Social Mining & Big Data Ecosystem

SoBigData

RESEARCH INFRASTRUCTURE

Project Acronym	***SoBigData***
Project Title	***SoBigData Research Infrastructure*** ***Social Mining & Big Data Ecosystem***
Project Number	***654024***
Deliverable Title	***SoBigData Evaluation Framework: Engagement and Sustainability Report***
Deliverable No.	***D11.4***
Delivery Date	***June 2019***
Authors	***Nino Antulov-Fantulin, Tian Guo***

SoBigData receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654024

DOCUMENT INFORMATION

PROJECT	
Project Acronym	SoBigData
Project Title	SoBigData Research Infrastructure Social Mining & Big Data Ecosystem
Project Start	1st September 2015
Project Duration	48 months
Funding	H2020-INFRAIA-2014-2015
Grant Agreement No.	654024
DOCUMENT	
Deliverable No.	D11.4
Deliverable Title	SoBigData Evaluation Framework: Engagement and Sustainability Report
Contractual Delivery Date	31 May 2019
Actual Delivery Date	10 June 2019
Author(s)	Nino Antulov-Fantulin (ETHZ), Tian Guo (ETHZ)
Editor(s)	Beatrice Rapisarda (CNR), Nino Antulov-Fantulin (ETHZ)
Reviewer(s)	Valerio Grossi (CNR), Beatrice Rapisarda (CNR)
Contributor(s)	K. Bontcheva (USFD), G. Gorrell (USFD), N. Andrienko (FRH), G. Andrienko (FRH), R. Trasarti (CNR), L. Pappalardo (CNR), G. Rossetti (CNR), R. Guidotti (CNR), F. Pratesi (CNR), C. Muntean (CNR), C. Boldrini (CNR), S. Cresci (CNR), M. Tesconi (CNR), F. Lillo (SNS), G. Curato (SNS), A. Sirbu (UNIPI), P. Ferragina (UNIPI), A. Facchini (IMT), G. Caldarelli (IMT), T. Squartini (IMT), M. Dumas (UT), J. Manuel Duran (TUDelft), A. Anand (LUH)
Work Package No.	WP11
Work Package Title	WP11 – NA5_Evaluation
Work Package Leader	ETHZ
Work Package Participants	ALL
Dissemination	Public
Nature	Report
Version / Revision	V1.1
Draft / Final	Final
Total No. Pages (including cover)	23
Keywords	Services, Indutrial Stakeholders, SoBigData Catalogue

DISCLAIMER

SoBigData (654024) is a Research and Innovation Action (RIA) funded by the European Commission under the Horizon 2020 research and innovation programme.

SoBigData proposes to create the Social Mining & Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by "big data". Building on several established national infrastructures, SoBigData will open up new research avenues in multiple research fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, re-use and integration of state-of-the-art big social data, methods, and services, into new research.

This document contains information on SoBigData core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as SoBigData Board members. Any reference to content in this document should clearly indicate the authors, source, organisation and publication date.

The document has been produced with the funding of the European Commission. The content of this publication is the sole responsibility of the SoBigData Consortium and its experts, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 27 member states of the European Union. It is based on the European Communities and the member states' cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (http://europa.eu.int/).

GLOSSARY

ABBREVIATION	DEFINITION
API	Application programming interface
GitHub	Web-based hosting service for version control of computer code using Git.

TABLE OF CONTENT

DELIVERABLE SUMMARY

This deliverable (D.11.4) reports about the activities in T11.3, and more specifically on the stakeholder engagement, commercial sponsorship, and sustainability activities. Here, we describe which services are available for engaging a range of stakeholders, going beyond the research community from T11.2, and towards the wider stakeholder community and funders.

EXECUTIVE SUMMARY

This deliverable (D.11.4) describes which services can be offered from each institution for commercial stakeholders. In particular, we describe each service in a language understandable to the non-research community, conditions for providing these services. Furthermore, we describe sustainability and engagement with industrial stakeholders. We leverage the innovation activities and stakeholder connections from WP5, to pursue sponsorship deals with relevant industry partners, funders and standardization bodies, to ultimately make this effort sustainable.

1 SERVICES AND STAKEHOLDER ENGAGEMENT

We list the potential catalogue of services for industrial stakeholders from different institutions in SoBigData project.

1.1 CNR SERVICES

The involved labs in CNR offer three categories of services as follows.

1.1.1 HPC LAB @ CNR SERVICES

The HPC Lab at ISTI-CNR can offer its expertise concerning learning to rank. In particular, the method, which was integrated into the SoBigData research infrastructure, is QuickRank, a C++ suite of **L**earning **t**o **R**ank algorithms (LtR), which is offered to **istella**, the Italian search engine for the web, which uses the ranking algorithms for ranking web pages in response to user queries.

