Project full title: *Orchestrating Information Technologies and Global Systems Science for Policy Design and Regulation of a Resilient and Sustainable Global Economy*

Contract no.: 611875

D2.2

Early version of social media based policy indicators

Workpackage:	WP2	Collective Intelligence for Nowcasting
Editor:	Luis Rei	JSI
Author(s):	Luis Rei	JSI
	Mario Karlovcec	JSI
	Gregor Leban	JSI
	Dunja Mladenić	JSI
	Marko Grobelnik	JSI
	João Pita Costa	JSI
	Inna Novalija	JSI
	Miha Papler	JSI
Authorized by	Silvano Cincotti	UNIGE
Doc Ref:		
Reviewer	Annarita Colasante	UNIVPM
Dissemination Level	PU	

SYMPHONY Consortium

No	Name	Short name	Country
1	UNIVERSITA DEGLI STUDI DI GENOVA	UNIGE	Italy
2	INSTITUTE OF COMMUNICATION AND COMPUTER SYSTEMS	ICCS	Greece
3	PlayGen Ltd	PlayGen	United Kingdom
4	GCF - GLOBAL CLIMATE FORUM EV	GCF	Germany
5	Athens Technology Center	ATC	Greece
6	UNIVERSITA POLITECNICA DELLE MARCHE	UNIVPM	Italy
7	INSTITUT JOZEF STEFAN	JSI	Slovenia
8	Germanwatch Nord-Sued-Initiative e.V.	GW	Germany
9	UNIVERSITAT JAUME I DE CASTELLON	UJI	Spain

Document History

Version	Date	Changes	Author/Affiliation
V0.01	14-03-2015	Added summary, introduction and event detection	Luis Rei/JSI
V0.02	16-03-2015	Added diffusion section	Mario Karlovcec/JSI
V0.03	20-03-2015	Added correlation and prediction sections	Gregor Leban/JSI
V0.04	22-03-2015	Added sentiment section	Luis Rei/JSI
V0.05	23-04-2015	Summary and introduction improved	Mario Karlovcec/JSI
V0.06	25-04-2015	Added discussion and conclusions sections	Luis Rei/JSI
V0.07	27-04-2015	Methodology	Mario Karlovcec/JSI
V0.08	28-04-2015	Small fixes in multiple sections	João Pita Costa/JSI
V0.09	30-04-2015	Improved discussion and nowcasting sections	Luis Rei/JSI
V0.10	01-05-2015	Use Cases	Luis Rei/JSI
V0.11	02-05-2015	JSI Review	Dunja Mladenić/JSI Marko Grobelnik/JSI Inna Novalija /JSI
V0.2	2-04-2015	Overall revision	Dunja Mladenic/JSI
V0.21	2-04-2015	Formatting Issues	Luis Rei/JSI
V0.3	7-04-2015	Review	Annarita Colasante/UNIVPM
V1.0	10-04-2015	Changes related to review	Luis Rei/JSI Mario Karlovcec/JSI

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

Executive Summary

Deliverable **D2.2 - Early version of social media based policy indicators** presents the work performed in SYMPHONY under WP2 Collective intelligence for Nowcasting, in particular in SYMPHONY Task 2.2 - Tracking Cross-Lingual Information and Opinion Diffusion and the early work under SYMPHONY Task 2.3 - Definition and Development of the SYMPHONY social media based expectations' indicators.

This report describes how we track information flow in social media along several dimensions: topic (event), geography, time, intensity and opinion. In this way a rich set of features that captures textual, structural and temporal dimension of social media and news data is obtained. This deliverable also describes how to use this information to improve social media based expectations' indicators and to provide social media based information for policy makers through the SYMPHONY integrated dashboard in Task 5.4 and to other SYMPHONY tasks.

Social media and news reflect the events, trends and opinions that are governing economic states and events in the World. Being aware of different types of signals that can be derived from social media and news, a major attention of this report is on methods for extracting different types of features from data. Social media and news are strongly intertwined in the real world. In section four we describe a method for matching these two sources. Section five addresses the issue of cross-lingual information and opinion diffusion, which uses the information about arrival of news and posts about different topic and also information about geographical location, to model the dynamics and geography of diffusion. In section seven we describe sentiment classifier, which is a method for determining if an opinion from social media post is positive, negative, or neutral.

In order to use collective intelligence for Nowcasting, we aim at creating a model based on a large collection of features obtained from social media and news data. We want to capture many different types of features to create the best possible regression model that can successfully tackle Nowcasting, automatic polling for Symphony use-cases, initialization of variables of SYMPHONY ABM model, and other related tasks.

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

Table of Contents

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

List of Figures

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

List of Tables

Abbreviations

CMA	Cumulative Moving Average
CRISP-DM	Cross Industry Standard Process for Data Mining
ER	Event Registry
MST	Minimum Spanning Tree
NER	Named Entity Recognition
SVM	Support Vector Machines

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

1 Introduction

This deliverable **D2.2 Early version of social media based policy indicators**, presents the work performed in SYMPHONY under **WP2 Collective intelligence for Nowcasting**. The deliverable consists of work described under **Task 2.2 - Tracking Cross-Lingual Information and Opinion Diffusion** and the early work described under **Task 2.3 - Definition and Development of the SYMPHONY social media based expectations' indicators**. This deliverable is focused on methods for extracting a rich set of features from social and news media, that exploit textual, structural and temporal data modality that will be later used in Nowcasting (predicting the present value of a given indicator).

In the first section of the deliverable we illustrate the importance of social media mining and show how to integrate signals from social and news media into SYMPHONY system. Next, we give an overview of the SYMPHONY social media mining infrastructure consisting of Twitter Observatory, Newsfeed and Event Registry. While Twitter Observatory is a social media mining tool developed specifically for Symphony as a part of D2.1 under **Task 2.1 - Establishing Social Media monitoring Data Infrastructure**, Newsfeed was partially supported by other EU projects RENDER, X-Like, PlanetData and MetaNet FP7 projects and is used in SYMPHONY as a source real-time large-scale source of news articles. Event Registry was developed under X-Like EU project and is used in SYMPHONY as the basis for developing a system for matching social media and news media events.

In real world, social media posts and news articles are tightly interconnected. In section four, we describe an approach for matching tweets to events derived from news articles. By combining both social media and standard news media, SYMPHONY can more comprehensive capture events and trends in the World.

In social media mining we are not only interested in content of posts and news articles, but also in the way the information diffuses and spreads. We cover the structure of diffusion dynamics and geographical spread of news articles and social media posts in section five. In this way we are enriching the set of signals captured from media with temporal and structural data modality.

We find sentiment an important aspect of social media. Knowing if a post is negative, positive, or neutral is a crucial feature for analyzing opinions. Sentiment is another type of signal used for correlation and Nowcasting problem. Sentiment identification is described in section six.

In section seven we describe a method for time series based correlations. Computing correlations is based on the Google correlate algorithm, but is applied on about 1.8 million unique concepts identified with Event Registry from real-time feed of news articles. We show how to use correlation for Nowcasting, using different types of signals derived from

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

matching social media posts to news articles, concepts and events, sentiment and information diffusion.

Finally, we present usage scenarios to how social media mining components fit to the SYMPONY system. In particular, this is illustrated using the usage scenarios of early version of the Regression and Nowcasting components.

2 The Role of Social Media Mining within SYMPHONY

Social media mining is one of the key components of integrated innovative SYMPHONY platform that will provide effective support for decision making.

Social media is a source of information, opinions, ideas or any kind of other expression of individuals or organizations, which arose with development of web and mobile internet and enables instant and open global communication. There is a long list of social media applications that includes Twitter, YouTube, Instagram, LinkedIn, Facebook, Snapchat, Tumblr and Flicker. Many of these applications are very intensively used by many people. For instance more than 350,000 tweets are send every minute, what makes 500 million tweets each day [1]. Our lives are increasingly interconnected with the Web. According to Internet Live Stats[1], 1,920 Instagram posts, 1,842 Tumblr posts and 48,196 Google searches are made each second. As NewInc [2]puts it, “we think in Tweets, see in Instagrams, and try to scroll on analog devices”. Alongside social media that is a new type of information source, development of technology revolutionized classical media such as print newspapers, newsmagazines, televisions and radio news broadcasters. Today many such media have online additions and blogs. Moreover, news media and social media are tightly interweaving, with social media posts as reactions to news article and social media trends as regular parts of news reports.

We perform social media mining in interconnection with mining electronic news media. In social data mining, we perform analysis across-languages and consider different aspects of the data including text, sentiment, geographical spread and diffusion dynamics. We do this to obtain a rich set of features which are used for correlations and in the final phase for predicting current actual values of macroeconomic indicators – Nowcasting. Results of social media integrate with other SYMPHONY components in several ways. Social media module will be a part of integrated SYMPHONY dashboard with Nowcasting functionality, which uses a rich set of features extracted with methods described in sections four, five and six of this deliverable. In the final phase, social media signals will be used to initialize the parameters of

[1] http://www.internetlivestats.com/one-second/

[2] http://www.newinc.org/blog-post/2015/2/4/newhive

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

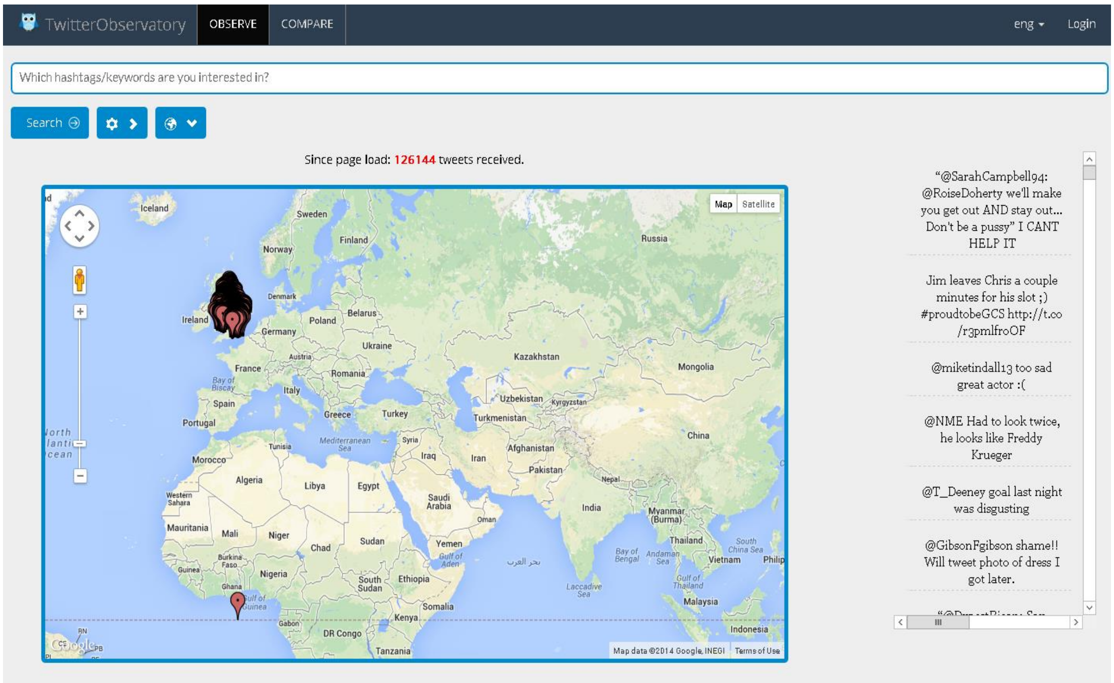
SYMPHONY ABM model. In this way social media mining is used to collect citizens' economic expectations and calibrate the expectations of artificial agents in the agent-based model.

3 Infrastructure Overview

Observing social media is performed within IJS NewsFeed and TwitterObservatory, and using Event Registry as the basis for matching tweets and news articles.

3.1 Twitter Observatory

TwitterObservatory[3] is a social media mining tool for data observation developed for SYMPHONY under WP2 Collective Intelligence for Nowcasting, in particular in SYMPHONY Task 2.1 - Establishing Social Media monitoring Data Infrastructure and described in SYMPHONY Deliverable **D2.1 Social media streams processing infrastructure**. TwitterObservatory crawls, stores and enriches tweets, provides search and analytics [2]. In addition, it provides a suitable user interface for these functionalities which can be seen in Figure 1.



Since page load: 126144 tweets received.

"@SarahCampbell94: @RoiseDoherty we'll make you get out AND stay out... Don't be a pussy" I CANT HELP IT

Jim leaves Chris a couple minutes for his slot ;) #proudtobeGCS http://t.co/r3pmlfroOF

@miketindall13 too sad great actor :(

@NME Had to look twice, he looks like Freddy Krueger

@T_Deeney goal last night was disgusting

@GibsonFgibson shame!! Will tweet photo of dress I got later.

Figure 1: *Twitter Observatory showing UK tweets*

We use the geo coordinates from 4 EU countries: UK, Italy, Germany and Slovenia. For each country we selected the 10 largest cities by population and created geo coordinate bounding

[3] www.twitterobservatory.net

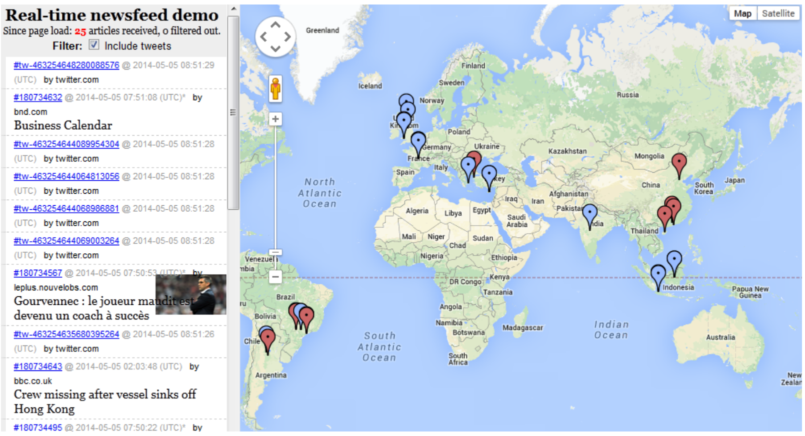
boxes around each city. These bounding boxes are supplied as parameters of Twitter API requests which then returns tweets geo-tagged with coordinates inside these boxes.

3.2 Newsfeed

IJS NewsFeed[4] is a tool for gathering data from news and social media, such as blogs and Twitter. Its development was supported in part by the RENDER[5], X-Like[6], PlanetData[7] and MetaNet[8] EU FP7 projects.

The IJS NewsFeed pipeline [3] is composed so that the tool periodically crawls RSS feeds, receives links to news articles and blogs, downloads and parses received inputs, taking care not to overload any of hosting servers. The new RSS sources mentioned in HTML are discovered and processed.

In addition to news and blog sources, IJS NewsFeed receives a 1% sample feed from Twitter, which gives us the possibility to effectively use this social media data in Symphony.



