Emotion-Aware Monitoring of Users' Reaction With a Multi-Perspective Analysis of Long- and Short-Term Topics on Twitter

Danilo Cavaliere[1], Giuseppe Fenza[1]*, Vincenzo Loia[1], Francesco Nota[2]

[1] Dipartimento di Scienze Aziendali e Management Information Systems (DISA-MIS), University of Salerno, Salerno (Italy)
[2] Defence Analisys & Research Institute, Center for Higher Defence Studies, 00165 Rome (Italy)

Received 28 October 2022 | Accepted 23 December 2022 | Published 1 February 2023

Abstract

Social networks, such as Twitter, play like a disinformation spread booster giving the chance to individuals and organizations to influence users' beliefs on purpose through tweets causing destabilization effects to the community. As a consequence, there is a need for solutions to analyse users' reactions to topics debated in the community. To this purpose, state-of-the-art methods focus on selecting the most debated topics over time, ignoring less-frequent-discussed topics. In this paper, a framework for users' reaction and topic analysis is introduced. First the method extracts topics as frequent itemsets of named entities from tweets collected, hence the support over time and RoBERTa-based sentiment analysis are applied to assess the current topic spread and the emotional impact, then a time-grid-based approach allows a granule-level analysis of the collected features that can be exploited for predicting future users' reactions towards topics. Finally, a three-perspective score function is introduced to build comparative ranked lists of the most relevant topics according to topic sentiment, importance and spread. Experiences demonstrate the potential of the framework on IEEE COVID-19 Tweets Dataset.

Keywords

Frequent Itemsets, Multi-perspective Topic Monitoring, Sentiment Analysis, Users' Reaction Prediction.

DOI: 10.9781/ijimai.2023.02.003

I. Introduction

In these days, fast and easy access to the Internet allows everyone to express their own ideas about different topics (politics, social events, etc.) by writing posts on social networks (e.g., FB posts, tweets) or even through their own blog or site that can reach millions of people worldwide. On one hand, this phenomenon certainly encourages the freedom of speech, but on the other hand, it leaves society vulnerable and helpless against possible news manipulation coming from multiple sources. Recent examples are the large-scale disinformation and misinformation related to Brexit and US presidential campaign in 2016 or the global infodemic associated to the Covid-19 pandemic and vaccine campaign. Social networks increase the effectiveness and scale of disinformation, that is tailor-made to manipulate users' beliefs on purpose by exploiting persuasive and propaganda techniques. Real and fake news attract users' attention by leveraging their psychological and emotional states leading them to react by expressing their own opinions on social networks. Among the most popular social networks, Twitter is vastly adopted to express reactions towards news and events, therefore, tweet mining could serve as a tool to assess the public opinion about news. Countering information disorder requires monitoring the evolution of trendy topics in order to early alert policymakers and governments about potential risks enabling them to mitigate the impact and harms. Therefore, automatic tools for preventing the spread of news are demanded to effectively contrast the phenomenon that can lead to strong disagreement among people and violent protests as an effect in the worst cases.

* Corresponding author.

E-mail address: gfenza@unisa.it (Giuseppe Fenza).

Canonical disinformation (e.g., fake news and hoaxes), as well as propagandistic tweets, appeal to user's sentiments to influence his/her opinions with the aim of creating confusion or satisfying the writer's intentions that may be of a different kind (e.g., political, social, economic, etc.). As a consequence, the analysis of people's emotional reactions can help monitoring strong reactions towards certain news and act as an early-stage signal to prevent massive disinformation spread. Since news is meant to provoke emotional reactions, the analysis of the emotional aspect is crucial. For this purpose, Sentiment Analysis provides tools to identify, extract, quantify, and study affective states and subjective information by using computational and linguistics techniques, including natural language processing, text analysis, computational linguistics, and biometrics [1].

Beyond the single-post linguistic analysis, time-sensitive, continuous, and heterogeneous information spreading should be constantly analysed since it represents one of the primary issues to the development of good-performing online information monitoring systems [2]. As a consequence, to fight against online fake news spreading, models and new monitoring tools are required to capture the dynamic nature of online information to promptly detect eventual attacks aimed at generating cognitive vulnerabilities and society destabilization effects.

For this purpose, a crucial point to online information monitoring is the analysis of topics debated over time in a community, by taking into consideration not only those ones discussed for long time, but also those topics debated in a restricted period of time, that may influence users' behaviours and thinking as well as the long-time-debated ones. This paper presents an emotion-aware solution to analyse users' reactions towards topics that have been constantly discussed over time (long-term topics) and topics that have been discussed in a specific brief period (short-term topics). The rationale behind the approach is to combine the emotional analysis of tweet content with the time frequent analysis of relevant topic itemsets and tweet spread to better evidence those topics that may have the strongest impact on the community. In other words, a mechanism is introduced to rank topics that may cause community destabilization effects by jointly considering topic sentiment, importance and spread. In detail, the approach collects tweets day-by-day in a reference period, extracts short- and long-term itemsets of topics from tweets, then evaluates topic mentions and extracts a sentiment score on four different emotional classes to depict users' reactions to topics. The approach is based on a granular time-grid-based data processing schema, that allows the emotion-aware analysis of the extracted short- and long-term topics which can be used for community monitoring, including the prediction of emotional users' reactions towards topics. The final and ultimate goal of the framework is a multi-perspective topic relevance analysis to provide ranked lists of topics in accordance with topic sentiment, importance, and spread.

In a nutshell, the paper contribution can be summarized in the following points:

- A report on multi-class emotional analysis of Twitter users' reactions showing that short-living topics, which are often discarded, may generally cause great emotional effects on community.
- The proposal of a time-grid-based approach to track topic mentions and their emotional impact over time, aimed at helping the detection of high-impact topics.
- The design of a time granular emotion-aware topic modeling to serve the collected information reuse for different tasks, including the prediction of eventual future users' reactions.
- The introduction of a score function combining topic sentiment, mention frequency and spread to perform a multi-perspective topic relevance analysis by comparing score-based topic impact ranked lists.

The paper is organized as follows: Section II provides the preliminaries to the research and discusses the related work, Section III describes the approach in detail, Section IV shows a case study related to the proposed approach, and Section V reports on the test conducted and conclusions close the paper.

II. Preliminaries

A. A preliminary Study

As a preliminary step, a careful analysis of Twitter users' emotional reactions has been carried out with the aim of finding out meaningful features of the topics debated, the intensity of the emotions they have aroused in the community and the durability of these emotional reactions. The rationale behind this preliminary study is to determine the effective impact that some short-term topics may have on the community and whether they are of interest for analysis of users' future behaviors and mitigation action planning.

