



*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870811*



## **D3.2 Semantic annotation of social curatorial products**

<b>Deliverable information</b>	
WP	WP3
Document dissemination level	PU Public
Deliverable type	Other
Lead beneficiary	CELI
Contributors	CELI
Date	30th April 2021
Document status	Final
Document version	1.0

***Disclaimer: The communication reflects only the author's view and the Research Executive Agency is not responsible for any use that may be made of the information it contains***

INTENTIONALLY BLANK PAGE

## Project information

**Project start date:** 1<sup>st</sup> of May 2020

**Project Duration:** 36 months

**Project website:** <https://spice-h2020.eu>

### Project contacts

#### Project Coordinator

**Silvio Peroni**

ALMA MATER STUDIORUM -  
UNIVERSITÀ DI BOLOGNA  
Department of Classical  
Philology and Italian Studies –  
FICLIT  
E-mail: [silvio.peroni@unibo.it](mailto:silvio.peroni@unibo.it)

#### Scientific Coordinator

**Aldo Gangemi**

Institute for Cognitive  
Sciences and Technologies of  
the Italian National Research  
Council  
E-mail: [aldo.gangemi@cnr.it](mailto:aldo.gangemi@cnr.it)

#### Project Manager

**Adriana Dascultu**

ALMA MATER STUDIORUM -  
UNIVERSITÀ DI BOLOGNA  
Executive Support Services  
E-mail:  
[adriana.dascultu@unibo.it](mailto:adriana.dascultu@unibo.it)

### SPICE consortium

No.	Short name	Institution name	Country
1	UNIBO	ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA	Italy
2	AALTO	AALTO KORKEAKOULUSAATIO SR	Finland
3	DMH	DESIGNMUSEON SAATIO - STIFTELSEN FOR DESIGNMUSEET SR	Finland
4	AAU	AALBORG UNIVERSITET	Denmark
5	OU	THE OPEN UNIVERSITY	United Kingdom
6	IMMA	IRISH MUSEUM OF MODERN ART COMPANY	Ireland
7	GVAM	GVAM GUIAS INTERACTIVAS SL	Spain
8	PG	PADAONE GAMES SL	Spain
9	UCM	UNIVERSIDAD COMPLUTENSE DE MADRID	Spain
10	UNITO	UNIVERSITA DEGLI STUDI DI TORINO	Italy
11	FTM	FONDAZIONE TORINO MUSEI	Italy
12	CELI	CELI SRL	Italy
13	UH	UNIVERSITY OF HAIFA	Israel
14	CNR	CONSIGLIO NAZIONALE DELLE RICERCHE	Italy

## Executive summary

The semantic annotator is an annotation service for the semantic enrichment of textual contents, targeting user generated contents as well as descriptions of museum artifacts. The service is multilingual and supports English, Finnish, Hebrew, Italian and Spanish. It consists of a natural language processing pipeline that performs sentiment analysis, emotion detection and entity linking; the service annotates the textual contents with respect to the ontological model developed in WP6 and stores the generated RDF graph in the linked data hub developed by WP4.

The document is structured as follows: first, it introduces the main aim of the annotation service and its general architecture, then it describes the different analysis components and finally presents the output format of the annotations as well as the service API details provided by the REST endpoint.

## Document History

Version	Release date	Summary of changes	Author(s) –Institution
V0.1	02/04/2021	First draft released	Alessio Bosca, Adriana Dematteis - CELI
V0.2	19/04/2021	Revised version released	Marilena Daquino – UniBO Rossana Damiano - UniTo
V0.3	23/04/2021	Final draft released	Alessio Bosca, Adriana Dematteis - CELI
V1	29/04/2021	Final version	Alessio Bosca, Adriana Dematteis - CELI

## Table of Contents

1 Introduction	7
2 Semantic Annotator Architecture and Infrastructure	7
2.1 Analysis Pipeline Components	9
3 SPICE Emotion and Sentiment Lexicon	10
3.1 State of the art	10
3.1.1 Linguistic corpora and emotion lexicons	10
3.1.2 Emoji sources and corpora for Emotion Detection	11
3.2 Creation of the Lexicon	12
3.2.1 Emoji	13
3.2.2 Lexicon	14
4 Analysis Pipeline Components Implementation	16
4.1 Language Analysis	16
4.2 Emotion Detection	17
4.3 Sentiment Analysis	18
4.4 Entity Linking	18
5 Semantic Annotator Service	19
5.1 Service Input	19
5.2 Service Output	20
7 References	23

## 1 Introduction

This deliverable document introduces the first release of SPICE Semantic Annotator. This component is used to semantically enrich textual contents with metadata linking content to concepts described in a knowledge graph and it targets user generated contents as well as descriptions of museum artifacts. Semantically tagged documents are easier to find, interpret, combine and reuse.

The result of this process consists in the automatic creation of metadata enriching the document (or specific fragments of it) with identifiers of concepts and entities mentioned in the text or relevant to it. Such references link the textual contents to the formal description of concepts/entities in a knowledge graph, and allow for further reasoning over the latter. In the context of SPICE project, reasoning over such semantic annotation allows for abstracting and generalizing input coming from museum visitors, finding commonalities between them and ultimately supporting the activity of users and communities modelling and the design of an advanced recommendation engine. More details can be found in **D3.1.1: *Prototype user and community modelling***.

The Semantic Annotator therefore is a key component within the SPICE infrastructure, since it provides a connection between the contents, coming from the User interfaces, developed in WP5 (more details on User Interfaces can be found in **D5.1.1: *Preliminary interfaces for interpretation***), and SPICE knowledge graph, designed in WP6 (more details on the knowledge graph can be found in **D6.3.1: *Initial ontology network specification***). The semantic annotations produced by this component are stored in the linked data hub developed in WP4 (more details on the Linked Data Hub infrastructure can be found in **D4.1 *Linked Data server technology: requirements and initial prototype***). Also, in WP4 the API was used as part of the dashboard for citizen curation activities analytics, further detailed in D4.1, and this may work as a *de facto* evaluation in a real-world application.

The Semantic Annotator currently performs Sentiment Analysis, Emotion Detection and Entity Linking; it is multilingual and supports all the languages used in museum use cases: English, Finnish, Hebrew, Italian and Spanish. The components providing Sentiment Analysis and Emotion Detection were specifically designed and developed for the project, while the components for the basic language analysis (e.g. lemmatization, PoS tagging) and entity linking were implemented reusing available Open Source resources and models.

