Twitter Vigilance: a Multi-User platform for Cross-Domain Twitter Data Analytics, NLP and Sentiment Analysis

Daniele Cenni, Paolo Nesi, Gianni Pantaleo, Imad Zaza
DISIT Lab, Distributed [Systems and internet | Data Intelligence and] Technologies Lab
Dep. of Information Engineering (DINFO), University of Florence
Florence, Italy, http://www.disit.org, http://www.disit.dinfo.unifi.it
paolo.nesi@unifi.it, daniele.cenni@unifi.it, gianni.pantaleo@unifi.it, imad.zaza@unifi.it

***Abstract*— The growth and diffusion of online social media have been enormously increased in recent years, as well as the research and commercial interests toward these rising sources of information as a direct public expression of the communities. Moreover, the depth and the quality of data that can be harvested by monitoring and analysis tools have evolved significantly. In particular, Twitter has revealed to be one of the most widespread microblogging services for instantly publishing and sharing opinions, feedbacks, ratings etc., contributing in the development of the emerging role of users as sensors. However, due to the huge amount of data to be collected and analyzed and limitations on data access imposed by Twitter public APIs, more efficient requirements are needed for analytics tools, both in terms of data ingestion and processing, as well as for the computation of analysis metrics, to be provided for deeper statistic insights and further investigations. In this paper, the Twitter Vigilance platform is presented, realized by the DISIT Lab at University of Florence. Twitter Vigilance has been designed as a cross-domain, multi-user tool for collecting and analyzing Twitter data, providing aggregated metrics based on the volume of tweets and retweets, users' influence network, Natural Language Processing and Sentiment Analysis of textual content. The proposed architecture has been validated against a dataset of about 270 million tweets showing a high efficiency in recovering Twitter data. For this reason it has been adopted by a number of researchers as a study platform for social media analysis, early warning, etc.**

Keywords*— *Social Media, Analytics, Metrics, Twitter, Web Crawling, Natural Language Processing, Sentiment Analysis.

I. INTRODUCTION

In recent years, the usage and popularity of social media (SM) has significantly increased. Meanwhile, a new role for human users is emerging, acting as smart sensors in the ubiquitous web and sharing observations and contents using mobile devices and the web, collecting, analyzing and disseminating information through more and more connected sensor networks. Actually, customers and users are increasingly used to: write opinions on different topics; publish reviews for sharing feedbacks about products, services or market trends; share information on social networks such as Twitter or Facebook. Therefore, a huge amount of unstructured and publicly available information is daily produced online, which is a matter of great interest both for research and commercial purposes. Social media analysis is the practice of gathering and analyzing data to gain new insights which could help to make informed decisions. Specifically, Social Media Analytics (SMA) is quite an emerging interdisciplinary research topic with the effort of combining, extending and adapting techniques and methods for social media data analysis, monitoring and visualization. Among the several social media solutions, Twitter is one of the most popular microblogging platform (counting about 300 million active users per month [1]), allowing users to have a personal news feed and followers attached to it. Twitter has emerged as one of the most widespread environment for social media analytics [2], with over 3 billion tweets and 15 billion API calls generated daily. Followers receive notifications connected to the actions performed by the users they follow. Typical actions of users can be: posting a message (tweet), commenting, expressing like/favorite, retweeting (the echo of some tweets by some users to the followers of the retweeting user). Therefore, tweets and retweets are exposed to other Twitter users, thus enhancing the chance of provoking their interests and reactions. Some of these mechanisms can generate viral processes that may lead to a huge diffusion of tweets in the user community.

A. Related Work

In the state of the art, numerous commercial and research tools for social media analysis have been presented. In particular, for analyzing Twitter data we have: Keyhole Social Media Analytics [3], Tweetreach [4], Brandwatch [5], Followewonk [6], TweetTracker [7], Twitris [8], OSOME [9] and SAS Sentiment Analysis [10]. They are examples of social media analytics platforms, which in some cases measure demographics, influential topics and sentiments.

Keyhole [3] social media analytics targets Twitter and Instagram and offers three services: text annotations tracking (keyword, URL or hashtags), account tracking and historical data retrieval. The text tracking service shows aggregate statistics such as the post volume containing text (in a three days' window), tweets sample, hashtags and word cloud, the number of users involved in the matching tweets, percentage of tweets, replies and retweet, sentiment score,. The account tracking service shows aggregate statistics as the number of total posts, followers, followees and the average counts of likes and retweets.

Union Metrics Tweetreach [4] targets Twitter and offers aggregate statistics such as volume of posts, tweet and retweet replies pie charts, top authors, retweet samples and timeline limited to 1.5M tweets and up to the 7 past days. Tweetreach offers also reach and exposure metrics. For instance, Reach metric is defined as the maximum number of unique Twitter accounts that received tweets containing keywords, hashtags or user citations specified in the search query during the 7-days limited time window; while Exposure metric is the total number of times tweets containing the search query key-terms were delivered to Twitter streams.

978-1-5386-0435-9/17/$31.00 ©2017 IEEE

Brandwatch Analytics [5] is a web-based social media monitoring platform designed to allow users to get the most out of the social media data relevant to their business. It is focused on providing a good customer experience. Its main use cases are: brand/reputation management, finding influences/advocates, market research, campaign, crisis management, community management, customer services, SEO and lead generation. Channels feature allows tracking any public Facebook page or Twitter account without the need for administrator rights.

Followewonk [6], as the name suggests, is a user network analysis centered Twitter analytics tool. Followewonk has five modules Search Twitter Bios, Compare Users, Analyze Followers, Track Followers and Sort Followers. The search feature allows to retrieve the Twitter users involved in the key terms contained in the search query. The compare users feature allows comparing Twitter users by followers. The analysis of followers allows to get information about followers such as mapped locations, most active hours, inferred gender, counts, etc. The Track Followers feature displays an interactive graph tracking changes in user gained and lost followers over certain time periods. The Sort Followers feature allows to sort all Twitter users' followers.

