UNIVERSITÀ
degli STUDI
di CATANIA

Dipartimento
di Fisica
e Astronomia
"Ettore Majorana"

DOTTORATO DI RICERCA IN SISTEMI COMPLESSI PER LE SCIENZE FISICHE,
SOCIO-ECONOMICHE E DELLA VITA

ALESSANDRO MARIO MUSCOLINO

NATURAL LANGUAGE PROCESSING SOLUTIONS FOR KNOWLEDGE EXTRACTION: *NetME* AND *EmotWion*

TESI DI DOTTORATO

RELATORE:
CHIAR.MO PROF. A. RAPISARDA

ANNO ACCADEMICO 2020/2021

Abstract

Natural language processing (*NLP*) is an area of computer science and artificial intelligence concerned with the interaction between computers and humans in natural language. With an exponential growth of big data in this era, the advent of *NLP* based systems has enabled us to access relevant information through a wide range of applications.

During my *PhD* I used these methodologies in two different domains, biomedical knowledge networks building and analysis of the contagion of emotions in social networks.

In biomedical domain, with the increasing volume and unstructured nature of scientific literature most of the information embedded within them are lost. The inference of new knowledge and the development of new hypotheses from current literature analysis are a fundamental processes for foundation of new scientific discoveries, and get knowledge about relations and interactions among biological elements, a very important case study in complex systems domain.

Knowledge Networks are helpful tools especially in the context of bio-

logical knowledge discovery and modeling, given the enormous amount of literature and knowledge bases available, and allow the researchers to obtain information on aspects already widely investigated by others researchers.

In emotion analysis domain, thanks to the social networks phenomenon, that deeply pervaded today's society, most of the communication paradigms have moved to online, hence there is a lot of social media data available which can be used for emotion analysis and classification. Emotion analysis is important because it affects our daily decision making capabilities, both socially or commercially context.

In this thesis I present *NetME*, a framework which I developed that combines *TAGME* annotation framework based on *Wikipedia* corpus and *NLP* methodology. It allows to build on-the-fly knowledge graphs starting from a subset of full texts obtained by a real-time query on *PubMed* and applying several syntactic analysis methodologies.

In this thesis I also describe another project, *EmotWion*, a framework which I developed that aims to study the contagion of emotions on complex networks like social networks and its duration over time.

Keywords: Knowledge Graph, Complex system, Complex network, Document Annotation, Syntactic Analysis Methodologies, Emotion Analysis Methodologies, Natural language processing, *spaCy*.

Acknowledgements

First of all, I would like to express my sincere thanks to my supervisor Prof. Andrea Rapisarda, for the consistent support, motivation and guidance during my PhD course and my research projects.

I would like to thank my family. My wife, Enrica supported me throughout this entire process and help me get to this point.

I am also thankful to the DMI's research group with which I collaborated, especially prof. Alfredo Ferro, prof. Alfredo Pulvirenti, Prof. Salvatore Alaimo, Dr. Antonio Di Maria, Drs. Valentina Rapicavoli.

Giarre 24/11/2021

Alessandro

Contents

Chapter 1

Introduction

The explosive growth of big data has created a rich source of knowledge and it is, currently, an hot topic in academia and industry. This term is used to describe a broad domain of concepts, ranging from extracting data from outside sources, storing and managing it, to processing such data with analytical techniques and tools.

Natural language processing (*NLP*) techniques can contribute to access relevant information from big data by offering automated means to do preprocessing, text classification, feature extraction, and topic modelling.

During my *PhD* I applied *NLP* techniques for big data analysis in biomedical domain and in social networks domain.

1.1 *NetME*

Due to the growing volume and unstructured nature of scientific literatures, most of the information embedded within them remain unusable. In particular, in research areas like biology or bio-medicine, thanks to fast-track

publication journals, the number of published papers increases significantly fast. Biomedical scientific publication databases such as *PubMed* are a valuable resource of literatures that include an enormous amount of information. Extract relevant information from resources of this size, need strenuous effort and careful examination of the literature.

On the other hand, network analysis has become a critical enabling technology to understand mechanisms of life, living organisms, and in general, uncover the underlying fundamental biological processes. Examples of applications include: (i) analyzing disease networks for identifying disease-causing genes and pathways [11]; (ii) discovering the functional interdependence among molecular mechanisms through network inference and construction [123]; iii) releasing Network-based inference models with application on drug re-purposing [59].

The availability of sizeable open-access article repositories such as *PubMed Central* [14], *arxiv* [48] *bioarxiv* [1] as well as ontology databases which hold entities and their relations [78], in the last few years motivated the implementation of text mining and and machine learning approaches to automatically extract biomedical knowledge from them.

Text mining [28], and *NLP* [77] tools employ information extraction methods to translate unstructured textual knowledge in a form that can be easily analyzed and used to build a functional network (i.e. a network in which the relations between two entities are not necessarily physical but can be indirect), or knowledge graphs [123, 38, 96].

This approaches to entity and relation extraction have shown an evolution from simple systems that rely only on co-occurrence statistics [23] to complex systems utilizing syntactic analysis or dependency parsing [47], and machine

learning algorithms [110].

These methodologies allow us to infer putative relations among molecules, such as understanding how proteins interact with each other or determining which gene mutations are involved in a disease. In the context of biology and biomedicine, the *Biological Expression Language (BEL)* [118], or *Resource Description Framework (RDF)* [85] have been widely applied to convert a text in semantic triplets having the following form: $< subject, predicate, object >$. The subject and object represent biological elements, whereas the predicate represents a logical or physical relationship between them [123, 58].

However, the implementation of biological text mining tools requires highly specialized skills in *NLP* and Information Retrieval. Therefore, several ecosystems and tools have been implemented and made available to the bio-science community.

Relevant tools include *PubAnnotation* [74], a repository of text annotations based on the "Agile text mining" concept; *PubTator* (*PTC*) [132], a web service for viewing and retrieving bio-concept annotations (for genes/proteins, genetic variants, diseases, chemicals, species, and cell lines) in full-text biomedical articles. This latter tool annotates all *PubMed* abstracts and more than three million full texts. The annotations are downloadable in multiple formats (XML, JSON, and tab-delimited) through the online interface, a RESTful web service, and bulk FTP.

Another interesting tool is *SemRep* [109], which extracts relationships from biomedical sentences in *PubMed* articles by mapping textual content to an ontology that represents its meaning. To establish the binding relation, *SemRep* relies on internal rules (called "indicator rules"), which map syntactic elements, such as verbs, prepositions, and nominalization, to predicates in the

Semantic Network.

Another relevant tool is *Hetionet* [59], an heterogeneous network of biomedical knowledge that unifies data from a collection of several available databases and millions of publications. Also, the edges are extracted from omics-scale resources and consolidated through multiple studies or resources.

Finally, in [139] authors propose an approach for knowledge graph construction with minimum supervision based on 24,687 unstructured biomedical abstracts. Authors included entity recognition, unsupervised entity and relation embedding, latent relation generation via clustering, relation refinement, and relation assignment to assign cluster-level labels. The proposed framework can extract 16,192 structured facts with high precision.

At chapther 4 I present *NetME* a web-based app, which is capable to extract knowledge from a collection of full-text documents (publicly available at https://www.netme.tk/ website, and https://github.com/alemuscolino/netme.git github repository) introduced in "*NetME On-the-fly knowledge network construction from biomedical literature*" presented at the Complex Networks 2020 conference [91], and, on August 2021, in a further work with new features, submitted in the *journal Applied Network Science*.

The tool orchestrates two different technologies:

- A customized version of the entity-linker *TAGME* [44] (called *Onto-TAGME*) for extracting network nodes (i.e., genes, drugs, diseases) from a collection of full-text articles.

- A software module, developed on top of *spaCy* [62] and *NLTK* [81] libraries, that derives relations (edges) between pair of nodes. Edges are weighted according to their frequency within the collection of full-texts

used to create the on-fly knowledge graph.

These inferred networks are handy in biomedicine, where it is essential to understand the difference between various components and mechanisms, such as genes and diseases, and their relations, such as up-regulation and binding. Therefore, the tool helps scientists fast identify reliable relations among the biological entities under investigation, based on their occurrences and mentions in *PubMed*'s articles.

The acronym *NetME* was designed to help understand how a knowledge network can be made from a series of annotations and relationships extracted by *OntoTAGME*, hence the union of *Network* and *OntoTAGME* generates the word *NetME*.

To the authors' knowledge, *NetME* is the first tool that allows to interactively synthesize biological knowledge-graphs on-the-fly starting from a *PubMed* query.

1.2 *EmotWion*

In social networks domain the way in which habits, behaviors and feelings spread within social networks websites has recently aroused considerable interest in the academic, politics and economics fields.

Social networks, which are complex networks, by their nature play a key role in the information dissemination process [95], so understand how the influence of a social object spreads over the web opens new horizons to a myriad of applications, especially, in advertising, amplifying the virality of marketing and the mechanism of product recommendations.

The social influence factor today constitutes a key that regulates human behavior and indicates the interaction between subjects in a virtual community that is reflected in the real community.

Interaction between users may lead to a change in thoughts, feelings or behaviors of people and numerous studies aimed at propagating of behaviors [27], of feelings [45] or more generally of communication [18] within social networks. It has been shown that the diffusion phenomenon loses its effect starting from the third, or, in some cases, the fourth degree of separation from the source of origin; in particular social information networks, allow to highlight and analyze the phenomenon in question.

Microblogging platforms like *Twitter* and *Facebook*, for example, offer users the ability to exchange huge amounts of information on the web based on the possibility of transmitting short text updates to the entire social network or to a select group of contacts.

Among all, *Twitter* is the most popular social microblogging network in the world, constantly growing and frequently used in research [56].

Compared to other social platforms with similar characteristics, *Twitter* content, called Tweet, is extremely short and users can follow other users in the network, based on the source of information which they prefer.

Numerous scientific researchers have emphasized that often transmitted tweets contain within them information on the emotional states of users [30], even when the user does not publish personal information directly related to their emotional profile.

At microscopic level, each tweet contains small parts of the individual user's mood, which, analyzed over a certain time window, can describe the general progress of the author's emotional state.

The emotional state of a user represented in written form is based exclusively on the articulation of words used to communicate the emotion of the author, and a sentimental analysis of the text can be done exclusively in an empirical way.

In the last decade, an approach to dimensional emotions has been proposed [112, 107], the circumflex model of emotions holds that affective states can be traced back to two main neurophysiological systems: the value of emotion, along a continuous line of pleasantness - unpleasantness, and the level of relative physiological activation. According to this model, every emotion can be summarized as the linear combination between two dimensions, varying in terms of valence and intensity of activation.

A very high number of empirical analyzes of emotional states are based on this or other models applied to big data collections generated by social platforms [125] or from sites or tools for reviews of goods or services [92, 55], in other cases they are based on the creation of dictionaries by direct comparison [89].

At chapter 5 I present *EmotWion*, a software that aims to analyze a process known as "emotional contagion" on *Twitter*. The software and analysis techniques described below are part of a research work to be submitted.

Research work analyzes: (i) source tweets published by social influencer users, with high number of followers, (ii) tweets from users who retweet source tweets.

The tool orchestrates four different subsystem:

- a *Twitter* content extraction system;
- a system for processing extracted tweet;

- a system for classifying the extracted tweets by calculating the overall emotion;
- a system for processing, comparing and displaying the emotional level of the extracted tweets.

It has been developed in *Python* programming language with support of *Python Twitter* and *NLTK* for text analysis.

Through *EmotWion* it is possible to analyze (i) the average quantity of emotionally contagion users and (ii) the average duration of the contagion for each user.

EmotWion acronym derives from union of two words, Emotion and Twitter; it enclosing the meaning of its functionality, the study of the contagion of emotions through twitter.

Results on the case studies analyzed by *EmotWion* highlight the importance of emotions in the processes of social dissemination of ideas and demonstrate the usefulness of methods based on social networks for the study of these phenomena.

They show how people are continuously exposed to emotional contagions through social networks, thus expanding the models of social influence and group polarization as the network components grow.

Chapter 2

Background and Related Work

2.1 *NetME*

The concept of semantic networks has a long history (Quillian, 1968) and opened up a basis for knowledge modeling and representation (Helbig, 2006) by providing an adaptable formal framework for scientific developments and applications [36].

These networks allow to model semantic relationships in patterns of interconnected labeled nodes and edges incorporating linguistic information that describes concepts or objects.

Semantic standards and technologies facilitate the combination of multiple biomedical association data across multiple domains, which can be highly heterogeneous, and enable the construction of knowledge networks composed of various biological entities, such as compounds, proteins, and genes [123].

Interactions among biological components is basically studied and analyzed through a rich spectrum of methods and metrics grounded on network

theory. Most of these methods focus on the comparison of two biological networks (e.g., control vs. disease) [67]. To support challenging biomedical data integration efforts, a computational technique should satisfy the following three basic requirements [31]:

- An inter-operable data model that is capable of capturing and modeling the observed biomedical networks.
- A data integration framework to map and merge network data across disparate data sources.
- A collection of computational services to analyze, discover and validate new associations in integrated biomedical networks.

Based on these requirements, computer scientist and in general life-science researchers are focusing their studies and research topics on Annotation echosystems and relations extraction frameworks.

2.1.1 Networks and graphs

A network is a collection of connected objects. Networks are often referred to graphs, graph theory is the area of mathematics concerning the study of graphs. In Computer Science, a graph is a data structure consisting of two components:

- objects of data structure, called nodes or vertices;
- connections between objects, called edges.

A graph G can be well described by the set of vertices V and edges E it contains, as $G = (V, E)$. Edges can be either directed or undirected, depending on whether there exist directional dependencies between vertices [4].

Networks can represent all sorts of systems in the real world. In some networks, not all nodes and edges are created equal, to model such difference, one can introduce different types of nodes and edges in the network, by using different colors and edge styles.

NetME works with directed and weighted multigraphs.

A directed graph is an ordered pair $G = (V, E)$ where E are not simple edge but arrows, or directed lines. If there are multiple edges with the same source and target nodes, the edge set is a multiset and these entities are addressed as directed multigraphs [5]. A weighted graph is a graph in which

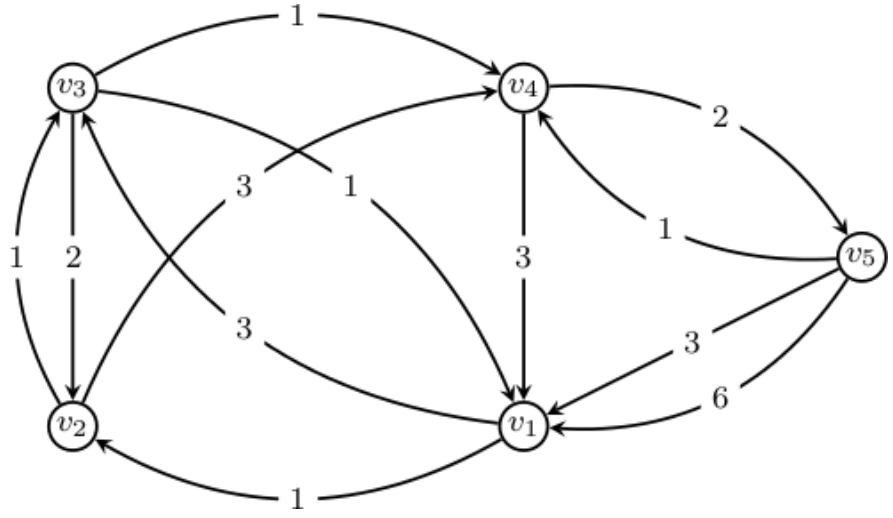


Figure 2.1: *Example of a simple directed and weighted multigraphs.* There are weighted arcs, with the same direction, which connect the same source and destination nodes (e.g. nodes v_1 - v_5).

each edge or node is given a numerical weight. An example of a directed and weighted multigraphs is shown in Figure 2.1.

2.1.2 Knowledge Graph

A Knowledge Graph is a systematic way to store both information and its meaning and connect information and data to knowledge. A knowledge graph is a directed labeled graph in which the labels have well-defined meanings and consists of nodes, edges, and labels. It represents a collection of interlinked descriptions of entities, real-world objects and events, or abstract concepts, obtained from several structured knowledge-bases such as ontologies $(O_1, O_2, ..., O_k)$.

An ontology is a formal description of knowledge as a set of concepts within a domain and the relationships that hold between them. To enable such a description, we need to formally specify components such as individuals (instances of objects), classes, attributes and relations as well as restrictions, rules and axioms. As a result, ontologies do not only introduce a sharable and reusable knowledge representation but can also add new knowledge about the domain [83].

The ontology data model can be applied to a set of individual facts to create a knowledge graph, a collection of entities, where the types and the relationships between them are expressed by nodes and edges between these nodes as shown in Figure 2.2. By describing the structure of the knowledge in a domain, the ontology sets the stage for the knowledge graph to capture the data in it. In some cases, a knowledge graph can contains multiple ontologies and inter-relations between them.

A semantic network, or frame network is a knowledge base that represents semantic relations between concepts in a network. This is often used as a form of knowledge representation. It is a directed or undirected graph $G = (E, R)$ consisting of vertices, which represent entity $e \in E$ coming from terminologies or ontologies, and edges $r \in R$, which represent semantic relations between them; both E and R are finite discrete spaces. In addition, every entity $e \in E$ can have some additional meta information defined with respect to the application of the knowledge graph [94].

There are a lot of examples of big knowledge graphs, *Google* announce its knowledge graph in 2012, however, there are very few technical details about its organization, coverage and size and there are also very limited means for

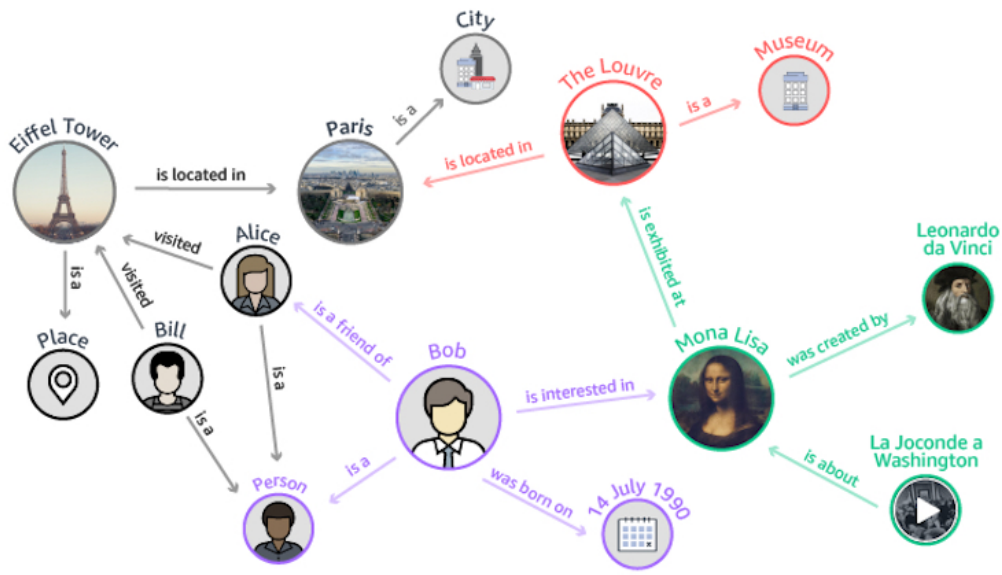


Figure 2.2: *Example of a simple knowledge graph*. Entity types are nodes and the relationships between them are labelled edges between these nodes.

using this knowledge graph outside *Google*'s own projects.

DBPedia is another example of big knowledge graph. This project leverages the structure inherent in the info boxes of *Wikipedia* to create an enormous dataset describes 4.58 million things, out of which 4.22 million are classified in a consistent ontology, including 1,445,000 persons, 735,000 places, 411,000 creative works, 241,000 organizations, 251,000 species and 6,000 diseases (https://wiki.dbpedia.org/about).

2.1.3 Documents repositories

For the researcher are available a considerable list of notable databases and search engines, useful in an academic setting for finding and accessing articles in academic journals, institutional repositories, archives, or other collections of scientific and other articles.

Thanks to the high availability of open-access articles repositories, and the continued growth of new entries and publications in the last few years, the research community has focused on text mining tools and machine learning algorithms to digest corpus and extract semantic knowledge. Due to the importance of these repositories, below is reported a brief explanation of the most commonly used documents repositories.

PubMed

PubMed [135] is a free resource supporting the search and retrieval of biomedical and life sciences literature with the aim of improving health–both globally and personally.

The *PubMed* database contains more than 32 million citations and abstracts

of biomedical literature. It does not include full text journal articles; however, links to the full text are often present when available from other sources, such as the publisher's website or *PMC*.

Available to the public online since 1996, *PubMed* was developed and is maintained by the National Center for Biotechnology Information (*NCBI*), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH).

The search for based-topic publications can be done through two different modes:

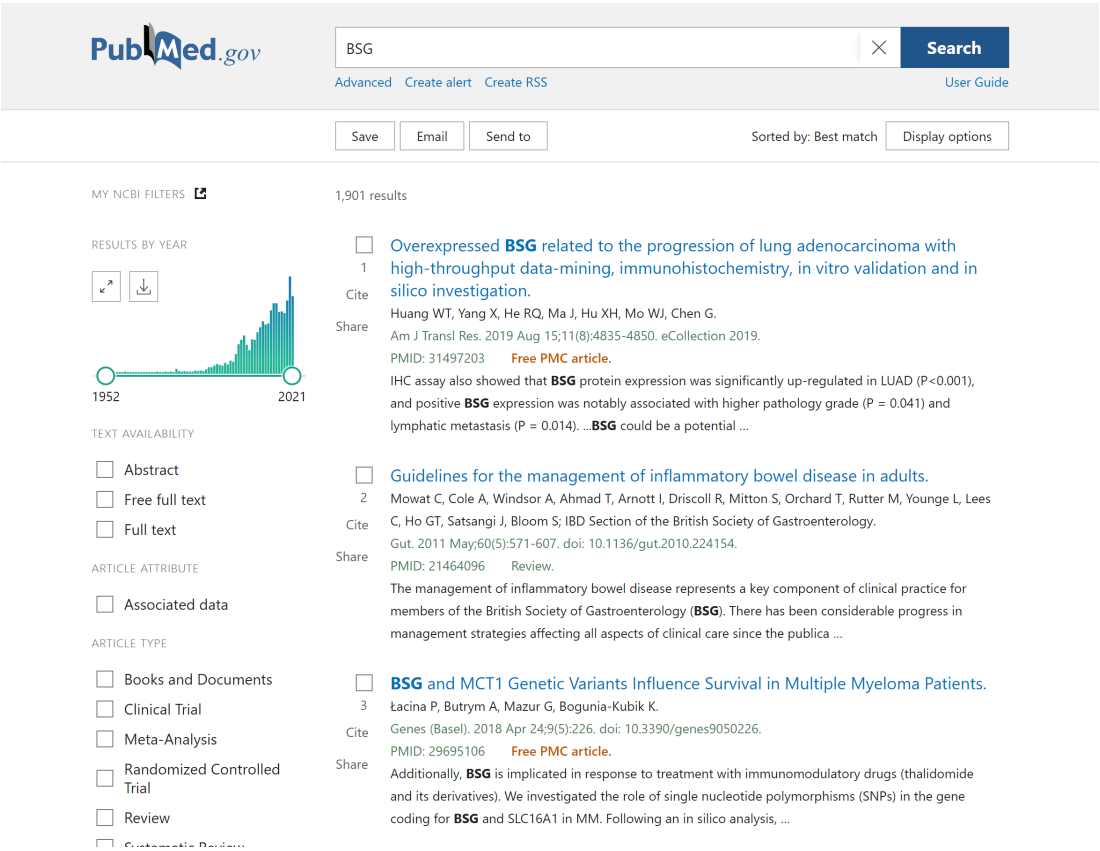


Figure 2.3: *Example of search through PubMed's website user interface*

Through user interface over the website https://pubmed.ncbi.nlm.nih.gov/ as shown in Figure 2.3.

Through *Entrez Programming Utilities (E-utilities)*, a set of eight server-side programs that provide a stable interface into the Entrez query and database system at the *National Center for Biotechnology Information* (*NCBI*). The E-utilities use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various *NCBI* software components to search for and retrieve the requested data. The E-utilities are therefore the structured interface to the Entrez system, which currently includes 38 databases covering a variety of biomedical data, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the biomedical literature. In *NetME* these APIs are used for both download the PMIDs list published from a search keyword, and to download title, abstract and content of the just downloaded PMIDs list.

PMC

PubMed Central (*PMC*) is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM). In keeping with NLM's legislative mandate to collect and preserve the biomedical literature, *PMC* serves as a digital counterpart to NLM's extensive print journal collection. *PMC* was developed and is managed by NLM's National Center for Biotechnology Information (*NCBI*).

Since its inception in 2000, *PMC* has grown from comprising only two journals, PNAS: Proceedings of the National Academy of Sciences and Molecular Biology of the Cell, to an archive of articles from thousands of journals.

Currently, *PMC* contains more than 6 million full-text records, spanning several centuries of biomedical and life science research (late 1700s to present). Content is added to the archive through. Just like *PubMed*, the *NCBI*'s E-utilities can be used to download full-text by REST API.

bioRxiv

bioRxiv [1] is a free online archive and distribution service for unpublished preprints in the life sciences. It is operated by Cold Spring Harbor Laboratory, a no-profit research and educational institution. By posting preprints on bioRxiv, authors are able to make their findings immediately available to the scientific community and receive feedback on draft manuscripts before they are submitted to journals.