Section 2 outlines the Semantic Annotator Architecture and Infrastructure, [Section 3](#) presents the SPICE multilingual Emotion and Sentiment Lexicon, while Section 4 describes the different analysis components. Finally, Section 5 presents the output document format as well as usage details of the service prototype, and service requests and responses examples.

## 2 Semantic Annotator Architecture and Infrastructure

This section describes the architecture of the Semantic Annotator and the interaction between the different analysis components.

The process of semantic annotation is realized by a Natural Language Processing Pipeline that includes different analysis modules, each one responsible for annotating the document with respect to a specific aspect: sentiment analysis, emotion detection, entity linking. The overall process is exposed by means of standard RESTful<sup>1</sup> APIs and produces a JSON-LD<sup>2</sup> document as output. JSON-LD is a JSON-based serialization for Linked Data that can be seamlessly stored in the Linked Data hub of WP4.

---

<sup>1</sup> <https://www.w3.org/2001/sw/wiki/REST>

<sup>2</sup> <https://www.w3.org/TR/json-ld11/>

The architecture (graphically represented in Figure 1) is designed to be modular and configurable in order to allow new analysis components to be included, or to easily replace any of them and experiment with different algorithms and models.

The RESTful service acts as the entry point of the annotation process. It receives as input the textual contents to be analysed along with some metadata (e.g., the contents language); textual contents and metadata are wrapped into a document object that is enriched with an analysis plan (detailing the different modules that should process the contents and in which order) and then it is submitted to the pipeline. An orchestration component within the pipeline is responsible for forwarding the document to the correct analysis modules; when the analysis plan is completed, the document is returned to the RESTful service that formats the document in the desired output (a JSON-LD document) and provides it as output.

The data exchange between the different modules of the pipeline (entry point, analysis component, orchestrator) happens by means of message queues. The queue system used is Kafka,<sup>3</sup> an open source, distributed messages streaming platform. Such a design choice allows us to isolate and decouple the different analysis modules that can be implemented with different technologies (e.g. Java, Python, R), and to exploit a wide variety of models and solutions available on the open source.

The whole pipeline is deployed on a Kubernetes<sup>4</sup> cluster with the replication of the analysis components managed by KEDA<sup>5</sup>. Kubernetes is an open-source system for automating deployment, scaling, and management of containerized components (e.g. Docker<sup>6</sup> images). KEDA instead is a single-purpose and lightweight component that can be added into any Kubernetes cluster and acts as a Kubernetes-based Event Driven Autoscaler; with KEDA it is therefore possible to configure the scaling of any container in Kubernetes based on the number of events needing to be processed.

Such architectural solutions were chosen in order to achieve horizontal scalability. In particular, we (1) increase the instances of a given component when the number of documents waiting to be processed in the queue exceeds a certain threshold and (2) decrease their number when they go below the threshold. This approach allows us to ensure service response time regardless of the system workload and to reduce costs by dismissing computational resources when they are not needed.

The whole system is deployed to AWS<sup>7</sup> cloud resources, on the servers in the European region.

---

<sup>3</sup> <https://kafka.apache.org/>

<sup>4</sup> <https://kubernetes.io/>

<sup>5</sup> <https://keda.sh/>

<sup>6</sup> <https://www.docker.com/>

<sup>7</sup> <https://aws.amazon.com/>



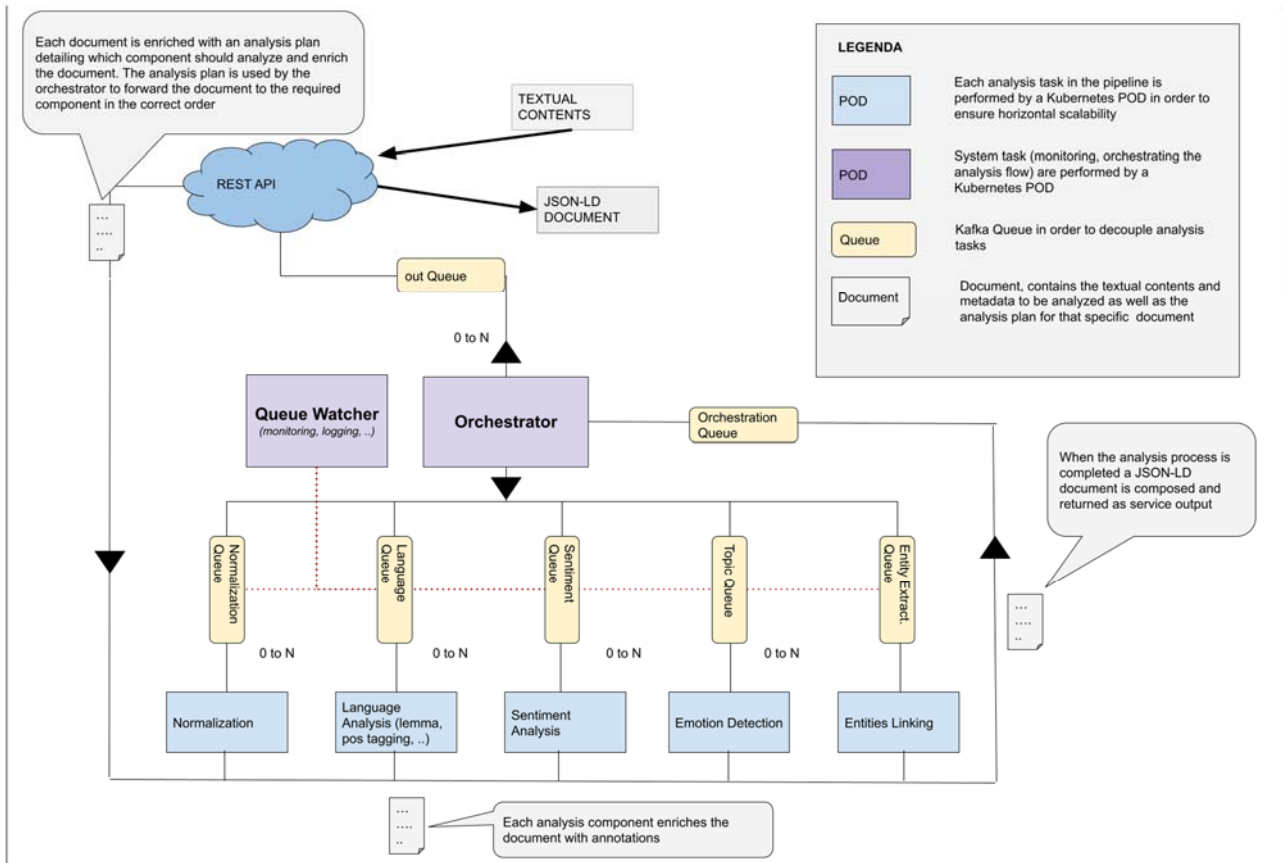


Figure 1. Semantic Annotator Architecture

## 2.1 Analysis Pipeline Components

The analysis pipeline includes the following modules:

- **Language Analysis**, with the goal of performing standard language analysis on the contents (e.g., lemmatization, PoS tagging). Such analysis will be exploited by the Emotion Detection and Sentiment Analysis components.
- **Emotion Detection**, with the goal of detecting textual expressions that can be linked to emotions, targeting terms of the Plutchik Emotion ontology, one of the emotions models specified by WP6 (more details on the ontological models used in SPICE for the representation of emotions can be found in the Deliverable Document **D6.3.1: Initial ontology network specification**). The decision to use Plutchik's Emotion model is due to several reasons: firstly, most of the emotive lexica available (which we take as resources for the SPICE lexicon) adopt this model; secondly, Plutchik's model covers a wide variety of emotions, because it considers low and high intensities of each of the eight basic emotions (giving as a result 24 emotions) and the interaction between the basic emotions, which determines 27 complex emotions. More information can be found in Section 3.
- **Sentiment Analysis**, with the goal of detecting textual expressions that carry a subjective information (e.g., like and dislike statements) along with its polarity: positive, negative or neutral. The conceptual framework used to model sentiment polarity is the MARL<sup>8</sup> ontology; Marl is a standardised data schema designed to annotate and describe subjective opinions expressed on the web or in particular Information Systems.
- **Entity Linking**, with the goal of detecting textual expressions that can be linked to relevant concepts and named entities in order to obtain a representation of the semantics of the contents through the detected entities and their types. Since the topics of user generated contents (as well

<sup>8</sup> <http://www.gsi.dit.upm.es:9080/ontologies/marl/>

as the subjects of museums use cases) cannot be restricted to a specific domain we decided to use DBPedia<sup>9</sup> as the target conceptual framework. The scope of the entity linker can be broadened in the following phases of the project if needed.

The components providing Sentiment Analysis and Emotion Detection were specifically designed and developed for the project, while the components for the basic language analysis (e.g., lemmatization, PoS tagging) and entity linking were implemented reusing available Open Source resources and models.

The Sentiment Analysis and Emotion Detection modules are rules based systems and they both exploit a multilingual lexicon resource created for the project and described in Section 3; Section 4 provides details on the different analysis components.

### 3 SPICE Emotion and Sentiment Lexicon

This section presents the resources and the methodology adopted in the creation of the SPICE emotion and sentiment multilingual lexicon.

#### 3.1 State of the art

This subsection describes the state of the art with respect to linguistic and emoji resources used in Emotion Detection.

##### 3.1.1 Linguistic corpora and emotion lexicons

Starting from a preliminary research on the state of the art on the emotion detection from texts and existing emotion lexicons, we refer to different linguistic resources both as benchmarks and as sources for the creation of our SPICE Emotion Lexicon.

One of the biggest multilingual lexicons of recent years is the NRC Word-Emotion Association Lexicon (EmoLex) by Mohammad, S. M., & Turney, P. D. (2013). The NRC Lexicon is a list of words manually annotated via crowdsourcing through Amazon's Mechanical Turk (a crowdsourcing marketplace) and their associations with Plutchik's eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two polarities (negative and positive). The source language is English, with other 104 languages obtained via automatic translation.

Two works address Italian lexicon, namely: the CELI Emotion Lexicon by CELI (2020) and the Italian EMotive Lexicon by Passaro et. al (2015). The CELI Emotion Lexicon for Italian has been created and updated during the years, performing social media monitoring, Sentiment Analysis and Emotion Detection. It contains 9,321 entries (single words and multiwords); each entry has: information about sentiment, i.e. polarity (positive or negative); one associated emotion; information if an entry is in its canonical or variant form (e.g. "antidepressivo" canonical vs. "anti depressivo" variant); domain of application. The lexicon comprises 15 tags of emotions or moods, which are: abandoned, anger, disgust, dislike, fear, happy, joy, like, love, worried, sadness, calmness, hopeful, surprise, peaceful. Not all the entries of the lexicon are associated with an emotion.

ItEM – Italian Emotive Lexicon (Passaro, L. et al., 2015) is an emotion lexicon for Italian in which each target term is provided with an association score with the basic emotions defined in the Plutchik's taxonomy: Joy, Sadness, Anger, Fear, Trust, Disgust, Surprise, Anticipation. Using an online elicitation paradigm, 60 Italian native speakers of different age groups, levels of education, and backgrounds were asked to list, for each of

---

<sup>9</sup> <https://www.dbpedia.org/>

the eight basic emotions, 5 lemmas for each of their Parts-of-Speech (PoS) of interest (Nouns, Adjectives and Verbs). Then, researchers calculated the emotion distinctiveness of each of the lemmas, taking into consideration only the terms evoked by a single emotion. They collected 555 emotive seeds.

A more recent resource for Italian is MultiEmotions-It (Sprugnoli, R., 2020), containing comments to music videos and advertisements posted on YouTube and Facebook manually annotated according to four different dimensions: i.e., relatedness, opinion polarity, emotions and sarcasm. For the annotation of emotions, they adopted the model proposed by Plutchik, taking into consideration both the eight basic emotions and the dyads, i.e., feelings composed of two basic emotions (e.g., Love is a combination of Joy and Trust).

### 3.1.2 Emoji sources and corpora for Emotion Detection

As for the linguistic data, we took into consideration various studies on the use of emojis to express emotions in texts (mainly on social networks), which we use as benchmarks and resources for the inclusion of emojis in the SPICE Emotion Lexicon.

Emojis are a widely used form of communication, especially in social media, and their importance for the specific purpose of the SPICE project was already highlighted in a first test carried out with GAMGame. GAMGame is a Web App, inspired by mobile apps, created in occasion of a Researchers Night event organized by GAM (Modern Art Gallery in Turin), UNITO, CELI in November 2020, during which classes received a presentation about SPICE and citizen curation methods, then made an online activity with the App GAMGame. In the game, among other activities, participants had to annotate works of art with emojis and text.

