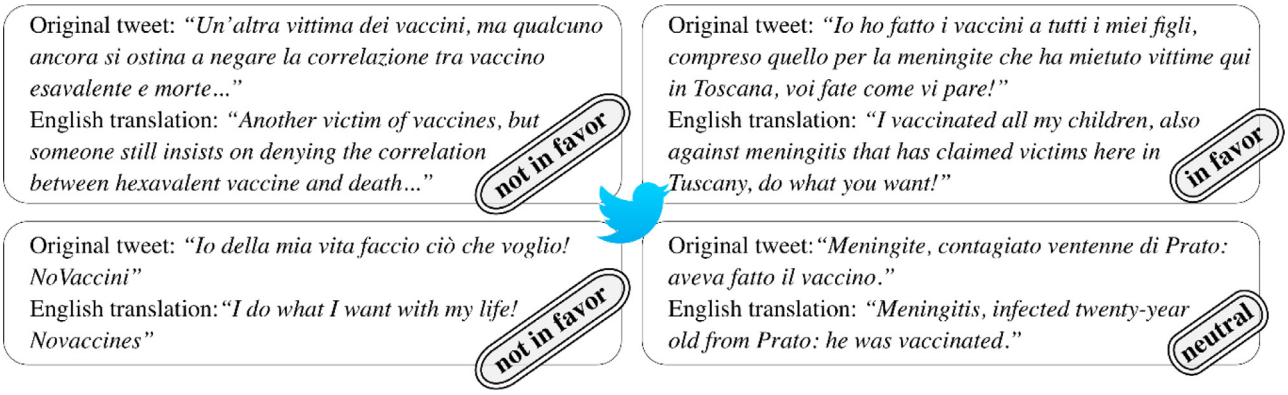

Fig. 3. Some example of manually labelled tweets.

about 624 tweets, and tested on about 69 tweets. We repeated the 10-fold stratified CV for two times, using two different seed values to randomly partition the data into folds. We recall that, for each fold, we consider a specific training set, which is used for learning the parameters for the text representation and the classification model. Indeed, all the compared schemes include a first phase for transforming texts into vectors of features and then a phase for the classification.

The first scheme that we experimented adopts the BOW for the text representation and classical machine learning classification models. We tried different BOW schemes, including different tokenization methods and the presence or the absence of the stemming stage. We also experimented with the following classification models: C4.5 decision tree (Quinlan, 1993), Naïve Bayesian (NB) (John & Langley, 1995), Multinomial NB (MNB) (Mccallum & Nigam, 1998), Random Forest (RF) (Breiman, 2001), Simple Logistic (SL) (Landwehr, Hall, & Frank, 2005), and SVM (Platt, 1999). The scheme discussed in Section 3 is the one that achieved the best results, thus it was considered in the comparison with the other approaches. During the training of the models, we identified on average $Q = 9529$ features, reduced to $F = 2000$ features after the feature selection step. In the following, we denote the two schemes as BOW + SVM_ALL and BOW + SVM_2000, respectively.

The second scheme taken into consideration is an extension of the previous one: the BOW representation, using *n*-grans as tokenization method, was extended by using word embeddings as extra features. We took inspiration from a similar approach that was recently adopted in Mohammad et al. (2017), where authors also compared a number of state-of-the art schemes for stance detection. In particular, the authors extended the BOW representation with word embeddings, achieving the best results in their experimental comparison. The word embedding model was obtained by means of a training stage with a domain related corpus containing tweets in English. As stated in Yang, Craig, and Iadh, (2017), in order to train a new word embedding model, millions of tweets may be necessary. Since we do not have such a huge amount of domain related tweets in Italian, despite of the work in Mohammad et al. (2017), we adopted three publicly available word embeddings, pre-trained on the Wikipedia corpus. We considered the pre-trained vectors from Fast-Text,[10] Glove[11] and Word2Vec.[12] We verified that some recent works have successfully adopted pre-trained word embeddings for text classification (Wang et al., 2016; Uysal & Yi Lu, 2017). Thus, we experimented three schemes that we denoted as BOW + FAST-TEXT + SVM and BOW + GLOVE + SVM and BOW + W2V + SVM, respectively. In each of the schemes, the dimension of word embedding space is equal to 300, thus the total number of adopted features is equal to 9829.

Finally, we also experimented two popular schemes for text representation and classification based on deep-learning. Both schemes adopt word embeddings for text representation. Convolutional Neural Networks (CNN) and Long Short-Term Memory Network (LSTM) are employed for the classification stage. The adopted models are inspired to the network architectures presented in Cliché, (2017). Albeit with different parametrizations, similar solutions have been exploited in recent works (Wang et al., 2016; Uysal & Yi Lu, 2017, Yang et al., 2017, Xiong et al., 2018).

The models were implemented using the Python Keras library.[13] The preprocessed tweets, as discussed in Section 3.1, were converted in sequence of tokens by using the Keras tokenizer and padded to a fixed length equal to 80 with a special pad token.

[10] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md
[11] http://hlt.isti.cnr.it/wordembeddings/glove_wiki_window10_size300_iteration50.tar.gz
[12] http://hlt.isti.cnr.it/wordembeddings/skipgram_wiki_window10_size300_neg-samples10.tar.gz
[13] https://keras.io/

Table 2
Definitions of the metrics employed.

Metric name	Definition
Accuracy	$Acc = \frac{TP+TN}{TP+FP+FN+TN}$
Precision	$\Pr ec = \frac{TP}{TP+FP}$
Recall	$Rec = \frac{TP}{TP+FN}$
F-measure	$Fmeasure = 2 \cdot \frac{Prec \cdot Rec}{Prec+Rec}$
AUC	$AUC = \left(Rec - \frac{FP}{FP+TN} + 1\right)/2$

Also in this case, we adopted pre-trained word embeddings considering Fast-Text, Glove and Word2Vec, which contain more or less 2500 words each. The dimension of the word embedding space is equal to 300 for all the three word embedding models. Since in our dataset we found 2776 different tokens, new words were initialized with random samples from a uniform distribution over [0,1), according to the findings presented in Yang et al. (2017). As regards the classification models, we carried out a deep experimental campaign for identifying the most suitable architectures and the most performing training parameters.

As regards CNN, we defined three convolutional layers (Conv1D with ReLU activation) characterized by three different filter sizes in {3, 4, 5} and 100 filtering matrices each. The number of hidden neurons was set to 30. A dropout layer was added after the pooling layer and the hidden fully connected layer with the aim of reducing overfitting.

As regards LSTM, the bidirectional layer consisted in two 100-cell LSTM layers. The 200 final hidden states were concatenated and fed into a fully connected layer of 30 units. Dropout was added in the LSTM layers and after fully connected hidden layer.

Both models were trained by minimizing the categorical cross-entropy loss at the output softmax layer, which consists of three neuronal units. We adopted the Adam optimizer with learning rate of 0.001 and batch size 128. In order to find the best hyperparameter configuration for LSTM and CNN models separately, we performed a grid search using different values of dropout rate, different number of epochs, different pre-trained word embedding. We selected the models that delivered the highest accuracy on our dataset, using a 10-fold cross validation. The best configuration for CNN was the one with Fast-Text word embedding, dropout rate equal to 0.2 and 40 epochs of training. The best configuration for LSTM was the one with Word2Vec word embedding, dropout rate equal to 0.4 and 60 epochs of training. In the following, we denote these two schemes as Fast-Text + CNN and W2V + LSTM. However, we will show also the best results achieved adopting the three word embedding schemes combined with the CNN and the LSTM. Thus, we will also consider the following schemes: GLOVE + CNN, W2V + CNN, Fast-Text + LSTM and GLOVE + LSTM.

We evaluated the models in terms of widely-used metrics (Forman, 2003), namely, accuracy, precision, recall, F-measure, and Area Under the Curve (AUC). Table 2 provides the definitions of the metrics employed. For the sake of simplicity, we will explain the metrics referring to the case of a binary classification (i.e., positive class vs. negative class), as the adaptation to a multi-class problem is straightforward. In fact, in this case, the metrics are computed for each class. First, we need to define a few elements of a classification task: (i) *true positives* (TP) is the number of real positive tweets correctly classified as positive; (ii) *true negatives* (TN) is the number of real negative tweets correctly classified as negative; (iii) *false positives* (FP) is the number of real negative tweets incorrectly classified as positive; (iv) *false negatives* (FN) is the number of real positive tweets incorrectly classified as negative. Thus, accuracy is the number of tweets correctly labeled, i.e., the sum of TPs and TNs, over the total number of tweets. Precision is the number of TPs, over the total number of tweets labeled as belonging to the class. Recall is defined as the number of TPs over the total number of tweets that actually belong to the class. The F-measure is the weighted harmonic mean of precision and recall. The AUC is the area underlying the Receiver Operating Characteristic (ROC) curve and is approximated with the equation in Table 2.