Articles are not peer-reviewed, edited, or typeset before being posted online. However, all articles undergo a basic screening process for offensive or non-scientific content and for material that might pose a health or biosecurity risk and are checked for plagiarism. An article may be posted prior to, or concurrently with, submission to a journal but should not be posted if it has already been accepted for publication by a journal.

In addition it is categorized as New Results, Confirmatory Results, or Contradictory Results.

- New Results describe an advance in a field.
- Confirmatory Results largely replicate and confirm previously published work
- Contradictory Results largely replicate experimental approaches used

in previously published work but the results contradict and/or do not support it.

Readers may comment the articles on bioRxiv and also contact authors directly to get clarification. Pre-prints deposited in bioRxiv can be cited using their digital object identifier (DOI). Such object identifier is the same for every version of the manuscript.

2.1.4 Frameworks for annotation and relationships extraction from biomedical literature

Automatic extraction of information from unstructured data sources aims to organize structured information existing in natural language, and storing it in a form that allows further usage to be made by software applications. Entities are typically short phrases representing a specific object; instead, relations are physical or logical interactions among the entities of the considered domain, biological in this thesis work [79].

Various approaches have been proposed to extract relations from biomedical literature, due to the inherent complexity of biomedical text, most of relation extraction systems work on sentence-based level. The approaches used in relation extraction systems change from the level of linguistic analysis to the way patterns or rules are being learned. Based on the techniques employed in these systems, approaches can be categorized into three groups, namely co-occurrence, pattern-based and Machine Learning based approaches.

The simplest approach is the detection of co-occurrences of entities from sentences or abstracts relying on clustering procedures. The central hypothesis is that similar entities tend to occur together in similar contexts. Generally,

the type and direction of the relation cannot be determined.

The pattern-based extraction approaches does not face the high variability of how a relation can be expressed in natural language. It generally extracts single word terms rather than well-formed and compound concepts [80].

The Machine Learning approach relies on a standard supervised classification strategy, several algorithms like Naive Bayes Classifiers [129], Decision Trees [10] and Support Vector Machines [124], are usually employed. For example, a vector-based approach is used to transform a span of text and candidate relations into a numerical vectors used during the classification procedure.

Such models suffer a main problems, the errors generated in the entities extraction step may propagate to the step of relation classification. For instance, if a drug or disease entity mention is incorrectly recognized, than the extraction of its related relationships will be incorrect.

PubAnnotation

PubAnnotation [74] is an ecosystem based on agile approaches to text mining and annotation procedures. It is composed of the following three components:

- An annotation model composed of three sub-components: (i). Storage, (ii). Accessors. and (iii). Processors
- A dictionary-based annotator.
- A storage component facilitates regression testing.

To facilitate the communication among the components, two models have been created: Passive communication model allows processors or accessors

to get annotation stored in the annotation server, or to push new and revised ones to a server. Active communication model enables an annotation storage for initializing communication to actively obtain new or update annotations.

PTC Central

PTC Central is a web service for viewing and retrieving bioconcept annotations in full text biomedical articles. *PTC* Central [132] provides automated annotations from state-of-the-art text mining systems for genes/proteins, genetic variants, diseases, chemicals, species and cell lines, all available for immediate download. *PTC* annotates *PubMed* and the *PMC* Text Mining subset.

The new *PTC* web interface allows users to build full text document collections and visualize concept annotations in each document. Annotations

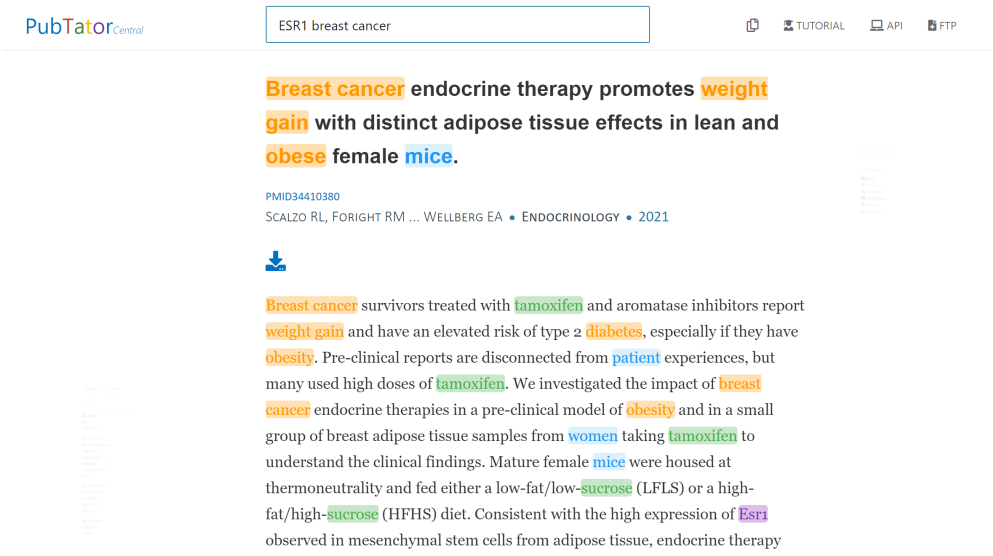


Figure 2.4: *Example of annotation with PTC Central.* Annotation are marked with different colors

are downloadable in multiple formats (XML, JSON and tab delimited) via the online interface, a RESTful web service and bulk FTP. Improved concept identification systems and a new disambiguation module based on deep learning increase annotation accuracy, and the new server-side architecture is significantly faster. *PTC* is synchronized with *PubMed* and *PMC*, with new articles added daily.

New articles in *PubMed* or *PMC* are first processed through a series of concept taggers to obtain annotations for each bioconcept type. In this manuscript, an annotation consists of a contiguous text span, a concept type, and an accession identifier.

The disambiguation module then resolves annotation conflicts (overlapping annotations). Annotated articles are subsequently stored in a MongoDB database, and made available to users via the new *PTC* Central web interface and the RESTful API for programmatic access.

To ensure consistency, the input/output text files for each step in the *PTC* Central processing pipeline are handled by BioC, a community-driven biomedical text processing data format for improved interoperability [100].

SemRep

SemRep [109] is a *NLP* system designed to recover semantic propositions from biomedical text using underspecified syntactic analysis and structured domain knowledge from the UMLS [65]. It extracts relationships from biomedical sentences in *PubMed* titles and abstracts by mapping textual content to an ontology which represents the meaning.

After the selection and tokenization of the text is completed, it is submitted

to an under-specified parser that relies on the syntactic information in the SPECIALIST Lexicon [25]. Part-of-speech ambiguities are resolved with the Xerox Part-of-Speech Tagger [33].

Basically, simple noun phrases are identified because these contain useful information about terms relationships. Prepositional phrases are treated as noun phrases whose first element is a preposition. Instead, other syntactic categories, including verbs, auxiliaries, and conjunctions are expressed through their part-of-speech label and put into a separate phrase. In order to show the returned annotations in a compact way, the semantic types are abbreviate. Therefore, "Diseases" or "Syndromes" are referred as "dsyn", "Organic Chemical" with (orch), etc.

The domain knowledge is obtained through Meta Map [9], a knowledge-based application that uses the SPECIALIST Lexicon and several rules to determine the best mapping between the text of a noun phrase and a concept. Thus, interpretation of semantic propositions depends on the under specified syntactic analysis enriched with domain knowledge.

The rules adopted allow to map syntactic indicators such as verbs, prepositions, and nominalization, to predicates in the Semantic Network. For example, exist a rule that links the "nominalization treatment" with the predicate TREATS. Further syntactic constraints on argument identification are controlled by statements expressed in a dependency grammar.

SemRep also addresses noun phrase coordination [109] by taking advantage of semantic types. Thus, on the basis of the under-specified syntax enhanced with domain knowledge, it is necessary to determine whether each coordinator is conjoining noun phrases or something other than noun phrases.

For a coordinator marked as conjoining noun phrases, the semantic type

of the noun phrase immediately to the right of that coordinator is examined. Then, the noun phrase immediately to the left of the coordinator, and noun phrases occurring to the left of that noun phrase (separated by another coordinator or a comma) are examined to understand whether they are semantically consonant.

Terms that share same meaning are classed under the same concept. Instead, each concept is assigned a semantic type, such as "Disease or Syndrome" or "Gene or Genome", so that similar concepts are clustered under the same semantic type. Once all text sentences are evaluated, the framework returns several tuples per each sentence that contain a subject, an object, and the relation that links them [63].

STRING

The *STRING* [123] database aims to collect and integrate this information, by consolidating known and predicted protein–protein association data for a large number of organisms. The associations in *STRING* include direct (physical) interactions, as well as indirect (functional) interactions, as long as both are specific and biologically meaningful.

Apart from collecting and reassessing available experimental data on protein–protein interactions, and importing known pathways and protein complexes from curated databases, interaction predictions are derived from the following sources: (i) systematic co-expression analysis, (ii) detection of shared selective signals across genomes, (iii) automated text-mining of the scientific literature and (iv) computational transfer of interaction knowledge between organisms based on gene orthology.

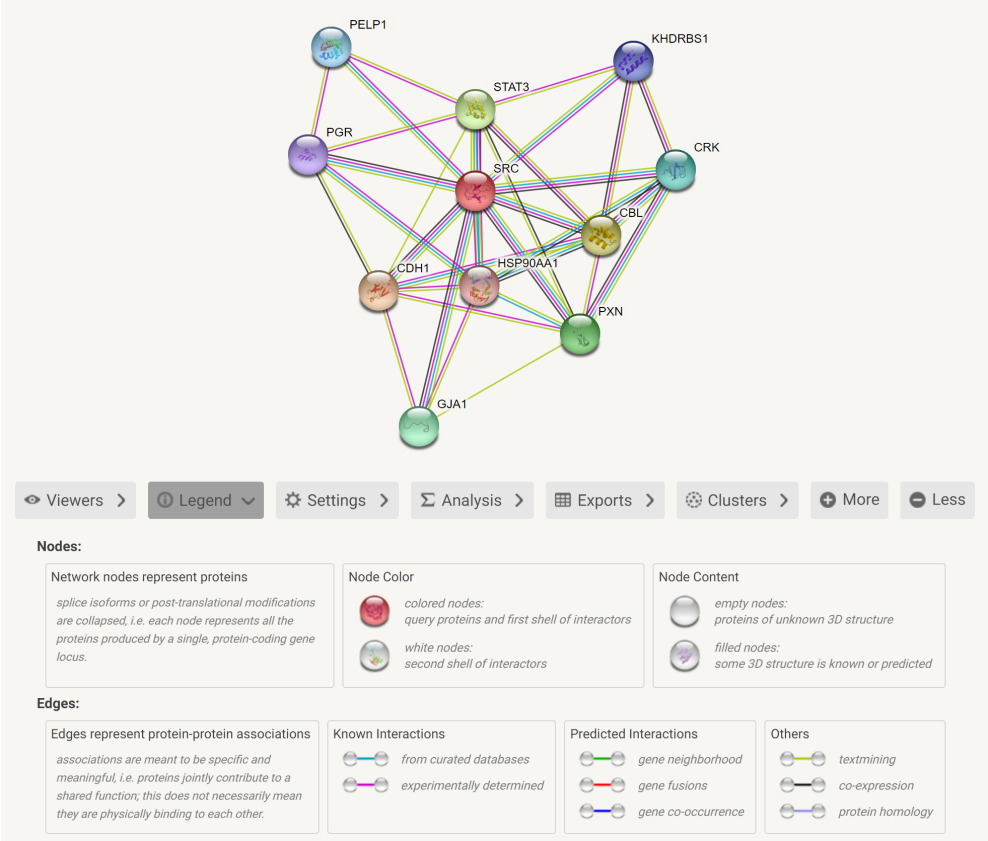


Figure 2.5: *Example of network generated with STRING from SRC gene*

In the latest version 10.5 of *STRING*, the biggest changes are concerned with data dissemination: the web frontend has been completely redesigned to reduce dependency on outdated browser technologies, and the database can now also be queried from inside the popular *CytoscapeJS* software framework.

Further improvements include automated background analysis of user inputs for functional enrichments, and streamlined download options.

Hetionet

A hetnet (short for heterogeneous information network) is a network where nodes and edges can be multiple types. This additional dimension allows a hetnet to accurately describe more complex data. Hetnets are particularly useful in biomedicine, where it's important to capture the conceptual distinctions between various components and mechanisms, such as genes and diseases, or upregulation and binding.

Hetionet [59] is a hetnet of biomedical knowledge. It encodes relationships

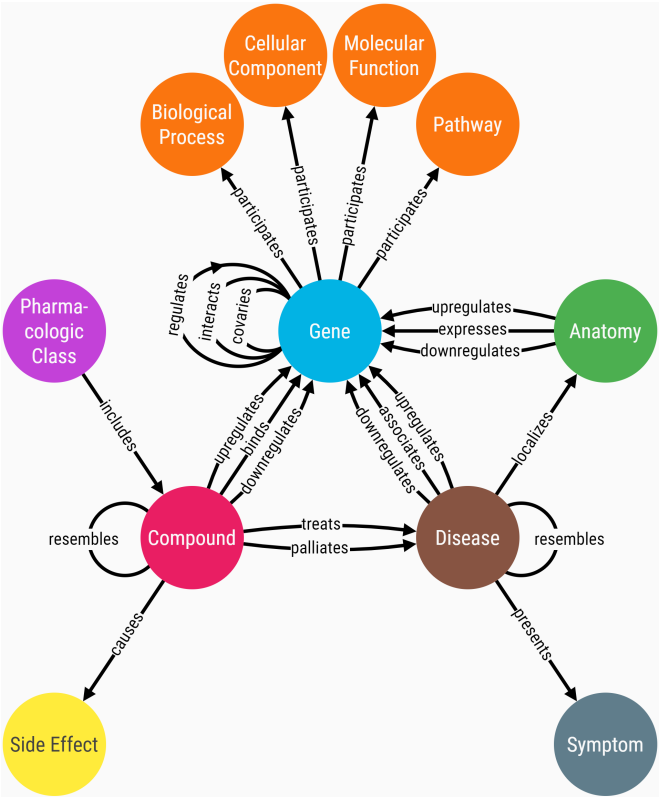


Figure 2.6: *Example of metagraph diagram generated with Hetionet*. It illustrates the connectivity between different types of nodes and edges in the network

uncovered by millions of studies conducted over the last half-century into a single resource. The network is constructed from a collection of publicly available databases, and is itself open-source and free to use, barring any upstream restrictions.

Hetionet enables scientists and biologists to formulate novel hypotheses, predictions, and other valuable insights by connecting an existing body of biomedical data across multiple levels and types in a convenient, accessible, holistic way.

Hetionet combines information from 29 public databases. The network contains 47,031 nodes of 11 types and 2,250,197 edges of 24 types.

Reactome

Reactome [31] is an open-source, open access, manually curated and peer-reviewed pathway database, a bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic and clinical research, genome analysis, modeling, systems biology and education.

Such network in *Reactome* can be seen as a directed graph, which consists of nodes and directed edges connecting ordered pairs of nodes.

Reactome stores pathway data in its natural form so that it does not require any transformation of data into a flat format. This approach reduces the complexity of the database and guarantees a easier access to the data.

Compared with other biological pathways databases, *Reactome* adopts *neo4j* as graph database to improve the data traversal performance and knowledge discovery. *Reactome* is available over the website: https://reactome.org.

2.2 *EmotWion*

Emotions govern our lives; they are an important part of the human experience, and they affect our decision-making. We tend to repeat actions that make us feel happy, but we avoid those that make us angry or sad.

Information spreads quickly via social networks, and as we know, emotions tend to intensify if left undealt with. Thanks to *NLP*, this subjective information can be extracted from written sources such as social posts, comments and conversations, allowing us to understand the emotions expressed by the author of the text and therefore act accordingly.

Emotion Detection will play a promising role in the field of Artificial Intelligence, especially in the case of Human-Machine Interface development, for Emotion Detection from an artificial intelligence different parameter should be taken into consideration. Various types of techniques are used to detect emotions from textual information. *NLP* techniques, machine learning, and computational linguistics are used.

Emotions detection provide observers with information regarding our current state and well-being. For businesses and individuals to be able to provide optimal services to customers, there is a need for them to identify the different emotions expressed by people and use that as the basis to provide bespoke recommendations to meet the individual needs of their customers

2.2.1 Emotion detection on social networks, related works

A lot of researchers have analyzed social networks such as *Twitter* and *Facebook* to assess the potential use of social media to detect behavioral disorders or in

the case of reactions to social phenomena.

Tomer, et al. [117] conducted some studies related to the use of *Twitter* during the terrorist attack on the Westgate shopping center in Kenya in September 2013 which led to a four-day siege, with 67 victims and 175 wounded dead; during the time of crisis, *Twitter* became a crucial communication channel between government, emergency responders and the public, facilitating the management of event emergencies. A total of 67,849 tweets were collected and analyzed, evaluating the hashtags and the propagation of the information based on the type.

Park et al. [99] have instead studied how to capture depressive moods on *Twitter*. The study was conducted on 69 subjects to understand how their depressive states are reflected in their status updates on the social network. The analysis was conducted in three phases: analyzing users to identify their level of depression, gathering users' tweets and at the last stage comparing users' depression levels with the content of their *Twitter* posts. The results showed that participants with depression showed an increase in the use of words related to negative emotions in their tweets.

Another work on depressive disorders in social networks with the aim of diagnosing them in a preventive way was carried out by Eichstaedt et al. [40] Starting from a dataset of 683 patients, 114 of whom were diagnosed with depression in their medical records; some predictors of depression including emotional, interpersonal and cognitive processes were found in the texts of these patients' *Facebook* posts.

Another analysis of emotions on *Twitter* data by Bollen et al. [24] who have tried to find a relationship between the general public mood and socio-economic or socio-cultural events. The studies conducted have led to the

extraction of six dimensions of mood, anger, tension, vigor, depression, confusion, fatigue using the psychometric instrument POMS, Profile of Mood States. It has been discovered that social, political, cultural and economic events have a significant and instantaneous effect on the various dimensions of mood.

The work carried out by Golder et al. [49], in which it was studied how individual mood varies from hour to hour, from day to day and through seasons and cultures, also measuring the positive and negative effect in *Twitter* posts was also very interesting.

2.2.2 Emotion models

Emotion models are the foundations of emotions detection systems; they define how emotions are represented. The models assume that emotions exist in various states thus the need to distinguish between the various emotion states. When undertaking any emotions detection related activity, it is imperative to initially define the model of emotion for use.

Various forms of representing emotions are identified [37]; the most important are the discrete and dimensional emotion models.

Discrete models of emotion

Discrete model of emotions involves placing emotions into distinct classes or categories. Prominent among them include:

- Paul Ekman model [41] distinguishes emotions based on six basic categories. The theory asserts that there exist six fundamental emotions that originate from separate neural systems as a result of how an experiencer

perceives a situation, thus emotions are independent. These fundamental emotions are happiness, sadness, anger, disgust, surprise, and fear. However, the synergy of these emotions could produce other complex emotions such as guilt, shame, pride, lust, greed, and so on.

- Robert Plutchik [103] model which as Ekman, postulates that there exist few primary emotions, which occur in opposite pairs and produces complex emotions by their combinations. He named eight of such fundamental emotions, that is, acceptance/trust and anticipation in addition to the six primary emotions posited by Ekman. The eight emotions in opposite pairs are joy vs sadness, trust vs disgust, anger vs fear, and surprise vs anticipation. According to Plutchik, for each emotion, there exist varying degrees of intensities that occurred as a result of how events are construed by an experiencer.

- Orthony, Clore, and Collins model [98] dissented to the analogy of "basic emotions" as presented by Ekman and Plutchik. They, however, agreed that emotions arose as a result of how individuals perceived events and that emotions varied according to their degree of intensity. They discretized emotions into 22, adding 16 emotions to the emotions Ekman posited as basic, thus spanning a much wider representation of emotions, with additional classes of relief, envy, reproach, self-reproach, appreciation, shame, pity, disappointment, admiration, hope, fears-confirmed, grief, gratification, gloating, like, and dislike.

Dimensional models of emotion

Dimensional models presupposes that emotions are not independent and that there exists a relation between them hence the need to place them in a spatial space. Thus dimensional models position emotions on a dimensional space depicting how related emotions are and usually, reflecting the two main fundamental behavioral states of good and bad [115].

Both unidimensional and multidimensional models are affected by relative degrees (low to high) of their occurrences; unidimensional models are rarely used but their fundamental idea permeates most multidimensional models.

- Russell [112] presents a circular two-dimensional model prominent in dimensional emotions representation called the circumplex of affect. The circumplex model of affect proposes that all affective states arise from cognitive interpretations of core neural sensations that are the product of two independent neurophysiological systems.

  This model stands in contrast to theories of basic emotions, which posit that a discrete and independent neural system subserves every emotion. Basic emotion theories no longer explain adequately the vast number of empirical observations from studies in affective neuroscience, the circumplex model suggest that a conceptual shift is needed in the empirical approaches taken to the study of emotion and affective psychopathologies.

  The circumplex model of affect is more consistent with many recent findings from behavioral, cognitive neuroscience, neuroimaging, and developmental studies of affect. Moreover, the model offers new theoretical and empirical approaches to studying the development of af-

fective disorders as well as the genetic and cognitive underpinnings of affective processing within the central nervous system.

This model suggests that emotions are distributed in a two-dimensional circular space, containing arousal and valence dimensions. Arousal represents the vertical axis and valence represents the horizontal axis, while the center of the circle represents a neutral valence and a medium level of arousal[108]. In this model, emotional states can be represented at any level of valence and arousal, or at a neutral level of one or both of these factors.

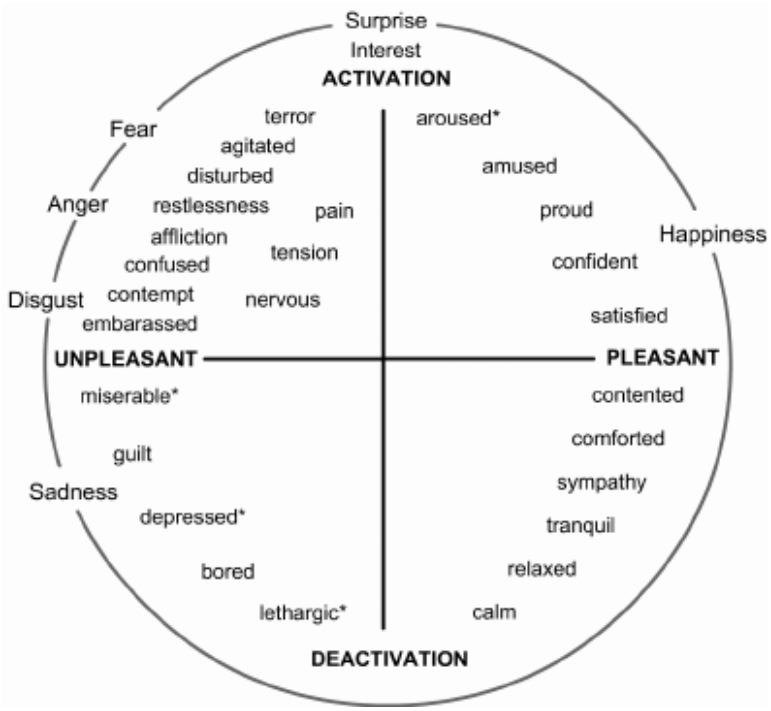


Figure 2.7: *A graphical representation of the circumplex model of affect*. The horizontal axis representing the valence dimension and the vertical axis representing the arousal or activation dimension.

The Circumplex model of Affect establishes that emotions are not independent but related as shown in Figure 2.7.

- Plutchik [103] presents a 2-dimensional wheel of emotions that shows Valence on the vertical axis and Arousal on the horizontal axis. The wheel shows emotions in concentric circles with the innermost emotions being derivatives of the eight fundamental emotions, then the eight fundamental emotions and finally combinations of the primary emotions on the outermost parts of the wheel.

  The wheel shows how related emotions are according to their posi-

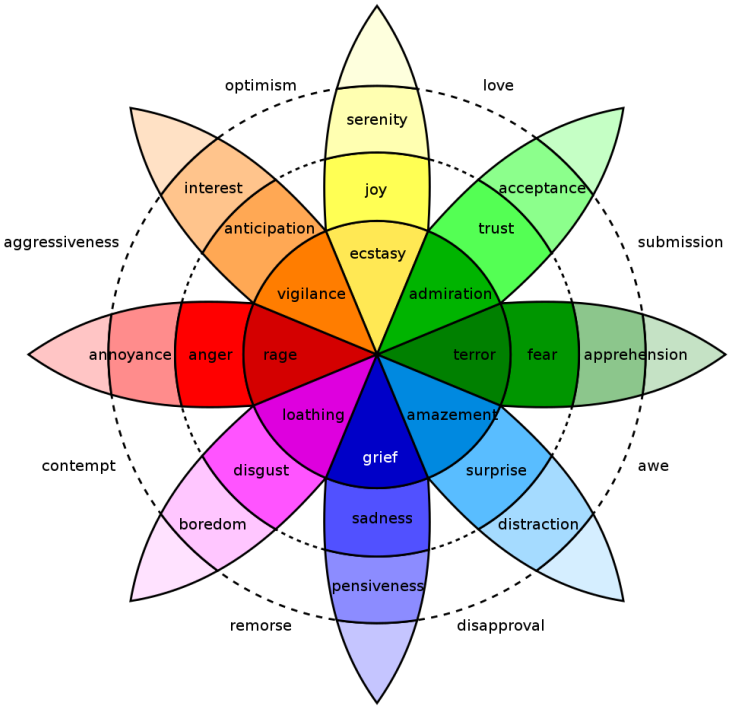


Figure 2.8: *A graphical representation of the Plutchik's whell of emotions.*

tions on the wheel. Figure 2.8 shows the wheel of emotions proposed by Plutchik.