The usefulness of emojis lies in the fact that they allow the user to express emotions in a more immediate and, above all, visual way. They are similar to a widely used jargon, especially for new generations, and that is why it is necessary to support them within a project like SPICE. It is a type of user-friendly communication that can be used to express impressions in a very intuitive and simple way, also by categories of users who may have difficulties in producing written text on technological devices (such as older people, people with disabilities or children generally who do not produce long and content-rich texts). For this reason, especially with certain types of interfaces, emojis can be functional to make the fruition process easier and, consequently, increase the engagement of different stakeholders and the incisiveness of the project.

First of all, a big source of emoji is Emojipedia<sup>10</sup>, an emoji reference website which documents the meaning and common usage of emoji characters in the Unicode Standard. Two interesting tasks about emoji are the Emoji Prediction task (Barbieri, F. et al., 2018) and the ITAmoji task (Ronzano, F. et al, 2018). In the Emoji Prediction task, participants had to predict, given a tweet in English or Spanish, its most likely associated emoji; the emoji were first removed from the original tweet. The dataset includes tweets that only contain one emoji, of the 20 most frequent emojis. Similarly, in the ITAmoji task, participants were invited to predict, given an Italian tweet, its most likely associated emoji.

Other interesting studies about emotion analysis through emoji are EmoTag (Shoeb, A.A et al., 2019) and the work by Wolny, W. (2016), both focused on the importance of Twitter as a source for sentiment and emotion analysis, because emoji express one's emotions regardless of his/her language. In EmoTag, researchers provide a method to quantify the emotional association of basic emotions such as anger, fear, joy, and sadness for a set of emojis.; they collected and created a corpus of 20 million emoji-centric tweets. Still, Wolny (2016) proposes to extend existing binary sentiment classification approaches using a multi-way emotions classification.

---

<sup>10</sup> <https://emojipedia.org/>

### 3.2 Creation of the Lexicon

For the creation of the multilingual lexicon "SPICE Emotion Lexicon", we chose a subset of the Italian Emotion Lexicon developed by CELI (9,321 entries) as the source dataset. This lexicon is then integrated with words taken from a subset (6,468 entries) of the NRC Word-Emotion Association Lexicon (Mohammad S. M., & Turney, P. D., 2013).

First of all, we created a subset of the NRC Lexicon, because we only needed linguistic data for English, Italian, Spanish, Hebrew and Finnish. We started from 14,183 entries of the NRC Lexicon; then, we deleted 98 languages and we only kept the languages of interest for SPICE. We created a column with the sum of the sentiment values and a column with the sum of the emotion values. We deleted all the entries that have neither sentiment nor emotions. We obtained a subset with 6,468 entries. The SPICE Lexicon is also integrated with a list of Italian emotive words (555 entries) contained in the ItEM lexicon (Passaro, L. et al., 2015).

Each term is associated with one or more emotions and one or more sentiment. The emotions belong to Plutchik's Emotion Model and they are: Anger, Anticipation, Calmness, Disgust, Disapproval, Fear, Joy, Interest, Love, Sadness, Surprise, Trust. The emotions are mapped on the emotion model that have been developed for the SPICE ontology by WP6.

Italian is chosen as the source language and CELI Lexicon as the source lexicon. Each Italian term has been translated into English, Spanish, Finnish and Hebrew. Firstly, we checked if each of the Italian words from the CELI Lexicon is also present in the NRC Lexicon; if so, we copied the translations from there, otherwise we did the translations ourselves. Not all of the Italian terms have a corresponding translation into English, Spanish, Hebrew or Finnish. Therefore, a different number of lexical entries for each language follows. If there is no corresponding translation for an Italian word in one of the target languages, we leave the cell blank in the target language column. If several translations in English, Spanish, Finnish or Hebrew correspond to an Italian term in NRC, we report the translations on several lines (Please note: we report only the translations of terms associated with emotions. If a term is associated with no emotions, we do not report the translation). A first quality control of the Hebrew and Finnish lexicons by the University of Haifa and Aalto University partners highlighted some mismatches and discrepancies in meaning between the terms of Hebrew and Finnish and those of the other three languages (English, Italian and Spanish). Taken as assumed that out-of-context translation may lead to mis-interpretations, we decide, for the moment, to keep Hebrew and Finnish lexicons separated, as singular resources, because, generally speaking, languages are culturally influenced, so that a given word can have a specific meaning in one language, but a completely different one in another; as a direct consequence, a different categorization of words in terms of emotion and / or sentiment evoked may follow. Many terms of the two languages, in this first phase of work, are not aligned with the corresponding meanings in the other languages, leading to incorrect interpretations. Therefore, a more accurate and deep translation work is necessary and it will be carried out in the next phases of the project for all the languages involved. The Hebrew and Finnish lexicons, then, have been exported as single lexicons.

The first version of the Hebrew lexicon contains 1,003 terms annotated with Plutchik's eight basic emotions, but it will be integrated with other linguistic resources in the next phases of the project.

For the Finnish lexicon, we referred to SELF (Sentiment and Emotion Lexicon for Finnish)<sup>11</sup>, by Öhman, E. (forthcoming) from The University of Helsinki. This lexicon is based on the NRC Emotion Lexicons and is a revised extension of the automatic translation done by the original creator. The first Finnish resource for SPICE contains 5,839 entries annotated with polarity (positive or negative) and emotions. The emotions are Plutchik's eight basic emotions. The temporary output is a multilingual aligned lexicon for English (1,865 entries), Italian (2,483 entries) and Spanish (1,795) and singular lexicons for Hebrew (1,003 entries) and Finnish (5,836 entries). All the individual

---

<sup>11</sup> Öhman, E., SELF and FEIL: Emotion Lexicons for Finnish, (Forthcoming)

linguistic resources will then be harmonized and integrated into the final multilingual lexicon in the continuation of the project.

expression EN	expression IT	expression ES	sentiment	emotions
	<abbacchiamento>		Negative	Sadness
<depressed>	<abbacchiato>		Negative	Sadness
<dazzle>	<abbacinare>		Neuter	Surprise
<abandon>	<abbandonare>	<abandonar>_VERB	Negative	Sadness, Fear
<abandoned>	<abbandonato>	<abandonado>	Negative	Sadness, Anger, Fear
<abandonment>	<abbandono>	<abandono>	Negative	Sadness
<fell>	<abbattere>	<abatir>_VERB	Negative	Sadness
<shot down>	<abbattuto>	<abatido>	Negative	Sadness
<dupe>	<abbindolare>	<embaucar>_VERB	Negative	Anger
<abundant>	<abbondante>	<abundante>	Positive	Joy
<abundance>	<abbondanza>	<abundancia>	Positive, Negative	Anticipation, Disgust, Joy, Trust
<hug>_VERB	<abbracciare>	<abrazar>_VERB	Positive	Joy
<brutalized>	<abbrutito>	<brutalizado>	Negative	Disapproval
<aberrant>	<aberrante>	<aberrante>	Negative	Disgust
<aberrate>	<aberrare>	<aberrar>_VERB	Negative	Disgust
<aberration>	<aberrazione>	<aberración>	Negative	Disgust
<abject>	<abietto>	<abyecto>	Negative	Disgust
<abjection>	<abiezione>	<abyección>	Negative	Disgust
<abysmal>	<abissale>	<abismal>	Negative	Sadness

Table 1. Multilingual Lexicon for English, Italian and Spanish Excerpt

When exporting the single lexicon of each language to create single lexicons, we check the lines for duplicates. In the SPICE Emotion Lexicon there are several lines with the same words, to which different emotions are associated; in that case, in the single lexicon, we merge the emotion labels.