Table 3 shows the average results achieved by the different methods discussed above. It is worth noting that with the methods based on deep learning we achieve the worst results. The remaining methods achieve similar results, even though BOW + SVM_ALL shows the highest accuracy. According to previous studies (Uysal & Yi Lu, 2017), this outcome is not completely unexpected. Indeed, deep architectures have proven to be a successful approach in many areas, but they typically require large training sets. In the type of application considered in the paper, also because of the limited number of tweets available in Italian and the known issues related to the small length of each data item, the generation of a training set suitable for deep architectures would be very tedious and almost unfeasible, and would make the application itself not very appealing.

In order to verify if there exist statistical differences among the values of accuracy achieved by the eleven classification models, we also performed a statistical analysis of the results. Similar to the analysis carried out in our previous work in D'Andrea et al. (2015), and as suggested in Derrac, Garcia, Molina, and Herrera, (2011), we applied non-parametric statistical tests: for each classifier we generated a distribution consisting of the 20 values of the accuracies on the test set obtained by repeating two times the 10-fold cross validation. We selected the BOW + SVM_All as control model and we statistically compared the results achieved by this model with the ones achieved by the remaining models. We applied the Wilcoxon signed-rank test (Wilcoxon, 1945), which detects significant differences between two distributions. In all the tests, we used $\alpha = 0.05$ as level of significance. Table 4 shows the results of the Wilcoxon signed-rank test: R+ and R− denote, respectively, the sum of ranks for the folds in which the first model outperformed the second, and the sum of ranks for the opposite condition. Whenever the p value is lower than the level of significance, we can reject the statistical hypothesis of equivalence. Otherwise, no statistical differences can be identified. Thus, BOW + SVM_All is statistically equivalent only to BOW + SVM_2000, BOW + FASTTEXT + SVM and BOW + W2V + SVM. On the other hand, BOW + SVM_All statistically outperforms the remaining models.

Since we aim to select the simplest scheme for text representation, we decided to embed the BOW + SVM_2000 scheme in our stance detection system. We can conclude that the selected scheme is the most suitable one for the task of detecting stance regarding vaccination discussions in Italy. Indeed, it is the simplest one (it adopts just 2000 features for text representation) and achieves results that are comparable (even slightly better) with the ones achieved by the recent state-of-the art method introduced in Mohammad et al. (2017). The authors of Mohammad et al. (2017) showed that their scheme, which adopts the BOW text representation extended with word embeddings and the SVM as classification model, is able to achieve better results, in the framework of stance detection, than the winner of the SemEval 2016 competition (Mohammad et al., 2016).

5. Online monitoring

In this section, first we show the outcomes of the real-time monitoring analysis on Twitter of the stance of people towards the vaccination topic in Italy. Then, since along the time the terms used to express stance about vaccination may change, we present

Table 3
Average results obtained by using the different approaches discussed in the text.

Classifier	Class	F-measure	Precision	Recall	AUC	Accuracy
BOW + SVM_All	*Not in favor*	0.60	62.6%	56.6%	0.73	65.4%
	In favor	0.65	64.5%	65.5%	0.74	
	Neutral	0.71	68.6%	74.0%	0.80	
BOW + SVM_2000	*Not in favor*	0.59	61.5%	56.2%	0.73	64.8%
	In favor	0.64	63.2%	63.9%	0.74	
	Neutral	0.72	69.4%	74.4%	0.81	
BOW + FASTTEXT + SVM	*Not in favor*	0.59	57.9%	60.3%	0.75	64.2%
	In favor	0.73	73.3%	72.6%	0.82	
	Neutral	0.61	62.1%	60.4%	0.72	
BOW + GLOVE + SVM	*Not in favor*	0.56	59.5%	53.0%	0.74	62.2%
	In favor	0.70	66.9%	72.1%	0.79	
	Neutral	0.61	59.9%	61.6%	0.71	
BOW + W2V + SVM	*Not in favor*	0.59	61.1%	56.6%	0.73	63.7%
	In favor	0.72	68.9%	74.9%	0.81	
	Neutral	0.60	60.7%	60.0%	0.72	
FASTEXT + CNN	*Not in favor*	0.57	57.8%	57.9%	0.69	62.9%
	In favor	0.63	64.3%	62.7%	0.70	
	Neutral	0.68	69.6%	68.0%	0.77	
GLOVE + CNN	*Not in favor*	0.55	54.8%	56.6%	0.67	60.5%
	In favor	0.63	64.5%	62.4%	0.71	
	Neutral	0.63	65.1%	62.2%	0.73	
W2V + CNN	*Not in favor*	0.57	57.2%	58.0%	0.69	62.5%
	In favor	0.62	62.9%	61.6%	0.70	
	Neutral	0.69	70.4%	68.1%	0.77	
FASTEXT + LSTM	*Not in favor*	0.55	54.6%	58.4%	0.67	61.2%
	In favor	0.63	61.5%	63.6%	0.70	
	Neutral	0.66	72.7%	61.1%	0.75	
GLOVE + LSTM	*Not in favor*	0.56	55.5%	58.5%	0.68	61.8%
	In favor	0.62	62.2%	63.2%	0.70	
	Neutral	0.67	73.3%	63.4%	0.76	
W2V + LSTM	*Not in favor*	0.57	56.6%	59.8%	0.68	61.9%
	In favor	0.59	59.3%	62.0%	0.69	
	Neutral	0.69	76.2%	63.9%	0.77	

Table 4
Results of the Wilcoxon Signed-Rank test on the accuracies obtained on the test set.

Comparison	R+	R-	p-values	Hypotesis
BOW + SVM_ALL vs. BOW + SVM_2000	27	18	0.528926	*Not-rejected*
BOW + SVM_ALL vs. BOW + FASTTEXT + SVM	35.5	19.5	0.386271	*Not-rejected*
BOW + SVM_ALL vs. BOW + GLOVE + SVM	55	0	0.003842	*Rejected*
BOW + SVM_ALL vs. BOW + W2V + SVM	30	15	0.343253	*Not-rejected*
BOW + SVM_ALL vs. FASTEXT + CNN	42	3	0.016172	*Rejected*
BOW + SVM_ALL vs. GLOVE + CNN	53	2	0.007267	*Rejected*
BOW + SVM_ALL vs. W2V + CNN	43	2	0.012851	*Rejected*
BOW + SVM_ALL vs. FASTEXT + LSTM	55	0	0.003842	*Rejected*
BOW + SVM_ALL vs. GLOVE + LSTM	41	4	0.022327	*Rejected*
BOW + SVM_ALL vs. W2V + LSTM	53	2	0.007267	*Rejected*
BOW + SVM_ALL vs. GLOVE + CNN	53	2	0.007267	*Rejected*
BOW + SVM_ALL vs. W2V + CNN	43	2	0.012851	*Rejected*
BOW + SVM_ALL vs. FASTEXT + LSTM	55	0	0.003842	*Rejected*
BOW + SVM_ALL vs. GLOVE + LSTM	41	4	0.022327	*Rejected*
BOW + SVM_ALL vs. W2V + LSTM	53	2	0.007267	*Rejected*

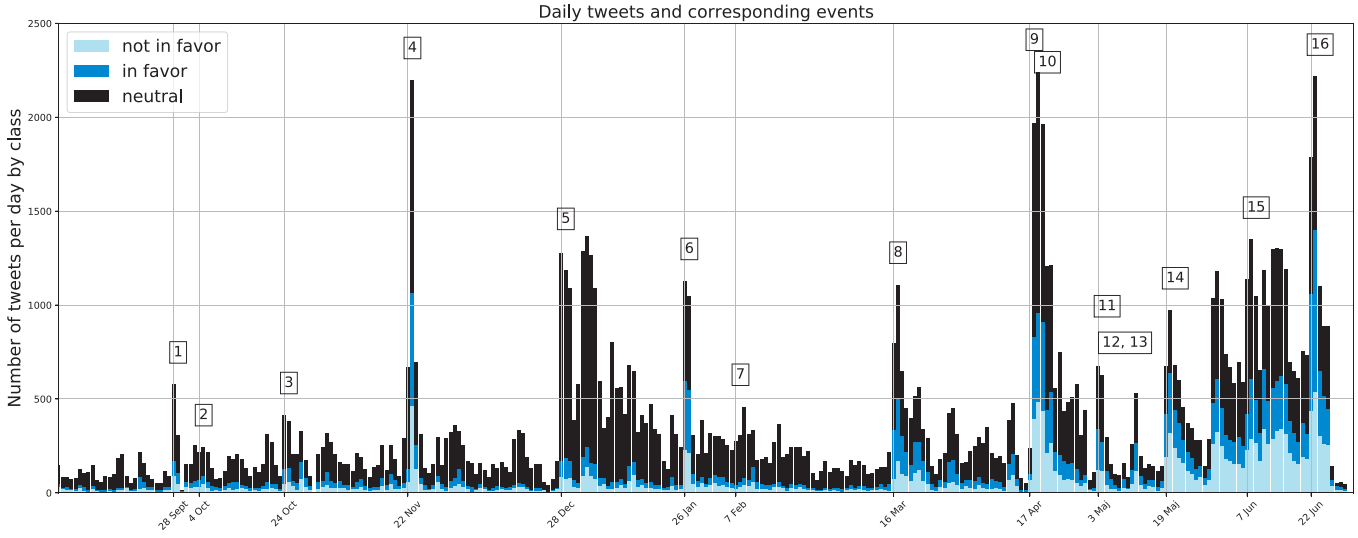

Fig. 4. Daily number of tweets by class (from September 1st, 2016 to June 30th, 2017).

