SOBIGDATA++
PLUS PLUS

Deliverable D8.3

Filling the gaps:
Emerging new analytical technologies 2

SoBigData Plus Plus receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 871042.

DOCUMENT INFORMATION

PROJECT	
PROJECT ACRONYM	SoBigData-PlusPlus
PROJECT TITLE	SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics
STARTING DATE	01/01/2020 (60 months)
ENDING DATE	31/12/2024
PROJECT WEBSITE	http://www.sobigdata.eu
TOPIC	INFRAIA-01-2018-2019 Integrating Activities for Advanced Communities
GRANT AGREEMENT N.	871042
DELIVERABLE INFORMATION	
WORK PACKAGE	WP8 JRA1 - Social Mining and Big Data Resource Integration
WORK PACKAGE LEADER	LUH, CNR
WORK PACKAGE PARTICIPANTS	CNR, USFD, UNIPI, UT, IMT, LUH, SNS, AALTO, ETHZ, CNRS, CEU, URV, BSC, UPF, UvA
DELIVERABLE NUMBER and TITLE	D8.3 Filling the gaps: Emerging new analytical technologies 2
AUTHOR(S)	Giulio Rossetti (CNR), Francesca Pratesi (CNR), Michela Natilli (CNR), Francesca Naretto (UNIPI)
CONTRIBUTOR(S)	
EDITOR(S)	Valerio Grossi (CNR), Michela Natilli (CNR)
REVIEWER(S)	Valerio Grossi (CNR), Michela Natilli (CNR), Roberto Trasarti (CNR), Ilaria Barsanti (CNR)
CONTRACTUAL DELIVERY DATE	30/06/2023
ACTUAL DELIVERY DATE	28/07/2023
VERSION	V1.2
TYPE	Report
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	26
KEYWORDS	Machine learning, AI, complex networks, human mobility

EXECUTIVE SUMMARY

This deliverable gives a complete overview of all research activities related to WP8, particularly on the completed and ongoing methods, tools, and datasets integration activities at Month 42. It provides an up-to-date description of the algorithmic and data resources that are - or are planned to be - integrated within the SoBigData++ research infrastructure. The provided description must be considered an incremental view of the resources available within the SoBigData++ research infrastructure that extends what is already reported in D8.2 – "Social mining services and application integration" and D8.3 – "Filling the gaps: Emerging new analytical technologies 1".

DISCLAIMER

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042.

SoBigData++ strives to deliver a distributed, Pan-European, multi-disciplinary research infrastructure for big social data analytics, coupled with the consolidation of a cross-disciplinary European research community, aimed at using social mining and big data to understand the complexity of our contemporary, globally-interconnected society. SoBigData++ is set to advance on such ambitious tasks thanks to SoBigData, the predecessor project that started this construction in 2015. Becoming an advanced community, SoBigData++ will strengthen its tools and services to empower researchers and innovators through a platform for the design and execution of large-scale social mining experiments.

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 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.

GLOSSARY

EC	European Commission
EU	European Union
H2020	Horizon 2020 EU Framework Programme for Research and Innovation
RI	Research Infrastructure

TABLE OF CONTENTS

1 Relevance to SoBigData++

1.1 Relevance to project objectives

The deliverable outlines a description of the newly integrated methods and datasets within the SoBigData++ research infrastructure and the planned/ongoing integration activities as of M36. The descriptions provided include pointers to related WP3 and WP10 activities and a discussion on the specific level of integration for each of the reported resources.

This document provides an incremental view of WP8 integration activities and an outlook on the future resources that will be made available within the SoBigData RI.

1.2 Relation to other work packages

Work package 8 is part of "social mining research infrastructure building," one of three axes the SoBigData++ work plan comprises. It is therefore strongly connected to the other work packages within the same axis, namely WP9 ("SoBigData e-Infrastructure and supercomputing network") and WP10 ("Exploratories"). They are aimed at building the project core and infrastructure as well as advancing research in social mining.

Additionally, WP8 is connected to work packages in the "community building" axis, such as WP2 ("Critical Data Literacy, Ethics, and Legal Framework"), WP3 ("Dissemination, Impact, and Sustainability"), and WP4 ("Training"), as they go hand in hand with the creation of the platform and infrastructure. Finally, WP8 maintains connections to the work packages in the "user accessibility" axis, WP6 ("Transnational Access") and WP7 ("Virtual Access"), as those dealing with providing access to the integrated resources.

This deliverable is intended to report on the methods/datasets resources that: are made available for exploitation in the connected WPs, have been integrated as a result of related WPs research activities, are outcomes of scientific publications within the consortium.

1.3 Structure of the document

The document is organized into four main sections:

- **Section 2:** describes how research activities are organized within the work package and introduces the different levels of methods/datasets integration within the RI.
- **Section 3:** reports on the methods and tools integrated so far and their relations with WP3 and WP10 activities.
- **Section 4:** reports on the datasets integrated so far within the RI.
- **Section 5:** describes ongoing and planned activities related to integrating analytical methods/datasets within the research infrastructure.

2 WP8 activities: organization and global statistics

WP8 focuses on the integration of algorithmic and data resources within the SoBigData++ research infrastructure. This section briefly describes the two main aspects that underlie the research activities carried out within this work package: micro projects (Subsection 2.1) and resources integration modalities (Subsection 2.2).

2.1 Micro-projects

To better organize the activities related to the integration of algorithmic resources and datasets within the research infrastructure, WP8 leverages the concept of micro-projects. A micro project is a commitment from a partner or more partners of the consortium over a period - typically 1-6 months - to produce a tangible outcome (dataset, method) to be made available to the community through the SoBigData platform. Micro projects allow (i) partners to plan and organize their efforts explicitly, and (ii) task/work package leaders to timely track the ongoing research activities. WP8 micro-projects target the integration of novel algorithmic resources/datasets within the RI and upgrade existing resource functionalities. Since their first introduction in January 2021, 85 micro-projects involving WP8 have been submitted, and 71 have been completed.

2.2 Integration modalities

As previously stated, WP8 micro projects' expected outcomes are methods/datasets to be integrated within the SoBigData RI. As underlined in D8.1 – "Social Data resources and Social media observatory", the integration of datasets will involve creating a dedicated entry in the Catalog.

However, algorithmic resources can be integrated following different modalities within the RI. We devise three different integration levels:

A. Base integration: entry in the SoBigData RI Catalog.
B. Experiment prototyping: integration within the SoBigData Jupyter Hub
C. Engine integration: integration within the SoBigData Method Engine.