### 3.2.1 Emoji

The lexicon also comprises a set of 122 emojis associated with emotions, which seem to be the most used and widespread on social media. The emojis are taken from Emojipedia. The association between the emoji and the emotion has been made both arbitrarily and on the basis of previous studies on emotion detection in the emoji field (Wolny, W. (2016); Shoeb, A.A. et al (2019); Arva, H. et al (2018)).














Emoji	Name (EN)	Nome (IT)	Nombre (EN)	Sentiment	Emotions
	Grinning Face	Faccina che sorride	Cara sorridente	Positive	Joy, Interest
	Grinning Face with Big Eyes	Faccina che sorride con occhi grandi	Cara sorridente con ojos grandes	Positive	Joy, Interest
	Grinning Face with Smiling Eyes	Faccina che sorride con occhi sorridenti	Cara sorridente con ojos sonrientes	Positive	Joy, Interest
	Beaming Face with Smiling Eyes	Faccina raggianti con occhi sorridenti	Rostro radiante con ojos sonrientes	Positive	Joy, Interest
	Smiling Face with Smiling Eyes	Faccina sorridente con occhi sorridenti	Cara sorridente con ojos sonrientes	Positive	Joy, Interest
	Face with Tears of Joy	Faccina con lacrime di gioia	Rostro con lágrimas de alegría	Positive	Joy
	Grinning Squinting Face	Faccina che ride e strabica	Cara sorridente entrecerrando los ojos	Positive	Joy
	Rolling on the Floor Laughing	Rotolare a terra ridendo	Rodando en el piso, riendo	Positive	Joy
	Smiling Face with Heart-Eyes	Faccina sorridente con occhi a cuore	Cara sorridente con ojos de corazón	Positive	Joy, Love, Interest
	Face Blowing a Kiss	Faccina che manda un bacio	Cara que sopla un beso	Positive	Joy, Love
	Kissing Face with Closed Eyes	Faccina che manda un bacio con gli occhi chiusi	Besar la cara con los ojos cerrados	Positive	Joy, Love
	Kissing Face with Smiling Eyes	Faccina che manda un bacio con gli occhi sorridenti	Besar la cara con ojos sonrientes	Positive	Joy, Love
	Hugging Face	Faccina che abbraccia	Abrazando la cara	Positive	Love, Joy, Interest

Table 2. *Emoji Lexicon Excerpt*

### 3.2.2 Lexicon

Terms of the lexicon are entered in two ways:

- **Lemmatized form:** the single term is inserted between angle brackets. A lemma is the basic form of the word which you find in the dictionary (the basic form comprises all the forms that a word can have). This is the case when a token/span changes its value if conjugated or declined, e.g.

EXPRESSION	EXAMPLE
<love>	1. I <b>have loved</b> you 2. He <b>loves</b> you

Table 3. *Lexicon entry with lemmatized form*

- **Superficial form:** the term is annotated without the use of brackets or other punctuation, meaning that the word in question can only have the specific form in which it was extracted. It is the case, for example, of idiomatic expressions or other expressions. In the following expression, for example, words are partly inserted in their lemmatized form and partly in their superficial form: the verb <be> can be conjugated, while the adjective “patient” cannot be declined.

EXPRESSION	EXAMPLE
<be> patient	I'm <b>patient</b>

Table 4. *Lexicon Entry with superficial form and lemma*

In some cases, a lemma can be ambiguous between different grammatical categories (e.g., <hug>\_verb and <hug>\_noun). To resolve the ambiguity, the lemma is associated with a label (*\_pos*, which indicates the grammatical category), which varies according to the part of the speech in question. Labels are attached outside the angle brackets. Taking up the previous example, the output is: <hug>\_VERB (or <hug>\_NOUN).

Below the list of *\_pos*:

**SOURCE:** <https://universaldependencies.org/u/pos/>

```

ADJ,          // adjective
ADP,          // adposition
ADV,          // adverb
AUX,          // auxiliary
CCONJ,        // coordinating conjunction
DET,          // determiner
INTJ,         // interjection
NOUN,         // noun
NUM,          // numeral
PART,         // particle
PRON,         // pronoun
PROPN,        // proper noun
PUNCT,        // punctuation
SCONJ,        // subordinating conjunction
SYM,          // symbol
VERB,         // verb
X,           // other

```

*Figure 2. Universal Dependencies Part of Speech Tagset*

Some words are syntactically ambiguous, that is, for example, some words can be both adjectives and past participles looking at their superficial form, especially in English. To give an example, the term “afflicted” can represent the past participle of the verb “to afflict” or the adjective “afflicted”. To resolve this kind of ambiguity, the morphological analyzer Stanza<sup>12</sup> is used. If a term is recognized by the morphological analyzer as an adjective and not as the participial form of a verb, then it is written as it is between the angle brackets in its base form (e.g. the term “afflicted” is recognised as an adjective, so it is inserted in the lexicon in his superficial form “afflicted”); contrariwise, if a word is recognised as the participle of a verb, then the base form of the verb is reported (e.g. the term “dented” is recognised as the participle of the verb “dent” and not as the adjective “dented”, so the term reported in the lexicon is <dent>).

At the moment the lexicon is a work in progress resource that we are enriching and harmonizing and when it will have reached a definitive form we will share it in the final deliverable of the project.

<sup>12</sup> <https://stanfordnlp.github.io/stanza/>

## 4 Analysis Pipeline Components Implementation

This section provides details on the different analysis components.