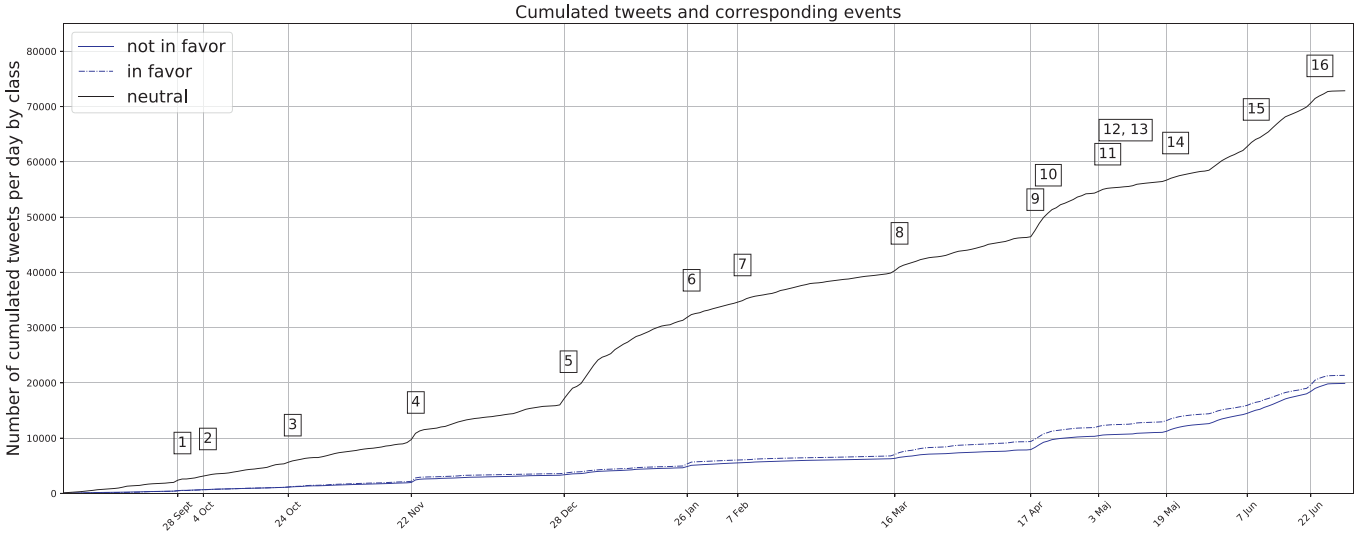

Fig. 5. Number of cumulated tweets by class (from September 1st, 2016 to June 30th, 2017).

an experimental study for detecting the possible presence of *concept drift*. Finally, we analyze the content of some tweets classified as belonging to the three classes considered in the paper.

5.1. Outcome of the online monitoring

The monitoring analysis lasted 10 months, from September 1st, 2016 to June 30th, 2017. During this time interval, we fetched $N = 112{,}397$ tweets using the same set of keywords presented in Table 1. From this set of tweets, we removed the training tweets. The remaining tweets were preprocessed and classified. As stated before, we selected the BOW + SVM_2000 scheme as the most suitable one for text representation and classification. The adopted scheme was trained using the entire training set extracted from September 1st, 2016 to January 31st, 2017.

In the following, we show and discuss the outcome of the online monitoring analysis over the 10 months. Further, we deepen the analysis in correspondence with local peaks of the daily number of tweets. We have verified that these peaks are related to particular events concerning vaccination (discussions in Parliament, approval of law establishing vaccination requirements, etc.) and therefore are very interesting to evaluate the effectiveness of our systems. Figs. 4–6 illustrate an overview of the number of tweets per day per class. More precisely, Fig. 4 depicts a stacked histogram of the number of *not in favor of vaccination*, *in favor of vaccination*, and *neutral* tweets classified by our system per day during the time span considered. Fig. 5 shows the cumulated value of tweets by class over the same time interval, and Fig. 6 compares the daily opinion trend limited to subjective tweets (*in favor of vaccination* and *not in favor of vaccination*). In all figures, we can easily see how the number of tweets increases in correspondence with some days (local peaks of the daily number of tweets). By analyzing news in these days, we have discovered the presence of specific events related to vaccination (the number upon the peak idicates the event). In particular, we have identified the following events:

(1) event #1: Cancellation of the projection of the documentary film "*Vaxxed: from cover-up to catastrophe*" in the Italian Republic Senate on September 28th, 2016[14];

[14] www.ilfattoquotidiano.it/2016/09/28/vaccini-senato-annulla-proiezione-%E2%80%AAdel-documentario-vaxxed-cover-catastrophe/3062895/, (accessed 16 October 2017).

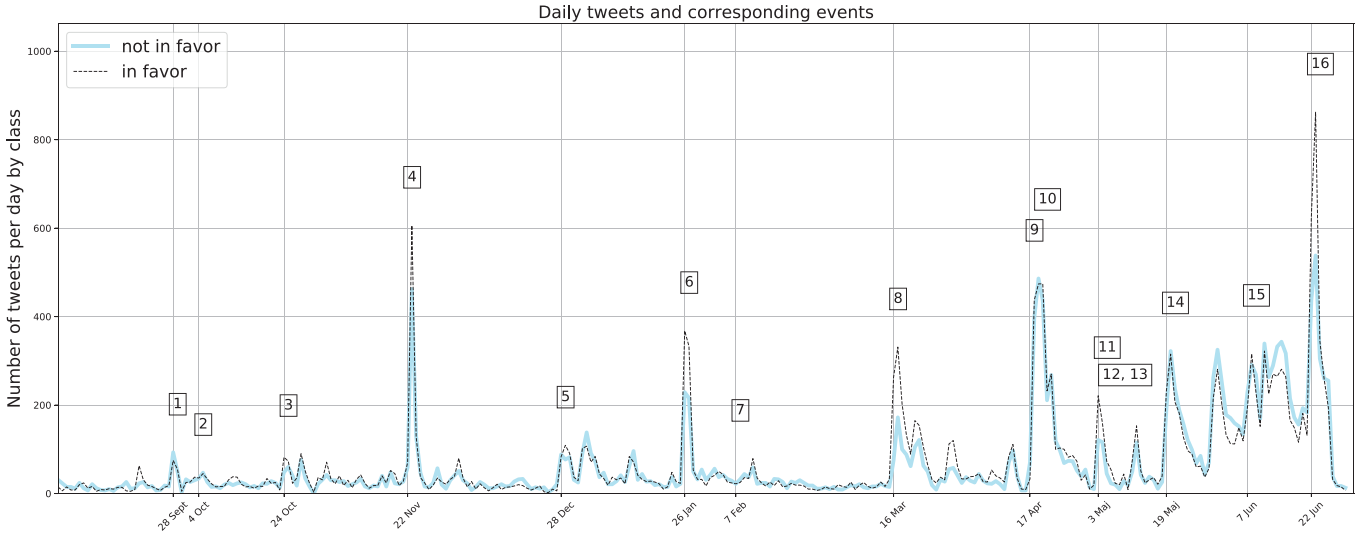

Fig. 6. Number of subjective tweets (*in favor of vaccination* vs. *not in favor of vaccination*) per day (from September 1st, 2016 to June 30th, 2017).