The "base integration" is the minimum requirement for a resource to be findable within the RI. It consists of a resource description through a fixed set of metadata and a link to its implementation and documentation.

The "experiment prototyping" integration level makes the algorithmic resource available for live experimental purposes within a SoBigData++ dedicated Jupyter Hub instance. This integration level allows RI users to prototype and execute their experiments using programming libraries developed within the consortium in a standard data science environment. Jupyter Hub integration is a novel feature for the SoBigData++ RI that has been introduced to ease the development of new social mining algorithms and methods. It provides access to a computing cluster that supports 80 concurrent users with 8 GB RAM per Jupyter notebook.

Finally, the "engine integration" level allows RI users to instantiate integrated methods using a visual interface - thus abstracting from code specificity - and run them on a dedicated experiment cluster.

In SoBigData++, we aim to integrate each resource into at least two of the levels mentioned above.

3 Concluded integration activities (Methods and Libraries)

In this section, we report, for each task, the resources integrated within the RI as of M42. For each method/library, a brief description is specified along with the related WP10 exploratory (if any), the integration level, and the main/relevant references (as a link toward dissemination and impact, WP3). Due to the incremental nature of this report, only integrations (reported by the relative task) that happened after M18 are reported.

3.1 Task 8.2: Social media observatory and crowd-sensing design and integration

Method	Type	Description	WP10 Exploratory	Integration Level	Main Reference	Related Papers
twittermonitor	Method	Twitter Monitor Library for JupyterHub	Societal Debates and Misinformation Analysis	A,B		
RetweetCascade	Method	Estimation of the retweet cascade graph	Societal Debates and Misinformation Analysis	A,B	[1]	
Data Inquiries	Other	A proposal for a new way to do research with data and to teach data literacy - shifting the attention away from data and technologies and to the social and political implications of data research	Societal Debates and Misinformation Analysis	A		
Data Sprint / Workshopping for adversarial publics	Other	A participatory design method to discuss and inquire about content moderation and the limits of online free speech	Societal Debates and Misinformation Analysis	A		

3.2 Task 8.3: Text and Social Media Mining services design and integration

Method	Type	Description	WP10 Exploratory	Integration Level	Main Reference	Related Papers
Classification of Wikipedia articles	Method	Identify the human-labeled high-quality articles, e.g., "featured" ones, and differentiate them from the popular and controversial articles.	Societal Debates and Misinformation Analysis	A	[2]	
Diversified approach for social media texts summarization	Method	Python module CLiQS (Cross-Lingual Query-based Summarization), designed for social media messages processing.		A	[3]	

Gate Cloud Annie Named Entity Recognizer	Method	English named entity recognition service. Identifies names of persons, locations, organizations, as well as money amounts, time and date expressions		A,C	[4]	
Gate Cloud Twitie Named Entity Recognizer For Tweets	Method	Named entity recognition service for Twitter data. Identifies person, location, organization etc. and also performs normalization of abbreviations and common shorthands (such as brb, gr8, 2day, etc.)		A,C	[5]	
Gate Cloud Measurement Expression Annotator	Method	Annotates numbers and measurement expressions in text. This method recognises many types of measurements including length, temperature, time and speed, and calculates their normalised values in the SI system of units		A,C	[6]	
Gate Cloud Rumour Veracity Classifier	Method	User generated content such as tweets often make claims that are unsubstantiated and possibly untrue. This service attempts to classify whether a text is discussing a rumour that is likely to be true, likely to be false, or if the rumour is unverified or the classification is unclear	Societal Debates and Misinformation Analysis	A,C	[7]	
Gate Cloud Hyperpartisan News Classifier	Method	Attempts to classify whether a text is hyper-partisan or not. This system won first prize in the Hyperpartisan News Detection task at SemEval 2019	Societal Debates and Misinformation Analysis	A,C	[8]	
Gate Cloud Toxic Language Classifier	Method	Classifies short texts such as Wikipedia comments based on whether they contain toxic language	Societal Debates and Misinformation Analysis	A,C		
Gate Cloud Offensive Language Classifier	Method	Classifies short texts such as Tweets based on whether they contain offensive language	Societal Debates and Misinformation Analysis	A,C		
Gate Cloud Covid19 Vaccine Text Categoriser	Method	A machine learning classifier trained to categorise text about COVID-19 vaccines into 6 categories - Liberty or Freedom; Development, Provision and Access; Safety, Efficacy and Necessity; Politics and Economics; Conspiracy; Morality, Religiosity and Ethics. The service returns the top two categories for each input text		A,C	[9]	
Gate Cloud Covid19 Misinformation Classifier	Method	A machine learning classifier trained to categorise claims about COVID-19 into 10 categories proposed by the Reuters Institute for the Study of Journalism - Public authority	Societal Debates and Misinformation Analysis	A,C	[10]	

		actions, policy, and communications; Community spread and impact; Medical advice and self-treatments; Claims about prominent actors; Conspiracy theories; Virus transmission; Virus origin and properties; Public preparedness, protests, and civil disobedience; Vaccines, medical treatments, and tests; Other. The service returns the most likely predicted class for each text				
Gate Cloud Brexit Tweet Analysis	Method	A pipeline designed to detect political topics, hashtags, URLs, user mention, and hashtag-based voting intentions, expressed in tweets about the UK referendum on membership of the EU	Societal Debates and Misinformation Analysis	A,C	[11]	
Gate Cloud Political Abuse Monitor	Method	A service that tags abusive utterances in any text with the type of abuse (sexist, racist, etc.) and whether the abuse was aimed at the addressee or some other party. This can be run on any English language text. You can check also which words or phrases were deemed potentially abusive via the SlurLookup, SensitiveLookup and OffensiveLookup output options. It will also tag UK members of parliament for the 2015, 2017 and 2019 general elections, candidates for the 2017 and 2019 elections, and members of the Irish Dail Eireann from 20th February 2020	Societal Debates and Misinformation Analysis	A,C	[12]	
Gate Cloud Hate And Abuse Detection	Method	A service that tags abusive utterances in any text with the type of abuse (sexist, racist, etc.) and whether the abuse was aimed at the addressee or some other party. This can be run on any English language text.	Societal Debates and Misinformation Analysis	A,C		
Ariadne English Archaeology Named Entity Recognizer	Method	Identifies terms and phrases in English for analysing archaeological text. The method delivers named entities of archaeological context, physical object, material, time appellation and structure, linked to concept labels of the National Cultural Heritage Thesauri (UK). This method was supplied by the Ariadne Infrastructure that integrates archaeological research data across Europe		A,C	[13]	
Ariadne Dutch Archaeology Named Entity Recognizer	Method	Identifies terms and phrases in Dutch for analysing archaeological text. The method delivers named entities of archaeological context,		A,C	[14]	