Services: QuickRank is an efficient Learning-to-Rank toolkit providing several C++ implementations of LtR algorithms. QuickRank was designed and developed with efficiency in mind.

The LtR algorithms currently implemented are:

- GBRT: J. H. Friedman. Greedy function approximation: a gradient boosting machine. Annals of Statistics, pages 1189–1232, 2001.
- LamdaMART: Q. Wu, C. Burges, K. Svore, and J. Gao. Adapting boosting for information retrieval measures. Information Retrieval, 2010.
- Oblivious GBRT / LamdaMART: Inspired to I. Segalovich. Machine learning in search quality at yandex. Invited Talk, ACM SIGIR, 2010.
- CoordinateAscent: Metzler, D., Croft, W.B.. Linear feature-based models for information retrieval. Information Retrieval 10(3), pages 257–274, 2007.
- LineSearch: D. G. Luenberger. Linear and nonlinear programming. Addison Wesley, 1984.
- RankBoost: Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. An efficient boosting algorithm for combining preferences. JMLR, 4, 933-969 (2003).
- DART: K.V. Rashmi and R. Gilad-Bachrach. Dart: Dropouts meet multiple additive regression trees. JMLR, 38 (2015).
- Selective: C. Lucchese, F. M. Nardini, S. Orlando, R. Perego and S. Trani. Selective Gradient Boosting for Effective Learning to Rank. ACM SIGIR, 2018. README

QuickRank also provides novel learning optimizations. Currently implemented optimizers are:

- CLEAVER: C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, F. Silvestri, S. Trani. Post-Learning Optimization of Tree Ensembles for Efficient Ranking. In Proc. ACM SIGIR, 2016. README
- X-CLEAVER: C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, F. Silvestri, S. Trani. X-CLEaVER: Learning Ranking Ensembles by Growing and Pruning Trees. Paper under revision.

Sustainability: This tool/framework is available on the SoBigData Research Infrastructure as well as on dedicated websites. It is open source and free use. The tool is an open source software in continuous development. The perspective for the use and maintenance of the service in the next few years are good. Potential collaborations with Industrials partners is promising the near future.

1.1.2 UBIQUITOUS INTERNET & WAFI @ CNR SERVICES

Services: Online social network analytics is provided for customer-oriented personalization. For many commercial stakeholders, availability of data is not a problem but availability alone does not guarantee a business advantage. Data analytics is necessary to bridge the gap between access to huge amounts of data, and the knowledge required to foster commercial applications such as product customizations. To this end, social analytics, applied to both internal commercial data and public social media data, can shed light on target customers' preferences and can help identify areas of commercial expansion. Focusing on products customization for a target customer, the offered service can monitor the public online conversations the customer and its friends engage with, and, relying on social computing techniques based on the ego network concept, identify products or characteristics of products that the customer might be interested in. This will also increase the loyalty of existing and future customers to the brand.

Big Data analytics on personal devices is available as well. Current data analytics is typically centralized but this might (i) impose excessive stress on wireless technologies to collect data, (ii) not be compliant with real-time requirements on the results of the analytics operation, and (iii) not be compliant with data ownership constraints requiring data not to be moved outside certain locations (e.g., the customers' smartphones). Distributed machine learning (DML) solutions, which have been vastly investigated in the last years, can solve or at least mitigate these problems. We offer consulting services to commercial stakeholders interested in adopting these solutions in their products.

For both services, the technologies behind them have been developed within the SoBigData project. These services are provided with price upon negotiation.

The Twitter Monitor is an interactive Web application designed to access the Twitter stream by exploiting the public TwitterStreaming APIs, allowing the user to manage concurrent monitors: with parallel listening sessions through an interactive Web interface. It features a set of functionalities aimed at minimizing the loss of data in case of network problems, recovering from error situations, providing alerts to system administrators.

Twitter Monitor is integrated and hosted in the SoBigData platform and is provided by IIT-CNR to SoBigData with free use. The support and evolution in case of commercial use offered by SoBigData will be collegially discussed inside the consortium.

Sustainability: As for the sustainability of the services, the main issue that the service might face is that the current big data analytics facilities may not scale up as desired by the commercial stakeholders. On the long run, this may require additional investments from CNR or from the commercial partners.

The strong point addressed by these two services is that there is currently a large, untapped potential for developing new affordable, big social data analytics products and services, since many companies in diverse areas (e.g. business intelligence, market research, campaign and brand reputation management, customer relationship management, enterprise search and knowledge management) are analysing and comparing big social data, often in a labour intensive and expensive manner.