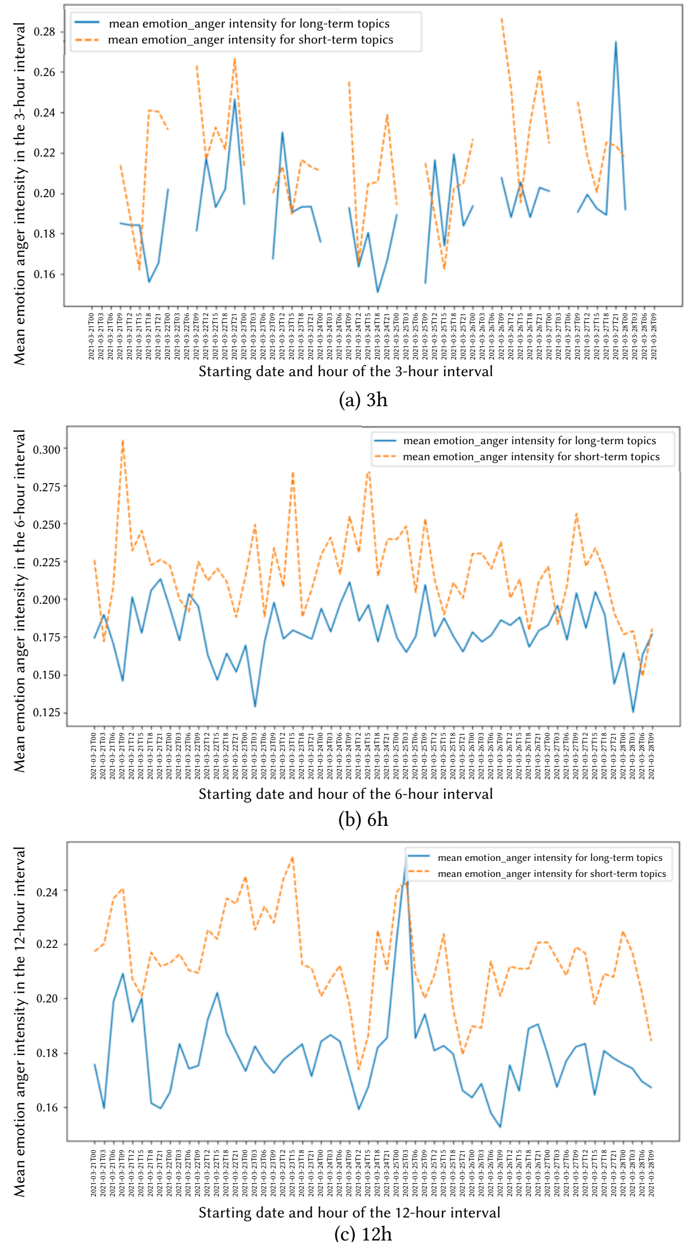


Fig. 1. Twitter users' averaged sentiment scores on anger class at different hour.

To analyse Twitter users' emotional reactions, tweets have been collected on a 3-months period from February 2021 to April 2021, hence tweet content has been extracted and processed with Sentiment Analysis to determine the emotional intensity (or score) on four emotions: sadness, anger, joy, and optimism. The sentiment scores have been analysed over time by using different time units, in detail the average sentiment score for each emotional class has been assessed each 3, 6, and 12 hours on short- and long-term topics, distinctively. This way, a time granular analysis of the emotional reactions has been conducted to evidence the emotional impact related to the debates on short-term topics.

Since from the analysis of the four emotional classes we can state analogous conclusions, for sake of simplicity results are reported for the emotional class anger in Fig. 1 on a period going from 03/04/2021 to 20/04/2021. The figure shows the average sentiment scores of the tweets associated with topics for the anger class on the short- and long-term topics as the 3-hour, 6-hour, and 12-hour analyses. As a general trend, let us say that short-term topics cause higher peaks in negative emotions (i.e., high curve peaks). In some emotional classes, the sentiment score of some emotions on short-term topics is constantly higher than the long-term topic sentiment score. For example, long-term topic mean anger intensity is statistically lower than short-term topics mean anger intensity (with a p-value less than 0.00001). This result proves that the undoubted influence of short-term topics on users' emotional reactions may lead the community to suffer from emotional destabilization phenomena and that monitoring emotions is fundamental to plan mitigation tools for economic and political analyses.

From the figure, let us notice that the 6-hour interval analysis uniformity seems to be a rational trade-off between the 3-hour analysis (excessive variability) and the 12-hour analysis (accentuated flatness).

B. Related Work

The spread of misinformation is generally related to hot topics (e.g., Covid-19 virus spread and vaccines), that are subject of different analyses aimed at finding out important trends, such as a decline in the number of vaccine supporters caused by the spread of fake news on vaccines [3]. Many works in the literature focus on misinformation and disinformation detection from texts [4]–[7]. In [4], the authors present a method to label a dataset on propagandistic text, run topic modelling, and then corpora imbalance assessment for propaganda detection. In [5], different techniques (e.g., GloVe, BERT, and LSTM) are combined to perform word representation, pretrain the model and detect persuasive text. Another approach [6] allows fragment-level text analysis which exploits tf/idf, word, and character n-grams to build a classifier for propagandistic text. In [7], the authors present a Machine Learning framework for article- and sentence-level persuasive text detection. Other works analyse disinformation on social networks to help find countermeasures, such in [8] where the authors propose a framework exploiting activity-connectivity maps based on network and temporal activity patterns to detect social influence among ISIS supporters.

Another predominant vein in literature is aimed at finding solutions to deal with fake news detection from social networks. In [9], the authors collect news from heterogeneous sources and test out different Machine Learning methods for fake news detection. Some works investigate sentiment analysis, such as the method proposed in [10], which explores sentiment analysis and determines the most relevant elements for fake news detection, including multilingualism, explainability, bias mitigation, and multimedia element treatment. In [11], the authors proposed an attention-based approach for multi-modal sentiment analysis. In [12], sentiment extracted from news is coupled with domain parameters to improve predictions. Several works focus on Twitter misinformation spread monitoring, such as the approach proposed in [13], which checks emotional valence in relation to false stories certified by Google Fact Checker API on Covid-19, and finds out that the emotional valence varies depending on the different topics. Another work [14] analyses changes in Twitter users' behaviours after a misinformation attack by analysing variations with respect to frequency and sentiment expressed in their tweets. The solution proposed in [15] performs topic identification, network analysis, and sentiment analysis to classify tweets into the six categories for misinformation analysis. It comes out that sentiment score is mainly influenced by government measures and public speeches from government officials, as well as news agencies and public figures. In detecting misinformation, Sentiment Analysis is applied to news and social network posts with different intents, such as the approach presented in [16], which assesses text sentiment-harmful news correlation and detects harmful news through sentiment analysis.

Some other works are focused on Twitter data analysis aimed at extracting meaningful patterns for users' behavior extraction or prediction, such as the work proposed in [17], which presents a Latent Dirichlet Allocation-based model to analyse people's reactions to Covid-19 from tweets, the method proposed in [18] that introduces a method to mine association rules from tweets and extract people's attitudes to topics, and the approach proposed in [19] that coupled topic identification and sentiment analysis to extract emotional people reactions from multi-lingual tweets. All these methods analyse exclusively general macro topics (e.g., politics, economic impacts, etc.) and focus on time evaluation that missed the momentary strong emotional impact that a topic may have on communities, and consequently fail in capturing a reliable topic relevance evaluation. To deal with this challenge, a granular time-grid hierarchical approach has been designed to process tweets, extract long-time and short-time-debated topics alongside their parameters (sentiment, topic mention frequency, tweet count), and evaluate them in time granules (grid cells) in order to achieve a better topic relevance assessment.