		physical object, material, time appellation and structure, linked to concept labels of the Rijksdienst vor het Cultureel Erfgoed (RCE) Thesauri (NL). This method was supplied by the Ariadne Infrastructure that integrates archaeological research data across Europe				
Ariadne Swedish Archaeology Named Entity Recognizer	Method	Identifies terms and phrases in Swedish for analysing archaeological text. The method delivers named entities of archaeological context, physical object, material, time appellation and structure. This method was supplied by the Ariadne Infrastructure that integrates archaeological research data across Europe		A,C	[15]	
Ariadne English Dendrochronology Entity Recognizer	Method	Identifies terms and phrases in English for analysing archaeological text. The method delivers named entities of archaeological elements, wood material, sample, and date, with all apart from dates linked to concept labels of the AAT Getty thesaurus. This method was supplied by the Ariadne Infrastructure that integrates archaeological research data across Europe		A,C	[16]	
Ariadne Dutch Dendrochronology Entity Recognizer	Method	Identifies terms and phrases in Dutch for analysing archaeological text. The method delivers named entities of archaeological elements, wood material, sample, and date, with all apart from dates linked to concept labels of the AAT Getty thesaurus. This method was supplied by the Ariadne Infrastructure that integrates archaeological research data across Europe		A,C	[17]	
Ariadne Swedish Dendrochronology Entity Recognizer	Method	Identifies terms and phrases in Swedish for analysing archaeological text. The method delivers named entities of archaeological elements, wood material, sample, and date, with all apart from dates linked to concept labels of the AAT Getty thesaurus. This method was supplied by the Ariadne Infrastructure that integrates archaeological research data across Europe		A,C	[18]	

Gate Cloud Chemical Entity Recognizer	Method	This service annotates chemical named entities using the open source OSCAR4 tagger. As well as the names of the detected entities the tagger also returns their structure in either the SMILES or InChI nomenclature		A,C	[19]	
Gate Cloud Multilingual News Similarity Service	Method	Service that takes two news article texts and attempts to determine whether they refer to the same underlying news story. The articles may be in the same or in different languages, and may optionally have been pre-tagged with named entities; if named entity annotations are present these are used as additional evidence when calculating the similarity score		A,C	[20]	
Gate Cloud Stance Classification English	Method	Service that processes threads of short English-language texts such as tweets or forum postings and attempts to determine the stance (support, deny, question or comment) of each reply towards the post it is replying to.	Societal Debates and Misinformation Analysis	A,C	[21]	
Gate Cloud Stance Classification Multilingual	Method	Service that processes threads of short texts such as tweets or forum postings and attempts to determine the stance (support, deny, question or comment) of each reply towards the post it is replying to. This method uses a multilingual BERT-based model and is able to classify threads where the reply and target posts are in the same or different languages	Societal Debates and Misinformation Analysis	A,C	[22]	
Gate Cloud Url Domain Analysis	Method	Service that takes a list of URLs and assigns to each information on what multiple organisations who analyse the credibility of online content have said about the domain (or sometimes, domain path) in the URL. Input is a CSV file with a column of URLs, and output is another CSV with a row for each item of credibility information that is available from the various sources. This may be several rows for a single URL if it has been rated by different organisations.	Societal Debates and Misinformation Analysis	A,C		

3.3 Task 8.4: Complex Network Analysis Mining services design and integration

Method	Type	Description	WP10 Exploratory	Integration Level	Main Reference	Related Papers
Generalized Network Dismantling	Method	Implements network dismantling problem: dismantle into isolated subcomponents, thereby disrupting the malfunctioning of a system or containing the spread of misinformation or an epidemic		A,C	[23]	
Human-Bot hybrid game	Method	Human-Bot coordination game played on a virtual network		A	[24]	[25]
Factor analysis methods for daily rhythms	Method	Factoring and classification of chronotype using geolocated communication data		A	[26]	
Tie-stabilty prediction using clustering	Method	Clustering methods to study the effect of long-distance residential move within the country on mobile phone communication	Migration Studies	A	[27]	
Fractal-network generator	Method	Genrating adjacency for regular fractal-like networks		A,B	[28]	
structify-net	Method	Structify_net is a python library allowing to create networks with a predefined structure, and a chosen number of nodes and links.		A,B	[29]	
Visualizing community detection in networks	Other	A series of experimental techniques to visualize relational communities and their blurred boundaries in networks		A		
Siblinarity antichains	Method	Implements community detection method for finding antichains (integrated within CDlib)		A,B	[30]	
Simulator SEIR dynamics on networks and recovering the true state of the epidemics from noisy biased samples	Method	Code for simulating SEIR dynamics on networks and recovering the true state of the epidemics from noisy biased samples	Societal Debates and Misinformation Analysis	A,C	[31]	
Stripe-Corrected Gravity Model	Method	The code implements the Stripe-Corrected Gravity Model, i.e. an extension of the Density-Corrected Gravity Model for layered, economic networks.	Economy & Finance 2.0	A	[32]	
Reciprocal Binary Configuration Model	Method	This code implements the Reciprocal Binary Configuration Model to be used for detecting binary, triadic network motifs.	Economy & Finance 2.0	A	[33]	[34]

Conditionally Reciprocal Weighted Configuration Model	Method	This code implements the Conditionally Reciprocal Weighted Configuration Model to be used for detecting weighted, triadic network motifs.	Economy & Finance 2.0	A	[35]	
Bipartite Configuration Model (binary and weighted)	Method	This code implements the Bipartite Configuration Model, both binary and weighted, on both Python ad R	Economy & Finance 2.0	A	[36]	
NetworkSNS	Method	Python library for: (I) the inference, simulation, and forecasting of several models of temporal networks; (ii) the computation of centrality measures in static and temporal networks and multiplexes. The library also includes the case of non-instantaneous link travel time, such as in transportation networks and multiplexes.		A,B	[37-42]	[43, 44]
conformity	Method	Multiscale feature-rich network assortativity measure		A,B	[45]	
Delta-conformity	Method	Time-aware multiscale feature-rich network assortativity measure		A,B	[46]	
XMark	Method	Synthetic network generator for benchmarking attributed community detection algorithms (integrated within CDlib)		A,B	[47]	