Regarding industrials partners, the research groups from CNR are involved in a number of National and Regional Competence Centers, some already active (such as the National Competence Centre on Advanced Robotics and Enabling Digital Technologies – ARTES 4.0, whose goal is to promote and accelerate innovation through public-private partnerships combining technological expertise, research programs, technology transfer and education on advanced robotics and associated enabling technologies) and some going to be activated soon (such as the Regional Competence Centre on BigData and Artificial Intelligence), that involve universities, research centres, companies, foundations and small and medium sized enterprises. In particular, at national level, ARTES 4.0 is considered strategic for the digital transformation of Italy,

promoted by the Ministry of Economic Development, through a new generation of ICT professionals and a new generation of software systems with embedded AI capabilities.

The main value of the Twitter Monitor is the ease of use, reliability, and performance. It depends on the Twitter APIs so any change in them over the time might require possible adjustments performed bu IIT-CNR.

1.1.3 KDD LAB @ CNR SERVICES

The KDD Lab at CNR offers its expertise concerning big data analysis.

Services: In particular, they provide services and methods for:

- Social Network Analysis.
- Managing and querying spatio-temporal data for understanding and studying human mobility, also including the study of migration flows.
- Sports analytics for supporting data scientists to describe performances by means of data, statistics and models.
- Investigate the changes in people's behaviour related to well-being indicators.

On the SoBigData Catalogue, we offer several tools dedicated to address common social network analysis modelling and analysis, among them:

- Community Discovery. Two algorithms, Demon and Tiles, tailored for clustering both static as well as dynamic network topologies. A python package to evaluate community partitions against ground-truth (F1-communities). A community discovery dedicated library (CDlib) will be released shortly;
- A library to model multiplex networks;
- A library to simulate and evaluate diffusive processes occurring on top of complex networks (NDlib) along with a dedicated remote experiment server (NDlib-REST)
- Additionally, it includes M-Atlas, Human Mobility tools, Migration Studies, Well-Being indicator methods and so on.

These services are available on the SoBigData Research Infrastructure as well as on dedicated websites.

Sustainability: The scarcity of data concerning the activity of financial institutions was, is and will be one of the main drivers of our research. The free availability of our services has allowed central bankers to include them in the aforementioned "horse races" between competing algorithms. As several tests of this kind have confirmed the good performance of our recipe(s), we expect their popularity, within the community of policy makers, to grow in the future. We also expect to strengthen ongoing collaborations (e.g. with the Dutch National Bank and with Bank of England).

1.2 UNIVERSITY OF SHEFFIELD SERVICES

Services: USFD provides three textual data related services as follows.

- Text analytics for social media
  Natural language processing algorithms were delivered as simple-to-use services, which companies can use to analyse social media content. There are currently 12 such services, spanning English, French, and German. The services offer language identification, tokenisation, part of speech

tagging, named entity recognition, and opinion mining. Further details about the services are available: https://cloud.gate.ac.uk/shopfront#tagged=Twitter

- Analysing online societal debates, e.g. posts around elections
  We offer some automated web services for analysing discussions on social media and newspaper articles, as well as consultancy services to help adapt the methods for analysis to the specific problem being addressed or help companies to apply the existing services to their data. Examples include studying public debates to understand which are the most discussed topics, follow the discussions around them, and track them through time and space. Analyses of debates around elections are a particular focus.

- Analysing online misinformation
  Similar services to those offered for analysing online societal debates.

These services are available on the SoBigData Research Infrastructure as well as on the GATE Cloud platform. They are offered on a freemium model, coupled also with consultancy and customisation service contracts, if required. Further details are available here:

https://sobigdata.d4science.org/web/societaldebates/data-catalogue

Sustainability: Regarding the sustainability of these services USFD is committed to offering these services for the next 5 years. They are being sustained from income generated by consultancy and knowledge transfer projects. Many of the tools have been made open source, to promote replicability and sustainability further.

Further development of the misinformation analysis methods and services is envisaged as part of the WeVerify H2020 project: https://weverify.eu/

Meanwhile, we have been working with the following users of our text analytics and other services:

- South London and Maudsley NHS Trust, London, UK
- TechCity UK
- Synaptica
- NHS Digital, UK
- A world-leading manufacturer of chemical products (NDA in place)
- IFPRI
- NIHR Innovation Observatory, UK
- BuzzFeed
- Press Association
- ITV
- Expert advice to policy makers in the UK, European parliament, and Singapore on online disinformation

1.3 FRAUNHOFER INSTITUTE IAIS SERVICES

FRH provides a visual analytics workflow for identification and semantic interpretation of individual and public places based on episodic mobility data (e.g. geo-referenced social media posts, call data records).

