Social Media Data Schemas for Crowdsourced Topics Observational Analytics

Ilias Dimitriadis, Vasileios G. Psomiadis, and Athena Vakali

***Abstract* - Crowdsourcing offers an invaluable toolkit for obtaining dynamic trends and insights from social media data analytics, enabling the capture of the wisdom of the crowds. The plethora of available platforms requires the appropriate definition of data schemas and techniques to allow for efficient knowledge extraction from unstructured social media user generated content and users' multilevel interactions. The present work addresses such challenges by designing an effective and flexible document-based data model that supports heterogeneous social media data integrations. This model is then exploited under a crowdsourced topics observatory that involves interactive visualization modules and advanced topic modelling methods. The proposed framework is implemented and demonstrated on a social innovation platform aiming to promote awareness on plastic waste revaluation and empower stakeholders of the plastics value chain.**

***Index Terms* - Information systems, Crowdsourcing, RESTful web services, Computing methodologies, Topic modeling, Social media analytics, Thematic detection, Heterogeneous data sources, Dynamics and trends discovery, Wisdom of the crowds, Visualization.**

II. INTRODUCTION

Social media data are constantly produced and shared, offering the ground for knowledge extraction and crowd's trends and opinions detection. Social media analytics has become a valuable approach to harvest such knowledge from openly circulated data over multiple popular platforms (such as Twitter, Facebook, Instagram, Flickr). The knowledge derived from the evolving crowd-driven ideas, identified as "the wisdom of the crowds", is heavily dependent on multiple data sources which should be integrated under appropriate and adaptive data schemas. Data produced in abundance in online social media sources offer a fertile ground for harvesting most popular topics expressed openly in the form of opinions and views of users about key issues.

This work is motivated by the need to define appropriate approaches which will support effective social media data schemas designs, resulting in social media data topics analytics. Dealing with multiple data types in social media is a challenging task since each of the online platforms produces different data types and formats and has a different scope and origin. For example, data produced in Twitter include a source (tweet text) and a metadata (retweet, time, user-id, geo-location, etc.) part, data in Facebook include such parts but also a so called social graph (with users friendship information), data in Flickr or Instagram focus on mostly multimedia type of User Generated Content (UGC), etc.

Capturing the actual wisdom of the crowds, ideally demands social media data integration from multiple platforms, thus harvesting knowledge over multiple crowdsourced data threads and types requires novel approaches and solutions. The exploitation of social media resources leads to the delivery of innovative and more human-centred services by leveraging the collective intelligence of the crowd. In this context, social media data mining as a collective intelligence approach, has evolved tremendously during the last decade. It involves extracting latent information and insights from: (i) unstructured social media user generated content (UGC); and (ii) users' interaction in social media, such as communities or groups with similar interests and behaviour.

The availability of these large volume and heterogenous data streams pose significant challenges to typical data mining algorithms. Existing relevant approaches mainly focus on general categorization of social media content [Zubiaga15]. NLP approaches have been utilized for assessing linguistic features [Duan12] and performing sentiment analysis [Kanavos14], [Santos14]. Although a lot of work has focused on the prediction of emerging topics and the identification of current trends [Xie16], real-time topic and trend detection still remains an issue. However, in several tasks a more fine-grained context and location-specific categorization were deemed necessary [Dong15], [Unankard15]. Identifying experts on a certain field using crowdsourcing can further improve the efficiency of these tasks and has also been addressed extensively [Ghosh12], [Brem15], [Bozzon13]. Users classified as experts are expected to provide cleaner data input and thus allow the crowdsourcing analytics process to produce more solid results [Rjab16]. While using crowdsourcing for data-mining is quite popular, the quality control of its outcome still has some drawbacks [Xintong14].

A fine-tuned combination of the sources above under a new framework, that can provide an improved semantics-aware crowdsourced observatory for specific thematic related terms and qualitative identification of current trends and topics, is utilized in the present work. This approach places emphasis on identifying the proper data schemas which will sustain the components required for offering a complete crowdsourced topics observatory. The data schemas proposed support an effective topic observatory design and implementation with a set of inter-linked components. These components complement

a framework which can be easily accessed and which is human-friendly in terms of interpreting the derived knowledge.

The proposed crowdsourced topics observatory highlights patterns relevant with a given thematic and it is adaptive and customizable depending on a given set of terms and hashtags. Emphasis is also placed on appropriate visualization of social media driven 'wisdom of the crowds' with highlighting topics perception as it is expressed in social media interactions and content threads. Advanced data visualizations are proposed in this observatory (such as word clouds, geolocations maps, etc.) to allow users to gain an evidence of the social data topics correlations providing zoom in and filtering capabilities, topic-level details, etc. This topic observatory allows users to easily comprehend thematic relationships and topics inter-dependencies.

Particular technology mature solutions, in Machine Learning (ML) and Natural Language Processing (NLP) areas, are exploited and advanced to deliver the required components that support the proposed crowdsourced topics observatory. The developed framework aims at offering a global and open interoperable environment among multiple stakeholders since it provides access both to topics summaries and to an open API, enabling various users' interactions. At the same time, it is easily customizable per thematic case and it follows a robust social media content processing pipeline to enable spotting of simple and sophisticated thematic correlations, trends and phenomena. By identifying qualitative content (in terms of readability and interestingness) and further classify it in a set of context-specific thematic categories, the topics correlations are evident and well aligned with their intensity and dynamics in the underlying collected crowdsourced dataset.

The remaining of this paper is organized as follows: Section 2 discusses the related work with emphasis on social media data driven topics and thematic detection and analysis. Section 3 summarizes the proposed social data schemas and objects design, while Section 4 discusses the proposed components and their data schemas functionality. Section 5 demonstrates the results of the proposed topics observatory testing with emphasis on a thematic related with social innovation. Finally, Section 6 includes the conclusions of the article and future potentials.