(2) event #2: Expected projection of the documentary film "*Vaxxed: from cover-up to catastrophe*" in the Italian Republic Senate on October 4th, 2016[15];
(3) event #3: Speech by President of Italian Republic about vaccines on October 24th, 2016[16];
(4) event #4: Approval of the law establishing vaccination requirements for school children in Emilia Romagna Region, Italy, approved on November 22nd, 2016[17];
(5) event #5: Death of a school teacher for meningitis in Rome, Italy, news of December 28th, 2016[18];
(6) event #6: Agreement between Italian Health Minister and Italian Regions about vaccinations requirement on January 26th, 2017[19];
(7) event #7: Cancellation of the projection of the documentary film "*Vaxxed: from cover-up to catastrophe*" at the European Parliament on February 7th, 2017[20];
(8) event #8: Increase of 230% cases of measles in Italy, news of March 16th, 2017[21];
(9) event #9: Italian TV show *Report* focusing on vaccines cause controversy on April 17th, 2017[22];
(10) event #10: Fake vaccinations in the Italian city of Treviso, news of April 19th, 2017[23];
(11) event #11: Fake vaccinations in the Friuli Region, Italy, news of May 3rd, 2017[24];
(12) event #12: NY Times against Italian political party against vaccines, news of May 4th, 2017[25];
(13) event #13: 5 times increase in measles cases in Italy in April 2017, news of May 4th, 2017[26];
(14) event #14: Approval of the decree on vaccinations requirement (12 vaccines) in Italian kindergartens on May 19th, 2017[27];
(15) event #15: President of Italian Republic signs the decree about 12 vaccinations requirement in Italian schools on June 7th, 2017[28];
(16) event #16: Kid sick of leukemia died for measles in Monza, Italy, news of June 22nd, 2017[29].

From Fig. 4 we can observe that, in absence of context-related events, the number of tweets per day is quite low (e.g., about 100–200 tweets per day until September 28th, 2016). This value rapidly grows when context-related events occur (e.g., it exceeds 500 on September 28th, 2016 when event #1 occurs). Further, by observing Figs. 4 and 5, we can see that some events produced a higher spike in the daily number of tweets, i.e., higher than 2000. These spikes occur immediately after November 22nd, 2016, April 17th, 2017, and June 22nd, 2017. In correspondence with these dates, we can identify the following triggering events: i) event #4 on November 22nd, 2016; ii) event #9 on April 17th, 2017; iii) event #10 on April 19th, 2017; and iv) event #16 on June 22nd, 2017.

The effect of a triggering event may be more or less emphasized depending on the flow of the event itself, and on the perception of the event by Twitter users. Further, the effect of the event (in terms of number of shared tweets) may be observable almost immediately, as it typically happens with viral news, or some hours/days later. E.g., the spike corresponding to event #4 actually occurs the day after, i.e., on November 23rd, 2016. Further, events very close in time may contribute to the same spike. E.g., the spike

[15] www.lastampa.it/2016/09/28/italia/film-contro-i-vaccini-in-senato-la-polemica-cancella-levento-LnDCe2j3uTq8KukEEey8uJ/pagina.html (accessed 10 Sept. 2018).
[16] www.repubblica.it/salute/medicina/2016/10/24/news/mattarella_sconsiderato_chi_critica_vaccini-150471038/, (accessed 16 October 2017).
[17] www.repubblica.it/salute/prevenzione/2016/11/22/news/vaccini_obbligatori_emilia_romagna_immunita_gregge-152543276/, (accessed 16 October 2017).
[18] www.ilpost.it/2016/12/28/meningite/, (accessed 16 October 2017).
[19] www.huffingtonpost.it/2017/01/26/vaccini-obbligatori-accordo-storico_n_14417108.html, (accessed 16 October 2017).
[20] www.repubblica.it/salute/prevenzione/2017/02/07/news/vaccini_il_film_vaxxed_sull_autismo_al_parlamento_ue_lorenzin_scrive_a_tajani-157788259/, (accessed 16 October 2017).
[21] www.ilfattoquotidiano.it/2017/03/16/morbillo-i-dati-del-ministero-della-salute-preoccupante-aumento-dei-casi-230-in-un-anno-e-colpa-del-rifiuto-dei-vaccini/3456211/, (accessed 16 October 2017).
[22] www.repubblica.it/cronaca/2017/04/19/news/tra_inchieste_e_bufale-163328161/, (accessed 16 October 2017).
[23] www.repubblica.it/cronaca/2017/04/19/news/treviso_infermiera_fiale_vaccini-163380333/, (accessed 16 October 2017).
[24] www.repubblica.it/cronaca/2017/05/03/news/friuli_venezia_giulia_fingeva_vaccini_oltre_20mila_dosi_dubbie-164506727/, (accessed 16 October 2017).
[25] www.repubblica.it/politica/2017/05/03/news/nyt_contro_i_5_stelle_loro_negazione_populista_sull_efficacia_dei_vaccini_aumenta_la_diffusione_di_gravi_malattie_-164484892/, (accessed 16 October 2017).
[26] www.repubblica.it/salute/prevenzione/2017/05/04/news/morbillo_casi_aumento_2017-164596470/, (accessed 16 October 2017).
[27] www.repubblica.it/salute/2017/05/19/news/vaccini_oggi_testo_in_cdm_boschi_no_scherzi_su_salute-165815370/ , (accessed 16 October 2017).
[28] www.ilfattoquotidiano.it/2017/06/07/vaccini-mattarella-firma-il-decreto-su-obbligo-per-iscrizione-a-scuola-bastera-autocertificazione-o-la-prenotazione/3642713/, (accessed 16 October 2017).
[29] milano.repubblica.it/cronaca/2017/06/23/news/morbillo_vaccini_bambino_morto_monza_dubbi_ospedale-168926326/, (accessed 16 October 2017).

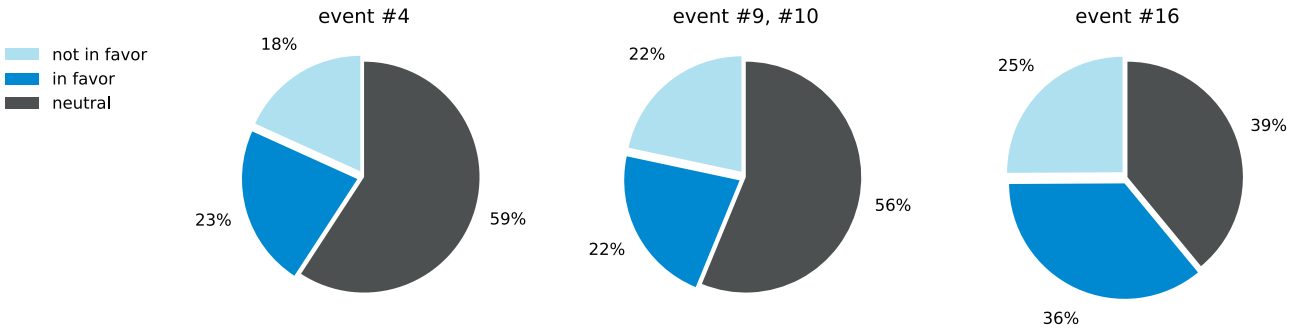

Fig. 7. Distribution of opinion polarity over the classes by event. The opinion polarity for event #4 regards November 22nd–24th, 2016, that for event #9 and #10 (considered together) regards April 17th–20th, 2017, and that for event #16 regards June 22nd–24th, 2017.

after April 17th 2017 may be caused by the event itself (event #9), but also event #10, occurring just two days later, can likely be a cause. Thus, events #9 and #10 may be merged into a single aggregated event, for the purpose of the analysis.

Further, in addition to the total number of tweets per day, independently of the class, we can observe (in Figs. 4–6) also spikes (i.e., sudden changes) in the number of tweets of a given class. More precisely, spikes of *neutral* tweets are the most notable, both in terms of frequency and of amplitude. However, *neutral* tweets, for the most part, correspond to news tweets and indicate users talking about vaccines or sharing news in correspondence with the event, whereas sometimes correspond also to personal objective texts without a clear opinion.

In order to understand better the opinion polarity after the occurrence of a context-related event, we can study the distribution of tweets over the three classes. Let us consider event #4. The event causes a main spike in the number of tweets on November 23rd, 2016, and two minor spikes (i.e., with more than 500 daily tweets) on November 22nd and November 24th, 2016. Thus, we can aggregate the opinions shared during these three days in order to analyze the effects of event #4. In the time interval November 22nd–24th, 2016, 3566 tweets were shared, about 59% of tweets were classified as *neutral*, about 23% as *in favor of vaccination*, and about 18% as *not in favor of vaccination*. Similarly, we can repeat this analysis for event #16, and for events #9 and #10 considered together. More precisely, the effects of event #4 span on November 22nd–24th, 2016, those of event #9 and #10 together span on April 17th–20th, 2017, and those of event #16 span on June 22nd–24th, 2017. The number of days to consider in order to study the opinion polarity in correspondence with each event, depends on the number of daily tweets, and was decided by visually inspecting Fig. 4. Fig. 7 summarizes the distribution of opinion polarity for the events considered. We can observe that the stance is overall *neutral* for aggregated events #9 and #10. It is considerably biased towards *in favor of vaccination* for event #16 (+11%) and slightly biased towards *in favor of vaccination* for event #4 (+5%). By taking into account the overall time span of 10 months, the distribution of opinion over the three classes is: *in favor of vaccination* for about 19%, *neutral* for about 64%, and *not in favor of vaccination* for about 17%. Having the majority of tweets classified as neutral is a common behavior in social networks (Ghiassi, Skinner, & Zimbra, 2013). Further, by taking into account only subjective tweets (i.e., by discarding *neutral* tweets), we can state that, overall during the 10 months, about 52% of opinions are *in favor of vaccination,* whereas about 48% are *not in favor of vaccination.* Hence, the opinion is slightly biased towards the *in favor of vaccination* class. Obviously, this is an aggregated result, which may hide the variations occurring during the time span. Hence, we made a monthly analysis. Figs. 8 and 9 show the distribution of tweets over the three classes per month, and the number of tweets shared per month, respectively. From Fig. 8, we can notice that the number of tweets expressing a subjective opinion increased significantly in Spring 2017. More precisely, the amount of *neutral* tweets, initially over 70%, decreases to around 50% in May 2017, making May the month with the highest percentages of subjective opinions. Further, we can observe from Fig. 4 that also the total number of daily tweets has grown during the time span. These facts suggest that the number of people talking about vaccination increased, as a consequence of vaccine-related events.