Twitris [11] is a social platform project developed at Kno.e.sis which is tailored to specific events such politics and elections, social movements and uprisings, crisis and disasters, entertainment, and environment. The main dashboard shows a map overlaid with markers indicating the spatial locations from which tweets were gathered for each event loaded in the platform. Moreover, each marker is enriched with related news, sentiment score, multimedia (images and videos) and Wikipedia articles by using explicit semantic information from DBPedia and SPARQL over metadata extracted from the tweets.

OSoMe [9] is a multi-modular architecture accessible through web interface and API. The system provides time series plots of the number of tweets (defined by one or multiple queries specified as one or multiple terms), spreads of specific hashtags through the social network via retweets and mentions, exploration of information diffusion through geographic space and time. The available API takes as input a time interval and a list of tokens (hashtags and/or usernames) and returns the following data in the same time interval: a list of tweet IDs mentioning at least one of the inputs, a count of the number of tweets mentioning each input token, a count of tweets matching any of the input tokens, a list of user IDs mentioning any of the tokens and a count of matching tweets produced by each user.

SAS Sentiment Analysis [10] is a general data source analytics platform which gathers any online data, for instance Facebook and Twitter, blogs, and review sites (e.g., TripAdvisor and Priceline). The tool performs sentiment analysis in real-time as the data is being retrieved, generates related reports with given locations where the topic is being discussed, and quantifies perceptions as positive, negative, and neutral. The provided graphic interface permits developing and managing sentiment analysis models (e.g., create, refine, upload).

In addition to the above-mentioned solutions, Sentiment Analysis of social media is typically used for assessing consumer feelings [12], predicting financial and market results [13], predicting election outcomes [14], providing early detection and warning for adverse medical issues [15], as well as for disaster response surveillance systems [16]. Research techniques and solutions in the field of Sentiment Analysis are divided into three main approaches [17]:

- Unsupervised learning: these methods rely usually in creating a sentiment lexicon in an unsupervised way, and using NLP linguistic-based rules to estimate the sentiment polarity;
- Supervised learning: input text is modeled as a feature vector representing its main characteristics to be estimated by mean of machine learning algorithms, typically using classifiers trained on collections of a priori annotated corpora;
- Concept-level Sentiment Analysis: these solutions are applied at concept level instead of word level, trying to extract multi-word expressions from text, conveyining specific semantics and sentics [18].

B. *Paper Aims and Overview*

This work presents the Twitter Vigilance platform, designed and developed by the DISIT Lab of the University of Florence, Italy. Twitter Vigilance is a multi-user tool for collecting Twitter data, producing/viewing analytics and several kinds of metrics in real time, creating personal dashboards, as well as studying events and trends on Twitter data, daily and in real time. Twitter Vigilance is much more effective for creating research studies on Twitter data with a high level of recall for the collected tweets. In addition, it presents a number of innovative features with respect to the state of the art related solutions, such as efficiency on recall, modeling channels, full faceted search, etc. The paper is structured as follows. In Section II, the most relevant requirements for Twitter data analytic platforms are discussed: Section III is devoted to provide an overview of the Twitter Vigilance architecture; subsequently, some real case studies are presented in Section IV. Finally, conclusions are drawn on Section V.

II. FUNCTIONAL REQUIREMENTS

Effective social media measurement and assessment require a combination of metrics derived by cross-domain techniques, and several different metrics must be combined together in order to extract relevant information [19]. Besides, the complexity of analysis may significantly vary, according to different business or research contexts.

In a Twitter data analytics platform, the visualization must permit data manipulation and filtering in an interactive way, to quickly identify anomalies and outliers. Moreover, it must be able to visualize streaming data at different time resolutions (per years, months, weeks, days, hourly, in real time). Processed data must permit also predictive analytics supporting decision processes, such as regression, predictions, clustering, machine learning. There must be the possibility to extract processed data from the platform, so that they can be processed later independently to identify how they work on analyzing, predicting or early warning of events based on social media data.

A social media analytics platform must deal with network analysis. The tool must be able to identify engagement and influence. Ruhi in [19] also proposes some specific use cases and requirements to be considered, such as the need to implement efficient methods to identify new or emerging relevant topics or themes to be analyzed, as well as for the identification of specific segments of social media audience as potential targets. Twitter analytics platforms must deal with tasks such as Twitter data retrieval (e.g., tweets, users' info,

etc.), store crawled data, implements efficiently text data analysis (e.g., Natural Language Processing, Sentiment analysis) and network analysis algorithms. Although social media analytics platform stores a minimal subset of tweets, every second, on average, around 6,000 tweets are tweeted on Twitter which corresponds around 200 billion tweets per year [20], therefore a social media analytic platform must deal with large data store.

The social media analytics platform must adopt convenient computation strategies according to the performed task. Building the network of retweets for a given hashtag will take more time and computational resources than just counting the number of posts containing the hashtags. Moreover, it should be considered as a relevant aspect the fact that tools available for researchers are far from being ideal. They are usually provided with uncomplete access to raw data without analytic tools, which have to be designed and implemented from scratch [21].