Services: At FRH, researchers propose a visual analytics workflow and a collection of prototype tools for visually-driven identification and interpretation of personal and public spaces. The proposed approach can be used for privacy-preserving analysis of mobility data in semantic spaces. Details of the approach are

described in http://dx.doi.org/10.1177/1473871615581216. A possible further step of analysis is construction of state transition graphs, see http://dx.doi.org/10.1177/1473871617692841.

Aforementioned service is currently provided for free use.

Sustainability: The methods and tools are developing further within a series of research projects. The strong point of the method is involvement of human intelligence through interactive visual interfaces. Human participation enables deeper understanding. The week point is involvement of human experts in analysis, requiring their time and efforts.

1.4 ETHZ SERVICES

ETHZ can offer its experience in modelling complex social systems and data science.

In particular:

1) Analysing the social signals in crypto-currency and block-chain domain:

The ability to track and monitor relevant and important news in real-time is of crucial interest in multiple industrial sectors. We focus on the set of crypto-currency news, which recently became of emerging interest to the general and financial audience. In order to track relevant news in real-time, we (i) match news from the web with tweets from social media, (ii) track their intraday tweet activity and (iii) explore different machine learning models for predicting the number of the article mentions on Twitter within the first 24 hours after its publication.

Beck, Huang, Lindner, Guo, Zhang, Helbing, Antulov-Fantulin, Sensing Social Media Signals for Crypto-currency News, ACM WWW '19 Conference, MSM'19 Workshop

2) Analysis of short-term volatility estimates in crypto-currency markets

Ability to understand which factors drive the fluctuations of the crypto-currency price and to what extent they are predictable is interesting both from theoretical and practical perspective. We study the problem of the short-term volatility forecasting by exploiting volatility history and order book data. Order book, consisting of buy and sell orders over time, reflects the intention of the market and is closely related to the evolution of volatility. We have developed the temporal mixture models capable of adaptively exploiting both volatility history and order book features for short-term volatility forecasting.

T. Guo, A. Bifet and N. Antulov-Fantulin, "Bitcoin Volatility Forecasting with a Glimpse into Buy and Sell Orders," *2018 IEEE International Conference on Data Mining (ICDM)*, Singapore, 2018, pp. 989-994.

3) Assessing the robustness of the structure of socio-technical systems

Conditions: Part of services are available on the SoBigData Research Infrastructure as well as on dedicated websites. Other services are available with a negotiation of conditions.

Sustainability: Weakness: scaling and maintaining the financial support for computing infrastructures.

Strong point: state-of-the-art research in the Fintech domain.

Support letters from industry: **SIX** (**S**wiss **I**nfrastructure and E**x**change), which is a financial service provider that operates the infrastructure of Switzerland's financial center.

1.5 IMT SERVICES

Services: The IMT School for Advanced Studies can offer its expertise concerning network analysis, which have been already employed by policy makers working in central banks.

- MaxAndSam Network Reconstruction Method

This method aims at reconstructing economic and financial networks (be they monopartite or bipartite) taking as input nodes fluxes (e.g. assets and liabilities, exports and imports) and the total number of observed links. The reconstruction provided by our method has been compared with the performance of similar algorithms: remarkably, our method is “the clear winner” among the ensemble algorithms [1,2,3,4].

[1] Cimini, Giulio, et al. "Systemic risk analysis on reconstructed economic and financial networks." Scientific reports 5 (2015): 15758.

[2] Squartini, Tiziano, et al. "Enhanced capital-asset pricing model for the reconstruction of bipartite financial networks." Physical Review E 96.3 (2017): 032315.

[3] Squartini, Tiziano, et al. "Reconstruction methods for networks: the case of economic and financial systems." Physics Reports (2018).

[4] Mazzarisi, Piero, and Fabrizio Lillo. "Methods for reconstructing interbank networks from limited information: A comparison." Econophysics and Sociophysics: Recent Progress and Future Directions. Springer, Cham, 2017. 201-215.

- DebtRank Systemic Risk Estimation Method

This method aims at providing a measure of distress of financial institutions. DebtRank is an iterative method quantifying the impact of subsequent (financial) shockwaves on the entities constituting the network under analysis; it complements usual stress tests since it also estimates their “closeness” to default. DebtRank has recently gained increasing attention, being employed by the ECB to monitor TARGET2 [5].