III. The Methodology

The framework has been designed as a three-tier model allowing sentiment-aware topic extraction, time granular emotion-aware analysis of users' reaction to topics and, a three-perspective topic relevance analysis. The complete pipeline is shown in Fig. 2, where the first tier includes tasks to perform *Tweet collection and topic extraction*, which lead to building datasets of tweets collected, pre-processing their content to extract relevant entities by running Named Entity Recognition (NER), and extract topics as frequent word itemsets from tweets assessing their mention frequency over time. Then, the second tier allows *Users' reaction modeling* by accomplishing several tasks, including topic sentiment analysis aimed at depicting users' reaction towards itemsets and a time-granular emotion modeling that can serve the prediction of users' future reactions towards topics.

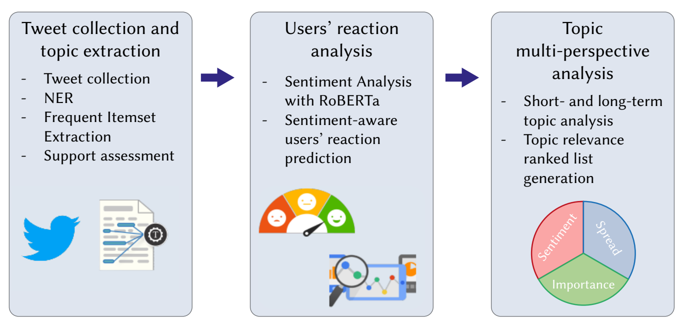


Fig. 2. Architecture pipeline.

The third tier is in charge of *Topic multi-perspective analysis* by running several tasks aimed at analysing the relevance of short- and long-term topics in the community by considering topic sentiment, importance, and spread in the community.

A. Tweet Collection and Topic Extraction

Tweets are collected to build a tweet dataset representing a multiplicity of topics. To polish the data acquired, widely-used Python libraries are exploited to perform state-of-the-art pre-processing tasks, including stop word, special characters and punctuation removal, tokenization, and Part-of-Speech (POS) tagging by using nltk[1], spaCy[2] and stanza[3]. The preprocessing task is closed by removing URLs from texts. To detect the effective entities debated in tweets, Named Entity Recognition (NER) is performed. Therefore, Stanza NER is then applied to hashtags and keywords to extract relevant names of people, organizations and locations discussed in tweets.

[1] https://www.nltk.org/

[2] https://spacy.io/

[3] https://stanfordnlp.github.io/stanza/

Then, topics are extracted and discriminated into two specific categories: short- and long-term topics; the former refers to those topics that are constantly debated over time (high mention frequency over time), and the latter represents those topics that have been debated within a limited time interval (i.e., they generally last few hours, or a day at maximum, or they are debated only a few times in different time periods). Since short-term topics seem to cause the strongest emotional reactions among Twitter users, as has been found out in the preliminary study (Section A), the main rationale behind short-term/long-term discrimination is to detect those events causing the strongest users' reactions that, as a consequence, may lead to community destabilization effects.

In order to define topics, frequent word itemsets are built as composed of hashtags, keywords, mentions, and NEs extracted from tweets. In detail, itemsets are defined as 2-grams and 3-grams representing debated topics in the community by using the Python library fpgrowth. Then the support, defined as the occurrence frequency of a pattern in a dataset, is calculated. In this work the patterns are 2-grams and 3-grams, therefore the support is their occurrence frequency in the tweets dataset [20].

To allow a more robust topic analysis over time, the features of the extracted itemsets (e.g., support, tweet frequency, etc.) have been analysed by means of time granules. For this purpose, a time grid schema is defined to analyse the itemset support with respect to time at this stage, therefore, the time interval considered for tweet collection is represented as a grid having cells of fixed size expressed in hours. For instance, a 6-hours-cell in the time grid will report the time support for each itemset decomposed in each of the 6 hours of the time interval considered. The grid helps focus on the most debated topics in a specific moment in time (i.e., itemsets with high support in specific grid cells). To discriminate between long- and short-term itemsets/topics, a support-based filtering function is applied to itemsets separating those itemsets debated over a long time from those that have a localized time occurrence. The support-based filtering function is reported as Algorithm 1; it takes the topic *T*, a counter *C* and the time grid cell length in hours (*H*), hence it discards *T* if it has a low support (lines 3-5) otherwise it increases *C* by 1 or decreases it by **H/12** depending on *T* support in each cell (lines 6-12). Then, it checks whether *T* is a short- or a long-term topic by applying the sigmoid function to the base two logarithm of counter *C* (lines 13-17). The rationale behind the logarithmic function is that it grows slowly allowing topics to be considered as long-term only if they are present (i.e. meaning their support is high) in many time-grid cells. Moreover, when a long-term topic becomes less present in tweets, the logarithm ensures a slow and gradual decrease when applied on the counter. Therefore, the longer the former long-term topic was discussed, the more time it will take to become short-term due to the decrease strategy. Conversely, if a topic has been debated for a while, but not for so long (e.g. one month), when it is not debated anymore it will come back faster to the state of short-term topic (or absent state in case it is not referred to at all). The sigmoid function helps achieve a clearer interpretation of the logarithmic function applied to the counter by forcing the score to lie within the range 0.5 to 1. For this reason, a threshold fixed to the half of the range (i.e. 0.75), has been used to discriminate between short- and long-term topics.

Algorithm 1. Support-based filtering on itemsets

1: Let T be a topic, C a counter and H the number of hours of the interval

Require: $T \geq 0$, $C \geq 0$ and $H \geq 3$

Ensure: `T is a short-term or long-term topic`

2: **function** SUPPORT_FILTER(T,C)

3:   **if** SUPPORT(T) ≤ 0.001 **then**

4:     `T is discarded`

5:   **end if**

6:   **for** `each time grid cell` **do**

7:     **if** SUPPORT(T) > 0.005 then

8:       $C \leftarrow C + 1$

9:     **else if** SUPPORT(T) ≤ 0.005 **then**

10:       $C \leftarrow C - \frac{H}{12}$

11:     **end if**

12:   **end for**

13:   **if** SIGMOID($\log_2 C$) ≤ 0.75 **then**

14:     `T is a short-term topic`

15:   **else if** SIGMOID($\log_2 C$) ≥ 0.75 **then**

16:     `T is a long-term topic`

17:   **end if**

18: **end function**

B. Users' Reaction Prediction

The second tier allows topic sentiment analysis and the prediction of users' emotional reactions towards the extracted topics.

For this purpose, first, Sentiment Analysis is exclusively applied to tweets related to topics selected through Algorithm 1. The sentiment analysis outcome on these tweets will be used to perform the sentiment analysis of the selected itemsets. To perform Sentiment Analysis, our framework employs RoBERTa [21], the robustly optimized pretraining approach of the famous NLP transformer-based machine learning technique Google BERT. RoBERTa-based model allows text Sentiment Analysis in terms of polarity classification and single emotional class classification, namely optimism, joy, sadness and anger. Once, RoBERTa Sentiment Analysis has been applied to tweets in a specific time-grid cell of duration *H* (e.g. 6 hours), the sentiment score for a topic *T* is calculated as the mean of the sentiment scores achieved for each tweet in which *T* is present. This is done for each of the four RoBERTa emotional classes, hence for each of the four emotions describing users' emotional reactions and for each time-grid an intensity score is associated for topic *T* . For example, the intensity scores are 160 if there are 4 time grid cells, which length is 6-hours each, in a total period of 10 days considered for all the 4 emotions (i.e. 4 x 10 x 4 = 160).