Twitter provides many different modalities to access Twitter data, among which the most efficient for recovering tweets are the Search API and the Streaming API. Since version 1.1 of Twitter API, it is necessary to log into Twitter by using OAuth protocol for all requests. Both Twitter APIs types return data in JSON format. Search API presents a limited number of requests every 15 minutes. The Streaming APIs give developers a low latency access to Twitter's global stream, but limited access to the whole tweets. Twitter offers different streaming endpoints customized for use type: public, user and site. Both the Search and the Streaming APIs present some limitations in terms of maximum number of tweets per hour, and any of them do not guarantee that all tweets which are on Twitter.com could be obtained for the analysis. For example, The Search API is not offering a complete index of all tweets, but instead an index of recent Tweets (published in the past 7 days). Total rate limits of Streaming API are quantitatively not well documented (the documentation say up to 1% of the full firehose of tweets). Therefore, none of the two types of APIs allows a complete recovery of the entire Twitter corpus for a given topic and/or search key. Furthermore, the maximum number of request calls the Twitter APIs can handle in a given period is limited (Twitter limits each API call to collect at most 100 tweets per request). This fixed threshold proves to be not suitable, since the 100 tweets limit per request seems to be well balanced when monitoring long-term events, hence compromising the capability of recovering all tweets and retweets for very fast events, producing high number of tweets/retweets per second. Other important factors, such as the huge number of tweets that can be produced and collected for certain cases, the complexity of social relationships among users, the limited size of tweets (140 characters), and the fact that historical Twitter data are not accessible via the Twitter APIs force the developers to set up specific architectures for collecting tweets, while attempting to retrieve them with sufficient reliability [22]. Furthermore, companies such as Twitter are both restricting free access to their data and licensing their data to commercial data resellers, such as Gnip and DataSift [21]. On the other hand, when monitoring explosive/fast events (typically with a high volume rate, in the order of hundreds thousands of tweets per day/hour), the same limit is too restrictive, thus affecting the precision of tweets gathering. It is worth noting that such an aspect is often underestimated by social media analytics tools at the state of the art, or at least no guarantee or realistic measure on their effectiveness is provided in this sense. In addition, as mentioned before, none of the Twitter APIs may guarantee the full access to the tweets.

Challenges and advances to be expected and outlined in the field of Sentiment Analysis are the adoption of improved models (e.g., moving from word-level techniques to concept-level techniques). Moreover, in order to develop a more expressive model for sentiment, multi-dimensional frameworks have been adopted recently, which aim at going beyond the widely applied positive/negative polarity model, such as the Hourglass of Emotions [23], inspired by the earlier Plutchik's model [24].

Geolocation based on users' position is not yet widespread enough. Some solutions have been presented in order to identify geolocated events and resources in smart cities [25], [26], as well as extracting geolocation information from unstructured text by using NLP techniques [27].

According to the above analysis of the state of the art, a set of requirements have been identified and thus a feature comparison among the reviewed solutions and the proposed Twitter Vigilance framework is showed in **Table I**. The features used for the comparison are defined in the following:

- Twitter Metrics: insights about tweets volume metrics (tweets, retweets over time etc.);
- Sentiment Analysis: techniques and solutions adopted to extract a quantitative measure of users' sentiment from published textual contents;
- Natural Language Processing (NLP): to extract linguistic information from textual contents, useful for further investigations (e.g., Sentiment Analysis);
- API availability: accessibility via API;
- User network analysis and, generally, users' based metrics;
- Data analysis based on geolocation (e.g., geographic clustering);
- Real -time analysis: the capability of the system to perform Twitter data analysis on a real-time basis;
- Full faceted search: in order to have further views of data analyzed on tweets collection it is mandatory the exploitation of full faceted search [28], which enable users to browse tweets by choosing from a pre-determined set of categories (tweet author, message, url, geolocation etc.); Metrics for quantitatively assessing the efficiency of message recovery, which is an important aspect to assess the system capability of recovering all the available tweets, due to the several limitations imposed by the Twitter APIs;Minimization of searches on Twitter: query optimization is a critical part in the functional flow of a Twitter analytics platform. Query parser must take account of Twitter API limit and issue an evaluation plan to minimize twitter API request.

TABLE I. COMPARISON OF STATE OF THE ART SOLUTIONS (NA: NOT AVAILABLE)

Service	Twitter Metrics (e.g. # of tweets, retweets over time)	Sentiment analysis	NLP Analysis	API availability	User network analysis	Data analysis based on geolocation	Real time Analytics	Full faceted Search	Metrics for assessing recall	Minimization of searches to Twitter
SAS	N	Y	N	Y	N	N	Y	N	N	na
Keyhole	Aggre -gate	N	N	N	Aggre -gate	Y	N	N	N	na

Tweetreach	Aggre -gate	N	N	N	Aggre -gate	Y	N	N	N	na
Brandwatch	N	N	N	N	Y	Y	Y	N	N	na
Followewonk	N	N	N	N	Y	Y	Y	N	N	na
Twitris	N	Y	N	N	N	Y	Y	N	N	na
OSoMe	Y	N	N	Y	Y	Y	Y	N	N	na
Twitter Vigilance	**Y**	**Y**	**Y**	**Y**	**Y**	**N**	**Y**	**Y**	**Y**	**Y**

III. OVERVIEW OF TWITTER VIGILANCE ARCHITECTURE

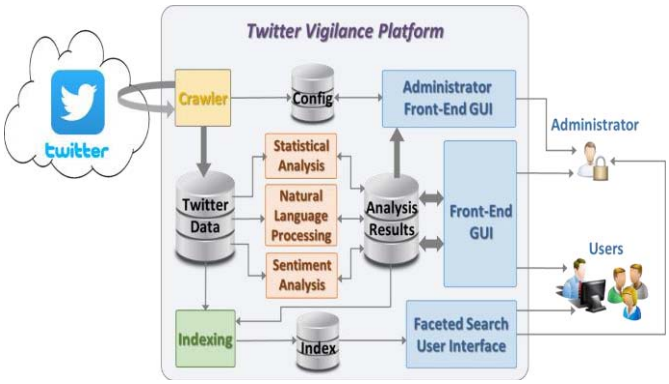
The Twitter Vigilance platform has been designed as a multipurpose comprehensive tool providing different tasks and metrics suitable for Twitter Search API, with the aim of Twitter data collection, analysis and monitoring for research purpose. Search API has been chosen, since it offers a greater flexibility and chance of filtering query results, given the number of parameters it provides. The architecture is depicted in Figure 1.

