Università degli Studi di Salerno

Dipartimento di Scienze Aziendali – Management
& Information Systems

Dottorato di Ricerca in Big Data Management
Ciclo 33 – a.a 2019/2020

TESI DI DOTTORATO / PH.D. THESIS

Knowledge Extraction in Big Data era: Veracity and Value Challenges

MARIACRISTINA GALLO

TUTOR: **PROF. VINCENZO LOIA**
CO-TUTOR: **PROF. GIUSEPPE FENZA**

PHD PROGRAM DIRECTOR: **PROF. VALERIO ANTONELLI**

Abstract

With the Big Data explosion, companies have the opportunity to access a massive amount of data that can improve their efficiency in terms of decision-making, adopted solutions, customer care, and so on. By conveniently structuring the Knowledge Extraction processes, companies can easily convert information into opportunities. However, in continuously evolving contexts, a significant analysis should be dedicated to the data quality assessment to deal with unreliable information. Furthermore, designed decision-making solutions should be aware of data drift and (re)adapt themselves along their lifecycle.

In this sense, the thesis work proposes Data Mining methodologies that take into account Veracity and Value challenges underlying Big Data. The meaning of Veracity in the context of Big Data concerns with the truthful of a data set and how trustworthy the data source, type, and processing is. However, the Value of Big Data is strictly related to the Veracity (or quality) of treated data. In fact, integrity awareness about data and its sources is crucial if we are trying to extract information from huge amounts of data. Some of the main achievements of this thesis work are summarized following:

- The application of the well-known theory of Formal Concept Analysis and its variants for extracting conceptualization models from different data streams contents (i.e., social media, papers, etc.).

- The definition and experimentation of a method for cross-relating data sources, with different velocity, size, and credibility levels, by joining conceptualization models to support information trustworthiness (i.e., Veracity) and enable an information filtering system.

- The definition and experimentation of a drift-aware deep learning model based on LSTM for adaptively recognizing and distinguishing evolving

energy consumption behaviors pruning the risk of false-positive alarm about frauds.

- The definition of a consistency measure based on Fuzzy Consensus model, a method widely used in Group Decision Making, to support the training data value assessment before applying a machine learning algorithm for extracting a predictive model.

Presented methodologies are supported by the application and experimentation on several real-world application scenarios giving an idea of their applicability and effectiveness. Faced problems include recommendations, anomaly detection, fake news detection, pharmacovigilance, Emergency Department overcrowding, etc.

Acknowledgments

This thesis work represents the conclusion of my Ph.D. course, but it is the result of more than three years devoted to research.

Thanks to the passion that characterizes my supervisors' work, my research interest has grown over time. Prof. Giuseppe Fenza (Giuseppe, for me), and Prof. Vincenzo Loia, through him, introduced and guided me along this way. I am grateful for their teachings and their trust in my skills.

In nearly 12 years of work, I met a lot of friends: Carmen, Alberto, Teresa, Monica, Angela, ... everyone taught me something.

My special thanks to my children Marirosa and Giuseppe. They certainly made a lot of sacrifices during my long days of work. For that, I am deeply grateful to my parents and parents-in-law, who took care of them since they were babies. I would never have done it without them.

Finally, I would like to express my gratitude to Nicola, who always believed in me!

The value of a sentiment is the amount of sacrifice you are prepared to make for it.
(John Galsworthy)

To my family.

Contents

Chapter 1

Introduction

The Big Data era, characterized by the availability and accessibility of a large amount of data, represents a relevant resource for organizations in terms of value transformation. Since it is strictly related to the organizations' ability to extract knowledge and transform data into actionable information, Big Data opens a series of challenges related to their analysis: search, sharing, querying, visualization, storage, etc.

In the last decade, we assisted in exponential growth in the number of social media users from 970 million in 2010 to 3.81 billion in 2020. In parallel, the increasing amount of network-enabled devices as the Internet of Things (IoT) contributed to the spreading of data among devices and the tuning of systems devoting to real-time analytics, machine learning, and so on.

Available data is a valuable resource for private companies and public organizations. By conveniently processing data, it is possible to tune recommender systems, empower decision support systems through understanding scenarios, and, ultimately, support the experts during the decision-making process. It follows that, by adopting suitable methodologies, more adaptive models can be achieved, giving companies better awareness.

In this sense, this thesis work deals with Veracity and Value challenges linked to Big Data. The Veracity of data refers to the degree of reliability of content and its source and directly affects the Value of Big Data. A data trustworthiness evaluation is crucial before extracting information from a huge amount of data. It is not unusual that, for example, social media users publish posts about crisis before the official and more reliable sources; nevertheless, the contents of the posts cannot always be seen as highly credible. In this scenario, an efficient credibility assessment system should rapidly

recognize untrustworthy information.

Data is also a starting-point in decision-making. By analyzing graphs, reports, and dashboards, domain experts must make decisions that influence the companies' ability to enter markets, gain profitability and potentially win over the competitors. Systems should enable the *Human in the Loop* (HITL) paradigm giving an understandable snapshot of data and guide the expert in the best way.

Due to the increase of available datasets, machine learning algorithms significantly improved in terms of performance and adoption in many research areas. However, data quality intrinsically influences the quality of predictive modeling; besides, the velocity of data modification could rapidly nullify its performance. In this sense, by defining a measure of a dataset's quality, it becomes possible to: (1) determine the suitability of the training set; (2) evaluate the adaptability of the model to the evolution of data.

Mentioned scenarios prove an effective role of Veracity and Value in Big Data management. *Knowledge Extraction* from data and consequential predictions require credible and consistent data to derive learning models able to evolve together with data itself.

This thesis work describes methodologies sharing a Knowledge Extraction process creating knowledge from structured and unstructured sources. In particular, by leveraging data mining techniques, presented solutions try to measure the Veracity and Value of data contributing to the achievement of more efficient and reliable Knowledge Extraction methodologies.

1.1 Objectives

This thesis work is mainly devoted to the definition of Knowledge Extraction methodologies applicable to stream data (i.e., social media or IoT stream). The adoption and the orchestration of Data Mining techniques enable the realization of flexible models that consider the data quality and extract value from data. Essentially, realized solutions go towards the following objectives:

- By cross-relating different information sources with varying trustworthiness levels, a framework based on the conceptualization of contents tries to assign credibility to social media information and assists information filtering operations.

- A conceptualization achieved through a syntactic and semantic analysis

of contents that also consider context information, guides the realization of two different recommender systems focused on social media.

- An adaptive anomaly detection methodology able to distinguish between anomalies and changes (drifts) in normal behaviors is designed and applied to an energy theft scenario.

- A descriptive and statistical analysis conducted on accesses to Emergency Department (ED) tries to understand the role of users' social conditions on the number of inappropriate accesses. The study represents support for administrations in decision-making related to the improvement of territorial assistance and other health care related solutions.

- Regarding the "real" value of a dataset, experimentation was conducted to understand the dataset consistency's role during the deep neural network (DNN) training and the hypothesis to adopt the Group Decision Making consensus as consistency measure.

1.2 Research Questions

This thesis work aims to reply to subsequent research questions:

- How is it possible to measure the trustworthiness of so huge user-generated content shared, for instance, on social media, to enable the extraction of useful insights?

- What are the main features influencing data value? Can contextual ones empower, for instance, a recommender system performance?

- How is it possible to measure the consistency of the dataset that allows data scientists to know in advance the machine learning model's performance?

1.3 Results overview

Following there is an overview of the main results.

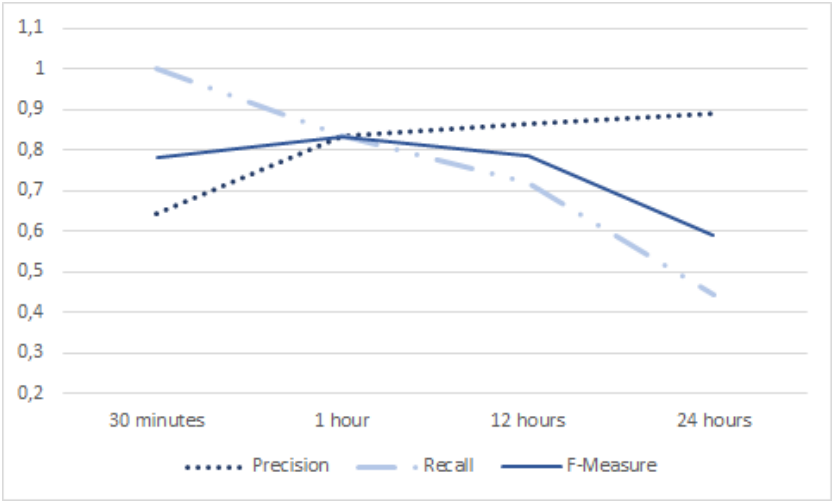

Figure 1.1: Precision, Recall and F-Measure of the Fake News' recognizer

The credibility assessment through the cross-relation of different information sources filters the data and supports fake news recognition with performance shown in Figure 1.1. Data coming in a fixed time interval (i.e., 30 minutes, 1 hour, 12 hours, 24 hours) is grouped and a credibility level assigned. Through a threshold establishment, information with low credibility is considered fake.

A recommender system built by adopting triadic timed Formal Concept Analysis based on the Semantics of tweets content, contextual dimensions like Time, and Location, achieves good performance as shown in Figure 1.2, especially in the afternoon hours. The adaptive anomaly detection methodology, applied to an energy theft scenario, reveals the capacity to distinguish between anomalies and consumption behaviour drift. Figures 1.3 shows the predictions during the consumption behaviour drift instead of the anomaly (depicted in Figure 1.4). In particular, experimental results achieve Precision and Recall values equal to 78% and 88%, respectively. The statistical study on the ED in Salerno city reveals an influence on inappropriate accesses from the patients' age, geographic origin, type of made diagnosis, and day of the week (see Table 1.1).

Finally, the Consensus Model widely used in the Group Decision Making problems is a useful consistency measure in data value assessment for Learning to Rank problems. The experimental results reveal the correlation between its evaluation over the training dataset and DNN's performance (see

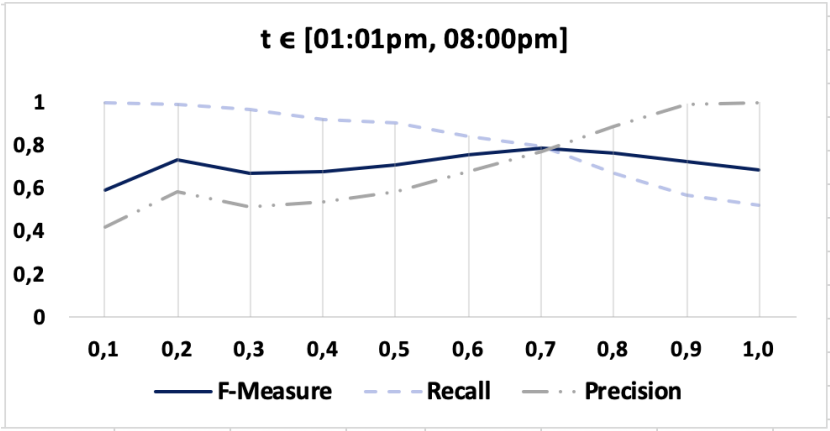

Figure 1.2: F-Measure evaluated by varying the level of threshold $\alpha \in [0, 1]$ in the time slot $[01:01pm - 08:00pm]$ for the recommender system.

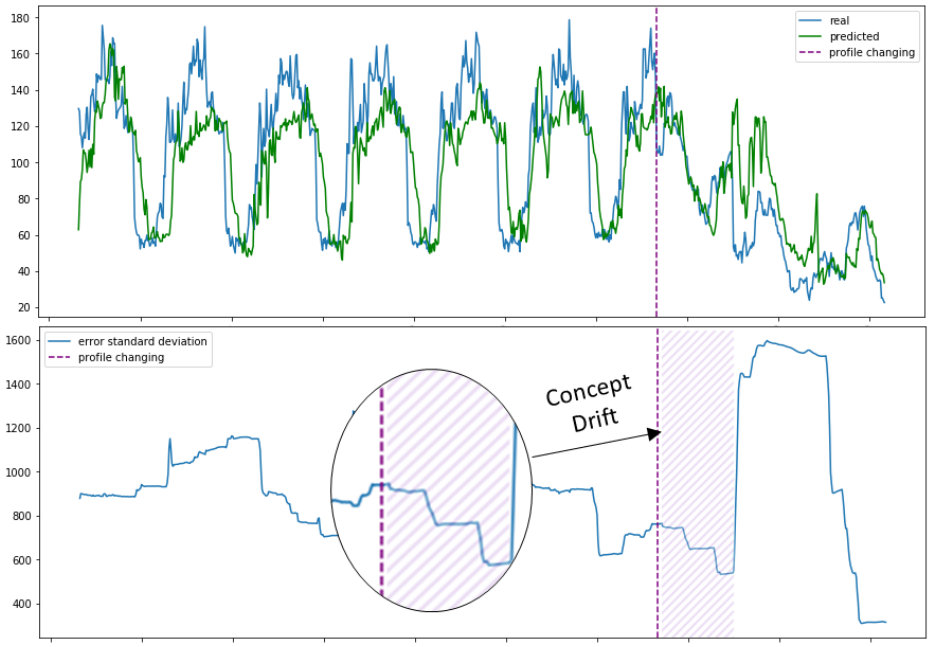

Figure 1.3: Example of concept drift recognition

Figure 1.5).

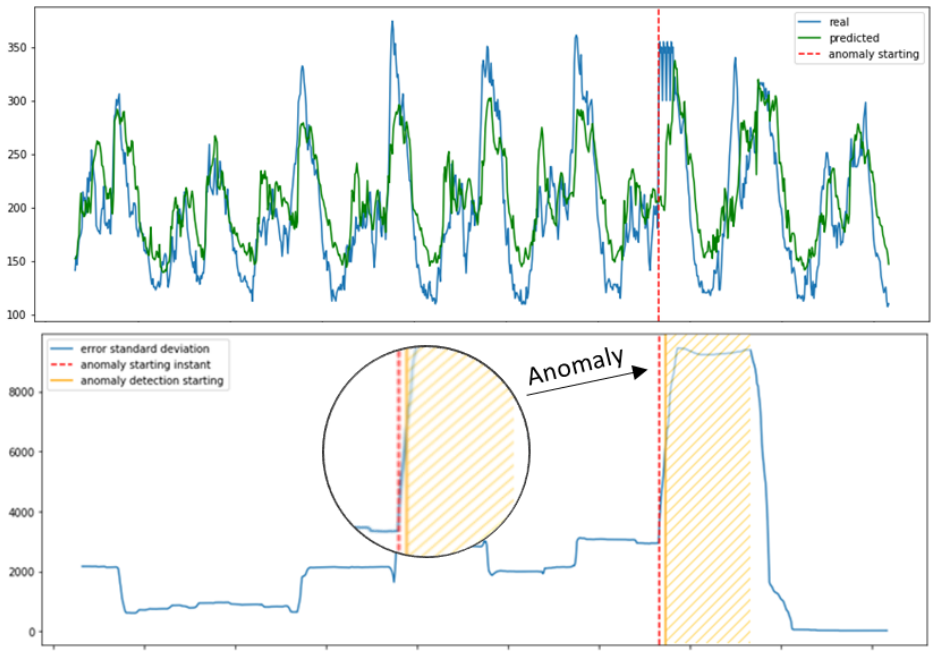

Figure 1.4: Example of anomaly detection recognition

Table 1.1: Regression Results.

	coef	*stderr*	*t*	$P > \|\|t\|\|$	[0.025	0.975]
Month	0.0204	0.080	0.255	0.799	-0.137	0.178
Day of Week	**-0.581**	**0.137**	**-4.225**	**0.000**	**-0.850**	**-0.311**
Day	0.027	0.031	0.880	0.379	-0.034	0.088
Age Class	**1.319**	**0.183**	**7.218**	**0.000**	**0.960**	**1.677**
Provenance	**4.386**	**0.131**	**33.389**	**0.000**	**4.128**	**4.643**
Diagnosis	**-0.006**	**0.001**	**-4.604**	**0.000**	**-0.009**	**-0.004**

1.4 Thesis Outline and Publications

This document consists of three main parts, structured as follows:

1. Part I describes the theoretical background related to treated aspects in this thesis. In particular, we present Data Mining techniques (Chapter 2) and theories of Group Decision Making (Chapter 3).

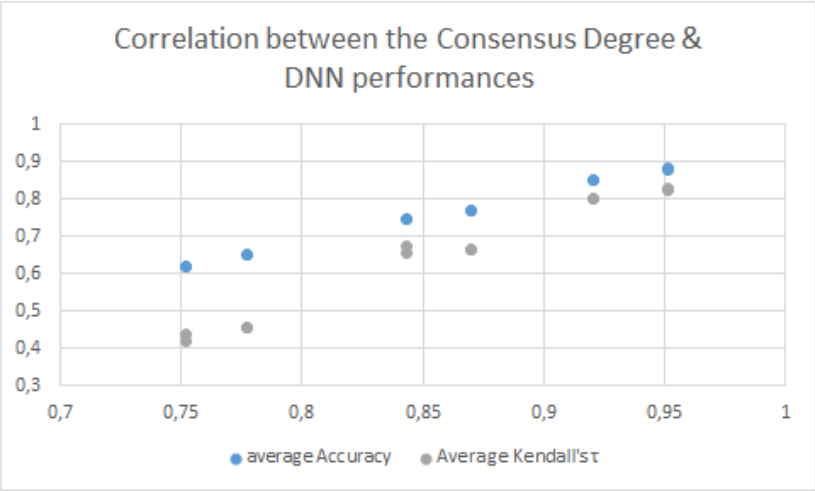

Figure 1.5: Correspondence among evolution of performance coefficients.

2. Part II presents the main methodologies introduced and experimented with during this thesis work. Chapter 4 describes a framework for knowledge extraction from text streams. An adaptive anomaly detection methodology is presented in Chapter 5. A method for the estimation of data quality in ranking problems is described in Chapter 6.