3.4 Task 8.5: Human Mobility Analytics services design and integration

Method	Type	Description	WP10 Exploratory	Integration Level	Main Reference	Related Papers
TraffiCO2	Method	Implementation of a simulation framework for assessing the impact of navigation services on CO2 emissions	Sustainable Cities for Citizens	A	[48]	
MoGAN	Method	AI model to generate synthetic od matrices	Sustainable Cities for Citizens	A	[49]	

3.5 Task 8.7: Visual Analytics services design and integration

Method	Type	Description	WP10 Exploratory	Integration Level	Main Reference	Related Papers
Visual media analysis for Instagram and other online platforms	Other	Collection of tools and techniques for visual analytics for social media research	Societal Debates and Misinformation Analysis	A	[50]	[51,52]
Gate Cloud Multilingual Image OCR	Method	Service that uses optical character recognition (OCR) to identify text contained within images. This is a multi-lingual service and not restricted to Latin scripts. It works in three stages. First it determines the bounding boxes of related text within the image. Secondly it extracts the text from within each bounding box, before finally determining the language of the extracted text. Two output CSV files are returned, one with one row per image giving the primary detected script for that image, and another with one row per bounding box giving the extracted text	Societal Debates and Misinformation Analysis	A,C		
How to Use Visual Media Analysis for Social Media Research	Other	How to guide to perform visual media analysis on social media platforms		A	[53]	
Imagetrendline	Method	A Python based tool for plotting images from urls for metric of choice and over time		C	[54]	
What is a meme technically speaking?	Other	Collection of techniques to create meme collection and how to study them, depending on the website of platform where they are sourced.		A	[55]	
How to Make Meme Collections	Other	How to guide to create meme collections and analyse them depending on the software environment where memes were sourced		A	[56-58]	

3.6 Task 8.8: Privacy Enhancing Technology and Discrimination preventing services design and integration

Method	Type	Description	WP10 Exploratory	Integration Level	Main Reference	Related Papers
k-anonymity via microaggregation	Method	k-anonymity is a method to protect the privacy of individuals in a dataset while it is preserved the utility of the anonymized data. Microaggregation is a natural approach to satisfy k-anonymity. Microaggregation consist of two steps: i) partition of data into clusters and ii) aggregation of the values of each cluster		A	[59]	
t-closeness through microaggregation	Method	t-closeness improves k-anonymity protecting the dataset against attribute disclosure (attribute disclosure occurs if the confidential attribute is too similar for all k individuals in a cluster). This method uses microaggregation to generate k-anonymous and t-close data sets.		A	[60]	
Differential privacy via individual ranking	Method	Differential privacy offers more robust privacy guarantees than k-anonymity and its extensions, at the cost of the utility of the anonymized data. To preserve the utility of the protected data, this method builds on microaggregation applied to each individual attribute. In this way, it is reduced the amount of noise needed to satisfy differential privacy		A	[61]	
Disclosure risk assessment via record linkage	Method	The performance of a protection method is measured in terms of the preserved utility and the disclosure risk. Record linkage is a realistic and practical disclosure risk assessment methodology applicable to every masking method.		A	[62]	[63]
Discrimination prevention method	Method	Discrimination consists of unfair treatment of people on the basis of their belonging to a specific group. Discrimination can be		A	[64]	

		direct or indirect if decisions are based, respectively, on sensitive attributes or non-sensitive attributes strongly correlated with biased sensitive attributes. Antidiscrimination techniques include discrimination discovery and prevention. This method evaluates and treats the data set removing direct and/or indirect discrimination biases preserving data quality.				

3.7 Task 8.9: Explainable AI services design and integration

Method	Type	Description	WP10 Exploratory	Integration Level	Main Reference	Related Papers
Focus Metric	Method	Metric to evaluate explainability methods and quantify their coherency to the task the model is solving.	Societal Debates and Misinformation Analysis	A	[65]	
XAILib	Method	A suite of methods for explainable AI for different Ai models		A,B		[66-69]

3.8 Task 8.10: Scalable machine learning services design and integration

Method	Type	Description	WP10 Exploratory	Integration Level	Main Reference	Related Papers
ALIR	Method	Asynchronous Training of Word Embeddings for Large Text Corpora		A,C	[70]	
Fast-Forward indexes	Method	A python library for efficient neural ranking		A,B	[71]	
Full Network Embedding	Method	Extract features from a pre-trained CNN for posterior use on other models. May be integrated with SVM for classification.		A	[72]	[73]

4 Concluded integration activities (Datasets)

In this section, we report, for each task, the datasets integrated within the RI as of M42. For each dataset, a brief description is specified along with the related WP10 exploratory (if any). Due to the incremental nature of this report, only integrations that happened after M18 are reported.

Dataset	Description	WP10 Exploratory
Crypto related tweets from 10.10.2020 to 3.3.2021	Dataset covering crypto currency related tweets	Economy & Finance 2.0
Bitcoin market microstructure data from 11-2017 until 02-2018	Dataset covering bitcoin market microstructure	Economy & Finance 2.0
MAMe	Image classification dataset with remarkable high resolution and variable shape properties.	Others
Academic Migration and Academic Networks	Datasets used and produced for and from the micro project titled: Academic Migration and Academic Networks: Evidence from Scholarly Big Data and the Iron Curtain"	Migration Studies
360° video footage from a bottom-up street-view survey in Pristina, Kosovo	This dataset contains 360° video footage from a bottom-up street-view survey in Pristina, Kosovo divided into major neighborhoods	City Of Citizen
Physical activity, quality of sleep, and quality of life in Italy: the long tail effect of the COVID-19 lockdown	This dataset could permit to evaluate the long tail effect of COVID-19 lockdown in Italy comparing the respondents' habits and well-being status of respondents in the months before (November 2019 and January 2020) and in the months long-after (November 2021 and January 2022) the confinement periods.	Economy & Finance 2.0
t Twitter dataset about two premier UK music festivals	The dataset contains twitter posts about two premier UK music festivals: Creamfields 2016 (on August 25th-28th) and VFestival 2016 (on August 20th-21st)	City Of Citizens
Conversational search dataset with labels	CAsT 2019 conversations from training set and from test set without qrel + ConvQ dataset. **Labels** - SE: classification label for utterances that are Self Explanatory (e.g., they do not need any rewriting) - FT: classification label for utterances referring to the First Topic in the conversation - PT: classification label for utterances referring to a Previous Topic in the conversation (different from the first topic)	Others
BioTAGME: A comprehensive platform for biological knowledge network analysis	This Network was built through BioTAGME, a system that combines TAGME, an entity-annotation framework based on Wikipedia corpus with a network-based inference methodology (i.e., DT-Hybrid).	Network Medicine
Thyroid-cancer patients	The data used originate from the web-based database of the Italian Thyroid Cancer Observatory (ITCO), opened in 2013 at the Thyroid Cancer Center of the Sapienza University of Rome and including 49 other thyroid cancer centers in Italy. The data collected on more than 10000 patients contain demographic and biometric information, circumstances of diagnosis, tumor pathology, surgical and radioactive iodine treatments, and results of periodic follow-up examinations.	Network Medicine
A large datast of treatments of type-2 diabetes patients		Network Medicine
Fake Twitter Accounts Dataset		Societal Debates
Covid infodemic in Italy -- Most retweeted accounts	Top 10 most retweeted accounts on Covid-related keywords, between Jan 30 and Mar 20, 2020.	Societal Debates
Twitter Dataset British MPs	This dataset contains the Twitter tweet_ids from the Timelines of 584 members of British Parliament (collected between 4th and 6th of March 2022).	Societal Debates