- Russell and Mehrabian [113] also present a 3-dimensional emotion model made up of Valence/Pleasure, Arousal, and Dominance as the third dimension. Arousal and Valence, as postulated in the 2-D, represent how pleasant/unpleasant or active/inactive an emotion is respectively.
  The third dimension of Dominance describes the degree to which experiencers had control over their emotions.

2.2.3 Emotion detection approaches

Rule construction approach

The rule-based approach outlines major grammatical and logical rules to follow in order to detect emotions from documents. Rules for few documents may be easily created; however, with large amounts of documents, complexities may result. The rule construction approach encompasses keyword recognition and lexical affinity methods.

The keyword recognition method deals with the construction or the use of emotion dictionaries or lexicons. There are numerous keyword recognition dictionaries, notable among them are the *WordNet-Affect* [122], *EmoSenticNet* [106], *DepecheMood* [121], *SentiWord Net dictionaries* [42] and the *NRC VAD Lexicon* [89]. These emotion lexicons contain emotion search words or keywords; the task is to find occurrences of these search words in a written text at the sentence level. Once the keyword is identified within the sentence, a label is assigned to the sentence.

This approach though simple and straightforward faces challenges, including the need for an emotion dictionary to contain reasonable number of emotion categories, since limited keywords can greatly affect the performance of the approach among ambiguity of keywords and the lack of linguistic information.

The lexical affinity method augments the keyword recognition method. This is because aside from the identification of keywords, random emotion words are assigned a probabilistic affinity. The lexical affinity is responsible for this second stage of assigning probabilistic affinities to the random emotion words. The word "good" for instance, may be assigned a probabilistic affinity of "positive," "angry" may be assigned a "negative" affinity, and so on.

The drawback associated with this probabilistic affinity assignment is that it does not fully represent the various categories of emotions but rather reduces them into two extremes states. Also, this approach can lead to inaccuracies in the classification of emotions depending on the context of the assigned words.

These drawbacks often necessitate the need for using other approaches for detecting emotions in texts.

The machine learning approach

The machine learning approach solves the emotion detection problem by classifying texts into various emotion categories through the implementation of machine learning algorithms. The detection is often carried out using a supervised or an unsupervised machine learning technique.

Canales et al. [22] showed that supervised machine learning algorithms

have been widely implemented in text-based emotion detection problems and have offered comparatively better detection rates than in problems where unsupervised machine learning techniques were implemented. However, throughout this survey, it's possible to observe that not only have the widely explored techniques been unsupervised but also traditional.

These traditional unsupervised machine learning techniques such as the support vector machine, naive Bayes, conditional random fields, and so on, are not as robust and do not explicitly extract the semantic information relevant for effectively detecting emotions in texts [57]. Recently, supervised deep learning models are being adopted as machine learning approaches to detect emotions from texts.

The implementation of these techniques to texts-based emotion detection problems has been seen to outperform techniques that implemented traditional unsupervised machine learning techniques [7].

Hybrid approach

The hybrid approach combines the rule-construction and the machine learning approaches into a unified model. Thus, drawing from the strength of both approaches while concealing their associated limitations, this approach has a higher probability of transcending the other two approaches individually.

However, in conducting this survey, it was identified that most systems that employed this approach implemented a rule engine together with a traditional unsupervised machine learning technique [26, 60] resulting in satisfactory results.

With unsupervised deep learning architectures performing better in recent

advancements, hybrid models implementing an unsupervised deep learning technique together with standard rules may perform comparatively better as highlighted by references [64].

However, there remains the need to obtain the most effective deep learning technique to enhance performance in the hybrid approach since the approach relies heavily on the particular type of deep learning technique used.

Chapter 3

Algorithms, Frameworks and Tools

In this chapter I describe algorithms, frameworks and tools, used in the research projects I worked on during my *PhD*.

In particular *spaCy* an open source framework for *NLP* used in both *NetME* and *EmotWion*, *TAGME* a powerful tool that is able to identify on-the-fly meaningful short-phrases in an unstructured text and *OntoTAGME* a custom *TAGME* version, which uses biological and biomedical ontologies suitable to perform annotations on text coming from scientific papers, which was used for the development of *NetME*.

3.1 *spaCy*

spaCy is an open source framework for *NLP* written in *Python*, designed specifically for production use and helps you build applications that process and "understand" large volumes of text. It can be used to build information extraction or natural language understanding systems, or to pre-process text for deep learning.

spaCy provides a variety of linguistic annotations to give you insights into a text's grammatical structure. This includes the word types, like the parts of speech, and how the words are related to each other. For English language there are trained models to download for a wealth of *NLP* tasks.

One of the ideas behind *spaCy* though is that their untrained models, consisting of deep, convolutional neural networks for the most common *NLP* tasks, can be used to train a model in any language [3]. Custom pipeline components, designed for the specific task at hand, can also be added to the pipeline, which was important for this work, since models from other frameworks were used and then wrapped by a custom *spaCy* component, thereby integrating these frameworks with the *spaCy* pipeline.

Figure 3.1 shows an image of the *spaCy* pipeline. It consists of preprocessing (tokenization), POS-tagging, dependency parsing and named entity recognition [2].

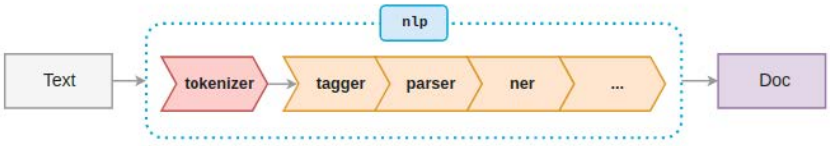


Figure 3.1: *spaCy Language Processing Pipeline.*

The advantages of having a single pipeline to separate models for each *NLP* task is that you get all the resulting data and features in one single *doc* object, which can then be used for customized tasks downstream, for instance relation extraction.

Processing raw text intelligently is difficult: most words are rare, and it's common for words that look completely different to mean almost the same

thing. The same words in a different order can mean something completely different. While it's possible to solve some problems starting from only the raw characters, it's usually better to use linguistic knowledge to add useful information. That's exactly what *spaCy* is designed to do, from raw text, and get back a Doc object, that comes with a variety of annotations. The functions used in the *NetME* project are described below.

Part of speech (POS) tagging

After tokenization, *spaCy* can parse and tag a given Doc. This is where the trained pipeline and its statistical models come in, which enable *spaCy* to make predictions of which tag or label most likely applies in this context.

A trained component includes binary data that is produced by showing a system enough examples for it to make predictions that generalize across the language, for example, a word following "the" in English is most likely a noun. Linguistic annotations are available as Token attributes; like many *NLP* libraries, *spaCy* encodes all strings to hash values to reduce memory usage and improve efficiency.

Morphology

Inflectional morphology is the process by which a root form of a word is modified by adding prefixes or suffixes that specify its grammatical function but do not change its part-of-speech; a lemma (root form) is inflected (modified/combined) with one or more morphological features to create a surface form.

For languages with relatively simple morphological systems like English,

spaCy can assign morphological features through a rule-based approach, which uses the token text and fine-grained part-of-speech tags to produce coarse-grained part-of-speech tags and morphological features.

- The part-of-speech tagger assigns each token a fine-grained part-of-speech tag. In the API, these tags are known as Token.tag. They express the part-of-speech, for example verb, and some amount of morphological information, for example that the verb is past tense.
- For words whose coarse-grained POS is not set by a prior process, a mapping table maps the fine-grained tags to a coarse-grained POS tags and morphological features.

Lemmatization

The Lemmatizer is a pipeline component that provides lookup and rule-based lemmatization methods in a configurable component. An individual language can extend the Lemmatizer as part of its language data.

For pipelines without a tagger or morphologizer, a lookup lemmatizer can be added to the pipeline as long as a lookup table is provided. The lookup lemmatizer looks up the token surface form in the lookup table without reference to the token's part-of-speech or context.

Dependency parser

spaCy features a fast and accurate syntactic dependency parser, and has a rich API for navigating the tree; the parser also powers the sentence boundary detection, and allows to iterate over base noun phrases, or "chunks".

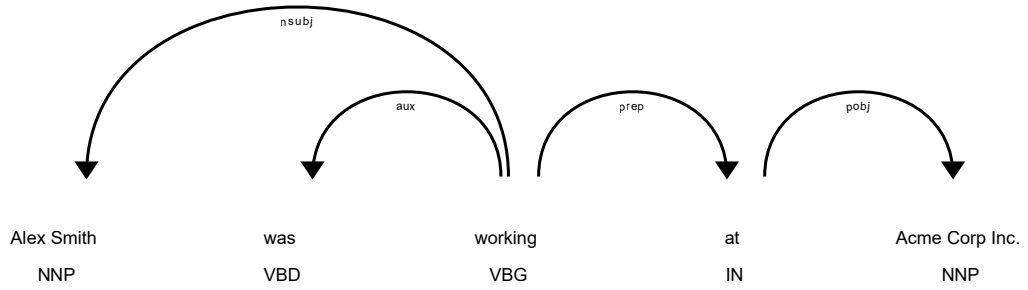


Figure 3.2: *spaCy example of syntactic dependency tree.*

spaCy uses the terms head and child to describe the words connected by a single arc in the dependency tree as described in Figure 3.2.

The term *dep* is used for the arc label, which describes the type of syntactic relation that connects the child to the head. As with other attributes, the value of *.dep* is a hash value.

Because the syntactic relations form a tree, every word has exactly one head, it's possible to iterate over the arcs in the tree by iterating over the words in the sentence.

3.2 *NLTK*

The Natural Language Toolkit, or more commonly *NLTK* [81], is a suite of libraries and programs for symbolic and statistical *NLP* for English written in the *Python* programming language.

NLTK is a leading platform for building *Python* programs to work with human language data and includes graphical demonstrations and sample data.

It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength *NLP* libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, *NLTK* is suitable for linguists, engineers, students, educators, researchers, and industry users alike.

The functions used in the *NetME* and *EmotWion* project are described below

Tokenization

Tokenizing means splitting the text. Tokenizing can split paragraphs into sentences, sentences into words, and words in characters. With *NLTK* we perform two types of tokenizing.

- Sentence tokenizing takes a text or paragraph string text as an argument and looks for the full stops punctuation symbols (.) in the text and split according to it.
- Word tokenizing split the text on basis of spaces, special type characters and etc.

Stopwords detection

Stopwords are words which does not add much meaning to a sentence they are contributed to it but had not much importance and during *NLP*, for better results, is recommended to remove they.

Example of some common stopwords are "the", "a", "he" and etc. To remove the stopwords from text or sentences there is a stopword class in *NLTK*.

Part of speech (POS) tagging

As already seen in the *spaCy* framework, POS tagging it's a method that assigned a label to every word in a text to indicate a part of speech e.g tenses, numbers, nouns, verb, adverbs, adjectives, pronouns, conjunction, and their sub-categories and etc.

POS tagging in *NLTK* is performed using the "pos_tag" function, which returns pairs of labeled tokens with the relative POS tag

3.3 *TAGME*

TAGME [44] annotates, on-the-fly and with high precision/recall, short text with pertinent hyperlinks to *Wikipedia* articles. *TAGME* uses *Wikipedia* anchor texts as spots and the pages linked to them in *Wikipedia* as their possible meanings.

The authors employ *Wikipedia* as corpus because of ever-expanding pages number. In addition, it enriches texts with explanatory links to provide structured knowledge about any unstructured fragment of text. So any task that is currently addressed with bags of words-indexing could use these links to draw a vast network of concepts and their inter-relations. The anchor for a *Wikipedia* page *p* is the text used in another *Wikipedia* page to point to p. All the anchors composed by one character or just numbers have been discarded because these could be unsuitable for annotation and probably misleading for disambiguation.

To solve ambiguity and polysemy, Tagme tries to disambiguate each anchor "a" in "AT" (set of all anchors occurring in the input text "T") by computing a score for each possible meaning "pa" of "a" through a new notation of "collective agreement" between "pa" and the possible meanings of all other anchors detected in "T".

Where "pa" in "$Pg(a)$" (set of all *Wikipedia* pages linked by "a"). Therefore, for each other anchor "b" in $AT - a$ is computed its vote to the annotation "$a" - > "pa$". Since "b" could have several meanings, *TAGME* compute the vote as the average relatedness between each meaning "pb" of the anchor "b" and the meaning "pa" that should be associated to the anchor "a".

However, not all possible meanings of "b" have the same significance, so *TAGME* weight the contribution of "pb" by means of its commonness "$Pr(pb|b)$". The formula used to compute the vote is:

$$\sum_{pb \in Pg(b)} rel(pb, pa) * Pr(pb|b)|Pg(b)| \tag{3.1}$$

When "b" is unambiguous, the vote $b(pa) = rel(pb, pa)$ because the product between "$Pr(pb|b)$" and "$|Pg(b)|$" is equal to one. But if "b" is polysemous, only the meanings "pb" related to "pa" will affect the vote "$b(pa)$". The total score $rel_a(pa)$ for the "goodness" of the annotation "$a - > pa$" is computed as the sum of the votes given to it by all other anchors "b" detected in the input text.

To disambiguate "a", the best annotation "$a - > pa$" is selected through two ranking algorithms combination: Disambiguation by Classifier (DC) and Disambiguation by Threshold (DT).

DC uses a classifier with the score "$rel_a(pa)$", and the commonness "$Pr(pa|a)$" as features to compute a "probability of correct disambiguation" for all meanings $pa \in Pg(a)$. Among all these meanings, DC selects the "pa" reporting the highest classification score.

Instead, DT recognizes a roughness in the value of "$rel_a(pa)$" among all $pa \in Pg(a)$, so it computes the top-best meanings p' in $Pg(a)$ according to their "$rel_a(p')$", and then annotates "a" with the means that obtains the highest commonness among them.

The set "$M(AT)$" of candidates produced by the Disambiguation Phase has to be pruned in order to remove improper anchors "a". The "bad anchors" are detected via a novel scoring function based on only two features: the link probability "$lp(a)$" of the anchor "a" and the coherence of its candidate annotation "$a- > pa$" with respect to the candidate annotations of the other anchors in "$M(AT)$".

Where "$lp(a)$" is defined as the ratio between the number of times the text "a" occurs as an anchor in *Wikipedia* ("$link(a)$") and the number of times the text a occurs in *Wikipedia* as an anchor or not ("$freq(a)$").

Instead, the coherence is the average relatedness between the candidate meaning "pa" for "a" and the candidate meaning "pb" for all other anchors "b" in "T". The objective is to keep all anchors whose link probability is high or whose assigned sense (page) is coherent with the senses (pages) assigned to the other anchors in $M(AT)$.

At the end of pruned procedure, we obtain the best anchors set to the Text T submit from the user.

3.4 *OntoTAGME*

TAGME [44] is a state-of-the-art entity linker for annotating *Wikipedia* pages mentioned in an input text. The tool searches for sequences of words (spots) that can be linked to pertinent *Wikipedia* pages (entities) that explain those words in that context. The use of *Wikipedia* as corpus allows to enrich texts with explanatory links in order to provide a structured knowledge for any unstructured fragment of the text. These links are then used for drawing a network of relationships among the extracted spots.

To mitigate ambiguity and polysemy, *TAGME* computes a ρ value $\in [0,1]$ for each Spot-Entity (Node) association, and keeps only those ones having the ρ value higher than an established user threshold. This value estimates

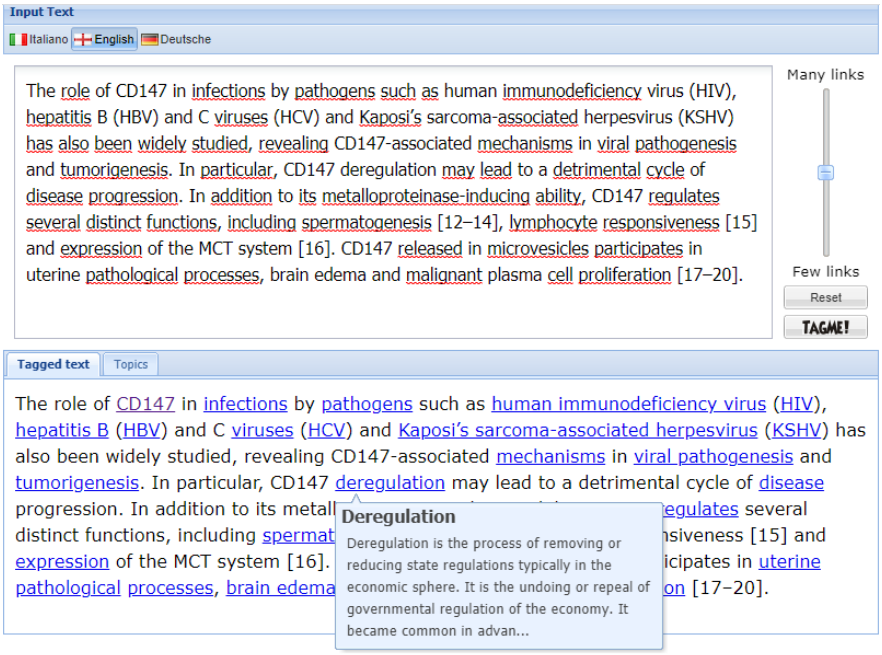


Figure 3.3: *An example of error encountered during TAGME annotation procedure*

the "goodness" of the annotation compared to other possible associations in the input text. A suitable use of ρ ensures the highest accordance among the extracted spots.

Due to the topics-generality of the *Wikipedia* corpus used by *TAGME*, several non-biological spots could be extracted during the annotation procedure as shown in Figure 3.3.

To overcome this limitation, we developed a customized version of *TAGME*, called *OntoTAGME*, which makes use of several ontology and literature databases, such as: *GeneOntology* (GO) [29], *DiseaseOntology* (DO) [116], *PathwayOntology* (PW) [101], *BRENDA* tissue /enzyme source (BTO) [53], *ProteinOntology* (PRO) [93], *Anatomical Entity Ontology* (AEO) [12], *Phenotype And Trait Ontology* (PATO) [133], *Cell Ontology* (CL) [34], *Cell Line Ontology* (CLO) [114], *DrugBank* [136], *DisGeNET* [102], *HGNC* [51], *ENSEMBL* [15], *CIViC* [54], and *PharmGKB* [134].

The usage of topic-specific ontology databases ensures reduced disambiguation errors and therefore yields highly reliable knowledge graphs inference.

The integration consisted of releasing a new intermediate *Python* layer (*Python* Parser in Figure 3.4), and a customized two-steps procedure (*Wikipedia* Adapter module in Figure 3.4) for converting ontology databases in a *Wikipedia-like* structure.

The *Python* layer transforms a generic ontology or database in a list of CSV files: *pages.csv*, *pageslink.csv* and *category.csv*. The *pages.csv* stores the name of each biological element, and all possible synonyms. The *pageslink.csv* contains all the relationships among the nodes of the ontology. Finally, the *category.csv* has the type of each element extracted from the ontology or database entry

(i.e Genes, Diseases, Drugs).

Next, a two steps procedure is triggered to convert each row of the page.csv file into an XML file containing a unique ID generated by our system, the name (title), type (category) and the description (page's body) of the considered biological element. Since an element j could have several linked pages "LPs" (i.e. DOID:0002116 is a DOID:10124), or redirected pages "RPs" due to synonyms (CD147 is a synonym of BSG), the process generates a tuple $\langle uniqueID_j, uniqueID_k \rangle$ for each element k belonging to LPs, and a tuple $\langle uniqueID_j, uniqueID_i \rangle$ for each element i belonging to RPs. These tuples are then stored in the SQL files "wiki-latest-pagelinks" and "wiki-latest-redirect", respectively.

Finally, the SQL and XML files are used to generate the complete *Onto-*

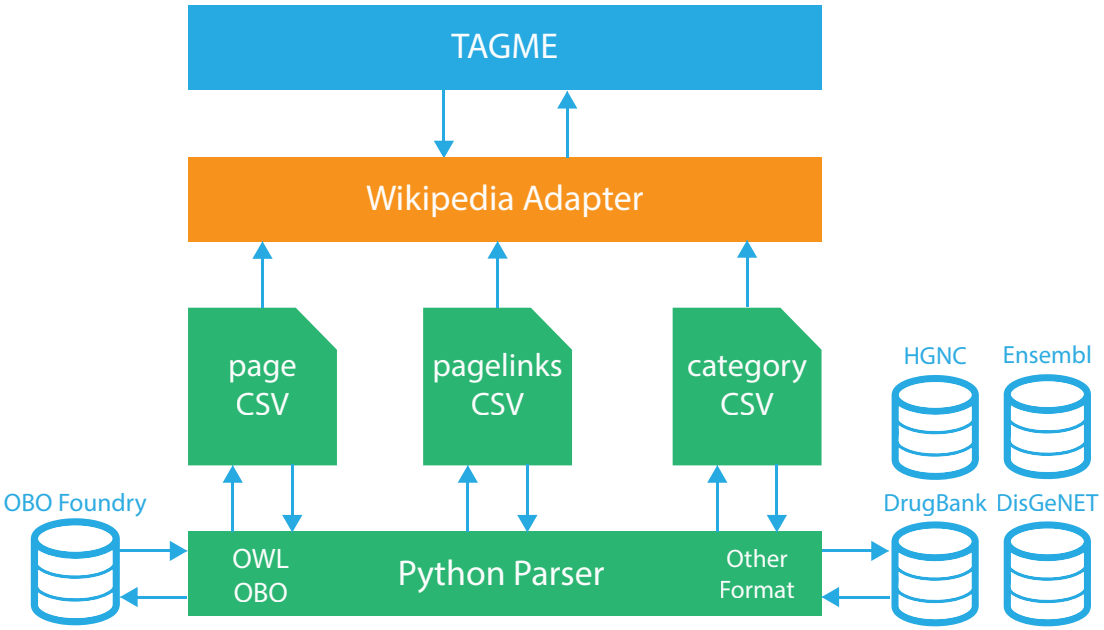


Figure 3.4: *OntoTAGME pipeline architecture*

TAGME network. It contains 331 thousand of main nodes, 700 thousand of synonyms, and 4 million of relationships.

Ontology Databases

In order to build the *OntoTAGME* annotation networks we used the following nine ontology and six bio-databases.

DrugBank [136] contains data about drugs name, drugs synonyms, drug-drug interaction, and other comprehensive drug-target information. The database release used in our project is the v5.1 which contains 13,367 drugs entries, including 2,611 approved small molecule drugs, 1,300 approved biotech (protein/peptide) drugs, 130 nutraceuticals and over 6,315 experimental drugs. Additionally, 5,155 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries.

HGNC (HUGO Gene Nomenclature Committee) [51] assigns unique and informative gene symbols and names to human genes. Standardized *HGNC* approved nomenclature is used in publications and biomedical databases to remove ambiguity and facilitate communication between researchers worldwide. The last database release contains more than 40,000 approved gene symbols of which over 19,000 are for protein-coding genes.

The *HGNC* also names a set of small and long non-coding RNA genes and pseudo-genes (659 since 2017). The genes are grouped on the basis of several shared characteristics such as homology, associated phenotype and encoded protein function.

ENSEMBL [15] contains genome annotation (i.e genes, variation, regulation and comparative genomics) across the vertebrate sub-phylum and key model organisms. This tool is also able to compute multiple alignments, predicts regulatory function and collects disease data. The last complete version of the *ENSEMBL* database has been downloaded through their FTP service, and then integrated in *OntoTAGME* thanks to *Python* Parser layer. All data in *ENSEMBL* are used in combination with those coming from *HGNC* to detect Genes name and symbols within a text.

DisGeNET [102] contains collections of genes and variants associated with human diseases. It integrates data from scientific literature, GWAS catalogues, expert curated repositories and animal models. Additionally, several original metrics are provided to assist the prioritization of genotype–phenotype relationships. *DisGeNET* releases two types of databases, Gene-Disease Associations and Variant-Gene Associations.

CIViC [54] is an expert-crowd-sourced knowledge-base for Clinical Interpretation of Variants in Cancer describing the therapeutic, prognostic, diagnostic and predisposing relevance of inherited and somatic variants of all types. *CIViC* is committed to open-source code, open-access content, public application programming interfaces (APIs) and provenance of supporting evidence to allow for the transparent creation of current and accurate variant interpretations for use in cancer precision medicine.

PharmGKB [134] is an interactive tool for researchers investigating how genetic variation affects drug response. It displays genotype, molecular, and

clinical knowledge integrated into pathway representations and Very Important Pharmacogene (VIP) summaries with links to additional external resources. A user may search and browse the knowledge-base by genes, drugs, diseases, and pathways through the website: http://www.pharmgkb.org).

Obofoundry [120] is the Open Biological and Biomedical Ontology (OBO) Foundry. It provides well-formed and scientifically accurate ontology thanks to the collaboration of ontology developers. They contribute to develop an evolving set of principles and common syntax based on ontology models that ensure the proper functioning of the system. In *NetME*, we use the following list of ontology:

- *GeneOntology* (GO) [29] project provides a uniform way to describe the functions of gene products from organisms across all kingdoms of life and thereby enable analysis of genomic data. it contains more than

ontology name	nodes number	edges number
go	43917	142086
doid	10862	29938
pr	326811	846366
pw	2619	6210
cl	10809	34410
clo	44712	91966
aeo	248	523
bto	6515	9378
pato	4610	13027

Table 3.1: *Number of nodes and edges for each ontology used in OntoTAGME*

44 thousand GO terms, 8 millions of annotations, 1.5 millions of gene products and nearly 5 thousand species.

- *Human Disease Ontology* (DO) [116] is a standardized ontology for human disease with the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease.

- *Pathway ontology* (PW) [101] is a controlled vocabulary for pathways that provides standard terms for the annotation of gene products.