3. Part III presents solutions adopting methodologies (or combinations of them) introduced in Part II. In particular, Chapter 7 summarizes two real scenarios related to credibility assessment on text streams. Chapter 8 describes two recommender systems; three solutions about the decision-making process are presented in Chapter 9.

Presented solutions consists of subsequent publications:

- The method in Section 7.2 was published in 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS) with the title "Cross-relating heterogeneous Text Streams for Credibility Assessment" [1].

- The method in Section 7.3 was published in Future Generation Computer Systems with the title "Pharmacovigilance in the era of social media: Discovering adverse drug events cross-relating Twitter and PubMed" [2].

- The method in Section 8.2 was published in Soft Computing with the title "Who and where: context-aware advertisement recommendation on Twitter" [3].
- The method in Section 8.3 was published in Future Generation Computer Systems with the title "Time-aware adaptive tweets ranking through deep learning" [4].
- The method in Section 9.1 was published in IEEE Access with the title "Drift-Aware Methodology for Anomaly Detection in Smart Grid" [5].
- The method in Section 9.2 was published in Journal of Data, Information and Management with the title "Etiology of emergency department overcrowding: descriptive analytics of inappropriate accesses at Salerno hospital in Italy" [6].
- The method in Section 9.3 was published to Applied Soft Computing with the title "Data set quality in Machine Learning: Consistency measure based on Group Decision Making" [7].

Part I

Theoretical Background

Chapter 2

Data Mining

Data Mining, in the Knowledge Extraction process, aims to detect inherited knowledge from large amounts of data [8]. The main objective of a data mining algorithm is to create a mathematical model representing data and its inherent regularities.

Data Mining techniques can be divide into *Predictive modeling* and *Descriptive modeling*, basing on the kind of information (attributes) known and the type of knowledge sought from the data-mining model. Predictive modeling is used to estimate the value of a particular target attribute. There are sample training data for which values of that attribute are known (i.e., Classification and Regression). In the Descriptive modeling, the proper groups are not known in advance; the patterns discovered by analyzing the data are used to determine the groups (i.e., Clustering).

At the end of the data mining application, generated patterns could be numerous. However, only the interesting ones represent real *knowledge*. Generally, a pattern is interesting if it is (1) easily understood by humans, (2) valid on new data with some degrees of certainty, (3) potentially useful, and (4) novel. A pattern is also interesting if it validates a hypothesis that the user sought to confirm.

Data mining incorporates techniques from many other domains, such as statistics, machine learning, pattern recognition, information retrieval. The following sections describe the adopted Data Mining technologies to implement methodologies presented in Part II of this thesis.

2.1 Natural Language Processing

The processing of text stream, aiming to give a feature representation of contents, is mainly done through syntactic and semantic tools. The syntactic analysis studies the text structure; while, the semantic analysis supports the content understanding. Following, adopted tools are described.

2.1.1 Stanford CoreNLP

Stanford CoreNLP[1] is a tool offering a series of Natural Language Processing tasks for numerous human languages. It is mainly adopted for identifying token and sentence boundaries, parts of speech, named entities, numeric and time values, dependency and constituency trees, and so on. The available java library offers a web API server, as well as standalone access. The core of the library is the "pipeline": it takes the text and returns full annotation objects. An example of analysis results is shown in Figure 2.1.

2.1.2 Wikipedia Miner

Wikipedia is a free online encyclopedia, constantly update by volunteers around the world. It contains millions of articles available for a large number of languages. The Wikipedia Miner toolkit is an open-source software system designed to integrate Wikipedia's semantics into applications [9]. The toolkit creates a local database copy of Wikipedia's contents (available in different languages) and offers a Java API to access them. The most interesting provided services are:

- ExploreArticle: takes in input a title or an id of an article and returns details about the corresponding page (definition, links, labels).
- ExploreCategory: corresponds to ExploreArticle service for categories.
- Compare: given a pair of terms or ids, returns a measure of correlation among corresponding pages.
- Wikify!: given a portion of text or a hypertext document, returns the same text annotated with links to the relevant Wikipedia articles. The

[1] https://stanfordnlp.github.io/CoreNLP/index.html

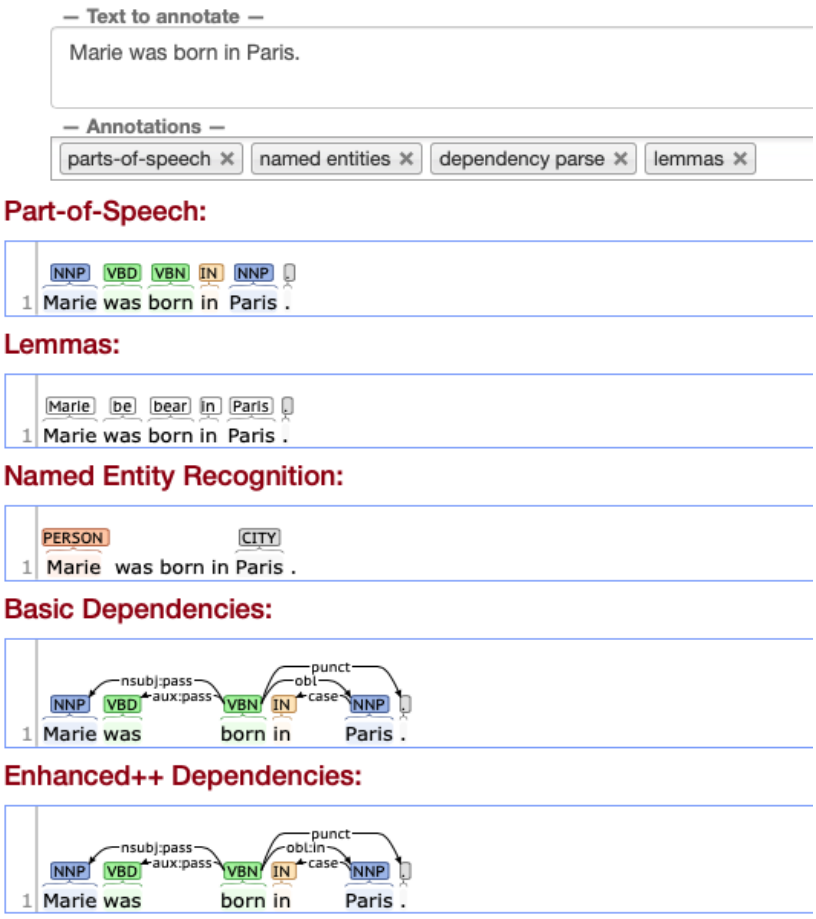

Figure 2.1: Example of CoreNLP's text annotation.

Wikify! service (also known as "text wikification") is useful for automatic keyword extraction and word sense disambiguation [10].

Figure 2.2 illustrates an example of Wikification: words contained into a tweet text are disambiguated and linked to Wikipedia pages (i.e., Topics).

2.1.3 DBpedia Spotlight

DBpedia Spotlight is an open-source tool for automatically detecting DBpedia resources in texts and providing a disambiguation task [11]. It can be

Figure 2.2: Wikification example

asked through REST-based web services. The service highlights text concerning a concept and links it to the concept unique identifier. It also associates a similarity score to the disambiguation task. The disambiguation algorithm is based upon cosine similarities and modified TF-IDF weights: the higher is the match score, the more reliable is the disambiguation result.

An example of the disambiguation task applied to a tweet is shown in Figure 2.3.

DBpedia Spotlight is particularly suitable for subsequent characteristics [12], [13]:

- DBpedia Spotlight allows users to configure the annotations to their specific needs through the DBpedia Ontology and quality measures such as prominence, topical pertinence, contextual ambiguity, and disambiguation confidence.

- DBpedia Spotlight is shared as open source and deployed as a Web Service freely available for public use.

- DBpedia Spotlight annotates text documents with DBpedia URIs. DBpedia [14] is developing into an interlinking hub in the Web of Data that enables access to many data sources in the Linked Open Data

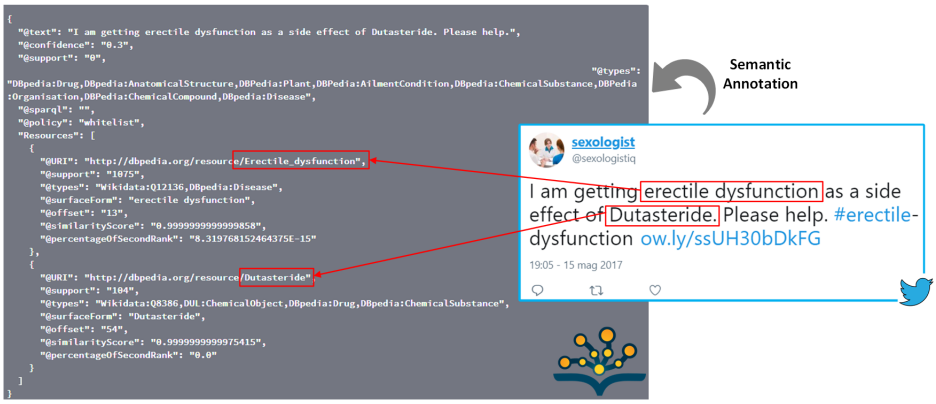

Figure 2.3: Example about semantic annotation of tweet text by means of DBPedia Spotlight.

cloud. It contains encyclopedic knowledge from Wikipedia for about 3.5 million resources. Several tools have been built on top of it (e.g., Open Calais, Zemanta, LODr, and TopBraid Composer).

2.2 Pearson correlation coefficient

In statistics, the *correlation* measures statistical relationships between two random variables. Generally speaking, the correlation evaluates how much two variables are linearly dependent. One of the most popular metrics of correlation measurement is the **Pearson Correlation Coefficient** [15]. Pearson correlation coefficient (ρ) is the covariance of the two variables divided by the product of their standard deviations. Formally, given a pair of random variables (X, Y), the Pearson correlation coefficient is evaluated as follows:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{2.1}$$

where *cov* is the covariance, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y. Equally, the equation can be written in terms of mean and expectation:

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{2.2}$$

where σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y, μ_X is the mean of X, μ_Y is the mean of Y, and $\mathbb{E}$ is the expectation.

The correlation coefficient ranges from -1 to 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of -1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables.

2.3 Classification and Regression

Classification is the problem of identifying a new item's belonging category based on a training dataset in which categories are known. Regression is a similar process in which the target value is continuous instead of categorical. Classification and regression, as well as clustering, are Machine Learning (ML) methodologies. In machine learning, specific artificial neural networks are designed to learn hidden data properties. Artificial neural networks try to make predictions for new unseen data after having experienced a training dataset. Deep learning is a branch of machine learning in which a layered structure characterizes neural networks and which layers aim for improving efficiency, trainability, and understandability.

2.3.1 Learning to rank

Learning to Rank is a machine learning technique aiming to construct a ranking model able to rank a list of items optimally. It can be implemented by different approaches [16]: Pointwise, Pairwise, and Listwise, briefly described following.

In the Pointwise method (e.g., PRank [17]), the ranking problem is approximated by a classification, regression, or ordinal classification problem. Given a query-document pair, the model must predict its score.

In the Pairwise approach (e.g., Ranking SVM (RankSVM) [18], RankNet [19]), the ranking problem is approximated by a binary classification problem where, given a query and a pair of documents, the model must predict the best one for such query.

The Listwise approach (e.g., ListNet [20], PermuRank [21]) takes the overall ranking list during the learning and prediction processes. For that reason, it can result in more complex than other approaches.

For the first two approaches existing algorithms can be applied; for the last one, ad-hoc algorithms must be implemented.

2.3.2 Long Short Term Memory network

Long Short Term Memory [22] is a particular type of Recurrent Neural Network (RNN). RNN is a type of neural network designed to understand the data's sequential nature by using history. It is technically achieved by connecting the Artificial Neural Network nodes and creating a directed graph along a temporal sequence. The architecture of LSTM consists of one input layer, one hidden (i.e., LSTM) layer, one output layer, and some memory blocks. Each memory block has multiple cells that have recurrent connections among them, an input gate (i), a forget gate (f), and an output gate (o). The input gate learns information to store in the memory. The forget gate learns the length of stored information, and the output gate learns when the stored information should be used. An example of a single memory block is shown in Figure 2.4.

The purpose of an LSTM network consists of identifying a correlation between an input sequence $x = (x_1, \ldots, x_T)$ and an output sequence $y = (y_1, \ldots, y_T)$. It is done by calculating the network unit activations by processing the following equations iteratively [24]:

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i) \tag{2.3}$$

$$f_t = \sigma(W_{fx}x_t + W_{mf}m_{t-1} + W_{cf}c_{t-1} + b_f) \tag{2.4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \tag{2.5}$$

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_t + b_o) \tag{2.6}$$

$$m_t = o_t \odot h(c_t) \tag{2.7}$$

$$y_t = \phi(W_{ym}m_t + b_y) \tag{2.8}$$

where the W terms denote weight matrices (e.g., W_{ix} is the matrix of weights from the input gate to the input), the b terms denote bias vectors, σ is the logistic Sigmoid function, m is the cell output activation vector which have the same size of i, f, o, and c (the cell activation vectors). $\odot$ is the element-wise product of the vectors, g and h are the cell output and input activations that use hyperbolic tangent activation functions, ϕ is the softmax output activation function.

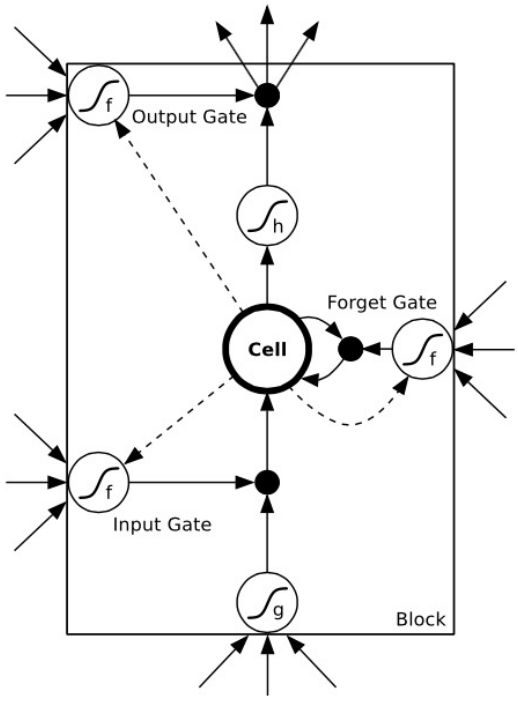

Figure 2.4: LSTM memory block [23]

2.3.3 Multiple Linear Regression

Multiple Linear Regression (MLR) is an extension of simple linear regression. It helps understand the role of multiple independent variables with respect to a single dependent variable by modeling the linear relationship between them. The MLR is evaluated through Equation 2.9.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon \tag{2.9}$$

For $i = n$ observations, y_i is the dependent variable, x_i are the independent variables, β_0 is a constant called intercept, β_p are the regression coefficients of each x variable, and ϵ is the model's error term (i.e., the residuals).

The goodness of fit about the linear regression is usually measured by the *coefficient of determination* (i.e., R^2 or $R-squared$). It is a statistical metric used to calculate how much of the variation in outcome can be explained by the independent variables' variation.

2.4 Clustering

Clustering is a data mining technique aiming to group items based on their characteristics. A group (i.e., cluster) contains items more similar to each other than those in other groups.

Differently from classification and regression that need a labeled training dataset, clustering is an unsupervised learning technique. It makes groups based on a defined similarity measure evaluated among items.

2.4.1 K-Means Clustering

The K-means algorithm [25] is considered the most popular and simplest clustering algorithm. Starting from a given number of clusters K, it arranges the input objects in K clusters based on their attributes through a set of iterations. At each iteration, the algorithm chooses a centroid for each cluster and links each object to the nearest centroid based on their distance. The algorithm ends when the last iteration does not make any change in the location of the centroids.

More formally, let $X = \{x_1, \ldots, x_n\}$ be the set of observations to be clustered into a set of K clusters, $C = \{C_1, \ldots, C_K\}$ whose centroids are $\{c_1, \ldots, c_k\}$. The algorithm aims to minimize the sum of the squared distances of each object (x_i) with respect to its centroid (c_k):

$$\min \sum_{k=1}^{K} \sum_{x_i \in C_k} || \, x_i - c_k \, ||^2 \tag{2.10}$$

where $\sum_{x_i \in C_k} || \, x_i - c_k \, ||^2$ is the Euclidean distance between the centroid and all observations in C_k. Each centroid is evaluated as:

$$c_k = \frac{\sum_{x_i \in C_k} x_i}{n_k} \tag{2.11}$$

where n_k is the number of observations belonging to the cluster C_k.

The algorithm works as follows:

1. Select K observations as the initial set of centroids (generally in a random fashion or by heuristics).

2. Assign each observation to the cluster having the closest centroid.

3. Recalculate the centroids when all observations are arranged.

4. Repeat Steps 2 and 3 until centroids do not change.

2.4.1.1 Heuristics for determining the "K" clusters number

Among available methods, solutions adopted during this thesis work apply: the Elbow Method and the Silhouette Coefficient [26].