Real-time newsfeed demo
Since page load: 25 articles received, 0 filtered out.
Filter: ☑ Include tweets

#tw-463254648280088576 @ 2014-05-05 08:51:29 (UTC) by twitter.com

#180734632 @ 2014-05-05 07:51:08 (UTC)* by bnd.com
Business Calendar

#tw-463254644089954304 @ 2014-05-05 08:51:28 (UTC) by twitter.com

#tw-463254644064813056 @ 2014-05-05 08:51:28 (UTC) by twitter.com

#tw-463254644068986881 @ 2014-05-05 08:51:28 (UTC) by twitter.com

#tw-463254644069003264 @ 2014-05-05 08:51:28 (UTC) by twitter.com

#180734567 @ 2014-05-05 07:50:53 (UTC)* by leplus.nouvelobs.com
Gourvennec : le joueur maudit est devenu un coach à succès

#tw-463254635680395264 @ 2014-05-05 08:51:26 (UTC) by twitter.com

#180734643 @ 2014-05-05 02:03:48 (UTC) by bbc.co.uk
Crew missing after vessel sinks off Hong Kong

#180734495 @ 2014-05-05 07:50:22 (UTC)* by

Figure 2: *IJS Newsfeed Demo*

As a constantly evolving tool, IJS NewsFeed gives its users the possibility to observe the upcoming news, blogs and tweets through the real-time newsfeed demo (Figure 2). In Figure 3, the article locations are showed on the map, and more detailed information on the inputs (date and time of the article/blog/tweet, title, accompanied images, publisher etc.) is provided on the left panel of the page.

[4] http://newsfeed.ijs.si

[5] http://render-project.eu

[6] http://xlike.org

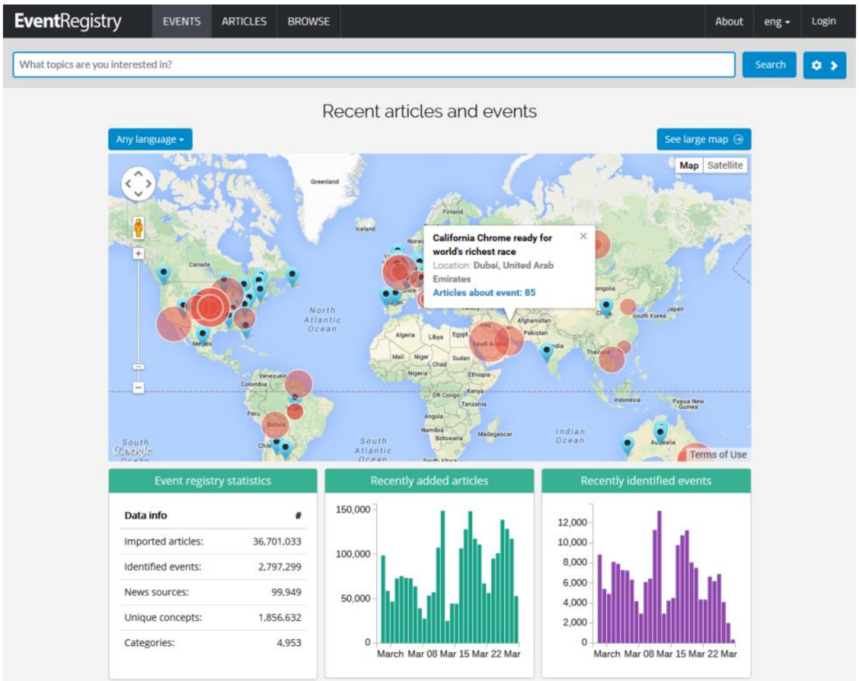
[7] http://www.planet-data.eu

[8] http://www.meta-net.eu

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

3.3 Event Registry

Event Registry[9] is a system that identifies world events by analyzing news articles. The system, described in [4] and [5] is able to identify groups of articles that describe the same event. It can identify groups of articles in different languages that describe the same event and represent them as a single event. From the articles in each event it can then extract event's core information, such as event location, date, who is involved (NER) and what is it about (categorization, keyword extraction). A user interface, shown in Figure 3, is available and it allows users to search for events using extensive search options, to visualize and aggregate the search results, to inspect individual events and to identify related events.



California Chrome ready for world's richest race
Location: Dubai, United Arab Emirates
Articles about event: 85

Data info	#
Imported articles:	36,701,033
Identified events:	2,797,299
News sources:	99,949
Unique concepts:	1,856,632
Categories:	4,953

Figure 3: *Event Registry*

Effectively, as the name implies, Event Registry is effectively a Knowledge Base for world events. Some Event Registry data metrics are displayed in **Errore. L'origine riferimento non è stata trovata.**. The concepts counted in **Errore. L'origine riferimento non è stata trovata.** can either be people, locations, organizations or non-entities (such as "accident", "bank", "money", etc.). For each of these concepts, the system maintains the number of times the concept was mentioned on each day since the system was started. The data in ER spans from the middle of December 2013 until now.

[9] http://eventregistry.org

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

Table 1: Event Registry metrics (at the time of writing, February 2015)

Data	Count
Articles	36,972,624
Events	2,823,398
News sources	100,132
Unique concepts	1,864,284
Categories	4,956

4 Event Detection

In the context of SYMPHONY we are particularly interested in identifying events that are related to the SYMPHONY objectives of preventing and mitigating economic and financial crises and fostering economically and ecologically sustainable growth. These are hypothesized to be big society level events rather than events that happen to a single individual. These events are also likely to be reported on by mainstream media. Because Event Registry already identifies events from news articles, effectively acting as a knowledge for event, we can reduce the problem of identifying events in social media to matching social media messages to events in Event Registry. The downside is that events that occur only on social media will not be identified, the upside is that the many irrelevant events in social media are also ignored.

There are several significant advantages in using Event Registry matching, depicted in Figure 4, as our basis for event detection:

- Event Registry is already cross-lingual with events (and extracted information) in English, German, Spanish Catalan, Portuguese, Italian, French, Russian, Arabic, Turkish, Chinese, Slovene, Croatian and Serbian;

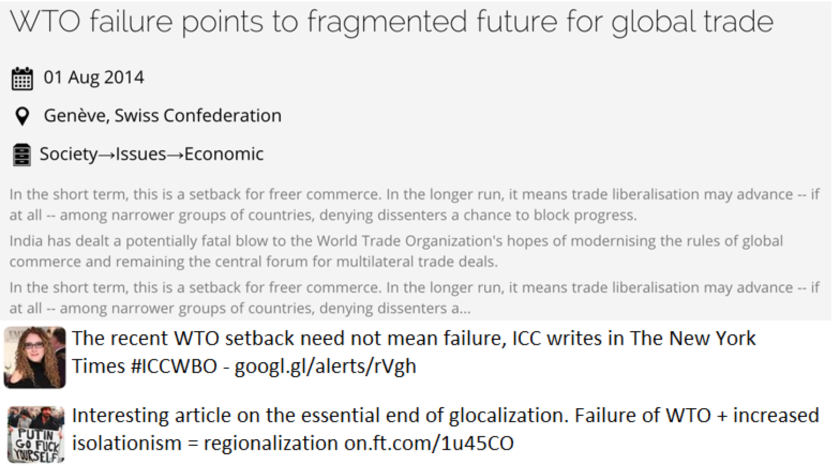
- it contains additional information about each event from Mainstream Media and is additionally linked with Wikipedia which provides additional context that may not be available on social media;

- automatically extracted information such as concepts (named entities and keywords) and Geographic Information;

- provides an importance filter for relevant events (mentioned above);

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

- Event Registry data is Open Data available to the general public via the graphical interface and to researchers via an API and open source tool[10].



WTO failure points to fragmented future for global trade

01 Aug 2014

Genève, Swiss Confederation

Society→Issues→Economic

In the short term, this is a setback for freer commerce. In the longer run, it means trade liberalisation may advance -- if at all -- among narrower groups of countries, denying dissenters a chance to block progress.

India has dealt a potentially fatal blow to the World Trade Organization's hopes of modernising the rules of global commerce and remaining the central forum for multilateral trade deals.

In the short term, this is a setback for freer commerce. In the longer run, it means trade liberalisation may advance -- if at all -- among narrower groups of countries, denying dissenters a...

The recent WTO setback need not mean failure, ICC writes in The New York Times #ICCWBO - googl.gl/alerts/rVgh

Interesting article on the essential end of glocalization. Failure of WTO + increased isolationism = regionalization on.ft.com/1u45CO

Figure 4: *Tweets matched to an event*

In the context of SYMPHONY, we are particularly interested in developing approaches that extend easily into different languages. The approaches detailed in this section are implemented to support the four languages for which we are collecting tweets: English, German, Italian and Slovenian. The source code for the software related to this task is available at https://github.com/lrei/er_match.

4.1 Matching Tweets to Events

When multiple news publishers report on the same event, their respective articles are clustered together [4]. Thus for each event in its database, Event Registry has the articles (title and body) that reported it as well as their respective metadata such as time of publication and URL.

Event Registry adds somewhere between 5000 and 40000 events to its database every day. Considering also the daily volume of tweets, techniques implemented should be computationally inexpensive.

4.2 URL Based Matching

In URL based matching we look for URLs in tweets and compare them to URLs in Event Registry. If a tweet contains a URL that matches the URL of an article in an event, we can say that the tweet is related to that article and thus to the event that contains the article.

[10] http://github.com/gregorleban/event-registry-python

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

This task is made slightly more challenging than simple string matching by the fact that the relationship between an article and a URL is often one-to-many. The most visible case is when URL shorteners are employed, where a different shorter domain is used in conjunction with a short code which then redirects to the longer URL[11]. Another case is when the URL for the article changes and the old URL redirects to a new one using the HTTP response status code 301 Moved Permanently [6]. It is also common for URLs to contain tracking query strings such as *http://www.example.org/1?utm_campaign=ex* or query strings that specify viewing options such as *http://www.example.org/1?page=1*. Furthermore publishing software often makes the same content available under different URLs based on the category structure, e.g. *http://example.com/politics/new-economic-policy/* and *http://example.com/economy/new-economic-policy/* . As a final example, several string level differences can be configured to be ignored by web server software to allow users to reach the right content even when they do not enter the exact string into their browsers, e.g. http://www.example.com/article can be the point to the same article as http://www.example.com/article/ .

To avoid being penalized in search engines rankings for content duplication, many publishers implement either HTTP redirection or the canonical link element [7]. The canonical link element commonly referred to as the canonical tag, is an HTML *<link>* element with the attribute *rel="canonical"* that can be inserted into the <head> section of an article (or any web page) e.g. <link rel="canonical" href="http://example.org/article/new-economic-policy/" />. It is also possible for the canonical link to be present in the HTTP headers instead of the HTML source. It is also common for publishers to implement the Open Graph Protocol [8] which allows for better integration with Facebook and requires a *<meta>* tag with the property *og:url* containing the article's canonical URL.

For each URL we make a request to it, taking not of any redirection, analyzing the headers and processing the HTML response body to obtain the canonical URL. The order of precedence is 1) canonical tag/header, 2) open graph *og:url* property 3) redirection 4) the original URL.

4.3 Content Based Matching

Intuitively (and experimentally) we know that not all tweets that refer to an event will include a URL to a news story about the event. Thus a different strategy is necessary to match those tweets to events. This is accomplished based on textual similarity between the tweet and events.

[11] For example, *http://alj.am/1OAQ9K9* directs the user who clicks on that link to *http://america.aljazeera.com/articles/2015/3/29/nashvilles-boom-pricing-out-middle-and-lower-class.html*.

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

We restrict content based matching of tweets to events within a time window: 3 days before and 3 days after the tweet has been published. Beginning with the oldest event within this window, once a tweet is matched to an event it is not tested against any more events. Each event within Event Registry is a cluster of news articles we select the medoid news article for each language in the event, i.e. the most representative news article for a given language, and use its text to compare to the tweet in that language. If the tweet does not have a language identification attribute we consider it to be the language of the country from which it was collected. If the event does not have a news article in the tweet's language, it is not considered for matching.

The problem of matching an article and a tweet is treated as a supervised binary classification problem where given an article and a tweet our classifier must answer if they 'match' or not. In binary classification, a classifier can be evaluated in terms of its precision and recall where

$$Precision = \frac{True\ Positives}{True\ Positves + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

In our case a *True Positive* corresponds to a tweet being correctly matched to an event, a *False Positive* corresponds to a tweet being incorrectly matched to an event and a False Negative corresponds to a tweet not being matched to its corresponding event. Thus, the numbers of True or False, Positives or Negatives always relate to tweet-event pairs that we have matched.

In the development of our classifier, precision was considered of paramount importance since showing a user a tweet in an event it was clearly not related too could potentially undermine confidence in the entire system. Improving precision can often require lowering recall. While this is certainly undesirable in our case, the cost of discarding tweets (i.e. false negatives) can be somewhat mitigated by redundancy which is common in social networks i.e. multiple users of a social network will create very similar content.

4.3.1 Text Preprocessing and Feature Generation

Text in articles and tweets is preprocessed similarly. Each document is preprocessed according to the following steps:

1. converted to lower case;
2. all URLs are removed;
3. all non-alphanumeric characters are removed (including punctuation and the hashtag symbol);

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

4. all characters are converted to their Unicode normal form [9];
5. the text is tokenized based on whitespaces;
6. stopwords are removed.

All tweets which after this preprocessing have less than 4 tokens are discarded. Once a document has been preprocessed, we generate its unigrams, bigrams trigrams and quadgrams i.e. its n-grams where *n ∈[1, 4]*. For news articles, the title and the body are processed separately.

For each article-tweet pair and for each *n ∈[4, 1]* we generate a similarity vector containing different measures of similarity between their ngrams:

1. the Jaccard similarity between the title of the article and the tweet;
2. the number of common terms between the tweet and the body of the article multiplied by logarithm of the number of terms in tweet [10];
3. the Jaccard similarity between the body of the article and the tweet;
4. the cosine similarity between the body of the article and the tweet.

4.3.2 Dataset Generation

In order to treat our problem as a supervised classification problem we must first create a supervised dataset. The URL matching described in 0 was used on historical data to create the positive examples dataset. The negative examples are generated by pairing tweets that have been URL matched to an event with a different event. We discarded any article-tweet pair with a 0 similarity vector in both positive and negative examples except for 1 in the negative examples. The number negative examples generated matches the number of positive examples used i.e. the dataset is balanced. The total number of examples the dataset we generated was 32372.

The dataset generation supports our goal of obtaining high precision in the sense that the classifier is trained almost exclusively with the hard cases: negative examples that share some similarity with the article. In practice these are actually an extremely small minority of all possible negative examples, since most tweets do not share any similarity with a given article. It also underlies the fact that by relying on simple text similarity between a tweet and a single article in an event ensures that many true positives are disregarded since they will also have a 0 similarity vector.

4.3.3 Tuning and Evaluation

We used a linear SVM as our binary classifier and then performed a random 50-50 split on the dataset into development and test subsets. Using precision as our scoring function, we performed parameter tuning using grid search with 5 fold cross evaluation on the development set for the penalty parameter (often called C) and class weight hyper-parameters, arriving at C=10 and a positive class weight 0.6 (negative class weight was kept fixed at 1). The positive class weight value multiplies C, since it is lower than 1 it allows the

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

SVM to learn a decision function that makes more misclassifications of positive examples. In particular, more false negatives.

Next we plotted the Precision-Recall curve (Figure 5) and the Precision-Recall vs Threshold curves (Figure 6).