- *Protein Ontology* (PRO) [93] defines taxon-specific and taxon-neutral protein-related entities in three major areas: proteins related by evolution; proteins produced from a given gene; and protein-containing complexes.

- *BRENDA* tissue / enzyme source (BTO) [53] is a structured controlled vocabulary for the source of an enzyme comprising tissues, cell lines, cell types and cell cultures.

- *Anatomical Entity Ontology* (AEO) [12] is an ontology of anatomical structures that expands *CARO*, the Common Anatomy Reference Ontology, to about 160 classes using the is_a relationship; it thus provides a detailed type classification for tissues. The AEO is useful in increasing the amount of knowledge in anatomy ontology, facilitating annotation and enabling interoperability across anatomy ontology.

- *Phenotype And Trait Ontology* (PATO) [133] is used in conjunction with other ontologies such as GO or anatomical ontology to refer to pheno-

types. Examples of qualities are red, ectopic, high temperature, fused, small, edematous and arrested.

- *Cell Ontology* (CL) [34] is designed as a structured controlled vocabulary for cell types. This ontology covers cell types from prokaryotes to mammals. However, it excludes plant cell types. One of the main uses of the CL is to describe samples used in transcriptomic and functional genomics studies, such as FANTOM5, ENCODE and LINCS.

- *Cell Line Ontology* (CLO) [114] is a community-driven ontology that is developed to standardize and integrate cell line information and support computer-assisted reasoning.

The data relating to the number of nodes and relationships extracted from each mentioned ontology have been listed in table 3.1

3.5 *NRC VAD Lexicon*

The *NRC Valence, Arousal, and Dominance Lexicon* is used as a comparison dictionary in the *EmotWion* project for the classification of emotions.

It includes a list of more than 20,000 English words and their valence, arousal, and dominance scores, manually annotated. For a given word and a dimension (V/A/D), the scores range from 0 (lowest V/A/D) to 1 (highest V/A/D).

To build this lexicon was used a comparative annotation technique called Best-Worst Scaling *(BWS)* [82] to obtain fine-grained scores and address issues of annotation consistency that plague traditional rating scale methods of annotation. In *NRC VAD Lexicon* are annotated commonly used English terms,

especially terms that denotate or connotate emotions and terms common in *Twitter*'s tweet. Specifically, the sources used are:

- All terms in the NRC Emotion Lexicon [90]. It has about 14,000 words with labels indicating whether they are associated with any of the eight basic emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.
- All 4,206 terms in the positive and negative lists of the General Inquirer [52].
- All 1,061 terms listed in ANEW [17].
- All 13,915 terms listed in the Warriner et al. [131] lexicon.
- 520 words from the Roget's Thesaurus categories corresponding to the eight basic Plutchik emotions.
- About 1000 high-frequency content terms, including emoticons, from the Hashtag Emotion Corpus (HEC) [88]

The union of the above sets resulted in 20,007 terms that were then annotated for valence, arousal, and dominance.

To annotate the words, questionnaire were created in which the annotating user was asked to select the word with the highest and lowest valence, arousal or dominance value from a set of 4 words.

The questionnaire uses a set of paradigm words that signify the two ends of the valence dimension. The paradigm words were taken from past literature on VAD [17, 112], e.g. [Q1: Which of the four words below is associated

with the MOST valence value happiness / pleasure / positiveness / satisfaction / contentedness / hopefulness OR LEAST valence value unhappiness / annoyance / negativeness / dissatisfaction / melancholy / despair? (Four words listed as options.)]

Authors showed that *NRC VAD Lexicon* has split-half reliability scores of 0.95 for valence, 0.90 for arousal, and 0.90 for dominance. These scores are markedly higher than that of existing lexicons.

Dimension	Word	H Score	Word	L Score
valence	love	1.000	toxic	0.008
	happy	1.000	nightmare	0.005
	happily	1.000	shit	0.000
arousal	abduction	0.990	mellow	0.069
	exorcism	0.980	siesta	0.046
	homicide	0.973	napping	0.046
dominance	powerful	0.991	empty	0.081
	leadership	0.983	frail	0.069
	success	0.981	weak	0.045

Table 3.2: *A set of terms from NRC VAD Lexicon.* In this table are reported terms with the highest (H) and lowest (L) valence (V), arousal (A), and dominance (D) scores.

Chapter 4

NetME

This chapter is organized as follows.

Section 4.1 introduces *NetME* system together with its components.

Section 4.2 provides the technical details of the back-end and the front-end of *NetME*.

Section 4.3 reports three different case studies that allow evaluating *NetME*'s prediction qualitatively.

The first one is a comparison with the web application *Hetionet*, using *SemRep* software as ground truth.

The second case study is focused on: (i) recovering known gene interactions; (ii) avoid false-negative ones. For this purpose, we selected a subset of gene-gene interactions in *KEGG*/*Reactome* [72, 71, 70, 43] by making use of *STRING* API. More precisely, such interactions were obtained by selecting 100 random gene-gene interactions (manually curated in *KEGG* or *Reactome* database) for each of the following *STRING* text-mining score intervals: 500-600,600-700, 700-800, 800-900, $>=$ 900. Next, we selected the first 100 pairs

of non-interacting genes from the Negatome 2.0 database [16, 119] in order to understand if *NetME* can avoid false-negative interactions. The experiment yielded accuracy values from 58% when the *STRING* text-mining score is in [500, 600] interval, to 84% when the value of such a score is higher than 900.

Whereas, the third case study is focused on building a CD147-genes interaction network through selected papers containing valuable information about CD147 gene. We compared the network returned by *NetME* against a manually-curated network derived from these selected papers. The experiment yielded 98% sensitivity and 100% specificity. Therefore, both experiments clearly showed the high reliability of *NetME* inferred networks.

Moreover, we have also assessed the *NetME* performance for inferring CD147-diseases interactions by selecting 100 random interactions from *DisGeNET*, and the same abstracts used by *DisGeNET* for inferring these interactions. *NetME* detected 63 True Positive values out of 100, revealing a sensitivity of 63%

4.1 The *NetME* Model

A Knowledge Graph (also known as a semantic network) is a systematic way to connect information and data to knowledge. It represents a collection of interlinked descriptions of entities, real-world objects, and events, or abstract concepts, obtained from knowledge-bases such as ontologies $(O_1, O_2, \cdots, O_k)$. Basically, a semantic network is defined as a graph $G = (V, E)$ where entities are in V, and relationships in E. Each relation represents a connection between entities of one (intra-relationship) or more (inter-relationship) ontologies [94]. Therefore, there might exist a relation $e = (v_1, v_2) \in E$ where $v_1 \in O_i$ and

$v_2 \in O_j$ with $i \neq j$.

An ontology is a formal description of knowledge as a set of domain-based concepts in relationships among them. As a result, the ontology does not only introduce a shareable and reusable knowledge representation, but it can also provide new knowledge about the considered domain [137].

NetME builds a biomedical knowledge graph starting from a set of n documents obtained through a query to the *PubMed* database. Papers can be sorted by relevance (default) or publication date. Users can also provide a list of PMCID/PMID or a set of PDF documents. The inferred network contains biological elements (i.e., genes, diseases, drugs, enzymes) as nodes and edges as possible relationships.

In Figure 4.1 we outline the architecture of *NetME*. The user provides the query terms to perform the search on *PubMed*, and she may directly provide PDFs or PMCIDs/PMIDs of other pertinent documents.

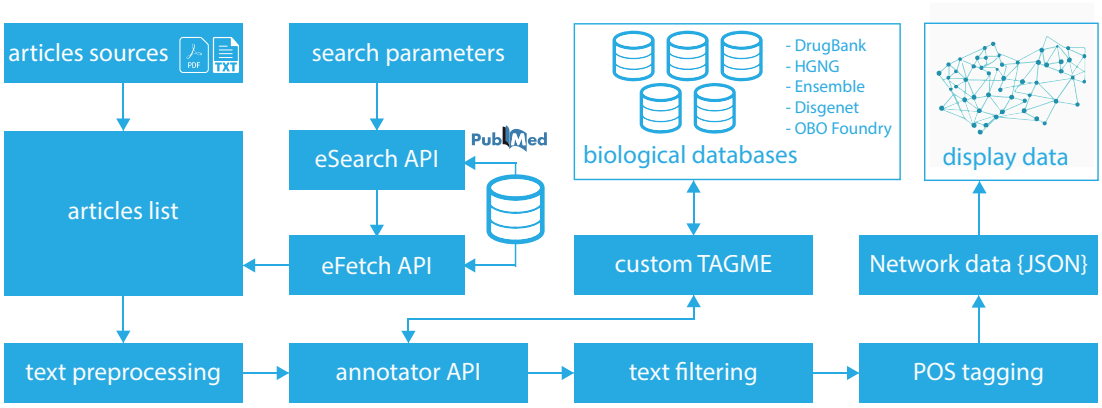


Figure 4.1: *NetME pipeline architecture*

Then *NetME* begins to create the network as follows:

1. First, *OntoTAGME* converts the full-text of the input documents into a list of entities (nodes) using literature databases and ontologies (such as *GeneOntology* [29], *DrugBank* [136], *DisGeNET* [102], and *Obofoundry* [120]) as corpus. These entities will be the knowledge graph nodes. Note that, *Obofoundry* contains a several ontologies, but only the following have been currently used in our model: GO, DO, PW, BTO, PRO, AEO, PATO, CL and CLO.

2. Next, in the first release of *NetME* [91] a *NLTK* [81] bottom-up and top-down approach were employed for building the syntactic tree of each document sentence, and to infer the relations among nodes. Starting from second release of *NetME*, an *NLP* model based on *Python spaCy* [62], and *NLTK* libraries, is executed to infer the relations among nodes entity-nodes belonging to the same sentence (S_i) or to the adjacent ones (S_i, S_{i+1}) of the same document. Such relationships indicate disease treatment, genes regulations, molecular functions, gene-gene interactions, gene-disease interactions, gene-drug interactions, drug-disease interactions, disease-disease interactions and drug-drug interactions.

The final network will contain both directed and undirected edges according to the predictions made by the model. At the end of the process, the network will be rendered through *CytoscapeJS*. The following two subsections provide the details of these two phases.

4.1.1 Network edge inference

Once the network nodes have been extracted the system will annotate their position and their main characteristics within the text. The system capture the significant elements in each sentence, by making use of the parts of speech (POS tags), then through a syntactic analysis it verify the coherence of the extracted elements. Indeed, sentences have an internal organization that can be represented using a tree. Solving a syntax analysis problem for a sentence consists of looking for predefined syntactic forms which, like a tree, branch out from the single words.

The main syntactic form is the sentence (S) which contains noun phrases (NP) or verb phrases (VP) that are formed by further elementary syntactic forms such as nouns (N), verbs (V), determiners (DET), etc (see Table 4.2). All these information will be used by the textual analysis phase to infer relations between them.

In the first release of *NetME* [91] a *NLTK* left-corner parser approach, which integrates both the bottom-up and the top-down approaches, were employed for building the syntactic tree of each document sentence. First, a left-corner parser pre-processes the context-free grammar to build a table where each row contains two cells. The first one holds a non-terminal category, and the second cell holds the collection of possible left corners of that non-terminal; e.g. in a grammar production like $S \rightarrow NP\ VP$, we store S as non-terminal category and NP as possible left corner. Next, it parses each phrase higher syntactic forms, filtering the results starting from the smallest text units.

Starting from the second version of *NetME*, were adopted a more per-

forming approach, to check the syntactic coherence and build the syntactic tree. A transition-based dependency parser is used. The dependency parser component inside the *spaCy* library jointly learns sentence segmentation and labelled dependency parsing. The parser uses a variant of the non-monotonic arc-eager transition-system [61], with the addition of a break transition to perform the sentence segmentation. Nivre's [97] pseudo-projective dependency transformation is also used to allow the parser to predict non-projective parses.

The parser is trained through an imitation learning objective. It follows the actions predicted by the current weights and, at each state, it determines which actions are compatible with the optimal parse that could be reached from the current state. The weights are updated in a way that the scores assigned to the set of optimal actions is increased, while scores assigned to other actions are decreased. Note that more than one action may be optimal for a given state.

Once *OntoTAGME* have extracted the set of nodes $n_1, \ldots, n_z$ from a list of N full-text documents $[p_1, p_2, \ldots, p_N]$, the edge inference module of *NetME* (developed on top of the *Python* library *NLTK* [81] and *spaCy* [62]) starts to establish any verbal relationships between those pairs of nodes. When two or more nodes are detected within a sentence or adjacent sentences, the syntactic analyzer extracts the parts of speech and syntactic dependencies within the sentence. For each sentence we then get a set of labelled tokens $lt_1, lt_2 \ldots, lt_{k_i}$. Each token is a tuple of the following form $\{token, POS, dependency\_label\}$, where POS and Dependency label are valued with the data present in Table 4.2.

Irrelevant POS are filtered out (stop-words, URLs, etc.), we keep only the

useful verb forms and the nodes which correspond to the noun parts. A final pruning phase is also executed in which we use: (i) POS tag labels and dependency labels to check if the syntactic link between the verb form and the annotations is correct and consistent, as described in the Figure 4.2; (ii) a dictionary of biological verb forms to check if they are pertinent. The surviving nodes and verb forms will allow to generate network edges.

In our final network, each edge $e = (a, b)$ is weighted with three parameters: the term frequency and inverse document frequency (tf.idf), the medium relatedness (*mrho*) and the biological degree (bio). More specifically, tf.idf is a measure of how much information the edge provides, namely if it is common or rare across all input documents. In formula, we compute

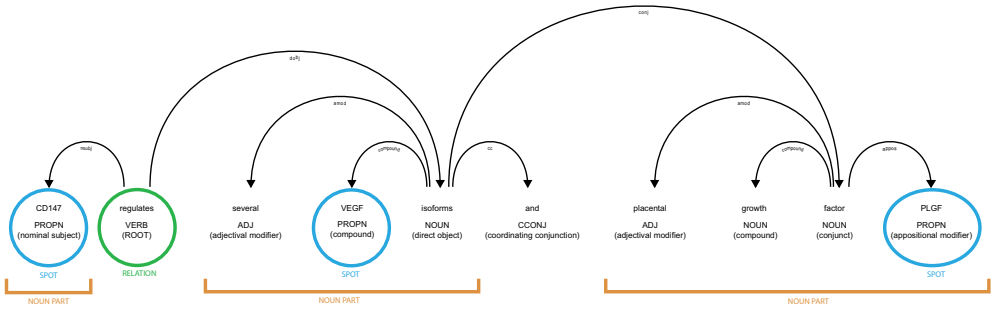


Figure 4.2: *NetME example of POS extraction and coherence checking*. Process described for the sentence *[...] CD147 regulates several VEGF isoforms and placental growth factor (PLGF), and it has unique effects on trophoblastic function.[...]*. Through *OntoTAGME* we detect the spots [*"BSG"*, *"VEGFA"*, *"PGF"*]. After the syntactic analysis, three noun parts are identified (the phrase spots, highlighted via orange segments): two of them ("VEGF" and "PLGF") have a joint relationship with the first ("CD147"). The verbal part is the root between the two pairs of nouns ("CD147" - "VEGF"), (CD147 - "PLGF").

$tf.idf(e, p, P) = tf(e, p) * idf(e, P)$.

Where, term frequency $tf(e, p)$ is the frequency of edge e, is defined as $tf(e, p) = f_{e,p}/\sum_{e' \in p} f_{e',p}$, with $f_{e,p}$ representing the number of times that edge e occurs in paper p. The inverse document frequency $idf(e, P)$ is a measure of how much information the edge e provides. It is defined as $idf(e, P) = log^N/|\{p \in P : e \in p\}|$, where N is the number of documents analyzed by the query such that $N = |P|$, and $|\{p \in P : e \in p\}|$ is the number of documents where the edge e appears. The parameter *mrho* measures the relatedness of the labels starting from the ρ value assigned by *OntoTAGME* to the two annotations involved, i.e. $mrho(e) = \frac{\rho_a * \rho_b}{2}$. The *bio*-parameter is the cosine similarity (having a value ranging from 0 to 1) between the inferred relationship and a set of biological verb forms (see Table 4.1). Figure 4.3 provides an example of such an annotation.

Verb Forms		
activate	downregulate	reduce
affect	enhance	regulate
associates	express	release
block	find	reveal
cause	inactivate	stimulate
contain	increase	trigger
control	induce	ubiquitination
decrease	interacts	upregulates
detect	overexpress	
display	produce	

Table 4.1: *List of biological verb forms used in NetME*

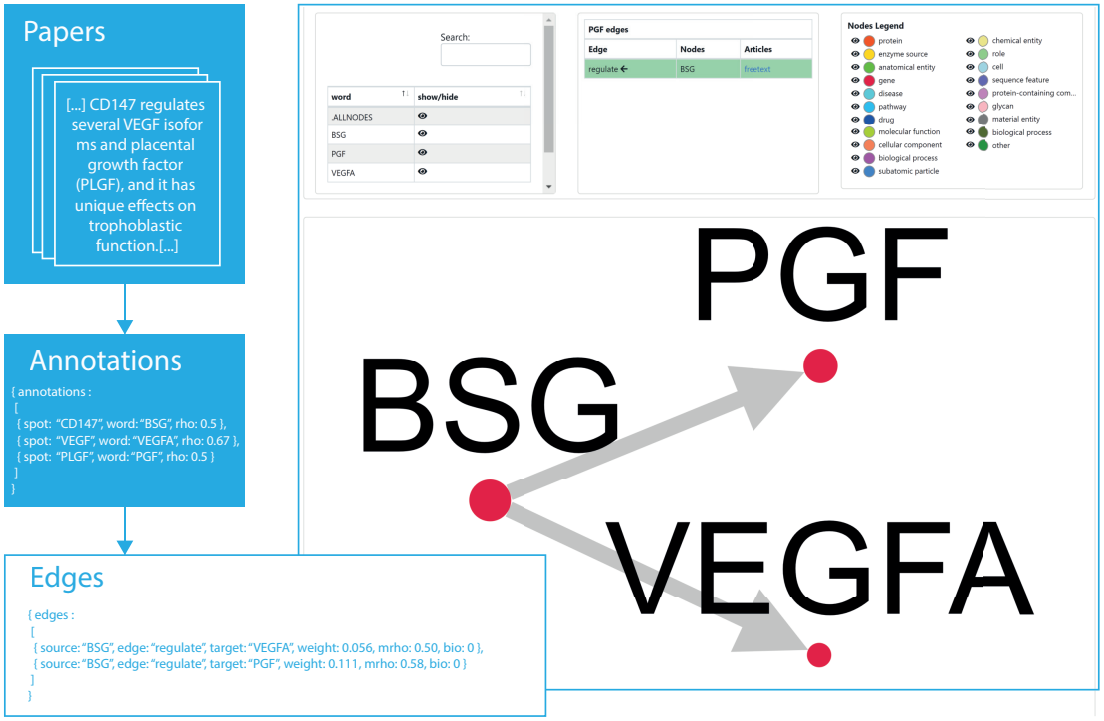


Figure 4.3: *Example of annotation of the sentence [...] CD147 regulates several VEGF isoforms and placental growth factor (PLGF), and it has unique effects on trophoblastic function.[...].* Through *OntoTAGME* we detect the spots [*"BSG"*, *"VEGFA"*, *"PGF"*], and after the syntactic analysis and noise reduction steps, we detect two valid edges: [*"BSG"*, *"regulate"*, *"VEGFA"*] and [*"BSG"*, *"regulate"*, *"PGF"*]. Note that "regulate" is a biological verb forms and it has bio parameter set to 0.

POS tag		Dependency label	
Symbol	Meaning	Symbol	Meaning
ADD	email	acl	clausal modifier of noun (adjectival clause)
AFX	affix	acomp	adjectival complement
CC	conjunction, coordinating	advcl	adverbial clause modifier
CD	cardinal number	advmod	adverbial modifier
DT	determiner	agent	agent
EX	existential there	amod	adjectival modifier
FW	foreign word	appos	appositional modifier
HYPH	punctuation mark, hyphen	attr	attribute
IN	conjunction, subordinating or preposition	aux	auxiliary
JJ	adjective	auxpass	auxiliary (passive)
JJR	adjective, comparative	case	case marking
JJS	adjective, superlative	cc	coordinating conjunction
LS	list item marker	ccomp	clausal complement
MD	verb, modal auxiliary	compound	compound
NFP	superfluous punctuation	conj	conjunct
NN	noun, singular or mass	csubj	clausal subject
NNP	noun, proper singular	csubjpass	clausal subject (passive)
NNPS	noun, proper plural	dative	dative
NNS	noun, plural	dep	unclassified dependent
PDT	predeterminer	det	determiner
POS	possessive ending	dobj	direct object
PRP	pronoun, personal	expl	expletive
PRP$	pronoun, possessive	intj	interjection
RB	adverb	mark	marker
RBR	adverb, comparative	meta	meta modifier
RBS	adverb, superlative	neg	negation modifier
RP	adverb, particle	nmod	modifier of nominal
SYM	symbol	npadvmod	noun phrase as adverbial modifier
TO	infinitival "to"	nsubj	nominal subject
UH	interjection	nsubjpass	nominal subject (passive)
VB	verb, base form	nummod	numeric modifier
VBD	verb, past tense	oprd	object predicate
VBG	verb, gerund or present participle	parataxis	parataxis
VBN	verb, past participle	pcomp	complement of preposition
VBP	verb, non-3rd person singular present	pobj	object of preposition
VBZ	verb, 3rd person singular present	poss	possession modifier
WDT	wh-determiner	preconj	pre-correlative conjunction
WP	wh-pronoun, personal	predet	None
WP$	wh-pronoun, possessive	prep	prepositional modifier
WRB	wh-adverb	prt	particle
		punct	punctuation
		quantmod	modifier of quantifier
		relcl	relative clause modifier
		xcomp	open clausal complement

Table 4.2: *List of spaCy POS tag and syntactic dependency labels.*

4.2 The annotation tool

NetME is provided with a front-end developed in PHP and Javascript, in which the network rendering is performed through the *CytoscapeJS* library [46]. Its back-end, which integrates *OntoTAGME*, is written in Java and communicates with both *Python NLTK* [81] and *spaCy* [62] libraries for the *NLP* module. *PubMed* search is performed with the Entrez Programming Utilities [39], a set of server-side programs providing a stable interface to the Entrez database and to the query system at the National Center for Biotechnology Information (*NCBI*).

NetME is equipped with an easy-to-use web interface providing three major functions (see Figure 4.4): (i) *PubMed* query-based network annotation; (ii) user-provided free-text network annotation; (iii) user-provided PDF documents network annotation.

In the *query-based network annotation*, the user provides a list of keywords, which are employed to run a query on *PubMed*, or a list of article ids. The top resulting papers are retrieved and then the network inference procedure is run. Several parameters can be set by the user (or left with default values) such as: the number of top article to retrieve from *PubMed*, and the criteria used to sort papers (relevance or date).

In the *user-provided free-text network annotation*, users provide a free text which is then input to the network inference procedure.

In the *user-provided PDF documents network annotation*, users give a set of PDF documents which are then input to the network inference procedure.

The result of the network inference procedure is a direct graph (network)

which shows all inference details in three main tables containing: the list of extracted papers, the list of annotations, and the list of edges together with their weight.

The user can then click on a node of the network to view all incoming and outgoing connections, or she can click on an edge to display its type and the verbal relation between the nodes it connects.