### 4.1 Language Analysis

Stanza is a Python natural language analysis package developed by the Stanford NLP Group, exploiting neural networks models built on top of the Pytorch<sup>13</sup> ecosystem. It contains tools which can be used in a pipeline, to (1) convert a string containing human language text into lists of sentences and words, (2) generate base forms of those words (lemmas), their parts of speech and morphological features, and to (3) give a syntactic structure dependency parse. The toolkit is designed to be parallel among more than 70 languages, using the Universal Dependencies<sup>14</sup> formalism.

More detail on Stanza implementation can be found in:

- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. In Association for Computational Linguistics (ACL) System Demonstrations. 2020.

In the SPICE analysis pipeline, the output of this component is used by Sentiment Analysis and Emotion Detection modules, as they are rule-based system, exploiting morphological and syntax features of textual contents.

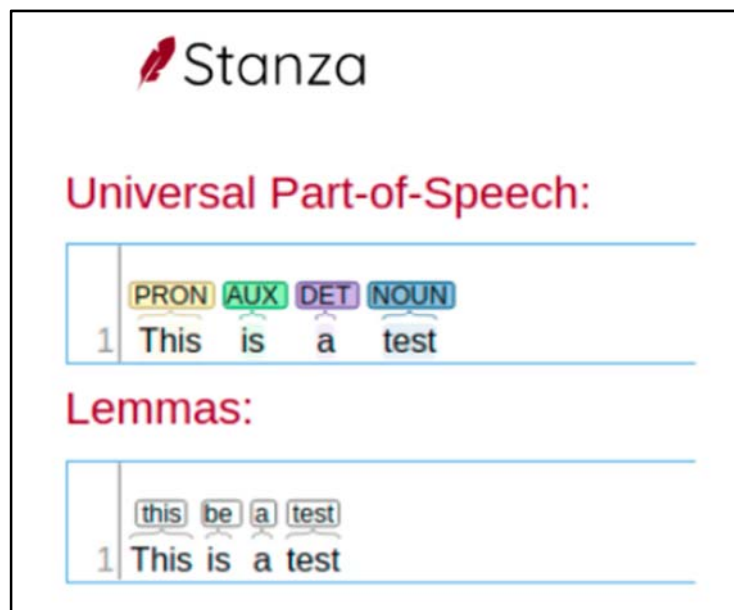


Figure 3. Stanza Analysis Output Example

<sup>13</sup> <https://pytorch.org/>

<sup>14</sup> <https://universaldependencies.org/>



## 4.2 Emotion Detection

The component for emotion detection is a rule-based system that identifies patterns in a sequence of words (enriched with morphological features, as lemma or part of speech). The rules are automatically generated combining the entries of an emotion lexicon (linking expression to emotions) with language specific rules (used for handling conjunctions, modifiers as adverbs, negations). Language specific rules were manually defined by linguists.

The rule engine used for applying these rules to the sequence of analyzed terms is Drools<sup>15</sup>. Drools is a rule management system with a forward and backward chaining inference-based rules engine, also known as a production rule system using an enhanced implementation of the Rete algorithm. More details on the Rete algorithm implementation used in Drools can be found in

- Proctor, Mark, et al. "Drools documentation." *JBoss 5.05* (2008): 2008.

---

<sup>15</sup> <https://www.drools.org/>

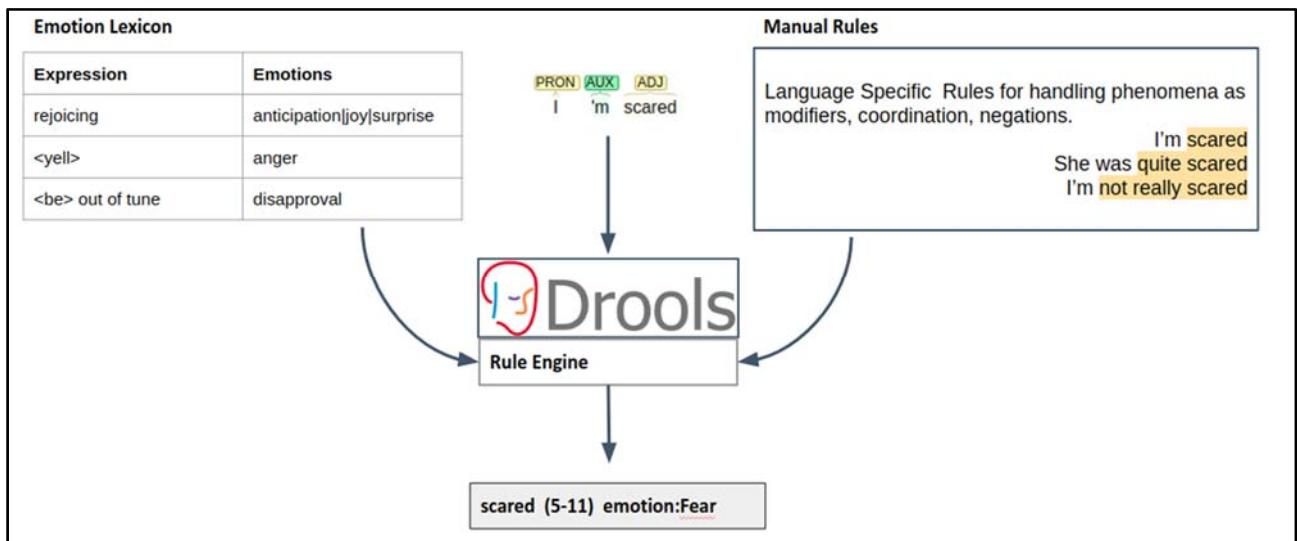


Figure 4. Emotion Detection Rule Based System

In the following phase of the project, as the generated contents from real users will start to be available, we plan to enrich the emotion recognition system, by adding to the rule based system a supervised sequence labelling system trained on contents coming from real users.

In order to test this approach, we are currently running some preliminary experiments on the GoEmotion<sup>16</sup> dataset, a public resource of textual contents (user comments from Reddit, linked to emotions) recently released by Google, with promising results.

### 4.3 Sentiment Analysis

The Sentiment Analysis Component is a rule-based system based on the same approach of the component for the Emotion Detection. The main difference between lies in the rules for handling negations; in the sentiment analysis component a negation modifies the polarity of an expression (e.g. very nice -> positive; not very nice -> negative). In the context of emotion, a negation removes an emotion expression from the analysis (e.g. I'm not really scared -> NOTHING).

In the following phase of the project, as the generated contents from real users will start to be available, we plan to enrich the sentiment analysis as well, by adding to the rule-based system a supervised sequence labelling system trained on contents coming from real users.