III. Related work

The challenges posed by the massive data sizes and low-quality content (e.g. poorly formatted, short textual entities, etc.) in social media platforms require advanced data modelling and analysis approaches. Existing relevant approaches focus on content classification (such as Twitter data threads) into categories such as substance, status style, social [Rampage10]. NLP approaches have been developed for assessing the expressed sentiment which may be more engaging for readers and can indicate their engagement capacity [Hoang13]. Moreover, the differences in privacy perceptions between users of different OSNs were studied and both Facebook and MySpace OSNs are concerned about privacy, yet this does not prevent them from sharing information online [Zhang15]. It is believed that often the perceived benefits of users outweigh the risk of personal privacy [Li14], [Debatin09]. Since users expose their views in open social platforms and in free forms, a more fine-grained and possible context-specific topics and thematic categorization is required. Such an approach requires the utilization of external domain specific knowledge provided by online sources (such as DBpedia and Wikipedia) or ontologies. For example, in [Gattani13] a Wikipedia-based global "real-time" knowledge base is utilized to automatically classify streaming data from Twitter. In addition, there are quite a few efforts for defining ontologies about products / brands that are used by such Linked Open Data efforts, such as The Product Types Ontology, the GoodRelations ontology, schema.org, and even the DBpedia and Freebase ontologies / folksonomies. The Product Types Ontology provides high-precision identifiers (ca. 300,000) for product or service types based on Wikipedia, extending the schema.org and GoodRelations standards for e-commerce mark-up. Therefore, social media topics related either to the company, the products and/or product types can be linked to these open classification standards to provide an improved semantics-aware repository for a thematic set of related terms.

Topic modelling techniques applied on social media content have also been researched in crowdsourcing applications. A two-stage hierarchical topic modelling system to address the clustering of noisy and sparse content from tweets [Wang2017], while topic modelling on Instagram hashtags was utilized to predict the subject of related images and improve image annotation [Argyrou2018].

The proposed work leverages on a terms-based crowdsourcing framework by building on advanced data modelling and mining approaches, to establish a robust social media content processing pipeline that will be able to identify qualitative content, and further classify topics in a set of context-specific categories. Analysis of the metadata and information around the collected UGC via topic modelling approaches are accompanied by ontology-based classification approaches to extract context along with features that will enable categorization in appropriate, predefined classes (such as crowdsourced ideas, thematic taxonomies, etc.). The proposed methodology for automatically collecting and identifying content relevant to specific topic areas acquires coarse-grained streams of content (such as posts and annotated multimedia and text content, as well as user reactions to them) via appropriate data collection mechanisms. Then, the identification and characterization of qualitative relevant information and the fine-tuning of the data collection parameters is followed exploiting and advancing the state of the art in the areas of qualitative crowdsourcing and on dynamic topic categorization. The proposed approach establishes a closed loop between a topics observatory taxonomy and a dynamic topic categorization model of social media content along with semantics annotations from existing available open data.

IV. Social Data schemas and objects design

The volume and velocity of data produced in Social

Networks increase at a very high pace and appropriate data objects and data models design are required. Due to the unstructured nature of social media posts, NoSQL database schemas are typically proposed. This work builds on a data schema which favors a document-based NoSQL database, since such technologies use an effective data model. The proposed data model considers that each record and its associated data forms an entity called "document". This document entity encapsulates all information related with the database object and it thus offers a generic model for multiple social media data integrations.

Data encoding should also follow an appropriate approach to match with this generic data model. Already some of the popular social media platforms have opened information about their preferred data models. For example, the Twitter API offers tweet data encoded by using Object Notation based on a particular technology (JavaScript – JSON, n.d.). JSON is based on key-value pairs, with named attributes and associated values. These attributes, and their state are used to describe objects. Such data object encodings match well with the NoSQL databases which perform very well in the handling of JSON encoded objects. The proposed social media knowledge observatory enables streaming data collection from Twitter and other social media and thus, the data schema proposed temporarily stores streaming data posts in accordance to such official Twitter schemas (e.g. as with Fig. 1 visualizing a Twitter object).

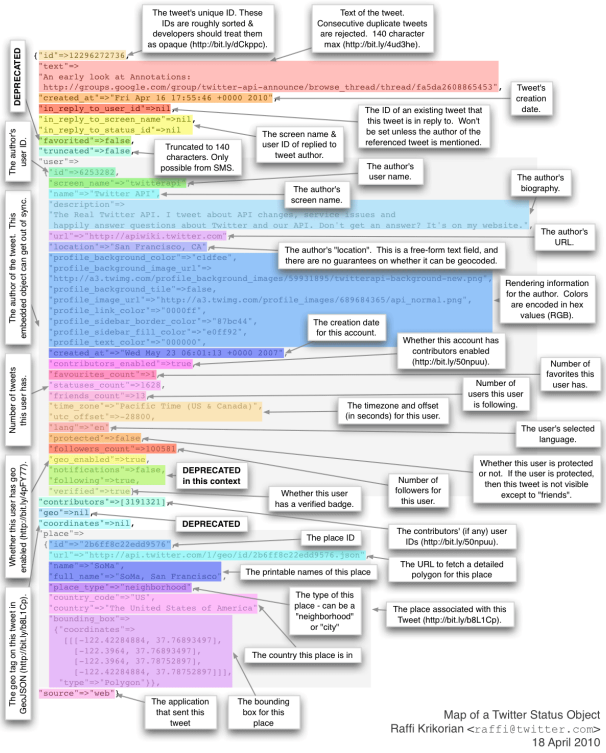

Figure 1: A sample of Twitter's JSON data object

The Twitter's JSON object contains (more than 150 attributes. However, only a small subset of these attributes is typically useful for any filtering process. In the proposed work we follow a flexible approach by using simple attributes with emphasis on those which embed information that could lead to topics detection. As depicted in Figure 2, social media data streams received by popular social media platforms are utilized by harvesting selected few and important attributes (such as geolocation information - Geo, Tweet timestamp, ids, hashtags etc.). These attributes are then used to support a filtering process which enables connections with specific and focused topic related components in the proposed observatory. The proposed observatory captures multiple users and content attributes to result in adaptive and effective data schemas which can support all of the proposed components.