To deal with the analysis of eventual users' reactions towards topics, a regressor module has been designed to process sentiment related to itemsets and accordingly predict how community users react to them. The regressor has been designed on the time grid introduced before (see Section A), where the support and sentiment on four classes are reported for each itemset in consecutive *s*-hour time-grid cells, where *s* is the number of hours considered as a step. The mean value of each emotional class and support is computed in each step. Then, given the mean sentiment value for each emotional class in the current and previous steps, the regressor goal is to predict the mean value of sadness, anger, optimism, and joy in the next step.

C. Topic Multi-Perspective Analysis

Since experts may be interested in analysing topic relevance from different perspectives, the last tier allows a synergistic multi-perspective analysis of the topics extracted. For this purpose, the topic is analysed by means of the three parameters: sentiment, importance and spread in the community. The sentiment is represented by the sentiment score assessed on the topic for a specific emotional class, as it has been defined in Section B, the importance is based on the topic

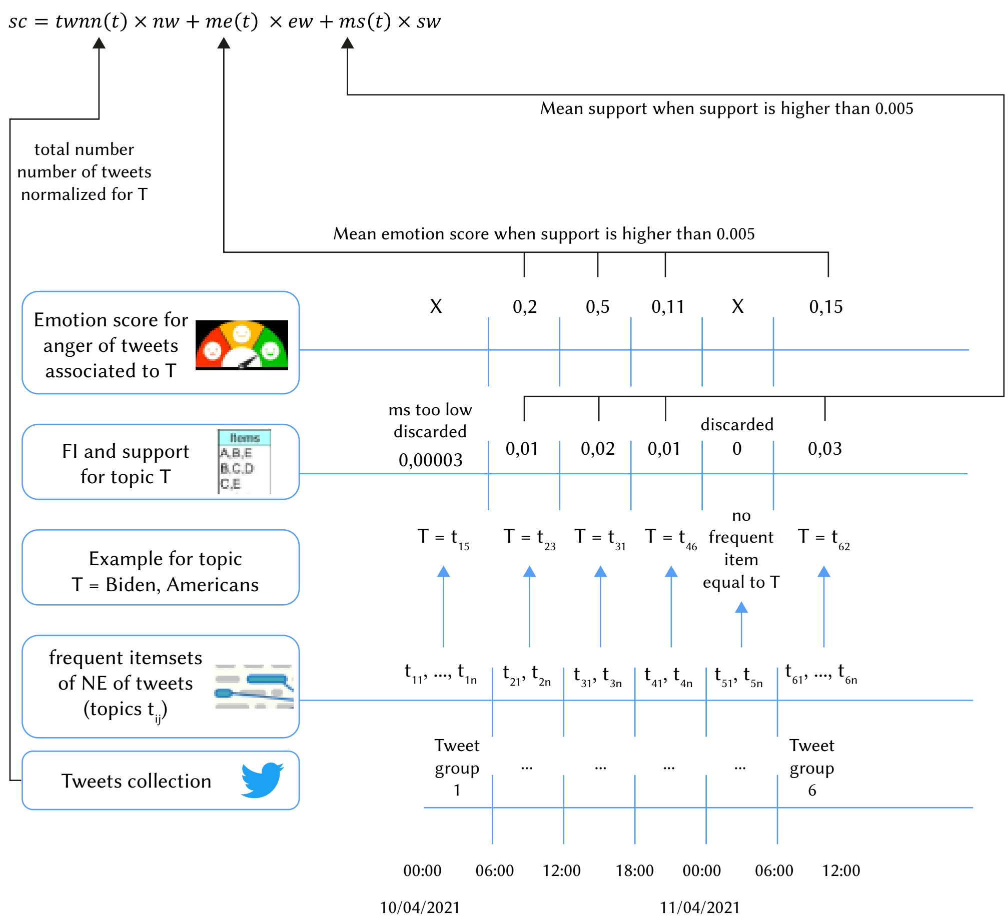


Fig. 3. Biden-americans: an example of the framework processing a topic.

support metric that represents the frequency of the topic over time compared to the other itemsets extracted from tweets (see Section A), and the topic spread evaluation is based on the number of topics in which the topic is present.

To provide a joint evaluation of topic relevance in the community, this tier employs a merged score function that combines the three above-mentioned parameters in a unique value depicting topic relevance. The merged score function is defined as a weighted function combining topic support, emotional class mean score, and the number of tweets in which the topic appears. Formally, let t be an itemset of terms, the merged score is assessed with the following weighted function:

$$sc = sw \cdot ms(t) + ew \cdot me(t) + nw \cdot twnn \quad (1)$$

where sw is the support weight, ms is mean support on all time-grids, ew is the weight associated to the average sentiment score, me is the average sentiment score over all time-grids, $twnn$ is the normalized number of tweets in which the t is present over all time-grids and nw is the weight associated to $twnn$. The average sentiment score (me) can be assessed for each of the four emotional classes considered by the framework.

The merged score is assessed for each topic among short- and long-term topics, hence a ranked list based on this score is generated to detect the most relevant short- and long-term topics among the extracted ones in the reference period. Moreover, experts can manually choose different weight values to generate and compare different views about the topic relevance. In fact, if experts want to focus on the emotional aspect, they can increase the weight associated with the mean emotional score (sw) and look out for those topics with the highest emotional impacts, without ignoring the effective importance, and spread of the topic. The merged score allows a better dynamic analysis of topic relevance in the community, as it will be shown in the next section.

To summarize the whole framework functioning, Fig. 3 shows an example of running the framework in a time reference period displaying how the score is assessed for a specific topic T. The blue lines represent the time grid schema for data processing and the black lines show the stage at which the three merged function parameters are calculated or retrieved. In detail, first, the tweets are collected and arranged in groups in each time interval of the grid, hence the normalized number of tweets is assessed from data so as to be used in the merged score calculation. In the second stage, named entities are extracted from each group in the reference interval and arranged in frequent itemsets (FIs) representing the topics, thus the support for a topic T is calculated in each time interval allowing the calculation of the mean support (ms) that will be used for the topic T merged score calculation. Then, the sentiment score is calculated for an emotional class (anger in this example) so that the mean of the sentiment scores (me) for topic T over the intervals considered can be assessed and used for the final merged score calculation.

IV. A Case Study on Long- and Short-Term Topics

To show the potential of the proposed framework, this section presents a real case scenario carried out by running the whole framework on tweets about the Covid-19 pandemic. For the sake of simplicity, we collected tweets with hashtags related to the Covid-19 general topic in the period going from 06/02/2021 to 14/02/2021. Then, data have been processed by following the three-tier pipeline in Fig. 2 and the frequent itemsets of named entities representing topics extracted from tweets have been generated.

Since the main aim of the case study is to show how the proposed framework allows a comparative analysis of the effective short- and long-term topic relevance, the topic impact in terms of support score and sentiment score of every single topic extracted is evaluated through the merged metric (Eq. 1) that has been defined in Section C. Thanks to the merged score function, the framework returns two ranked lists of topics, one for the short-term and the other one for the long-term topics that have caused the strongest reactions among users and have been among the most discussed ones in the reference period.