[5] Battiston, Stefano, et al. "Leveraging the network: a stress-test framework based on DebtRank." Statistics & Risk Modeling 33.3-4 (2016): 117-138.

These services are available on the SoBigData Research Infrastructure as well as on dedicated websites.

Sustainability: Regarding the Sustainability of these services, the scarcity of data concerning the activity of financial institutions was, is and will be one of the main drivers of our research. The free availability of our services has allowed central bankers to include them in the aforementioned “horseraces” between competing algorithms. As several tests of this kind have confirmed the good performance of our recipe(s),

we expect their popularity, within the community of policy makers, to grow in the future. We also expect to strengthen ongoing collaborations (e.g. with the Dutch National Bank and with Bank of England).

1.6 LUH SERVICES

Services: The L3S Research center at LUH can offer services related to Web Science as follows.

- BoilerNET: Automatic boilerplate removal (content extraction) from web pages

BoilerNET is a deep learning based tool for web content extraction. It can process arbitrary HTML web pages and highlight or extract the main content, discarding boilerplate like navigational items, menus, ads etc.

- ArchiveSpark: Processing archive collections using Apache Spark

ArchiveSpark [1] is an Apache Spark framework for easy data processing, extraction as well as derivation for archival collections. Originally developed for the use with Web archives, it has now been extended to support any archival dataset through Data Specifications.

[1] https://github.com/helgeho/ArchiveSpark

The Services can be made available on the SoBigData platform.

Sustainability: L3S would be able to provide some these services free for use in terms of trained models (for boilerNet). However, if the code is used for commercial purpose then there is potential to monetize it under the apache licences.

1.7 SNS SERVICES

Services: SNS can offer its expertise concerning network science and time series analysis with application to economics and finance. In particular the methods:

- Statistically Validated Networks

Many complex systems can be represented by networks, but the large dimensionality does not allow to identify easily the relevant or unexpected connections. To this end, the SVN package constructs a suitable null model and performs a rigorous statistical test to identify the unexpected connections. The method is scalable with the network size. Among the applications (already done) we mention community detection, fraud detection, etc. The method is implemented in the SBD platform.

- Bipartite Network Reconstruction for assessment of systemic risk due to fire sales

One of the main sources of systemic risk is fire sales spillover, i.e. coordinated massive sale of assets whose price drops affect the balance sheet of other institution forcing them to sell. Assessing the systemic risk due to fire sales is complicated, as it requires the knowledge of portfolio composition of all financial institutions. With this method, we propose an alternative estimation approach based on the Maximum Entropy Principle. The method is implemented in the SBD platform.

- Risk spillovers with Granger causality in tails

Identifying spillover of risk of financial assets, i.e. how an extreme event on one asset triggers an extreme event in another asset in the future is complicated due to the low frequency of extreme events. With this method, we propose to use Granger causality in tail to identify econometrically risk spillovers. A rigorous test is implemented and, when many assets are considered, the method returns a directed network of risk spillovers. The method is implemented in the SBD platform.

- Financial Consulting Services

Financial Consulting Services consist in joint applied activities on several aspects of finance on which SNS has consolidated experience and had collaborated with commercial partners in the past. These include

high frequency finance, trading algorithms, transaction cost analysis, portfolio optimization and asset allocation, credit scoring.

These services are available on the SoBigData Research Infrastructure as well as on dedicated websites. Financial Consulting Service is available with a negotiation of price.

Sustainability: As for the sustainability of these services, the abundance of financial data leads to the development of Financial Data Science, a new discipline with many applications in the industry and therefore a potentially large market value. Both regulators (e.g. central banks) and financial institutions (banks, hedge funds, and brokers) benefit enormously from these tools. Moreover, most of the Fintech industry is based on the application of Data Science methods to finance.

SNS Quantitative Finance group had several collaborations with industrial partners for joint projects and financial consulting (e.g. Unicredit, HSBC, List Group, etc.).

1.8 TU DELFT SERVICES

Services: Currently, TU Delft is not directly involved with commercial stakeholders. However, there is great potential for it with the following SoBigData services.

- SoBigData exploratories (mainly City of Citizens, Explainable Machine Learning, Societal Debates)

The aforementioned SoBigData exploratories are valuable for in-house research, which in turn can be offered to commercial stakeholders. For instance, Delft Design for Values offers research consultancy to different stakeholders. Currently, there are projects on AI & Ethics for The Dutch National Police, Sustainable Development for The Hague municipality and Micro Targeting & Digital Platforms for Dutch State Committee on Reforming the Parliamentary System

- SoBigData e-Learning Area

As for the SoBigData e-Learning Area, its importance lies in educating researchers on the platform, its benefits and limitations for a more successful approach to our stakeholders.