A dataset of journalists on Twitter	This dataset comprises the Twitter timelines of journalists belonging to 17 different countries from 8 different continental regions, downloaded in May 2018.	Societal Debates
A dataset of gamers on Twitter	This gaming-related dataset consists of 8932 users (labeled as gamers) engaging in game-related conversations. We have collected (June 2018) their timeline (the most recent 3200 tweets) using the Twitter Search API.	Others
Activity data from the Covid19 period	Activity data from Telia telecommunications company, Finland reports the number of people dwelling in area for a certain amount of time. More precisely, activity count calculates the number of unique persons with an activity per spatial and temporal unit. Here, the spatial unit is one Telia scalable grid cell, and time unit is a day.	Migration Studies
Learning to quantify: LeQua 2022 datasets	The aim of LeQua 2022 (the 1st edition of the CLEF "Learning to Quantify" lab) is to allow the comparative evaluation of methods for "learning to quantify" in textual datasets, i.e., methods for training predictors of the relative frequencies of the classes of interest in sets of unlabelled textual documents.	Other
Product Reviews for Ordinal Quantification	This data set comprises a labeled training set, validation samples, and testing samples for ordinal quantification. The data is extracted from the McAuley data set of product reviews in Amazon, where the goal is to predict the 5-star rating of each textual review.	Other
Cherenkov Telescope Data for Ordinal Quantification	This labeled data set is targeted at ordinal quantification.	Other
Ukraine-related Disinformation Dataset	It contains the tweet ids of all tweets containing disinformation and debunk links and of all the Ukraine-related debunk links.	Societal Debates
VaxxHesitancy: A Dataset for Studying Hesitancy Towards COVID-19 Vaccination on Twitter	Dataset of over 3,100 COVID-19 vaccine-related tweets labeled as one of four stance categories: pro-vaxx, anti-vaxx, vaxx-hesitant, or irrelevant.	Others
Multi-aspect Integrated Migration Indicators (MIMI) dataset	MIMI includes both official data about bidirectional human migration (traditional flow and stock data) with multidisciplinary variables and original indicators, including economic, demographic, cultural and geographic indicators, together with the Facebook Social Connectedness Index (SCI).	Migration Studies
Ego Networks of Words in Twitter	Anonymized ego-networks of four disjoint twitter populations: Journalists from the NYT, Science writers, Random users #1, Random users #2.	Other
Cross-Lingual Dataset of Crisis-Related Social Media		Other
Dataset for Evaluating Abstractive Summaries of Crisis-Related Social Media	The dataset created for evaluation of summaries generated from social media posted during five natural disasters. The dataset contains: ground truth reports created by human assessor based on ERCC Echo Flash reports (5 events); summaries generated by extractive and abstractive state-of-the-art with manual annotation of category-relevant and crisis-relevant claim.	Other
Interaction bias	Artificial Intelligence (AI) is increasingly used to build Decision Support Systems (DSS) across many domains. In our work, we conducted a series of experiments designed to observe human response to different characteristics of a DSS such as accuracy and bias, particularly the extent to which participants rely on the DSS, and the performance they achieve.	Explainable Machine Learning
UNI Fake Giveaway Dataset	Dataset related to a scam originated on Twitter that lured users into sending their Uniswap (UNI) tokens to a fake giveaway.	Economy & Finance 2.0

5 Conclusions - Ongoing/Planned integration activities

This deliverable reported the concluded methods and datasets integration activities involving the SoBigData RI. The list of available algorithmic resources and datasets is continuously updated throughout the project's lifetime. We conclude reporting the resources whose integration is ongoing/planned. For each resource, it is specified its nature (either method or dataset), a brief description, and the main/relevant references (as a link toward dissemination and impact, WP3).

Ongoing/Planned integrations are differentiated on the project they will be allocated to (namely, either SoBigData++ or SoBigData.it[1]).

5.1 SoBigData++

Microprojects	Type	Description
privlib	Method	Python software package to manage privacy risk and discrimination in tabular and sequential data. It comprises methods to assess privacy risk ad discrimination, anonymize data according to various algorithms and mitigate discrimination
Leveraging Ego Network Layers for Link Prediction	Method	We propose to leverage well-established social cognitive theories to improve link prediction performance. According to these theories, individuals arrange their social relationships along, on average, five concentric circles of decreasing intimacy. We postulate that relationships in different circles have different importance in predicting new links. To validate this claim, we focus on popular feature-extraction prediction algorithms (both unsupervised and supervised) and we extend them to include social-circles awareness. We validate the prediction performance of these circle-aware algorithms against several benchmarks (including their baseline versions as well as node-embedding- and GNN-based link prediction), leveraging two Twitter datasets comprising a community of video gamers and generic users. We show that social awareness generally provides significant improvements in the prediction performance, beating also state-of-the-art solutions like node2vec and SEAL, without increasing the computational complexity.
Risk forecasting with reservoir computers	Method	In the framework of recurrent neural networks, reservoir computing has been developed on the paradigm of training a simple readout mechanism, while reading the state of the recurrent neurons. The layer of the recurrent neurons defines the so-called reservoir, whose internal weights are fixed and randomly generated. Training only the readout layer of a reservoir computer is a remarkable simplification in terms of computational complexity. Nevertheless, the approximation properties are preserved for a large class of dynamic processes, once a few hyper-parameters are optimized. Interestingly, reservoir computers as approximants do not rely necessarily on the specific choice of mean squared error as a loss function for the readout training. In fact, they can be devised to predict any moment of a dynamic process, in particular any quantile of the conditional distribution of a random process. To this end, the readout can be trained with the quantile loss function introduced for quantile regression. A forecasting study of the Value at Risk of several stocks for the US financial market shows the merit of the devised methodology.