The topics extracted by the framework have been ranked by using the merged score for the anger emotional class and shown in Tables I, II, and III for long-term topics and Tables IV, V and VI for short-term topics. Three different ranked lists are generated for each class of topic (i.e., short- and long- term topics) by setting different weight values for the three parameters, including anger emotion weight (*ew*), support weight (*sw*) and normalized number of tweets weight (*nw*).

Let us notice some difference between ranked lists of long- and short-term topics. For what concern the long-term ones, the highest-scored topic is *biden-americans* on most of the lists (see Tables II and III), which is a very general topic debated in 5,986 tweets, and to which users react with low anger (0.20). Since this topic is very general, it is not easy to associate it to news released in the reference period. However, the wide spread of the topic means that slight changes in negative emotions could influence a high number of people.

TABLE I. Long-term Topic Ranked List Focused on Sentiment ($sw = 0.2$, $ew = 0.6$, $nw = 0.2$)

keys	*sc*	*me*	*ms*	*twnn*
republicans-rwpusa	0.545	0.815	0.014	0.266
senate-republicans	0.521	0.747	0.015	0.349
biden-americans	0.322	0.199	0.013	1.0
senate-eugene goodman fri-michaelart123	0.245	0.406	0.006	0.0
rt-rand paul	0.245	0.405	0.007	0.0

Legend: *sc* is the merged score, *me* is the average sentiment score over all time-grids, *ms* is the mean support over all time-grids, *twwn* is the normalized number of tweets

TABLE II. Long-term Topic Ranked List Focused on Support ($sw = 0.2$, $ew = 0.6$, $nw = 0.2$)

keys	*sc*	*me*	*ms*	*twnn*
biden-americans	0.248	0.199	0.013	1.0
senate-republicans	0.228	0.747	0.015	0.349
republicans-rwpusa	0.225	0.815	0.014	0.266
rt-rand paul	0.085	0.405	0.007	0.0
senate-eugene goodman fri-michaelart123	0.085	0.406	0.006	0.0

TABLE III. Long-Term Topic Ranked List Focused on the Number of Tweets in Which the Topic Is Present ($sw = 0.2$, $ew = 0.2$, $nw = 0.6$)

keys	*sc*	*me*	*ms*	*twnn*
biden-americans	0.642	0.199	0.013	1.0
senate-republicans	0.362	0.747	0.015	0.349
republicans-rwpusa	0.326	0.815	0.014	0.266
senate-eugene goodman fri-michaelart123	0.082	0.406	0.006	0.0
rt-rand paul	0.082	0.405	0.007	0.0

TABLE IV. Short-term Topic Ranked List Focused on Sentiment ($sw = 0.2$, $ew = 0.6$, $nw = 0.2$)

keys	*sc*	*me*	*ms*	*twnn*
china-junta	0.634	0.955	0.018	0.29
asian-iamcindychu-non-asian	0.631	0.914	0.006	0.406
2021-junta	0.628	0.953	0.024	0.259
china-2021	0.624	0.944	0.024	0.263
china-myanma-2021	0.611	0.952	0.02	0.181
angelarayner-british	0.573	0.907	0.005	0.141
asian-asian americans	0.549	0.869	0.008	0.131
vp-asian	0.549	0.868	0.008	0.132
vp-asian americans	0.547	0.864	0.008	0.133
trump-covid	0.52	0.575	0.007	0.871
covid-billienomxtes	0.49	0.695	0.009	0.355
dwuhlfelderlaw-ron desantis	0.483	0.746	0.006	0.171
biden-florida	0.478	0.701	0.006	0.284
biden-trump	0.401	0.397	0.008	0.803
super bowl-dwuhlfelderlaw-tampa	0.377	0.585	0.008	0.125
super bowl-tampa	0.367	0.564	0.008	0.136
trump-gop	0.354	0.434	0.007	0.462
doctorpisspants-20s	0.353	0.444	0.011	0.422
trump-americans	0.344	0.401	0.008	0.511
0-daliagebrial	0.322	0.435	0.01	0.296
kylegriffin1-biden	0.303	0.295	0.013	0.616
senate-rand paul	0.3	0.436	0.008	0.184
senate-rt	0.293	0.42	0.008	0.196
a year ago today-gtconway3d	0.285	0.404	0.007	0.205
emekamba-nigeria	0.278	0.411	0.014	0.141

TABLE V. Short-Term Topic Ranked List Focused on Support ($sw = 0.6$, $ew = 0.2$, $nw = 0.2$)

keys	*sc*	*me*	*ms*	*twnn*
trump-covid	0.293	0.575	0.007	0.871
iamcindychu-asian	0.268	0.914	0.006	0.406
china-junta	0.26	0.955	0.018	0.29
2021-junta	0.257	0.953	0.024	0.259
china-2021	0.256	0.944	0.024	0.263
biden-trump	0.245	0.397	0.008	0.803
china-myanma	0.238	0.952	0.02	0.181
covid-billienomxtes	0.216	0.695	0.009	0.355
angelarayner-british	0.213	0.907	0.005	0.141
asian-vp-asian americans	0.205	0.869	0.008	0.131
vp-asian	0.205	0.868	0.008	0.132
vp-asian americans	0.204	0.864	0.008	0.133
biden-florida	0.201	0.701	0.006	0.284
kylegriffin1-biden	0.19	0.295	0.013	0.616
trump-americans	0.187	0.401	0.008	0.511
dwuhlfelderlaw-ron desantis	0.187	0.746	0.006	0.171
trump-gop	0.184	0.434	0.007	0.462
doctorpisspants-20s	0.18	0.444	0.011	0.422
potus-america	0.178	0.161	0.008	0.706
0-daliagebrial	0.152	0.435	0.01	0.296
super bowl-dwuhlfelderlaw-tampa	0.147	0.585	0.008	0.125
super bowl-tampa	0.145	0.564	0.008	0.136
gop-americans	0.145	0.316	0.007	0.386
three weeks ago-america	0.135	0.206	0.01	0.439
supe-anildash	0.135	0.36	0.009	0.291

TABLE VI. Short-term Topic Ranked List Focused on Sentiment (sw = 0.2, ew = 0.6, nw = 0.2)

keys	*sc*	*me*	*ms*	*twnn*
trump-covid	0.639	0.575	0.007	0.871
biden-trump	0.563	0.397	0.008	0.803
potus-america	0.457	0.161	0.008	0.706
kylegriffin1-biden	0.431	0.295	0.013	0.616
asian-non-asian	0.428	0.914	0.006	0.406
trump-americans	0.388	0.401	0.008	0.511
china-junta	0.369	0.955	0.018	0.29
trump-gop	0.365	0.434	0.007	0.462
covid-billienomxtes	0.354	0.695	0.009	0.355
china-2021	0.351	0.944	0.024	0.263
2021-junta	0.35	0.953	0.024	0.259
doctorpisspants-20s	0.344	0.444	0.011	0.422
biden-florida	0.312	0.701	0.006	0.284
three weeks ago-america	0.307	0.206	0.01	0.439
myanma-2021	0.303	0.952	0.02	0.181
potus-covid	0.301	0.134	0.007	0.455
gop-americans	0.296	0.316	0.007	0.386
biden-gop	0.279	0.286	0.008	0.367
barbados-india	0.27	0.182	0.006	0.387
india-justintrudeau	0.269	0.141	0.007	0.4
angelarayner-british	0.267	0.907	0.005	0.141
0-daliagebrial	0.267	0.435	0.01	0.296
asian-asian americans	0.254	0.869	0.008	0.131
vp-asian americans	0.254	0.864	0.008	0.133
vp-asian	0.254	0.868	0.008	0.132

Now, let us pick the short-term topics with a higher merged score (0.63) in Table IV, which are about Myanmar junta, China and year 2021 (e.g., *china-junta*, *china-myanma*, *china-myanma-2021*, etc.). Even though these topics are discussed for a short time and do not have high mean support, Twitter users react to them by expressing considerable anger (me= 0.95). Moreover, these topics identify specific news released in that period about the Myanmar military junta that took power with a coup d'etat considering the Covid-19 pandemic restrictions as a violation imposed by the State Counsellor of Myanmar, and then killed and tortured hundreds of civilians, including children[4].