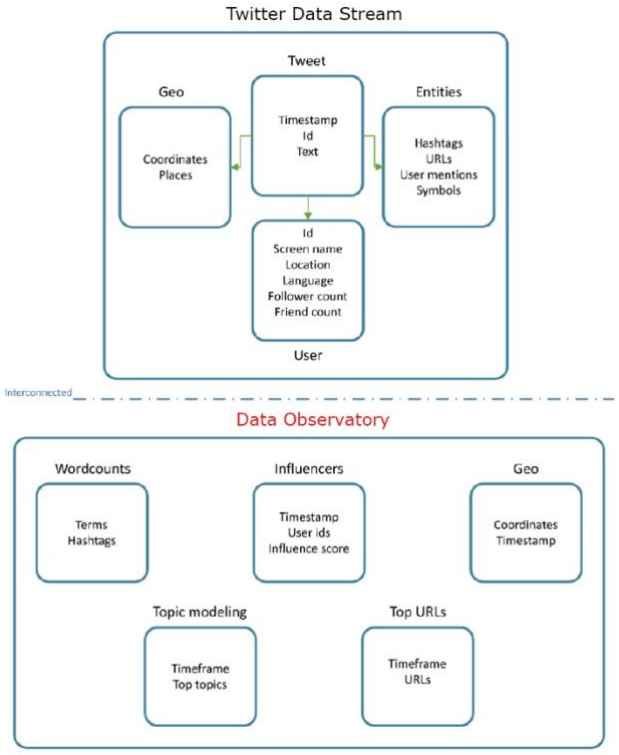

Figure 2: Data streams and their selected attributes for the Topics Observatory

V. Data Driven Components for a Crowdsourced Topics Observatory

The proposed design for a crowdsourced topics observatory is based on a flexible data schema model which involves specific sub components that are inter-connected but each with its own data schema used for storing its data filtering process results. All of the available components are detailed in the next subsection.

A. Topic Observatory components

The topics observatory components focus on content visualization (Wordclouds, Locations, Topic Modelling), influential content identification (Top Posts, Top URLs, Influencers) and discovery of crowdsourced open-licensed media related with a specific thematic (Repositories).

Wordclouds

The Wordclouds component is proposed and designed to visualize the most frequent hashtags and terms in the collected

data. Word clouds for both terms and hashtags are calculated on a frequent (e.g. daily) basis, for each keyword category. The results are stored in a JSON document, which contains only limited and necessary attributes. The data schema of the word clouds results is presented in Fig. 3 and includes the following attributes:

Figure 3: Wordclouds data schema

- Timestamp from [tweets with timestamp equal or higher than]
- Timestamp to [tweets with timestamp lower than]
- Language
- Keyword category (which category of keywords is responsible for collecting this tweet)
- All terms in this period and their frequency
- All hashtags in this period and their frequency

The terms and hashtags attribute actually represent a key – value pair, where the key is the term or hashtag and the value is the number of appearances of the specific term/attribute in this date range. The "timestamp from" – "timestamp to" period specify the duration of the analysis (e.g. they would vary 1 day if all word clouds are calculated in a daily basis). When a user performs a query for a certain date range, the system returns an aggregation of all the documents within this period. For example, if a user requests for the word clouds in the date range of July 1st – July 15th, the hashtags and terms word clouds included in the documents that are within this date range sum up and a final word cloud with all the hashtags and terms is generated. In general, the proposed schema of these attributes is depicted in Fig. 4.

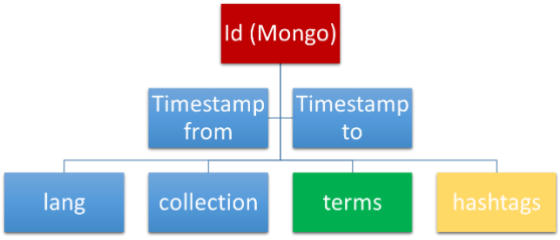

Figure 4: JSON schema for Wordclouds

Locations

The Locations component covers two tasks providing: a) a map visualization of the places with the highest crowd posting activity (such as tweeting) regarding the topic of interest and b) information about the terms and hashtags used more frequently in an area selected by the observatory user. For these tasks the data schema requires the following attributes:

- Geographical coordinates (Latitude, Longitude)
- Hashtags included in the post (if available)
- Terms included in the post's text
- Keyword category (which category of keywords is responsible for collecting this post)
- Timestamp
- Language

In task a) the user may have the option of selecting a keyword category, a certain time range and a specific language, while in task b) the user may also select a certain area in the map and retrieve the terms and hashtags with the highest frequency in it. The data schema used to store the results for this component has been designed in such a way, so that a single response to task a) also fulfills the requirements of task b) and vice versa. It is obvious that the data schema remains the same, compared to the one used in the Wordclouds component, with an addition of an array that contains the geographical coordinates of the tweet (Latitude, Longitude). The Locations data schema is described in Fig. 5.

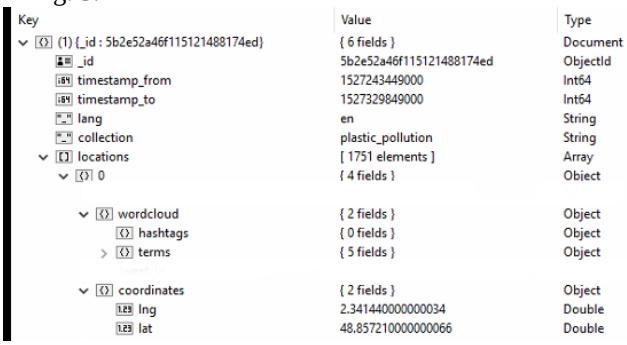

Figure 5: Locations data schema

Topic Modelling

Topic modelling is used to extract related sub-topics under the umbrella of a main theme. In our approach we utilized the probabilistic topic modelling method called Latent Dirichlet Allocation (LDA)[Blei2003] such that the data schemas do not store anything more than the time period in which the training has been done, the language, the keyword category and the path to the model that has been produced. LDA uses a bag of words approach to transform user's corpus into a vector of word counts. After the post's text is pre-processed for cleanup, a stemming process is applied to reduce the words in their base root. Then the training of the model occurs, while the next and final step is the evaluation of the results. A topics-to-words matrix for each collection is generated and the end-user can examine the words with the highest weight inside a topic. The component's data schema is provided in Fig. 6.