### 4.4 Entity Linking

DBpedia Spotlight<sup>17</sup> is a well-known Open Source library for automatically annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia. Pretrained models<sup>18</sup> for several languages are available on the project page, including English, Finnish, Italian and Spanish. It is possible as well to train new models<sup>19</sup> in order to support other languages, (as Hebrew, not present in the pretrained models) or to update existing models for including new entities added in DBpedia in the meantime.

The algorithm used by DBPedia Spotlight consists of a 4 steps approach:

- **Spotting:** Identification of surface forms substrings of the original input that may be entity mentions
- **Candidates Selection:** Selecting a set of surface forms from step 1 along with the DBpedia resources that are candidate meanings for those surface forms
- **Disambiguation:** Deciding on the most likely candidate resource for each selected surface form
- **Filtering:** Adjusting the annotations to task-specific requirements according to user-provided configuration

More detail on DBpedia Spotlight implementation can be found in:

- Daiber, J., Jakob, M., Mendes, P.N. *Improving Efficiency and Accuracy in Multilingual Entity Extraction*. Proceedings of the 9th International Conference on Semantic Systems (I-Semantics), 2013.

## 5 Semantic Annotator Service

This section describes the prototype service detailing about its input, output and usage. The service is exposed through standard REST APIs behind a Basic Authentication<sup>20</sup> scheme. The service can be accessed at the URL:

- <https://sophia-cluster-dev.aws.celi.it/<LANGCODE>/spice/analysis>

<LANGCODE> is a path parameter and it is used to specify the language content, the supported values are: **en, es, fi, it, he**

### 5.1 Service Input

The service can be accessed with:

- GET requests: accepting “content” as the only parameter
- POST requests: accepting a json document as input, with the following properties:
  - content: mandatory - the textual contents to be analyzed
  - ns\_prefix: optional - the prefix used for representing the textual content in the JSON-LD response document, default value is “spice”
  - ns\_uri: optional - the URI of the ontology used for representing the textual contents in the JSON-LD document, default value is “https://w3id.org/spice/resource/”
  - collection: optional - a textual label representing the collection/museum/use case, default value is “spice”

An example service request, using curl<sup>21</sup>:

- `curl --user USR:PWD -XPOST https://sophia-cluster-dev.aws.celi.it/en/spice/analysis -d '{"content": "I love the Mona Lisa painting"}'`

<sup>16</sup> <https://research.google/pubs/pub49131/>

<sup>17</sup> <https://www.dbpedia-spotlight.org/>

<sup>18</sup> <https://sourceforge.net/projects/dbpedia-spotlight/>

<sup>19</sup> <https://github.com/dbpedia-spotlight/model-quickstarter>

<sup>20</sup> <https://tools.ietf.org/html/rfc7617>

<sup>21</sup> <https://curl.se/>

USR and PWD should be substituted with a real authentication

## 5.2 Service Output

The Semantic Annotator exposes the NLP pipeline analysis results as a JSON-LD<sup>22</sup> document. JSON-LD is a method of encoding linked data using JSON. Linked Data is structured data which is interlinked with other data so it becomes more useful through semantic queries. It builds upon standard Web technologies such as HTTP, RDF and URIs. More details on the Linked Data Hub designed and deployed by WP4 can be found in **D4.1 Linked Data server technology: requirements and initial prototype**.

The JSON-LD document contains two main sections:

- **Context:** detailing the ontologies used to describe data along with their prefix (used for compact notations in the graph section)
- **Graph:** containing a set of RDF triples represented as JSON objects; in our case the textual contents along with some metadata, followed by a set of annotations referencing the textual spans that can be linked to an emotion, a sentiment value or an entity (within DBPedia knowledge graph)

The following picture represents the service output for the input: *"I love the Mona Lisa painting"*

---

<sup>22</sup> <https://json-ld.org/>

```

{
  "@context":{
    "spice":"https://w3id.org/spice/resource/",
    "owl":"http://www.w3.org/2002/07/owl#",
    "dbr":"http://dbpedia.org/resource/",
    "earmark":"http://www.essepuntato.it/2008/12/earmark#",
    "xsd":"http://www.w3.org/2001/XMLSchema#",
    "rdfs":"http://www.w3.org/2000/01/rdf-schema#",
    "semiotics":"http://ontologydesignpatterns.org/cp/owl/semiotics.owl#",
    "emotion":"https://w3id.org/spice/SON/PlutchikEmotion/",
    "marl":"http://www.gsi.upm.es/ontologies/marl/ns#"
  },
  "@graph":[
    {
      "@id":"spice:sa_1617608245103",
      "@type":"earmark:StringDocuverse",
      "@language":"en",
      "earmark:hasContent":"I love the Mona Lisa painting"
    },
    {
      "@id":"ex:anno_1_emotion_2-6",
      "@type":"earmark:PointerRange",
      "rdfs:label":"love",
      "semiotics:denotes":{
        "@id":"ex:anno_1_emotion_2-6_love",
        "@type":"emotion:Love"
      },
      "earmark:refersTo":{
        "@id":"ex:docuverse"
      },
      "earmark:begins":{
        "@type":"xsd:nonNegativeInteger",
        "@value":2
      },
      "earmark:ends":{
        "@type":"xsd:nonNegativeInteger",
        "@value":6
      }
    }
  ],
  "@id":"ex:anno_2_sentiment_2-6",
  "@type":"earmark:PointerRange",
  "rdfs:label":"love",
  "semiotics:denotes":{
    "@id":"marl:Positive"
  },
  "earmark:refersTo":{
    "@id":"ex:docuverse"
  },
  "earmark:begins":{
    "@type":"xsd:nonNegativeInteger",
    "@value":2
  },
  "earmark:ends":{
    "@type":"xsd:nonNegativeInteger",
    "@value":6
  }
},
{
  "@id":"ex:anno_3_entity_11-20",
  "@type":"earmark:PointerRange",
  "rdfs:label":"Mona Lisa",
  "semiotics:denotes":{
    "@id":"dbr:Mona_Lisa",
    "@types":[
      "http://dbpedia.org/ontology/Work",
      "http://dbpedia.org/ontology/Artwork"
    ]
  },
  "earmark:refersTo":{
    "@id":"ex:docuverse"
  },
  "earmark:begins":{
    "@type":"xsd:nonNegativeInteger",
    "@value":11
  },
  "earmark:ends":{
    "@type":"xsd:nonNegativeInteger",
    "@value":20
  }
}
]
}

```

Figure 5. Example of JSON-LD response document from Semantic Annotator

The main element of the graph section contains a unique identifier of the textual contents and the content itself. The following *PointerRange* elements specify character offsets (with the properties *earmark:begins* and *earmark:ends*) that identifies an expression within the text, while the property *semiotic:denotes* contains the semantic connotation of the element along with its value and type.