5.2. Concept drift detection and analysis

When dealing with continuous classification of data streams along the time, the issue of concept drift should be analyzed. Indeed, the classification models are usually trained using data extracted in a specific time interval. Then, such models are used for classifying the new instances received in streaming. Since the characteristics of the phenomenon under observation can change along time, the performance of the classification models may deteriorate, due to this concept drift. Thus, once in the classification system the presence of concept drift is detected, appropriate strategies for reducing it have to be possibly applied (Gama, Žliobaitė, Bifet, Pechenizkiy, & Bouchachia, 2014).

In the context analyzed in this work, in which we carry out a real-time classification of stance about vaccination in Italy from tweets, users of Twitter may change over time the words and/or phrases used for expressing their opinion. For this reason, we decided to carry out an additional experimental analysis for detecting the presence of concept drift along the time span under observation. To this aim, we analyzed the tweets belonging to 7 local peaks of the daily number of tweets, which correspond to seven, namely #4, #5, #6, #8, #10, #14 and #16, out of the sixteen events described in Section 5.1. We randomly read several tweets and manually labelled around 60 tweets for each event, trying to identify 20 tweets for each class. We limited the analysis to a subset of the sixteen events just because the labelling task is quite tedious and time expensive.

To detect the presence of drift, we evaluated the F-measure, the precision, the recall, the AUC per class and the overall accuracy, obtained classifying the labelled tweets of each selected peak. First, we evaluated the performances obtained on each event by the classification model trained using the initial training set, extracted from September 1st, 2016 to January 31st, 2017 (we checked that the tweets of the events #4, #5, and #6 were not included into the training set). Then, before evaluating the classification performances on a specific event, we re-trained the classification model extending the training set by considering the labelled tweets of the previously analyzed events. As an example, when we analyzed event #10, we re-trained the classification models considering the initial training set extended with the labelled tweets of events #4, #5, #6 and #8. This incremental learning procedure was proposed in Costa, Silva, Antunes, and Ribeiro, (2014), where three

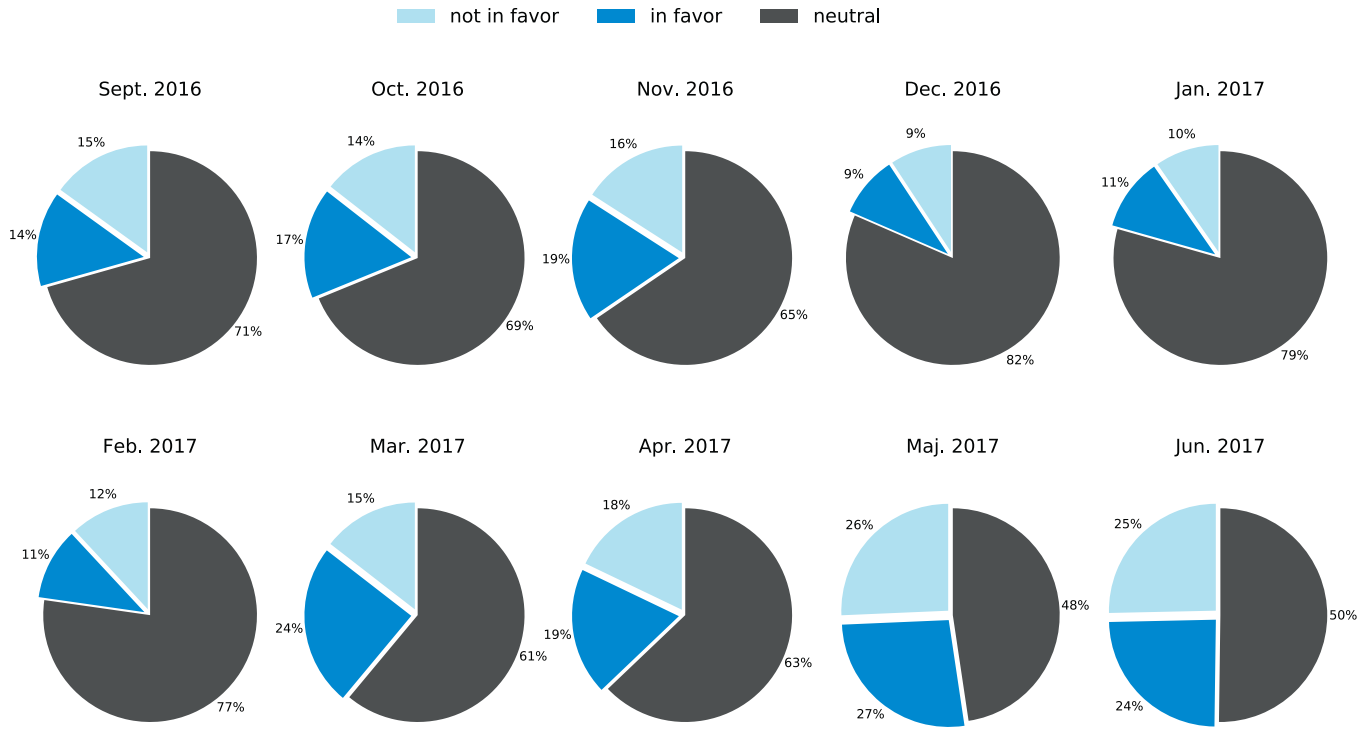

Fig. 8. Distribution of opinion polarity over the classes by month.

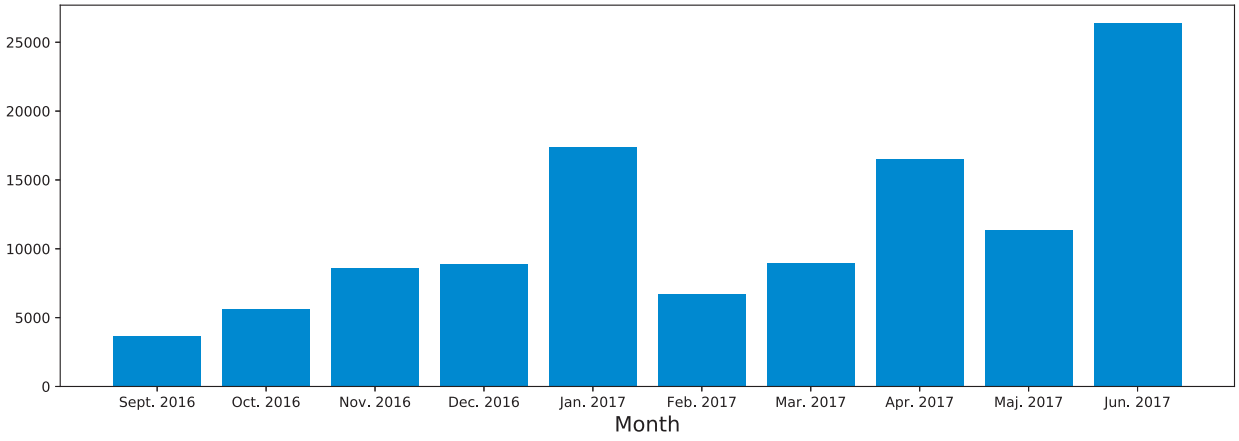

Fig. 9. Number of tweets per month.

different solutions were compared for approaching the problem of handling concept drift in the classification of Twitter streams. The procedure resulted to be the one with the best performance. In the following, we refer to the incrementally trained classifier as BOW + SVM_2000_INC.

Tables 5 and 6 show the results of the concept drift analysis for the BOW + SVM_2000 and BOW + SVM_2000_INC classifiers, respectively. Furthermore, Fig. 10 shows a plot of the trends of the accuracy over the time span of the selected events for both classification schemes. In the figure, the blue dotted line and the continuous black line show the trend of the accuracy achieved, respectively, by BOW + SVM_2000 and BOW + SVM_2000_INC.