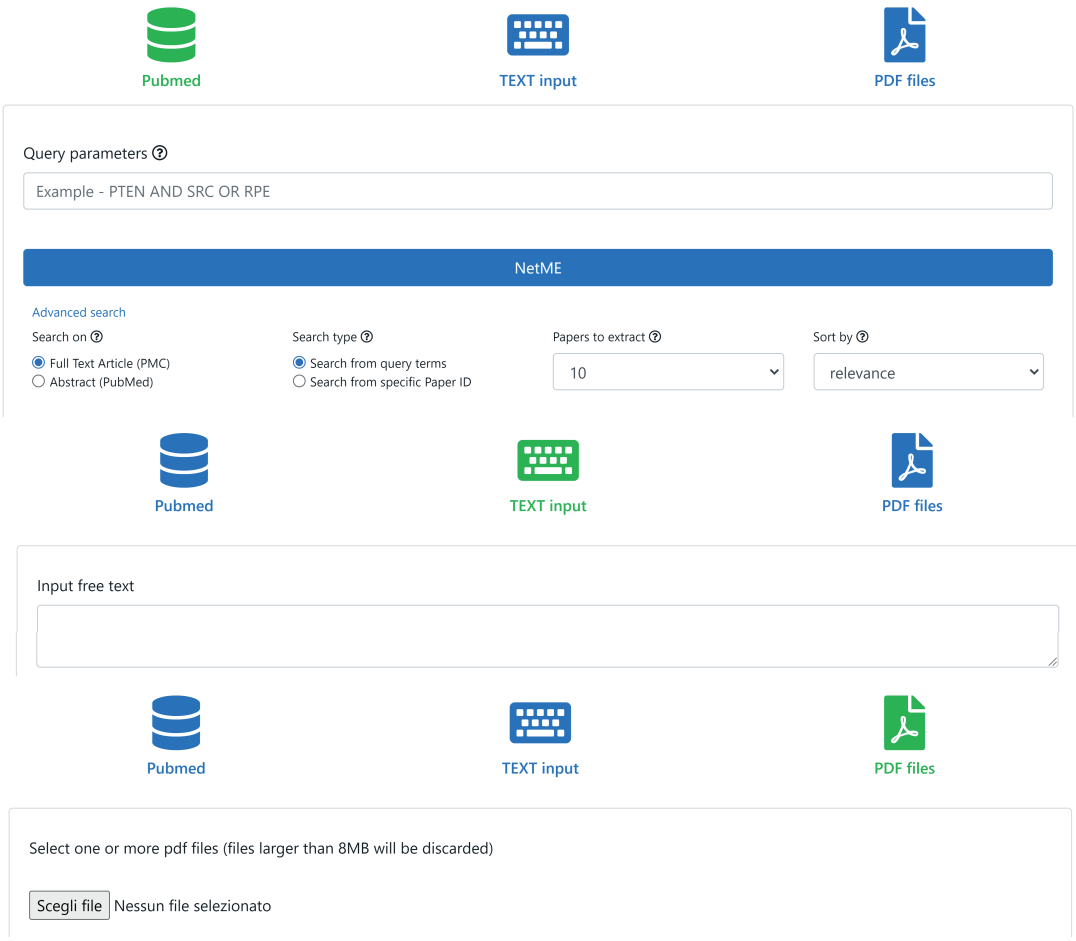


Figure 4.4: *NetME web interface*

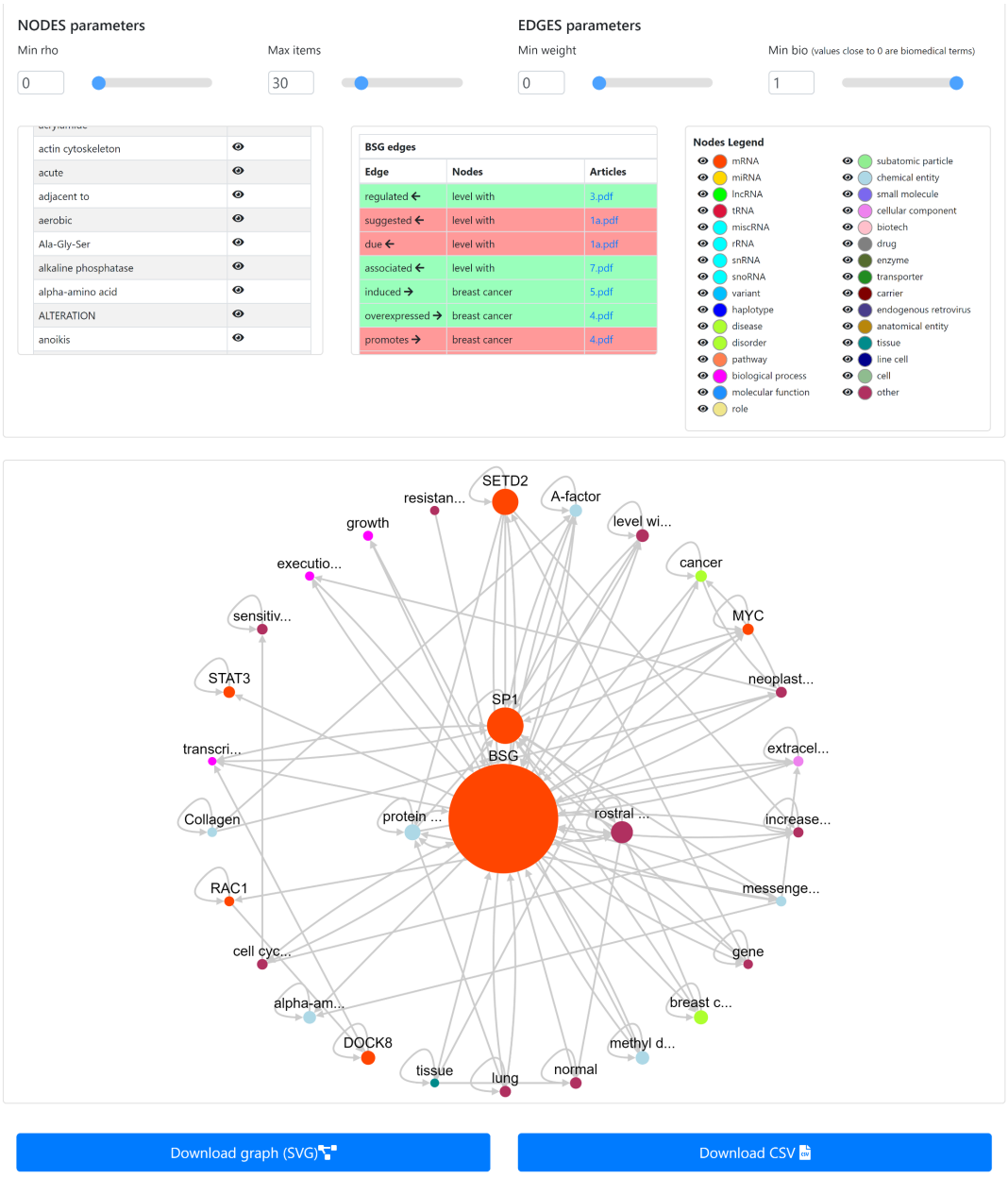


Figure 4.5: *Network example generated by NetME.* In the first top panel there are the filtering parameters of the nodes and edges of the network. On the top right there is a panel to enable/disable nodes, in center a panel for more information on the selected nodes or edges, on the right instead a panel to enable/disable the node categories.

4.3 Experimental Analysis

To analyze the reliability of *NetME* knowledge graphs, we performed three case studies.

The first one, conducted with the previous version of *NetME* [91], is a comparison with the web application *Hetionet*, using *SemRep* software as ground truth. We built a network of 200 nodes from the Top-20 *PubMed* articles by using the query *SRC* (Proto-oncogene tyrosine-protein kinase).

The second case study, conducted with the second version of the *NetME*, aims to providing a comprehensive analysis of *NetME* performance by checking its ability to predict known relations between genes drawn from Kyoto Encyclopedia of Genes and Genomes - *KEGG* [72, 71, 70] or *Reactome* [32, 69, 31] pathways and, on the other hand, its ability to avoid inferring false connections between proteins by using the Negatome 2.0 database [16, 119].

The third case study, also conducted with the second version of *NetME*, is more specific and focuses on building a network based on some selected publications that contain valuable information specific to the CD147 gene. Such a network is then compared against a manually-curated one derived from the same papers by a bio-expert. In both cases, the performance of *NetME* has been measured in terms of a precision/recall curve.

4.3.1 Case study 1

In the fisrt case study we built a network by using the query *SRC* (Proto-oncogene tyrosine-protein kinase), as shown in Figure 4.6.

We choose to build a network of 200 nodes from the Top-20 *PubMed* arti-

cles. ρ has been set to 0.3 which are suitable to perform an interactive test.

We can observe several interesting edges between the nodes. For each node having the gene *SRC* as source or destination we performed a comparison with HetioNet. The results are reported in table 4.4.

For a more systematic comparison using SemRep as ground-truth, we built a network of 200 nodes from the Top-20 *PubMed* articles using the previous query. ρ has been set to 0.3. Results from SemRep were obtained using the same list of *PubMed* articles obtained with our query.

For each node and edge in SemRep results, we checked whether *NetME* was able to infer it. The same analysis has been done using HetioNet. In this

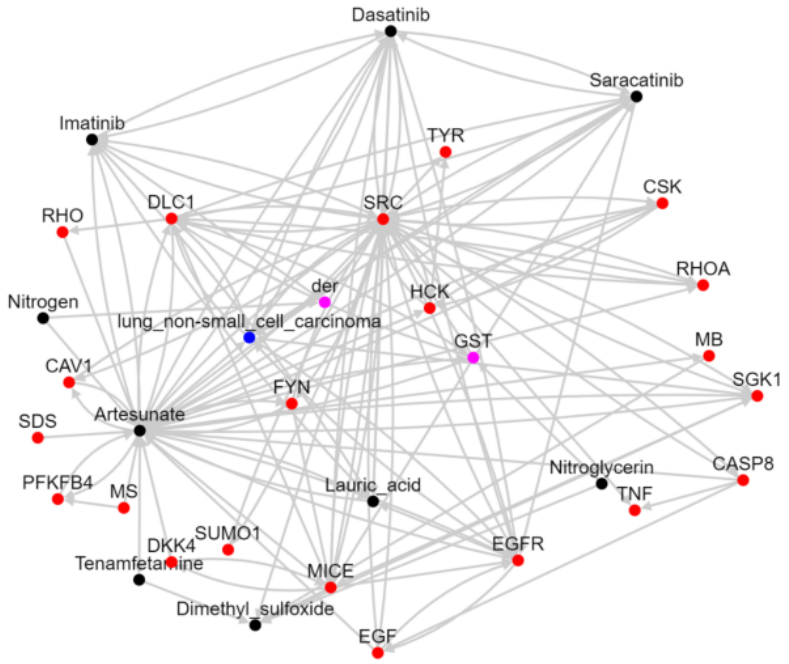


Figure 4.6: *A sample of NetME with Proto-oncogene tyrosine-protein kinase (Src).* The partial list of edges between the nodes is reported in Table 4.4

analysis edge labels were ignored.

Finally, all comparisons were performed in terms of Precision and Recall. All results are showed in Table 4.3.

Our analysis, although preliminary, clearly shows that *NetME* has a comparable precision with *Hetionet* in terms of inferred nodes and an higher recall. *Hetionet* precision is slightly higher although *NetME* recall is much higher in terms of inferred edges.

	NETME	HetioNet	SemRep (Ground truth)
Detected nodes	200	112	189
Valid nodes (TP)	123	61	189
Wrong nodes (FP)	77	51	–
Missing nodes (FN)	66	128	–
Precision	61.5%	54.5%	–
Recall	65.1%	32.3%	–
Detected edges	495	56	292
Valid edges (TP)	178	34	292
Wrong edges (FP)	317	22	–
Missing edges (FN)	114	258	–
Precision	56.2%	60.7%	–
Recall	61%	11.6%	–

Table 4.3: *NetME vs Hetionet comparison performance results*. *NetME* has a comparable precision with *Hetionet* in terms of inferred nodes and an higher recall. Concerning inferred edges, *Hetionet* precision is slightly higher although *NetME* recall is much higher.

Nodes	incoming edges' labels		outgoing edges' labels	
	NetME	HetioNet	*NetME*	HetioNet
Artesunate	activate		regulates, cause	
CSK	regulates	interact	regulates, interact	regulate
DLC1	interact, reactivate		activate	interact
TNF	activate			regulates
CAV1	includes, binds	interact	interact	interact
EGFR	associate, include	interact	includes	regulates
SGK1	decrease		regulates	
RHOA	regulates	regulates		interact
Dasatinib	binds	binds	inhibit	binds
CASP8	inhibit		activate	interact
EGF	upregulates	regulates	activate	
Nsclc	treat		activate, include	
FYN	activate	interact		interact

Table 4.4: *List of nodes and labeled edges connected with SRC*

4.3.2 Case study 2

The second case study focuses on assessing *NetME* performance through its capability to recover known gene interactions. For this purpose, we selected a subset of gene-gene interactions from *KEGG/Reactome* by making use of *STRING* API. More precisely, such interactions were obtained by selecting 100 random gene-gene interactions as a true-positive set, for each of the following *STRING* text-mining score intervals: 500-600, 600-700, 700-800, 800-900, $>=$ 900 (listed in Supplementary 1.json, Supplementary 2.json, Supplementary 3.json, Supplementary 4.json, Supplementary 5.json, files respectively. Files are available at https://netme.tk/supplementary_material.zip).

Next, we selected 100 random pairs of non-interacting genes from the Negatome 2.0 database as a true-negative set (listed in Table 4.5). For each

interacting gene-pairs, we queried *NetME* with the papers used by *STRING* to infer the interactions. On the other hand, to annotate non-interacting genes, we queried *NetME* with the pair of genes of interest, selecting the top 20 papers from *PubMed*. Accuracy, sensitivity, specificity and PPV values, detected by *NetME*, are listed in Table 4.6.

The results clearly show that *NetME* produces reliable results when the

Non-interacting genes from Negatome 2.0							
SOURCE	TARGET	SOURCE	TARGET	SOURCE	TARGET	SOURCE	TARGET
AKT1	TSC1	CTNND1	APC	MAD2L2	MAD1L1	TANK	RBCK1
ARAF	BCL2L1	CTNND1	CTNNA1	NCK1	EGFR	TBC1D7	TSC2
ARAF	BCL2	CTNND1	CTNND1	OSM	LIFR	TFDP1	CDK2
BCL10	BIRC3	CTNND1	CTNNB1	PARD3	LIMK1	TFDP1	CCNA1
BCL2L1	MAVS	DKK1	WNT1	PDGFC	FLT1	TICAM1	TLR4
BMPR1A	TGFB1	DKK1	SOST	PFN4	ACTB	TJAP1	F11R
BMPR1A	BMP5	DVL1	TSC1	PGF	KDR	TJAP1	CLDN1
BMPR1A	BMP6	EIF3I	ACVR2A	PIAS3	STAT1	TJAP1	TJP1
BMPR1B	TGFB1	EIF3I	ACVR1	PIK3CG	PIK3R2	TNF	EGFR
BMPR1B	BMP5	EIF3I	TGFBR1	PKN1	RPS6KA1	TRADD	TNFRSF10A
BMPR1B	BMP6	EP300	CD44	PKN1	RPS6KA3	TRADD	TNFRSF10B
BMPR2	BMP2	ERBB2	PIK3R2	PKN1	MAP3K2	TRAF6	IRF3
CCND1	MCM2	ETS1	CREBBP	PKN2	RPS6KA1	TSC1	CDKN1B
CCR3	CCL3	FOXO1	TSC1	PKN2	RPS6KA3	VAV1	SHC1
CCR3	CCL4L2	GRAP2	SOS1	PKN2	MAP3K3	VEGFB	KDR
CD274	CD28	GRAP2	CBL	RB1	SMAD3	VEGFB	FLT4
CD274	CTLA4	HDAC2	RELA	RBL2	SMAD3	VEGFC	FLT1
CD274	ICOS	HIPK2	MDM2	RIPK1	TNFRSF10A	VIPR2	RAMP1
CD3G	ZAP70	HSPA4	BAX	RIPK1	TNFRSF10B	VIPR2	RAMP2
CD74	NOTCH1	IGF2	IGF1R	SFN	TSC1	VIPR2	RAMP3
CDKN1B	TSC1	IL15	IL2RA	SH3KBP1	TNFRSF14	VWF	F8
CSF2	IL3RA	IL1A	EGFR	SMAD1	ANAPC10	YWHAB	TSC1
CTNNB1	HSP90AA1	IL22	IL10RA	SMAD4	ANAPC10	YWHAE	TSC1
CTNNB1	DDIT3	IL4R	IL13	SOCS3	JAK2	YWHAZ	TSC1
CTNND1	IL2	KDR	FLT1	STIM1	TRPC6	NFKBIA	CREB3L2

Table 4.5: *List of the first 100 pairs of non-interacting genes from the Negatome 2.0 database.* The column "SOURCE" indicates the starting gene, instead the column "TARGET" indicates the gene to which the action of the source gene is directed.

annotations are performed on top of relevant literature (*STRING* text-mining score higher than 700). On the other hand, when the *STRING* text-mining score is lower than 700, the *NetME* performances degrade in accordance with *STRING* predicted confidence as highlighted by their score. The reason behind such a behaviour is due: (i) not enough literature about these interactions; (ii) the interactions have been inferred by human curators as a combination of other interactions occurring in the text.

Furthermore, when the text-mining score is small, *STRING* predictions could be wrong. In fact, as reported in [123], a score of 500 would indicate that roughly every second term of an interaction might be erroneous (i.e., a false positive). Therefore, the computed value of accuracy, sensitivity, specificity and PPV could be incorrect.

text-mining score interval	accuracy	sensitivity	specificity	PPV
500-600	58.5%	31%	86%	68.8%
600-700	66.5%	47%	86%	77.05%
700-800	72.5%	59%	86%	80.8%
800-900	73.5%	61%	86%	81.3%
$>= 900$	84%	82%	86%	85.4%

Table 4.6: *Metrics on NetME's ability to predict known interactions (from KEGG/Reactome) and non-interactions (from Negatome 2.0) between genes.*

4.3.3 Case study 3

Many tools [6] and computational models rely on existing network databases, such as *KEGG* [72, 71, 70] and *Reactome* [32, 69, 31]. However, despite the

enormous amount of available data, these databases are still incomplete and therefore have partial information [86]. As an example, *KEGG* includes approximately one-third of the known genes.

In this case study, we have chosen CD147, also known as Basigin (BSG) or EMMPRIN, as a starting point for the gene-gene interactions network construction. This gene represents an example of a biological element that should be supplemented to the *KEGG* network since it is not currently described in their pathways. Among the bibliography consulted to build the network manually, we have carefully selected 11 papers containing a significant amount of helpful information for our purpose. On the other hand, in this case study, we have also assessed the capabilities of *NetME* in inferring CD147-diseases relations. For this purpose we selected 100 random interactions from *DisGeNET* [102], as well as the same abstracts used by *DisGeNET* for inferring such interactions (listed in Supplementary 6.json, file is available at https://netme.tk/supplementary_material.zip).

CD147 is a transmembrane glycoprotein of the immunoglobulin superfamily, expressed in many tissues and cells, which is known to participate in several high biological and clinical relevance processes and is a crucial molecule in the pathogenesis of several human diseases [138]. Recently Wang et al. [128] discovered an interaction between host cell receptor CD147 and SARS-CoV-2 spike protein, together with Angiotensin-Converting Enzyme 2 (ACE2), as an entry point for SARS-CoV-2.

In this direction, CD147 is an example of how a missing crucial gene within a biological network can compromise scientists' efforts to understand certain molecular phenomena. In literature, there are many valuable tools [59, 58] to integrate the missing information into bio-databases, such as *KEGG*. How-

ever, the most reliable approach in terms of accuracy and updated information remains the manual curation of such networks through careful and time-consuming literature analysis. On the other hand, a manually constructed network provides partial information due to the limited number of articles that a scientist could read.

Our second case study affords this issue by providing a practical example of how *NetME* can create valuable networks by analyzing quickly and automatically larger sets of publications. The set of 11 selected papers, described in Figure 4.7, was analyzed by a bio-expert to derive a CD147-genes interactions network manually. This process resulted in 50 genes and 64 interactions, as shown in Figure 4.7.

Next, by using the same set of papers, we run *NetME* with no upstream filter. The automatically generated network consisted of 86 genes and 139 relationships between them (see Figure 4.7 - 4.8). As the manually curated network consists of genes and proteins, only elements from these two categories were selected for the evaluation. This was performed by considering edges with the lowest "bio" score for each node pair.

Qualitatively, this network includes most of the interconnections mentioned in the papers, thus providing a reliable and comprehensive overview of the molecular function of Basigin. Quantitatively, *NetME* achieved an accuracy of 98.99%, a sensitivity of 100%, a specificity of 98.98%, and a positive predicted value of 46.32%.

NetME shows that CD147 is a potent inducer of metalloproteinases (MMPs) such as MMP2, MMP14 and MMP9 as reported in [138, 111, 35]. Furthermore, the overexpression of CD147, which results in increased phosphorylation of PI3K(PIK3CA), Akt(AKT1), leads to the secretion of vascular en-

dothelial growth factor (VEGFA) in several biological contexts such as KSHV infection [138] [111]. In addition to its ability to induce MMPs, CD147 regulates spermatogenesis, lymphocyte reactivity and MCT system, in particular MCT1 and MCT4 (MCTS1 and SLC16A4) expression [138] [75]. Our results also show that CD147 can increase the expression of ATP-binding cassette transporter G2 (ABCG2) protein, regulating its function as a drug transporter, as mentioned by Xiong et al. for MCF-7 cells [138]. *NetME* identifies also BSG as an upstream activator of STAT3, highlighting its involvement in tumor development in agreement with the literature [130]. As summarized by our knowledge network, CD147 is regulated by various inflammatory mediators, such as RANKL (TNFSF11), denoting its involvement in inflammatory processes [50] [111]. Among the potential activators of BSG, *NetME* also find the transcription factor c-Myc (MYC) [76]. Figure 4.9 - 4.10 - 4.11 depicts the precision/recall curve (AUC 0.997), the sensitivity/specificity curve and the True positive rate/False Positive Rate one. The construction of the curves considered all possible gene-pairs and their edges.

Finally, we queried *NetME* with the selected 100 random CD147-diseases interactions in *DisGeNET*, selecting the same *PubMed* abstract used by *DisGeNET* for inferring those interactions. *NetME* detected 63 True Positive values out of 100, revealing a sensitivity of 63%

It is essential to stress that *NetME* allows us to extract a satisfactory and valid amount of information in a few minutes, compared to a manual search that may take days or weeks. We also believe that this case study is significant because, in the evaluation, we considered not only the presence of a link between two nodes but even more closely the type of edge, hence the adequacy and specificity of the annotated edge in its biological context.

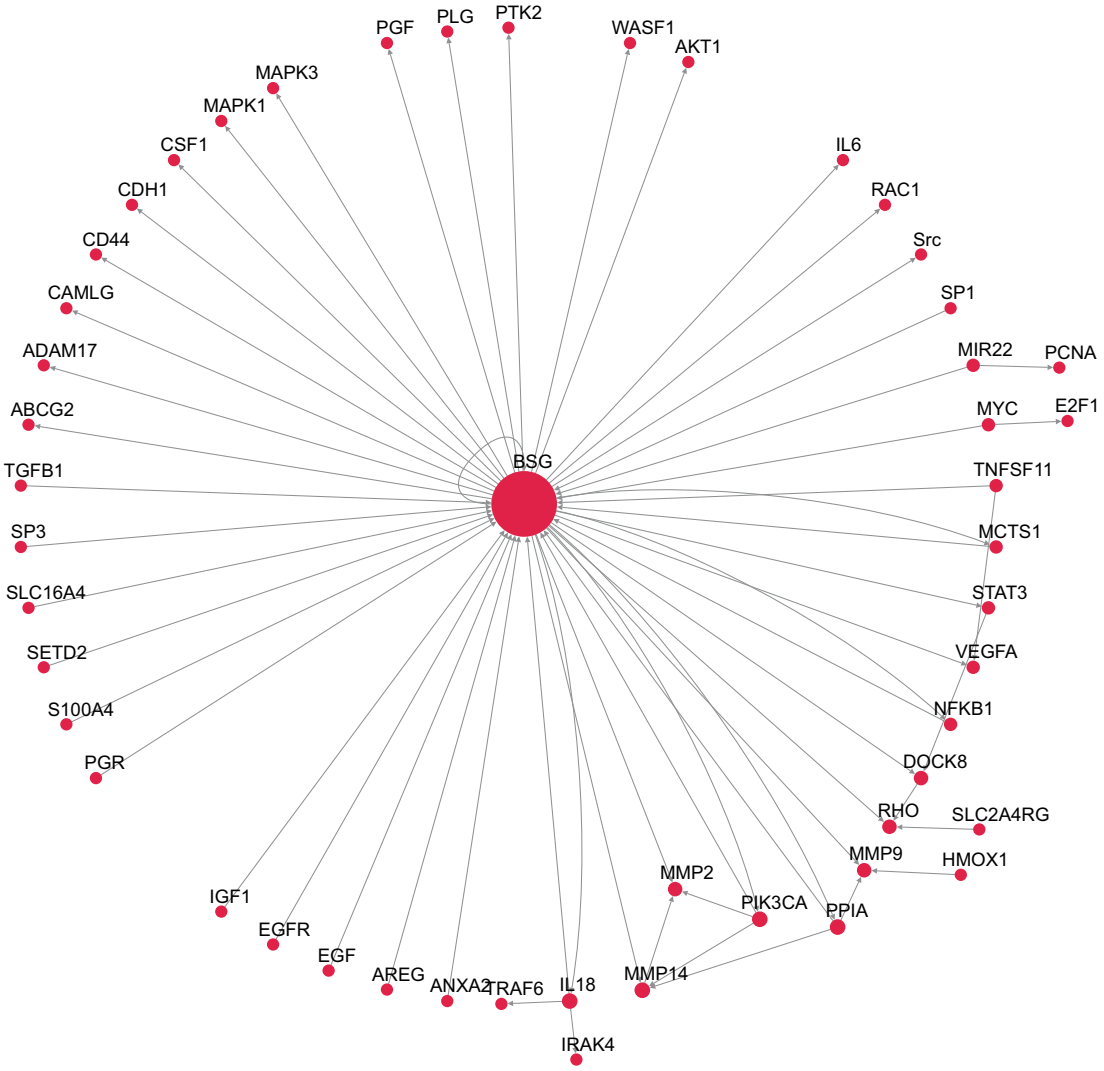


Figure 4.7: *Handmade Pathway.* Pathway constructed by hand from the selected papers [68, 76, 73, 50, 138, 111, 35, 126, 130, 76, 75], with CD147 (BSG) as the central node.