TU Delft can provide these services with either condition: free of use and at a price (both fixed and upon negotiation). What determines the conditions are the type of service provided.

Sustainability: As for the sustainability of the services, TU Delft is expanding at a fast pace, and we anticipate that in the near future, we will be integrating more strongly the SoBigData platform into our research network. Currently, we are recruiting qualified researchers for promoting SoBigData and SoBigData++ in research as well as with stakeholders.

1.9 UNIPI SERVICES

Services: Researchers at UNIPI developed a suite of text analytics software that allow state-of-the-art semantic annotation and enrichment of linguistic texts to be used in downstream applications or other IR tools to empower the processing of textual big data.

The suite is available at: https://sobigdata.d4science.org/web/tagme and it consists of four main services, which are described in the next paragraph:

- TagMe performs end-to-end entity linking by annotating a natural language text with proper Wikipedia entities. This system has been specifically designed to work well both on well-structured (e.g., news and Web pages) and noisy (e.g., tweets) documents
- WAT is a state-of-the-art entity linking for large-scale entity annotation which improves quality of the annotations performed by TagMe by making it significantly faster (20x of speedup) and more accurate (+10% in accuracy over a benchmark constituted of 12 datasets) upon well-formed texts.
- SMAPH uses WAT for the detection of Wikipedia entities in Web queries and it is currently the state-of-the-art in this domain. Different from standard text, a query is usually very short (composed by very few characters), irregular and noisy (with misspelling errors commonly present in these kinds of texts). The linking of the input text with Wikipedia entities is performed with a novel algorithmic technology that piggybacks a Web search engine in order expand the context of the query as well as to alleviate the irregularities present in its text.
- SWAT enriches the annotations offered by WAT with proper salience scores, which measure the relevance of the linked entities with respect to the main topics described in the input text. This allows to automatically remove erroneous or not relevant entities that could have been actually annotated by WAT; and it can be used as a more modern alternative to the tf-idf scores on which classical bag-of-words paradigms are based on.

It is able to provide above services as free use under a free registration via the SoBigData platform.

Sustainability: As for the sustainability of the services, the strongest point relies on the fact that all our services implement state-of-the-art algorithms. Their robustness has been tested upon a large and variegate number of datasets, as well as it services has been queried about 800 million times by a variegate set of users. A number of scientific results have also spurred out in order to show the effectiveness of our entity annotators for improving both efficiency and quality of the returned results.

The weakness resides in its scalability and accuracy; they might be both improved in order to annotate more precisely terabyte of textual data, possibly specializing the software over some verticals. This is actually a weakness of all known academic tools, among which ours are the fastest and more accurate, and thus these issues deserve more research and algorithm engineering, which we will pursue in the next years.

Industrials collaborations include Bloomberg (one research grant 2017), Google (two research faculty awards), Spazio Dati (their tool Atoka has been derived from TagMe), Tiscali-Istella and Treccani (part of the search interface at http://www.treccani.it/ has exploited TagMe), and IBM Haifa (several papers make use of our TagMe). Meanwhile, aforementioned services received support in terms of awards, grants from companies, and industrial papers that mention the use of our suite. Few of them are listed below, as an example of impact of the suite onto some industrial prototypes:

- *Company: Workday, Inc* - Michael D. Conover, Matthew Hayes, Scott Blackburn, Pete Skomoroch, Sam Shah. Pangloss: Fast Entity Linking in Noisy Text Environments. Procs. KDD, 2018.
- *Company: Microsoft Research* - Chenyan Xiong, Zhengzhong Liu, Jamie Callan, Tie-Yan Liu. Towards Better Text Understanding and Retrieval through Kernel Entity Salience Modeling. Procs. ACM SIGIR, 2018.

- *Research Institute: AllenAI* - Waleed Ammar, et al. Construction of the Literature Graph in Semantic Scholar. Procs. NAACL-HLT, 2018.
- *Company: Google Zurich* - Marco Cornolti, Paolo Ferragina, Massimiliano Ciaramita, Stefan Rüd, Hinrich Schütze: SMAPH: A Piggyback Approach for Entity-Linking in Web Queries. ACM Transaction of Information Systems, 2019.
- *Company: IBM Haifa* - Yuan Ni, et al. Semantic Documents Relatedness using Concept Graph Representation. Procs. ACM WSDM, 2016.
- *Company: Google Zurich* - Marco Cornolti, Paolo Ferragina, Massimiliano Ciaramita: A framework for benchmarking entity-annotation systems. Procs. WWW, 2013.
- *Company: Google Zurich* - Ugo Scaiella, Paolo Ferragina, Andrea Marino, Massimiliano Ciaramita: Topical clustering of search results. Procs. ACM WSDM, 2012.