Elbow Method
Starting with $K = 2$ and increasing it in each step by 1, this method evaluates the clustering and its goodness employing the sum of squares distance inside each cluster. More precisely, if we plot the average sum of squares distance inside of each cluster (i.e., W_k), and the number of the clusters, we can see that the first clusters will add much information (explain a lot of variances) and, for a particular K, the graph begins to flatten significantly. This point is named "elbow"and is the value we are looking for. W_k is calculated as follows:

$$W_k = \sum_{r=1}^{k} \frac{1}{n_r} D_r \tag{2.12}$$

where k is the number of the clusters, n_r is the number of points in cluster r, and D_r is the sum of distances between all points in a cluster, evaluated as follows:

$$D_r = \sum_{i=1}^{n_r - 1} \sum_{j=i}^{n_r} \left\| d_i - d_j \right\| \tag{2.13}$$

Silhouette Coefficient
The Silhouette Coefficient measures the difference between the within-cluster tightness and separation from the rest. In particular, the silhouette width $s(o)$ for object $o \in O$ is defined as:

$$s(o) = \frac{b(o) - a(o)}{\max(a(o), b(o))} \tag{2.14}$$

where $a(o)$ is the average distance between o and all other objects of the cluster to which o belongs, and $b(o)$ is the minimum of the average distances between o and all the objects in each other cluster. Values of silhouette width range between -1 and 1. If all the silhouette width values are close to 1,

it means that the set O is well clustered. Clustering can be characterized by the average silhouette width S of individual objects. The largest average silhouette width, over different K, indicates the best number of clusters.

2.4.2 Fuzzy Formal Concept Analysis Theory

The Formal Concept Analysis (FCA) is a theoretical framework which supplies a basis for conceptual data analysis [27]. It is based on the theory of lattices [28] and enables representing relations between objects and attributes of a domain through the formal context. FCA sets up a mathematical model of knowledge (i.e., the concept lattice) more informative than traditional tree-like conceptual structures.

Fuzzy FCA is a combination of fuzzy logic with the FCA theory that enables the representation of uncertainty through a membership value range. In particular, Fuzzy FCA deals with fuzzy relations between objects and their features, considering membership varying in $[0, 1]$.

Following, some definitions about Fuzzy FCA are given.

Definition 1. *A* **Fuzzy Formal Context** *is a triple $K = (G, M, I)$, where G is a set of objects, M is a set of attributes, and $I = \big((G \times M), \mu\big)$ is a fuzzy set.*

Being I a fuzzy set, each pair $(g, m) \in I$ has a membership value $\mu(g, m)$ in $[0, 1]$. In the following, the fuzzy set function μ will be denoted by μ_I.

Definition 2. Fuzzy Representation of Object. *Each object O in a fuzzy formal context K can be represented by a fuzzy set $\Phi(O)$ as $\Phi(O)$=$\{A_1(\mu_1)$, $A_2(\mu_2)$,..., $A_m(\mu_m)\}$, where $\{A_1, A_2, \ldots, A_m\}$ is the set of attributes in K and μ_i is the membership of O with attribute A_i in K. $\Phi(O)$ is called the fuzzy representation of O.*

Fuzziness enables modeling relations between objects and attributes in a more smoothed way, ensuring more precise representation and uncertainty management.

Taking into account Fuzzy Formal Context, the Fuzzy FCA algorithm can identify Fuzzy Formal Concepts and subsumption relations among them. More formally, the definition of Fuzzy Formal Concept and order relation among them are given as follows.

Given a fuzzy formal context $K = (G, M, I)$ and a confidence threshold χ, for $G' \subseteq G$ and $M' \subseteq M$, we define $G^* = \{m \in M \mid \forall g \in G',\ \mu_I(g, m) \geq \chi\}$ and $M^* = \{g \in G \mid \forall m \in M',\ \mu_I(g, m) \geq \chi\}$.

Definition 3. Fuzzy Formal Concept. *A fuzzy formal concept (or fuzzy concept) C of a fuzzy formal context K with a confidence threshold χ, is $C = (I_{G'}, M')$, where, for $G' \subseteq G$, $I_{G'} = (G', \mu), M' \subseteq M, G^* = M'$ and $M^* = G'$. Each object g has a membership $\mu_{I_{G'}}$ defined as*

$$\mu_{I_{G'}}(g) = min_{m \in M'} \left(\mu_I(g, m)\right) \tag{2.15}$$

where μ_I is the fuzzy function of I.

Note that if $M' = \emptyset$ then $\mu_I(g) = 1$ for every g. G' and M' are the extent and intent of the formal concept $(I_{G'}, M')$, respectively.

In addition, let us define the *Fuzzy Formal Concept Support* (briefly, *Support*).

Definition 6: *Fuzzy Formal Concept Support*. *Let $K = (G, M, I)$ be a fuzzy formal context, the support of a Fuzzy Formal Concept $C' = (I_{G'}, M')$ is given by*

$$Supp(C') = \frac{|G'|}{|G|} \tag{2.16}$$

Chapter 3

Group Decision Making

Group Decision Making (GDM) is a model by which, given a finite set of alternatives, a group of experts expresses individual preferences. Then, the alternative with the higher number of preferences becomes the choice for the group. The process can produce a consensus degree among decision-makers and a shared ranking of preferences.

In the thesis scope, GDM is the process adopted to measure a dataset's consistency: the higher the consensus degree among decision-makers, the higher the consistency. In particular, we propose a soft consensus model [29] dealing with vague or imprecise experts' opinions during the consensus process, as detailed following.

3.1 Consensus evaluation process

Given a set of decision makers $DM = \{dm_1, dm_2, \ldots, dm_m\}$, $(m \geq 2)$ and a set of alternative branches $AB = \{ab_1, ab_2, \ldots, ab_n\}$, $(n \geq 2)$, we assume that experts provide their opinions in terms of fuzzy preferences values $ab^k_{dm_i} \in [0, 1]$ that enable the evaluation of fuzzy preference relations as described in [30].

Definition 4. *A fuzzy preference relation P on a set of alternatives X can be represented by a fuzzy set on the product set $X \times X$, i.e., it is characterized by a membership function $\mu_P : X \times X \to [0, 1]$.*

Preference relations are usually represented by a $n \times n$ matrix $P = (p_{ik})$ where $p_{ik} = \mu_P(x_i, x_k)$ $(\forall i, k \in 1, \ldots, n)$ is the preference degree of the alternative x_i over x_k. In other words, $p_{ik} = \frac{1}{2}$ indicates an equal preference

degree between x_i and x_k while, $p_{ik} = 1$ indicates that x_i is absolutely preferred to x_k, and $p_{ik} \geq \frac{1}{2}$ indicates that x_i is preferred to x_k. Consequently, values on the diagonal line are always equal to $\frac{1}{2}$ (i.e., $p_{ii} = \frac{1}{2}\ \forall i \in 1, \ldots, n$).

Definition 5. *Let be $X = \{x_1, ..., x_n\}$ the set of alternatives, and $ab^i_{dm_k}$ the evaluation of alternative x_i by the expert dm_k. Then, the intensity of preference of alternative x_i on alternative x_j, $p^k_{i,j}$, for the expert dm_k, is evaluated by the following transformation function:*

$$p^k_{ij} = \varphi(ab^i_{dm_k}, ab^j_{dm_k}) = \frac{1}{2} \cdot (1 + ab^i_{dm_k} - ab^j_{dm_k}) \tag{3.1}$$

3.2 Consensus degree evaluation

Starting from the fuzzy preference relation matrices P^{dm} constructed for each decision-maker $dm \in DM$, it is possible to calculate the consensus degree, reached by the involved decision-makers, at each iteration.

Firstly, the moderator constructs the *similarity matrix* $SM^{kl} = (sm^{kl}_{ij})$ for each pair of decision-makers $\{dm_k, dm_l\}$, where $sm^{kl}_{ij} = 1 - |p^k_{ij} - p^l_{ij}|$, and $p_{ij} = \mu_P(x_i, x_j)$ is the intensity of the alternative x_i over x_j. The consensus degree is evaluated through the following equation:

$$cm_{ij} = \phi(sm^{kl}_{ij}), k = 1, ..., m-1, l = k+1, ..., m \tag{3.2}$$

where ϕ is an aggregation function (e.g., arithmetic mean, median).

Then, global consensus degree *co* is evaluated by aggregating consensus degrees of each alternative branch, as defined in [31]:

$$co = \frac{\sum_{i=1}^{n} ca_i}{n} \tag{3.3}$$

where:

$$ca_i = \frac{\sum_{j=1, j\neq i}^{n} (cm_{ij} + cm_{ji})}{2(n-1)} \tag{3.4}$$

The global consensus degree, *co*, identifies the level of agreement among all the decision-makers about the alternative branch. The higher is *co* (i.e., closer to 1), the higher the consensus among decision-makers. Usually, if the consensus degree is higher than a given threshold, the process ends; otherwise, some strategies can be implemented in terms of decision-makers' weight to increase the consensus degree level in the next round of consensus.

Part II

Methodologies & Applications

Chapter 4

Knowledge Extraction from Text Stream

Non-structured information (e.g., text stream) needs an intense pre-processing to be analyzed. In this sense, one of the leading research objectives attempted and described in this thesis work regards the Pattern Recognition from text guided by a Context-Aware Knowledge Extraction (CA-KE). Free text is acquired through ad-hoc scraping or querying from data sources (e.g., Twitter, Google News, PubMed's abstracts). It is then extracted and analyzed syntactically and semantically, obtaining features representing it and adopted to make a conceptualization of contents. Finally, a matching criterion between concepts is defined.

The methodology is composed of:

1. CA-Feature Extraction phase, aiming to extract syntactic and semantic features from the text;
2. CA-Concepts Mining phase, which applies the Fuzzy Formal Concept Analysis and defines a hierarchy formally representing evaluated text;
3. CA-Match-Making that correlates concepts of different lattices generated during the previous phase.

4.1 CA-Feature Extraction

This phase aims to identify the best representative feature for text content. The adopted methodology involves syntactic and semantic analysis. Syn-

tactic analysis is done by leveraging Natural Language Processing (NLP) tools and methods and aims to represent the text's structure. The semantic analysis aims to "understand" text meaning. As mentioned in Section 2.1, adopted NLP methodologies are POS Tagging and Named-Entity Recognition (NER), and the adopted tool is CoreNLP. Two possible ways for the semantic analysis are available: (1) a content wikification, or (2) the DBpedia Spotlight web service application. Both services link text portions to the corresponding resources of the referenced Knowledge Base (i.e., Wikipedia, and DBpedia, respectively).

Based on the final objective of knowledge extraction, different features can be selected. It is often useful to identify and extract: Named-Entities, verbs, and their form (i.e., affirmative or negative). Together with context information (e.g., temporal, geographic, etc.), these features populate a Fuzzy Formal Context. In particular, it includes:

- sentences, tweets, or portions of them as *objects*;
- features extracted from the text and features concerning the context as *attributes*.

4.2 CA-Concepts Mining

Features extracted during the previous phase, in the form of Fuzzy Formal Context, are input for the Fuzzy FCA lattice construction. It includes formal concepts grouping together tweets (or sentences) with the same meaning arranged into a hierarchy (see Figure 4.1). The hierarchy is adopted in the subsequent step of *CA-Match-Making* and is useful in decision-supports, marketing strategies, and so on.

4.3 CA-Match-Making

In the third phase of the methodology, a *join* function among lattice concepts extracts a matching set of concepts between the available conceptualizations (i.e., lattices). The join function can be an intersection between concepts [3] or a threshold applied to the results of a *Similarity Measure* [1] (depicted in Figure 4.2). Furthermore, found matches can be weighted by evaluating the *Fuzzy Formal Concept Support* of the involved formal concepts [2].

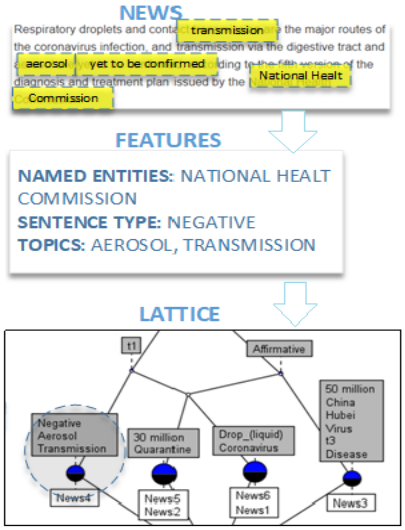

Figure 4.1: CA-Concepts Mining example.

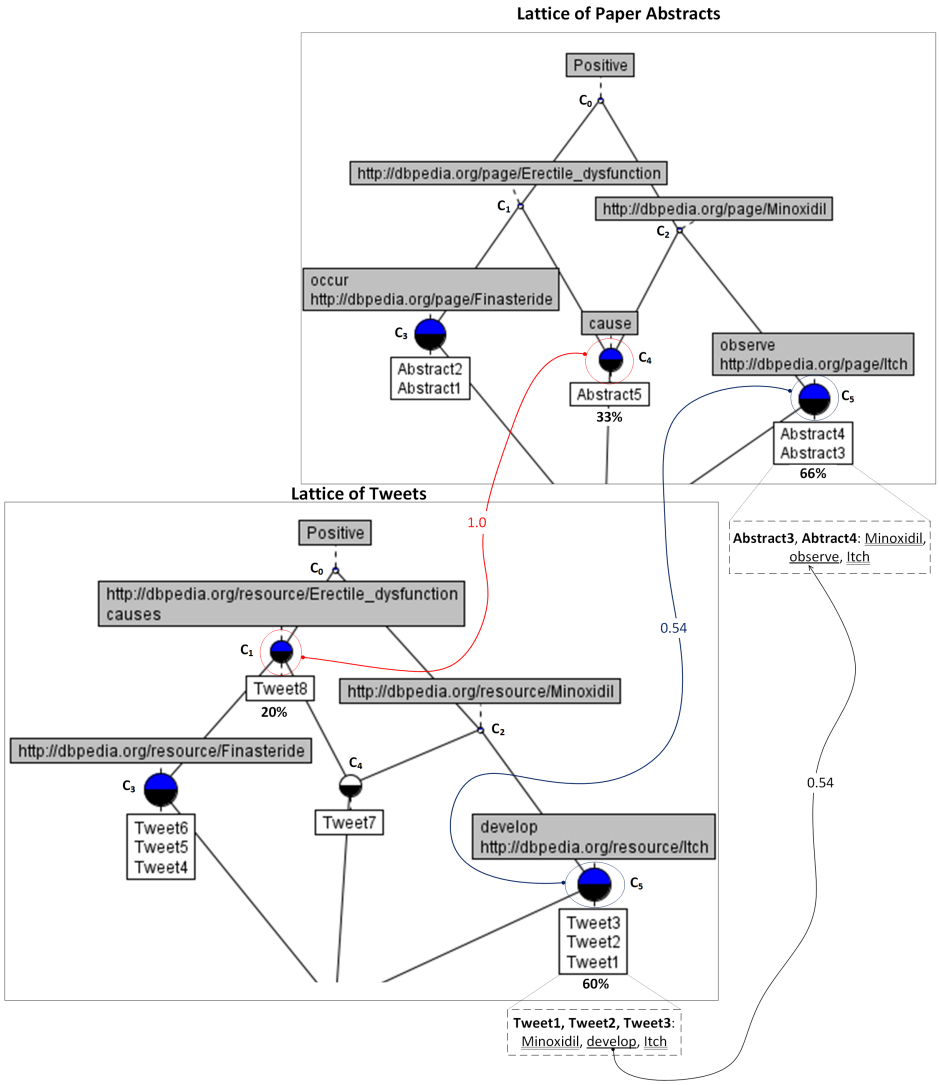

Figure 4.2: CA-Match-Making example.

Chapter 5

Adaptive Anomaly Detection

In dynamic contexts, like Big Data, where continuous changes characterize data, prediction models must consider the concept-drift problem [32]. It identifies the changes in the statistical variables that the model tries to predict and can induce prediction accuracy losses. Models must adapt itself basing on the context along time.

In this chapter, a methodology for adaptive anomaly detection is described. It consists of a noise-tolerant concept-drift and time-aware system, mainly designed to reduce false positives in fraud detection problems. The idea consists of continuous monitoring of data to evaluate the difference (i.e., the error) between the predicted value (by a regression model) and the real observed value. The estimated error is then compared with the previous error "trend". When this trend changes significantly, an alert is produced.

5.1 Methodology

In the first phase of the methodology, the neural network is trained by a set of representative profiles extracted through historical series clustering. Each centroid, as the best delegate for its cluster, contributes to the learning model training (Figure 5.1) and its capacity to recognize different tendencies subsequently.

In the subsequent phase, during the model adoption, the system collects prediction errors and analyze its trend. When the standard deviation of the prediction error (during a specific time interval, e.g., the last week) exceeds the range $(-2\sigma, 2\sigma)$, an anomaly is pointed out (Figure 5.2). Setting the

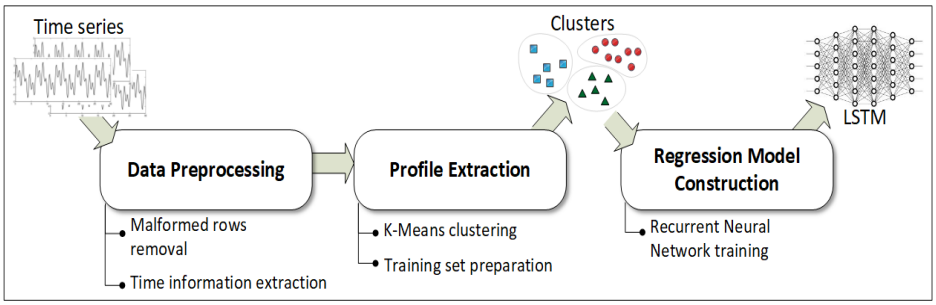

Figure 5.1: Process - First phase

range based on the standard deviation was inspired by the empirical rule of a normal distribution from which the 95% of values lie within roughly in range $(-2\sigma, 2\sigma)$.

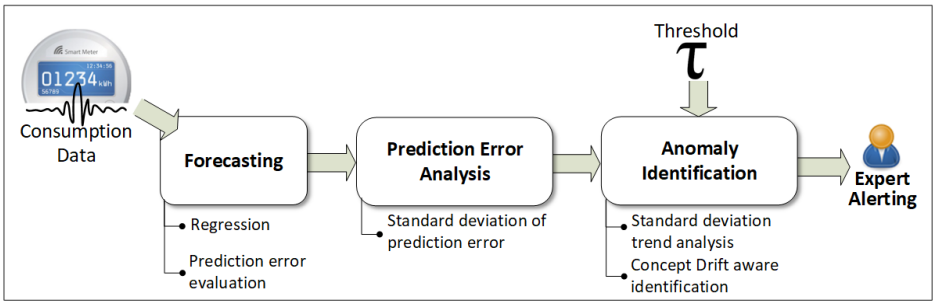

Figure 5.2: Process - Second phase

Summing up, given a threshold τ (i.e., 2σ), at each timestamp t, the proposed anomaly detection algorithm executes the process exposed in Algorithm 1.