Fig. 1. Overview of Twitter Vigilance architecture.

Twitter vigilance is a modular architecture composed of:

- TV: Twitter Vigilance main tool (http://disit.org/tv/), collecting and analyzing tweets daily;
- RTTV: Real-time twitter Vigilance (http://disit.org/rttv/), collecting and analyzing tweets in real time;
- TVSolr: Twitter Vigilance Advanced search (http://tvsolr.disit.org/), indexing tweets and faceted search

In the two former Twitter Vigilance tools, a multithread crawler performs data gathering and extraction by using Twitter Search APIs. The data acquisition approach is based on the concept of Twitter Vigilance Channel, consisting of a set of simple and complex search queries, which can be defined by a registered user by combining keywords, hashtags, user's IDs, citations, etc., in a structured logical syntax, according to the search syntax of Twitter. TVSolr is a Twitter Vigilance Advanced search tools which permits to browse efficiently the pool of gathered tweets indexed using a four-shards Apache Solr cluster which allows full faceted search and filtering; the final visualization interface has been realized by using a HUE graphical interface.



In the two former Twitter Vigilance tools, a multithread crawler performs data gathering and extraction by using Twitter Search APIs. The data acquisition approach is based on the concept of Twitter Vigilance Channel, consisting of a set of simple and complex search queries, which can be defined by a registered user by combining keywords, hashtags, user's IDs, citations, etc., in a structured logical syntax, according to the search syntax of Twitter. TVSolr is a Twitter Vigilance Advanced search tools which permits to browse efficiently the pool of gathered tweets indexed using a four-shards Apache Solr cluster which allows full faceted search and filtering; the final visualization interface has been realized by using a HUE graphical interface.

Therefore, Twitter Vigilance platform provides a large number of features with respect to other state of the art solutions as depicted in Table I. In addition, it:

- allows registered users (researchers) to create and edit customized channels as a collection of searches on API;
- crawls tweets, computes metrics, and shows results of Twitter Data, as: volume metrics about tweets, retweets and user statistics, NLP and Sentiment Analyses based metrics;
- provides public access to metric results computed on channels and search analysis;
- provides user based metrics such as the number of users per channel/searches, number of posts per users per channel, percentage distribution per search, detailed user information such profile creation, number of favorites tweets, number of followers, number of following, number of list, location, number of tweets retweets;
- Allows the researchers to download resulting metrics values (through API service) over time for further analysis.

The search queries associated with each Twitter Vigilance Channel are posed to the Twitter platform via a crawler. Both configuration parameters and statistical results are accessible from the front-end interface for the user. Collected tweets are processed by the back-office processes, which implement statistical analysis, natural language processing (NLP) and sentiment analysis (based on distributed NLP on Hadoop [29]), as well as general data indexing. The metrics resulted by the back-office processes are periodically computed and stored on a dedicated database. This approach allows to make them accessible to the front-end graphic user interface (see Figure 2 as an example) with the needed performance of the front-end server. The solution allows customizing the search query, dashboards, reports and file export computed metrics on tweeter data (e.g., to Excel or CSV format) for visual analytics, temporal trends and time series visualizations, data results navigation, Twitter user's statistics and analysis.

Specifically designed strategies have been implemented in order to make the whole back office processing more efficient and to maximize consistency of retrieved data, such as: optimized search strategies for collecting tweets (e.g., avoiding repetitions of more inclusive search queries, on the basis of user defined logical rules for searches associated at each channel); incremental processing and caching in order to avoid analysis of search queries which are shared among different channels (typically also on different temporal windows); adaptive real-time processing to optimize the usage of API calls, due to the restrictions and limitations imposed by Twitter previously discussed in Section II.

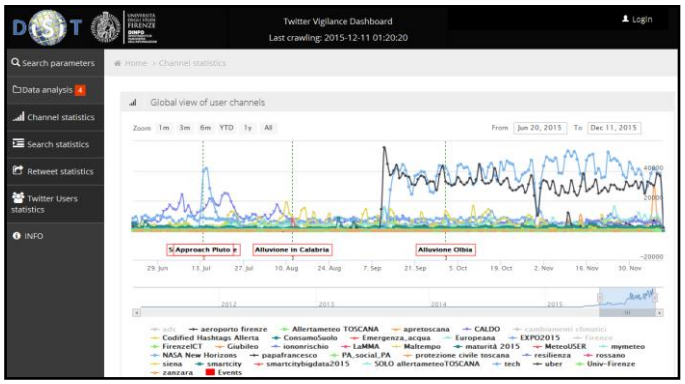


Fig. 2. TwitterVigilance front-end graphic user interface: http://www.disit.org/tv

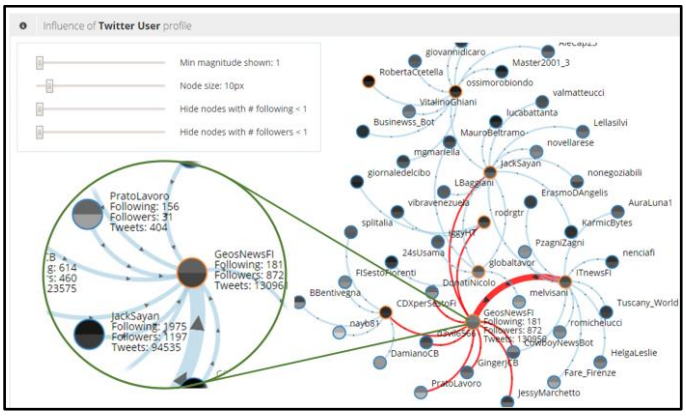


Fig. 3. User Influence Network.