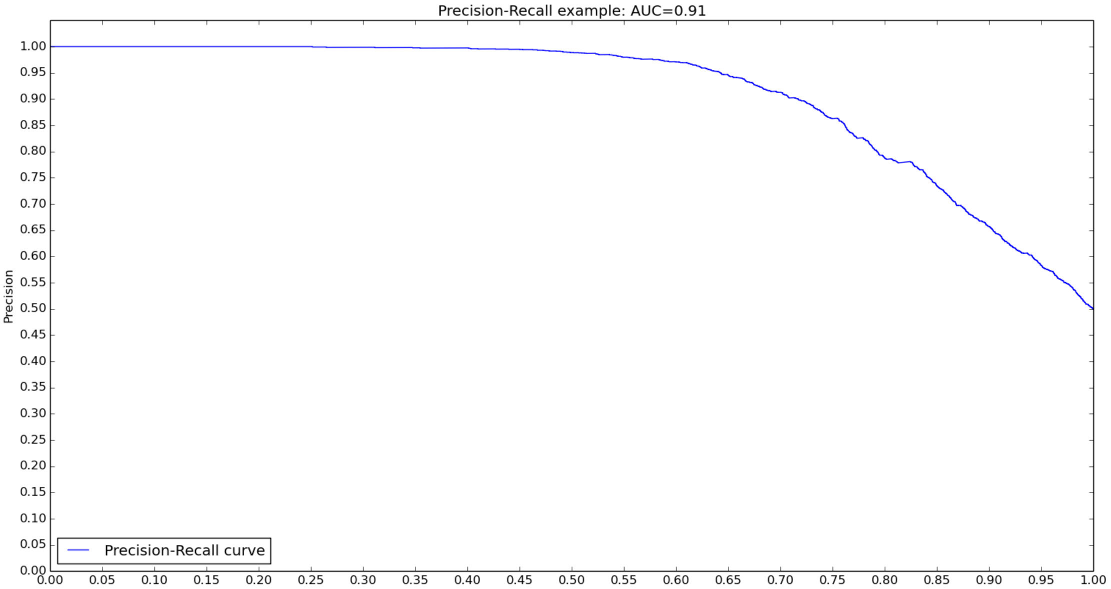


Figure 5: *Precision-Recall Curve*

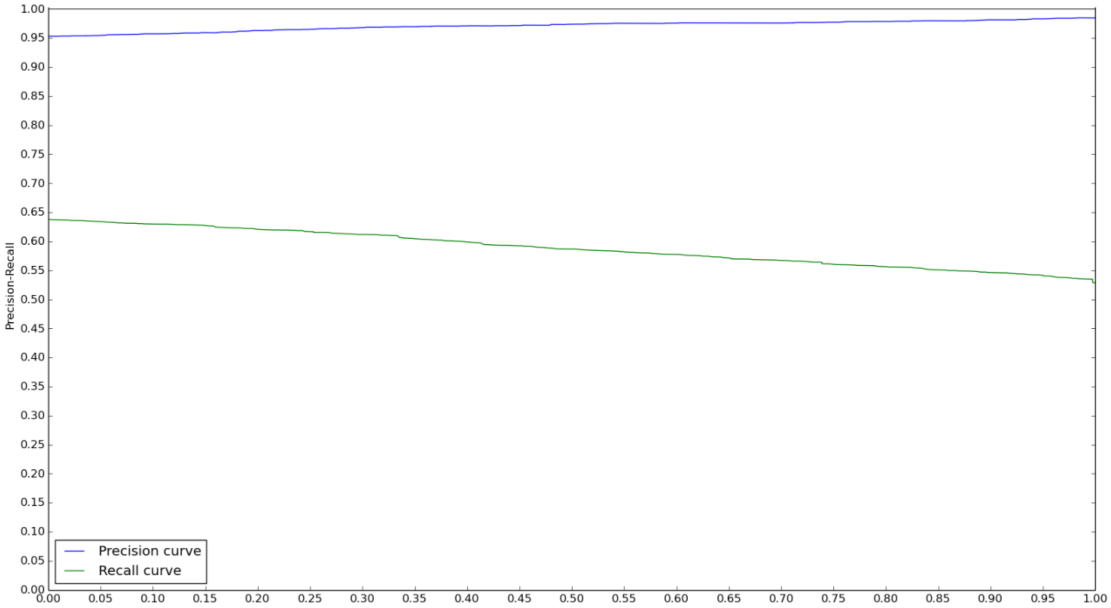


Figure 6: *Precision-Recall vs Threshold*

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

In particular, we can see that if we chose a threshold near 1, our classifier has nearly 100% precision while lowering our classifiers recall to nearly 55%. We consider this an acceptable trade-off in our case.

The procedure for generating our dataset introduces a huge bias into our evaluation: in reality, we will have many more false negatives since many true positives have 0 similarity vectors and will thus become false negatives (those case were discarded in our dataset). Real recall can be expected to be much lower than the recall on our test dataset.

The approach to content based event detection relies on 3 language specific parts:

- Whitespace based tokenization - which is expected to work equally well across all European languages;
- Stopword lists - which are available for all of the most widely spoken European languages and could be replaced, albeit probably with slightly worse results, with a automatically generated lists of the highest frequency words if expert generated lists were not available;
- Unicode normalization - which can yield worse results for certain languages certain languages where this normalization corresponds to a lesser degree (or not all) with the shortcuts people make when writing on social networks.

Future Work

Future work in this subtask will focus primarily on improving recall.

For URL based matching In case there are no redirects or any other indicators of a publisher provided canonical URL, it is still possible to follow the guidelines put forward for URL normalization [11] to reduce some of the issues in comparing URLs. This technique can however introduce false positives [12]. Common query parameters e.g. ?utm_campaingn could also be safely removed.

Associating a certain hashtag with a single event for a short period of time based on frequency in URL matched tweets could also further increase recall.

5 Cross lingual opinion and information diffusion

Bradley Greenberg's classic study of media contagion [13] from 1964 showed that 50% of the public learned about the Kennedy assassination via interpersonal ties. Since this news was of such importance, it is clear that everybody would get the information from media eventually, but the word-of-mouth had an important role in spreading the information faster [14]. Today, because of development of internet technologies and World Wide Web, advances in mobile devices and spread of applications based on social networks, the role of interpersonal ties in information spread is enormous. But not only had this kind of

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

information spread changed in the last 50 years, the technological advancements equally revolutionized the news media. Many classical print newspapers, newsmagazines, televisions and radio news broadcasters have online newspapers and blogs. Online electronic versions of news media can be instantly updated as soon as some event occurs. News agencies can distribute the information to subscribing news organizations much faster and news organizations can instantly share or copy information between them. Another factor important for information spread is interweaving of news media and social media, where a post on social media is a reaction on a news article, or a news article references social media. With such infrastructure of news and social media, the spread of information was never faster. In addition to providing news, benefits of advances in news and social media from a computer science perspective are in tractability of news articles and posts. Since we can detect time and location of large number of news articles and social media post and automatically determine a lot about the content, we can measure and analyze information diffusion.

Information diffusion is a research domain centered on developing techniques and models to capture, analyze and predict information diffusion in online social and information networks. Information diffusion tries to answer research questions like: which pieces of information or topics are popular and diffuse the most; how, why and through which paths information is diffusing and will be diffused in the future; and which members of the network play important roles in the spreading process [15].

As a part of this deliverable we use multilingual events and match social media posts with news media to capture both opinion and news diffusion. We capture structure of diffusion dynamics and compute geographical spread. In rest of this section of deliverable we describe method for analyzing information cascade dynamics and method for measuring geographical spread.

5.1 Information cascades

Here we discuss information spread and modelling of the spread as temporal graph. Figure 6 shows an illustrative example of an event and possible processes and interactions that spread the information about the event.

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

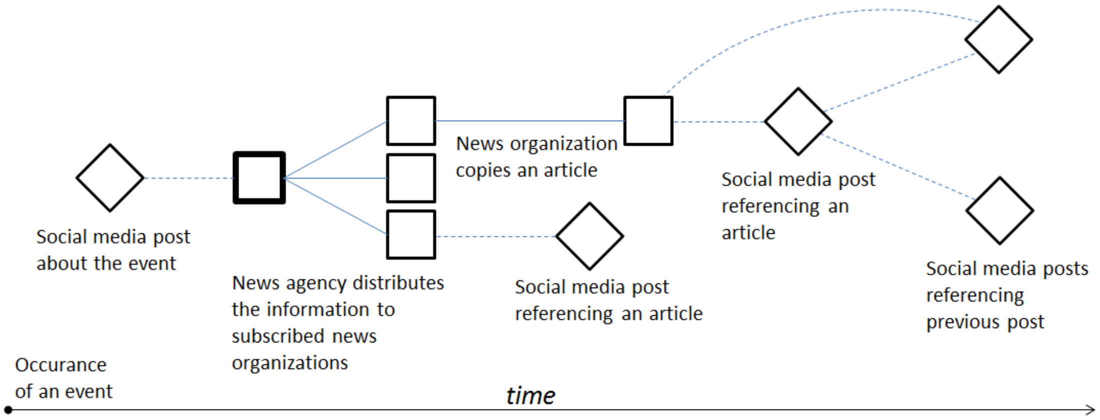

Social media post about the event
News agency distributes the information to subscribed news organizations
News organization copies an article
Social media post referencing an article
Social media post referencing an article
Social media posts referencing previous post
Occurance of an event
time

Figure 7: *Illustrative hypothetical example of information spread about an event.*

In the example we assume the first information about the event was published in social media by a user who witnessing the event first hand. Later, a news agency collects details about the event including the information from the first post, and distributes the information to several subscribed news media organizations. At this point in time the news about the event is quite spread and a new post in social media occurs. Next, a news organization copies an article about the event from another news organization. Finally, there is a social media post referencing a news article and later two more posts, one referencing another post and one referencing a post and an article. The illustrative example shows that information can travel between different types of sources in different order. For example, information can travel from social media post to a news article and vice versa. A source can use information from another source either referencing it, or not.

Observing how the information spreads we can conclude that duration of event, number of articles and posts, their arrival rate and geographical spread are important features of an event. In the next subsections we describe how we detect event and diffusion based on adaptable sliding time window that creates a temporal graph of articles and posts. We also describe method for measuring the geographical spread of events.

5.1.1 Event detection

Modelling diffusion as temporal graph using adaptable sliding time window is conducted on events that are reported in news articles and social media posts. This means that the first step in modelling diffusion is event detection. For this task we are reusing event detection infrastructure developed on X-Like [12]project and is available with Event Registry [13] [4] web

[12] X-Like: Cross-lingual Knowledge Extraction (Call for proposal: FP7–ICT–2011–7, Project reference: 288342)

[13] http://eventregistry.org/

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

service. Event detection pipeline is based on Newsfeed [14]service input data that contains more than 36 million articles (in the end of March 2015) in 40 languages and monitoring 75.000 RSS feeds, on average adds 72.000 articles added every day. Each new article undergoes pre-processing step that includes (a) detection and disambiguation of entities and topics, (b) identification of mentions of dates, (c) detection of mention of explicit event location, (d) identification of similar articles in other languages, and (e) detection of article duplicates. After pre-processing and extracting information, individual news articles are clustered into events using online multi-threaded clustering approach [16]. Since the event detection infrastructure handles cross-language processing, articles of the events can be in different languages, enabling cross-lingual information diffusion analysis.

In order to include social media and opinion diffusion analyses we developed a classifier that matches social media posts to events extracted from news media. This component was developed specifically for the needs of Symphony project and is described in detail in section 4.1 of this deliverable.

5.1.2 Generating temporal graphs using adaptable sliding window

After we have individual news articles or social media posts of a multilingual event, we model the diffusion by constructing a temporal graph for the event. Nodes of the graph are individual news articles, or social media post, that are connected with edges in case the nodes fall inside of the borders of sliding time window. To avoid fixed sizes of sliding windows for different events and for different stages of events, we developed a method for adapting the size of sliding time window using CMA (cumulative moving average). In this way the sliding window size adapts with the rate of article or post arrival, and connects the nodes that arrive in relatively fast rate and disconnects the nodes that arrive in relatively slow rate. Here relatively refers to difference between events (fast rate for one event can be slow rate for another event) and to different stages of an event (what was fast rate in the beginning of an event can be slow rate in the end of an event). The purpose generating temporal graphs that are based on node arrival rate is to be able to capture properties of diffusion dynamics. For example there can be events with news articles and posts arriving in constant rate and those that have several disconnected burst. Clearly the dynamics of these two types of events differ and we want to capture this by computing structural properties of temporal graphs.

Figure 2 shows a temporal graph constructed for an event called “Solar Financing Continues to Evolve as Manufacturer Upsolar Gets into Loans”. The event has 83 news articles that arrive in different points in time.

[14] http://newsfeed.ijs.si/

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

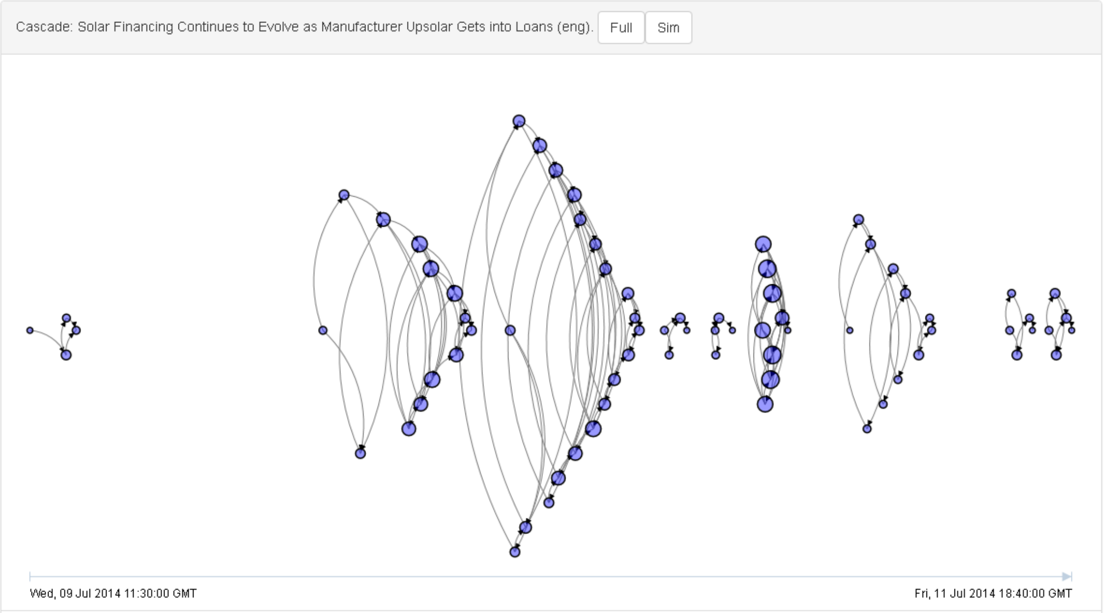

Cascade: Solar Financing Continues to Evolve as Manufacturer Upsolar Gets into Loans (eng).
Full
Sim
Wed, 09 Jul 2014 11:30:00 GMT
Fri, 11 Jul 2014 18:40:00 GMT

Figure 8: *Temporal graph of "Solar Financing Continues to Evolve as Manufacturer Upsolar Gets into Loans" based on adaptable sliding window.*

Nodes of the graph are news articles with horizontal position corresponds to the arrival time. The time difference between the first article of the event on leftmost position and the last article on the rightmost position is 2 days, 7 hours and 10 minutes. The edges of the graph are determined using sliding window adaptable with CMA as illustrated on Figure 9. On the x-axis of the graph in Figure 9 are labels of the temporal graph nodes and the y-axis holds time values in hours. Double red line on the Figure 9 shows sizes of sliding window with different number of nodes. For example when node with label 22 arrives, the size of the sliding window that corresponds to CMA is 1.22 h. Double red line on the graph shows delay of the node from preceding node. For example, node 22 arrives 45 minutes after node 21. Every time the full line is above the double line, the arrival of the new node took more time than it is covered with the sliding window starting from the previous node, which means that the new node gets disconnected from the rest of the nodes that arrived before it. If the full line matches or is below the red double line, the newly arrived node connects at least with the previous node. Looking at the temporal graph on Figure 2 we can see that the graph has several disconnected components and that components have different density. We compute these and other properties of the graphs to capture event diffusion dynamics.