More formally, at each instant t, being $S = \{s_1, s_2, \ldots, s_7\}$ the set of standard deviations evaluated on 24 hours for each day of the week that precedes t (i.e., 96 observations at each day), we define the anomaly as follows:

$$\forall s \in S \; s_t > 2 \times s \tag{5.1}$$

where s_t is the standard deviation assessed in the 24 hours that immediately precede t.

As experimented in the real case study described in Section 9.1, in the case of drift-aware, the model is able to adjust itself in a relatively short time interval by producing an error that goes back before the alert creating.

Algorithm 1: Anomaly Detection Algorithm

Begin
Training the network until instant t;
Take the real consumption at $t+1$: CR_{t+1};
Forecast the next value: CP_{t+1};
Evaluate the prediction error: $\delta_{t+1} = (CR_{t+1} - CP_{t+1})^2$;
Evaluate $\sigma(\Delta_k), \Delta_k = \{\delta_{t+1}, \delta_t, \delta_{t-1}, \ldots, \delta_{t-94}\}$;
Check if: $\sigma(\Delta_k) > \tau$;
end

Chapter 6

Data Quality Estimation

In the Artificial Intelligence area, prediction models that can learn "by examples" are heavily influenced by datasets used at the training stage. In this sense, it can be useful to measure the "quality" of these datasets. Data quality can see from different points of view. Among others, the quality of training data relies on consistency, that is, the degree to which the labeler's annotations agree. Consistency prevents random noise by ensuring that labels are correct or incorrect in a consistent manner.

In this perspective, the proposed methodology suggests the adoption of the Group Decision Making consensus as a consistency measure for ranking datasets. By equating similar input (e.g., users, query, etc.) to groups of experts, the methodology evaluates the degree of consensus about output representing preferences (e.g., items, documents for queries, etc.).

6.1 Process

The process starts with the definition of a similarity measure used for the construction of equivalence classes among experts. It essentially depends on features that characterize experts. In general, it can be defined as follows:

$$sim(u_i, u_j) = \sum_F \frac{1}{n} sim(f_{u_i}, f_{u_j}) \tag{6.1}$$

where f_{u_i} is the value of feature f for the user u_i, and $sim(f_{u_i}, f_{u_j})$ is the scaled similarity between f_{u_i} and f_{u_j}, evaluated based on the feature type.

For example, if we consider experts' ages, it can be assessed as:

$$sim(age_{u_i}, age_{u_j}) = 1 - \left| age_{u_i} - age_{u_j} \right| \tag{6.2}$$

When the similarity exceeds an established threshold sim_τ, experts are considered similar and grouped into the same class. The consensus process appli-

Table 6.1: Example of Information Retrieval dataset

Query ID	***Document ID***	***Relevance***
$query_1$	doc_1	5
$query_1$	doc_2	4
$query_1$	doc_n	3
$query_2$	doc_1	1
$query_2$	doc_2	2
...	...	...
$query_2$	doc_n	3
...	...	...
$query_m$	doc_n	rel_k

cation for consistency evaluation of a ranking dataset needs a data alignment: "experts" and "alternatives" must be identified. If we treat a document retrieval problem, queries can be considered experts while possible document rankings as preferences. The consensus process should measure the consistency among queries and suggested documents in the labeled dataset. In other words, more similar document rankings for similar queries, more consistent is considered the dataset (i.e., higher is the consensus degree). Alternatively, if we face a recommendation problem, experts can be equated to users, and rankings of items to suggest are the preferences.

Secondly, data regarding the same experts need to be aggregated: a row of the GDM dataset must represent the expert's preferences for each available alternative. For example, assuming a document retrieval problem, each row regarding a query must give each available document an associated ranking position. In this sense, let considering have an information retrieval dataset as one represented in Table 6.1. It provides a relevance score for each couple query-document. The corresponding GDM dataset is exemplified in Table 6.2: each row summarizes each document's relevances for the specific query.

Table 6.2: Example of GDM dataset

Query	***Documents***			
	doc_1	doc_2	...	doc_n
$query_1$	5	4	...	3
$query_2$	1	2	...	3
$query_m$	...	...	...	rel_k

Finally, a group decision-making process (as described in Section 3.1) can be applied to the GDM dataset formatted in the just explained way for each identified expert's class.

Part III

Case Studies

Chapter 7

Credibility Assessment on Text Streams

This chapter presents two case studies concerning the application of the Context-Aware Knowledge Extraction (CA-KE) methodology (introduced in Chapter 4) to guide the evaluation of the credibility of Twitter streams and their eventual filtering. In particular, the chapter starts with related works about credibility assessment on text streams (Section 7.1); then, Section 7.2 presents experimentation made for discriminating a set of fake news diffused at the beginning of Coronavirus spreading in China in January - February 2020, through social media. Section 7.3 describes the methodology application for drugs' adverse effects discovering and validation on social media.

7.1 Related Works

Research works about credibility can be divided into three main areas of interest: (1) credibility of the content, (2) credibility of the source and, (3) credibility assessment systems.

One of the first research focusing on content credibility was proposed by Castillo et al. [33]. Authors build a classifier that predicts a degree of credibility of tweets related to a set of trending arguments by adopting a labeled dataset of tweets. Two groups evaluate each item: the first considers its newsworthiness while the second evaluates the level of credibility. Similarly, Ikegami et al. [34] discuss a way to realize two classification models for Twit-

ter: topic-based and opinion-based. Authors first identify the topic using LDA and then group various opinions and similar ones using sentiment analysis to understand each one's polarity. Finally, they compute a frequency ratio for each group of opinions to evaluate the related credibility.

Other approaches assess both the credibility of content and source. Gupta et al.'s research work [35] applies regression analysis to understand the set of source-based features relevant to predict the credibility of the information (e.g., characteristics of the author). Mendoza et al. [36] studied Twitter activity after high-impact events and showed how reliable and misleading information spread differently.

Finally, credibility assessment systems deal with the design and development of systems able to assign a label or a rating to a user and tweets. Lorek et al. [37] developed a bot that computes the credibility score of the tweet directly on top of the Twitter UI. Gupta et al. [38] developed a plug-in that calculates a rating related to the user account and evaluates it in real-world situations.

Our approach tries to evaluate the credibility of tweet contents by comparing them with official sources in a time and context-aware manner.

7.2 Fake News

In the last decade, the spread of network-enabled devices allowed the surge of social media. Social media users vary from the oldest to the youngest generations giving everyone a spot in the *knowledge society*. The final user is the new reporter that embraces social media tools to spread novelty and breaking news about specific subjects. On multiple occasions, it has been proved that this collective knowledge is beneficial in case of emergency and damaging events like heartquakes and hurricanes [39]. It is not unusual that users' posts spread crisis news before the official sources. Nevertheless, the contents of the posts cannot always be seen as highly credible. Intentionally or accidentally, users could create information bias that will be magnified by the exposure that the user has on his social media accounts, contaminating other users' perceptions. This domino effect makes it challenging to separate facts from fiction. For this reason, the credibility assessment is becoming more critical than ever, while official sources are shifting towards a reliability-centric approach; social media users cannot easily contribute in this sense.

The proposed method tries to cross-relate heterogeneous text streams

with different reliability levels to assess tweets' degree of credibility. Specifically, the approach (depicted in Figure 7.1) is mainly inspired by the CA-KE methodology already described in Chapter 4 and consists of the following phases:

1. Data Acquisition. Users' posts are collected through the official Twitter Streaming API. For Google News, a spider crawls, collects, and scrapes news articles and the related meta-data.

2. Feature Extraction (i.e., CA-Feature Extraction). Features are extracted by applying semantic and syntactic analysis and collecting context information.

3. Concepts Mining (i.e., CA-Concepts Mining). Through an incremental implementation of the Fuzzy Formal Concept Analysis, two different lattices are created to represent two information sources.

4. Credibility Assessment (i.e., CA-Match-Making). An established similarity measure is used to identify the most similar concept in the lattice built on trustworthy news derived from authentic sources (namely, the *Twin Concept*). The resulting similarity score provides a measure of the reliability of the relations among entities falling in the same concept of Twitter's lattice.

7.2.1 Experimentation

The proposed framework has been validated through tweets and Google News referring to the *Coronavirus* epidemic subject. About 9k tweets were collected. Contemporary, Google News have been harvested and filtered based on a specific white-list of reliable and trustworthy media, and by removing the ones that contain words like "misinformation" and "fake". Finally, considered news articles were about 300.

The validation of the framework calculates the goodness of extracted degrees of credibility by comparing low credible tweets to a set of declared fake news. In particular, by establishing (empirically) a similarity threshold of 0.8, we consider as not credible (i.e., fake) tweets belonging to a concept of Twitter's lattice not having a twin concept in the News' lattice with a similarity value higher than the defined threshold. Then, the set of tweets

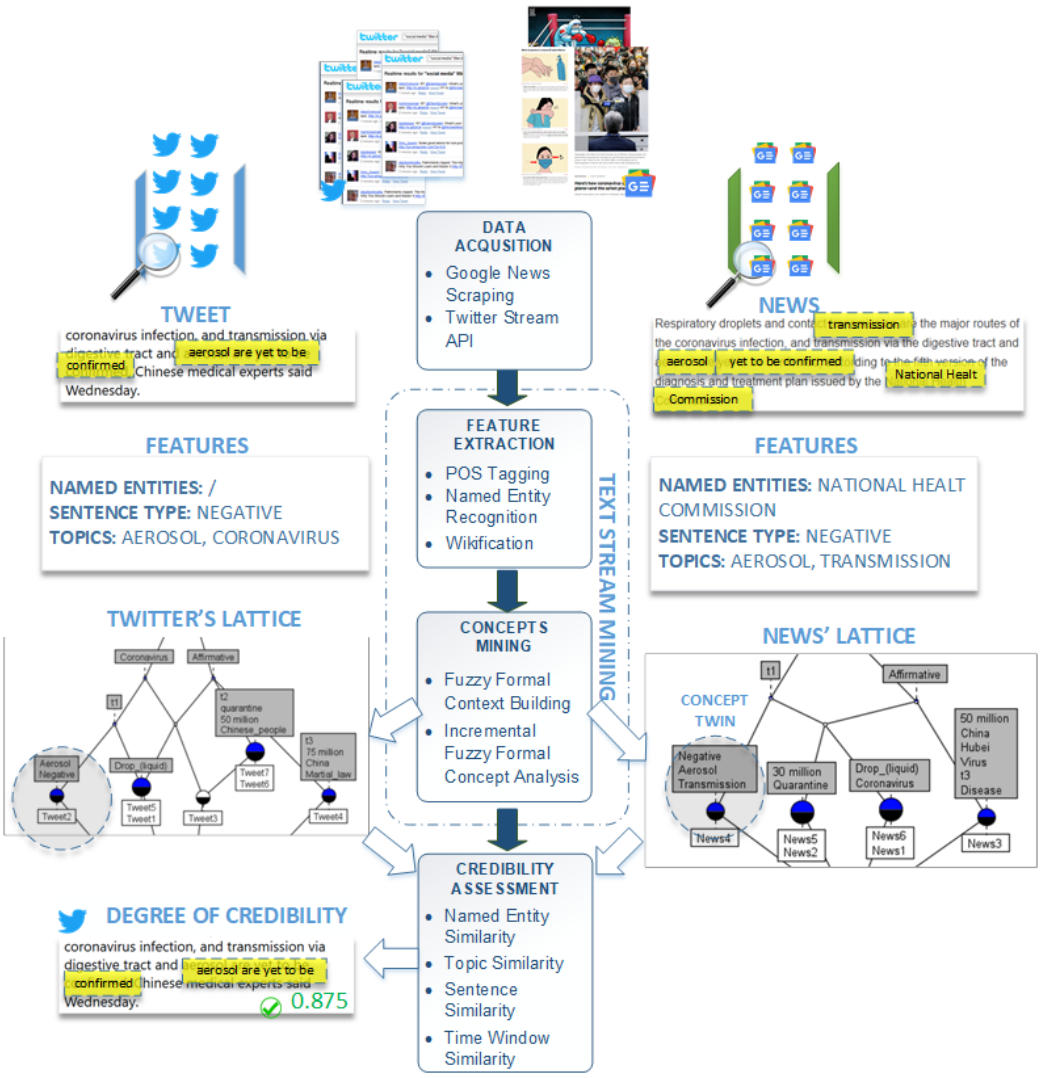

Figure 7.1: Overall Workflow.

recognized as fake from the system is compared to an oracle. In this sense, we selected as reference an official report containing 25 fake tweets[1].

Precision, Recall, and F-Measure (detailed in Appendix A) have been computed to assess the system (see Figure 7.2). Four time-window have

[1]https://www.buzzfeednews.com/article/janelytvynenko/coronavirus-disinformation-spread

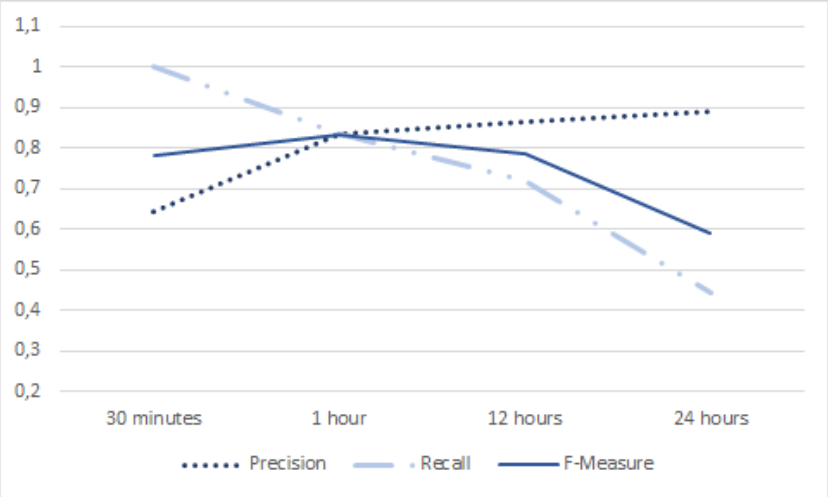

Figure 7.2: Precision, Recall and F-Measure

been selected to understand the respective behaviors and locate the best one (i.e., 30 *minutes*, 1 *hour*, 12 *hours*, and 24 *hours*). What emerges is that the time-window of 1 *hour* achieves higher performance, while other time-windows perform worst due to the following reasons:

- a smaller time-window has too many false positives due to the lag between social and official reporting. By reducing the time-window, we create more distinct concepts, and finding twins become unlikely;
- a larger time-window produces too many false negatives: the same concept groups together more different tweets (that could represent different information) so find a twin become easier.

7.3 Pharmacovigilance

Pharmacovigilance focuses on the monitoring of drug-related adverse events. The post-marketing surveillance systems are almost based on spontaneous reporting systems (SRS); however, it is estimated that more than 90% of adverse drug reactions are under-reported, demonstrating the limits of SRSs' effectiveness [40]. Moreover, the local management of the monitoring process implies a lack of integration among disclosures related to the same drugs but coming from users belonging to different countries. In this perspective, it

emerges the need to carry out real-time and free accessible information to overcome the main complexity deriving from traditional monitoring systems. In recent years, the increasing development of social networks makes them an essential reference for users sharing, among others, a wide variety of personal medical experiences.

The proposed solution employs the CA-KE methodology to represent adverse drug events in Twitter and PubMed and, then, search a similarity among concepts of two lattices. Once the system identifies a correlation between drug and side effect on Twitter with support t_1, assessed on the official site "sideeffects.embl.de", it evaluates the corresponding support t_2 on PubMed. Then, the system keeps track of the difference between t_1 and t_2, called *residual threshold*, to tune itself and understanding, in the absence of official information, if a correlation deserves further in-depth human analysis or can be considered reliable.

7.3.1 Overall Workflow

The methodology (inspired by the CA-KE introduced in Chapter 4) consists of a quantitative and semantic analysis of tweets and medical paper abstracts to evaluate the correlation between drugs and their side effects. The study consists of calculating the frequencies of cited side effects, assessing the correlation (in terms of co-citations) with the drug itself, and then comparing the rates reported on the official site (*sideeffects.embl.de*). More in detail, the process consists of subsequent steps (also depicted in Figure 7.3):

1. Selection of drugs matter of the research. We choose some medications commonly used for hair loss: *Dutasteride*, *Finasteride*, and *Minoxidil*. Selection criteria consider drugs on the market for a long time and are popular enough to have enough information from Twitter.

2. Data collection. The query, consisting of the list of brand names and active ingredients of drugs combined by a logic OR, was submitted, in turn, to Twitter and PubMed.

3. CA-Feature Extraction applied to tweets' descriptions and PubMed's abstract contents.

4. CA-Concepts Mining. Two different concept lattices are extracted to formalize two information contents.

5. CA-Match-Making. The Support of concepts is evaluated for both Twitter and PubMed and subsequently compared. Then, we check if the correlation between the drug and the side effect represented by the concept is recognized on the official site. The distance between two concepts of two lattices helps to find a *residual threshold* that establishes a final level of reliability for Twitter in the area of pharmacovigilance.