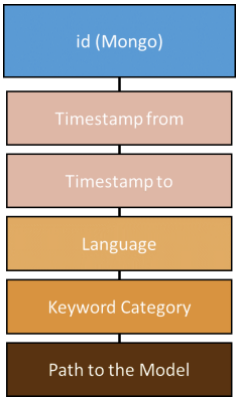

Figure 6: Topic Modelling data schema

Top Posts and Top URLs

The proposed topic observatory also includes a component which produces a list with the most propagated URLs (included in social media posts) and a list of the top posts (based on the number of retweets in the case of Twitter). The official Twitter JSON document contains an object called "entities" which includes potential URLs added to the original tweet. A process running in the back-end keeps track of the URLs posted each day. Following a similar approach to the one demonstrated in the Wordclouds component, the results of the Top URLs use the data schema depicted in Fig. 7.

Key	Value	Type
(4) {_id : 5b6ac4956f1151319c2a2ddb}	{ 6 fields }	Document
_id	5b6ac4956f1151319c2a2ddb	ObjectId
timestamp_from	1521139464000	Int64
timestamp_to	1521225864000	Int64
lang	en	String
collection	innovations	String
top_urls	[0 elements]	Array

Figure 7: Top URLs data schema

The *top_urls* is supported by a key-value pairs array, where key is the URL and the value is the number of appearances in the time period defined by the timestamp from to the timestamp to values. The official Twitter JSON document also contains a *retweeted_status* object containing, amongst others, an attribute that counts the number of times the original post has been retweeted. Following the same methodology described earlier in the Top URLs component we end up with a similar schema for the Top Posts, with the only difference in the case of Twitter being that the "number of retweets" is calculated in a cumulative manner. Thus, instead of storing each occurrence we only store the highest to avoid duplicates.

Influencers

The Influencers component is designed to identify the top k (top 100 for example) influencers in a given crowdsourced dataset. To discover these users, a social graph model is proposed based on several metadata (such as the retweets, mentions and replies) among all users for whose posts been collected. To generate a social graph that can produce reliable results regarding the detection of expert users, a large number of nodes and edges is enabled. Thus, the influencer detection algorithm runs for a period of time (e.g. one month) and in an accumulative manner (e.g. two months, three months, etc.). This process runs periodically in the back-end part of the designed crowdsourcing observatory. The results of the influencer identification process follow the next data schema (see Fig. 8):

Key	Value	Type
(14) {_id : 5b6acfbc6f11510d587d5e96}	{ 6 fields }	Document
_id	5b6acfbc6f11510d587d5e96	ObjectId
timestamp_from	1532108664000	Int64
timestamp_to	1532972664000	Int64
lang	en	String
collection	innovations	String
influencers	[12 elements]	Array
0	2327035620	String
1	122030887	String
2	462152975	String
3	2438302135	String
4	3004320047	String
5	626261903	String
6	2327277326	String
7	824984548677185536	String
8	258710092	String
9	1205042994	String
10	16900850	String
11	932594007489826816	String

Figure 8: Influencers data schema

The data schema includes the following attributes:

- Timestamp from
- Timestamp to
- Language
- Keyword category
- List of identified influencers

Repositories

In this component various media streams, associated with the specific thematic under examination and knowledge extracted previously, are captured and presented to the end-users in real time utilizing the official APIs of the respective services. Social media platforms such as Thingiverse (hosting open 3D printer designs) and Flickr (photos sharing) are utilized. Additionally, an Instagram crawler locates posts associated with predefined hashtags. These hashtags are associated with activities that are immediately linked with the thematic of interest for the observatory offering handpicked high-quality content to the user.

B. Topic Observatory Open API

To further facilitate the use of the proposed topic observatory and strengthen its interoperability, an open API is designed and implemented exploiting REST API technologies. Registered users of the observatory are able to make calls to this open crowdsourcing API and receive the information in response to their request. The user is enabled to make API calls to certain endpoints of the API and retrieve a JSON object, according to the called endpoint. Figure 9 presents all available endpoints and the description of the retrieved JSON object response provided back to the caller.

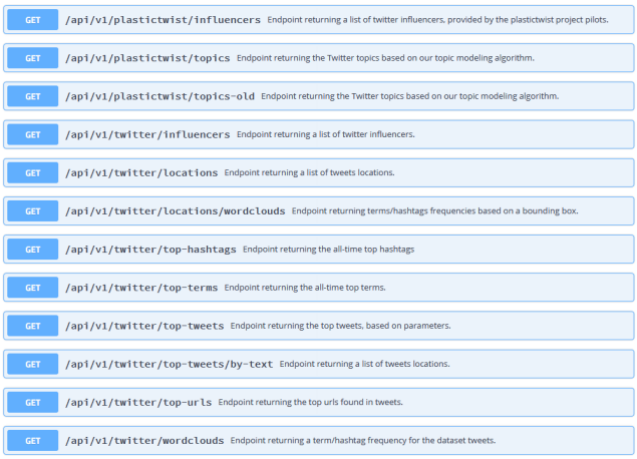

Figure 9: Crowdsourcing API endpoints

Our endpoints were semantically mapped based on the social media data that users would like access to. All the endpoints follow a structure that can be easily matched by the user to the corresponding developed components of the observatory. For example:

- /api/v1/twitter/top-urls
- /api/v1/flickr/top-photos
- /api/v1/thingiverse/top-things

Since the present work handles social media data which are semi-structured and not model-centric, there was not a need for complex frameworks with various drivers and tools built-in. Instead we opted for a lightweight solution ensuring high performance and flexibility. Therefore, the development of the API was completed using Python's Flask with the Flask-Restful extension.

VI. Experimentation on a Social Innovation Platform

The proposed components have been developed and tested under a topics observatory platform which focuses on the sensitive issue of plastic waste revaluation, aiming to promote collective awareness and social innovation. Indicative datasets outcomes are presented next with a focus on the topics related with plastic pollution and reuse.

Wordclouds

Figures 10-11 showcase some results from the wordclouds visualizations regarding plastic innovations and plastic pollution from prosts in Twiteer in a two month period.