The assignment of different weights to the merged score function contributes to build three different views on topics focusing on a specific concept but taking into consideration the others at the same time. This mechanism allows a comparative analysis that can indeed serve the detection of topics to pay attention to for avoiding destabilization effects:

1. Higher emotion weight: by giving higher weight to the sentiment score, the ranked lists (Tables I and IV) highlight those topics to which users react in the strongest way that may differ from the most-discussed topics (i.e., ranked lists focused on support and number of tweets). In fact, the long-term topic that caused the strongest users' emotional reactions is related to senate and republicans (see Table I), even though the most debated long-term topic is the already-mentioned *biden-americans* (see Tables II and III). The same goes for the short-term topic; the *china-junta-2021*, causing the strongest users' reactions (Table IV), is not discussed as much as the *trump-covid* and *biden-trump* which are the most debated topics(Tables V and VI).
2. Higher normalized tweet number weight: a normalized-tweet-number-focused ranked lists depict the importance of a topic. For instance, let us notice that *biden-trump* and *china-junta* are the topics with 80% and 29% of normalized number of tweets in the reference period (Table VI). Therefore, *biden-trump* has more importance than *china-junta*.
3. Higher support weight: the analysis mainly based on support gives a score to *biden-trump* that is less than the score of *china-junta*, meaning that, in the periods in which the topics are discussed, *china-junta* is more frequent than *biden-trump* and may be considered more important from this perspective (Table V). In other words, even if *china-junta* may be discussed within a smaller number of time-grid cells, during this time its mean frequency is considerably higher than the frequency of *biden-trump*.

[4] https://www.voanews.com/a/myanmar-junta-violations-may-amountto-crimes-against-humanity/6242469.html

The merged score allows to always consider all the parameters in a topic relevance evaluation even though a higher weight may be assigned to one of them. In fact, in the support-focused analysis (Table V), *china-junta* has higher ranking than *biden-trump* that does not only depend on support but also on the emotional score (i.e., *china-junta* causes definitely stronger emotional reactions than *biden-trump*).

V. Experimentation

To test how much the proposed granular time-based framework is good for monitoring users' emotional reactions, a time-based test has been carried out to check out how long the framework has good performance at predicting users' reactions.

A. The Dataset

A Tweet dataset has been considered for tests: Coronavirus (COVID-19) Tweets Dataset [22]. This dataset is composed of CSV files including IDs and sentiment scores of the tweets related to the COVID-19 pandemic. The dataset includes 2,023,557,636 tweets in English, covering a global geographic area and a long period starting with the date of the first tweet on the topic which dates back to October 01, 2019. The real-time Twitter feed is monitored for coronavirus-related tweets using 90+ different keywords and hashtags that are commonly used while referencing the pandemic. For efficiency purposes, a subset of this dataset has been acquired, containing 2,000,000 tweets.

B. Methods

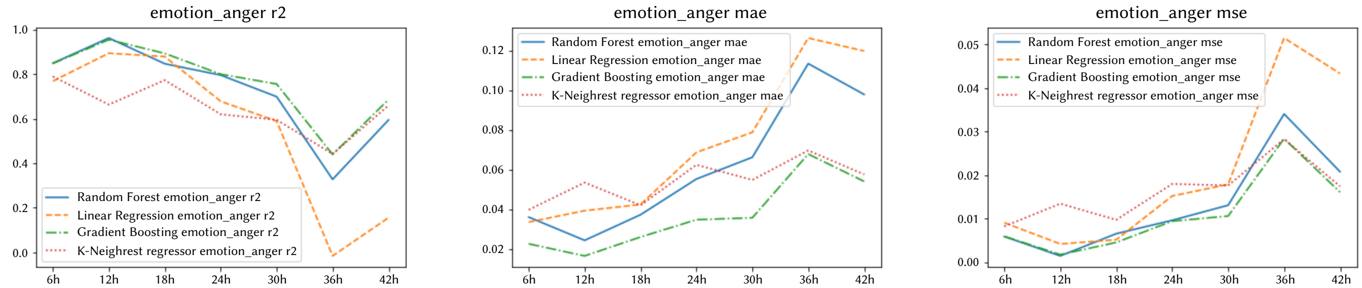
To check the feasibility of using our framework for users' reaction prediction, several state-of-the-art regression methods have been run and compared on the COVID-19 Tweets Dataset. Details on methods are reported below.

- **Linear regression** is a regression model consisting of a predictor variable and a dependent variable related linearly to each other.
- **Random Forest** is an ensemble learning method that averages the predictions made by multiple decision trees to perform regression.
- **Gradient boosting** is a predictive model combining an ensemble of weak prediction models for accomplishing regression.
- **K-nearest regressor** is a non-parametric method that approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

C. Metrics

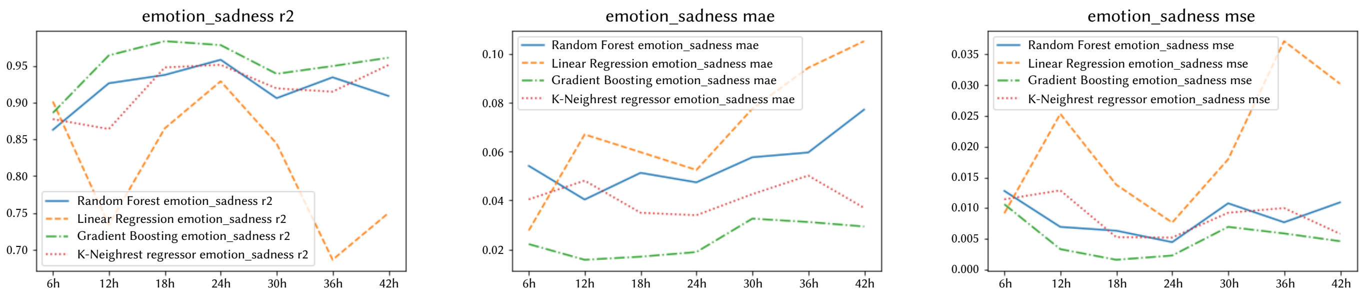
Several metrics have been employed for tests:

- *MSE*. Mean Squared Error (*MSE*) refers to minimizing the mean squared error between predictions and expected values. It is calculated as the mean or average of the squared differences



(a) r2 score (b) mae score (c) mse score

Fig. 4. Test results for the anger class.