1.10 UTARTU SERVICES

Services: UT offers the following services.

- Corporate training on business data analytics

  We have developed a course on business data analytics, which has been delivered both in a 32-hour and in a 64-hours format. We foresee that the 32-hours version could be offered for corporate training, for example to groups of 10-12 business analysts and managers who wish to acquire data analytics skills.

- Consultancy services on use of social media datasets in business data analytics

  We have developed the social media analysis datasets, particularly those available in the SoBigData infrastructure. By combining this with our existing expertise in business data analytics and social network analysis, we are in a position to offer consultancy services to companies seeking to extend their internal business data analytics pipelines with external datasets.

The Institute of Computer Science at University of Tartu regularly offers seminars to industry, particularly in the field of data science. We also have an established process for delivering on-site training courses to companies at rates ranging from 800 to 1200 euros per day. Our business development managers have started to advertise to companies our capability to deliver courses in the field of business data analytics. We expect that this capability will be deployed in 2019-20.

Consultancy engagements are negotiated on a case-by-case basis based on requests for expertise channelled via the University's commercialisation service.

Sustainability: Regarding the sustainability of services, the main weakness is the lack of Estonian-speaking staff, which is necessary to target about 60-70% of the Estonian companies. We are currently forming Estonian-speaking Masters students via our existing course offers (one of them developed in the SoBigData project). We expect these students will become ready to cooperate with us in the delivery of services in 2021.

The University will launch a Masters of Data Science with a professional orientation in 2020. The first graduates will come up in 2022. We expect about 50 graduates per year. As these graduates join companies, we expect to be able to generate leads for consultancy and corporate training services in the field of business data analytics well into the late 2020s.

As for industrial collaboration, UT has on going cooperation with the Estonian Competence Centre in Data Science (STACC). We have engaged in one consultancy agreement and two research and development

projects with them in the past 3 years. We have not yet had a contract directly related to SoBigData services, but STACC's business development team is informed of our capability in the area of social media analytics services and is ready to channel customer leads in this area as they arise.

2 INDUSTRY SUPPORT LETTERS

SEACOM
the leading open source architects

> Email & Collaboration
> Big Data Management

To: Dr. Fosca Giannotti
Coordinator of SoBigData – Research Infrastructure on Social Mining & Big Data Ecosystem
ISTI-CNR – Information Science and Technology Institute of the Italian National Research Council
Via Moruzzi 1
56124 Pisa, Italy

Subject: Support to RI proposal **"SoBigData"**(H2020 Call INFRAIA-01-2018-2019, – Research Infrastructure – Integrating Activities for Advanced Communities)

Dear coordinator of the SoBigData research infrastructure proposal,

Seacom provides consultancy and services in the field of Big Data, with a strong skill on ElasticSearch.

Seacom is highly interested in following the scientific initiative of the SoBigData research infrastructure, whose aim is to create a rich environment for data scientists facilitating the exchange of competencies as well as the access to Big Data and related technologies. We acknowledge the increasingly central role of big data in society and business, and the tremendous skill gap in the rapidly blossoming field of big data analytics. We therefore welcome the SoBigData initiative, that focus on this challenging area of interdisciplinary research at all levels, bringing world leading experts together (on data mining, machine learning, big data, social simulation, complex network analysis, smart cities, ethics, media) into a scientific/educational/communication program of highest rank, that will give access to large scale experiments of social mining and big data analytics to many actors that would be excluded otherwise.

Seacomis interested to interact closely with SoBigData once it will be funded by the EC, and to find ways ofleveraging the research infrastructure, to the aim of exploiting the innovation opportunities that will certainly emerge. To these purpose, Seacomis interested to take part at the activities of the projects that are of common interest, and to explore potential follow-ups and collateral projects.

Seacom is also interested to become member of the SoBigData Association that will be constructed and become active part in the SoBigData community.

In conclusion, Seacom strongly supports the SoBigData proposal and invites the EC to consider this initiative with extreme interest.