Figure 10: Tagcloud for the plastic innovations category

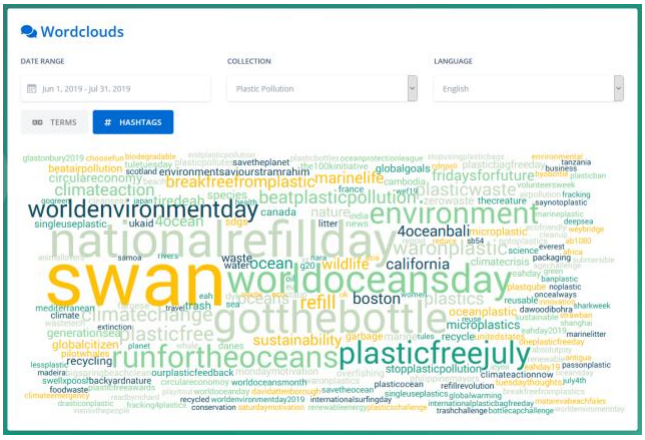

Figure 11: Tagcloud for the plastic pollution category

The observatory allows users to chose all the desired parameters (date range, collections, language) present in the data schema of the Wordclouds component to output the requested visualizations. Moreover, users can proceed with selecting words of their interest that are present in the generated cloud to explore the top social media posts that contain those terms (see Fig. 12). Thus, exploiting an interlink between various components, Wordclouds and Top Posts in this case, that encourages further user engagement and awareness.

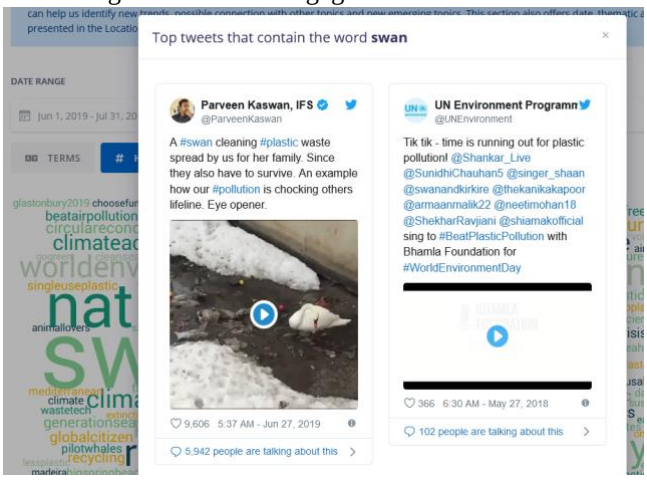

Figure 12: Interconnection between Wordclouds & Top Posts

Locations

Social media posts collected in the Locations sub component originate from users that share either their exact or approximate location or they state their location in their social network profile. In our case, where Twitter is utilized as data source, tweets that can not be associated with any geographical information are discarded. The total number of tweets collected, along with their geo-information grouped at city level, is aggregated and the results are displayed over a world map where users can pan and zoom in real time (Fig. 13). Again, users can define all the parameters (date range, collections, language) present in the data schema of the Locations component to generate the desired output.

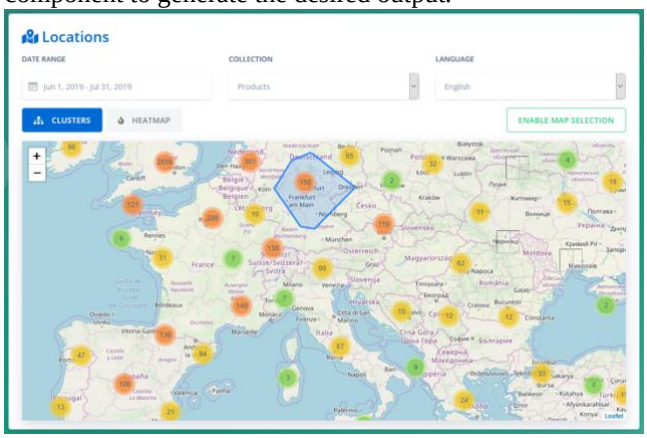

Figure 13: Dynamic map displaying terms density over geographical regions

This module can also interact with the Worldclouds component when a user selects a specific geographical region on the map to view the popular terms and hashtags of that location (Fig. 14).

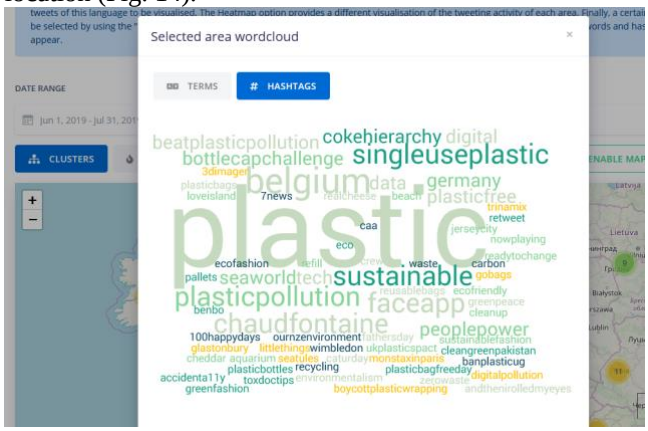

Figure 14: Tag cloud generated for a user selected geographical region

Topic Modelling

In this component users will gain an overview of the social media content and what they should expect to find in each of the available collections. After selecting the collection, they are interested in, they can explore the discovered topics in an interactive graph that also displays information regarding the frequency of the terms for each topic structured in a histogram (Figures 15-16). Topic Modelling covers the entire time span of the collected social media data.