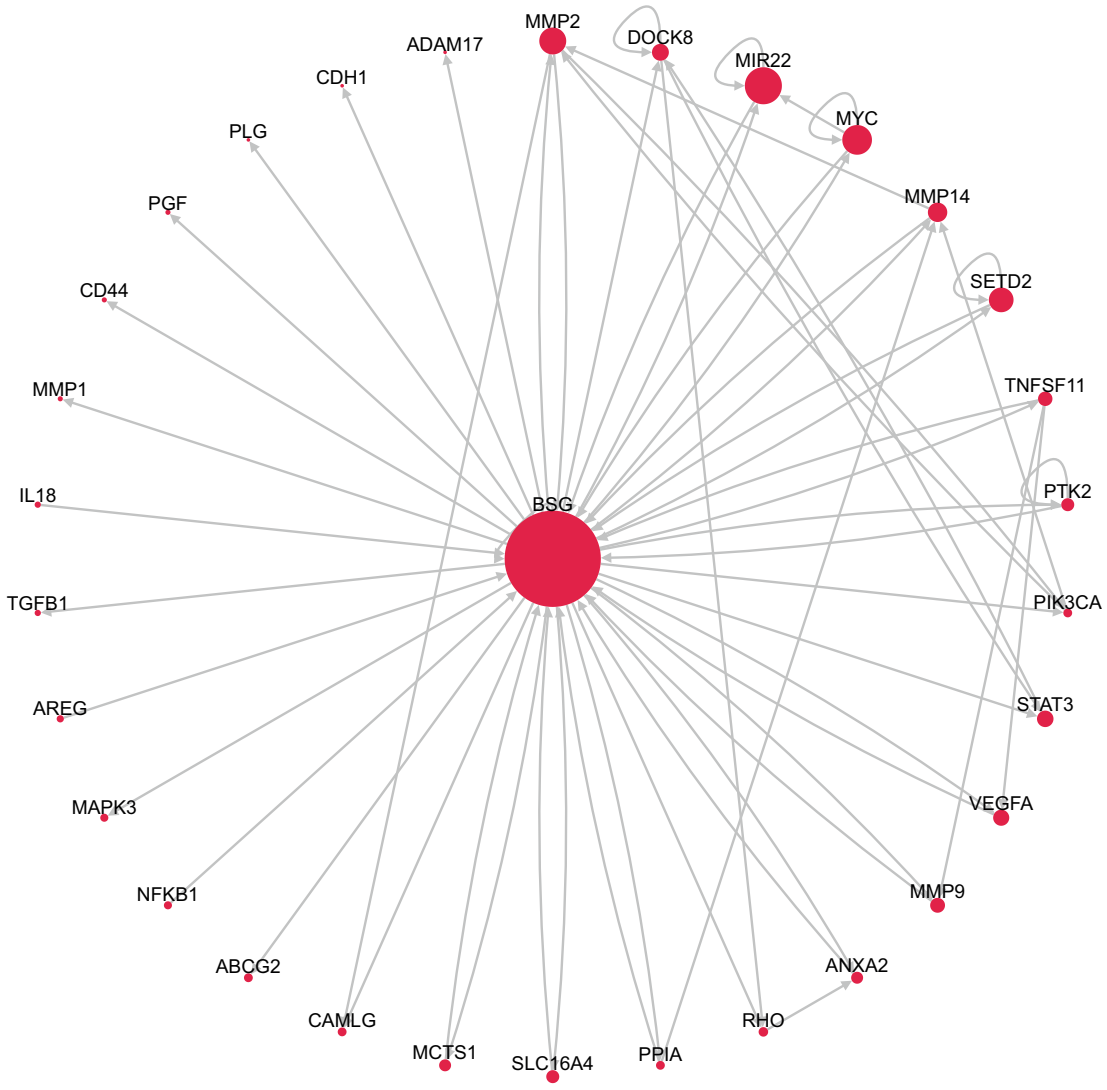


Figure 4.8: *NetME BSG network.* Molecular mechanisms summarised in the knowledge network developed by *NetME* in accordance with the same papers used in Figure 4.7.

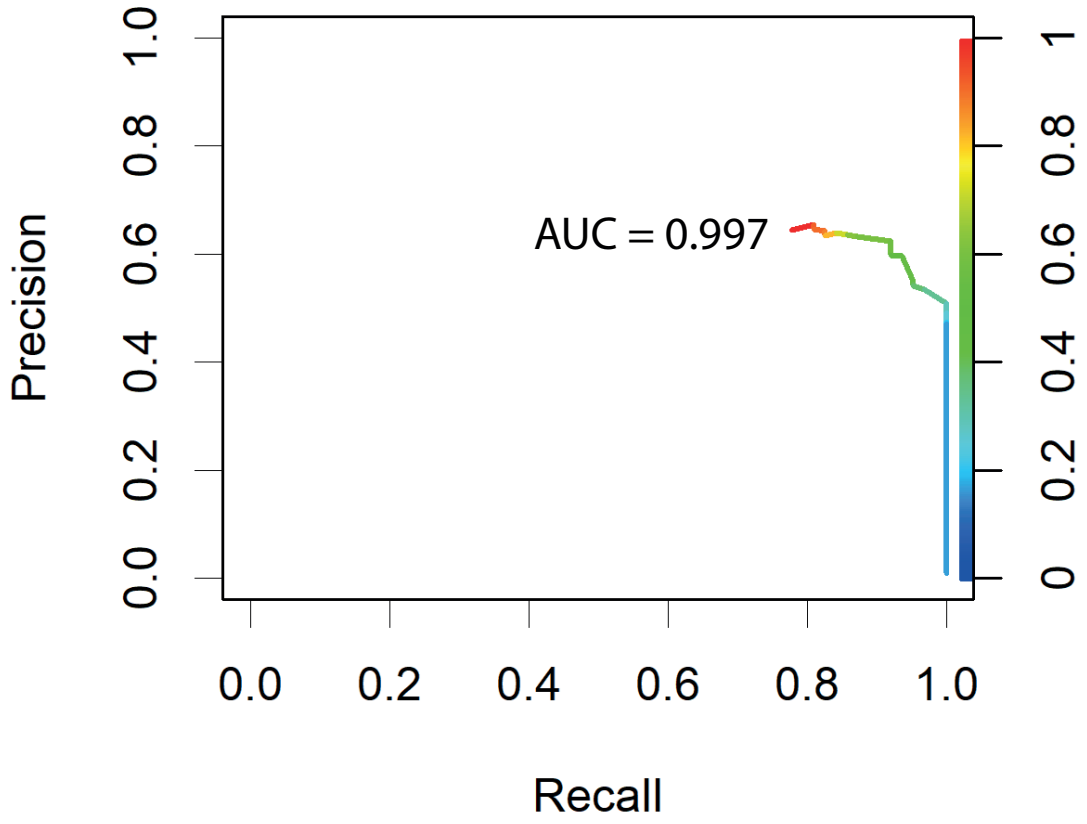


Figure 4.9: *Metrics of BSG-network performed by NetME.* The plots show Precision/Recall curve

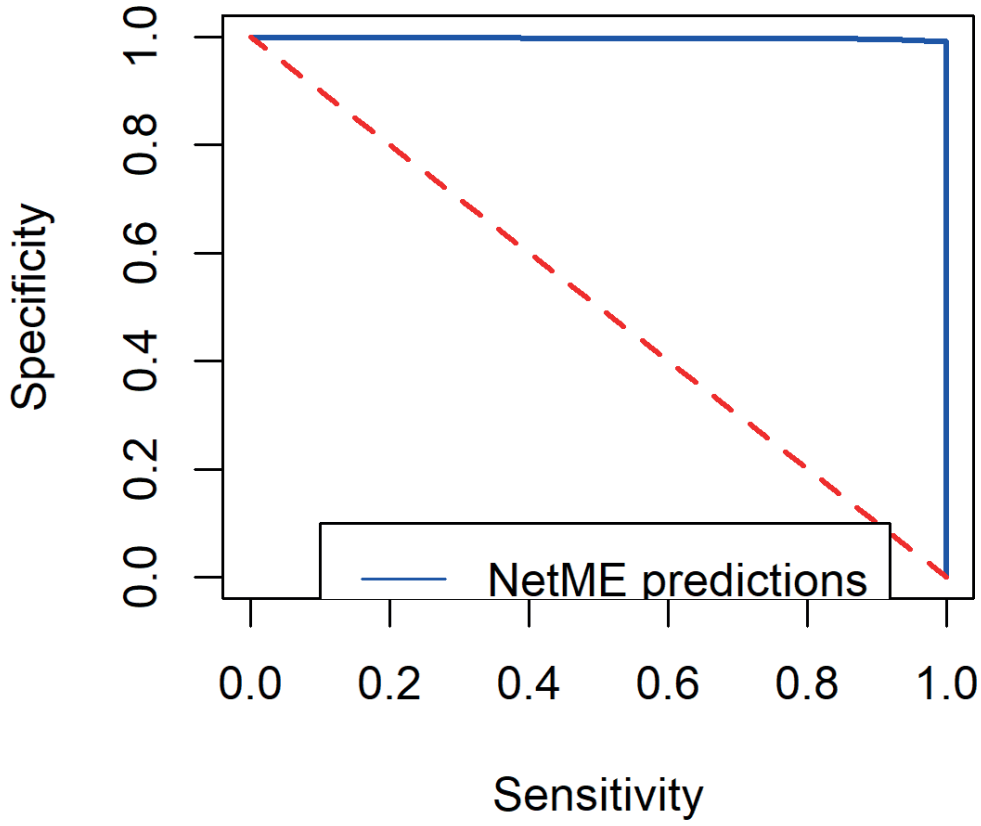


Figure 4.10: *Metrics of BSG-network performed by NetME.* The plots show Sensitivity/Specificity. The red dashed line indicates the expected result if the used method was random that is any method which, given a pair of nodes, elects whether between them there is a link with a probability of 0.5.

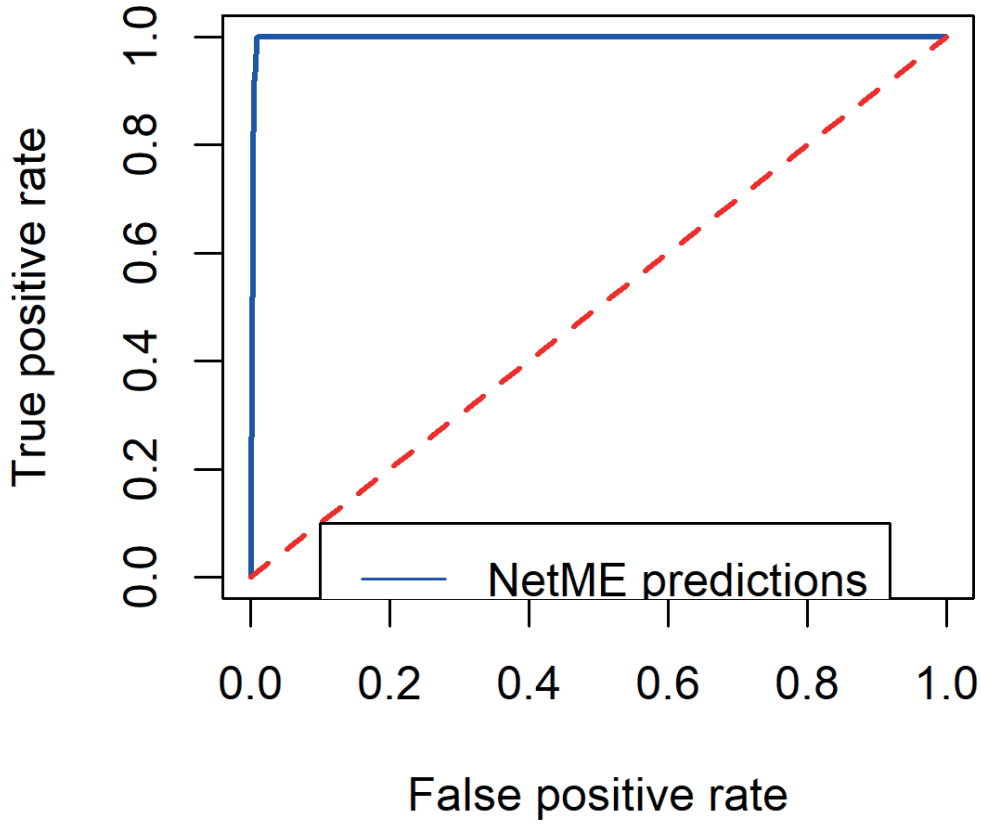


Figure 4.11: *Metrics of BSG-network performed by NetME.* The plots show True positive rate/False Positive Rate. The red dashed line indicates the expected result if the used method was random that is any method which, given a pair of nodes, elects whether between them there is a link with a probability of 0.5.

Chapter 5

EmotWion

This chapter is organized as follows.

Section 5.1 introduces and describe *EmotWion* pipeline with its components.

Section 5.2 provides the technical details of the topic and gender detection system inside *EmotWion*.

Section 5.3 describes the tweets elaboration process, all tweets are processed through 5 distinct processes, tokenization, noise elimination, POS tagging, semanticanalysis, frequency analysis.

Section 5.4 provides the technical details of the algorithm developed for the classification of emotions from tweets.

Section 5.5 reports a list of different case studies to analyze through *EmotWion*: (i) the average quantity of emotionally contagion users; (ii) the average duration of the contagion for each user.

For this purpose, it has been selected a series of 10 sources of tweets with politics topic and 10 sources of tweets with COVID-19 pandemic topic and for

each of these we have analyzed the tweets of users who have retweeted the source tweet in a time window of 12 days, the 6 days before and 6 days after the retweet date.

The experiments yielded, in the 72 hours following the retweet of the source tweet, an average percentage of contagion from 73.9% to 83.6% for tweets classified with high arousal values.

5.1 The *EmotWion* model

Emotion classification is a widely debated topic in psychology [13]. There are two main models about emotions: the first posits a discrete and finite set of emotions, while the second suggests that emotions are not independent and that exists a relation between them, hence the need to place them in a spatial space. Research in *NLP* has been focused mostly on Ekman's model of emotion [41] which posits the existence of six basic emotions: anger, disgust, fear, joy, sadness and surprise.

In this work, we focus on the most popular dimensional model of emotion: the circumplex model introduced in "A circumplex model of affect" [112].

This model suggests that all affective states are represented in a two-dimensional space with two independent neurophysiological systems: one related to valence (a pleasure–displeasure continuum) and the other to arousal, or alertness.

Every affective experience is a linear combination of these two independent systems, which is then interpreted as representing a particular emotion. For example, fear is a state involving the combination of negative valence and high arousal [107].

The approach to detect emotions used is a hybrid type as seen in paragraph 2.2.3. A rule-based approach based on the *NRC VAD lexicon* is combined with machine learning techniques to weigh and link the extracted terms to the context of the sentences.

EmotWion aims to analyze the contagion of emotions on *Twitter* through a multi-stage pipeline.

- Initially, from a defined discussion topic, tweets (called source tweets) published by users with a high number of followers (social influencers) are classified according to their emotion.

- Then all tweets of users who retweeted the relative source tweet, are extracted, retweets represent a very reliable index of approval of tweet [87], therefore a source of contagion with very high probability. For each user who has retweeted the source tweet, all tweets published are examined in a time window of 12 days, 6 days before and 6 days after, the date of publication of the source tweet.

- After, all extracted tweets are classified according to their emotion, they are, also, anonymized and tagged according to gender.

- Once the tweets of all users have been classified, the contagion of emotions can be estimated by comparing the emotional levels of the source tweet with the relative emotional levels of the retwitters' tweets.

The system pipeline according with the above procedures is composed by four processing systems Fig.5.1:

1. a *Twitter* content extraction system that guarantees a constant data flow;

2. a system for processing extracted tweets;

3. a system for estimating emotional metrics and for classifying the extracted tweets by calculating the overall emotion;

4. a system for processing, comparing and displaying the emotional level of the extracted tweets by comparing it with the relative source tweet.

The programming language used for tweets extraction and analysis is *Python* with support of some code libraries. *Python Twitter* to interact with the *Twitter* API, *NLTK* [81] for text analysis, *Matplotlib* to plot data in graphs. To store collected data was used a MySQL database.

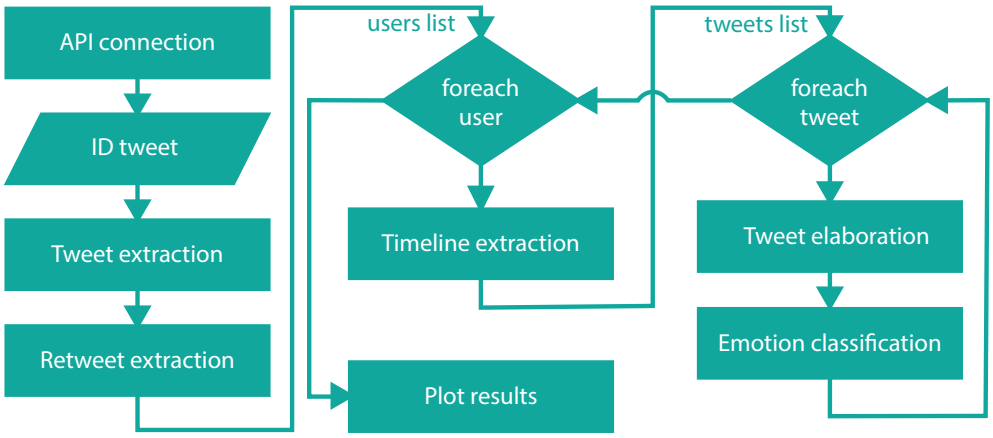


Figure 5.1: *EmotWion pipeline architecture.* Through an API request, the details of the source tweet are obtained (information on the tweet, retweet, etc.) All users who have retweeted are iterated first, then the list of tweets published by them. The tweets are finally classified.

5.2 Topic and gender detection

In order to determine discussion's topics, a classification algorithm, showed in Figure 5.2, based on *Word2vec* model and a synonyms dictionary was implemented. *Word2vec* is a simple two-layer artificial neural network designed to *NLP*.

The algorithm requires a text corpus as input and returns a set of vectors

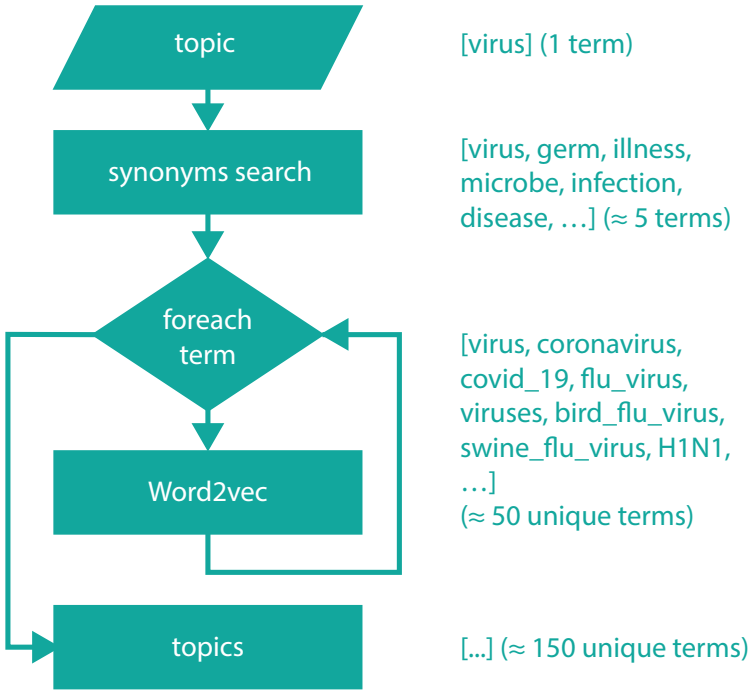


Figure 5.2: *Architecture of topic detection procedure.* Starting from a topic word, other words are extracted through a dictionary of synonyms. These words are given as input to the *Word2vec* algorithm which will return top semantically close words.

that represent the semantic distribution of words in the text. For each word contained in the corpus, a vector is uniquely constructed in order to represent it as a point in the multidimensional space created. In this space words will be closer if recognized as semantically more similar.

Once the corpus is built, the main topic of discussion is inserted as input of the algorithm, from which the synonyms are extracted by dictionary, subsequently for each synonym found the semantically closest terms through *Word2vec* model.

User's gender detection based on direct comparison between user's name and the list of names contained in the library nltk.corpus.names, which contains 7944 names labeled by gender.

5.3 Tweets elaboration

All tweets collected, both the source tweet to be analyzed and the lists of tweets extracted by the users who retweeted the source tweet, are processed through 5 distinct processes: tokenization, noise elimination, POS tagging, semantic analysis and frequency analysis as shown in Figure 5.3.

If tweet is made up of several sentences, it will be divided, each sentence will be processed according to the elaboration process described below.

5.3.1 Tokenization

Initially the tweet is subjected to a tokenization process that constitutes a preliminary step to any computational processing of the text [127].

Tokenize a text means to divide the sequences of characters into minimum

units of analysis which are called "tokens". Tokens can be simple entities such as words, punctuation, numbers or can be structurally complex entities such as dates that will be assumed as basic unit for subsequent processing levels. Depending on the type of language and writing system, tokenization is an extremely complex task, in this case the nltk.tokenize package belonging to the *NLTK* library for text segmentation was used.

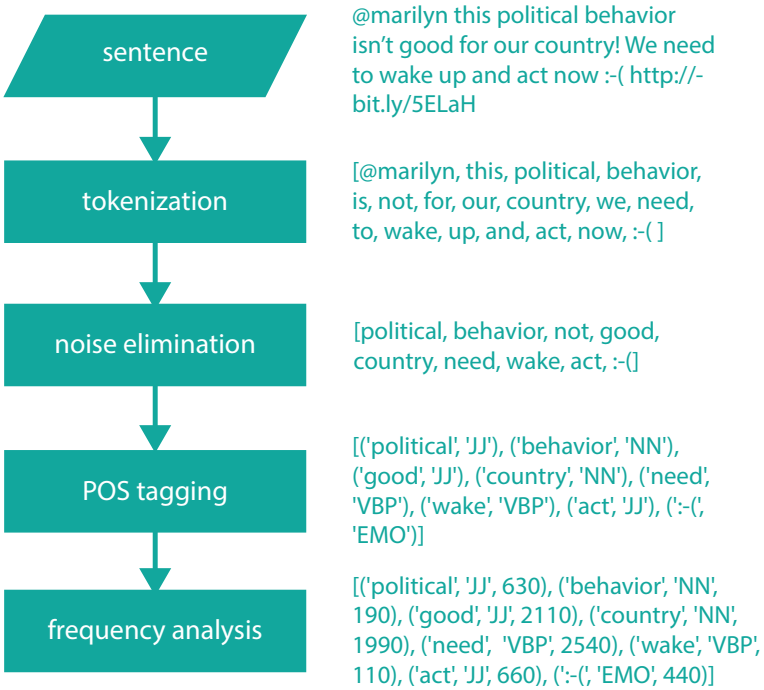


Figure 5.3: *Architecture of tweets elaboration procedure.* Starting from a sentence, the text is tokenized, so obtaining an array of terms, that are analyzed by the noise eliminator, after this terms are labeled using the POS tagger, and finally a frequency analysis of the terms is performed.

5.3.2 Noise elimination

Once the tokens are obtained, it is necessary to eliminate the noise present in the sequence to avoid producing misleading results in the analysis phase. In the case of irrelevant, redundant information or noisy data, the performance of the system could be considerably reduced [8].

Tweets contain usernames that start with the @ symbol, all words that start with the @ symbol are discarded through regular expressions. Many tweets contain URLs, which are discarded through regular expressions. Some tweets can contain numerical values that will be discarded because the goal is to analyze textual data. The tweets may contain special characters such as *, /,.,> which are removed. From each token, through the *strip* method, white spaces are removed at the beginning and at the end. Hashtags are not removed because they may contain important information about the tweet topic, only the # symbol is removed.

5.3.3 POS tagging

The noise elimination procedure from the text is followed by the POS (Part of Speech) tagging phase, also called grammatical tagging. It is a process that associates words with the corresponding POS based on its definition, its context, and its relation with adjacent and related words in a sentence or paragraph. Using the features of the nltk.tag package, in the *NLTK* library, the POS are detected and individual headwords are tagged with constants (ADJ – adjective, ADV – adverb, NOUN – noun, VERB – verb, etc.) [84].

In the following phases, only significant parts of the period for emotion classification will be considered.

5.3.4 Frequency analysis

Finally, output object from POS tagging is used as input for the terms frequency analysis function, through an *NLTK* library function, so we obtain the frequency distribution of terms for all tweets from the same user timeline.

5.4 Emotion classification

The terms extracted from each tweet are classified by direct comparison with a dictionary of words classified according to Russell's circumflex model of affect. According to this model, each emotion can be summarized in two dimensions, varying in value and intensity of activation [107].

For the classification was used the *NRC VAD Lexicon* vocabulary [89], consisting of 20,007 words classified by valence (emotion value) and arousal (activation intensity), original variables have values $\in [0, 1]$, they were normalized to values $\in [-1, 1]$. The set of words is composed of 25.1% adjectives [19], 57.5% nouns, 14.5% verbs and 2.9% elsewhere in the speech.

The output of the tweet processing phase previously seen in Figure 5.3, is an array of terms (array of objects), which is characterized by a POS tag and a frequency value within the text.

Through classification phase we associate to each term object, which is in the vocabulary, a value pair $(valence, arousal)$. The result of these classification rules will be a new array of term objects enriched with these two new values; therefore we get a result: $\{term, POS, frequency, (valence, arousal)\}$ as shown in Figure 5.4.

After valence and arousal, for each term of a sentence in a tweet is weighted

with term frequency and inverse document frequency $tf.idf$ parameters as described in the Eq. 5.1 and Eq. 5.2.

More specifically, tf.idf is a measure of how much information the term provides, namely, if it is common or rare across all input tweets.

In formula, we compute $tf.idf(i,u,T) = tf(i,u) * idf(i,T)$, where, term frequency $tf(i,u)$ is the frequency of term i, is defined as $tf(i,u) = f_{i,u}/\sum_{i' \in u} f_{i',u}$, with $f_{i,u}$ representing the number of times that term i occurs in a tweet of a specific user timeline u.