(a) r2 score (b) mae score (c) mse score

Fig. 5. Test results for the sadness class.

between predicted and expected target values in a dataset:

$$MSE = \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N} \tag{2}$$

where y_i is the i^{th} expected value in the dataset and $\hat{y}_i$ is the i^{th} predicted value.

- *RMSE.* It is an extension of MSE that returns an error in the same unit of the target value. MSE allows to punish large errors by squaring the error, while RMSE reverses this operation through the square root:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N}} \tag{3}$$

- *MAE.* Mean Absolute Error (MAE) does not give more or less weight to different types of errors, contrary to MSE and RMSE which punish larger errors more than smaller errors. MAE is calculated as the average of the absolute error values:

$$MAE = \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)}{N} \tag{4}$$

- *R^2.* The coefficient of determination or R squared is statistics to evaluate how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. It is calculated by relating the residual sum of squares and the total sum of squares:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2} \tag{5}$$

where $\overline{y} = \frac{\sum_{i=1}^{N} y_i}{N}$.

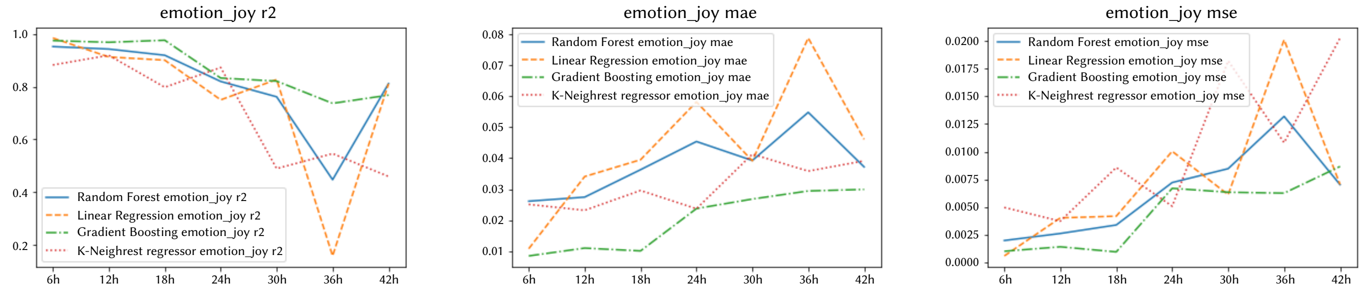
D. Test Results

The framework has been tested with each of the four methods introduced in Section B, then the results have been compared. The regressors have been applied to each of the four emotional classes. The selected dataset has been divided by 66% for training and 33% for test validation phases, hence all the regressors have been applied to predict the mean sentiment score in the next step by considering the mean sentiment score in the previous four cells of the grid. Let us consider a prediction window as the number of cells after which the regressor provides users' sentiment prediction, the test has been designed as an incremental scheme that calculates the four metrics (Section C) on each regressor by incrementally increasing the prediction window by one at each test run. In other words, the regressor performance is evaluated firstly when it predicts the sentiment after one cell, then after two cells, then after three cells, and so on. The maximum number of cells to which regressor prediction accuracy is evaluated is 8. Since each interval is fixed to 6 hours, 8 cells correspond to 48 hours, therefore considering a scale of 8 cells, the regressor will be evaluated at predicting users' reactions after 6 h, 12 h, 18 h, 24 h, 30 h, 36 h, 42 h, 48 h. This time-based test allows to evaluate how long is the regressor good at predicting future users' emotional reactions.

From the figures, let us notice that, in general, Gradient boosting regressor outperforms all the other methods, followed by Random forest and k-nearest neighbor methods. Linear regression is definitely the method with the worst performance. The prediction accuracy over time changes with the emotional class, in detail, let us look Fig. 4c for the anger emotional class, r^2 score decreases after 18 hours, with the best performing method (gradient boosting) going down from 0.90 to 0.60, Random forest and k-nearest neighbor reaching 0.50 after the 30 hours, while Linear regression registers some negative steep dips. Regressors have definitely better performance on the other emotional classes, for sadness the r^2 score of the best-performing methods (e.g., Gradient boosting and Random forest) is constantly high lying in the range r0.9, 1s (Fig. 5c); for joy, r^2 score for three of the methods is high for 30 hours, after that values decrease a bit but lying in the range r0.6, 0.8s (Fig. 6a); for optimism, there is more variability, with the best-performing three regressors decreasing a bit after 12 hours but keeping high accuracy (r^2 score around 0.8) and going higher after 18 hours, but then having a steep dip after the 24 hours.

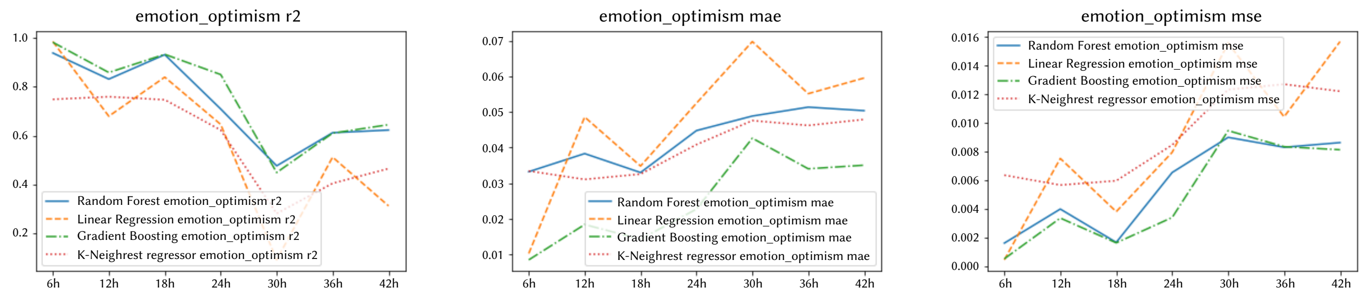
VI. Conclusion

The paper presented an emotion-aware framework for the analysis of long- and short-term topics over time and users' reaction to



(a) r2 score (b) mae score (c) mse score

Fig. 6. Test results for the joy class.



(a) r2 score (b) mae score (c) mse score

Fig. 7. Test results for the optimism class.

topics with the aim of supporting experts and institutions to keep disinformation on social networks under monitoring. This study introduced several important contributions to build monitoring systems for disinformation fighting in online social communities, including:

- A preliminary study showing the relevance of short-time-discussed topics in causing strong negative users' reactions.
- A time-grid-based approach to track topic frequency and emotional impact for the analysis of long- and short-time- debated topics.
- An emotion-aware topic modeling to support monitoring activities over time, including users' future reaction prediction.
- A score function combining topic frequency, sentiment and spread to support a robust multi-perspective topic relevance evaluation.

Future research intents will be focused on automatising some processes, including an automatic weight assignment depending on the context (e.g., social, economical, political, etc.) and the analysis goal to find out the most relevant parameters for the topic relevance evaluation. Future research directions are also targeted at studying echo chamber effects in order to extend the developed short- and long-term topic detection model to help community analysers fight radicalization phenomena.