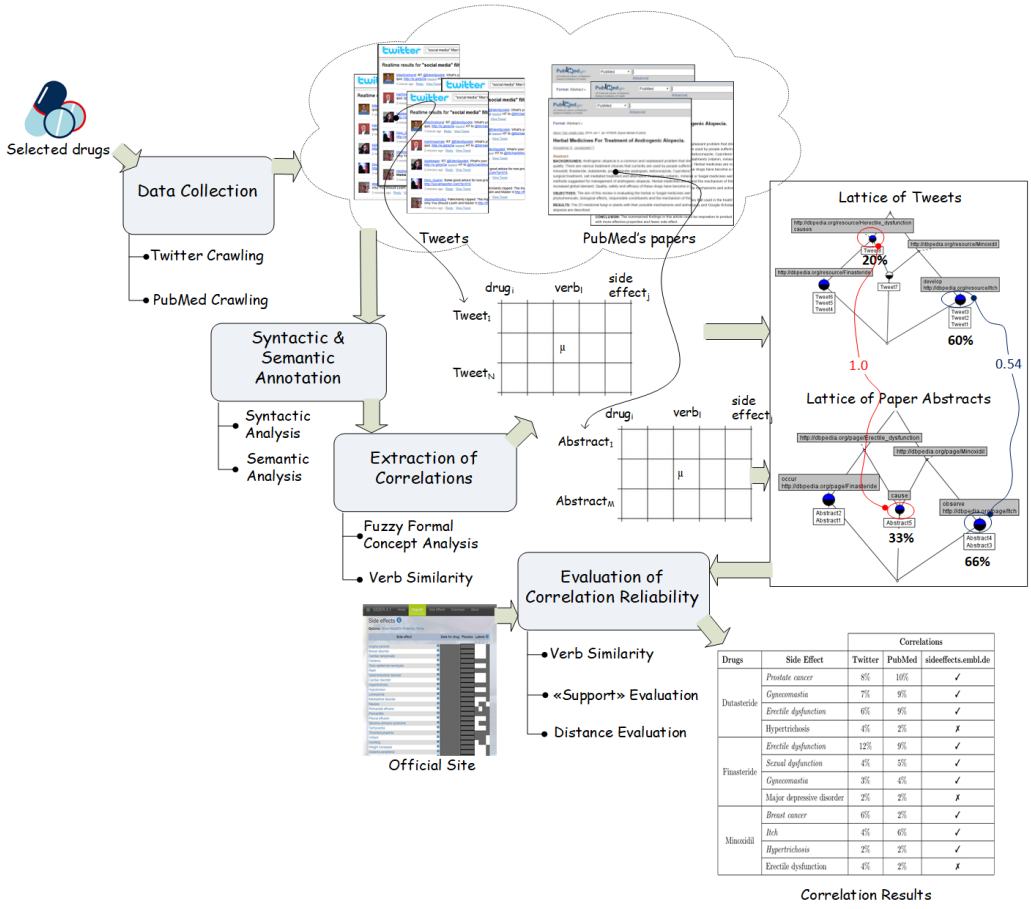

Figure 7.3: Overall Workflow.

7.3.2 Experimentation

The experimentation starts from a total of about 20 thousand tweets citing selected drugs in the period 2009 - 2019. Through Europe PMC services,

the research in PubMed has found nearly 4 thousand papers whose abstracts cite at least one of the selected drugs, from 1974 to 2019.

Data distribution is as summarized in Table 7.1.

Drug	Twitter	PubMed
Dutasteride	3767	776
Finasteride	9012	2401
Minoxidil	7781	1604

Table 7.1: Distribution of data among investigated drugs.

Table 7.2 shows some examples of correlation for each of the studied drugs. Rows written in italic font are confirmed correlations, while those in normal font are probably wrong ones (i.e., not directly confirmed by the official site "sideeffects.embl.de").

		Correlations		
Drug	**Side Effect**	**Twitter**	**PubMed**	**OfficialSite**
Dutasteride	*Prostate cancer*	8%	10%	✓
	Gynecomastia	7%	9%	✓
	Erectile dysfunction	6%	9%	✓
	Hypertrichosis	4%	2%	✗
Finasteride	*Erectile dysfunction*	12%	9%	✓
	Sexual dysfunction	4%	5%	✓
	Gynecomastia	3%	4%	✓
	Major depressive disorder	2%	2%	✗
Minoxidil	*Breast cancer*	6%	2%	✓
	Itch	4%	6%	✓
	Hypertrichosis	2%	2%	✓
	Erectile dysfunction	4%	2%	✗

Table 7.2: Some correlation results.

The experimentation shows that, in most cases, correlations between

drugs and adverse effects extracted from Twitter posts are confirmed by ones declared on the official site. In most of all other cases, the difference between correlation in Twitter and one in Pubmed is shallow.

Experiment results show that online discussions are useful providers of information on drug side effects. By fixing the *residual threshold* of the difference between the correlations percentage in Twitter and PubMed, to ±4%, we observed that 91% of extracted correlations for all studied drugs are considered reliable (i.e., they are included in the official site). In particular, as expressed in Figure 7.4, we compare correlation levels in Twitter and PubMed and evaluate how many correlations are confirmed by the official site. We notice that with a residual correlation of ±5%, the reliability is lesser than 70%. Otherwise, from the *residual threshold* of ±4%, the level of confirmed correlations starts to grow significantly.

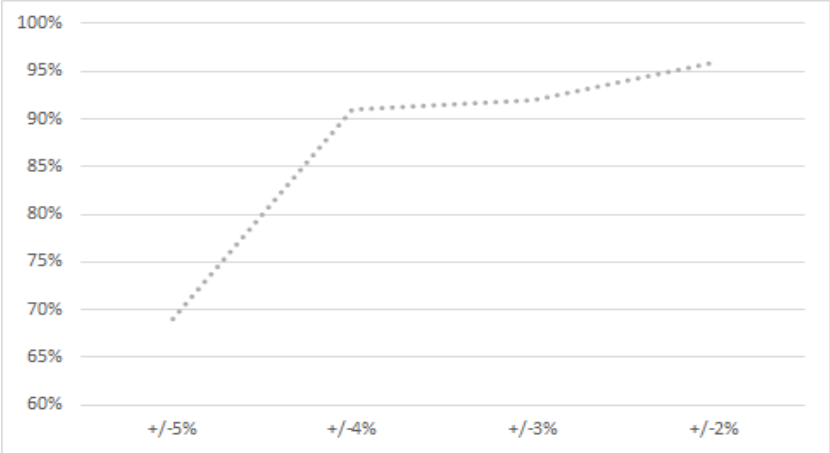

Figure 7.4: Residual plot

7.4 Conclusions

Research works described in this chapter demonstrate that there are good assumptions in cross-relating heterogeneous streams for credibility assessment. However, there are many weak points addressable in the future. Among others, we point out that the most impacting limits of the method are affecting on: the latency in the rise of credibility for a tweet because the most reliable

information maybe not yet published when the users post; the system misses the authorship management for evaluating the credibility of the posted tweet.

In the future, it should be interesting to enrich the set of features used for assessing the credibility, including the authorship, and, eventually, make comparisons between our approach and existing ones. On the other hand, besides its practical utility, in the future, the proposed framework could be integrated with a process that allows us to automatically extract an evolving Knowledge Graph from Unstructured Data Stream, guaranteeing the trustworthiness of links among its nodes.

Chapter 8

Context-Aware Recommender Systems

Recommender systems try to predict users' preferences about an item and are generally adopted for commercial purposes. This chapter proposes two different solutions as recommender systems for social media (i.e., Twitter). Both consider context information about tweets and users for making suggestions. In particular, Section 8.1 introduces some related works in the area of recommendation systems for Twitter; Section 8.2 presents a solution that, by exploiting the CA-KE methodology, infers users' interests and provides an attractive advertisement for them. Section 8.3 adopts a deep learning solution for the learning-to-rank problem (described in Section 2.3.1) that returns an adaptive tweet ranking guided by user preferences.

8.1 Related Works

The growth of accessibility to live streaming information supported the spread of original Twitter recommendation and ranking methods. Most researches define recommendation algorithms for suggesting tweets, hashtags, advertisements, users to follow, and so forth. In [41], a people-to-people recommender system is proposed considering the users' interests, sentiments, and attitudes extracted from the tweets' contents. In [42], the authors present a novel followee ranking scheme using a latent factor model to leverage implicit users' feedback, including tweet content and social relation information for recommending high-quality top-k followees over microblogging systems.

Location-related statuses are used for supporting information delivery in [43]. A hashtags recommendation technique is proposed in [44] by applying topic models and collaborative filtering techniques to help users retrieve contents of interest.

Regarding user's interests, most of the works mentioned above use topic models to project high-dimensional words into low-dimensional latent topics extracted from users' tweets, and words are used to infer users' interests. In [45], the authors present a collaborative ranking model by considering tweet topics, social relation aspects, quality (i.e., using some content-based measures) of the tweet, publisher authority, etc., as features for recommending useful tweets to the users.

Contributions presented in this thesis add to user and tweet related features, local and temporal ones, for generating a context-aware ranking model. The idea is that the user's engagement concerning the tweet content depends on the time slot when the tweet pops out and his/her location.

8.2 Context-aware advertisement system

Advertising is becoming a business on social networks. Billions of people around the world use social media, and fastly, it has become one of the defining technologies of our time. Social platforms like Twitter are one of the primary means of communication and information dissemination and can capture potential customers' interest. Therefore, it is crucial to select suitable advertisements for users in specific times and locations for grabbing their attention profitably.

Given a topic-focused timestamped tweet stream, by exploring the geographic, temporal, and semantic dimensions of tweets, the proposal provides context-aware personalized services (e.g., an advertisement), through the definition and adoption of the Triadic Timed Formal Concept Analysis. Triadic concept analysis (TCA) is an extension of the Formal Concept Analysis (dyadic case) introduced by Wille and Liehman in [46]. It is based on a formalization of the triadic relation connecting objects, attributes, and conditions, under which objects may have specific attributes. In particular, two types of Triadic Timed Formal Concept Analysis are defined. The first one focuses on users' location dimension data for uncovering the social location-focused online communities; the second one focuses on Topics to arrange resources (i.e., tweets) into a hierarchy of time-dependent concepts.

The final process is achieved, taking into account the tweets' locations and semantics to personalized advertising recommendations.

8.2.1 Framework overview

The proposed framework aims to provide custom context-aware (i.e., location and time-based) services to identify targeted advertisements. Specifically, the framework defines the Triadic Timed Formal Concept Analysis methodology to perform geographic, temporal, and conceptual data analysis of social media, starting from the CA-KE methodology introduced in Chapter 4.

Let $U = \{u_1, u_2, \ldots, u_n\}$ being the set of users, $T = \{t_1, \ldots, t_k\}$ the set of range of time (e.g., morning, afternoon, weekend, etc.), $URI_i = \{URI_{i_1}, \ldots, URI_{i_m}\}$ the set of topics URIs extracted from the $i-th$ user's tweet and $M = \{m_1, m_2, ..m_L\}$ the set of locations where users have checked in. The objective is a methodology that takes in input the features T, M, U, and $URIs$ and retrieves groups of potential users located in a specific area interested in a particular event, topic, etc., and recommends a personalized advertisement.

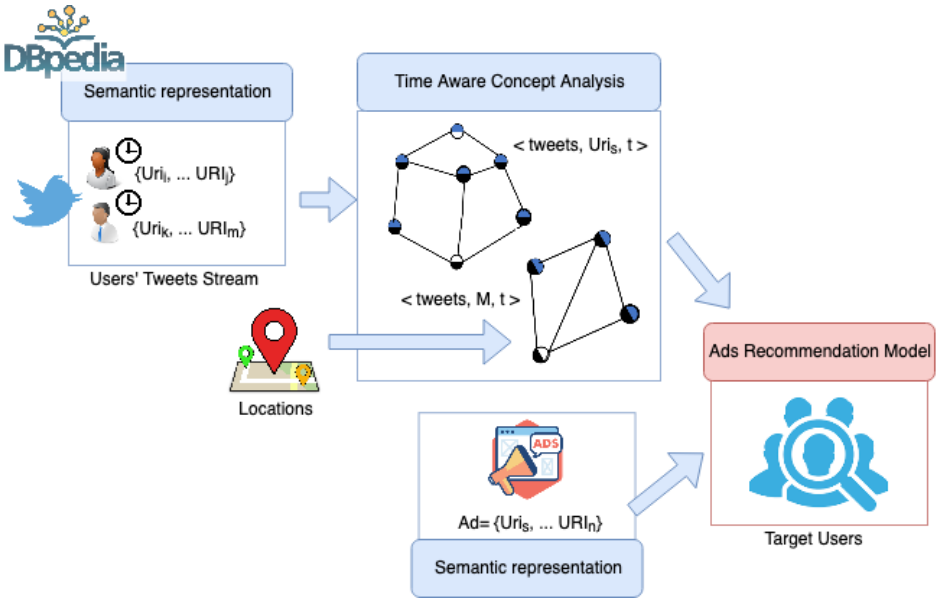

Figure 8.1: Overall Approach

Figure 8.1 shows the overall process of the proposed system consisting of the following macro-phases:

- *Semantic Representation.* Given the input data (i.e., tweet stream or advertisements), this step performs text annotation through DBpedia Spotlight to detect the meaning of the text and perform ad-hoc term weighting (i.e., CA-Feature Extraction).

- *Time-Aware Concept Analysis.* The application of the triadic concept-based approach on two fronts: on the one hand, to discover the evolution of the frequent user locations; on the other, to classify during the timeline, users based on their social content (i.e., the CA-Concepts Mining).

- *Ads Recommendation Model.* The methodology analyzes data and selects the target users interested in a specific advertisement at a given time, based on users, times, semantic representation of tweet stream, locations, and advertisings given in input (i.e., CA-Match-Making).

8.2.2 Experimental Results

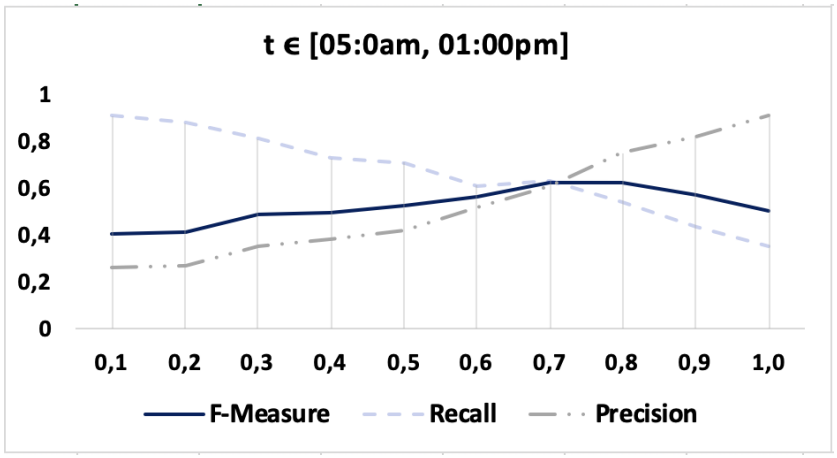

Figure 8.2: F-Measure evaluated by varying the level of threshold $\alpha \in [0, 1]$ in two time slot $[05:00am - 01:00pm]$

Conducted experiments, assessing the proposed approach, adopt real-world Twitter data. Using the Twitter API, we acquired the tweets during April 2019 posted by 31 users in 29 different locations, and we selected 5 tweets as branding ads.

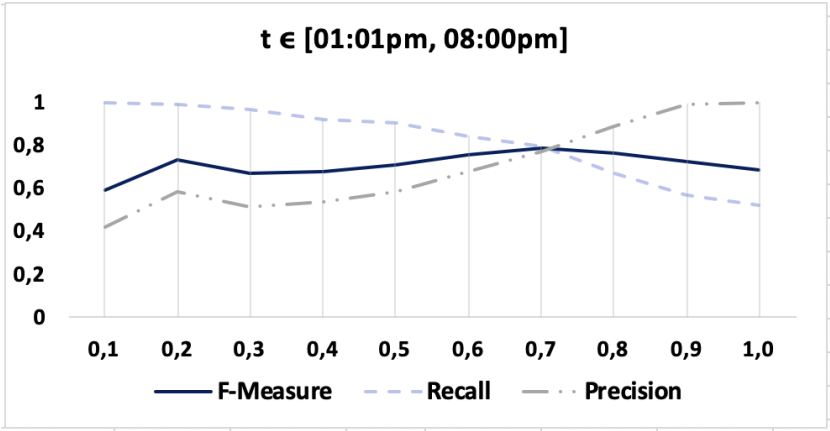

Figure 8.3: F-Measure evaluated by varying the level of threshold $\alpha \in [0, 1]$ in the time slot $[01:01pm - 08:00pm]$

The performances were evaluated in terms of *F-score* (for detail, see Appendix A). As we can see in Figure 8.2 and Figure 8.3, the performances were tested for two time slots [05:00am - 01:00pm] and [01:01pm - 08:00pm] with different thresholds $\alpha \in [0.0, 1.0]$. The framework reveals the best performance with a threshold $\alpha \in [0.65, 0.75]$. Furthermore, as we can see, better results are obtained in the second time slot $[01:01pm - -08:00pm]$ because it has a higher intensity of the posted tweets compared to other time slots. This more flow allows an enriched classification of users that provides more successful matching.

8.3 Context-aware tweets ranking

Generally, tweets about brands, news, and so forth are mostly delivered to the Twitter user in reverse chronological order, choosing among those twitted by the so-called followed users. Recently, Twitter is facing information overload by introducing new filtering features, such as "while you are away", to show only a few tweets summarizing the posted ones and ranking the tweets considering the quality, in addition to timeliness. Trivially enough, we state that the strategy to rank the tweets to maximize the user engagement and, why not, augmenting the tweet and re-tweet rates, is not unique. Several dimensions affect the ranking, such as time, location, semantic, publisher authority, quality, and so on. We point out that the tweet ranking model

should vary according to the user's context, interests, and how those change along the timeline, cyclically, weekly, or at a specific date-time when the user logs in.

In this work, we introduce a deep learning method attempting to re-adapt the tweets' ranking by preferring those more likely interesting for the user. User's interests are extracted by mainly considering previous user re-tweets and replies, and when they occurred.