[1] SoBigData.it is supported by European Union – NextGenerationEU – National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) – Project: "SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics" – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021.

Automatic Generation of Explanation Dashboards	Method	Creation of a dashboard that interacts with the current version of the XAI Library and beautifies the returned explanations and makes them interactive when possible.
Build LORE Yourself	Method	Re-implementation of the LORE explainer in a modular way such that users can personalize how it works, returns explanations, the surrogate model adopted, and the data processing used.
Social AI gossiping simulator	Method	n this micro-project, we focus on the design and development of a simulator for accomplishing a decentralized learning task on a social network and we study what models emerge from the combination of local models, where combination considers the social relationships between heterogeneous nodes (humans associated with the AI).
ranking-utils: A library for deep neural ranking models	Method	Python library for training deep neural ranking models. The library will support dataset parsing and pre-processing as well as training, validating and testing ranking models. The training procedure will integrate seamlessly with PyTorch and PyTorch Lightning.
Diversified approach for social media texts summarization	Method	The result of the project is the Python module CLiQS (Cross-Lingual Query-based Summarization), which can be used in further projects for processing social media messages.
Uncovering the Behavior of New Twitter Accounts through Early Detection and Monitoring	Dataset	There is a significant body of literature concerning the analysis of Twitter accounts, yet the behavior of newly created accounts remains relatively unexplored. In this context, we developed a novel technique to detect new Twitter accounts right after registration, which allowed us to identify and monitor over 500,000 new accounts. In this microproject, we aim to share the results of the analysis of this dataset with the SoBigData++ community. Each account was monitored for 21 days by sampling profile information and timelines at scheduled intervals, retrieving over 8 million tweets. An additional sample of profile information was collected approximately two years after creation, in May 2022.

5.2 SoBigData.it

Microprojects	Type	Description
MANILA	Application	Low-Code Web Application for the Quality-Based Development of Machine Learning Systems
ASH	Method	Attributed Stream Hypergraph python library
PyGM	Method	Library of index data structures enabling fast searches in arrays of billions of items using orders of magnitude less space than traditional indexes.
NetME	Application	NetME can create biological networks on-the-fly based on user queries.
OntoTagME	Application	OntoTagME is an entity linker which specializes TAGME onto the biological field, by deploying a different knowledge base that comprehends a custom-made biological filtering of Wikidata.
CDlib v0.3.0	Method	Community Detection library major release
DyNetX v0.3.2	Method	Dynamic Network Modeling library major update

GlocalX	Method	Python library to explain machine learning models by hierarchically aggregating single explanations of its predictions.
TRIPLEX	Method	TRIPLEX is an explainability package for Transformer-based models fine-tuned on Natural Language Inference, Semantic Text Similarity, or Text Classification tasks.
Angel	Method	Integration in the Method Engine of the Angel overlapping Community Detection algorithm

References

[1] Zola P, Cola G, Mazza M, Tesconi M. Interaction Strength Analysis to Model Retweet Cascade Graphs. *Applied Sciences*. 2020; 10(23):8394. https://doi.org/10.3390/app10238394

[2] Ogushi F., Kertész J., Kaski, K. K. et al. Ecology of the digital world of Wikipedia. Sci Rep 11, 18371 (2021). https://doi.org/10.1038/s41598-021-97755-w

[3] Vitiugin, Fedor, and Carlos Castillo. "Cross-Lingual Query-Based Summarization of Crisis-Related Social Media: An Abstractive Approach Using Transformers." Proceedings of the 33rd ACM Conference on Hypertext and Social Media. 2022.

[4] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[5] K. Bontcheva, L. Derczynski, A. Funk, M.A. Greenwood, D. Maynard, N. Aswani. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013).

[6] M. Agatonovic, N. Aswani, K. Bontcheva, H. Cunningham, T. Heitz, Y. Li, I. Roberts and V. Tablan. Large-scale, Parallel Automatic Patent Annotation. In Proceedings of 1st International CIKM Workshop on Patent Information Retrieval - PaIR'08, Napa Valley, California, USA, 30 October, 2008.

[7] Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can Rumour Stance Alone Predict Veracity?. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3360–3370, Santa Fe, New Mexico, USA. Association for Computational Linguistics

[8] Y.Jiang, J. Petrak, X. Song, K. Bontcheva, D. Maynard. Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network . Proc. of SemEval 2019 - the International Workshop on Semantic Evaluation, 6-7 June 2019, Minneapolis, US.

[9] Y. Li, C. Scarton, X. Song, K. Bontcheva. Classifying COVID-19 vaccine narratives. arXiv pre-print, DOI 10.48550/arXiv.2207.08522

[10] X. Song, J. Petrak, Y. Jiang, I. Singh, D. Maynard, K. Bontcheva. "Classification aware neural topic model for COVID-19 disinformation categorisation." PloS one 16, no. 2 (2021): e0247086.