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

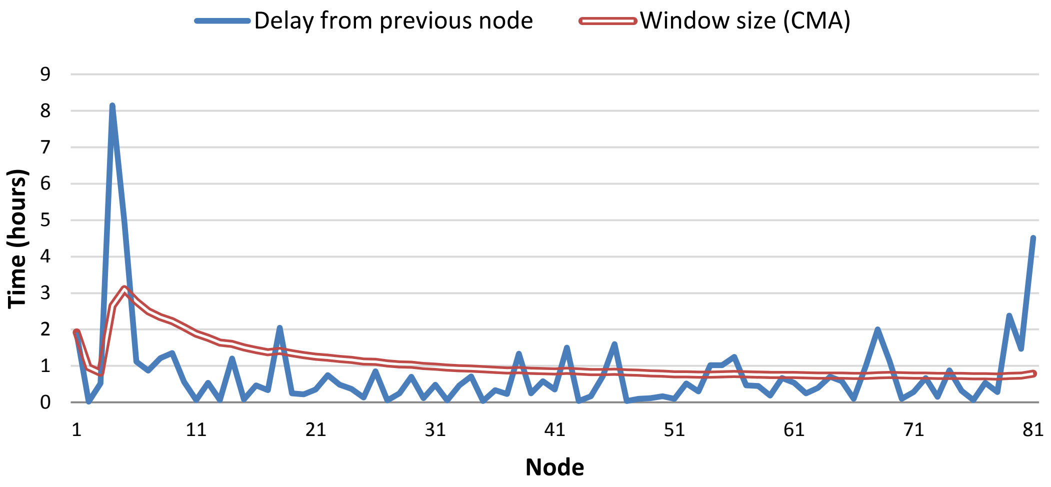


Figure 9: *Size of the sliding time window calculated as CMA (comulative moving average) for the event "Solar Financing Continues to Evolve as Manufacturer Upsolar Gets into Loans".*

5.1.3 Computing structural graph properties

The set of graph properties that will in further experimentation be used as diffusion dynamics features for nowcasting includes:

- **Number of nodes** is simply the frequency of articles or posts that correspond to the event. Bigger number of nodes can indicate more popular events.
- **Number of edges** shows how many pair of nodes got connected with our method with adaptable sliding window. Smaller number of nodes indicates that article or post arrival rate was mostly decreasing with time. A borderline case where no edges would be created is if the articles or posts would arrive with continuously increasing delay. Bigger number of edges indicates that the nodes were arriving in intensive burst or that the arrival rate was increasing with time, what could mean that event was increasingly more popular.
- **Duration** is simply the time difference between first and last node in the graph. Longer duration can mean that the event is more important.
- **Node per minute** is number of nodes normalized with time. Higher nodes per minute rate can means that arrival of news or post was more intensive.
- **Graph density** is the number of existing edges in the graph, divided by number of possible edges in the graph. For undirected simple graph density is calculated with the following formula:

$$density = \frac{E}{\frac{N^2 - N}{2}}$$

where N is number of nodes and E is number of edges.

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

- **Clustering coefficient** of a graph is a measure that shows to which degree nodes in a graph tend to cluster into tightly knit group [16]. It is real number between zero and one, with zero when there is no clustering, and one for maximal clustering, which happens when the network consists of disjoint cliques. Clustering coefficient can be calculated as ratio between number of closed triplets and number of connected triplets of nodes. In our model clustering coefficient indicates similar node arrival dynamics as graph density. If the arrival rate is stable throughout the event, the clustering coefficient will be low. If the arrival rate changes and nodes arrive in bursts, the clustering coefficient will be high.
- **Number of connected components** tells us how many connected subgraphs exist in the graph. Larger number of connected components can indicate that there where many longer pauses between burst of news articles or posts. On the other hand a single connected component would occur if the arrival rate was very uniform.
- **Size/share of the largest connected component** tells us what is the number and share of nodes that are contained in the largest connected component. It can happen that our temporal graph has many connected components, but most of the nodes can be concentrated in only one component. This situation would happen in case most of articles or posts arrived in a single isolated burst.
- We call a period from the earliest to the latest node of a connected component a chain and calculate **sum and share of all chain durations**. In this we capture the empty space between connected components when there were no articles or social media posts.

5.1.4 Examples of event dynamics

Here we show examples of temporal graphs and their properties (Table 1) on four events. Graph of the first event (Figure 10) has 17 connected components which are not densely connected. Regular, but not intensive bursts result with relatively high clustering coefficient and lower density.

Table 2. Properties of temporal graphs

Event	Nodes	Edges	Connected components	Clustering coefficient	Density
Solar Financing Continues to Evolve as Manufacturer Upsolar Gets into Loans	83	132	17	0.579	0.0387
Jordan To Commission 1800 MW	59	62	22	0.412	0.0362

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

Renewable Energy Capacity By 2018					
US economy grew at 3.9 percent rate in 3rd quarter	131	827	31	0.648	0.0971
Rise in renewable energy being facilitated by developing nations	28	17	15	0.25	0.04497

The temporal graph of the second event (Figure 11) has similar properties as the first event. The graph of second event has 22 connected components, from which two connected components have are larger than others. The two larger components are not dense and the other components are relatively small, what results with lower density and clustering coefficient of the graph.

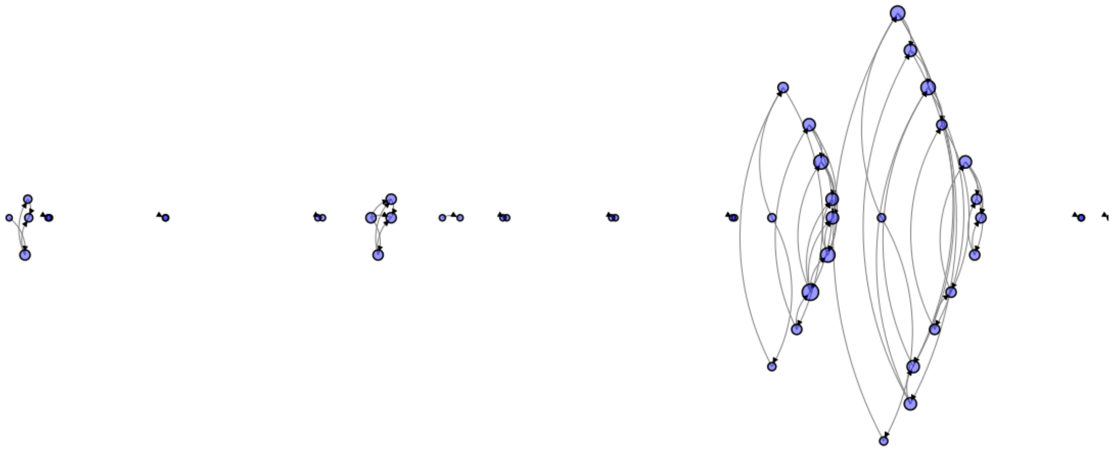


Figure 10: *Temporal graph of "Jordan To Commission 1800 MW Renewable Energy Capacity By 2018" based on adaptable sliding window.*

The graph of the third event (Figure 11) has much higher density, which is caused by intensive bursts towards the middle of the event. The intensive burst resulted with large and dense connected components. This graph has higher density and clustering coefficient from the other examples.

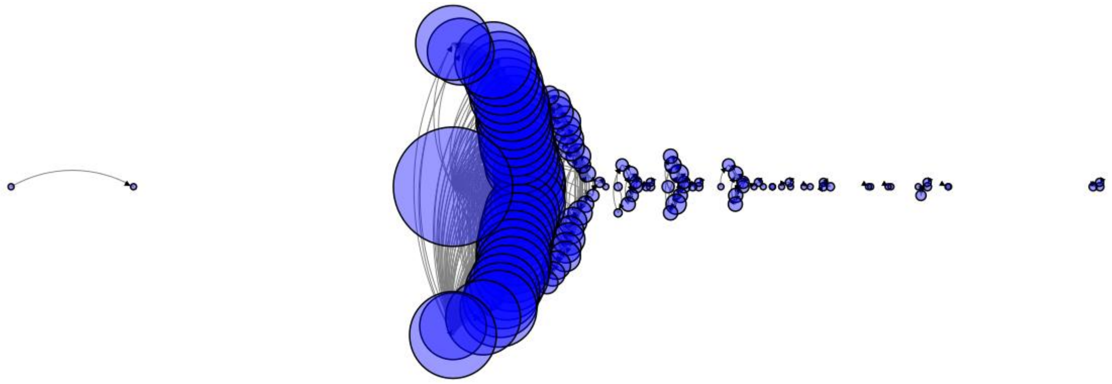


Figure 11: *Temporal graph of " US economy grew at 3.9 percent rate in 3rd quarter" based on adaptable sliding window.*

Temporal graph of the fourth event is characterized with small number of nodes and edges. The event has 15 small connected components. Since this event did not have intensive bursts, clustering coefficient of the graph is very low.

Figure 12: *Temporal graph of "Rise in renewable energy being facilitated by developing nations" based on adaptable sliding window.*

5.2 Geographical spread

An event can have different importance in different parts of World. For example, two events can have the same number of news articles, one being very global with articles distributed across the world, and the other one being very local with articles concentrated in one smaller area. Here we measure the geographical spread of news articles and social media posts that report about events. This is possible because both news articles and social media posts have assigned geographical information of publishing.

5.2.1 Method for determining geographical spread

The method for determining geographical spread can be divided into the following steps:

- Creating Euclidean graph with node positions determined with geographical coordinates
- Creating a complete graph with edges corresponding to haversine distances
- Creating Delaunay triangulation to reduce the density of the graph
- Computing minimum spanning tree of Delaunay triangulation

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

- Sum of edge weights of the MST corresponds to geographical spread of event

Geographical spread is measured by creating a Euclidian graph, where positions of nodes correspond to the geographical coordinates of publishing locations. Since we are using geographical latitude and longitude, we compute the distances between the nodes using haversine formula. Haversine formula for calculating distance between two points defined with latitude and longitude is based on spherical model of earth. Since the earth is slightly ellipsoidal using a spherical model gives errors typically up to 0.3%, but this is acceptable for our purpose. Distance between two points is calculated as:

$$a = \sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right), c = 2 \cdot \tan^{-1} 2\left(\sqrt{a}, \sqrt{(1-a)}\right), d = R \cdot c,$$

where φ is latitude, λ is longitude, R is Earth's radius (mean radius of 6,371km was used) and d is the distance between two points.

In order to compute the geographical spread we need to sum the geographical distances between the nodes, so the key questions is how to connect nodes of the graph. We do this by computing minimum spanning tree of the graph. Minimum spanning tree is a subgraph of the graph that connects all the nodes of the graph in such a way that the sum of the edge distances is minimal. The Euclidian MST problem is different from the typical graph problem because all the edges are implicitly defined. If we construct a complete graph with N nodes and $N(N-1)/2$ edges, we can use Prim's algorithm for finding MST in time proportional to N^2 [17]. This solution is generally to slow so we use the fact that a graph known as Delaunay triangulation contains the MST by definition and has number of edges proportional to N. Delaunay triangulation can be computed in time proportional to $N\ log\ N$ and greatly speeds up the MST finding which we conduct using Prim's algorithm [18].

5.2.2 Examples of geographical spread

Here we illustrate geographical spread calculated on four different events. We compare number of news articles that appear in events, duration of events and geographical spread.

Table 3. Geographical spread of events.

Event	Nodes	Duration (hours)	Spread (km)
Obama orders sanctions after Russia takeover	26	20	29 626
Greece presents its debt plan to skeptical Eurozone	500	900	69 713
Paris shootings: Prime suspect previously jailed for terror	2740	179	107 557
Berlin: Energieverbrauch in Deutschland fällt[15]	17	23	1 540

[15] eng. "Berlin: Energy comsumption in Germany falls"

The event called “Obama orders sanctions after Russia takeover” has 26 nodes duration of almost one day and geographical spread of more than 29 thousand kilometers (Figure 13). The event does not have very high frequency and spreads mostly in USA, but the geographical spread is high because it includes few nodes in Asia and one in South Africa. High geographical spread indicates that this event is globally important, event thou just the frequency would maybe not indicate this.

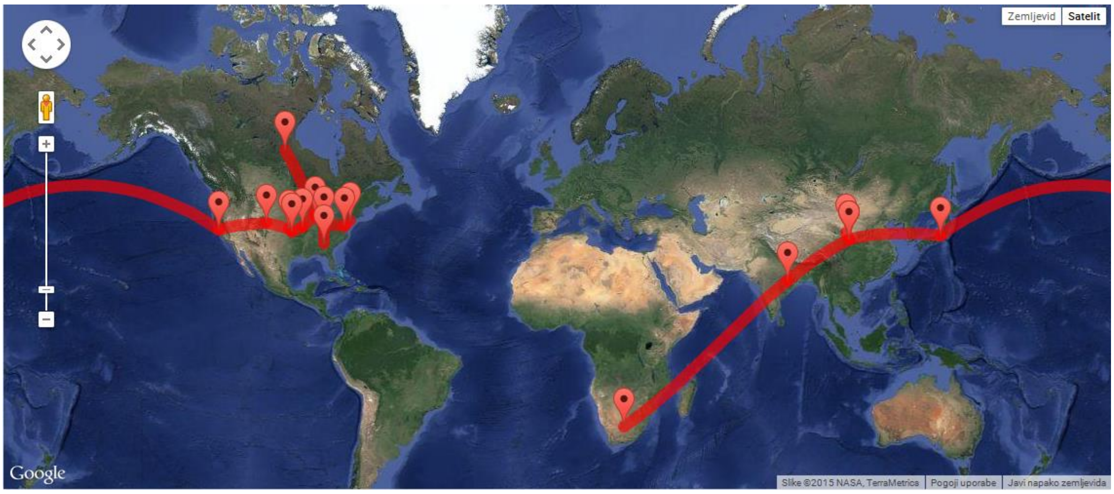


Figure 13*: Geographical spread for "Obama orders sanctions after Russia takeover" event.*

The second event (Figure 14) is about Greek bailout proposal which is very important for the Eurozone, but also for United States and other Europe trading partners. Global importance of this event is indicated by the geographical spread of almost 70 thousand kilometers.