The inverse document frequency $idf(i,T)$ is a measure of how much infor-

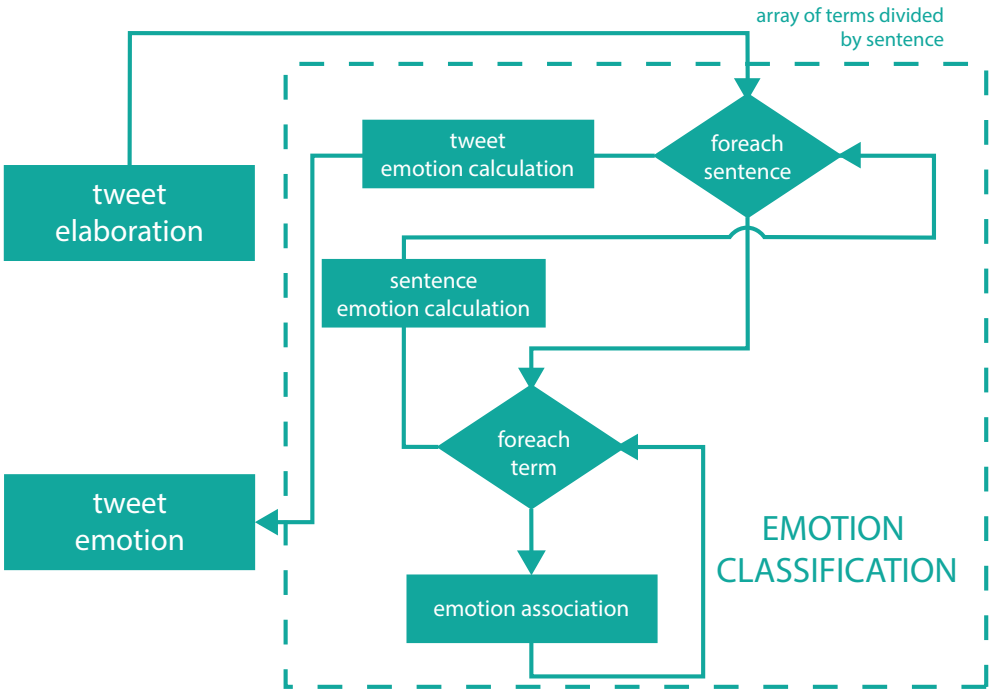


Figure 5.4: *Pipeline of emotion classification section.* As input to the pipeline goes an array of terms for each sentence of the tweet, the emotion is first calculated at sentence level and then at tweet level.

mation the term i provides. It is defined as $idf(i,T) = log^{N}/_{|\{u \in T : i \in u\}|}$, where N is the number of user timelines analyzed starting from a tweet source such that $N = |T|$, and $|\{u \in T : i \in u\}|$ is the number of tweets where the term i appears.

$$\bar{v}_{sentence} = \frac{\sum_{i=0}^{n} v_i * tfidf_i}{\sum_{i=0}^{n} tfidf_i} \qquad (5.1) \qquad \bar{a}_{sentence} = \frac{\sum_{i=0}^{n} a_i * tfidf_i}{\sum_{i=0}^{n} tfidf_i} \qquad (5.2)$$

After the valence and arousal calculation for each sentence of the tweet, the overall emotion of the tweet is calculated through an average of the values obtained for each sentence, as reported in the equation 5.3 and 5.4, where k is the number of sentences in the tweet.

$$\bar{v}_{tweet} = \frac{\sum_{i=0}^{k} v_{i-sentence}}{k} \qquad (5.3) \qquad \bar{a}_{tweet} = \frac{\sum_{i=0}^{k} a_{i-sentence}}{k} \qquad (5.4)$$

For a more systematic comparison of the proposed classification system we use *Emobank* [20, 21] as ground-truth.

Emobank is a large-scale text corpus manually annotated with emotion according to the psychological valence - arousal scheme. It was built at *JULIE Lab, Jena University* and containing 10,000 sentences. Each *Emobank* sentence is classified with *EmotWion*, an extract of classification results is shown in the table. Valence and arousal values in *Emobank* are included in the interval $[1,5]$, to be compared with *EmotWion* they have been normalized in the interval $[-1,1]$.

Δt	Precision %	Δt	Precision %	Δt	Precision %
0.20	94.29	0.15	86.28	0.10	71.44

Table 5.1: *Precision value obtained from the comparison with Emobank for different* Δt *range.*

The average value of precision obtained from the comparison with *Emobank* is reported in Table 5.1; Δt is considered a parameter that indicates a tolerance deviation of the valence and arousal values to evaluate true positives (TP) and false positive (FP) cases. Fixed the value of Δt, for each *Emobank* corpus sentence, if $(valence, arousal)$ value pair of the sentence has an euclidean distance less than or equal to the $(valence, arousal)$ value pair estimated with *EmotWion*, so it is a true positive (TP), otherwise it is a false positive (FP). Precision value in Table 5.1 is obtained through the formula: $precision = {}^{TP}/_{TP+FP}$

The average error found in the classification of the 10,000 sentences, evaluated as the average Euclidean distance between *Emobank*'s $(valence, arousal)$ value pairs and those classified with *EmotWion*, is equal to 0.13.

Table 5.2 shows an extract of data from the comparison between *EmotWion* and *Emobank*. This table shows an excerpt of 20 texts out of the 10,000 of *Emobank*, column ID is a unique value of the data record, column *Text* contains the sentences of *Emobank*, columns V_{eb} and A_{eb} contain respectively *Emobank*'s valence and arousal values for the corresponding sentence, columns VN_{eb} and AN_{eb} contain V_{eb} and A_{eb} normalized values in the interval $[-1, 1]$, columns V_{ew} and A_{ew} contain valence and arousal values estimated by *EmotWion* for the corresponding sentence, column error contains the classification error,

i.e. the euclidean distance between (VN_{eb}, AN_{eb}) and (V_{ew}, A_{ew}) points.

ID	Text	V_{eb}	A_{eb}	VN_{eb}	AN_{eb}	V_{ew}	A_{ew}	error
110CYL068_702_771	Goodwill finds jobs for people with mental and physical disabilities.	3.29	3.00	0.15	0.00	0.17	-0.05	0.05
112C-L013_1054_1129	Then the task was to help children who lost parents in the Civil War.	3.00	3.00	0.00	0.00	0.01	-0.07	0.07
118CWL048_1437_1492	Early adolescence is the most vulnerable age for youth.	3.00	2.90	0.00	-0.05	0.05	0.00	0.07
SemEval_1006	North Africa feared as staging ground for terror	2.10	3.50	-0.45	0.25	-0.49	0.20	0.06
SemEval_1027	Archaeologists find signs of early chimps' tool use	3.00	2.80	0.00	-0.10	0.08	-0.08	0.08
SemEval_1046	Damaged Japanese whaling ship may resume hunting off Antarctica	2.86	3.00	-0.07	0.00	-0.12	0.06	0.07
SemEval_1062	Taliban seize rural district in southwest as police flee	3.00	3.00	0.00	0.00	-0.07	0.02	0.07
SemEval_1066	Iran vs. North Korea: not all enemies are equal	2.88	2.75	-0.06	-0.13	-0.03	-0.16	0.05
SemEval_1094	Russian bird flu outbreak is deadly 'Asian strain'	2.00	3.60	-0.50	0.30	-0.42	0.29	0.08
SemEval_1153	'Jackass' star marries childhood friend The secrets people reveal	3.50	3.37	0.25	0.19	0.27	0.13	0.06
SemEval_159	New Indonesia Calamity, a Mud Bath, Is Man-Made	2.90	2.80	-0.05	-0.10	0.01	-0.11	0.06
SemEval_504	Closings and cancellations top advice on flu outbreak	2.60	2.90	-0.20	-0.05	-0.17	-0.03	0.04
SemEval_52	Britain to restrict immigrants from new EU members	2.70	3.20	-0.15	0.10	-0.17	0.06	0.04
SemEval_53	Confusion Reigns In the Expanding Digital World	2.90	3.00	-0.05	0.00	0.04	-0.00	0.09
SemEval_614	Terai protesters hail 'historic' victory over government	3.50	3.30	0.25	0.15	0.33	0.19	0.09
SemEval_643	Bush authorized Iranians' arrest in Iraq, Rice says	3.00	2.90	0.00	-0.05	0.05	-0.03	0.05
SemEval_65	Doctors Seeing Patients Who Think They're Internet Addicted	3.00	2.90	0.00	-0.05	0.08	-0.09	0.09
SemEval_664	Super Bowl ads of cartoonish violence, perhaps reflecting toll of war	2.70	3.30	-0.15	0.15	-0.10	0.16	0.05
SemEval_797	Bill would strip convicted legislators of pensions	2.78	3.22	-0.11	0.11	-0.13	0.08	0.04

Table 5.2: *Example of comparison test between EmotWion and Emobank.* A set of 20 texts extracted from *Emobank* comparison, column *ID* is a unique value of the data record, column *Text* contains the sentences of *Emobank*, columns V_{eb} and A_{eb} contain respectively *Emobank*'s valence and arousal values for the corresponding sentence, columns VN_{eb} and AN_{eb} contains the values V_{eb} and A_{eb} normalized in the value interval $[-1, 1]$, columns V_{ew} and A_{ew} contain valence and arousal values estimated by *EmotWion* for the corresponding sentence, column error contains the classification error, i.e. the euclidean distance between (VN_{eb}, AN_{eb}) and (V_{ew}, A_{ew}) points.

5.5 Case studies

In the proposed case studies, we analyze through *EmotWion*: (i) the average quantity of emotionally contagion users; (ii) the average duration of the contagion for each user.

For this purpose, we have selected a series of 10 sources of tweets with politics topic and 10 sources of tweets with COVID-19 pandemic topic, and for each of these we have analyzed the tweets of users who have retweeted the source tweet in a time window of 12 days, the 6 days before and 6 days after the retweet date.

Each tweet is processed and classified by *EmotWion* as described in sections 5.3 and 5.4. After classifying all tweets for each user, results are represented in a graph, in which are annotated heatmaps, through a *Python* library for graphs *Matplotlib* [66]. Heatmap graph requires that data pairs are discrete categories, then $(valence, arousal)$ value pairs were discretized to 0.2 interval, obtaining 100 squares with different $(valence, arousal)$ values.

On each square in the heatmap, tweet count value is written. Color map from blue (low number of users) to yellow (high number of users) allows to visually evaluate the variations within the space.

In figures 5.5 and 5.6 an example of analysis of source tweets is shown.

In Figure 5.5 data of the source tweet (text of the tweet, date of publication, etc.) is shown on top, while in the graph the red dot represents the $(valence, arousal)$ value pair of the source tweet in the circumplex space model, blue dots are some sample reference $(valence, arousal)$ value pairs.

In Figure 5.6 are shown the heatmaps in the 12 days of analysis, 6 days

before (first row) and 6 days after (second row) the retweets of each single user who retweeted.

In the example the source tweet has an emotional level with a medium-high valence and medium arousal values, starting from the same day, that users retweeted the source tweet, and for the following 3 days there is an increase of the number of users around the $(valence, arousal)$ values of the original tweet (colors with yellow shades). This indicates an emotional contagion from the original tweet to the tweets of users who retweeted it, on the same topic.

Tweet	Valence	Arousal	PC(-6)	PC(-5)	PC(-4)	PC(-3)	PC(-2)	PC(-1)	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)	PC(6)	AC
001	0,38	0,01	31,7	37,3	33,1	36	32,8	37,8	50,7	55,2	56	51,7	46	36,3	54
002	0,85	0,42	29,6	32,4	37,6	34,2	35	29,2	71,2	84	81,4	78,5	74,7	41,5	78,9
003	0,42	0,39	33,2	29,5	30,8	36,4	38,9	33,7	73	76,4	72,2	75	64,3	31,2	73,9
004	0,78	-0,34	32,2	35,4	35,8	37,5	40,6	34,6	66	57,9	61,9	51,4	52,7	36,7	61,9
005	0,64	-0,35	31,7	33,8	35,4	37,7	40,5	36,2	63,1	61,4	68,5	66	54,9	41,5	64,3
006	0,74	0,39	39	34,2	37	38,9	41	33,8	82,7	81,6	77,5	76,1	51,6	35,4	80,6
007	0,39	-0,31	27,9	33,1	29,4	42	34	37,2	73,6	62	69,3	66,5	44,1	36,9	68,3
008	0,45	-0,18	38,9	37,7	41,3	41,2	43,8	38,4	66,1	74,3	69,4	69,2	54,5	38,6	69,9
009	0,69	0,31	37,3	39	39,9	39	40,9	36,1	77,1	79,3	80,7	69,3	44,4	33,6	79,1
010	0,24	-0,17	31	38,6	29,5	35,4	44,6	37,9	56,6	54,2	51,8	59,3	47,3	31	54,2

Table 5.3: *Daily trend of the percentage of users with the same emotional values pairs $(valence, arousal)$ related to their tweet source, with a global COVID-19 pandemic topic.*

Tables 5.3 and 5.4 show the results of the experiments carried out on 20 source tweets (10 on world pandemic from COVID-19 and 10 on a politics topic).

The columns of the table show: (i) progressive number of the source tweet, (ii) valence and arousal of the source tweet, (iii) percentage of users with the same value of valence and arousal at different days (PC), from 6 days prior to the retweet to 6 days after the retweet, with a tolerance of 0.2 units of valence

Tweet	Valence	Arousal	PC(-6)	PC(-5)	PC(-4)	PC(-3)	PC(-2)	PC(-1)	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)	PC(6)	AC
001	-0,71	0,36	41	37,2	34	39,9	44	35,8	83,7	85,6	81,5	75,1	55,6	41	83,6
002	-0,81	-0,12	38,2	31,4	38,8	37,5	42,6	32,6	60	66,9	60,9	55,4	52,7	32,9	62,6
003	-0,54	0,51	35,2	29,5	31,8	35,4	39,9	29,7	71	80,4	76,2	79	65,3	42,2	75,9
004	-0,8	0,14	43,3	38	42,9	40	45,9	36,1	81,1	83,3	79,7	69,3	46,4	44,1	81,4
005	0,48	0,13	36,7	38,3	30,1	34	33,8	37,8	48,7	55,2	57	52,7	49	37,6	53,6
006	0,34	0,06	39,9	38,7	40,3	40,2	41,8	40,4	64,1	77,3	72,4	67,2	54,5	32,5	71,3
007	-0,74	0,1	29	36,6	33,5	31,4	49,6	36,9	57,6	52,2	51,8	60,3	48,3	39	53,9
008	-0,68	-0,18	36,7	36,8	41,4	35,7	43,5	33,2	61,1	63,4	70,5	69	55,9	41,3	65
009	-0,67	-0,05	33,9	36,1	28,4	39	31	41,2	54,6	63	67,3	69,5	46,1	33,2	61,6
010	-0,61	0,78	27,6	34,4	34,6	37,2	39	34,2	75,2	88	79,4	76,5	75,7	38,4	80,9

Table 5.4: *Daily trend of the percentage of users with the same emotional values pairs* $(valence, arousal)$ *related to their tweet source, with politics topic.*

or arousal (iv) average (AC) of the percentages of users (PC) in the 72 hours after the retweet.

Figure 5.7 shows the graphs of the data reported in Table 5.3 - 5.4, red lines represent the trend of (PC) for high arousal values ($0.35 < a < 1$).

The experiments yielded, in the 72 hours following the retweet of the source tweet, an average percentage of contagion (AC) from 73.9% to 83.6% for tweets classified with high arousal values ($0.35 < a < 1$) (highlighted in red).

Results show that emotions like anger (low valence - high arousal) spreads fastest through users and lasts longer over time. Same results with emotion like joy (high valence - high arousal).

On the other hand, emotion like sadness, boredom or disgust, with medium or low $(valence, arousal)$ values, are even more private emotions, for which there is no emotional contagion through users.

topic	**World pandemic**
source tweet	The Media should view this as a time of unity and strength. We have a common enemy, actually, an enemy of the World, the CoronaVirus. We must beat it as quickly and safely as possible. There is nothing more important to me than the life & safety of the United States!
date	**11/03/2020**
valence	**0.38**
arousal	**0.01**
nearest emotion word	**determinated**

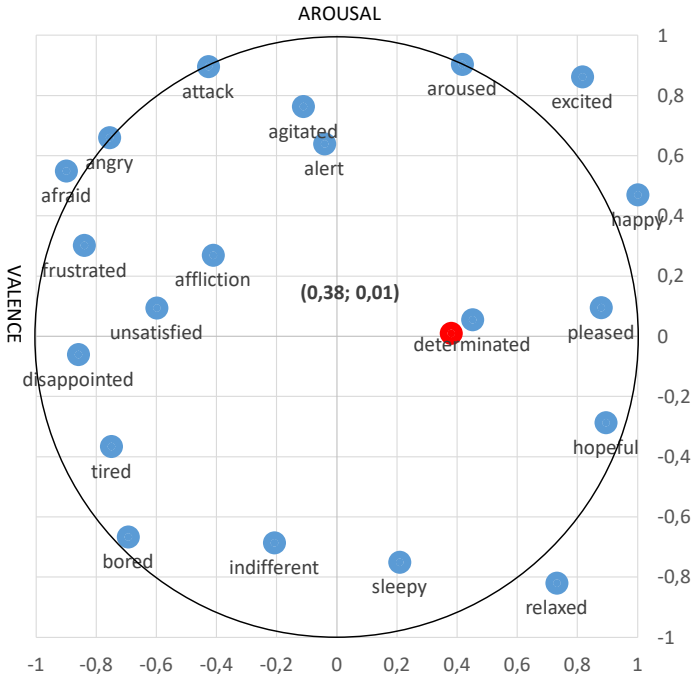


Figure 5.5: *Analysis of a tweet source with world pandemic topic.* Table shows information of the source tweet (tweet text, publication date, valence and arousal values, etc.). In the circumplex model graph, blue dots are some sample reference $(valence, arousal)$ value pairs, red dot is the $(valence, arousal)$ value pair of the source tweet.

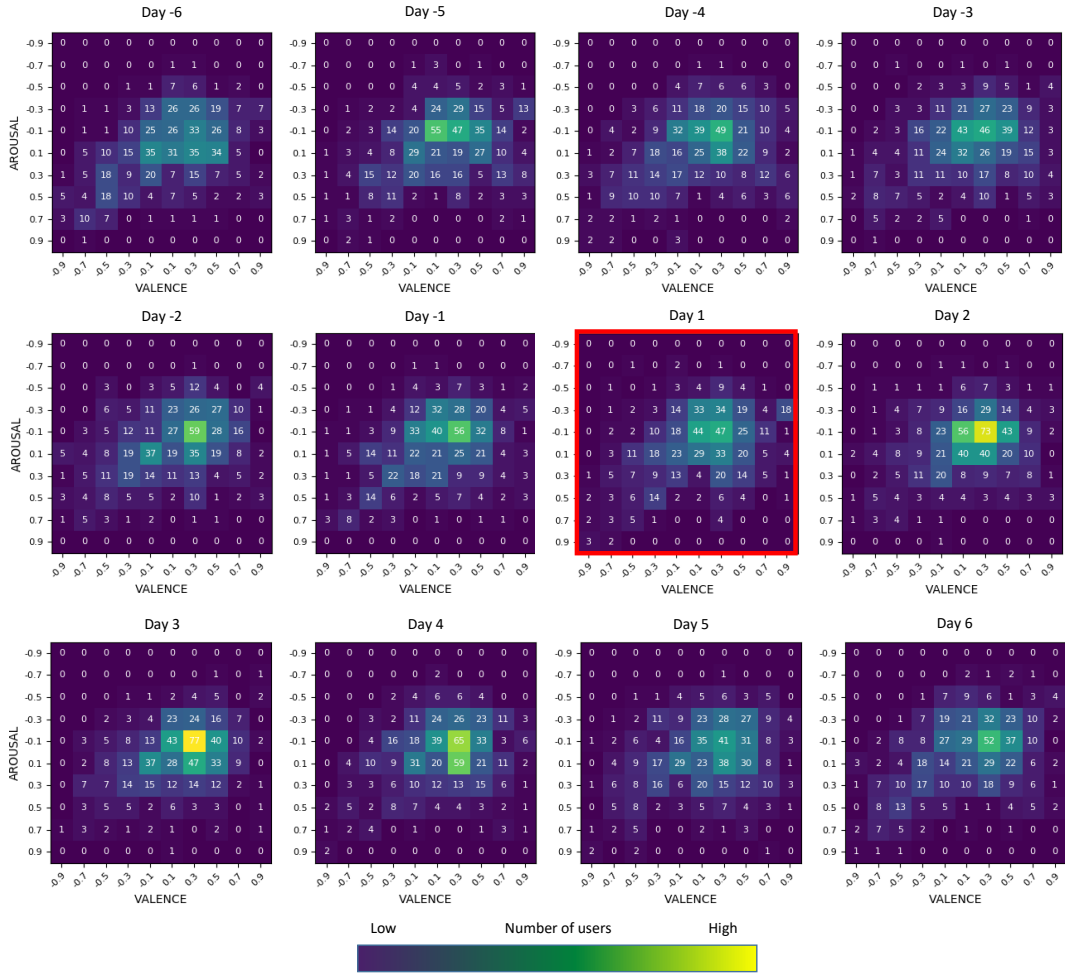
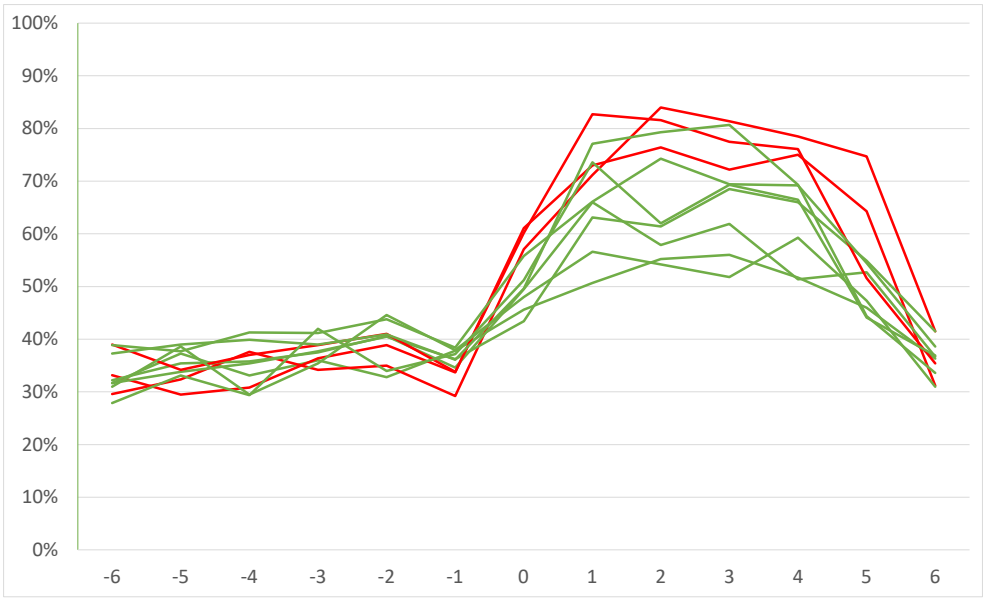
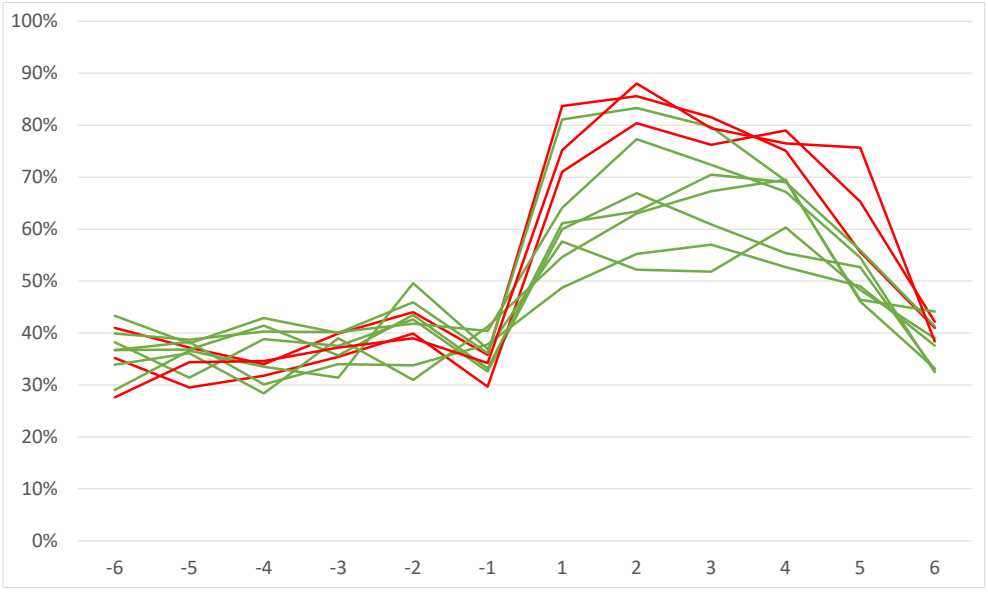


Figure 5.6: *Heatmap chart for source tweet shown in figures 5.5.* A heatmap is plotted for each of 12 time window observation days (6 days before the retweet - 6 days after the retweet). Retweet day is outlined in red. Each heatmap is divided into 100 squares with steps of 0.2 (valence - arousal), each square contain number of users with the same $(valence, arousal)$ value pair.



(a)



(b)

Figure 5.7: *Daily trends of the percentage of users with the same emotional values pairs (valence, arousal) related to their tweet source, with global COVID-19 pandemic topic in Figure 5.7a and with politics topic in Figure 5.7b.* Graphs show data reported in Table 5.3 - 5.4 respectively. The days are shown on the abscissa and the percentages of contagion on the ordinate.

Conclusions

In this thesis I introduce *NetME* and *EmotWion*, two frameworks that use *NLP* technologies for text mining.

NetME is a framework, which I developed, that infer on-the-fly knowledge-graphs from a collection of full-text papers obtained from *PubMed* or user-provided, to get knowledge about relations and interactions between elements in the biomedical field, which lead to emerging phenomena studied in the complex systems domain.