References

[1] E. Tonkin, "Chapter 2 - a day at work (with text): A brief introduction," in *Working with Text*, E. L. Tonkin, G. J. Tourte Eds., Chandos Information Professional Series, Chandos Publishing, 2016, pp. 23–60, doi: https://doi.org/10.1016/B978-1-84334-749-1.00002-0.

[2] X. Zhang, A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, no. 2, p. 102025, 2020, doi: https://doi.org/10.1016/j.ipm.2019.03.004.

[3] S. Loomba, A. Figueiredo, S. Piatek, K. de Graaf, H. Larson, "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa," *Nature Human Behaviour*, vol. 5, pp. 1–12, 2021, doi: 10.1038/s41562-021-01056-1.

[4] Y. Kirill, I. G. Mihail, M. Sanzhar, M. Rustam, F. Olga, M. Ravil, "Propaganda identification using topic modelling," *Procedia Computer Science*, 9th International Young Scientists Conference in Computational Science, YSC2020, 05-12 September 2020, vol. 178, pp. 205–212, 2020, doi: https://doi.org/10.1016/j.procs.2020.11.022.

[5] J. Dao, J. Wang, X. Zhang, "YNU-HPCC at SemEval- 2020 task 11: LSTM network for detection of propaganda techniques in news articles," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online), Dec. 2020, pp. 1509– 1515, International Committee for Computational Linguistics.

[6] V. Ermurachi, D. Gifu, "UAIC1860 at SemEval- 2020 task 11: Detection of propaganda techniques in news articles," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online), Dec. 2020, pp. 1835–1840, International Committee for Computational Linguistics.

[7] V.-A. Oliinyk, V. Vysotska, Y. Burov, K. Mykich, V. B. Fernandes, "Propaganda detection in text data based on nlp and machine learning," in *MoMLeT+DS*, 2020.

[8] E. Ferrara, "Contagion dynamics of extremist propaganda in social networks," *Information Sciences*, vol. 418-419, pp. 1–12, 2017, doi: https://doi.org/10.1016/j.ins.2017.07.030.

[9] S. Khan, S. Hakak, N. Deepa, B. Prabadevi, K. Dev, S. Trelova, "Detecting covid-19-related fake news using feature extraction," *Frontiers in Public Health*, vol. 9, 2022, doi: 10.3389/fpubh.2021.788074.

[10] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, J. Vilares, "Sentiment analysis for fake news detection," *Electronics*, vol. 10, no. 11, 2021, doi: 10.3390/electronics10111348.

[11] M. Huddar, S. Sannakki, V. Rajpurohit, "Attentionbased multi-modal sentiment analysis and emotion detection in conversation using RNN," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 112-121, 2021, doi: 10.9781/ijimai.2020.07.004.

[12] C.-H. Chen, P.-Y. Chen, J. Lin, "An ensemble classifier for stock trend prediction using sentence-level chinese news sentiment and technical indicators," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 3, pp. 53–64, 2022, doi: 10.9781/ijimai.2022.02.004.

[13] M. Charquero-Ballester, J. G. Walter, I. A. Nissen, A. Bechmann, "Different types of covid- 19 misinformation have different emotional valence on twitter," *Big Data & Society*, vol. 8, no. 2, p. 20539517211041279, 2021, doi: 10.1177/20539517211041279.

[14] Y. Wang, R. Han, T. Lehman, Q. Lv, S. Mishra, "Analyzing behavioral changes of twitter users after exposure to misinformation," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, New York, NY, USA, 2021, p. 591–598, Association for Computing Machinery.

[15] N. Kalantari, D. Liao, V. G. Motti, "Characterizing the online discourse

in twitter: Users' reaction to misinformation around covid-19 in twitter," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 4371–4380.

[16] S.-Y. Lin, Y.-C. Kung, F.-Y. Leu, "Predictive intelligence in harmful news identification by bert-based ensemble learning model with text sentiment analysis," *Information Processing & Management*, vol. 59, no. 2, p. 102872, 2022, doi: https://doi.org/10.1016/j.ipm.2022.102872.

[17] S. Kwon, A. Park, "Understanding user responses to the covid-19 pandemic on twitter from a terror management theory perspective: Cultural differences among the us, uk and india," *Computers in Human Behavior*, vol. 128, p. 107087, 2022, doi: https://doi.org/10.1016/j.chb.2021.107087.

[18] P. Koukaras, C. Tjortjis, D. Rousidis, "Mining association rules from covid-19 related twitter data to discover word patterns, topics and inferences," *Information Systems*, vol. 109, p. 102054, 2022, doi: https://doi.org/10.1016/j.is.2022.102054.

[19] K. Garcia, L. Berton, "Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa," *Applied Soft Computing*, vol. 101, p. 107057, 2021, doi: https://doi.org/10.1016/j.asoc.2020.107057.

[20] J. Han, J. Pei, Y. Yin, "Mining frequent patternswithout candidate generation," *SIGMOD Rec.*, vol. 29, p. 1–12, may 2000, doi: 10.1145/335191.335372.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019.

[22] R. Lamsal, "Coronavirus (covid-19) tweets dataset," 2020. [Online]. Available: https://dx.doi.org/10.21227/781w-ef42, doi: 10.21227/781w-ef42.

Danilo Cavaliere

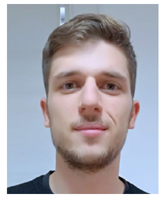
Danilo Cavaliere received both the master degree cum laude in computer science and the Ph.D. degree from the University of Salerno (Italy) in 2014 and 2020, respectively. He works now as postdoctoral researcher for the same institution. Dr. Cavaliere is in the editorial board of Neurocomputing journal and in the Program Committee for IEEE Symposium on Intelligent Agents, in the conference IEEE SSCI since 2019. His research interests include artificial and computational intelligence, data science, intelligent agents, data mining and knowledge discovery, scientific fields on which he has published many papers.

Giuseppe Fenza

Giuseppe Fenza (Member, IEEE) received the degree and Ph.D. degrees in computer sciences from the University of Salerno, Italy, in 2004 and 2009, respectively. He is currently an Associate Professor of computer science at the University of Salerno. The research activity concerns computational intelligence methods to support semantic-enabled solutions and decision-making. He has over 60 publications in fuzzy decision making, knowledge extraction and management, situation and context awareness, semantic information retrieval, service oriented architecture, and ontology learning. More recently, he worked in automating open source intelligence and big data analytics for counterfeiting extremism and supporting information disorder awareness.

Vincenzo Loia

Vincenzo Loia (Senior Member, IEEE) received the degree in computer science from the University of Salerno, Italy, in 1985, and the Ph.D. degree in computer science from the Université Pierre Marie Curie Paris VI, France, in 1989. He is currently a Computer Science Full Professor at the University of Salerno, where he worked as a Researcher, from 1989 to 2000, and as an Associate Professor, from 2000 to 2004. He is the Co-Editor-in-Chief of Soft Computing and the Editor-in-Chief of Journal of Ambient Intelligence and Humanized Computing. He serves as an editor for 14 other international journals.



Francesco David Nota

Francesco David Nota (Member, IEEE) received the master's degree in Business Innovation and Informatics from the University of Salerno, Italy, in 2020. He is currently pursuing the Ph.D. degree in innovation sciences for defence and security–digital transformation and cybersecurity with the Center for Higher Defence Studies (CASD). His research interests include CognitiveWarfare.