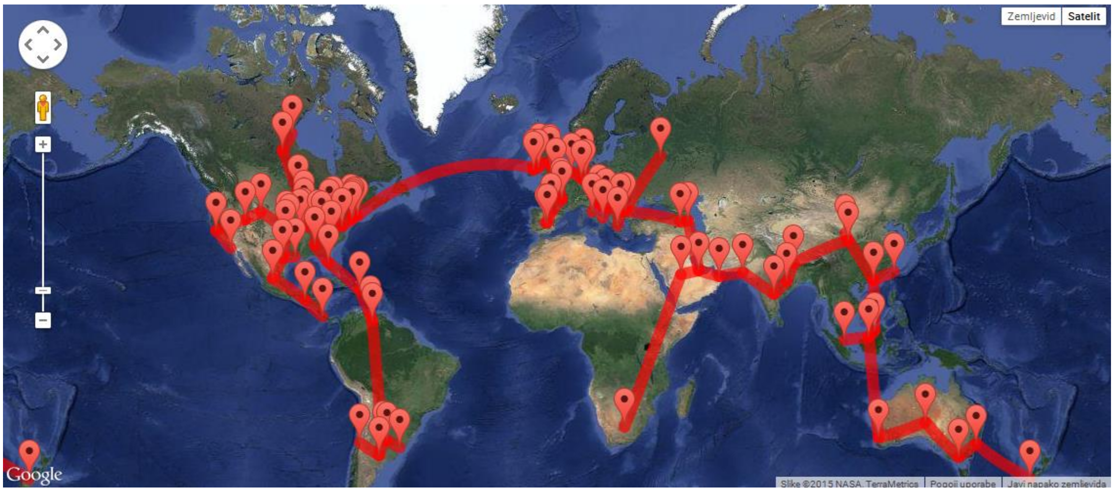


Figure 14: Geographical spread for “Greece presents its debt plan to skeptical Eurozone” event.

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

Third event is about terrorist shooting in Paris (Figure 15) and has much higher number of events (2 740), which create a spanning tree with spread of 107 thousand kilometers. This event is and extreme in both number of events and in its spread, which could mean that the event occurred unexpectedly and was very important for the whole world. The first two events are also globally important, but are parts of many related events that are developing for longer period of time. The third event was sudden and unexpected, what can be an explanation for the large frequency and geographical spread. Paris shooting event is a good example of sudden, shocking and globally important event that is big news around the globe.

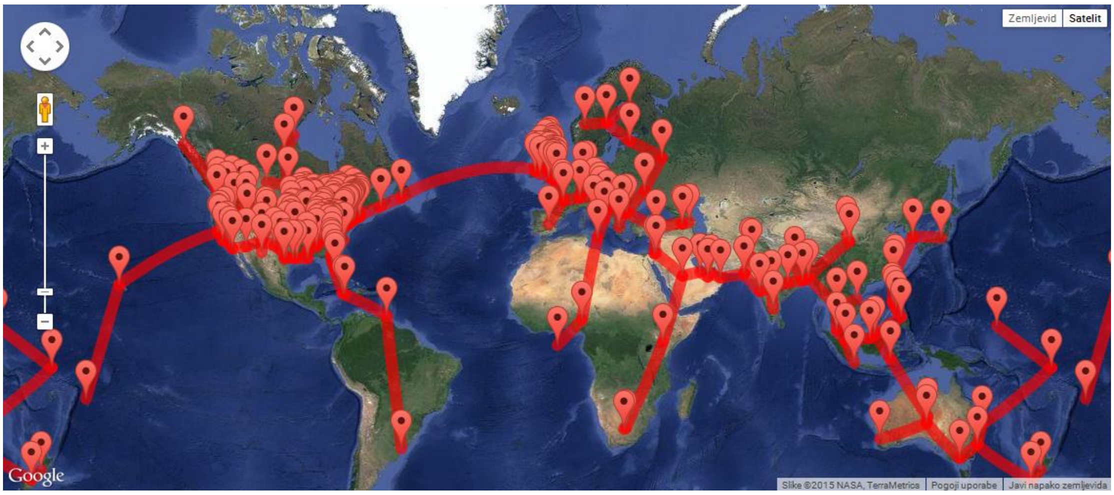


Figure 15: *Geographical spread for "Paris shootings: Prime suspect previously jailed for terror" event.*

The fourth event (Figure 16) “Berlin: Energieverbrauch in Deutschland fällt” (eng. Berlin: Energy consumption in Germany falls), has 17 nodes and geographical spread of only 1.5 thousand kilometers. This event has about twice as less nodes than the first two events, but its geographical spread is much smaller. This indicates that the event is important for smaller geographical scope, which is in this case national level (Germany).

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

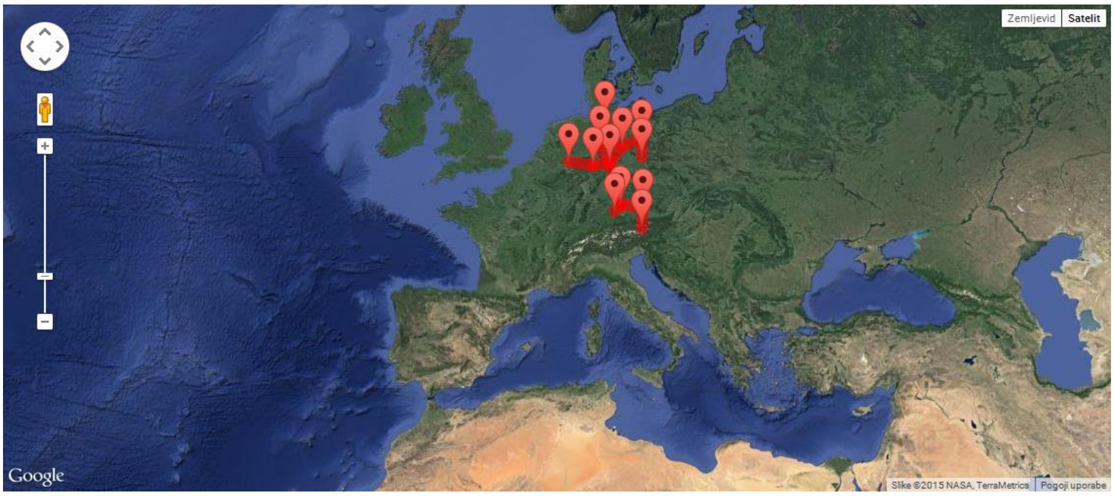


Figure 16: *Geographical spread for "Berlin: Energieverbrauch in Deutschland fällt auf niedrigsten Stand seit 1990" event.*

The four examples of events show that geographical spread gives information related to scope of importance of events.

6 Sentiment Classification

On a personal level, people have traditionally considered "what other people think" relevant. For decision makers, especially elected officials, finding out what people or particular subgroups of people (e.g. a politician's electorate) think about certain particular issues or policies is a priority and can have a very strong influence on the decision making process. Traditionally, only techniques such as telephone or mail surveys were available. The proliferation of user generated content such as blogs, online forums and social media on the internet has changed this. While automatic sentiment analysis can aid traditional techniques such as the analysis of responses to mail based political questionnaires, it is actually plausible that it could replace them entirely in certain cases or significantly supplement them [19].

Within SYMPHONY, we plan to use sentiment polarity and frequency as a high level signal for correlation analysis and nowcasting. There is evidence that supports the hypothesis that sentiment is relevant to this task, for example, the fact that financial markets are affected by the social mood was documented in [20].

We also plan to make an automatic polling tool available through the SYMPHONY dashboard (WP5 Task 5.4) and include the social media sentiment based sentiment visualizations available to the public through Event Registry. Finally we plan to explore integration with other SYMPHONY tools such as using sentiment polarity and term frequency to initialize certain ABM variables (e.g. confidence in the European Central Bank).

Message level sentiment (polarity) classification consists of a assigning to an entire message (e.g. a tweet) a polarity label: *Positive*, *Neutral* or *Negative*. Different authors and datasets provide different guidelines for labelling messages as *Neutral*. Since we're concerned with sentiment analysis on twitter, we chose to adopt the guidelines used in the Semeval-2013: Sentiment Analysis in Twitter shared task[16]. The dataset was described in [21].

For this task we implemented the *Paragraph Vector* based sentiment classifier described in [22]. This classifier yielded state of the art results without requiring any other tools such as lexicons, part-of-speech classifiers, twitter text normalizations, etc, making it ideal for the multi-lingual requirements of SYMPHONY and the noisy data of social media. This sentiment classifier works by generating a vector for each new document and then feeding it to a previously trained logistic regression classifier. A *Paragraph Vector* is a fixed length representation of a document such as part of a sentence, a tweet or a blog post. In this sense, it is similar to the bag-of-words method. It however overcomes the two major weaknesses of the latter: the loss of word order (a "bag" or multiset is unordered by definition) and the fact that ignores word semantics e.g. the words "king", "queen" and "cheese" are equally different or *distant,* ignoring the semantic similarity between the first two words.

The source code for the software related to this task is available at https://github.com/lrei/nlk.

6.1 Generating Paragraph Vectors

Generally speaking, a *Paragraph Vector* algorithm could be created from any Neural Network language models algorithm that learns a vector representation for words in a similar manner to the Neural Probabilistic Language Model [22]. That is, a model that learns word vectors while predicting a missing word (or sequence of words) given a context. Typically this has been a model that learns to predict the word at position *T* in the sequence from the previous *T-1* words in that sequence i.e. the typical formulation of a language model where the objective is to maximize

$$\frac{1}{T}\sum_{t=k}^{T-k} \log(w_t | w_{t-1}, \dots, w_{t+k})$$

A simple algorithm for learning such a language model is depicted in Figure 17. Every word is mapped (or projected) to a unique vector, represented by a column in a matrix *W*. The linear layer represented by *W* is commonly called a lookup table, where the vector representation of the word is indexed by its position in a vocabulary. The concatenation, sum or average of

[16] http://www.cs.york.ac.uk/semeval-2013/task2/

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

the vectors that form the context is then used as the feature to predict the next word in the sequence by feeding it into a softmax layer. Errors are back propagated to *W*, generating the word vectors.

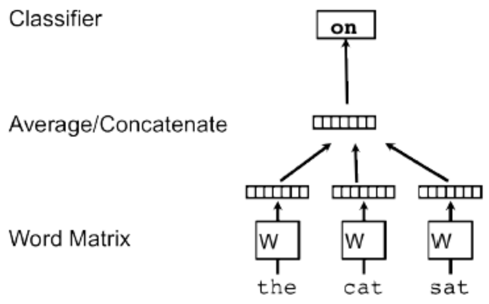


Figure 17: *Learning Word Vectors* [23]

In [23], the algorithms presented, Distributed Memory Model of Paragraph Vectors (PV-DM) and Distributed Bag of Words Model of Paragraph Vectors (PV-DBOW) (Figure 18) are based on Continuous Bag-of-Words (CBOW) and Skip-gram models [24] (Figure 19), respectively.

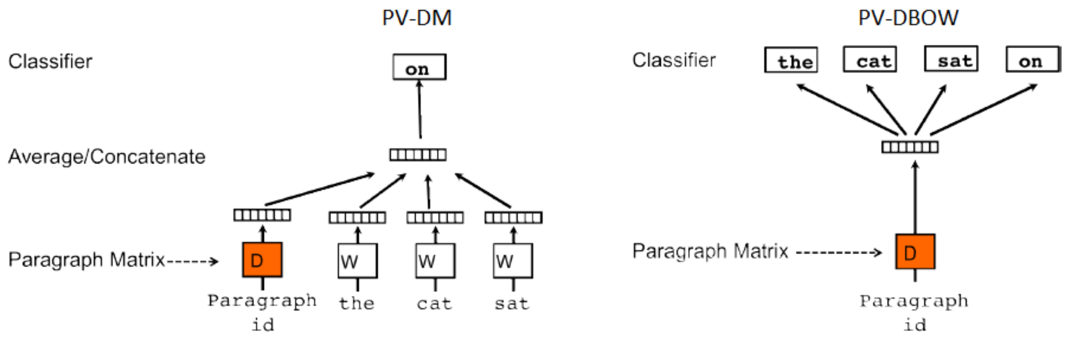


Figure 18: *PV-DM and PV-DBOW models* [23]

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

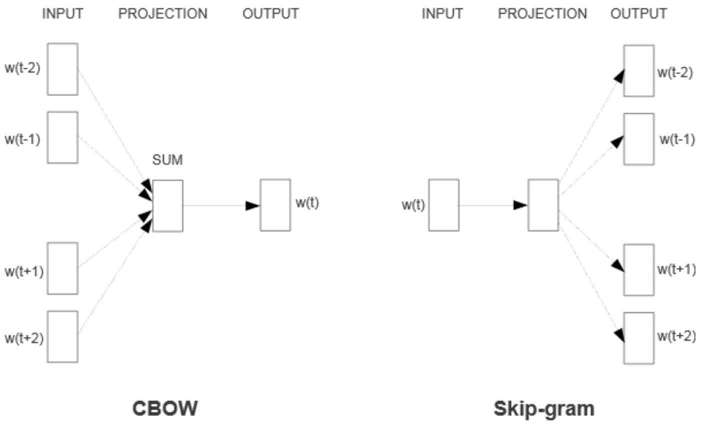


Figure 19: *CBOW and Skip-gram models* [24]

In PV-DM, the *Paragraph Vector* can be thought of as another word that is always present in the context for a given document, representing what is missing from it. In PV-DBOW, the *Paragraph Vector* can be seen as the sole feature for predicting all of the documents contexts.

Training consists of learning the word and paragraph vectors are learned simultaneously by stochastic gradient descent and backpropagation. This process updates the parameters for softmax as well as *W* and *D,* where *D* is the equivalent of *W* for paragraphs (a paragraph lookup table). To perform prediction after the model is trained, the algorithm is the same except only *D* is updated while the remaining parameters remain fixed. This allows *Paragraph Vectors* to be learned (generated) for new, unseen, documents. Once this vectors are generated, they can be used as features (representations) of their respective documents, similarly to how a bag-of-words representation would be used. A document can be represented by a *Paragraph Vector* created by concatenating the document’s PV-DM and PV-DBOW vectors.

6.2 Sentiment over Paragraph Vectors

By transforming a *Paragraph Vector* for each sentence (or message) in a supervised sentiment polarity corpus, the problem of sentence (or message) level sentiment can be seen as a classification problem over *Paragraph Vectors* and a standard classifier, such as logistic regression, can be used to learn sentiment classification using a *Paragraph Vector* as input (Figure 20).

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

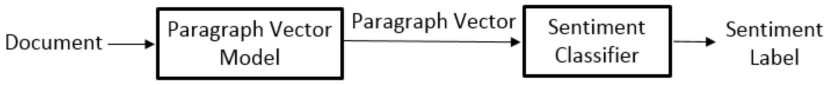


Figure 20: *Sentiment Classification over Paragraph Vectors*

We followed the same experimental protocol described in [23] for sentiment classification on the Stanford sentiment treebank dataset[17], in order to validate our implementation.