8.3.1 Approach

The proposed method implements a pairwise preference learning introduced in Section 2.3.1. Let us assume that the user's preference for a tweet is expressed by posting a re-tweet or a reply. In particular, this tweet is preferred to ignored ones in the interval that goes from the tweet posting to the instant the user's re-tweet or reply is posted. Besides, the user's features allow us to personalize the resulting ranking model.

The resulting comparative model is used in a classical sorting algorithm to rank the tweets for an arbitrary user when he/she logs on Twitter. A similar network configuration has also been trained when we exclude some features in order to understand their singular contribution.

8.3.2 Experimentation

The adopted testset consists of tweets timely adjacent to the stream used for training the model. We tested the resulting ranking model for a specific user, evaluating the top-ranked tweets obtained by varying input time slots. The input time slot represents the instant in which the user logs on Twitter.

The adopted ranking measures are MAP [45] that averages on values of precision at n (P@n) [47], and NDCG [48] (for detail, see Appendix A).

Tables 8.1 and 8.2 show the number of users, tweets, and corresponding re-tweets/replies grouped by time slot. Table 8.1 details the training set, and Table 8.2 refers to the test set used for evaluating the system. Four different timeslots are considered: *Morning*, *Afternoon*, *Evening*, and *Night*.

As shown in Figure 8.4, our dataset suffers from data sparsity: the number of users that frequently re-tweet/reply to another tweet is very low. Since this aspect should negatively influence the resulting model, we considered some more contextual features during deep network training to generalize the training data as much as possible, as studied in [45]. For instance, we

Table 8.1: Training-set statistics: tweets collected from 26/01/2017 to 10/02/2017

Training set					
	Morning	*Afternoon*	*Evening*	*Night*	***Total***
Re-tweets/ Replies	1.595	1.619	2.055	1.453	**6.722**
Tweets	15.872	16.444	22.259	15.352	**62.987**
Users	314	316	340	275	**656**

Table 8.2: Test-set statistics: tweets collected from 10/02/2017 to 12/02/2017.

Test set					
	Morning	*Afternoon*	*Evening*	*Night*	***Total***
Re-tweets/ Replies	17	15	28	69	**129**
Tweets	137	76	232	408	**835**
Users	13	12	18	29	**53**

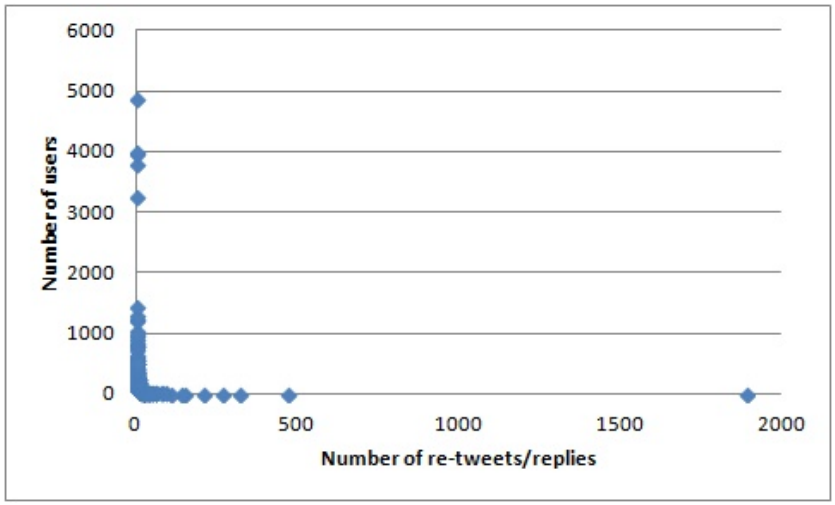

Figure 8.4: Sparsity of dataset in terms of number of re-tweet/reply(s) and number of users.

have included in the training tuples user's information to represent not only individual (re)tweeting users but a class of them generalizing the resulting model. Analogously, we use tweet topics, the number of re-tweets, and so on for representing the tweets.

8.3.2.1 Experimental Results

Experimentation consists of comparing the list of ranked tweets with the set of re-tweeted/replied tweets. The results in terms of $P@n$ are shown in Figure 8.5, while in terms of $MAP@10$ and $NDCG@10$ are in Figure 8.6. Both figures highlight a very good precision in the first and last time slot (i.e., *Morning* and *Night*), while a discrete difficulty in the other two (especially in the *Evening* slot). Such a result is probably due to an undersized training set in terms of contained week-ends. Since the users' behavior can change a lot between holidays and weekdays, the deep model cannot accurately identify the user's interests in all slots. Results of the proposed approach have been

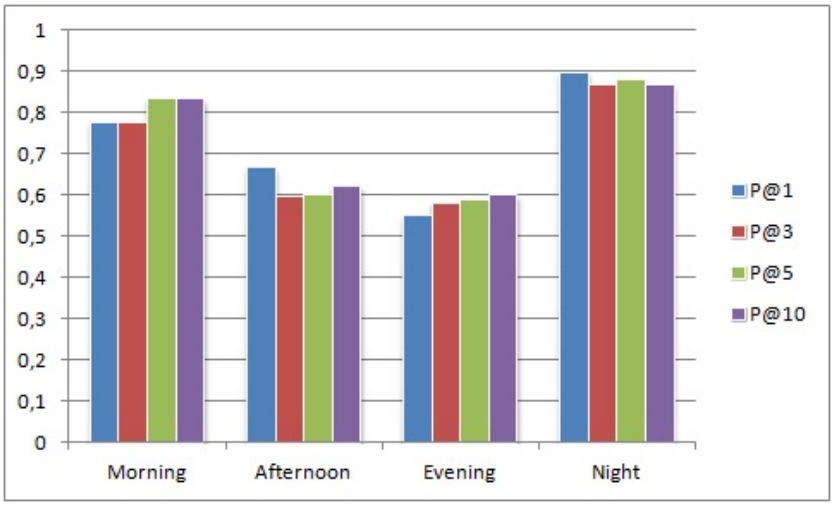

Figure 8.5: Precision of our approach measured in terms of $P@n$ in the different daily slots.

compared with the following ranking criteria:

- Reverse Chronological Order (RCO): Tweets list is ranked in reverse chronological order (from the most recent to the oldest one).

- User's Interests Score (UIS): Tweets are ordered according to a score calculated by multiplying the characteristic vector corresponding to the tweet content and the vector of frequencies of topics representing the level of interest of a user concerning the fixed set of categories.

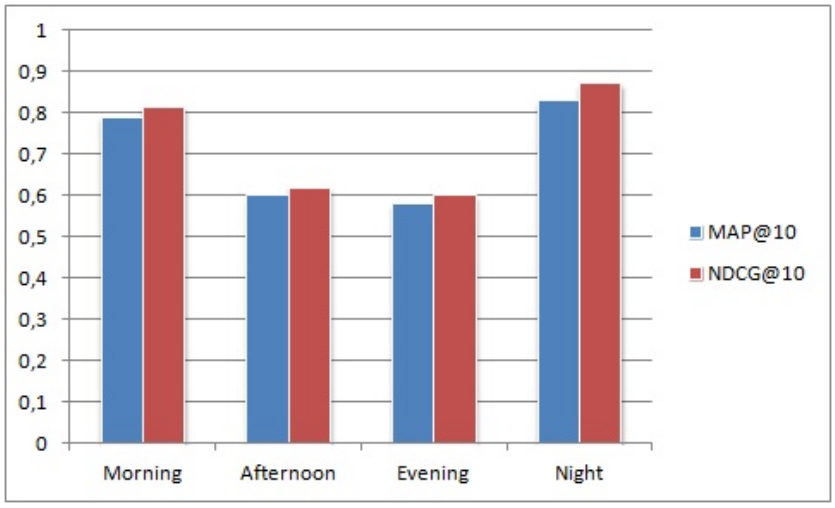

Figure 8.6: Precision of our approach measured in terms of MAP@10 and $NDCG$@10.

With respect to other approaches, the proposed one, named Time-aware Adaptive Tweet Rank (TATR), inherits the advantages of a timed representation of the user's profile.

In order to validate selected features and the role of the time, we also evaluate the performance with some different configurations of TATR. In particular, we add three tests:

- TATR − Publisher's Authority features (TATR - PA): the deep model has been trained with a copy of the original dataset whose are removed features relative to the publisher's authority.
- TATR − Social Relation features (TATR - SR): the deep model has been trained with a copy of the original dataset, with fewer features relative to the social relation between the user and the and the tweet's author.
- TATR − Time-aware features (TATR - T): the deep model has been trained without information about time.

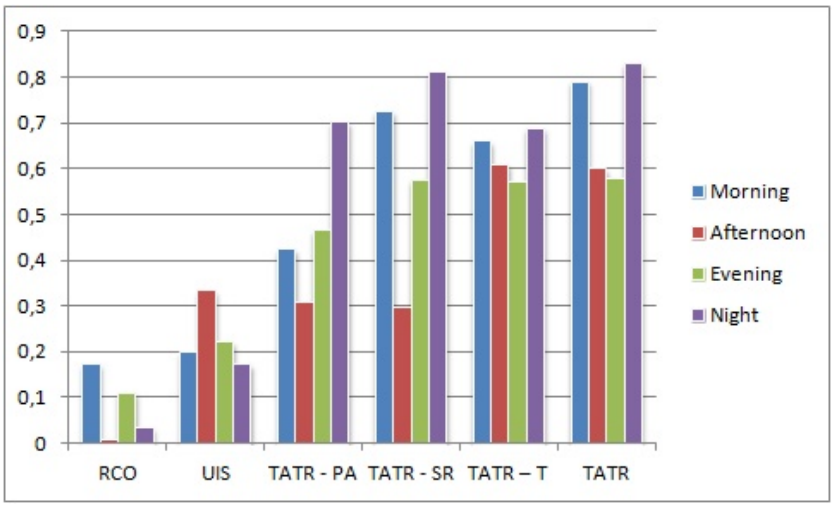

Figure 8.7: Values of MAP@10 of different approaches in the daily slots.

A summary of the performance is shown in Figure 8.7. The experimental results reveal two aspects: the importance of time-awareness as a feature aiming to adapt the ranking model for each user, and the influence of some features on the neural network. In fact, as shown in Figure 8.7, values for RCO and UIS approaches turn out very low, and TATR has better performances also with respect to TATR-PA and TATR-SR in all considered slots, with values of MAP@10 that vary from 0.58 and 0.83. In particular, results show that among tweet evaluated features, "Publisher's Authority" (PA) ones have a better impact on performances. In fact, MAP@10 values substantially decrease when we omit PA features, while when we miss "Social Relation" (SR) features, the performance drop is less critical. It follows that users give more importance to the author's reputation than a social relation existing with them. Regarding time-awareness, instead, results show a tendency of TATR-T to have worse performances with respect to TATR and almost constant values in different slots, highlighting the importance of setting time features in a recommender system of this kind.

8.4 Conclusions

This chapter presents two different approaches as solutions to recommender problems. The first one adopts the methodology related to constructing

multiple lattices to represent information and matching them. The second one is a deep learning solution to the learning-to-rank problem. Both strongly consider context information and reveal good performances.

In the feature, also to compare our results with existing solutions, it could be interesting to merge one of the presented solutions with methodologies expressed in the previous chapter. The idea is to adopt the CA-KE methodology to extract a quality feature about content and add it to the features set to improve the ranking results.

Chapter 9

Decision Support Systems

Decision Making is a cognitive process regarding a problem-solving activity by identifying and analyzing alternatives. Computer and data scientists support this activity by designing *Decision Support Systems* (DSSs) to give a comprehensive vision of the problem and possible solutions. Nowadays, Decision Support Systems are extensively adopted in organizations and businesses. Through the analysis of massive amounts of data, they offer comprehensive information reports to domain experts, guiding them into the best decisions. The success of an intelligent DSS relies on its capability to process large amounts of data and extract useful knowledge from it by also considering the veracity and value of processed data.

This chapter presents some proposals related to the decision-making activities from different points of view: (1) a monitoring and alerting system in the fraud detection area; (2) a support in organization management decisions; (3) the adoption of the decision-making process for the evaluation of data quality. In particular, Section 9.1 describes a drift-aware methodology for anomaly detection in smart grids adopting the method introduced in Chapter 5. Section 9.2 studies trends and causing of Emergency Department overcrowding at an Italian hospital. The study leverages a multiple linear regression analysis described in Section 2.3.3. In Section 9.3, the consensus measure used in Group Decision Making is adopted for evaluating dataset consistency (see Chapter 6). Then, a correlation between the training dataset consistency and the learning model performance is measured through Pearson's correlation coefficient.

9.1 Drift-Aware Methodology for Anomaly Detection in Smart Grid

Electricity thefts cause a loss of about $96 billion every year for utility companies worldwide. Global energy consumption will increase much according to the estimates for 2040, and the impact on our economy and our planet may also be smoothed by reducing energy consumption in buildings.

Moving toward smart cities and smart energy [49], the implementation of devices like Advanced Metering Infrastructure (AMI) allows us to provide value-added services for addressing monitoring consumptions, early alerting about anomalies, and, eventually, improve the discovery of energy thefts. AMI enables two-way communication between utilities and customers as an integrated system of smart meters, communications networks, and data management systems. Smart Meters provide near real-time data about power consumption that can be exploited to evaluate trends and eventually point out anomalies to address the aforementioned aspects. In this sense, combining it with Artificial Intelligence results promising in terms of anomaly detection [50].

The proposed methodology adopts a Long Short Term Memory (LSTM) network to profile and forecast the consumers' behavior based on their recent past consumptions. The continuous monitoring of the consumption prediction errors allows us to distinguish between possible anomalies and changes (drifts) in normal behaviors that correspond to different error motifs.

9.1.1 Related Works

Considering anomaly detection solutions, a methodology for non-technical loss (NTL) detection that exploits smart meter data together with auxiliary databases with contextual information is presented in [51]. Chen et al. [52] propose using fractional-order self-synchronization error-based Fuzzy Petri nets (FPNs) to detect non-technical losses and outage events in micro-distribution systems. Ford et al. [53] demonstrate the effectiveness of an artificial neural network as a technique for modeling consumers' energy utilization and identifying anomalies. In [54], two linear regression-based algorithms aiming to (i) model consumers' energy consumption and (ii) evaluate potential energy theft caused by meter tampering are presented. The work in [55] presents a theft detector based on the combination of meter audit

logs of physical and cyber events with consumption data. Another approach based on the combination of text mining, neural networks, and statistical techniques for the detection of NTLs is presented in [56]. Finally, some other approaches aiming to reduce NTLs, propose lower-level solutions such as a GSM-based Energy Recharge system [57] or a state estimation based method for distribution transformer load estimation [58].

Concerning concept drift, in the literature, its applications regard different tasks: monitoring and control, information management, and diagnostics. Together with varying target applications, concept drift can be applied to multiple data types: sensor streams, time-stamped documents, relational data tables, and so on [59]. Ogundele et al. described the importance of capturing time drifting patterns in user preferences [60]. They presented two leading recommender techniques: factor modeling and item-item neighborhood modeling. The time feature is considered crucial also in POI [61] and service [62] recommendation through the evaluation of a time-aware user similarity; and in mining customer preference in physical stores [63]. In [64], authors adopt clustering and time impact factor matrix to predict user interest drifts through linear regression.

In terms of the combination between two presented issues (i.e., anomaly detection and concept drift), a very recent work [65] illustrates a time series anomaly detection system based on Recurrent Neural Networks (RNNs), which is updated from time to time after an anomaly detection. Since the approach is not thought for the energy domain, it does not fit our goal because of the risk that a fraud seen as an anomaly becomes a network parameter and influences subsequent predictions negatively. So, what emerges is the lack of systems able to recognize anomalies without the influence of concept drift events.

9.1.2 Overall Workflow

The proposed solution consists of a classification technique based on a regression model exploiting sensor data from the smart grid. This model aims to help an expert in identifying energy consumption anomalies by monitoring the differences between the predicted and real consumptions. As already described in Chapter 5, the framework consists of two main stages:

1. Regression model preparation phase. Historical series are clustered, and centroids produce the training set.

2. Anomaly detection phase. With the learning model's adoption, the system collects prediction errors and analyzes its trend to distinguish anomalies from drifts.

9.1.3 Experimentation

The dataset used to validate the proposed framework is *ElectricityLoadDiagrams20112014*[1], publicly available on the UCI repository. It comprises observations of about 370 users from 2011 to 2014, with an observation every 15 minutes. Involved users are clients with different economic activities such as offices, factories, hotels, restaurants, and so on. A further dataset description is available in [66].

A pre-processing of data removes empty rows and extracts needed additional information (e.g., timeslice, previous average consumption, etc.).

After the extraction of 4 clusters of users' consumptions during 2014, the model's training was conducted on the extracted centroids between January and November. The testset included consumptions of the same consumers in December of the same year (300 users). Concerning the anomaly detection, we identify anomalous profiles (i.e., time-series which presents some irregularities) for a total of 14% of the overall testset and create profile changes (i.e., concept drifts) by combining consumptions belonging to multiple clusters, for a total of 13% of the testset. In this case, we are assuming the hypothesis in which a customer changes his tendencies passing from a cluster to another (e.g., a hotel launches services of a health center).

The testset execution carried out Precision and Recall values equal to 78% and 88%, respectively[2].

9.1.4 Discussion

An emerging limitation of the proposed framework is the inability to recognize anomalies in the first week of the system execution. In fact, the anomaly detection algorithm compares the actual prediction error trend with the previous one needing observations about a whole week to understand the consumer profile. The anomaly is highlighted when the actual standard deviation (i.e., regarding the last 24 hours) about the prediction error double

[1] http://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014

[2] See Appendix A for details about adopted measures.

exceeds the standard deviation of all days of the previous week or, analogously, is lower than its half. So, the prerequisite is one week of observation. Furthermore, the practical realization of this type of system is conditioned to the availability of streaming data about different user's profiles, constraints about privacy, and computational capabilities.