[11] D. Maynard, I. Roberts, M. A. Greenwood, D. Rout and K. Bontcheva. A Framework for Real-time Semantic Social Media Analysis. Web Semantics: Science, Services and Agents on the World Wide Web, 2017 (in press). https://doi.org/10.1016/j.websem.2017.05.002

[12] Gorrell, G., Bakir, M. E., Roberts, I., Greenwood, M. A., & Bontcheva, K. (2020). Which politicians receive abuse? Four factors illuminated in the UK general election 2019. EPJ Data Science, 9(1), 18

[13] Vlachidis, A., & Tudhope, D. (2016). A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain. Journal of the Association for Information Science and Technology, 67 (5), 1138-1152. doi:10.1002/asi.23485

[14] Vlachidis, A., Tudhope, D., & Wansleeben, M. (2021). Knowledge-Based Named Entity Recognition of Archaeological Concepts in Dutch. Metadata and Semantic Research. MTSR 2020. Communications in Computer and Information Science, 1355. Cham, Switzerland: Springer Verlag. doi:10.1007/978-3-030-71903-6_6

[15] Binding, C., Tudhope, D., & Vlachidis, A. (2019). A study of semantic integration across archaeological data and reports in different languages. Journal of Information Science, 45 (3), 364-386. doi:10.1177/0165551518789874

[16] Binding, C., Tudhope, D., & Vlachidis, A. (2019). A study of semantic integration across archaeological data and reports in different languages. Journal of Information Science, 45 (3), 364-386. doi:10.1177/0165551518789874

[17] Vlachidis, A., & Tudhope, D. (2022). A Method for Archaeological and Dendrochronological Concept Annotation using Domain Knowledge in Information Extraction. International Journal of Metadata, Semantics and Ontologies, 192-203. doi:10.1504/IJMSO.2021.123042

[18] Binding, C., Tudhope, D., & Vlachidis, A. (2019). A study of semantic integration across archaeological data and reports in different languages. Journal of Information Science, 45 (3), 364-386. doi:10.1177/0165551518789874

[19] Jessop, D.M., Adams, S.E., Willighagen, E.L. et al. OSCAR4: a flexible architecture for chemical text-mining. J Cheminform 3, 41 (2011). https://doi.org/10.1186/1758-2946-3-41

[20] Iknoor Singh, Yue Li, Melissa Thong, and Carolina Scarton. 2022. GateNLP-UShef at SemEval-2022 Task 8: Entity-Enriched Siamese Transformer for Multilingual News Article Similarity. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1121–1128, Seattle, United States. Association for Computational Linguistics.

[21] Li, Y., & Scarton, C. (2023). Evaluating the Role of Target Arguments in Rumour Stance Classification. arXiv preprint arXiv:2303.12665.

[22] Li, Y., & Scarton, C. (2023). Evaluating the Role of Target Arguments in Rumour Stance Classification. arXiv preprint arXiv:2303.12665.

[23] Ren, X. L., Gleinig, N., Helbing, D., & Antulov-Fantulin, N. (2019). Generalized network dismantling. *Proceedings of the national academy of sciences*, *116*(14), 6554-6559.
[24] Bhattacharya, Kunal, et al. "Group formation on a small-world: experiment and modelling." Journal of the Royal Society Interface 16.156 (2019): 20180814.
[25] Takko, T., Bhattacharya, K., Monsivais, D., & Kaski, K. (2021). Human-agent coordination in a group formation game. Scientific Reports, 11(1), 1-10.
[26] Roy, C., Monsivais, D., Bhattacharya, K., Dunbar, R. I., & Kaski, K. (2021). Morningness-eveningness assessment from mobile phone communication analysis. Sci Rep 11, 14606 (2021). https://doi.org/10.1038/s41598-021-93799-0
[27] Fudolig, M. I. D., Monsivais, D., Bhattacharya, K., Jo, H. H., & Kaski, K. (2021). Internal migration and mobile communication patterns among pairs with strong ties. EPJ Data Science, 10(1), 1-21.
[28] Monsivais-Velazquez, D., Bhattacharya, K., Barrio, R. A., Maini, P. K., & Kaski, K. K. (2020). Dynamics of hierarchical weighted networks of van der Pol oscillators. Chaos: An Interdisciplinary Journal of Nonlinear Science, 30(12), 123146.
[29] Cazabet, R., Citraro, S., & Rossetti, G. (2023). Structify-Net: Random Graph generation with controlled size and customized structure. *arXiv preprint arXiv:2306.05274*.
[30] Vasiliauskaite, V., Evans, T.S. Making communities show respect for order. *Appl Netw Sci* **5**, 15 (2020). https://doi.org/10.1007/s41109-020-00255-5
[31] Vasiliauskaite, V., Antulov-Fantulin, N., & Helbing, D. (2022). On some fundamental challenges in monitoring epidemics. *Philosophical Transactions of the Royal Society A*, *380*(2214), 20210117.
[32] Ialongo, L. N., de Valk, C., Marchese, E., Jansen, F., Zmarrou, H., Squartini, T., & Garlaschelli, D. (2022). Reconstructing firm-level interactions in the Dutch input–output network from production constraints. *Scientific Reports*, *12*(1), 1-12.
[33] Squartini, T., & Garlaschelli, D. (2011). Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics*, *13*(8), 083001.
[34] Di Vece, M., Pijpers, F. P., & Garlaschelli, D. (2023). Commodity-specific triads in the Dutch inter-industry production network. *arXiv preprint arXiv:2305.12179*.
[35] Di Vece, M., Pijpers, F. P., & Garlaschelli, D. (2023). Commodity-specific triads in the Dutch inter-industry production network. *arXiv preprint arXiv:2305.12179*.
[36] Caruso, T., Clemente, G. V., Rillig, M. C., & Garlaschelli, D. (2022). Fluctuating ecological networks: A synthesis of maximum-entropy approaches for pattern detection and process inference. *Methods in Ecology and Evolution*, *13*(11), 2306-2317.
[37] Oliver E Williams, Fabrizio Lillo and Vito Latora. Effects of memory on spreading processes in non-Markovian temporal networks. New Journal of Physics, Volume 21, https://iopscience.iop.org/article/10.1088/1367-2630/ab13fb/meta.
[38] Oliver E. Williams, Fabrizio Lillo and Vito Latora. How auto- and cross-correlations in link dynamics influence diffusion in non-Markovian temporal networks.https://arxiv.org/abs/1909.08134v1
[39] Michele Benzi, Paola Boito (2020), Matrix functions in network analysis, GAMM-Mitteilungen, Volume 43, Issue 3, https://onlinelibrary.wiley.com/doi/abs/10.1002/gamm.202000012.
[40] Michele Benzi, Christine Klymko (2013) Total communicability as a centrality measure, Journal of Complex Networks, Volume 1, Issue 2, Pages 124–149, https://doi.org/10.1093/comnet/cnt007.
[41] Silvia Zaoli, Piero Mazzarisi & Fabrizio Lillo (2019), Trip Centrality: walking on a temporal multiplex with non-instantaneous link travel time, scientific Reports, 9, Article number: 10570, https://www.nature.com/articles/s41598-019-47115-6.
[42] Silvia Zaoli, Piero Mazzarisi & Fabrizio Lillo (2021), Betweenness centrality for temporal multiplexes, Scientific Reports volume 11, Article number: 4919, https://www.nature.com/articles/s41598-021-84418-z.
[43] P. Mazzarisi, P.Barucca, F.Lillo, D.Tantari. A dynamic network model with persistent links and node-specific latent variables, with an application to the interbank market. European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2019.07.024.
[44]Michele Benzi, Isabel Chen, Howard H. Chang, Vicki S. Hertzberg (2017), Dynamic communicability and epidemic spread: a case study on an empirical dynamic contact network, Journal of Complex Networks, Volume 5, Issue 2, https://doi.org/10.1093/comnet/cnw017.
[45] Rossetti, G., Citraro, S., & Milli, L. (2021). Conformity: a path-aware homophily measure for node-attributed networks. *IEEE Intelligent Systems*, *36*(1), 25-34.
[46] Citraro, S., Milli, L., Cazabet, R., & Rossetti, G. (2022). Δ-Conformity: multi-scale node assortativity in feature-rich stream graphs. *International Journal of Data Science and Analytics*, 1-12.
[47] Citraro, S., & Rossetti, G. (2021). X-Mark: A benchmark for node-attributed community discovery algorithms. *Social Network Analysis and Mining*, *11*(1), 99.
[48] Cornacchia, G., Böhm, M., Mauro, G., Nanni, M., Pedreschi, D., & Pappalardo, L. (2022, November). How routing strategies impact urban emissions. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems* (pp. 1-4).
[49] Mauro, G., Luca, M., Longa, A. *et al.* Generating mobility networks with generative adversarial networks. *EPJ Data Sci.* **11**, 58 (2022). https://doi.org/10.1140/epjds/s13688-022-00372-4