6.3 Future Work

The final version of the described sentiment classifier will be provided in Deliverable **D2.4 Final version of social media based policy indicators**.

6.3.1 Dataset

Sentiment classification as a supervised classification problem requires labeled training data. While many standard corpus are publicly available for English, including the SemEval-2013 Task 2 corpus previously mentioned, the situation is not the same for other languages. Because we needed to cover German, Italian and Slovenian and wanted in-domain (i.e. tweet) corpus we opted to create it using the technique described in [25] based on the procedure previously described in [26] and [27] for each language:

Collect tweets with happy emoticons such as “:-)”, “:)”, “=)”, “:D” to use as positive polarity examples;

Collect tweets with sad emoticons such as “:-(”, “:(”, “=(”, “;(” to use as negative polarity examples;

Collect tweets from news and magazine publishers to serve as neutral/objective polarity examples.

This approach has the benefit of extending naturally to any language.

6.3.2 Evaluation

While the dataset described is good enough to learn, rigorous evaluation should be made using a human expert annotated dataset. We plan to use the SemEval-2013 dataset previously made to evaluate English language results and create small evaluation datasets for the other languages.

[17] http://nlp.stanford.edu/sentiment/

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

7 Correlate and Nowcast

Governmental institutions and international organizations such as the International Monetary Fund and the World Bank periodically release economic indicators. However, these are typically made with very low frequency (e.g. weekly or monthly) and a large lag (i.e. the data made available provides for the previous week or month). It would be helpful to have more timely forecasts of these economic indicators. In Nowcasting, we want to predict the present, specifically, the present value of an indicator which has not been released. This same goal was attempted previously by Google Trends with some success [28] using search queries entered into google search and thus it seems like a natural starting point for our efforts within SYMPHONY.

While in Google Trends the signal is simply an aggregate number of queries grouped by a category, we can make use of many different signals, all of which can be represented as a time series of values:

- The number of mentions in mainstream media for a concept or aggregate category of concepts (**Errore. L'origine riferimento non è stata trovata.**);
- The number of social media messages matched to a concept or aggregate category of concepts (Section **Errore. L'origine riferimento non è stata trovata.**);
- Diffusion metrics, such as geographical spread, applied to mainstream and social media (Section **Errore. L'origine riferimento non è stata trovata.**);
- A ratio between positive and negative social media matched to a concept or aggregate category of concepts (Section **Errore. L'origine riferimento non è stata trovata.**)

Our goal is identification of which among all these millions of signals be used to predict the present value (nowcast) of any given time series. In order to do this, we need to perform 2 steps for any given time series (e.g. monthly unemployment numbers in Italy):

1. Identify which signals show potential predictive power;
2. Test if they have predictive power.

To identify which signals show potential predictive power for a given time series, among the millions of possibilities, we compute an approximate correlation between the given time series and all the possible signals using the Google Correlate algorithm [29].

At this point we can take the highest correlated signals and attempt to show the predictive power for each of them individually using an approach based on [28]:

1. Create an autoregressive model for the given variable: let y_t be the ***log*** of the variable at time t, than the model $y_t = \sum_{i=1}^{p} b_i y_{t-i} + \varepsilon_t$ where p controls the

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

number of lagged variables, b represents the parameters of the model and ε_t is white noise. This model serves as a baseline;

2. For each potential signal, create a model where it is added to the regression;
3. Use a rolling window forecast where the models are trained first with 70% of the data and then tested on the remaining 30%.
4. Calculate Mean Absolute Error (MAE) for all models tested: models which show lower MAE than the baseline and their respective signals are selected as predictors.

The prototype built for this work and the features described in this section can be accessed at http://beta.eventregistry.org/correlate/.

7.1 Data Preparation

A naïve approach in computing top correlations with a given data curve would be to directly compute correlation of the data curve with each of the concepts. Since the number of concepts is approximately 1.8 million and each concept has several signals associated with it and potentially hundreds of data point per signal (e.g. a daily time series could have 500 data points going back to the start of ER), the computations would take too long.

In order to be able to efficiently calculate top correlations we first have to prepare the training data. First we normalize the time series for each individual concept. The normalization that we perform includes subtracting the mean value and dividing by the standard deviation. The resulting time series therefore has mean value 0 and standard deviation 1. In the next step we divide the time series into smaller chunks. Each chunk contains a consecutive number of intervals (e.g. days). In our experiments we used chunks of size 7, although we could also use a larger value. For each chunk we then perform the following. All time series are clustered based on their similarity. We use k-means clustering with k=256. In this way, each chunk for a concept can be described with one byte that represents the id of the cluster. For each cluster we also compute a centroid as the mean value of all members of the cluster.

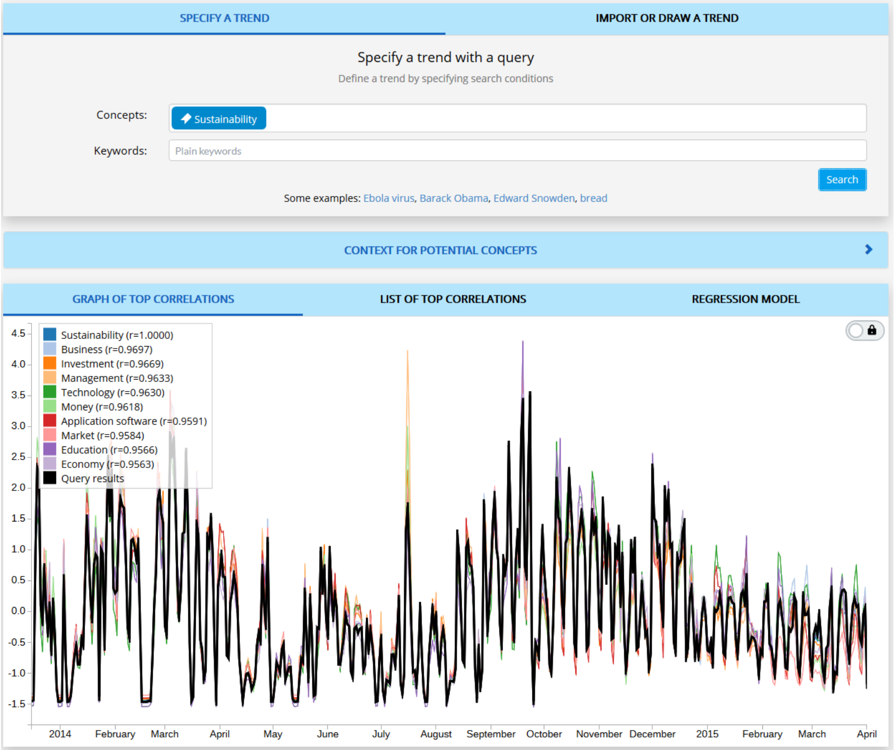
7.2 Identifying top correlations

When a testing time series is provided we do not compare it to the time series of individual concepts. Instead we first normalize it in the same way as we normalized the concepts – by removing the mean value and dividing by the standard deviation. Next, we split it into the same chunks as did for concepts. For each data chunk we compare the distance of the test curve to each of the cluster centroids. By knowing the membership of the concepts in the clusters we can compute for each concept the approximate distance of its curve to the data curve. The distance is only approximated since we are using the curves of the cluster centroids instead of the actual concept curve. After computing the approximate distances for all concepts we can sort the concepts by increasing distance. For first N concepts we can then compute an exact correlation of the concept curve with the provided time series. If we wish to provide the user with top m correlations we set $N = 5 * m$. Our experiments show

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

that despite using the approximations, the expanded window of *5*m* concepts is sufficient for obtaining almost 100% recall in identifying top *m* concepts that correlate the most with the provided time series.

Correlation between a given time series and the signals, between different signals or between a query and the signals can already provide valuable information to a user (e.g. a policymaker, a researcher or a journalist). We provide access to this via a user interface on ER show in Figure 21 and Figure 22.



Specify a trend with a query

Define a trend by specifying search conditions

Concepts: Sustainability

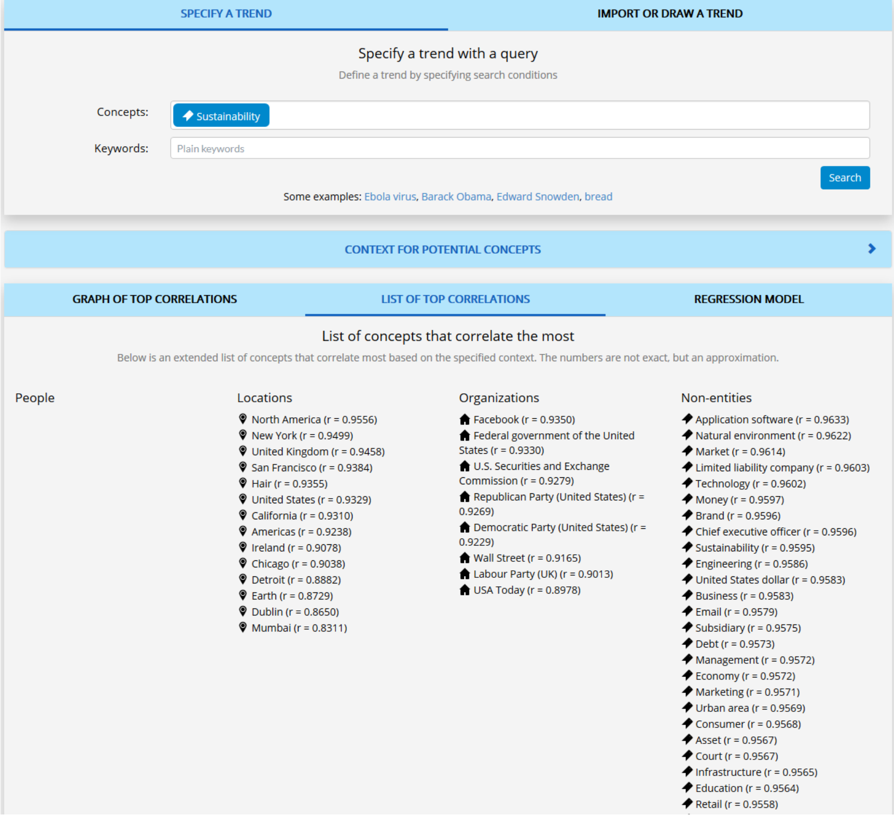
Keywords: Plain keywords

Some examples: Ebola virus, Barack Obama, Edward Snowden, bread

🟦 Sustainability (r=1.0000)
🟦 Business (r=0.9697)
🟧 Investment (r=0.9669)
🟧 Management (r=0.9633)
🟩 Technology (r=0.9630)
🟩 Money (r=0.9618)
🟥 Application software (r=0.9591)
🟥 Market (r=0.9584)
🟪 Education (r=0.9566)
🟪 Economy (r=0.9563)
⬛ Query results

Figure 21: *Correlate in Event Registry - curves of top concepts that correlate with the concept "Sustainability"*

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688



Specify a trend with a query

Define a trend by specifying search conditions

Concepts: Sustainability

Keywords: Plain keywords

Some examples: Ebola virus, Barack Obama, Edward Snowden, bread

List of concepts that correlate the most

Below is an extended list of concepts that correlate most based on the specified context. The numbers are not exact, but an approximation.

People

Locations

- North America (r = 0.9556)
- New York (r = 0.9499)
- United Kingdom (r = 0.9458)
- San Francisco (r = 0.9384)
- Hair (r = 0.9355)
- United States (r = 0.9329)
- California (r = 0.9310)
- Americas (r = 0.9238)
- Ireland (r = 0.9078)
- Chicago (r = 0.9038)
- Detroit (r = 0.8882)
- Earth (r = 0.8729)
- Dublin (r = 0.8650)
- Mumbai (r = 0.8311)

Organizations

- Facebook (r = 0.9350)
- Federal government of the United States (r = 0.9330)
- U.S. Securities and Exchange Commission (r = 0.9279)
- Republican Party (United States) (r = 0.9269)
- Democratic Party (United States) (r = 0.9229)
- Wall Street (r = 0.9165)
- Labour Party (UK) (r = 0.9013)
- USA Today (r = 0.8978)

Non-entities

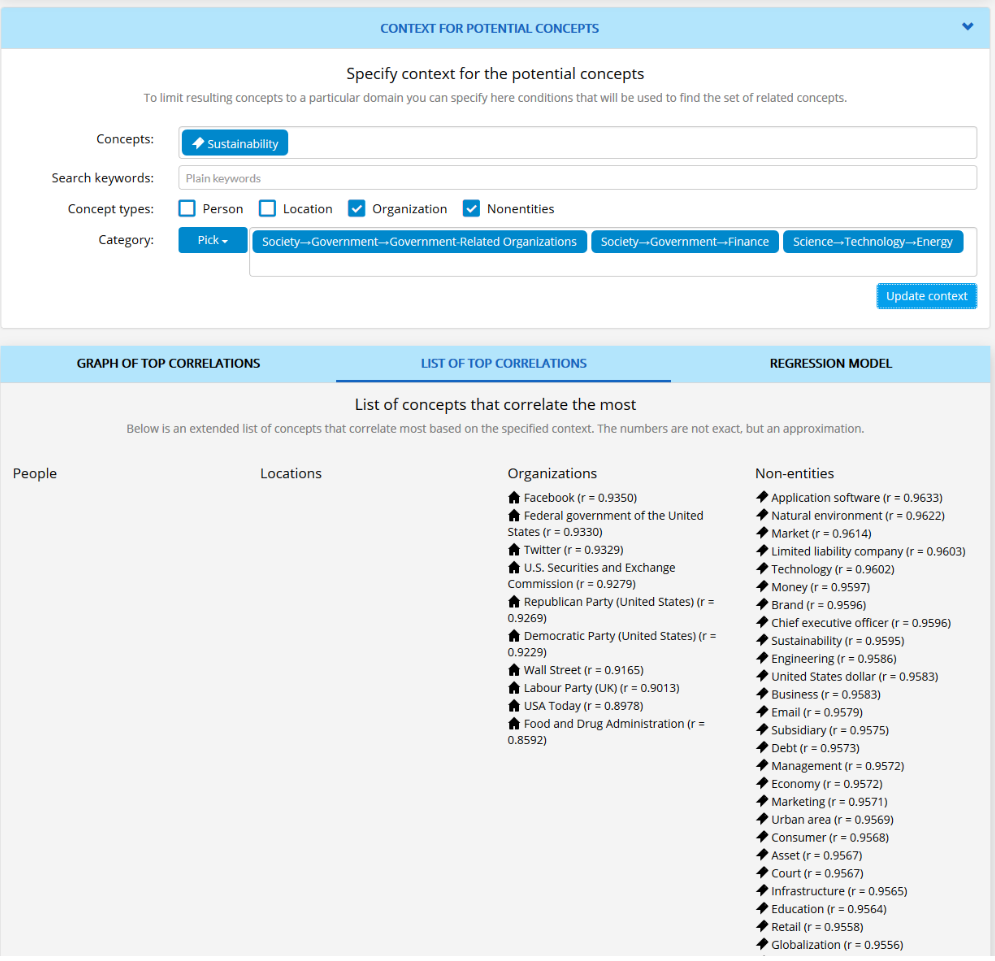
- Application software (r = 0.9633)
- Natural environment (r = 0.9622)
- Market (r = 0.9614)
- Limited liability company (r = 0.9603)
- Technology (r = 0.9602)
- Money (r = 0.9597)
- Brand (r = 0.9596)
- Chief executive officer (r = 0.9596)
- Sustainability (r = 0.9595)
- Engineering (r = 0.9586)
- United States dollar (r = 0.9583)
- Business (r = 0.9583)
- Email (r = 0.9579)
- Subsidiary (r = 0.9575)
- Debt (r = 0.9573)
- Management (r = 0.9572)
- Economy (r = 0.9572)
- Marketing (r = 0.9571)
- Urban area (r = 0.9569)
- Consumer (r = 0.9568)
- Asset (r = 0.9567)
- Court (r = 0.9567)
- Infrastructure (r = 0.9565)
- Education (r = 0.9564)
- Retail (r = 0.9558)

Figure 22: *Correlate in Event Registry - list of all top concepts correlated with the concept "Sustainability"*

7.3 Avoiding spurious correlations

Since the space of possible concepts is huge it is common that we identify in the data some correlations that are unexpected and likely invalid. An example is the time series of Barack Obama that highly correlates with concepts “Music” (r=0.845) and “Rape” (r=0.844). In order to avoid these unlikely concepts we implemented a way for providing the meaningful context. In the interface, the user is able to provide one or more search conditions that determine the domain of possible concepts and thus, their associated signals. Using the search conditions we perform a search for articles in the Event Registry. From the articles that match the criteria we extract the concepts mentioned in the articles. By aggregating over all matching articles we can obtain the list of top N concepts mentioned in these articles. N can be larger or smaller, depending on the size of the context that we wish to use. In our experiments we use N=10,000. Only the 10,000 concepts that are most related to the provided context are therefore used as the candidate concepts on which we compute the correlations. This can be seen in Figure 23.