The mentioned analyses are performed at both Twitter Vigilance Channel level and at single search level. In the specific, the following information, metrics and features can be retrieved.

Original Tweet Retrieval: Twitter Vigilance takes care of the original tweet of retweets. With the term "original tweet" we define a tweet which has at least one retweet. For each retweet, the original tweet is identified, checked whether these tweets have already been collected or not by the system. This may happen using the Search API. Twitter search API accepts a maximum of 100 IDs per request. Due to this limit, even though this process is scheduled once a day, it can sometimes occur that not all original tweets are recovered. Thus subsequently, the missing tweets have to be requested to Twitter to have a global view. The request is made through a specifically designed periodic process, which on the basis of the original tweets ID retrieves the missing original tweets. The coverage of original tweets is important in order to identify sources of potential viral events and topics and to make content-based data processing more efficient.

User influence network and other users' metrics: Twitter Vigilance performs a user centered statistic and users relationships analysis. The user statistics include the number of users per channel, the number of user per searches, the number of posts per users per channel, the percentage distribution per search, detailed user information such profile creation, the number of favorites tweets, the number of followers, the number of following, the number of lists, location, the number of tweets and retweets. The relations between users are described as a diffusion network (see Figure 3), where nodes represent users and an edge drawn between any two connected nodes indicates an exchange of information between those two users, i.e. the retweets. Edges have a weight to represent the number of messages connecting two nodes. Convenient information such as the number of tweets, followers and followees are present as label near the nodes. Also, a visual graphic semantics is adopted: each node is divided in two parts ranging from light gray to black, which indicate the number of followers and followees, respectively. Darker color indicates high values. To preserve graph clarification, graphical editing features are provided such as min magnitude, node size, hide following or follower.

Volume based metrics: number of tweets (TW) and retweets (RTW) at channel and search level, Aggregated measures at different time resolution (daily, on hourly basis, every 5 minutes for real time processing etc.) are computed and stored on a dedicated database to be quickly accessed and made available for graphical scripts. In addition, higher level metrics are computed, such as the ratio RTW/TW, which provides a measure of users' reactivity to specific topics or events. The ratio RTW/TW may lead to large values, for instance if the event under monitoring becomes strongly viral.

NLP-based metrics: extraction of Part-of-Speech (POS) tagged keywords and computation of keyword occurrence at different time resolution (daily, on hourly basis, every 5 minutes for real-time processing on RTTV), grouped by nouns, adjectives, verbs, hashtags and user citations/mentions. This kind of information allows the users to analyze and identify emerging trends or firing/**explosive** events regarding specific keywords, hashtags (to understand which are the most used, emerging, evolving) or mentions (to detect potential influencers, pushers, emerging citations etc.).

Sentiment Analysis based metrics: sentiment polarity extraction for each single tweet; this kind of information can be useful to assess and estimate the general sentiment of the Twitter community regarding a specific channel or search (i.e. regarding specific topics of interest). In addition, the sentiment polarity is extracted at a more fine-grained level, that is for each single noun, adjective and verb. These lower level metrics are used to weight keyword occurrences (computed as NLP-based metrics, as previously described) in order to evaluate the most influential keywords for sentiment analysis, as well as identifying possible sources and reasons to explain or interpret specific sentiment trends. A more detailed description of the Sentiment Analysis is provided later in the section. Aggregated measures of all these are computed at different time resolution (daily, on hourly basis, every 5 minutes for real time processing etc.), and stored in a dedicated database to be quickly accessible and available for graphical scripts.

The derived metrics and information can be useful to understand which are the most widely used or emerging hashtags, as well to detect which are the most influential in determining the polarity (the positive/negative signature) of the sentiment, in order to better tune the collected tweets, and for pre-computing basic metrics that can be useful for the researcher to make further analysis for different domains, communication and media, predictive models etc. It can be a useful tool for

identifying reasons for positive/negative tweets, as well as the reaction of the community. NLP and Sentiment Analysis related tasks are performed in a multi-language fashion, currently supporting English and Italian. The computation of Sentiment Analysis metrics has been performed by exploiting SentiWordNet [30], a semantic knowledge base specifically designed for Sentiment Analysis. SentiWordNet assigns sentiment scores to each extracted keyword in order to estimate the general sentiment polarity of the collected tweets. In SentiWordNet independent positive, negative, and neutral sentiment values (i.e., real numbers varying in the interval from -1 to 1) are associated with about 117 thousand synsets. In order to perform the analysis in both English and Italian languages, the SentiWordNet lexicon (originally designed in English) has been automatically translated to an Italian version, on the basis of MultiWordNet [31], a resource which aligns WordNet English synsets to the Italian ones. For each single tweet/retweet, its overall polarity score is given by the sum of all the sentiment weighted keywords extracted in it. A detail of the Sentiment Analysis and NLP manager interface is shown in Figure 4.

(a)

(b)

Fig. 4. Sentiment Analysis and NLP manager interface: (a) trend of the most relevant sentiment analysis metrics and (b) detail of Top-Sentiment rated Italian adjectives for a single channel.

IV. Case Studies and Validation

In this section, outcomes and performance results of the usage of the Twitter Vigilance platform are reported. As a qualitative evaluation, Twitter Vigilance provides several different functionalities that outperform the state-of-the-art solutions in terms of features. Exploiting all these features jointly (such as the computation of posts' volume and users' metrics, NLP and Sentiment Analysis, API availability, real-time analysis, full-faceted search, and optimized strategies to minimize Twitter API calls) allowed the platform to be used to collect and analyze Twitter data in many different real cases. Some examples are reported in the following.