The analysis of Table 5 and Fig. 10 shows that the incremental solution outperforms the BOW + SVM_2000 scheme only in events 6 and 8. On the other hand, in the worst case the accuracy of BOW + SVM_2000_INC deteriorates down to a value below 60%. Conversely, the accuracy of BOW + SVM_2000 remains quite stable along the time window. Thus, we can affirm that our selected solution does not look to be particularly affected by concept drift. We can conclude that adopting an incremental learning does not lead to an effective reduction of the concept drift. Indeed, although in the first events the incremental learning produces a gain in accuracy, in the subsequent events we observe a decrease. This trend may be due to a strong dependency of the accuracy on specific words, which are used in some events, but not in all. Anyway, given the complexity of the overall scenario and the relatively small amount of labelled data available for the re-training procedure, the incremental solution does not provide very effective improvements.

We believe that an average accuracy of 62.75% on the selected and labelled tweets (a test set of around 420 tweets) can be considered a good result. Indeed, a recent work discussed in Dey et al. (2018) obtains, on the SemEval 2016 stance detection Twitter task dataset,[30] a best-case accuracy of 60.2% on the test set. On the other hand, the BOW + SVM_2000 scheme may be

[30] http://saifmohammad.com/WebPages/StanceDataset.htm

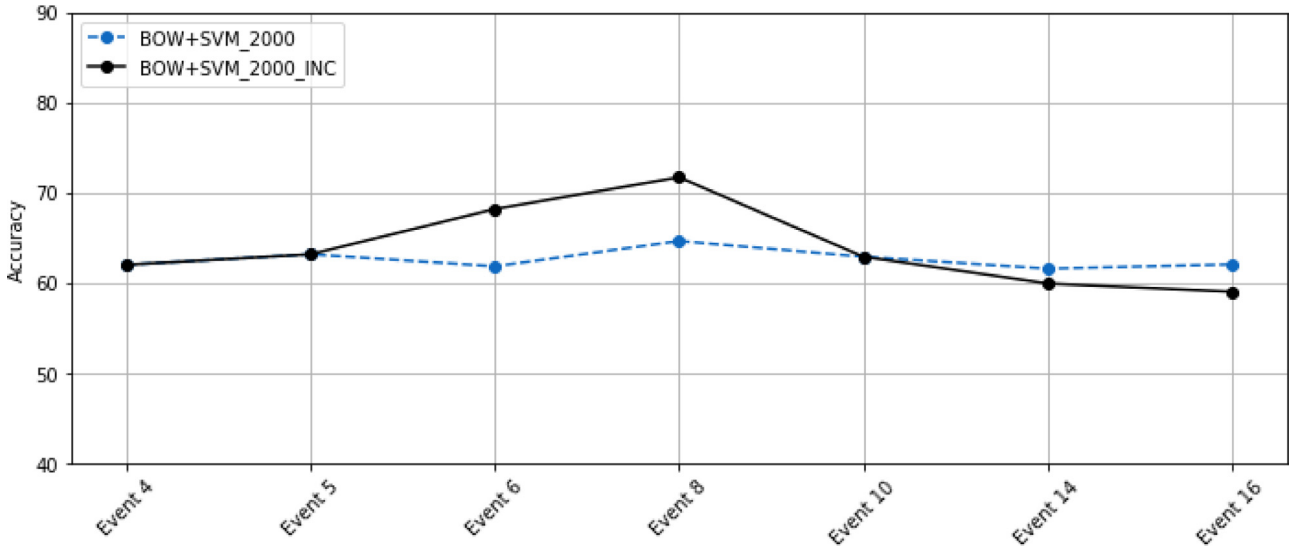

Fig. 10. Accuracy trends over the time span of seven peaks of the daily number of tweets.

Table 5
Results of the concept drift analysis for BOW + SVM_2000.

Event	Class	F-measure	Precision	Recall	AUC	Accuracy
#4	*Not in favor*	0.51	52.9%	50.0%	0.64	62.1%
	In favor	0.63	66.7%	60.0%	0.73	
	Neutral	0.70	65.2%	75.0%	0.76	
#5	*Not in favor*	0.55	72.7%	44.4%	0.68	63.2%
	In favor	0.51	61.1%	44.0%	0.70	
	Neutral	0.75	61.5%	96.0%	0.80	
#6	*Not in favor*	0.53	64.3%	45.0%	0.65	61.9%
	In favor	0.55	48.0%	63.2%	0.68	
	Neutral	0.75	75.0%	75.0%	0.84	
#8	*Not in favor*	0.53	63.2%	46.2%	0.79	64.7%
	In favor	0.64	67.9%	61.3%	0.77	
	Neutral	0.73	63.2%	85.7%	0.82	
#10	*Not in favor*	0.64	75.0%	56.3%	0.70	63.0%
	In favor	0.43	40.0%	46.2%	0.62	
	Neutral	0.73	70.4%	76.0%	0.77	
#14	*Not in favor*	0.50	55.6%	45.5%	0.65	61.6%
	In favor	0.66	66.7%	64.3%	0.70	
	Neutral	0.67	60.7%	73.9%	0.76	
#16	*Not in favor*	0.72	73.1%	70.4%	0.76	62.1%
	In favor	0.50	47.6%	52.6%	0.65	
	Neutral	0.62	63.2%	60.0%	0.81	
Average	*Not in favor*	0.57	65.3%	51.1%	0.70	62.6%
	In favor	0.56	56.9%	55.9%	0.69	
	Neutral	0.71	65.6%	77.4%	0.79	

Table 6
Results of the concept drift analysis for BOW + SVM_2000_INC.

Event	Class	F-measure	Precision	Recall	AUC	Accuracy
#4	*Not in favor*	0.51	52.9%	50.0%	0.64	62.1%
	In favor	0.63	66.7%	60.0%	0.73	
	Neutral	0.70	65.2%	75.0%	0.76	
#5	*Not in favor*	0.60	75.0%	50.0%	0.72	63.2%
	In favor	0.49	62.5%	40.0%	0.69	
	Neutral	0.74	60.0%	96.0%	0.79	
#6	*Not in favor*	0.65	70.6%	60.0%	0.74	68.3%
	In favor	0.60	57.1%	63.2%	0.72	
	Neutral	0.78	76.0%	79.2%	0.85	
#8	*Not in favor*	0.64	71.4%	57.7%	0.80	71.8%
	In favor	0.70	68.8%	71.0%	0.78	
	Neutral	0.80	75.0%	85.7%	0.86	
#10	*Not in favor*	0.69	68.8%	68.8%	0.73	63.0%
	In favor	0.44	42.9%	46.2%	0.65	
	Neutral	0.69	70.8%	68.0%	0.75	
#14	*Not in favor*	0.53	55.6%	50.0%	0.70	60.0%
	In favor	0.67	68.4%	65.0%	0.73	
	Neutral	0.61	56.5%	65.0%	0.74	
#16	*Not in favor*	0.61	58.6%	63.0%	0.68	59.1%
	In favor	0.43	44.4%	42.1%	0.63	
	Neutral	0.72	73.7%	70.0%	0.82	
Average	*Not in favor*	0.60	64.7%	57.1%	0.72	63.9%
	In favor	0.57	58.7%	55.4%	0.70	
	Neutral	0.72	68.2%	77.0%	0.80	

adapted/updated when an appreciable concept drift is detected. This can happen when new keywords become of interest, or in the case that the way of writing or the opinion of users actually changes with respect to the keywords.

5.3. Analysis of a selected set of tweets on major events

Lastly, we discuss the outcome of the classification of a few tweets fetched in correspondence with the major events mentioned above, i.e., events #4, #9, #10, and #16. Table 7 shows 22

Table 7
A few examples of tweets classification for some major events.