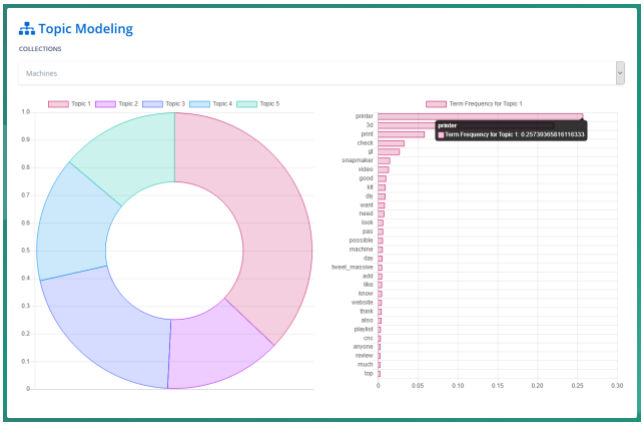

Figure 15: Identified topics from the plastic machines collection

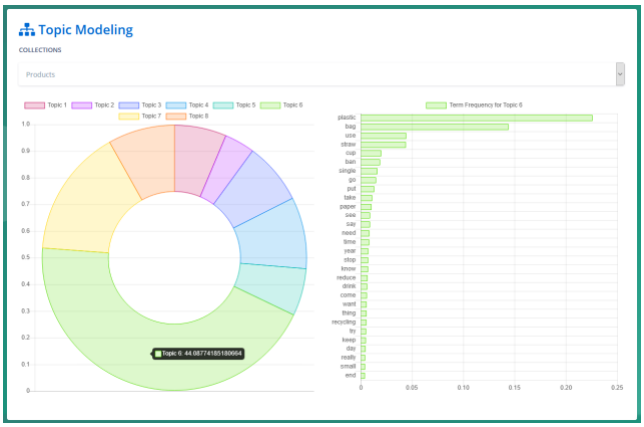

Figure 16: Identified topics from the plastic products collection

In Fig. 15 the prevailing identified topic of the *plastic machines* collection includes terms related with 3D printers and Snapmaker[1], popular among the makers community super-compact 3-in-1 machine supporting 3D printing, laser-engraving and CNC carving. These results are highly correlated with the discovered topic since 3D printing is one of the most used technology in plastic upcycling.For the *plastic products* collection the most significant topic is related with single-use products like plastic bags, straws and cups (Fig. 16). Single-use plastics, or disposable plastics, are a major concern in the discussion surrounding plastic pollution since their useful lifecycle is very short and they end-up in the environment as plastic waste really quick.

Top Posts

The quality of social media content can be tracked by filtering out posts based on other users' interactions. For Twitter, these interactions refer to the number of times a post has been retweeted, replied or marked as favorite. To select the top collected tweets, a ranking by retweets, favorites and replies count is performed and presented to users (Fig. 17).

[1] https://snapmaker.com/

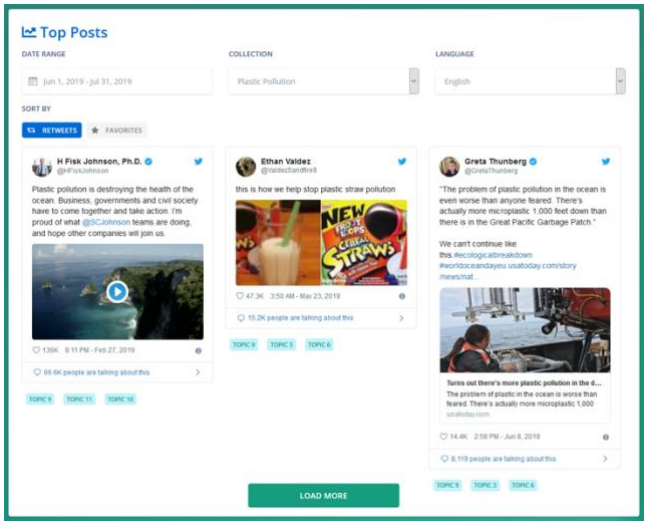

Figure 17: Top posts by retweets in the plastic pollution collection

The displayed top posts are also associated with the discovered topics from the Topic Modelling component. Pressing the buttons under each post will transfer the user to the respective identified topic for further exploration of the correlated terms under the specific topic.

Top URLs

Since social media posts often contain web links, these are also filtered by calculating their appearance frequency in the collections of every month and then sorted in descending order to determine the top URLs (Fig. 18).

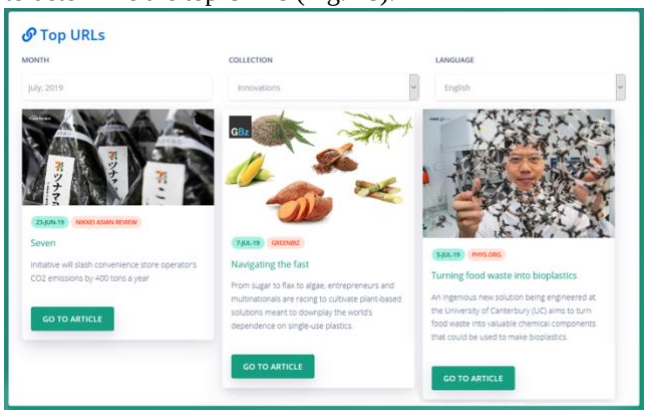

Figure 18: Top URLs in the plastic innovations collection

Influencers

This component supports the detection of influencers (users that diffuse information related to the subject of interest more efficiently throughout the network) among the Twitter social network using either the NetShield algorithm [Tong10] or the betweenness centrality metric [Kitsak10]. Users select the criteria they prefer (time period monthly based, collection, language and algorithm) and are presented with a list of the most influencing Twitter accounts (Fig. 19).