A. Case Studies

DISIT Twitter Vigilance platform has been adopted in several studies as a tool to monitor city services, critical events and conditions, user behavior, assessing appreciation of services, city response to events, providing predictive capabilities [32] to support decisional processes. It is also adopted in smart city projects as Sii-Mobility SCN http://www.sii-mobility.org, REPLICATE and RESOLUTE EC H2020 project http://www.resolute-eu.org, for monitoring as an early warning solution to detect critical conditions (inception of critical events on the city, new reaction on drugs, etc.).

Twitter Vigilance currently contains more than 100 channels created by about 30 users which are mainly local Public Administrations and research groups to collect information and monitor different topics and trends related to different domains of interest, such as: traffic and public transportation (e.g.: "TPL", "TPL2", "Uber"), weather monitoring (e.g.: "Maltempo", "Meteo Firenze", "Codified Weather Hashtags", "ARPAT", "LaMMA", "PAMeteoNews"), resilience and early warning for natural disasters and other accidents (e.g.: "PAProtezione Civile", "Protezione Civile Toscana", "Allerta Meteo", "Allerta Meteo Toscana", "Mugnone 2016"), tourism and events in Florence and Tuscany (e.g.: "Turismo Firenze", "Firenze"), monitoring large attendance public events and TV shows (e.g.: "EXPO2015", "EXPO2015Toscana", "Xfactor9", "Xfactor10", "Pechino Express"), terrorism, adverse drug reaction, etc. In addition, in May 2016 Twitter Vigilance TV and RTTV platforms were used in a joint-venture flood dealing simulation composed of civil protection, fire fighters and other local institutions (http://protezionecivile.comune.fi.it/?p=7822) [33]. For this purpose, the channel "Mugnone 2016" (from the name of a creek that flows in the city) has been created for real-time monitoring. Presently a real-time dashboard for monitoring Twitter data is provided (http://www.km4city.org/dashboard/tv.html).

In [34], the "Codified Weather Hashtags" channel has proven to be an effective tool to convey useful information on Twitter, with formal and informal sources regarding weather related events and alerts. The authors collected tweets published within a 30 days' period, identified by three codified hashtags (in the three different Italian regions, i.e., Liguria, Tuscany and Piedmont). The aim was to assess whether codified hashtags could represent an effective way to align formal and informal sources of information during weather related emergencies. In this perspective, the use of codified hashtags may potentially improve the performance of systems, for automatic information retrieval and processing during disasters. In [35], the twitter data have been demonstrated to be strongly correlated with respect to the temperature of the hot period in the Tuscany region.

B. Validation on Collecting Tweets Performance

In this section, the validation of the system in terms of efficiency in getting Twitter data is presented, to put in evidence, its capabilities to recover the correct number of tweets and related metadata, despite the Twitter Search API usage limitations, as well as to adapt itself to changes in the set of recoverable tweets. Therefore, to perform a rigorous assessment, two validation metrics have been defined:

- The *coverage percentage of original tweets, $CoTW_O$,* provides a qualitative parameter on the system's ability to recover all the available tweets. Being TW_O the number of original tweets, and MTW_O the number of missing original tweets, the coverage of original tweets $CoTW_O$ is given by the ratio of successfully retrieved original tweets to the total of identified original tweets, that is:

TABLE II -- SYSTEM EFFICIENCY FOR ORIGINAL TWEETS COVERAGE AND TWITTER SEARCH API CALL SATURATION ON OVERALL MESSAGE RETRIEVAL.

Posts Volume (Tweets + Retweets) Range	*# Recovered Original Tweets*	*# Missing Original Tweets*	*% Original Tweets Coverage ($CoTW_O$)*	*# Twitter Search API requests*	*# Saturations on Twitter Search API requests*	*% Saturations on Twitter Search API requests ($S_{\%}$)*	*% Not-Saturated Twitter Search API requests (1-$S_{\%}$)*
< 10k	18571	2033	89,05%	124299	1	0,00%	100,00%
[10k, 50k)	130051	13716	89,45%	399170	100	0,03%	99,97%
[50k, 100k)	96171	10278	89,31%	123804	165	0,13%	99,87%
[100k, 500k)	997833	86755	91,31%	849062	1589	0,19%	99,81%
[500k, 1M)	930646	61632	93,38%	439956	1998	0,45%	99,55%
[1M, 5M)	6454463	439628	93,19%	2787485	31585	1,13%	98,87%
> 5M	14714124	899035	93,89%	4509184	64284	1,43%	98,57%

$$CoTW_0 = \frac{TW_O - MTW_O}{TW_O}$$

- The *percentage of saturations,* $S_{\%}$, indicates the system aptitude to adapt to changes in the number of recoverable tweets, that is a measure of the overall system efficiency in recovering messages. Saturation is defined as the inability of the system to accommodate the growth in the number of messages that are available to a given search: since the twitter Search API limits to 100 tweet IDs per request, saturation occurs when a request through a Twitter Search API call produces 100 new posts of 100 returned. The saturation percentage $S_{\%}$ is computed as the ratio of the number of saturations S to the total number of Twitter requests Req_{TW} made by the system crawler:

$$S_{\%} = \frac{S}{Req_{TW}}$$

The assessment has been performed on the whole dataset collected in the last 24 months, since the activation of the platform. The assessed data set consists of about 270 million of Twitter messages (of which about 47% tweets and about 53% retweets) grouped into 102 channels (of which 78 are still active) and a total of 1407 searches (of which 1194 are active). Inactive searches are those which are not associated with any active channel. The dataset has been clustered into 7 categories (see Table II) according to the total posts volume, given by the sum of collected tweets and retweets for each channel, in order to identify and analyze possible different system adaptive behaviors and results for increasing volumes of posts. The first subset contains all collected posts belonging to channels with tweets plus retweets volume of less than 10 thousand. The second subset is formed by collected posts from channels with number of tweets plus retweets between 10 and 50 thousand, and so on. Results are shown in Table II, while the performance trend is summarized in Figure 5. The system shows very low saturation levels (i.e., very high percentage of not-saturated Twitter Search API calls, as depicted in Figure 5) for each cluster. Results show a very good overall performance of the proposed system in recovering Twitter messages, with an efficiency percentage never dropping below 98%. Saturation percentage slightly increases with increasing number of retweet, as expected, since higher numbers of retweets to be collected may be associated with rapidly variable and/or viral events, which usually lead to a delay in the system adaptation and a consequent higher number of saturations of Twitter requests.

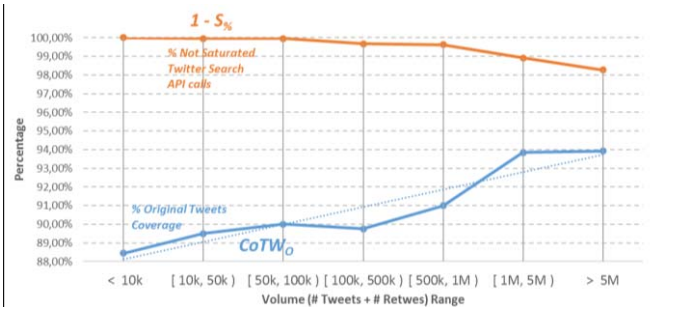


Fig. 5. Original Tweets coverage and Twitter Search API call saturation percentages on overall message retrieval. The analyzed dataset has been clustered into seven different categories according to the total volume of posts (sum of tweets and retweets).

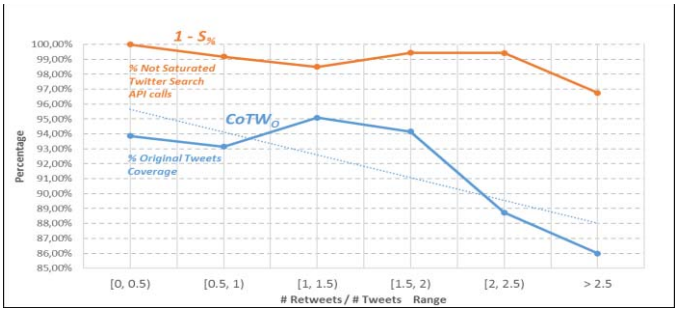


Fig. 6. Original Tweets coverage and Twitter Search API call saturation percentages on overall message retrieval. The analyzed dataset has been clustered into six different categories according to the ratio of retweets to tweets.

Furthermore, a generally increasing trend for coverage of original tweets can be observed (for increasing retweet ranges, as outlined by the dotted trend line in Figure 5), showing a very good original tweet recover efficiency, from about 89% to 94%. A descending trend should be expected in this case as well. However, the increasing trend can be explained by the dedicated process for recovering original tweets. A different behavior is observed when computing the same validation metrics upon a different clustering of the same test dataset, that is based on the ratio of retweets to tweets. The ratio of retweets to tweets better describes the virality of analyzed events and topics, rather than considering the absolute tweets and retweets volume. This means that the original tweet recover efficiency decreases with increasing virality.

V. CONCLUSIONS

In this paper, the Twitter Vigilance framework has been presented, as a multipurpose comprehensive tool for Twitter data collection and analysis. The platform provides several different solutions that outperform the state-of-the art solutions in terms of features and recall capability. It is suitable for research purposes (based on volume of tweets and retweets, users' influence network and Natural Language Processing and Sentiment Analysis of posts) in a cross-domain, multi-user environment, which is capable of handling millions of Twitter related data. In addition, the architecture design has been improved with computational strategies to make big data handling and computation more efficient, and some quite novel features and insights have been proposed, such as NLP and Sentiment Analysis at POS-level (considering nouns, adjectives, and verbs), user influence network. One of the most interesting feature presented is the capability of recovering original tweets, to identify sources of potentially viral trends, topics or events, as well as to make data processing more efficient. The solution for reliable Twitter data collection and analysis has been validated against a dataset of about 270 million Twitter posts (collected by the system itself), showing very good capabilities in retrieving original tweets, as well as a very high efficiency in overall message recovering.

ACKNOWLEDGMENT

The authors would like to thank for funding in the context of RESOLUTE H2020 project, which received funding from the European Commission under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 653460).

REFERENCES

[1] Stieglitz S., Dang-Xuan, L., Bruns, A. and Neuberger, C., Social Media Analytics. in Business & Information Systems Engineering, Vol. 6(2), pp. 89-96 (2014).

[2] Abbasi A, Hassan A, Dhar M., Benchmarking Twitter sentiment analysis tools. In 9th international conference on language resources and evaluation (LREC), 2014.

[3] Keyhole. Web site: http://keyhole.co/

[4] Tweetreach. Web site: https://tweetreach.com/

[5] Brandwatch. Web site: https://www.brandwatch.com/

[6] Followewonk. Web site: https://moz.com/followerwonk/

[7] TweetTracker. Web site: http://tweettracker.fulton.asu.edu/

[8] Twtitris. Web site: http://twitris.knoesis.org/

[9] Davis CA, Ciampaglia GL, Aiello LM, Chung K, Conover MD, Ferrara E, Flammini A, Fox GC, Gao X, Gonçalves B, Grabowicz PA, Hong K, Hui P, McCaulay S, McKelvey K, Meiss MR, Patil S, Peli Kankanamalage C, Pentchev V, Qiu J, Ratkiewicz J, Rudnick A, Serrette B, Shiralkar P, Varol O, Weng L, Wu T, Younge AJ, Menczer F. (2016) OSoMe: The IUNI observatory on social media. PeerJ Preprints 4:e2008v1 doi: 10.7287/peerj.preprints.2008v1.