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688



Specify context for the potential concepts

To limit resulting concepts to a particular domain you can specify here conditions that will be used to find the set of related concepts.

Concepts: Sustainability

Search keywords: Plain keywords

Concept types: ☐ Person ☐ Location ☑ Organization ☑ Nonentities

Category: Pick ▾ Society→Government→Government-Related Organizations | Society→Government→Finance | Science→Technology→Energy

List of concepts that correlate the most

Below is an extended list of concepts that correlate most based on the specified context. The numbers are not exact, but an approximation.

People

Locations

Organizations

- Facebook (r = 0.9350)
- Federal government of the United States (r = 0.9330)
- Twitter (r = 0.9329)
- U.S. Securities and Exchange Commission (r = 0.9279)
- Republican Party (United States) (r = 0.9269)
- Democratic Party (United States) (r = 0.9229)
- Wall Street (r = 0.9165)
- Labour Party (UK) (r = 0.9013)
- USA Today (r = 0.8978)
- Food and Drug Administration (r = 0.8592)

Non-entities

- Application software (r = 0.9633)
- Natural environment (r = 0.9622)
- Market (r = 0.9614)
- Limited liability company (r = 0.9603)
- Technology (r = 0.9602)
- Money (r = 0.9597)
- Brand (r = 0.9596)
- Chief executive officer (r = 0.9596)
- Sustainability (r = 0.9595)
- Engineering (r = 0.9586)
- United States dollar (r = 0.9583)
- Business (r = 0.9583)
- Email (r = 0.9579)
- Subsidiary (r = 0.9575)
- Debt (r = 0.9573)
- Management (r = 0.9572)
- Economy (r = 0.9572)
- Marketing (r = 0.9571)
- Urban area (r = 0.9569)
- Consumer (r = 0.9568)
- Asset (r = 0.9567)
- Court (r = 0.9567)
- Infrastructure (r = 0.9565)
- Education (r = 0.9564)
- Retail (r = 0.9558)
- Globalization (r = 0.9556)

Figure 23: ***List of concept that correlate with the concept of "Sustainability" using a user-provided reduced context***

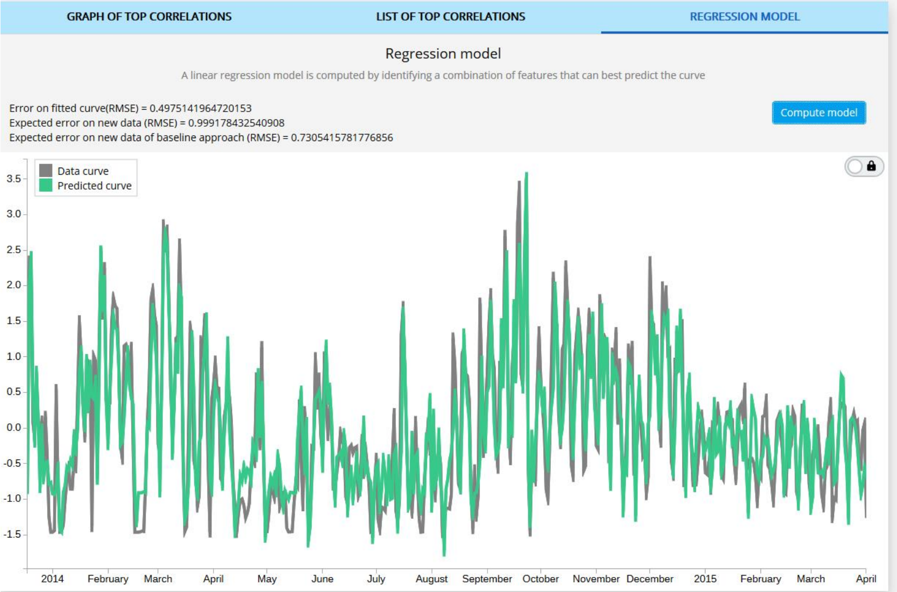
7.4 Nowcasting

In this task, the goal is to answer the following question: Given a time series data for an indicator or any other regression variable, *y*, until time *t*, can we use the available information to predict the value of *y* at the next time point y_{t+1}. For the early work we begin by implementing a simple linear regression.

We treat the given task as a learning problem in which we have a set of *N* learning features and the value y_{t+1} that we wish to learn to predict. The learning features can be anything that would provide potential information about the regression variable. The features can however contain only information available at time *t* or sooner since information at time *t+1* is not available yet. As features we use the value y_t, and the latest trend of *y* – was the curve steady, going up or down relative to y_{t-1} i.e. the value $y_{t-1} - y_{t-2}$. In addition, we also want to include information that is provided by the concept time series. In the same way as for the values of *y*, we add for each signal two relevant features – the value of the signal at *t* and the latest trend of the signal. Given a set of m concepts each with *s* signals we generate $2 + 2 * s * m$ features.

$$\hat{y}_{t+1} = \sum_{i=1}^{p} b_i y_{t-i} \quad + \sum_{j=1}^{m} \sum_{k=1}^{s} a_{jk} x_{jk_t} + c_{jk}(x_{jk_t} - x_{jk_{t-1}}) + \varepsilon_t$$

Where $\hat{y}_{t+1}$ is our prediction for the regression variable *y* at time *t+1*, y_{t-i} represents the true value of *y* at time *t-i* i.e. historical values of *y*, x_{jk_t}represents the value of the signal *k* at time *t* for the concept *j*. *a*, *b* and *c* are the learned weights of the model. The result of this regression model can be seen in Figure 24.



Regression model

A linear regression model is computed by identifying a combination of features that can best predict the curve

Error on fitted curve(RMSE) = 0.4975141964720153
Expected error on new data (RMSE) = 0.999178432540908
Expected error on new data of baseline approach (RMSE) = 0.7305415781776856

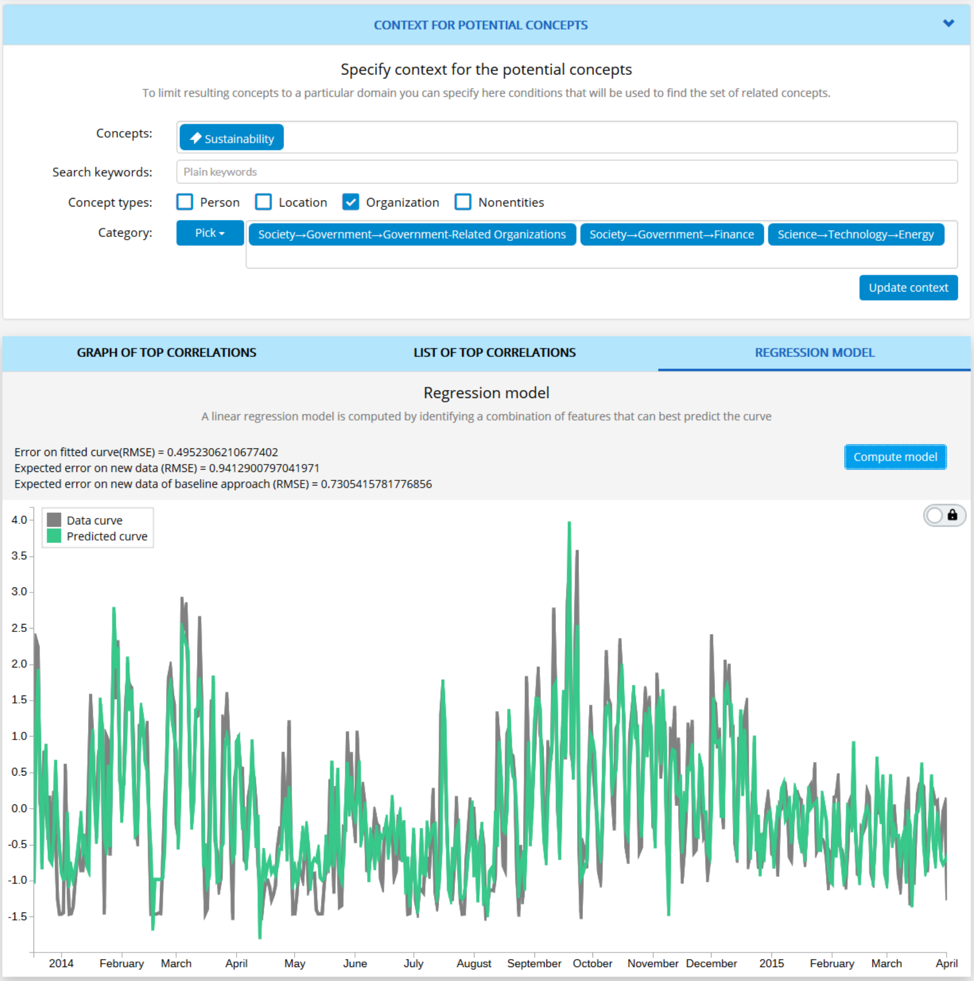
Figure 24: *Linear regression model result*

In our experiments we used only the signal from the number of mentions in mainstream media for the concepts since the beginning of ER data which corresponds to 500 data points per concept modeled as a data series. We also tested with daily time series with 500 points (e.g. historical financial market data limited to the same date as the ER data). The space of the learning features is however large – if we were to use all 1.8 million concepts there would be 3.6 million features. This represents a problem for the linear regression algorithm we used for this early prototype since the model can easily overfit the data.

In order to avoid overfitting to the data we employ two mechanisms. The first is that we reduce the space of the learning features. We can be certain that most of the concepts don't provide valuable information for a given data curve. We can leverage the knowledge of the

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

user to limit the number of irrelevant features used. The intuition is that if, for example, the time series provided by the user in order for us to predict the next value represents the stock value of some technology company, than the potentially relevant concepts are those which are related to the company itself and domain of technology. Using the developed user interface for providing the domain context, the user can provide one or more search conditions that can be used to reduce the space of concepts to the range of a few thousands. This feature is shown in Figure 25.



CONTEXT FOR POTENTIAL CONCEPTS

Specify context for the potential concepts

To limit resulting concepts to a particular domain you can specify here conditions that will be used to find the set of related concepts.

Concepts: Sustainability

Search keywords: Plain keywords

Concept types: □ Person □ Location ☑ Organization □ Nonentities

Category: Pick ▾ Society→Government→Government-Related Organizations Society→Government→Finance Science→Technology→Energy

Update context

GRAPH OF TOP CORRELATIONS | LIST OF TOP CORRELATIONS | REGRESSION MODEL

Regression model

A linear regression model is computed by identifying a combination of features that can best predict the curve

Error on fitted curve(RMSE) = 0.4952306210677402
Expected error on new data (RMSE) = 0.9412900797041971
Expected error on new data of baseline approach (RMSE) = 0.7305415781776856

Compute model

Figure 25: *Linear Regression model result fitted with a user-reduced context*

Since the number of features can still be larger than the number of training examples (number of data points) we used the linear regression with l_2 regularization. This method computes the target value using a function that is a weighted linear combination of the learning features. The l_2 regularization is used to penalize the large weights linear function.

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

This in essence forces the algorithm to identify a simple model – that is, a model, where most of the features have weights close to 0.

Besides providing the predicted curve, we also compute the error of the model compared with the actual data curve. The measure used for computing the error with this algorithms is the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{t=k}^{n}(\hat{y}_{t+1} - y_{t+1})^2}{n-k}}$$

Where $\hat{y}_{t+1}$ is the value of the regression variable at time *t+1* which we predict at time *t*, y_{t+1} is the true value of the regression variable at time *t*. *n-k* represents the total number of predictions we made starting at time *k* and ending at time *n*.

It is important to note that the obtained error is computed on the training data and is an optimistic estimate of the error that would be obtained if running the model on the new, unseen data. To obtain a more unbiased error estimate we also train and test models using cross-validation. In cross-validation, data is split into *n* folds (in our experiments we used 10 folds). All but one fold is then used for training the model and the left-out fold is then used to test the model. The procedure is repeated *n* times, each time with a different left-out fold. On each tested model we can compute the RMSE and then report the average RMSE as the error that we would expect to obtain on the new data.

7.5 Future Work

Firstly we need to add the ability to create and test the autoregressive models with the top correlated signals and test predictive power. Then we need to use only these signals, possibly selected by the user, as features for the linear regression algorithm we have implemented. We also intend to provide the option to use the simpler autoregressive models with multiple signals manually selected. These changes further address the issue with attempting to use too many variables in the linear regression algorithm.

This work deals with lag in the release of indicators e.g. a monthly indicator being released 1 week after the beginning of the following month. Next we need the ability to handle dealing with values for the current time e.g. predict the value of a monthly indicator during the month it refers to. To do this we plan to spread the provided indicator time series data and spread the changes in the value over smaller time periods through linear interpolation. For example, if an indicator is a monthly value e.g. the monthly unemployment numbers for Italy, we can spread the change between each month to each week or each day of that month. We can than treat the monthly time series as a weekly or daily time series and perform correlation and prediction as described previously.

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

8 Usage Scenario

This section is dedicated scenarios that illustrate usage of social media mining results for SYMPHONY use-cases and for the integrated SYMPONY system.

8.1 Automatic Polling and Initialization of ABM Variables

In the automatic polling scenario which we plan to add to the SYMPHONY Dashboard (Task 5.4) we provide the option to visualize the number of mentions and social media sentiment regarding a particular query or concept.