9.2 Emergency Department overcrowding monitoring

In Italy, health management is a relevant aspect of a regional administration in terms of costs, reputation, and population health. Although the governments establish different practices during the years, also h24, Emergency Departments (EDs) are often full, falling into overcrowding, a worldwide problem [67]. It may be due to an underestimation of available infrastructures added to many other causes:

- Patients have a wrong perception of their problem, with much anxiety;
- More trust, satisfaction, and, especially, velocity in treatment achievement with respect to primary assistance offer;
- For some people, access to primary assistance is difficult due to a lack of registration to the National Health Service (e.g., foreigners).

EDs' overcrowding represents a waste source in terms of money and offered service [68], [69]. In this study, the objective is to highlight useful insights by analyzing clinical data about accesses to Emergency Departments to potentially support decision-making concerning health management. In particular, we argue that it is possible to reduce the number of inappropriate accesses for dealing with the ED overcrowding problem. We define inappropriate accesses that may also be treated outside the ED (e.g., by general practitioners, GP).

This study analyzes the clinical data about access to EDs for evaluating how many of them may be classified as appropriate or inappropriate and carrying out some descriptive analytics, for instance, about when peaks of demand happen. Inappropriate accesses have a vastly negative impact on the overall health management system, with many concrete negative consequences, such as long waits for patients, high pressure on health personnel,

difficulty in treating patients with greater critical issues, significant increases in costs, etc.

This study aims to identify the most impacting causes of the ED overcrowding problem and how much the inappropriate accesses impact it. We perform descriptive and statistical analytics for identifying the characteristics of patients that access the ED.

9.2.1 Related Works

The topic of Emergency Department overcrowding lies in companies' management support. It is a common problem all around the World. In the Italian context, Riva et al. [70] analyzed the rate of attendance at EDs from the pediatric population in the Lombardy Region. They adopted a statistical analysis focused on the classification of accesses. Analogous research is described in [71]. It analyzes the pediatric emergency room (ER) activities in Italy. What emerges is that only in 56% of the cases, ER Units have a dedicated pediatrician. The research highlights the heterogeneity of the Italian reality, with great possibilities for improvement, especially in southern regions. In [72], authors declare that a percentage, varying from 8.7 to 9.9%, of non-urgent patients, uses EDs to skip gatekeeping, creating overcrowding. In [73], authors studied accesses from the elderly population to a hospital in "Rome" while authors of [74] focused on the situation at "Lecco" and "Monza e Brianza" cities. Cremonesi et al. [75] studied patient flows and overcrowding to define an average per-patient cost according to the severity of his health condition. To the best of our knowledge, there are no similar studies focused on the southern area of Italy, where almost all regions are undergoing a return plan established by the Italian Ministry of Health.

At the international level, Unwin et al. [76] studied non-urgent patients' accesses to ERs in a regional hospital in Tasmania, Australia. The objective is principally understanding their decision-making process. Among motivations, authors found convenience and perceived need. In [77], authors randomly selected 1 million people from all beneficiaries of Taiwan's National Health Insurance claim database in 2010 and analyzed them to estimate the distribution of ED visits among ED users. Research highlights significant ED visit associations with factors such as socio-demographics, health care utilization, and comorbidity. Another similar study but with different results is described in [78]. The authors studied ED demand in Western Australia (WA) from 2007 to 2013. Results highlight an increasing rate in metropolitan

WA, mostly dues to an increasing in people with urgent and complex care needs and not a shift (demand transfer) from primary care. A survey among European neurotrauma centers is described in [79].

9.2.2 Data Analysis

Data analysis consists of two separate dataset analysis. In the first one, a descriptive analysis tries to understand distributions. Then, the regression analysis proves some evidence emerged previously.

Descriptive and statistical analysis show that inappropriate accesses mainly regard the children and the elderly population. The flu period, which can be considered an unexpected emergency, moves many EDs (often uselessly). Moreover, patients from small districts seem to have more trust in hospital facilities than territorial assistance.

These analyses could help health experts and managers in adopting solutions in the short and long term. From our point of view, suggestions consist of improving the availability, in terms of daily and weekly opening of primary assistance on the territory (particularly for the pediatric population). Moreover, enhancing trust in this type of aid could dissuade ED access.

In terms of managing unexpected emergencies, a (partial) solution could be improving patients' awareness through transparent information, tuning dedicated sources of information (e.g., web-sites), and physicians' remote availability.

In a more innovative vision, it should be useful to encourage the diffusion and the adoption of telehealth solutions that could lighten ED workers.

9.3 Consensus Model as Consistency Measure for Dataset in Learning To Rank

Machine Learning and Deep Learning, as well, are strongly characterized by the capability to learn data properties. The performance and the applicability of these methods heavily depend on the dataset used at the training stage. In the last decade, the Big Data explosion had a positive impact on deep learning dissemination. The availability of a huge amount of data enables machine learning and deep learning adoption in several domains. Nevertheless, training deep learning models on massive amounts of data is time-consuming and requires high data processing capabilities. Additionally,

many questions related to the optimization of hyperparameters to get the best performance arise. In this sense, understanding the input data may be useful at the training stage for perceiving: (1) When a learning model should be considered outdated; (2) Training dataset filtering needs; (3) The existence of a method to assess dataset quality and its suitability to train a model.

This research work stresses the urgent need to define methods for determining input data quality. The quality of training data relies on consistency, which is the degree to which the labeler's annotations (human or machine) agree with one another. In general, the dataset's' consistency for binary or multi-class classifiers is measured through a consensus value that may be trivially calculated by dividing the sum of agreeing on labels by the total number of labels per asset. This study focuses on evaluating training data consistency for ranking problems that are mainly solved using learning to rank algorithms (LTR) [16]. Among others, learning to rank algorithms are successfully applied to domains such as recommender systems [80][4][81], and Information Retrieval [82][83][84]. These algorithms assume that similar labelers are agreed, not only about the highest-ranked alternative but also on a ranked list of possible options. The idea is measuring the consistency among similar inputs (i.e., input features) with respect to their outputs (i.e., labels). By grouping data based on input characteristics, we evaluate how coherent their outputs are. Then, we try to validate the hypothesis of a statistical relationship between the dataset consensus level and the performance of learning to rank model implemented through a Deep Neural Network (DNN) [85].

9.3.1 Related Works

In the area of Machine Learning, training data quality assumes an important role. Since the learning process is usually expensive in terms of time and resources, starting from suitable data is mandatory. Data quality can see from different points of view. In general, the quality of data depends on the quality of contained information [86]. In a training set, the quality depends on errors [87], noise [88], or inconsistency [89] in labeled data. In this perspective, authors in [90] try to give a comprehensive definition of quality by considering multiple dimensions, characteristics, and indexes. Analogously, Merino et al. [91] present a quality assessment model based on Contextual Adequacy, Operational Adequacy, and Temporal Adequacy. Ardagna et al. [92] formalize a data quality assessment service and evaluate its incidence on the loss

of accuracy after its application on only a portion of the dataset.

With the evaluation of data quality level, domain-specific techniques must be employed to deal with it. A data cleaning system based on data association and repairing is described in [93]. Krishnan et al. [94] formalize a machine learning approach for dirty data detection. A noise reduction corresponding to specific types of inconsistencies is described in [95].

The research community is spending a big effort discovering the best solution to detect and repair inconsistencies. Nevertheless, it remains an open issue [96].

9.3.2 Experimentation

A deep neural network aiming to resolve the learning to rank is trained and tested. The system collects performance measures (i.e., prediction accuracy and Kendall's τ coefficient) to compare with consensus degrees through Pearson's correlation coefficient evaluation.

The experimentation regards two different macro-tests: the first one on a randomly generated dataset; the second one on a real dataset, described in the following sub-sections.

9.3.2.1 Correlation evaluation on the random generated dataset

Tests on the randomly generated dataset started from establishing two similarity thresholds: 0.7 and 0.6, resulting in 34 and 25 equivalence classes and 26 and 37 users per class, respectively[3].

Table 9.1 reports average accuracy and Kendall's τ coefficient corresponding to the three generated consensus degrees (i.e., high, medium, and low) obtained through a 50 epochs training. Table 9.2 shows the resulting Pearson's correlation coefficients for each test, while Figure 9.1 shows trends of consensus degrees and DNN performances on the same data. Both highlight a significant statistical relationship between Accuracy and Kendall's τ coefficient achieved by the DNN and the consensus degree.

[3]Let us note that by further augmenting the threshold (e.g., to 0.8) we obtained classes with a too low cardinality that make DNN training difficult.

Test	*Average Consensus*	*Average Accuracy*	*Average Kendall's τ*
1	0.92	0.85	0.8
	0.84	0.75	0.65
	0.75	0.62	0.42
2	0.95	0.85	0.8
	0.87	0.75	0.65
	0.78	0.63	0.43

Table 9.1: Performances at 50 epochs' training

Table 9.2: Pearson's correlation coefficients

Test	**Consensus degree Vs. Accuracy**	**Consensus degree Vs. Kendall's τ**
1	0.86	0.84
2	0.89	0.89

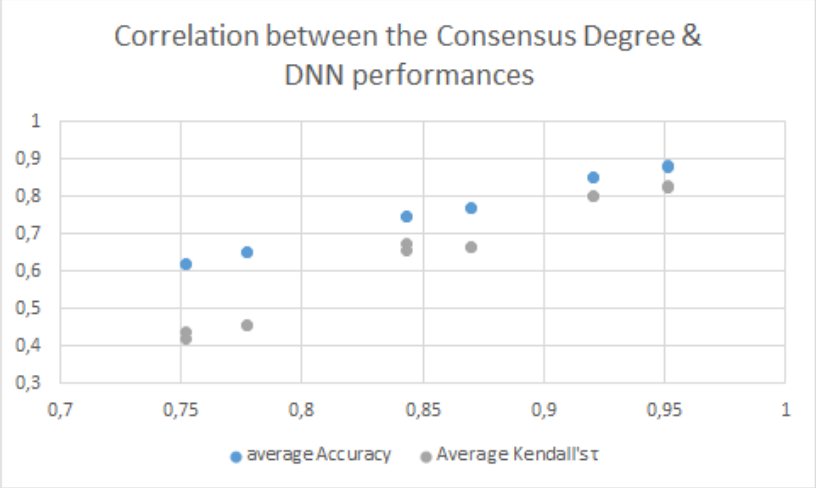

Figure 9.1: Correspondence among evolution of performance coefficients.

9.3.2.2 Correlation evaluation on MovieLens dataset

To show a real application of the methodology, in this section, we add and explain two further experiments made on portions of the dataset MovieLens.

As expressed in Table 9.3, we select two sets of data with high and low average consensus among classes, respectively, and group users by a sim_τ equal to 0.7.

Table 9.3: Tests on real datasets.

Test	*Average Consensus*	*Average Accuracy*	*Average Kendall's* τ
1	0.88	0.88	0.77
2	0.67	0.35	0.11

By applying the overall process and training the model by entire datasets, we obtain performances in line with the previous tests, with a significant dependency between consensus degree and DNN performances. Trends of evaluated measures are also visible in Figure 9.2.

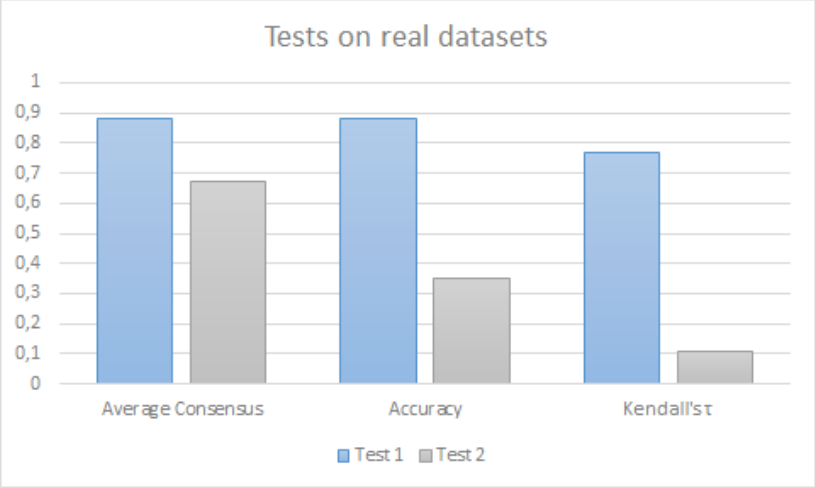

Figure 9.2: Performance results of methodology application on real datasets.

9.4 Conclusions

Decision Support Systems help experts in their business or organizational activities. This chapter highlights two main aspects: (1) both machine learning algorithms and regression statistics are valid solutions to give experts crucial information driving their decision-making activities in terms of management, operations, etc. (2) Decision-making could also be crucial in other contexts.

Besides their effectiveness, ML models are often costly in terms of resource and time needs. The availability of a method able to foresee the achievable performance of a learning model through the training dataset characteristics can reveal valid information.

For each of the proposed solutions, there are some possible improvements. The anomaly detection could enhance in terms of delay between the anomaly and its detection; we could define a dashboard, including predictions useful in guiding resource planning and distribution in terms of ED management. Finally, the dataset consistency could be measured through a multi-criteria group decision-making process; and it could be interesting to establish criteria to decide when appropriate to include new items for re-training a machine learning model.

Part IV

Conclusions

Chapter 10

Conclusion and Future Work

This chapter closes the thesis work by providing a short summary and defining possible future challenges.

10.1 Summary

This thesis work stresses the concept of data quality in knowledge extraction. By defining different solutions for Fraud Detection, Recommender Systems, Decision Support Systems, and so on, it highlights the importance of characteristics as value and veracity of data. The objective is giving to companies or public governments adaptive solutions that could reflect data and its evolutions. In this sense, three main methodologies are presented:

1. *Knowledge Extraction from Text Stream.* It defines a conceptualization process for text stream and a correlation rule between different lattices concepts deriving from various information sources.

2. *Adaptive Anomaly Detection.* It consists of a noise-tolerant concept-drift and time-aware system, mainly designed to reduce false positives in fraud detection problems.

3. *Data Quality Estimation.* It aims to identify a level of consistency for a dataset to understand its adoption for learning purposes.

The above methodologies are applied to three main research areas:

1. *Credibility Assessment on Text Streams*: evaluation of information trustworthiness by cross-relating different information sources. The

main objective is to understand the reliability of the information contained in social media.

2. *Context-Aware Recommender Systems*: recommender systems leveraging context information such as geographical and temporal are demonstrated to be more accurate.

3. *Decision Support Systems*: systems aim to support experts during their activities and give researchers a way to address machine learning tasks in terms of data suitability.

10.2 Future Work

In terms of extensions of the proposed methodologies, there are a lot of feasible improvements:

- The credibility assessment could be integrated into a more complex system that automatically extracts an evolving Knowledge Graph from Unstructured Data Stream, guaranteeing the trustworthiness of links among its nodes.

- The recommender system could adopt the credibility assessment methodology to improve ranking results by taking into account the reliability of a single item.

- The data quality estimation could be evaluated through a multi-criteria group decision-making process and adopted to make decisions about the inclusion of new items for re-training a machine learning model.

Appendix A

Performance Measures

This additional chapter defines performance measures adopted in experimentations described in Part III of this thesis.

A.1 Root Mean Squared Error (RMSE)

The *Root Mean Squared Error (RMSE)* metrics, a widely used statistical method to calculate the difference between real and forecasted values, is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (y'_t - y_t)^2}{n}} \tag{A.1}$$

where y'_t is the prediction value, y_t is the real value in the testset, and n is the number of instances.

A.2 Precision and Recall

Precision and *Recall* are defined as following:

$$Precision = \frac{\left|A_{gold} \bigcap A_{ret}\right|}{|A_{ret}|} \tag{A.2}$$

$$Recall = \frac{\left|A_{gold} \bigcap A_{ret}\right|}{\left|A_{gold}\right|} \tag{A.3}$$

where $A_{gold} = \{a_{gold_1}, a_{gold_2}, \ldots, a_{gold_m}\}$ and $A_{ret} = \{a_1, a_2, \ldots, a_n\}$ are, respectively, the gold set, and the result set retrieved by the system.

F-Score is defined as follows:

$$F - Score = 2 \cdot \frac{|Precision \cdot Recall|}{|Precision + Recall|} \tag{A.4}$$

The F-score can provide a more realistic measure of a test's performance using Precision and Recall.

A.3 Mean Average Precision

Given a user u, let us define $P@n$ that measures the relevance of the top n results of the ranking list:

$$P@n = \frac{relevant\ tweets\ in\ top\ n\ results}{n} \tag{A.5}$$

Given a user u, the average of the precision $P@n$ measured for all re-tweets/replies is Average Precision (AP_u) defined as follows:

$$AP_u = \frac{\sum_{n=1}^{N} P@n \cdot rel(n)}{N_u} \tag{A.6}$$

where N and N_u are, respectively, the number of tweets and re-tweets/replies for the user u, $rel(n)$ is a function that has value 1 if the $n-$th tweet in the ordered list has been re-tweeted/replied by u, 0 otherwise. Thus, AP_u averages the values of $P@n$ over the positions n of the relevant tweets. Finally, the MAP value is computed as the mean of AP_u over the set of all users.

A.4 Normalized Discount Cumulative Gain

The Normalized Discount Cumulative Gain (NDCG) considers the relevance of returned tweets in the resulting list and is calculated, for each user, as follows:

$$NDCG_u@n = Z_n \sum_{i=1}^{n} \frac{rel(i)}{\log_2(i+1)} \tag{A.7}$$

where n is the evaluated position, $rel(i)$ is analogous of $rel(n)$ in the previous measure, and Z_n is a normalization factor given by an ideal ordering (i.e., all $rel(i)$ with value 1). In this thesis, we refer to $NDCG@n$ as the mean of $NDCG_u@n$ over the set of all users.