[10] Sas Sentiment Analysis. Web site: https://www.sas.com/

[11] Sheth, Amit, et al. "Twitris: A system for collective social intelligence." *Encyclopedia of Social Network Analysis and Mining*. Springer New York, 2014. 2240-2253.

[12] Smith, A. N., Fischer, E., and Yongjian, C., How does brand-related user-generated content differ across YouTube, Facebook, and Twitter? In Journal of Interactive Marketing, Vol. 26(2), pp. 102-113, 2012.

[13] Bollen, J., Mao, H., and Zeng, X., Twitter mood predicts the stock market, In Journal of Computational Science, Vol. 2(1), pp. 1-8, 2011.

[14] Skoric, M., Poor, N., Achananuparp, P., Lim, E. P., and Jiang, J., Tweets and votes: A study of the 2011 Singapore general election, In Proceedings of the 45th Hawaii International Conference on System Science (HICSS), pp. 2583—2591, 2012.

[15] Abbasi, A., Fu, T., Zeng, D., and Adjeroh, D., Crawling Credible Online Medical Sentiments for Social Intelligence, In Proceedings of the ASE/IEEE International Conference on Social Computing, 2013.

[16] Goodchild, M. F. and Glennon, J. A., Crowdsourcing geographic information for disaster response: a research frontier, In International Journal of Digital Earth, Vol. 3(3), pp. 231-241, 2010.

[17] Chikersal P., Poria S., Cambria E., Gelbukh A., Siong C.E., Modelling Public Sentiment in Twitter: Using Linguistic Patterns to Enhance Supervised Learning. In Computational Linguistics and Intelligent Text Processing. CICLing 2015. Lecture Notes in Computer Science, vol 9042. Springer, Cham, 2015.

[18] Cambria E. and Hussain A.: Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis. Springer, Cham, 2015.

[19] Ruhi U., Social Media Analytics as a Business Intelligence Practice: Current Landscape & Future Prospects, In Journal of Internet Social Networking & Virtual Communities, 2014.

[20] Retrieved from http://www.internetlivestats.com/twitter-statistics/

[21] Batrinca B. and Treleaven P.C. Social media analytics: a survey of techniques, tools and platforms, In Journal of AI & Society, Vol. 30: 89, 2015.

[22] Oussalah M, Bhat F, Challis K, Schnier T. A software architecture for Twitter collection, search and geolocation services, In Knowledge-Based Systems. Vol. 37, pp.105-120, 2013.

[23] Cambria, E. and Hussain, A., Sentic Computing: Techniques, Tools, and Applications, Springer, Dordrecht, Netherlands, 2012.

[24] R. Plutchik, The nature of emotions, American Scientist, Vol. 89(4), pp. 344–350, 2001.

[25] Anantharam, Pramod, et al. "Extracting City Traffic Events from Social Streams." ACM Transactions on Intelligent Systems and Technology 9.4 (2014).

[26] Doran, Derek, Swapna Gokhale, and Aldo Dagnino. "Human sensing for smart cities." Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ACM, 2013.

[27] P. Nesi, G. Pantaleo, M. Tenti, "Geographical Localization of Web-Visible Human Activities by employing Natural Language Processing, Pattern Matching and Clustering Based Solutions", Journal: Engineering Applications of Artificial Intelligence, Elsevier. 10.1016/j.engappai.2016.01.011.

[28] Tunkelang, D., "Faceted search." Synthesis lectures on information concepts, retrieval, and services 1.1, pp. 1-80, 2009.

[29] P. Nesi, G. Pantaleo and G. Sanesi, A Hadoop Based Platform for Natural Language Processing of Web Pages and Documents, accepted for publication on JVLC, Journal of Visual Languages and Computing, Elsevier. 11-112015, http://dx.doi.org/10.1016/j.jvlc.2015.10.017.

[30] Esuli, A. and Sebastiani, F. Sentiwordnet: A publicly available lexical resource for opinion mining. In Proc. of the 5th Conference on Language Resources and Evaluation (LREC'06), pp. 417–422, Genova, IT, 2006.

[31] Pianta, E., Bentivogli, L. and Girardi, C. MultiWordNet: developing an aligned multilingual database. In Proc. of the First International Conference on Global WordNet, Mysore, India, January 21-25, 2002.

[32] A. Crisci, V. Grasso, P. Nesi, G. Pantaleo, I. Paoli, I. Zaza, "Predicting TV programme Audience by Using Twitter Based Metrics", in Multimedia Tools and Applications, Springer, 2017.

[33] Grasso V, Zaza I, Zabini F, Pantaleo G, Nesi P, Crisci A. (2016) Weather events identification in social media streams: tools to detect their evidence in Twitter. PeerJ Preprints 4:e2241v1 https://doi.org/10.7287/peerj.preprints.2241v1.

[34] Grasso, Valentina, and Alfonso Crisci. "Codified hashtags for weather warning on Twitter: an Italian case study." PLoS currents 8 (2016).

[35] V. Grasso, A. Crisci, P. Nesi, G. Pantaleo, I. Zaza and B. Gozzini, "Public crowd-sensing of heat-waves by social media data", 16th EMS Annual Meeting & 11th European Conference on Applied Climatology (ECAC), 12–16 September 2016 | Trieste, Italy, CE2/AM3, Delivery and communication of impact based forecasts and risk based warnings.