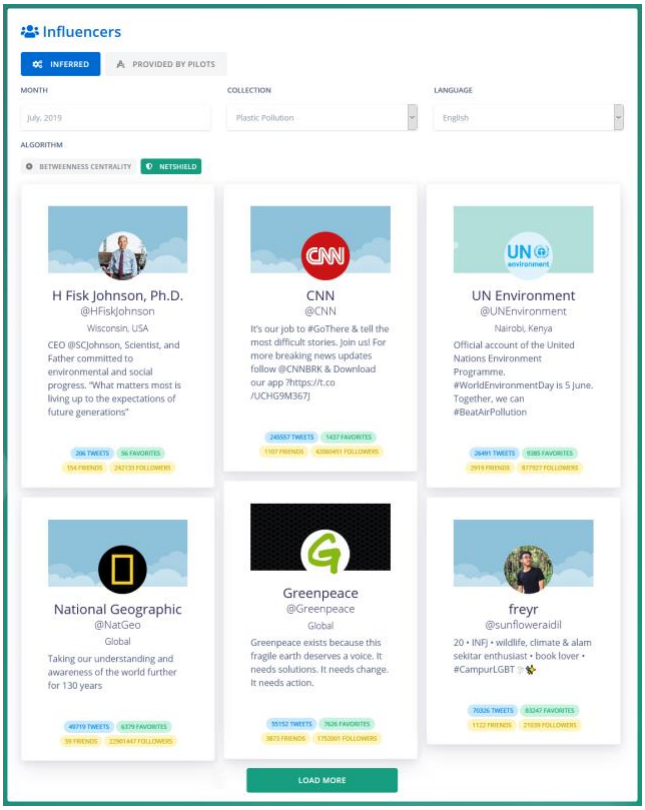

Figure 19: Top influencers regarding plastic pollution by NetShield

Repositories

The following Figures 20-23, contain examples of related external media content that is discovered and presented to the users via the crowdsourced observatory. For Thingiverse poplar, featured and new open 3d printer designs can be displayed while for Flickr users are asked to select one of the top hashtags, as determined by the observatory, or provide their own to discover related photos.

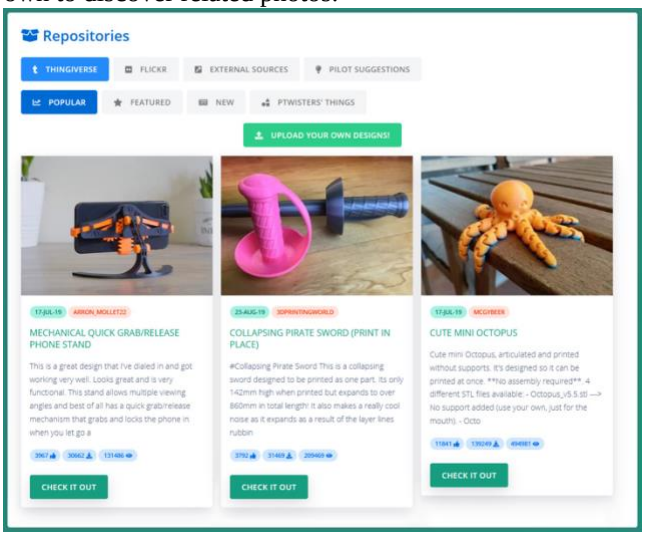

Figure 20: Popular open 3d printer designs from Thingiverse

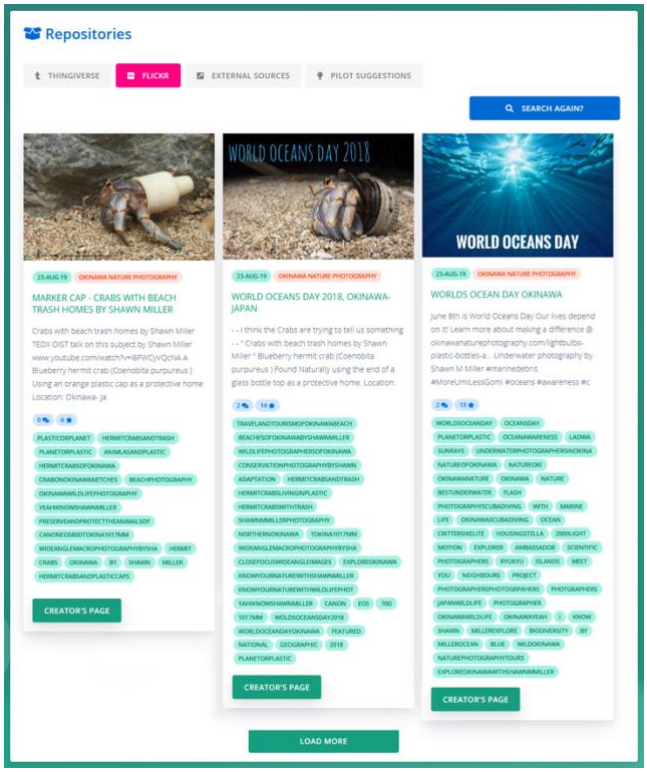

Figure 21: Flickr photos related with the #worldoceansday hashtag

Posts with predefined hashtags related with the observatory's thematic are collected and visualized through the Instagram feed (Fig. 22).

VII. Conclusions and Future Work

In the present work appropriate data schemas and a flexible and effective data model have been specified in order to successfully incorporate crowdsourced content from heterogenous social media sources. The presented document-based data model has allowed us to address challenges due to the high volume and velocity of social media data and has been exploited successfully towards the developing of a crowdsourced topics observatory. This dynamic observatory, that utilizes interactive visualization and advanced topic modelling methods, supports effectively multiple social media data schemas under a single thematic. Emphasis was placed to the design of the data schemas to successfully describe the unstructured social media content and users' interaction in social networks while allowing for efficient interoperability with the observatory's visualization applications. The fine-tuned combination of the available social media resources under various appropriately adapted approaches allows to efficiently capture the dynamics, trends and patterns relevant with the given thematic, depending on an initial set of keywords and a given hierarchy. The implementation of a streamlined user interface with advanced visualizations offers a high-level experience and customization capabilities where the user is able to explore and delve into the collective intelligence of the crowd offered.