Best regards,
Navacchio (PI) 08/02/2019

Seacom s.r.l.
Via Gramsci n° 5
56023 NAVACCHIO (PI)
P. IVA 01310070493

SEACOM Srl - Via Gramsci n.5 56023 Navacchio (PI)
Tel. +39 050 751 9999 - Fax +39 050 751 9998
www.seacom.it
info@seacom.it

STACC
Ülikooli 2, 5th floor
Tartu 51003
Estonia
http://www.stacc.ee/en

To: Dr. Fosca Giannotti
Coordinator of SoBigData – Research Infrastructure on Social Mining & Big Data Ecosystem
ISTI-CNR – Information Science and Technology Institute of the Italian National Research Council
Via Moruzzi 1
56124 Pisa, Italy

Subject: Support to RI proposal "SoBigData++: integrated research infrastructure for Social Mining and Big data Analytics" (H2020 INFRAIA-01-2018-2019 – Research Infrastructure: Integrating Activities for Advanced Communities)

Dear coordinator of the SoBigData research infrastructure proposal,

STACC is the leading data science and machine learning competence center in Estonia. We provide a portfolio of machine learning solutions for companies doing business in the field of e-commerce and online media to enhance business performance and increase competitiveness. In addition, STACC solutions are also contributing towards building up preventive, predictive and participatory health system in Estonia. We also work in the domain of process mining which is becoming widely adopted by the most competitive manufacturing industries, because it enables to collect and analyze the massive amount of big data in manufacturing systems to gain insight into existing business processes, identify problems such as bottlenecks, and find ways to improve overall operational workflow.

STACC has collaborated with researchers in the SoBigData research infrastructure in the period 2015-2019 in the context of the Social Impact datathon held in Tartu in November 2017 and the sTARTUp AI Day 2019. We also maintain research collaboration with SoBigData researchers in the field of predictive analytics.

STACC is interested in following the scientific initiative of the SoBigData research infrastructure, whose aim is to create a rich environment for data scientists facilitating the exchange of competencies as well as the access to Big Data and related technologies. As an active company in this space, we are witnessing the increasingly central role of big data in society and business, and the tremendous skill gap in this field. We welcome the SoBigData initiative. We expect it will give us access to datasets, tools, and expertise to help us further develop our capacity in Big Data analytics.

STACC will continue to interact closely with SoBigData once it will be funded by the EC, and to find ways of leveraging the research infrastructure, to the aim of exploiting the innovation opportunities that will certainly emerge. To these purpose, STACC is interested to take part into the **industrial advisory board of SoBigData**, to take part in activities that are of common interest, and to explore potential follow-ups and collateral industrial projects.

STACC is also interested to become member of the SoBigData Association that will be constructed and become active part in the SoBigData community.

In conclusion, STACC supports the SoBigData proposal and invites the EC to consider this initiative with high interest.

Yours,

Kristjan Eljand
CEO

Page **2** of **2**

11.03.2019

Prof. Dirk Helbing
Project Investigator at ETH, project partner of SoBigData

Subject: Support to RI proposal "SoBigData+: integrated research infrastructure for Social Mining and Big data Analytics"

SIX is a leading securities exchange, financial information and banking services provider mainly in Switzerland.

SIX is highly interested in following the scientific initiative of the SoBigData research infrastructure, whose aim is to create a rich environment for data scientists facilitating the exchange of competencies as well as the access to Big Data and related technologies. In particular, into the research and innovation in finance and cryptocurrency domain.

We acknowledge the increasingly central role of big data in society and business, and the tremendous skill gap in the rapidly blossoming field of big data analytics. We therefore welcome the SoBigData initiative, that focus on this challenging area of interdisciplinary research at all levels, bringing world leading experts together into a scientific and technological program of highest rank, that will give access to large scale experiments of social mining and big data analytics to many actors that would be excluded otherwise.

SIX is interested to **interact** closely with the ETHZ partner of SoBigData once it will be funded by the EC. SIX will help find ways of leveraging the research infrastructure and the ETHZ's expertise on Data Science, cryptocurrency markets and social media analysis to the aim of exploiting the innovation opportunities that will certainly emerge.

In conclusion, SIX strongly supports the SoBigData proposal and invites the EC to consider this initiative with great interest.

Sincerely,

Andreas Sprock
Head Innovation Management

SIX Management AG
Innovation & Digital
Hardturmstrasse 201
CH-8021 Zürich

3 CONCLUSIONS

In this deliverable, we have described the main services that the institutions from SoBigData can provide to the industrial sector. As the majority of the partners on SoBigData are primarily academic institutions, we have tried to adjust the outputs for the industrial sector, going beyond the research community. For each service, we have tried to quantify the strengths and weaknesses of sustainability. The major difficulty lies in the maintenance of the infrastructure and updates of the software products with a limited number of academic personals. This is also the reason why the majority of services are open-sourced or the conditions of use are negotiated only upon request from industrial partners. Finally, we have shown a few examples of support letters from industrial partners to demonstrate applicability for commercial usage.