Bibliography

[1] C. De Maio, G. Fenza, M. Gallo, V. Loia, and A. Volpe, "Cross-relating heterogeneous text streams for credibility assessment," in *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*. IEEE, 2020, pp. 1–8.

[2] M. De Rosa, G. Fenza, A. Gallo, M. Gallo, and V. Loia, "Pharmacovigilance in the era of social media: Discovering adverse drug events cross-relating twitter and pubmed," *Future Generation Computer Systems*, 2020.

[3] C. De Maio, M. Gallo, F. Hao, and E. Yang, "Who and where: context-aware advertisement recommendation on twitter," *Soft Computing*, pp. 1–9, 2020.

[4] C. De Maio, G. Fenza, M. Gallo, V. Loia, and M. Parente, "Time-aware adaptive tweets ranking through deep learning," *Future Generation Computer Systems*, vol. 93, pp. 924–932, 2019.

[5] G. Fenza, M. Gallo, and V. Loia, "Drift-aware methodology for anomaly detection in smart grid," *IEEE Access*, vol. 7, pp. 9645–9657, 2019.

[6] V. Andretta, G. Fenza, M. Gallo, and V. Loia, "Etiology of emergency department overcrowding: descriptive analytics of inappropriate accesses at salerno hospital in italy," *Journal of Data, Information and Management*, vol. 2, no. 3, pp. 111–120, 2020.

[7] G. Fenza, M. Gallo, V. Loia, F. Orciuoli, and E. Herrera-Viedma, "Data set quality in machine learning: Consistency measure based on group decision making," *Applied Soft Computing*, p. 107366, 2021.

[8] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[9] D. Milne and I. H. Witten, "An open-source toolkit for mining wikipedia," *Artificial Intelligence*, vol. 194, pp. 222–239, 2013.

[10] R. Mihalcea and A. Csomai, "Wikify! linking documents to encyclopedic knowledge," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 233–242.

[11] J. Flisar and V. Podgorelec, "Document enrichment using dbpedia ontology for short text classification," in *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*. ACM, 2018, p. 8.

[12] S. Hellmann, C. Stadler, J. Lehmann, and S. Auer, "Dbpedia live extraction," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2009, pp. 1209–1223.

[13] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "Dbpedia spotlight: shedding light on the web of documents," in *Proceedings of the 7th international conference on semantic systems*, 2011, pp. 1–8.

[14] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "Dbpedia-a crystallization point for the web of data," *Journal of web semantics*, vol. 7, no. 3, pp. 154–165, 2009.

[15] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[16] T.-Y. Liu, *Learning to rank for information retrieval*. Springer Science & Business Media, 2011.

[17] K. Crammer and Y. Singer, "Pranking with ranking," in *Advances in neural information processing systems*, 2002, pp. 641–647.

[18] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon, "Adapting ranking svm to document retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 186–193.

[19] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 89–96.

[20] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 129–136.

[21] J. Xu, T.-Y. Liu, M. Lu, H. Li, and W.-Y. Ma, "Directly optimizing evaluation measures in learning to rank," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 107–114.

[22] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Presses universitaires de Louvain, 2015, p. 89.

[23] W. De Mulder, S. Bethard, and M.-F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling," *Computer Speech & Language*, vol. 30, no. 1, pp. 61–98, 2015.

[24] M. B. Kamal, G. J. Mendis, and J. Wei, "Intelligent soft computing-based security control for energy management architecture of hybrid emergency power system for more-electric aircrafts," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 4, pp. 806–816, 2018.

[25] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[26] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in k-means clustering," *International Journal of Advance Research inComputer Science and Management Studies*, vol. 1, no. 6, pp. 90–95, 2013.

[27] C. De Maio, G. Fenza, V. Loia, and F. Orciuoli, "Distributed online temporal fuzzy concept analysis for stream processing in smart cities," *Journal of Parallel and Distributed Computing*, vol. 110, pp. 31–41, 2017.

[28] B. Ganter and R. Wille, *Formal concept analysis: mathematical foundations*. Springer Science & Business Media, 2012.

[29] C. De Maio, G. Fenza, V. Loia, F. Orciuoli, and E. Herrera-Viedma, "A framework for context-aware heterogeneous group decision making in business processes," *Knowledge-Based Systems*, vol. 102, pp. 39–50, 2016.

[30] E. Herrera-Viedma, F. Herrera, and F. Chiclana, "A consensus model for multiperson decision making with different preference structures," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 32, no. 3, pp. 394–402, 2002.

[31] F. J. Cabrerizo, J. A. Morente-Molinera, I. J. Pérez, J. López-Gijón, and E. Herrera-Viedma, "A decision support system to develop a quality management in academic digital libraries," *Information Sciences*, vol. 323, pp. 48–58, 2015.

[32] J. C. Schlimmer and R. H. Granger, "Incremental learning from noisy data," *Machine learning*, vol. 1, no. 3, pp. 317–354, 1986.

[33] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684.

[34] Y. Ikegami, K. Kawai, Y. Namihira, and S. Tsuruta, "Topic and opinion classification based information credibility analysis on twitter," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2013, pp. 4676–4681.

[35] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," in *Proceedings of the 1st workshop on privacy and security in online social media*, 2012, pp. 2–8.

[36] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we rt?" in *Proceedings of the first workshop on social media analytics*, 2010, pp. 71–79.

[37] K. Lorek, J. Suehiro-Wiciński, M. Jankowski-Lorek, and A. Gupta, "Automated credibility assessment on twitter," *Computer Science*, vol. 16, no. 2), pp. 157–168, 2015.

[38] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "Tweetcred: Real-time credibility assessment of content on twitter," in *International Conference on Social Informatics*. Springer, 2014, pp. 228–243.

[39] M. Martinez-Rojas, M. del Carmen Pardo-Ferreira, and J. C. Rubio-Romero, "Twitter as a tool for the management and analysis of emergency situations: A systematic literature review," *International Journal of Information Management*, vol. 43, pp. 196–208, 2018.

[40] L. Hazell and S. A. Shakir, "Under-reporting of adverse drug reactions," *Drug safety*, vol. 29, no. 5, pp. 385–396, 2006.

[41] D. F. Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti, "Temporal people-to-people recommendation on social networks with sentiment-based matrix factorization," *Future Generation Computer Systems*, 2017.

[42] H. Chen, X. Cui, and H. Jin, "Top-k followee recommendation over microblogging systems by exploiting diverse information sources," *Future Generation Computer Systems*, vol. 55, pp. 534–543, 2016.

[43] D. Namiot and M. Sneps-Sneppe, "Social streams based on network proximity," *International Journal of Space-Based and Situated Computing*, vol. 3, no. 4, pp. 234–242, 2013.

[44] F. Zhao, Y. Zhu, H. Jin, and L. T. Yang, "A personalized hashtag recommendation approach using lda-based topic model in microblog environment," *Future Generation Computer Systems*, vol. 65, pp. 196–206, 2016.

[45] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu, "Collaborative personalized tweet recommendation," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012, pp. 661–670.

[46] R. Wille, "The basic theorem of triadic concept analysis," *Order*, vol. 12, no. 2, pp. 149–158, 1995.

[47] N. Craswell, "Precision at n," in *Encyclopedia of database systems*. Springer, 2009, pp. 2127–2128.

[48] K. Järvelin and J. Kekäläinen, "Ir evaluation methods for retrieving highly relevant documents," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2000, pp. 41–48.

[49] N. Mohamed, J. Al-Jaroodi, I. Jawhar, S. Lazarova-Molnar, and S. Mahmoud, "Smartcityware: a service-oriented middleware for cloud and fog enabled smart city services," *Ieee Access*, vol. 5, pp. 17 576–17 588, 2017.

[50] M. Zanetti, E. Jamhour, M. Pellenz, M. Penna, V. Zambenedetti, and I. Chueiri, "A tunable fraud detection system for advanced metering infrastructure using short-lived patterns," *IEEE Transactions on Smart Grid*, 2017.

[51] M.-M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Transactions on Smart Grid*, 2018.

[52] S.-J. Chen, T.-S. Zhan, C.-H. Huang, J.-L. Chen, and C.-H. Lin, "Nontechnical loss and outage detection using fractional-order self-synchronization error-based fuzzy petri nets in micro-distribution systems," *IEEE Transactions on smart grid*, vol. 6, no. 1, pp. 411–420, 2015.

[53] V. Ford, A. Siraj, and W. Eberle, "Smart grid energy fraud detection using artificial neural networks," in *Computational Intelligence Applications in Smart Grid (CIASG), 2014 IEEE Symposium on*. IEEE, 2014, pp. 1–6.

[54] S.-C. Yip, K. Wong, W.-P. Hew, M.-T. Gan, R. C.-W. Phan, and S.-W. Tan, "Detection of energy theft and defective smart meters in smart grids using linear regression," *International Journal of Electrical Power & Energy Systems*, vol. 91, pp. 230–240, 2017.

[55] S. McLaughlin, B. Holbert, A. Fawaz, R. Berthier, and S. Zonouz, "A multi-sensor energy theft detection framework for advanced metering infrastructures," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 7, pp. 1319–1330, 2013.

[56] J. I. Guerrero, C. León, I. Monedero, F. Biscarri, and J. Biscarri, "Improving knowledge-based systems with statistical techniques, text mining, and neural networks for non-technical loss detection," *Knowledge-Based Systems*, vol. 71, pp. 376–388, 2014.

[57] B. Omijeh and G. Ighalo, "Modeling of gsm-based energy recharge scheme for prepaid meter," *IOSR Journal of Electrical and Electronics Engineering*, vol. 4, no. 1, pp. 46–53, 2013.

[58] S.-C. Huang, Y.-L. Lo, and C.-N. Lu, "Non-technical loss detection using state estimation and analysis of variance," *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 2959–2966, 2013.

[59] I. Žliobaitė, M. Pechenizkiy, and J. Gama, "An overview of concept drift applications," in *Big data analysis: new algorithms for a new society*. Springer, 2016, pp. 91–114.

[60] T. J. Ogundele, C.-Y. Chow, and J.-D. Zhang, "Socast*: Personalized event recommendations for event-based social networks: A multi-criteria decision making approach," *IEEE Access*, vol. 6, pp. 27 579–27 592, 2018.

[61] H.-T. Zheng, Y. Zhou, N. Liang, X. Xiao, A. K. Sangaiah, and C. Zhao, "Exploiting user mobility for time-aware poi recommendation in social networks," *IEEE Access*, 2017.

[62] L. Qi, X. Xu, W. Dou, J. Yu, Z. Zhou, and X. Zhang, "Time-aware ioe service recommendation on sparse data," *Mobile Information Systems*, vol. 2016, 2016.

[63] Y. Chen, Z. Zheng, S. Chen, L. Sun, and D. Chen, "Mining customer preference in physical stores from interaction behavior," *IEEE Access*, vol. 5, pp. 17 436–17 449, 2017.

[64] B. Sun and L. Dong, "Dynamic model adaptive to user interest drift based on cluster and nearest neighbors," *IEEE Access*, vol. 5, pp. 1682–1691, 2017.

[65] S. Saurav, P. Malhotra, V. TV, N. Gugulothu, L. Vig, P. Agarwal, and G. Shroff, "Online anomaly detection with concept drift adaptation using recurrent neural networks," in *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. ACM, 2018, pp. 78–87.

[66] F. Rodrigues and A. Trindade, "Load forecasting through functional clustering and ensemble learning," *Knowledge and Information Systems*, pp. 1–16, 2018.

[67] C. Morley, M. Unwin, G. M. Peterson, J. Stankovich, and L. Kinsman, "Emergency department crowding: A systematic review of causes, consequences and solutions," *PloS one*, vol. 13, no. 8, p. e0203316, 2018.

[68] S. Di Somma, L. Paladino, L. Vaughan, I. Lalle, L. Magrini, and M. Magnanti, "Overcrowding in emergency department: an international issue," *Internal and emergency medicine*, vol. 10, no. 2, pp. 171–175, 2015.

[69] R. Salway, R. Valenzuela, J. Shoenberger, W. Mallon, and A. Viccellio, "Emergency department (ed) overcrowding: evidence-based answers to frequently asked questions," *Revista Médica Clínica Las Condes*, vol. 28, no. 2, pp. 213–219, 2017.

[70] B. Riva, A. Clavenna, M. Cartabia, A. Bortolotti, I. Fortino, L. Merlino, A. Biondi, and M. Bonati, "Emergency department use by paediatric patients in lombardy region, italy: a population study," *BMJ paediatrics open*, vol. 2, no. 1, 2018.

[71] R. Longhi, R. Picchi, D. Minasi, and A. D. C. Merlone, "Pediatric emergency room activities in italy: a national survey," *Italian journal of pediatrics*, vol. 41, no. 1, p. 77, 2015.

[72] R. Levaggi, M. Montefiori, and L. Persico, "Speeding up the clinical pathways by accessing emergency departments," *The European Journal of Health Economics*, pp. 1–8, 2019.

[73] J. M. Legramante, L. Morciano, F. Lucaroni, F. Gilardi, E. Caredda, A. Pesaresi, M. Coscia, S. Orlando, A. Brandi, G. Giovagnoli *et al.*, "Frequent use of emergency departments by the elderly population when continuing care is not well established," *PLoS One*, vol. 11, no. 12, 2016.

[74] E. Amodio, L. C. d'Oro, E. Chiarazzo, C. Picco, M. Migliori, I. Trezzi, S. Lopez, O. Rinaldi, and M. Giupponi, "Emergency department performances during overcrowding: the experience of the health protection agency of brianza," *AIMS public health*, vol. 5, no. 3, p. 217, 2018.

[75] P. Cremonesi, E. di Bella, M. Montefiori, and L. Persico, "The robustness and effectiveness of the triage system at times of overcrowding and the extra costs due to inappropriate use of emergency departments," *Applied health economics and health policy*, vol. 13, no. 5, pp. 507–514, 2015.

[76] M. Unwin, L. Kinsman, and S. Rigby, "Why are we waiting? patients' perspectives for accessing emergency department services with non-urgent complaints," *International emergency nursing*, vol. 29, pp. 3–8, 2016.

[77] M. Ko, Y. Lee, C. Chen, P. Chou, and D. Chu, "Prevalence of and predictors for frequent utilization of emergency department: a population-based study," *Medicine*, vol. 94, no. 29, 2015.

[78] P. Aboagye-Sarfo, Q. Mai, F. M. Sanfilippo, D. B. Preen, L. M. Stewart, and D. M. Fatovich, "Growth in western australian emergency department demand during 2007–2013 is due to people with urgent and complex care needs," *Emergency Medicine Australasia*, vol. 27, no. 3, pp. 202–209, 2015.

[79] K. B. Velt, M. Cnossen, P. P. Rood, E. W. Steyerberg, S. Polinder, and H. F. Lingsma, "Emergency department overcrowding: a survey among european neurotrauma centres," *Emerg Med J*, vol. 35, no. 7, pp. 447–448, 2018.

[80] C. Pei, Y. Zhang, Y. Zhang, F. Sun, X. Lin, H. Sun, J. Wu, P. Jiang, J. Ge, W. Ou *et al.*, "Personalized re-ranking for recommendation," in *Proceedings of the 13th ACM Conference on Recommender Systems*, 2019, pp. 3–11.

[81] C. Park, D. Kim, J. Oh, and H. Yu, "Improving top-k recommendation with truster and trustee relationship in user trust network," *Information Sciences*, vol. 374, pp. 100–114, 2016.

[82] E. Ghanbari and A. Shakery, "Query-dependent learning to rank for cross-lingual information retrieval," *Knowledge and Information Systems*, vol. 59, no. 3, pp. 711–743, 2019.

[83] O. A. S. Ibrahim and D. Landa-Silva, "An evolutionary strategy with machine learning for learning to rank in information retrieval," *Soft Computing*, vol. 22, no. 10, pp. 3171–3185, 2018.

[84] D. Seyler, P. Chandar, and M. Davis, "An information retrieval framework for contextual suggestion based on heterogeneous information network embeddings," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 953–956.

[85] Y. Bengio, *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[86] E. Herrera-Viedma, G. Pasi, A. G. Lopez-Herrera, and C. Porcel, "Evaluating the information quality of web sites: A methodology based on fuzzy computing with words," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 4, pp. 538–549, 2006.

[87] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[88] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie, "Learning from noisy large-scale datasets with minimal supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 839–847.

[89] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artificial intelligence*, vol. 151, no. 1-2, pp. 155–176, 2003.

[90] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data science journal*, vol. 14, 2015.

[91] J. Merino, I. Caballero, B. Rivas, M. Serrano, and M. Piattini, "A data quality in use model for big data," *Future Generation Computer Systems*, vol. 63, pp. 123–130, 2016.

[92] D. Ardagna, C. Cappiello, W. Samá, and M. Vitali, "Context-aware data quality assessment for big data," *Future Generation Computer Systems*, vol. 89, pp. 548–562, 2018.

[93] H. Liu, A. K. Tk, J. P. Thomas, and X. Hou, "Cleaning framework for bigdata: An interactive approach for data cleaning," in *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2016, pp. 174–181.

[94] S. Krishnan, M. J. Franklin, K. Goldberg, J. Wang, and E. Wu, "Activeclean: An interactive data cleaning framework for modern machine learning," in *Proceedings of the 2016 International Conference on Management of Data*, 2016, pp. 2117–2120.

[95] C. Chuck, M. Laskey, S. Krishnan, R. Joshi, R. Fox, and K. Goldberg, "Statistical data cleaning for deep learning of automation tasks from demonstrations," in *2017 13th IEEE Conference on Automation Science and Engineering (CASE)*. IEEE, 2017, pp. 1142–1149.

[96] B. Saha and D. Srivastava, "Data quality: The other face of big data," in *2014 IEEE 30th International Conference on Data Engineering*. IEEE, 2014, pp. 1294–1297.