### 5.3 Preliminary assessments

To verify the accuracy of the Semantic Annotator, a first test for Italian was carried out on an extract (40 sentences) of the “GamGame Lab UNITO” corpus (which can be found in the shared repository [at this link](#)). The corpus contains the results of the GamGame task presented to the students of the University of Turin (UNITO) during the seminar “Digital Humanities: Museum, Art, Emotion” on March 19, 2021. During the task, students had to choose three artworks from a list of twenty-six artworks of the GAM (Gallery of Modern Art of Turin) and for each of the works they had to answer to the following questions: “Now choose one or more emotions to associate with the work, or write an emotion yourself if it is not present in the list”; “What strikes me about this work?”; “What does this work remind me of?”; “How does this work make me feel?”; “Write tags that you think could be associated with this work”.

The results of the first test tend to be good, with the predicted emotions and sentiment correctly annotated, but some critical issues have arisen. In all cases (3 occurrences) in which an adjective indicating an emotion (e.g., “triste”, ‘sad’) is preceded by a negation, thus going to indicate the opposite emotion and sentiment (where “non triste”, ‘not sad’, indicates an emotion and a positive sentiment), the system does not recognize the rule and annotates the word with the emotion and sentiment with which it is labelled in the original lexicon (therefore “triste”, ‘sad’, with the emotion “sadness” and negative sentiment). This leads us to revise the rule on negation by which the Semantic Annotator was instructed.

An important critical point highlighted by the test concerns significant and recurring words in the context of emotions and sensations evoked in the artistic field, such as “curiosity” and “nostalgia”, which are not labelled by the SA with any emotion. This is due to the fact that in the original lexicon those words were not labelled with any emotion or polarity.

Similarly, some words are not annotated with important emotions, but with secondary ones, e.g., “solitudine”, ‘loneliness’, is labelled only with the emotion “fear”, but not with “sadness”. Finally, there is only one case in which, for the same adjective, in the grammatical difference between male and female gender, the emotions and sentiment labelled differ: the male adjective “amato”, ‘loved’, is annotated with positive sentiment and “love” emotion, while the female adjective “amata”, ‘loved’ is not labelled with any emotion or sentiment.

To resolve these issues, the future intention is therefore to test the entire “GamGame Lab UNITO” corpus for the Italian and the corpora of the other languages and case studies involved in SPICE to find the single recurring words in each language that the system does not annotate and then manually annotate them in the lexicons to improve their quality.

## 6 Conclusions and future works

In the next phases of the work, we expect to continue the revision of the lexicons, in order to make the resources for each language aligned and harmonious, creating a single multilingual emotional lexicon. For this purpose, on the one hand, the search for new linguistic sources will be carried out to assist the improvement of individual lexicons from a linguistic and emotion categorization point of view, and, on the other, the study on updates on emotional lexicons and emotion detection in the field of art perception will continue.

Furthermore, again with a view to expanding resources, multiword (expressions which are made up of at least 2 words and which can be syntactically and/or semantically idiosyncratic in nature; they act as a single unit) and idiomatic expressions will be added; a possible misalignment between languages is expected, especially with regard to idiomatic expressions, which will therefore require careful translation and linguistic confrontation with the partners involved in SPICE.

As for the lexicons, the revision and improvement of the grammar rules underlying the system, especially regarding negations and conjunctions, will be carried out.

At the same time, other tests on the Semantic Annotator will be performed with corpora from each of the SPICE languages, in order to test the accuracy of the system in a multilingual way and be able to improve its functioning. To make this possible, in addition to the collection of corpora produced by individual SPICE case studies, the intention is to create a multilingual corpus that contains user-generated content in all SPICE languages; to achieve this goal, we think about the development of a test to be submitted to users where, for each case study (and language), the user must choose between a number (tbd) of works (previously selected by the curators) from the relative museum and answer some questions related to each of the chosen works (e.g. "how does this work make you feel?"). In this way, it will be possible to collect material containing reflections with an emotional/descriptive background, which may then be analysed by the Semantic Annotator. However, the steps for creating the multilingual test and corpus are yet to be specifically defined.

## 7 References

Arva, H., Halénus, P., Herkevall, J., Lindblad, P., Rahlén, S., Tronde, J. E., May, W. M., 30 (2018). Emotional Emoji. Introducing the concept of emotion analysis for emoji.

Barbieri, F., Camacho-Collados, J., Ronzano, F., Espinosa-Anke, L., Ballesteros, M., Basile, V., Patti, V., Saggion, H. (2018). SemEval-2018 Task 2: Multilingual Emoji Prediction, *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pp. 24-33

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3), 436-465.  
*EmoLex*

Öhman, E. (Forthcoming). SELF and FEIL: Emotion Lexicons for Finnish.

Passaro, L., Pollacci, L., & Lenci, A. (2015). ItEM: A vector space model to bootstrap an italian emotive lexicon. In *Second Italian Conference on Computational Linguistics CLiC-it 2015* (pp. 215-220). Academia University Press.

Ronzano, F., Barbieri, F., Pamungkas, E. W., Patti, V., & Chiusaroli, F. 2018. Overview of the EVALITA 2018 Italian Emoji Prediction (ITAMoji) Task. In Caselli, T., Novielli, N., Patti, V., & Rosso, P. (Eds.), *EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples*. Torino: Accademia University Press. doi:10.4000/books.aaccademia.4460

Shoeb, A.A., Raji, S., & Melo, G.D. (2019). EmoTag - Towards an Emotion-Based Analysis of Emojis. *RANLP. Proceedings of Recent Advances in Natural Language Processing*, pages 1094–1103

Sprugnoli, R. (2020). MultiEmotions-it: A new dataset for opinion polarity and emotion analysis for Italian. In *7th Italian Conference on Computational Linguistics, CLiC-it 2020* (pp. 402-408). Accademia University Press.

Wolny, W. (2016). Emotion Analysis of Twitter Data That Use Emoticons and Emoji Ideograms. *ISD*.

Proctor, Mark, et al. "Drools documentation." *JBoss 5.05* (2008): 2008.

ZHANG, Yuan, and Qingguo XIA. "JAVA Rules Engines Based on Rete Algorithm [J]." *Science Technology and Engineering* 11 (2006).