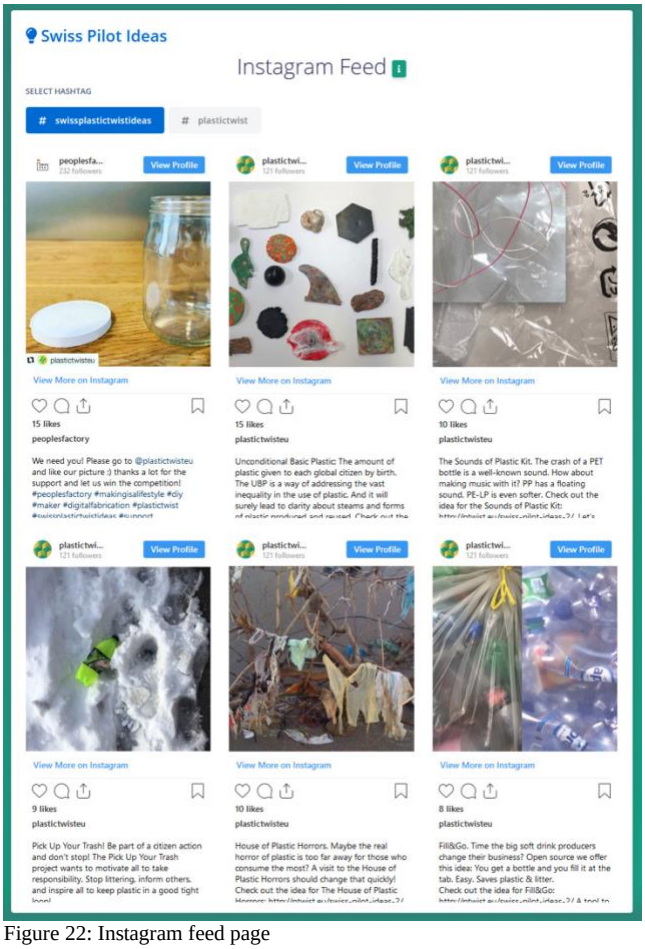

Figure 22: Instagram feed page

ACKNOWLEDGMENTS

Parts of this work have been supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreements No. 780121 and No. 871403.

REFERENCES

[1] Argyrou, A., Giannoulakis, S. and Tsapatsoulis, N. Topic modelling on Instagram hashtags: An alternative way to Automatic Image Annotation? *13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Zaragoza, 2018, pp. 61-67.

[2] Blei, David M., Ng, Andrew Y. and Jordan, Michael I. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, (3/1/2003), 993–1022.

[3] Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M., and Vesci, G. 2013. Choosing the right crowd: expert finding in social networks. *In Proceedings of the 16th International Conference on Extending Database Technology*, ACM, March 2013, pp. 637-648.

[4] Brem, A., and Volker B. 2015. The search for innovative partners in co-creation: Identifying lead users in social media through netnography and crowdsourcing. *Journal of Engineering and Technology Management* 37 (2015): 40-51.

[5] Debatin, B., Lovejoy, J. P., Horn, A. K., and Hughes, B. N. 2009. Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences. *Journal of Computer-Mediated Communication* 15:83–108.

[6] Dong, X., Mavroeidis, D., Calabrese, F., and Frossard, P. 2015. Multiscale event detection in social media. *Data Mining and Knowledge Discovery*, 29(5), 1374-1405.

[7] Duan, Y, Chen, Z., Wei, F., Zhou, M., and Shum, H.-Y. 2012. Twitter Topic Summarization by Ranking Tweets using Social Influence and Content Quality. *Proceedings of COLING 2012*: 763-780.

[8] Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., Subramaniam, S., Rajaraman, A., Harinarayan V., and Doan, A.H. 2013. Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach. *Proc. VLDB Endow*. 6, 11 (August 2013), 1126-1137.

[9] Ghosh, S., Sharma, N.K., Benevenuto, F., Ganguly, N., and Gummadi, K.P. (2012). Cognos: crowdsourcing search for topic experts in microblogs. SIGIR.Proceedings of the 35th international ACM conference on research and development in information retrieval.

[10] Hoang, T.-A., Cohen, W.W., Lim, E.-P., Pierce, D., and Redlawsk D. P. 2013. Politics, sharing and emotion in microblogs. *In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM '13). ACM, New York, NY, USA, 282-289.

[11] Kanavos, A., Perikos, I., Vikatos, P., Hatzilygeroudis, I., Makris, C. and Tsakalidis, A. 2014. Modeling ReTweet Diffusion using Emotional Content. *IFIP International Conference on Artificial Intelligence Applications and Innovations* (AIAI 2014), Rodos, Greece.

[12] Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., and Makse, H.A. 2010. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11), 888–893.

[13] Li, Y. and Smeaton, A.F. 2014. From Smart Cities to Smart Neighborhoods: Detecting Local Events from Social Media, *ECIR'14 Information Access in Smart Cities Workshop*, (i-ASC).

[14] Rampage, D., Dumais, S., and Liebling, D. 2010. Characterizing microblogs with topic models. *In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[15] Rjab, A. B., Kharoune, M., Miklos, Z., and Martin, A. 2016. Characterization of experts in crowdsourcing platforms. *In International Conference on Belief Functions*, September 2016, pp. 97-104. Springer, Cham.

[16] Santos, dos C., and Gatti, M. 2014. Deep convolutional neural networks for sentiment analysis of short texts. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, August 2014, 69–78.

[17] Tong, H., Prakash, B. A., Tsourakakis, C., Eliassi-Rad, T., Faloutsos, C. and Chau, D. H. 2010. On the vulnerability of large graphs. *In Proceedings - IEEE International Conference on Data Mining*, ICDM, 1091-1096.

[18] Unankard, S., Li, X., and Sharaf, M. A. (2015). Emerging event detection in social networks with location sensitivity. *World Wide Web*, 18(5), 1393-1417.

[19] Wang, B., Liakata, M., Zubiaga, A., Procter, R. 2017. A Hierarchical Topic Modelling Approach for Tweet Clustering. *In: Ciampaglia G., Mashhadi A., Yasseri T. (eds) Social Informatics. SocInfo 2017*. Lecture Notes in Computer Science, vol 10540. Springer, Cham

[20] Xie, W., Zhu, F., Jiang, J., Lim, E. P., and Wang, K. 2016. Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, 28(8), 2216-2229.

[21] Xintong, G., Hongzhi, W., Song, Y., and Hong, G. 2014. Brief survey of crowdsourcing for data mining. *Expert Systems with Applications*, 41(17), 7987-7994.

[22] Zhang, J. 2015. Voluntary information disclosure on social media. *Decision Support Systems* 73: 28-36.

[23] Zubiaga, A., Spina, D., Martínez, R., and Fresno, V. 2015. Real-Time Classification of Twitter Trends. *Journal of the Association for Information Science and Technology, 66(3), 462–473.*