It has been implemented upon a customized version of *TAGME*, called *OntoTAGME*, in connection to a syntactic analysis module developed on top of the *Python NLTK* and *spaCy* libraries. Our results clearly show that *NetME* allows extracting reliable knowledge graphs in a few minutes or a few hours compared to a manual search that could take several days or weeks. The completeness of the extracted knowledge increases when the documents used by *NetME* comprehensively describe the desired topic under study.

To evaluate *NetME* we performed three case studies.

The first one compared *NetME* with the web application *Hetionet*, using

SemRep software as ground truth. The experiment showed that *NetME* has a comparable precision with *Hetionet* in terms of inferred nodes and an higher recall. Concerning inferred edges, *Hetionet* precision is slightly higher although *NetME* recall is much higher.

The second case study tested the ability of *NetME* in recovering relationships between genes. The experiment yielded accuracy ranging from 58%, when using low reliable relations (i.e. False Positives) from *STRING*, to 84%, when such *STRING* relations are very reliable.

At the same time, the third case study tested the ability of *NetME* in integrating knowledge about genes starting from a selected set of scientific papers. The experiment yielded 98% sensitivity and 100% specificity. Therefore, both experiments clearly showed the high reliability of *NetME*'s inferred networks.

Future work will include: (i) the construction of knowledge-graphs from all the open-access papers stored in *PMC*; (ii) the integration of all *Obofoundry* ontology within *OntoTAGME*; (iii) the design of a more effective algorithm to select the pertinent scientific papers on which *NetME* has to be applied [105, 104]; and finally, add a methodology that allows to extract context-based relationships.

EmotWion is a framework, which I developed, that aims to analyze the contagion of emotions on complex network like *Twitter*, from a tweet (called source tweet) published by an user with an high number of followers (social influencer), to users who retweet source tweet.

It has been implemented upon (i) a *Twitter* content extraction system, (ii) a system for processing extracted tweets, (iii) a system for classifying the extracted tweets by calculating the overall emotion and (iv) a system for pro-

cessing, comparing and displaying the emotional level of the extracted tweets.

It has been developed in *Python* programming language with support of *Python Twitter* and *NLTK* for text analysis.

To evaluate *EmotWion* classification system we use *Emobank*, a large-scale text corpus manually annotated, as ground-truth, obtaining a precision from 94.29% with a 0.20 units of tolerance to 71.44% with a 0.10 units of tolerance.

In the proposed case studies, we analyze through *EmotWion*: (i) the average quantity of emotionally contagion users; (ii) the average duration of the contagion for each user.

The experiments yielded, in the 72 hours following the retweets of the source tweets, an average percentage of contagion from 73.9% to 83.6% for tweets classified with high arousal values.

Results show that emotions like anger spreads fastest through users and lasts longer over time. Same results with emotion like joy. On the other hand, emotion like sadness, boredom or disgust, with medium or low $(valence, arousal)$ values, are private emotions for which there is no emotional contagion through users.

Future work will include: (i) improve the algorithm for the syntactic analysis of the texts; (ii) testing additional dictionaries for emotions classification; (iii) creation of an usable interface for on-the-fly classification of texts.

Bibliography

[1] *bioRxiv*. Available at `https://www.biorxiv.org/`.

[2] *Language Processing Pipeline*. Available at `https://spacy.io/usage/processingpipelines`.

[3] *What's New in v3.0*. Available at `https://spacy.io/usage/v3#summary`.

[4] Introduction to graph theory. In *Applied Combinatorics*, pages 147–262. Chapman and Hall/CRC, jun 2009.

[5] Introduction to graphs. In *Graphs & Digraphs*, pages 13–66. Chapman and Hall/CRC, oct 2010.

[6] S. Alaimo, R. V. Rapicavoli, G. P. Marceca, A. La Ferlita, O. B. Serebrennikova, P. N. Tsichlis, B. Mishra, A. Pulvirenti, and A. Ferro. Phensim: Phenotype simulator. *bioRxiv*, 2020.

[7] F. M. Alotaibi. Classifying text-based emotions using logistic regression. *VAWKUM Transactions on Computer Sciences*, pages 31–37, apr 2019.

[8] G. Angiani, L. Ferrari, T. Fontanini, P. Fornacciari, E. Iotti, F. Magliani, and S. Manicardi. A comparison between preprocessing techniques for sentiment analysis in twitter. In *KDWeb*, 2016.

[9] A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, may 2010.

[10] A. T. Azar and S. M. El-Metwally. Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 23(7-8):2387–2403, oct 2012.

[11] A. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, Dec. 2010.

[12] J. B. L. Bard. The AEO, an ontology of anatomical entities for classifying animal tissues and organs. *Frontiers in Genetics*, 3, 2012.

[13] L. F. Barrett, M. Gendron, and Y.-M. Huang. Do discrete emotions exist? *Philosophical Psychology*, 22:427 – 437, 2009.

[14] J. Beck. Report from the field: PubMed central, an XML-based archive of life sciences journal articles. In *Proceedings of the International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML*. Mulberry Technologies, Inc.

[15] E. Birney. An overview of ensembl. *Genome Research*, 14(5):925–928, May 2004.

[16] P. Blohm, G. Frishman, P. Smialowski, F. Goebels, B. Wachinger, A. Ruepp, and D. Frishman. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Research*, 42(D1):D396–D400, Nov. 2013.

[17] M. M. Bradley and P. J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology . . . , 1999.

[18] J. J. Brown and P. H. Reingen. Social ties and word-of-mouth referral behavior. *Journal of Consumer research*, 14(3):350–362, 1987.

[19] M. Brysbaert, B. New, and E. Keuleers. Adding part-of-speech information to the subtlex-us word frequencies. *Behavior research methods*, 44(4):991–997, 2012.

[20] S. Buechel and U. Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 2017.

[21] S. Buechel and U. Hahn. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*. Association for Computational Linguistics, 2017.

[22] L. Canales and P. Martínez-Barco. Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*. Association for Computational Linguistics, 2014.

[23] E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman. Automated acquisition of disease-drug knowledge from biomedical and clinical documents: An initial study. *Journal of the American Medical Informatics Association*, 15(1):87–98, jan 2008.

[24] M.-Y. Chen and T.-H. Chen. Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena. *Future Generation Computer Systems*, 96:692–699, jul 2019.

[25] Q. Chen and G. chun Ge. A corpus-based lexical study on frequency and distribution of coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*, 26(4):502–514, jan 2007.

[26] R. Chopade. Text based emotion recognition: A survey. 2015.

[27] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

[28] A. M. Cohen. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, Jan. 2005.

[29] G. O. Consortium. The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(90001):258D–261, Jan. 2004.

[30] M. D. C. S. Counts and M. Gamon. Not all moods are created equal! exploring human emotional states in social media. In *Proc. Int. AAAI Conf. Web Social Media (ICWSM)*, pages 1–8, 2012.

[31] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477, Nov. 2013.

[32] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Database):D691–D697, Nov. 2010.

[33] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical part-of-speech tagger. In *Proceedings of the third conference on Applied natural language processing* -. Association for Computational Linguistics, 1992.

[34] A. D. Diehl, T. F. Meehan, Y. M. Bradford, M. H. Brush, W. M. Dahdul, D. S. Dougall, Y. He, D. Osumi-Sutherland, A. Ruttenberg, S. Sarntivijai, C. E. V. Slyke, N. A. Vasilevsky, M. A. Haendel, J. A. Blake, and C. J. Mungall. The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, 7(1), July 2016.

[35] P. Ding, X. Zhang, S. Jin, B. Duan, P. Chu, Y. Zhang, Z. Chen, B. Xia, and F. Song. Cd147 functions as the signaling receptor for extracellular divalent copper in hepatocellular carcinoma cells. *Oncotarget*, 8(31):51151–51163, May 2017.

[36] P. Drieger. Semantic network analysis as a method for visual text analytics. *Procedia - Social and Behavioral Sciences*, 79:4–17, jun 2013.

[37] J. J. Dunkin. Emotion, mind, and brain: The neuropsychology of emotion. j.c. borod (ed.). 2000. new york: Oxford university press. 511 pp., $69.50. *Journal of the International Neuropsychological Society*, 8(5):727–728, jul 2002.

[38] J. Dörpinghaus, A. Apke, V. Lage-Rupprecht, and A. Stefan. Data exploration and validation on dense knowledge graphs for biomedical research, 2019.

[39] S. E. *Entrez Programming Utilities Help*. Available at `https://www.ncbi.nlm.nih.gov/books/NBK25501/`.

[40] J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, D. Preoţiuc-Pietro, D. A. Asch, and H. A. Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, oct 2018.

[41] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6:169–200, 1992.

[42] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, 2006.

[43] A. Fabregat, K. Sidiropoulos, G. Viteri, O. Forner, P. Marin-Garcia, V. Arnau, P. D'Eustachio, L. Stein, and H. Hermjakob. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics*, 18(1), Mar. 2017.

[44] P. Ferragina and U. Scaiella. Tagme. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM 2010*. ACM Press, 2010.

[45] J. H. Fowler and N. A. Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *Bmj*, 337, 2008.

[46] M. Franz, C. T. Lopes, G. Huck, Y. Dong, O. Sumer, and G. D. Bader. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, page btv557, Sept. 2015.

[47] K. Fundel, R. Kuffner, and R. Zimmer. RelEx–relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, dec 2006.

[48] P. Ginsparg. *arXiv*. Available at `https://arxiv.org`.

[49] S. Golder, Y. K. Loke, and M. Bland. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: Methodological overview. *PLoS Medicine*, 8(5):e1001026, may 2011.

[50] G. D. Grass and B. P. Toole. How, with whom and when: an overview of cd147-mediated regulatory networks influencing matrix metalloproteinase activity. *Bioscience Reports*, 36(1), Jan. 2016.

[51] K. A. Gray, R. L. Seal, S. Tweedie, M. W. Wright, and E. A. Bruford. A review of the new HGNC gene family resource. *Human Genomics*, 10(1), Feb. 2016.

[52] B. F. Green, P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. The general inquirer: A computer approach to content analysis. *American Educational Research Journal*, 4(4):397, nov 1967.

[53] M. Gremse, A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling, and D. Schomburg. The BRENDA tissue ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Research*, 39(Database):D507–D513, Oct. 2010.

[54] M. Griffith, N. C. Spies, K. Krysiak, J. F. McMichael, A. C. Coffman, A. M. Danos, B. J. Ainscough, C. A. Ramirez, D. T. Rieke, L. Kujan, E. K. Barnell, A. H. Wagner, Z. L. Skidmore, A. Wollam, C. J. Liu, M. R. Jones, R. L. Bilski, R. Lesurf, Y. Feng, N. M. Shah, M. Bonakdar, L. Trani, M. Matlock, A. Ramu, K. M. Campbell, G. C. Spies, A. P. Graubert, K. Gangavarapu, J. M. Eldred, D. E. Larson, J. R. Walker, B. M. Good, C. Wu, A. I. Su, R. Dienstmann, A. A. Margolin, D. Tamborero, N. Lopez-Bigas, S. J. M. Jones, R. Bose, D. H. Spencer, L. D. Wartman, R. K. Wilson, E. R. Mardis, and O. L. Griffith. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, 49(2):170–174, Jan. 2017.

[55] N. Gupta, M. Gilbert, and G. D. Fabbrizio. Emotion detection in email customer care. *Computational Intelligence*, 29(3):489–505, 2013.

[56] M. Hasan, E. Rundensteiner, and E. Agu. Emotex: Detecting emotions in twitter messages. 2014.

[57] M. Hasan, E. Rundensteiner, and E. Agu. Automatic emotion detection in text streams by analyzing twitter data. *International Journal of Data Science and Analytics*, 7(1):35–51, feb 2018.

[58] D. S. Himmelstein and S. E. Baranzini. Heterogeneous network edge prediction: A data integration approach to prioritize disease-associated genes. *PLOS Computational Biology*, 11(7):e1004259, July 2015.

[59] D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, and S. E. Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6, Sept. 2017.

[60] R. Hirat and N. Mittal. A survey on emotion detection techniques using text in blogposts. 2015.

[61] M. Honnibal and M. Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[62] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo, 2020.

[63] C. T. Hoyt, D. Domingo-Fernández, and M. Hofmann-Apitius. BEL commons: an environment for exploration and analysis of networks encoded in biological expression language. mar 2018.

[64] J. Huang, Z. Lin, and X. Liu. Episodic memory network with self-attention for emotion detection. In *DASFAA*, 2019.

[65] B. L. Humphreys, D. A. B. Lindberg, H. M. Schoolman, and G. O. Barnett. The unified medical language system: An informatics research collaboration. *Journal of the American Medical Informatics Association*, 5(1):1–11, jan 1998.

[66] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[67] C. Huo, H. Wu, J. Xiao, D. Meng, S. Zou, M. Wang, P. Qi, H. Tian, and Y. Hu. Genomic and bioinformatic characterization of mouse mast cells (p815) upon different influenza a virus (h1n1, h5n1, and h7n2) infections. *Frontiers in Genetics*, 10, jun 2019.

[68] Z. Jiang, S. Hu, D. Hua, J. Ni, L. Xu, Y. Ge, Y. Zhou, Z. Cheng, and S. Wu. β3gnt8 plays an important role in CD147 signal transduction as an upstream modulator of MMP production in tumor cells. *Oncology Reports*, 32(3):1156–1162, June 2014.

[69] G. Joshi-Tope. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database issue):D428–D432, Dec. 2004.

[70] M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, Jan. 2000.

[71] M. Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11):1947–1951, Sept. 2019.

[72] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 01 2000.

[73] X. Ke, F. Fei, Y. Chen, L. Xu, Z. Zhang, Q. Huang, H. Zhang, H. Yang, Z. Chen, and J. Xing. Hypoxia upregulates cd147 through a combined effect of hif-1*alpha* and sp1 to promote glycolysis and tumor progression in epithelial solid tumors. *Carcinogenesis*, 33(8):1598–1607, June 2012.

[74] J. Kim, Y. Wang, T. Fujiwara, S. Okuda, T. J. Callahan, and K. B. Cohen. Open agile text mining for bioinformatics: the PubAnnotation ecosystem. *Bioinformatics*, 35(21):4372–4380, Apr. 2019.

[75] P. Kirk, M. Wilson, C. Heddle, M. Brown, A. Barclay, and A. Halestrap. CD147 is tightly associated with lactate transporters MCT1 and MCT4 and facilitates their cell surface expression. *The EMBO Journal*, 19(15):3896–3904, Aug. 2000.

[76] L.-M. Kong, C.-G. Liao, Y. Zhang, J. Xu, Y. Li, W. Huang, Y. Zhang, H. Bian, and Z.-N. Chen. A regulatory loop involving mir-22, sp1, and c-myc modulates cd147 expression in breast cancer invasion and metastasis. *Cancer Research*, 74(14):3764–3778, June 2014.

[77] M. Krallinger, R. A. Erhardt, and A. Valencia. Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today*, 10(6):439–445, Mar. 2005.

[78] P. Lambrix, H. Tan, V. Jakoniene, and L. Strömbäck. Biological ontologies. In *Semantic Web*, pages 85–99. Springer US.

[79] F. Li, M. Zhang, G. Fu, and D. Ji. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics*, 18(1), mar 2017.

[80] Q. Li, M. Jiang, X. Zhang, M. Qu, T. P. Hanratty, J. Gao, and J. Han. TruePIE. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, jul 2018.

[81] E. Loper and S. Bird. NLTK. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics* -. Association for Computational Linguistics, 2002.

[82] J. J. Louviere, T. N. Flynn, and A. A. J. Marley. *Best-Worst Scaling*. Cambridge University Press, 2015.

[83] X. Ma and L. Gao. Biological network analysis: insights into structure and functions. *Briefings in Functional Genomics*, 11(6):434–442, Nov. 2012.

[84] N. Madnani. Getting started on natural language processing with python. *XRDS: Crossroads, The ACM Magazine for Students*, 13(4):5–5, 2007.

[85] B. McBride. The resource description framework (RDF) and its vocabulary description language RDFS. In *Handbook on Ontologies*, pages 51–65. Springer Berlin Heidelberg, 2004.

[86] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabasi. Uncovering disease-disease relationships through

the incomplete interactome. *Science*, 347(6224):1257601–1257601, Feb. 2015.

[87] P. Metaxas. Retweets indicate agreement, endorsement, trust: A meta-analysis of published twitter research. *Arxiv. Retrieved from http://cs. wellesley. edu/˜ pmetaxas/WorkingPapers/Retweet-meaning. pdf*, 2017.

[88] S. Mohammad. # emotional tweets. In * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, 2012.

[89] S. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018.

[90] S. M. Mohammad and P. D. Turney. CROWDSOURCING a WORD-EMOTION ASSOCIATION LEXICON. *Computational Intelligence*, 29(3):436–465, sep 2012.

[91] A. Muscolino, A. Di Maria, S. Alaimo, S. Borzì, P. Ferragina, A. Ferro, and A. Pulvirenti. NETME: On-the-fly knowledge network construction from biomedical literature. In *Complex Networks & Their Applications IX*, pages 386–397. Springer International Publishing, 2021.

[92] A. M. Muscolino and S. Pagano. Sentiment analysis, a support vector machine model based on social network data. *International Journal of Research in Engineering and Technology*, 7, 2018.

[93] D. A. Natale, C. N. Arighi, J. A. Blake, J. Bona, C. Chen, S. Chen, K. R. Christie, J. Cowart, P. D'Eustachio, A. D. Diehl, H. J. Drabkin, W. D. Duncan, H. Huang, J. Ren, K. Ross, A. Ruttenberg, V. Shamovsky, B. Smith, Q. Wang, J. Zhang, A. El-Sayed, and C. H. Wu. Protein ontology (PRO): enhancing and scaling up the representation of protein entities. *Nucleic Acids Research*, 45(D1):D339–D346, Nov. 2016.

[94] D. Nettleton. Data representation. In *Commercial Data Mining*, pages 49–66. Elsevier, 2014.

[95] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[96] D. N. Nicholson and C. S. Greene. Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, 18:1414–1428, 2020.

[97] J. Nivre and J. Nilsson. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 99–106, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

[98] A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, jul 1988.

[99] S. Park, S. W. Lee, J. Kwak, M. Cha, and B. Jeong. Activities on facebook reveal the depressive state of users. *Journal of Medical Internet Research*, 15(10):e217, oct 2013.

[100] Y. Peng, C. O. Tudor, M. Torii, C. H. Wu, and K. Vijay-Shanker. iSimp in BioC standard format: enhancing the interoperability of a sentence simplification system. *Database*, 2014(0):bau038–bau038, may 2014.

[101] V. Petri, P. Jayaraman, M. Tutaj, G. Hayman, J. R. Smith, J. De Pons, S. JF Laulederkind, T. F. Lowry, R. Nigam, S. Wang, M. Shimoyama, M. R. Dwinell, D. H. Munzenmaier, E. A. Worthey, and H. J. Jacob. The pathway ontology – updates and applications. *Journal of Biomedical Semantics*, 5(1):7, 2014.

[102] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, Nov. 2019.

[103] R. Plutchik. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pages 3–33. Elsevier, 1980.

[104] M. Ponza, P. Ferragina, and S. Chakrabarti. On computing entity relatedness in wikipedia, with applications. *Knowledge Based Systems*, 188, 2020.

[105] M. Ponza, P. Ferragina, and F. Piccinno. Swat: A system for detecting salient wikipedia entities in texts. *Computational Intelligence*, 35(4):858–890, 2019.

[106] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, and G.-B. Huang. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems*, 69:108–123, oct 2014.

[107] J. Posner, J. A. Russell, and B. S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715, 2005.

[108] M. J. Power. The structure of emotion: An empirical comparison of six models. *Cognition & Emotion*, 20(5):694–713, aug 2006.

[109] T. C. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477, Dec. 2003.

[110] B. Rink, S. Harabagiu, and K. Roberts. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600, sep 2011.

[111] N. Rucci, D. Millimaggi, M. Mari, A. Del Fattore, M. Bologna, A. Teti, A. Angelucci, and V. Dolo. Receptor activator of nfkb ligand enhances breast cancer– induced osteolytic lesions through upregulation of extracellular matrix metalloproteinase inducer cd147. *Cancer Research*, 70(15):6150–6160, July 2010.

[112] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[113] J. A. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, sep 1977.

[114] S. Sarntivijai, Y. Lin, Z. Xiang, T. F. Meehan, A. D. Diehl, U. D. Vempati, S. C. Schürer, C. Pang, J. Malone, H. Parkinson, Y. Liu, T. Takatsuki, K. Saijo, H. Masuya, Y. Nakamura, M. H. Brush, M. A. Haendel, J. Zheng, C. J. Stoeckert, B. Peters, C. J. Mungall, T. E. Carey, D. J. States, B. D. Athey, and Y. He. CLO: The cell line ontology. *Journal of Biomedical Semantics*, 5(1):37, 2014.

[115] K. R. Scherer and H. G. Wallbott. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328, 1994.

[116] L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein, K. Bisordi, N. Campion, B. Hyman, D. Kurland, C. P. Oates, S. Kibbey, P. Sreekumar, C. Le, M. Giglio, and C. Greene. Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Research*, 47(D1):D955–D962, Nov. 2018.

[117] T. Simon, A. Goldberg, L. Aharonson-Daniel, D. Leykin, and B. Adini. Twitter in the cross fire—the use of social media in the westgate mall terror attack in kenya. *PLoS ONE*, 9(8):e104136, aug 2014.

[118] T. Slater. Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discovery Today*, 19(2):193–198, Feb. 2014.

[119] P. Smialowski, P. Pagel, P. Wong, B. Brauner, I. Dunger, G. Fobo, G. Frishman, C. Montrone, T. Rattei, D. Frishman, and A. Ruepp. The negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Research*, 38(suppl_1):D540–D544, Nov. 2009.

[120] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, Nov. 2007.

[121] J. Staiano and M. Guerini. Depeche mood: a lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2014.

[122] C. Strapparava and A. Valitutti. Wordnet affect: an affective extension of wordnet. In *LREC*, 2004.

[123] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, S. M., A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen, and C. Von Mering. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362–D368, Oct. 2016.

[124] Y. Tang, B. Jin, and Y.-Q. Zhang. Granular support vector machines with association rules mining for protein homology prediction. *Artificial Intelligence in Medicine*, 35(1-2):121–134, sep 2005.

[125] M. Thelwall, D. Wilkinson, and S. Uppal. Data mining emotion in social network communication: Gender differences in myspace. *Journal of the American Society for Information Science and Technology*, 61(1):190–199, 2010.

[126] H. Ulrich and M. M. Pillat. CD147 as a target for COVID-19 treatment: Suggested effects of azithromycin and stem cell engagement. *Stem Cell Reviews and Reports*, 16(3):434–440, Apr. 2020.

[127] S. Vijayarani, R. Janani, et al. Text mining: open source tokenization tools-an analysis. *Advanced Computational Intelligence: An International Journal (ACII)*, 3(1):37–47, 2016.

[128] K. Wang, W. Chen, Z. Zhang, Y. Deng, J.-Q. Lian, P. Du, D. Wei, Y. Zhang, X.-X. Sun, L. Gong, X. Yang, L. He, L. Zhang, Z. Yang, J.-J. Geng, R. Chen, H. Zhang, B. Wang, Y.-M. Zhu, G. Nan, J.-L. Jiang, L. Li, J. Wu, P. Lin, W. Huang, L. Xie, Z.-H. Zheng, K. Zhang, J.-L. Miao, H.-Y. Cui, M. Huang, J. Zhang, L. Fu, X.-M. Yang, Z. Zhao, S. Sun, H. Gu, Z. Wang, C.-F. Wang, Y. Lu, Y.-Y. Liu, Q.-Y. Wang, H. Bian, P. Zhu, and Z.-N. Chen. CD147-spike protein is a novel route for SARS-CoV-2 infection to host cells. *Signal Transduction and Targeted Therapy*, 5(1), Dec. 2020.

[129] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267, aug 2007.

[130] S.-J. Wang, H.-Y. Cui, Y.-M. Liu, P. Zhao, Y. Zhang, Z.-G. Fu, Z.-N. Chen, and J.-L. Jiang. CD147 promotes src-dependent activation of rac1 signaling through STAT3/DOCK8 during the motility of hepatocellular carcinoma cells. *Oncotarget*, 6(1):243–257, Nov. 2014.

[131] A. B. Warriner, V. Kuperman, and M. Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207, feb 2013.

[132] C. Wei, A. Allot, R. Leaman, and Z. Lu. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593, May 2019.

[133] O. T. WG. Phenotype and trait ontology.

[134] M. Whirl-Carrillo, E. M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C. F. Thorn, R. B. Altman, and T. E. Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, Oct. 2012.

[135] P. O. Williamson and C. I. J. Minter. Exploring PubMed as a reliable resource for scholarly communications services. *Journal of the Medical Library Association*, 107(1), jan 2019.

[136] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson. DrugBank 5.0: a major update to

the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, Nov. 2017.

[137] M. Xiaoke and G. Lin. Biological network analysis: insights into structure and functions. *Briefings in Functional Genomics*, 11(6):434–442, 11 2012.

[138] L. Xiong, C. Edwards, and L. Zhou. The biological function and clinical utilization of CD147 in human diseases: A review of the current scientific literature. *International Journal of Molecular Sciences*, 15(10):17411–17441, Sept. 2014.

[139] J. Yuan, Z. Jin, H. Guo, H. Jin, X. Zhang, T. Smith, and J. Luo. Constructing biomedical domain-specific knowledge graph with minimum supervision. *Knowledge and Information Systems*, 62(1):317–336, Mar. 2019.