Consider the scenario where a policymaker, researcher or journalist wants to quantify "Confidence in the European Central Bank". She can look up all tweets related to the concept "European Central Bank" for each day and look at the number and ratio of positive and negative sentiment. Thus arriving at a measure that could be a proxy for "Confidence in the European Central Bank". This number could be highly biased and very far from reality but it can provide an approximate answer to a question without any delay or cost that would be associated with traditional polling.

We picked this example also because the ABM used within SYMPHONY has an initial parameter named "Confidence in the European Central Bank", thus initializing this and other initial values with social media data seems interesting to explore.

8.2 Correlations

A policymaker or researcher could pose the question, what media signals are linked to a particular indicator? She could then upload the historical data for that indicator. Progressively she would reduce the context for the correlations using her expert knowledge. Eventually she would arrive at a set of signals which could form the basis for a more resource intensive study or to a result which could be the starting point for deeper research.

8.3 Nowcasting

While nowcasting can be used in the same use case as before in 8.2, it can also be used to inform a policymaker's decision on some issue. By having an accurate approximation of an indicator weeks in advance, the policymaker can make decisions much earlier or at least more informed decisions. This use case is at the basis of Work Package 2 and the value of this use case seems self-explanatory.

9 Discussion

In this section we discuss usage of data sources and bias, usage of existing components and modelling issues.

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

9.1 Methodology

For the social data mining tasks described in this deliverable we used CRISP-DM (Cross Industry Standard Process for Data Mining) methodology [29]. The methodology breaks the mining process into six following steps: business understanding, data understanding, data preparation, modelling, evaluation and deployment. Business understanding step of the methodology refers to understanding the basics of ABM business model that is being developed on SYMPHONY, as well as SYMPHONY business and sustainability use-cases. In data understanding phase we examined the structure of data, volume, arrival rate and basic statistics such as word frequency distribution. In data preparation step we perform data retrieval, on-line processing and efficient storage of data. In our case data was enriched with concepts extraction, sentiment detection, matching between social media posts and news articles, and computing diffusion features. After the data is prepared, machine learning models are built in the modelling phase. In evaluation phase the developed models are evaluated using some of the standard evaluation procedure. For the problem of twitter post to news events matching, we used random 50-50 split, where randomly chosen 50% of the data is used to train the model, and other 50% is used as testing examples as if the label is unknown. By counting the number of true positive, true negative, false positive and false negative samples from the testing set, we calculate precision and recall of the model. In the last step of the methodology, the developed model is deployed as a web service.

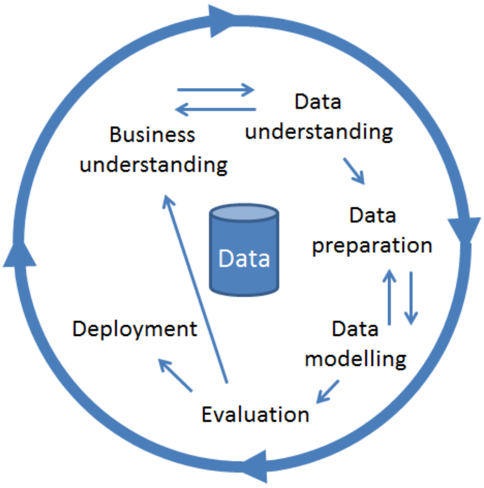


Figure 26: *CRISP-DM data mining methodology*

9.2 Data Sources and Bias

As described in Section 3.2, the source of mainstream news and blogs is IJS Newsfeed. As explained in [3], Newsfeed periodically crawls a list of RSS feeds and a subset of Google

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

News and obtains links to new articles. Upon download and parsing the articles, it extracts potential news sources from links to other sites. If the link destination has a valid RSS feed that is not already in the system, it is added to the list of RSS feeds to be crawled periodically. Additional sources were also identified manually and added to the list.

Our data collection process for tweets was described in the SYMPHONY Deliverable **D2.1 Social media streams processing infrastructure**: geographical bounding boxes around the top 10 cities by population for each country, namely, UK, Germany, Italy and Slovenia, are used to make queries to the twitter API which returns all user geo-tagged tweets published within those bounding boxes.

It is immediately obvious that this collection process introduces massive geographic bias into the data, not representing at all rural areas and limiting data to mostly big cities. The fact that internet access and the ownership of a computing device are required to use social media means that poor people are unlikely to be proportionally represented. It's also likely that there is an age bias, anecdotal bias shows that no grandmothers in our group of friends were twitter users (though many were on Facebook). Other more subtle bias are undoubtedly present. In using this data for automatic polling, the user is responsible for being aware of the biases.

Our hypothesis is that social media posts, however unrepresentative, will give us a signals that may be used to nowcast indicators. Since this part of the work is not based on polling and the methodology for proving that a signal (feature, variable, feature) has predictive power is independent of any bias, we consider that bias does not play a role when nowcasting.

9.3 Existing Components

In the SYMPHONY project we use some existing services whose development was partly supported by previous European projects. These are IJS Newsfeed and Event Registry, both introduced in Section 3 and previously in SYMPHONY Deliverable **D2.1**. Event Registry is also being augmented within the scope of this project by adding social media data as well as the work detailed in Section 7 Correlate and Nowcast.

We also use QMiner[18] for the work described in SYMPHONY Deliverable **D2.**1 and it was augmented in order to perform the work described in Section 5 Opinion Diffusion.

[18] http://qminer.ijs.si/

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

9.4 Terminology

In our work the terms *variable*, *feature* and *signal* are equivalent. The word *model* is used to describe a statistical or machine learning model. The use of the word *predict* and its variations means that the value of a variable was arrived at by a model **without** the model having seen that particular value. This means that the model was fitted without that value - sometimes called out of sample forecasting. This is the common procedure and usage of the word in computer science and engineering. We make this point since in other fields, in-sample forecasting is common, as is fitting a particular curve to in-sample data and stating that the curve predicts the variable. We also use the term predict instead of nowcast, since from the point of view of the algorithm, this is effectively a prediction in the sense we just described.

9.5 Modelling Issues

We would like to point out that we are aware of problems with our approach. In particular, some of the issues with Google Flu Trends[19] [30] described in [31], [32] and [33] are particularly relevant. In our work we address replicability and transparency – our data is publicly available (unlike Google query data) and the entire process of building and testing a model is entirely transparent to the user.

Algorithm dynamics issues occur when there are changes to the data generation process caused by engineering changes. The easy case is when we change data processing algorithms. For example, if we were to replace our NER algorithms with better ones that would surface many more occurrences of entities or if we significantly improved the recall of our tweet matching algorithm. This invalidates the previously built models but assuming the new algorithms are used to update the old data, new models can be trained without issue. The harder to address case is when the changes happen to the data collection process, for example, if we suddenly add a large number of news sources to Event Registry data or suddenly start collecting more tweets by changing how our social media collection process works. Then, even if we retrain the models, said models would have to deal with a sudden discontinuity in the data. We hypothesize that if the models are built some time steps after the changes, the training process would be able to fix the issues and that the correlation algorithm would eventually adapt also. This hypothesis however remains untested and how long it would take for both algorithms (correlation and prediction) to adapt is unknown and perhaps specific to individual data points and indicators. The final problem posed by this issue is that model invalidation should be automatic. This is effectively a distributed system with many people working on different parts of the system i.e. IJS Newsfeed, Event Registry, and the components developed for SYMPHONY are all developed and maintained somewhat

[19] https://www.google.org/flutrends/

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

independently: one person could, hypothetically, make a change without informing the others. Thus the only option is to attempt to automatically detect these changes and automatically invalidate models and warn end users of potential issues. Off course, presently models cannot be saved and must be retrained every time which also helps mitigate this issue, but in the future this assumption might not hold.

Variations in our data without underlying causes (changes to the related indicator) pose the biggest challenge to our approach. News and social media can suddenly have a high number of posts on a topic without the underlying indicator changing. Traditional media has echo effects, where stories, if popular, can be re-published by other sources and new articles created on the same topic do to unexpected popularity of the original articles in a bid to capture the attention of readers or to express agreement or disagreement with them. This can originate an explosion of articles on a given topic without necessarily anything new occurring. For social media the phenomena is very similar. An example would be if a couple of news articles about unemployment in Italy were published and generated interest even though unemployment numbers had remained mostly constant over the previous months. Assuming these articles were popular with readers, other publishers would re-publish these articles and create new ones on the same topic. The attention generated in mainstream media would spread to social media with a large number of social media posts on the topic. If we had built a model to predict unemployment in Italy using concepts mentioned in these posts, than our model would necessarily predict differently falsely assuming that the change in its input was indicative of a change in the underlying data. This would be similar to the phenomena reported during the pH1N1 pandemic of 2009 though the phenomena was reversed: less searches for the tracked queries were made than would have been made normally, resulting in underestimation of influenza-like illnesses [33]. At present, we are not aware of any automatic way to address this challenge.

10 Conclusions

In this deliverable we discussed the work done in the context of WP2 Collective intelligence for Nowcasting **Task 2.2 - Tracking Cross-Lingual Information and Opinion Diffusion** and part of **Task 2.3 - Definition and Development of the SYMPHONY social media based expectations' indicators**. We consider that all subtasks contemplated in Task 2.2 were successfully completed:

1. Event Detection (Section 1)
2. Opinion Diffusion (Section 5)
3. Sentiment Classification (Section 6)

Some additional work needs to be done regarding the evaluation of sentiment classification and improvements to event detection would be useful. We have also developed a very good prototype for Task 2.3 with the correlation component being finalized and the prototype

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

already being available to be used by others. Overall, the direction for our remaining work within Work Package 2 is clear and on track.

11 References

[1] K. Weil, "Measuring Tweets," Twitter Official Blog, 22 February 2010. [Online]. Available: https://blog.twitter.com/2010/measuring-tweets.

[2] I. Novalija, M. Papler and D. Mladenić, "Towards Social Media Mining: Twitterobservatory," in *SiKDD*, Ljubljana, Slovenia, 2014.

[3] M. Trampuš and B. Novak, "Internals Of An Aggregated Web News Feed," in *SiKDD*, Ljubljana, Slovenia, 2012.

[4] G. Leban, B. Fortuna, J. Brank and M. Grobelnik, "Cross-lingual detection of world events from news articles," in *The 13th International Semantic Web Conference*, Trentino, Italy, 2014.

[5] G. Leban, B. Fortuna, J. Brank and M. Grobelnik, "Event Registry – Learning About World Events From News," in *World Wide Web conference*, Seoul, Korea, 2014.

[6] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter and T. Berners-Lee, "Request for Comments: 2616: Hypertext Transfer Protocol -- HTTP/1.1," Network Working Group, The Internet Society, 1999. [Online]. Available: https://tools.ietf.org/html/rfc2616. [Accessed 2015 March 10].

[7] M. Ohye and J. Kupke, "Request for Comments: 6596," Internet Engineering Task Force (IETF), April 2012. [Online]. Available: http://tools.ietf.org/html/rfc6596. [Accessed 15 March 2015].

[8] Facebook, "The Open Graph Protocol," 20 October 2014. [Online]. Available: http://ogp.me/. [Accessed 10 March 2015].

[9] M. Davis and K. Whistler, "Unicode Standard Annex #15: Unicode Normalization Forms," 5 June 2014. [Online]. Available: http://www.unicode.org/reports/tr15/. [Accessed 6 March 2015].

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

[10] P. Saleiro, L. Rei, A. Pasquali, C. Soares, J. Teixeira, F. Pinto, M. Nozari, C. Felix and P. Stretch, "POPSTAR at RepLab 2013: Name ambiguity resolution on Twitter," in *CLEF 2013 Eval. Labs and Workshop Online Working Notes*, Valencia, Spain, 2013.

[11] S. H. Lee, J. Kim and S. Hong, "On URL normalization," in *Computational Science and Its Applications*, Singapore, 2005.

[12] T. Berners-Lee, R. Fielding and L. Masinter, "STD: 66: Uniform Resource Identifier (URI): Generic Syntax," Network Working Group, The Internet Society, January 2005. [Online]. Available: https://tools.ietf.org/html/rfc3986. [Accessed 10 March 2015].

[13] B. S. Greenberg, "Person-to-Person Communication in the Diffusion of News Events," *Journalism & Mass Communication Quarterly December,* vol. 41, pp. 489-494, 1964.

[14] E. Bakshy and C. Marlow, "The Role of Social Networks in Information Diffusion," in *In Proceedings of the 21st international conference on World Wide Web (WWW '12)*, New York, NY, USA, 2012.

[15] A. Guille, H. Hacid, C. Favre and D. A. Zighed, "Information diffusion in online social networks: a survey," *SIGMOD Rec,* vol. 42, no. 2, pp. 17-28, 2013.

[16] J. Brank, G. Leban and M. Grobelnik, "A High-Performance Multithreaded Approach For Clustering a stream of documents," in *SiKDD*, Ljubljana, 2014.

[17] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, 1994.

[18] R. Sedgewick, Algorithms in C++, Third Edition, Part 5: Graph Algorithms, Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2002.

[19] R. C. Prim, "Shortest Connection Networks And Some Generalizations," *Bell System Technical Journal,* vol. 36, no. 6, p. 1389–1401, 1957.

[20] B. O'Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," in *4th Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington, DC, United States of America, 2010.

[21] J. Nofsinger, "Social Mood and Financial Economics," *The Journal of Behavioral Finance,* vol. 6, pp. 144-166, 2005.

[22] P. Nakov, . Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov and T. Wilson, "SemEval-

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688

2013 Task 2: Sentiment Analysis in Twitter," in *Proceedings of the International Workshop on Semantic Evaluation, SemEval*, Atlanta, Georgia, United States of America, 2013.

[23] Y. Bengio, R. Ducharme, P. Vincent and C. Janvin, "A neural probabilistic language model," *The Journal of Machine Learning Research,* vol. 3, pp. 1137-1155, 2003.

[24] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *The 31st International Conference on Machine Learning*, Beijing, China, 2014.

[25] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *International Conference on Learning Representations (ICLR)*, Scottsdale, Arizona, United States of America, 2013.

[26] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," *Language Resources and Evaluation,* vol. 10, pp. 1320-1326, 2010.

[27] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in *ACL Student Research Workshop. Association for Computational Linguistics*, Ann Arbor, Michigan, 2005.

[28] A. Go, L. Huang and R. Bhayani, *Twitter sentiment analysis,* 2009.

[29] H. Choi and H. Varian, "Predicting the present with google trends," *Economic Record,* vol. 88, no. 1, pp. 2-9, 2012.

[30] D. Vanderkam, R. Schonberger, H. Rowley and S. Kumar, "Nearest Neighbor Search in Google Correlate," Google, 2013.

[31] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature,* vol. 457, no. 7232, pp. 1012-1014, 2009.

[32] D. Lazer, R. Kennedy, G. King and A. Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis," *Science,* vol. 343, no. 6176, pp. 1203-1205, 2014.

[33] D. Butler, "When Google got flu wrong," *Nature,* vol. 494, pp. 155-156, 2013.

[34] S. Cook, C. Conrad, A. L. Fowlkes and M. H. Mohebbi, "Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic," *PloS one,* vol. 6, no. 8, p. e23610, 2011.

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611688