[50] Rogers, R. (2021). Visual media analysis for Instagram and other online platforms. Big Data & Society, 8(1). https://doi.org/10.1177/20539517211022370
[51] Bainotti, L. & Rogers, R. (2022). How to Use Visual Media Analysis for Social Media Research. SAGE Research Methods: Doing Research Online. https://doi.org/10.4135/9781529608687
[52] Bainotti, L. & Rogers, R. (2022). Visual Media ; Analysis for Online Discourses: The Case Study of Deepfakes [Video]. SAGE Research Methods Doing Research Online. https://doi.org/10.4135/9781529608380
[53] Bainotti, L., & Rogers, R., (2022). *How to Use Visual Media Analysis for Social Media Research* [How-to Guide]. Sage Research Methods: Doing Research Online. https://doi.org/10.4135/9781529608687
[54] Bainotti, L. & Rogers, R. (2022). Visual Media Analysis for Online Discourses: The Case Study of Deepfakes [Video]. SAGE Research Methods Doing Research Online. https://doi.org/10.4135/9781529608380
[55] Rogers, R. & Giorgi, G. (2023). What is a meme, technically speaking? in Information, Communication & Society. https://doi.org/10.1080/1369118X.2023.2174790
[56] Giorgi, G., Rogers, R., & Omena, J.J. (2023). How to Make Meme Collections, in Sage Research Methods: Doing Research Online. https://doi.org/10.4135/9781529611267
[57] Giorgi, G., Rogers, R. & Omena, J.J. (2023). Making meme collections [Video]. SAGE Research Methods: Doing Research Online. https://doi.org/10.4135/9781529627732
[58] Rogers, R. & Omena, J.J (2023). Analysing Meme Collections with the Computer Vision Network Approach [video]. SAGE Research Methods Doing Research Online. https://doi.org/10.4135/9781529627756
[59] Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, *11*, 195-212.
[60] Soria-Comas, J., Domingo-Ferrer, J., Sanchez, D., & Martinez, S. (2015). t-closeness through microaggregation: Strict privacy with enhanced utility preservation. *IEEE Transactions on Knowledge and Data Engineering*, *27*(11), 3098-3110.
[61] Sánchez, D., Domingo-Ferrer, J., Martínez, S., & Soria-Comas, J. (2016). Utility-preserving differentially private data releases via individual ranking microaggregation. *Information Fusion*, *30*, 1-14.
[62] Domingo-Ferrer, J., & Torra, V. (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing*, *13*, 343-354.
[63] Martínez, S., Sánchez, D., Valls, A., & Batet, M. (2012). Privacy protection of textual attributes through a semantic-based masking method. *Information Fusion*, *13*(4), 304-314.
[64] Hajian, S., & Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, *25*(7), 1445-1459.
[65] A. Arias-Duart, F. Parés, D. Garcia-Gasulla and V. Giménez-Ábalos, "Focus! Rating XAI Methods and Finding Biases," 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Padua, Italy, 2022, pp. 1-8, doi: 10.1109/FUZZ-IEEE55066.2022.9882821
[66] Guidotti, Riccardo, et al. "Factual and counterfactual explanations for black box decision making." IEEE Intelligent Systems 34.6 (2019): 14-23.
[67] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.
[68] Lundberg, Scott, and Su-In Lee. "A unified approach to interpreting model predictions." arXiv preprint arXiv:1705.07874 (2017).
[69] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." International Conference on Machine Learning. PMLR, 2017.
[70] *Avishek Anand, Megha Khosla, Jaspreet Singh, Jan-Hendrik Zab, and Zijian Zhang. 2019. Asynchronous Training of Word Embeddings for Large Text Corpora. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19). Association for Computing Machinery, New York, NY, USA, 168–176. https://doi.org/10.1145/3289600.3291011*
[71] *Jurek Leonhardt, Koustav Rudra, Megha Khosla, Abhijit Anand, and Avishek Anand. 2022. Efficient Neural Ranking using Forward Indexes. In Proceedings of the ACM Web Conference 2022 (WWW '22). Association for Computing Machinery, New York, NY, USA, 266–276. https://doi.org/10.1145/3485447.3511955*
[72] *Garcia-Gasulla, D., Vilalta, A., Parés, F., Ayguadé, E., Labarta, J., Cortés, U. and Suzumura, T., 2018, November. An out-of-the-box full-network embedding for convolutional neural networks. In 2018 IEEE International Conference on Big Knowledge (ICBK) (pp. 168-175). IEEE.*
[73] *Garcia-Gasulla, D., Parés, F., Vilalta, A., Moreno, J., Ayguadé, E., Labarta, J., Cortés, U. and Suzumura, T., 2018. On the behavior of convolutional nets for feature extraction. Journal of Artificial Intelligence Research, 61, pp.563-592.*