Event	Text of tweet – [English translation]	Actual class	Assigned class	Note
#4	*"Meningite, donna morta a Firenze. Non era vaccinata contro meningococco di tipo C" – ["Meningitis, a woman died in Florence. It was not vaccinated against type- C meningococco"]*	*Neutral*	*Neutral*	Hit, news tweet
#4	*"Ai miei tempi varicella e morbillo si curavano in casa, a letto, con pozioni magiche della mamma! Ora solo vaccini contro il male di vivere!" – ["In my time, varicella and measles were cared for at home, in bed, with mom's magic potions! Now only vaccines against the evil of living!"]*	*Not in favor*	*Not in favor*	Hit
#4	*"Si ringraziano i criminali imbecilli che fanno propaganda contro i vaccini." – ["Thanks criminal idiots who make propaganda against vaccines."]*	*In favor*	*Not in favor*	Miss, due to irony/sarcasm
#4	*"Epidemia di meningite in Toscana, ma tranquilli continuate a non vaccinare i bambini, che tanto non succede niente!" – ["Meningitis outbreak in Tuscany, but be calm and still do not vaccinate children, nothing will happen!"]*	*In favor*	*Not in favor*	Miss, due to irony/sarcasm
#4	*"Dovessi fare io le leggi, proibirei a questa bella gente l'accesso a tutti i luoghi pubblici e imporrei la vaccinazione coatta." – ["If I had to do myself the laws, I would prohibit to such nice people the access to all public places and impose the forced vaccination."]*	*In favor*	*In favor*	Hit
#4	*"Trattamenti periodici e vaccinazioni del tuo cane. Ricordi tutte le scadenze? Scommettiamo di no…" – ["Periodic treatments and vaccinations for your dog. Do you remember all deadlines? We bet not…"]*	*Neutral*	*Neutral*	Hit
#4	*"Vergogna, votano contro la vaccinazione in Emilia Romagna che per fortuna passa. Oscurantismo e medioevo spacciati per progresso." – ["Shame! They vote against vaccination in Emilia Romagna, which fortunately was approved. Obscurity and Middle Ages passed off as progress."]*	*In favor*	*In favor*	Hit
#4	*"I vaccini funzionano, servono e la realtà lo ha confermato. Dibattito si può fare solo con prove che dicano diversamente." – ["Vaccines work, they are useful and reality has confirmed it. Debate can only be done only with evidence that show differently."]*	*In favor*	*In favor*	Hit
#9, #10	*"In Italia un caso di difterite: le conseguenze del calo dei vaccini." – ["In Italy a case of diphtheria: the consequences of the fall in vaccines."]*	*Neutral*	*Not in favor*	Miss, news tweet
#9, #10	*"Visto che i vaccini sono obbligatori, perchè non li fanno gratis? Evidentemente adesso vogliono fare arricchire qualche casa farmaceutica." – ["Since vaccines are mandatory, why they do not do it for free? Obviously now they want to enrich some pharmaceutical company."]*	*Not in favor*	*Not in favor*	Hit
#9, #10	*"Quindi, la mia opinione è che alcuni vaccini devono essere obbligatori, anche per tutela dei genitori." – ["So, in my opinion some vaccines must be mandatory, also for protecting the parents."]*	*In favor*	*In favor*	Hit
#9, #10	*"Non fate girare queste bufale assurde. Ormai è stato dimostrato da anni che l'autismo non ha nulla a che fare con i vaccini." – ["Do not spread these absurd fake news. It has been shown for years that autism has nothing to do with vaccines."]*	*In favor*	*Neutral*	Miss
#9, #10	*"Loro votano No ai vaccini, e a noi ci chiamano serial killer dei loro figli!" – ["They say No to vaccines, and call us serial killers of their children!"]*	*In favor*	*Neutral*	Miss, due to irony/sarcasm
#9, #10	*"Il primo effetto del vaccino contro l'influenza è la voglia di dormire altre ore." – ["The first effect of the vaccine against flue is the wish for sleeping extra hours."]*	*Neutral*	*Neutral*	Hit
#9, #10	*"Peccato non esista un vaccino contro la stupidita!" – ["What a pity there is no vaccine against stupidity!"]*	*In favor*	*In favor*	Hit
#16	*"Le aborrite case farmaceutiche farebbero molti piu profitti senza vaccini, cara signora. Sai quanti farmaci per le complicanze del morbillo?" – ["The abhorred pharmaceutical companies would earn more without vaccines, dear lady. Do you know how many medicines for the complications of measles?"]*	*In favor*	*Neutral*	Miss, due to irony/sarcasm
#16	*"Noi che siamo quasi uguali di età e abbiamo avuto varicella e morbillo, e siamo sopravvissuti… Perchè ora fanno i vaccini?" – ["We are almost the same age and we got varicella and measles, and we survived … Why now they make vaccines?"]*	*Not in favor*	*Not in favor*	Hit
#16	*"Contagiato dai fratelli non vaccinati per scelta della famiglia. No, non levamogliela la patria potestà…" – ["Infected by unvaccinated siblings by choice of the family. No, let's not take away their parental rights…"]*	*In favor*	*Not in favor*	Miss, due to irony/sarcasm
#16	*"Morto di morbillo il bimbo che non potendosi vaccinare, contava sull'immunita di gregge … Lo avete uccisovoi antivax!!" – ["Dead of measles the child who not being able to vaccinate, relied on the herd immunity… You antivax killed him!!"]*	*In favor*	*In favor*	Hit
#16	*"Un bambino leucemico non puo vaccinarsi e si becca il morbillo da qualcuno non vaccinato per scelta… Ecco il risultato… Chiaro?" – ["A leukemia child can not vaccinate and contracts measles from someone not vaccinated by choice … Here's the result… Clear?"]*	*In favor*	*In favor*	Hit
#16	*"Bruno, ok, i vaccini servono. Ma che senso ha l'obbligo di vaccini, con quelle sanzioni, pure per malattie non infettive come il tetano?" – ["Bruno, ok, vaccines are needed. But what is the sense of mandatory vaccines, with those sanctions, even for non-infectious diseases like tetanus?"]*	*Neutral*	*Not in favor*	Miss, due to ambiguity for discording opinions
#16	*"Vaccini, il decreto diventa piu morbido su sanzioni e patria potestà." – [Vaccines, the decree becomes softer on sanctions and parental rights."]*	*Neutral*	*Neutral*	Hit

randomly chosen tweets. For each tweet, we show the actual class label, the class label assigned by the system, and the outcome of the classification. More precisely, the system correctly classifies 14 tweets, whereas it misclassifies 8 tweets. Regarding the misclassified tweets, we can observe that 6 of such tweets contain irony or sarcasm expressions in the text, making the classification more difficult. In fact, whereas humans are easily able to detect irony or sarcasm in a text, in the field of sentiment analysis and opinion mining, irony detection is a challenging task, given that the presence of irony may completely reverse the text polarity (Giachanou & Crestani, 2016). Further, one misclassified tweet is ambiguous, that is, there are discording opinions within the same tweet.

In conclusion, the proposed intelligent system can be successfully employed for stance detection in the context of vaccination in Italy. Detecting users' opinions over vaccines or shifts of the public opinion concurrently with social context-related events may be important for Public Healthcare Organizations in order to promote actions aimed at avoiding outbreaks of eradicated diseases.

Further, the system may be employed to early detect the spread of fake/incomplete news (e.g., in the case of an unexpected rising negative opinion about vaccination, caused by the diffusion of a fake news).

6. Conclusions

In this work, we have discussed how to perform stance classification on Twitter with reference to the vaccination topic in Italy. The proposed approach fetches and pre-processes vaccine-related tweets and employs an SVM model to classify tweets as belonging to one among three classes, namely, *in favor of vaccination, not in favor of vaccination*, and *neutral*, with an accuracy of 64.84%. The results achieved are in line or outperform similar works in the literature. The aim is to monitor and track shifts of the Italian public opinion about vaccinations, with reference to social context-related events, which may influence the public opinion itself. In fact, an early detection of an opinion shift may be of the utmost importance for Public Healthcare Organizations in order to promote actions aimed at avoiding outbreaks of eradicated diseases. We have also shown the results of a monitoring campaign lasting 10 months, from September 2016 to June 2017. In particular, we have analyzed how the polarity of the public opinion changes in correspondence with the local peaks of the daily number of tweets. These peaks correspond to specific events related to the vaccination topic. Finally, we have shown that our system does not suffer particularly from concept drift.

Author contributions section

E. D'Andrea, P. Ducange and F. Marcelloni conceived the initial idea. E. D'Andrea and P. Ducange developed the overall approach, performed the experiments, and wrote the relative paper section. A. Renda, with the supervision of A. Bechini, dealt with the experiments on neural networks and wrote the corresponding part in the paper. A. Bechini took care of the Introduction and the state of the art section of the paper. Francesco Marcelloni supervised all the work, wrote some parts of the paper, and revised the final version of the paper. Thus, all authors contributed to the final manuscript.

Acknowledgements

This work was partially supported by the project funded by "Progetti di Ricerca di Ateneo- PRA 2017" of the University of Pisa.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media* (pp. 30–38).

Aisopos, F., Papadakis, G., & Varvarigou, T. (2011). Sentiment analysis of social media content using N-Gram graphs. In *Proc. of the 3rd ACM SIGMM Int. Workshop on Social Media* (pp. 9–14).

Alsaedi, N., Burnap, P., & Rana, O. (2017). Can we predict a riot? Disruptive event detection using twitter. *ACM Transactions on Internet Technology, 17*(2), 1–18 1826.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of International Conference on Language Resources and Evaluation *(*LREC) (pp. 2200–2204).

Balahur, A., & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language, 28*(1), 56–75. https://doi.org/10.1016/j.csl.2013.03.004.

Basile, V., & Nissim, M. (2013). Sentiment analysis on Italian tweets. In *Proceedings of 4th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*.

Bechini, A., Gazzè, D., Marchetti, A., & Tesconi, M. (2016). Towards a general architecture for social media data capture from a multi-domain perspective. In *Proceedings of 2016 IEEE International Conference on Advanced Information Networking and Applications (AINA)* (pp. 1093–1100). doi:10.1109/AINA.2016.75.

Becker, K., Moreira, V. P., & dos Santos, A. G. L. (2017). Multilingual emotion classification using supervised learning: comparative experiments. *Information Processing & Management, 53*(3), 684–704. https://doi.org/10.1016/j.ipm.2016.12.008.

Bello-Orgaz, G., Hernandez-Castro, J., & Camacho, D. (2017). Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems, 66*, 125–136.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*, 1–8.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Botsis, T., Nguyen, M. D., Woo, E. J., Markatou, M., & Ball, R. (2011). Text mining for the Vaccine Adverse Event Reporting System: Medical text classification using informative feature selection. *Journal of American Medical Informatics Association, 18*(5), 631–638.

Breiman, L. (2001). Random forests. *Machine Learning,, 45*(1), 5–32.

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems, 31*, 102–107.

Cambria, E., Havasi, C., & Hussain, A. (2012). SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In Proceedings of 25th Florida Artificial Intelligence Research Society Conf. (FLAIRS) (pp. 202–207).

Castellucci, G., Croce, D., Cao, D. D., & Basili, R. (2016). User Mood Tracking for Opinion Analysis on Twitter. In G. Adorni, S. Cagnoni, M. Gori, & M. Maratea (Eds.). In *AI*IA 2016 advances in artificial intelligence: 10037* (pp. 76–88). Cham: Springer. AI*IA 2016. Lecture Notes in Computer Science.

Pandey, Chandra,A, Singh, A., Rajpoot, D., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management, 53*, 764–779.

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of twitter: content analysis of tweets during the 2009 H1N1 outbreak. *PLoS ONE, 5*(11), 1–13.

Chien, C. C., & Tseng, Y.-D. (2011). Quality evaluation of product reviews using an information quality framework. *Decision Support Systems, 50*, 755–768.

Cliche, M. (2017). BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.

Costa, J., Silva, C., Antunes, M., & Ribeiro, B. (2014). Concept drift awareness in twitter streams. *Proc. IEEE 13th International Conference on Machine Learning and Applications (ICMLA)* (pp. 294–299). doi:10.1109/ICMLA.2014.53.

D'Andrea, E., Ducange, P., Lazzerini, B., & Marcelloni, F. (2015). Real-Time Detection of Traffic From Twitter Stream Analysis. *IEEE Transactions on Intelligent Transportation Systems, 16*(4), 2269–2283. https://doi.org/10.1109/TITS.2015.2404431.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th Int'l Conference on World Wide Web* (pp. 519–528).

Derrac, J., Garcia, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation, 1*(1), 3–18.

Dey, K., Ritvik, S., & Saroj, K. (2018). Topical stance detection for twitter: a two-phase LSTM model using attention. In *Proceedings of European Conference on Information Retrieval* (p. 2018).

Du, J., Xu, J., Song, H.-Y., & Tao, C. (2017). Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. *BMC Medical Informatics and Decision Making, 17*(Suppl 2)(69), 63–70. doi:10.1186/s12911-017-0469-6.

Ducange, P., Pecori, R., & Mezzina, P. (2017). A glimpse on big data analytics in the framework of marketing strategies. *Soft Computing, 22*(1), 325–342. https://doi.org/10.1007/s00500-017-2536-4.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3*, 1289–1305.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys (CSUR), 46*(4).

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems, 6*, 115–125.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications,, 40*, 6266–6282.

Giachanou, A., & Crestani, F. (2016). Like it or not: a survey of twitter sentiment analysis methods. *ACM Computing Survey, 49*(2), 1–28 2841.

Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012). Opinion mining and sentiment analysis on a Twitter data stream. In *Proceedings of International Conference on Advances in ICT for Emerging Regions (ICTer2012)* (pp. 182–188).

Gupta, V., Gurpreet, S., & Lehal, S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence., 1*(1), 60–76.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations Newsletter, 11*, 10–18.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Ji, X., Chun, S. A., Wei, Z., & Geller, J. (2015). Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining, 5*(1), 25. doi:10.1007/s13278-015-0253-5.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of.11th Conference on Uncertainty in Artificial Intelligence* (pp. 338–345).

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation, 13*(3), 637–649.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning, 95*(1-2), 161–205.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.
Li, Y.-M., & Li, T.-Y. (2013). Deriving market intelligence from microblogs. *Decision Support Systems, 55*, 206–217.
Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya, & F. J. Damerau (Eds.), *Handbook of natural language processing* (Second Edition). Taylor and Francis Group, Boca.
Liu, B. (2015). *Sentiment analysis: mining opinions, sentiments, and emotions*. New York, NY, USA: Cambridge University Press ISBN: 9781107017894.
Mccallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *AAAI Workshop on Learning for Text Categorization*.
Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: a survey. *Ain Shams Engineering Journal, 5*(4), 1093–1113.
Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*, 39–41.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).
Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*(10), 4241–4251.
Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 31–41).
Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT), 17*(3).
Nguyen, D. T., & Jung, J. E. (2017). Real-time event detection for online behavioral analysis of big social data. *Future Generation Computer Systems, 66*, 137–145.
Ortega, R., Fonseca, A., & Montoyo, A. (2013). SSA-UO: Unsupervised Twitter sentiment analysis. In *Proceedings of the Second Joint Conf. on Lexical and Computational Semantics*.
Ortigosa, J. M. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior, 31*, 527–541.
Park, J., Barash, V., Fink, C., & Cha, M. (2013). Emoticon Style: Interpreting Differences in Emoticons Across Cultures. In *Proceedings of the Seventh Int. AAAI Conf. on Weblogs and Social Media*.
Pennington, J., Socher, R., & Manning, C. (2014). Glove: global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing *(EMNLP)*.
Patil, L. H., & Atique, M. (2013). A novel feature selection based on information gain using WordNet. In *2013 Science and Information Conference* (pp. 625–629).
Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods: support vector learning* (pp. 185–208). Cambridge, MA: MIT Press.
Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.
Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Mateo, CA: Morgan Kaufmann.
Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science, 5*, 23.
Rosenthal, S., Noura, F., & Preslav, N. (2017). SemEval-2017 task 4: sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.
Saleh, Rushdi, M. , Martín-Valdivia, T. , M., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications, 38*(12), 14799–14804.
Saif, H., Yulan, H., & Alani, H. (2012). Alleviating data sparsity for twitter sentiment analysis. In *21st Int. Conf. on the World Wide Web* (pp. 2–9).
Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering, 25*(4), 919–931.
Salathé, M., Vu, D. Q., Khandelwal, S., & Hunter, D. R. (2013). The dynamics of health behavior sentiments on a large online social network. *EPJ Data Science, 2*(1), 1–12. doi:10.1140/epjds16.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*, 513–523.
Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 1555–1565). *Volume 1: Long Papers*.
Tellez, E. S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Siordia, O. S., & Villaseñor, E. A. (2017a). A case study of Spanish text transformations for twitter sentiment analysis. *Expert Systems with Applications, 81*, 457–471.
Tellez, E. S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Suárez, R. R., & Siordia, O. S. (2017b). A simple approach to multilingual polarity classification in Twitter. *Pattern Recognition Letters, 94*, 68–74.
Uysal, A. K., & Yi Lu, M. (2017). Sentiment classification: feature selection based approaches versus deep learning. In *Proceedings of 2017 IEEE International Conference On Computer and Information Technology (CIT)*.
Valdivia, A., Luzíón, M. V., & Herrera, F. (2017). Neutrality in the sentiment analysis problem based on fuzzy majority. In Proceeding of the IEEE International Conference on Fuzzy Systems, *Naples* (pp. 1–6).
Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A system for real-time Twitter sentiment analysis of 2012U.S. presidential election cycle. In *Proceedings of the ACL System Demonstrations (ACL)* (pp. 115–120).
Wang, P., Xu, B., Xu, J., Tian, G., Liu, C. L., & Hao, H. (2016). Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing, 174*, 806–814.
Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80–83.
Xiong, S., Hailian, L., Weiting, Z., & Donghong, J. (2018). Towards Twitter sentiment classification by multi-level sentiment-enriched word embeddings. *Neurocomputing, 275*, 2459–2466.
Yang, X., Craig, M., & Iadh, O. (2017). Using word embeddings in twitter election classification. *Information Retrieval Journal, 21*(2-3), 1–25.
Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *WIREs Data Mining Knowledge Discovery, 8*(4), E1253.
Zhou, X., Tao, X., Yong, J., & Yang, Z. (2013). Sentiment analysis on tweets for social events. In *Proceedings of the 2013 IEEE 17th International. Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 557–562).