UNIVERSITÀ DI PISA

Integration of Web-scraped Data in CPM Tools: the Case of Project SIBILLA

Alessio Bechini[1], Beatrice Lazzerini[1], Francesco Marcelloni[1], Alessandro Renda[1,2]

[1]University of Pisa, Dept. of Information Engineering, Pisa, Italy
[2]University of Florence, Dept. of Information Engineering, Florence, Italy

The SIBILLA project

*«Progettazione e sviluppo di un **SI**stema di **B**usiness Inte**LL**igence per **A**ziende Industria 4.0, con funzionalità di collaboration ed automatic interaction e di Big Data Analytics e machine learning per estrarre conoscenza e realizzare analisi predittive integrando big data acquisiti dal web e da architetture Internet of Things (IoT)»*

FABRICALAB co-make your business

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

TomorrowData

bsd

2014 2020 POR CreO REGIONE TOSCANA FONDO EUROPEO DI SVILUPPO REGIONALE

Regione Toscana

Design and development of a **Business Intelligence system** for Industry 4.0, with collaboration and automatic interaction functionalities, Big Data Analytics and machine learning to extract knowledge and perform predictive analysis by integrating **big data acquired from the Web** and from Internet of Things (IoT) architectures.

IN SUPREMÆ DIGNITATIS 1343

Motivation and Goals

Business Intelligence (BI) / Corporate Performance Management (CPM)
a tool that enables business *data* processing in order to generate *information*
that helps acquire *knowledge* useful in defining *strategies*

Traditional BI limitation:
it entirely relies on "ordinary" information gathered **within** the company domain

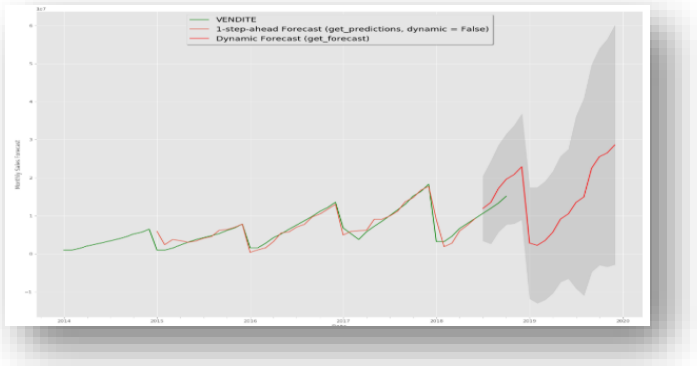
The SIBILLA challenge:
including other data sources in the company «information asset»:

- Web and Social Networks
- Internet of Things

Web Crawling and Data Mining (WCDM) module

Requirements:

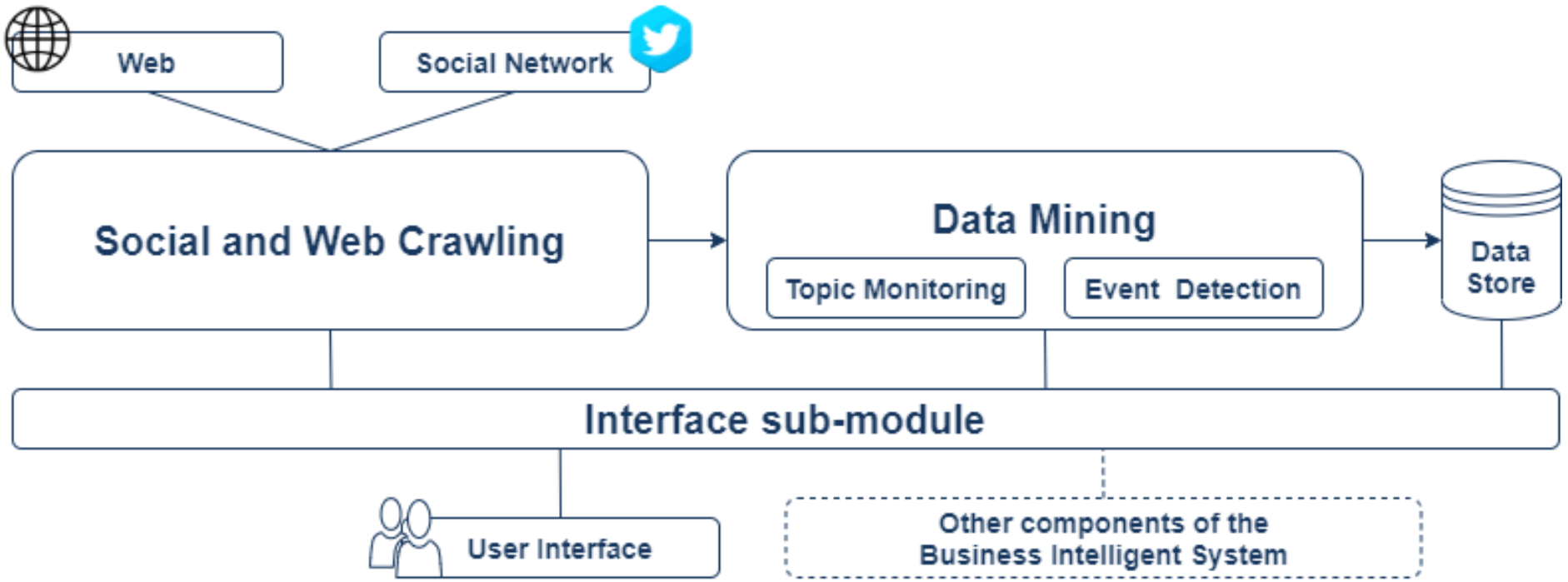
- To collect **large amount of data,** possibly adapting to the streaming rate
- To detect **real-world events** in real-time: handling such events can represent a competitive business advantage
- To obtain scores of **Sentiment and Opinion** for user's brand perception



Final Goal:

- Information enrichment to improve the performance of **predictive algorithms**

WCDM Module Architecture



Event Detection Sub-Module

Knowledge Extraction from Web Data

Step 1

Configuration

Language Setting
Seed Urls specification
Classes of Events specification

Step 2

Web Scraping

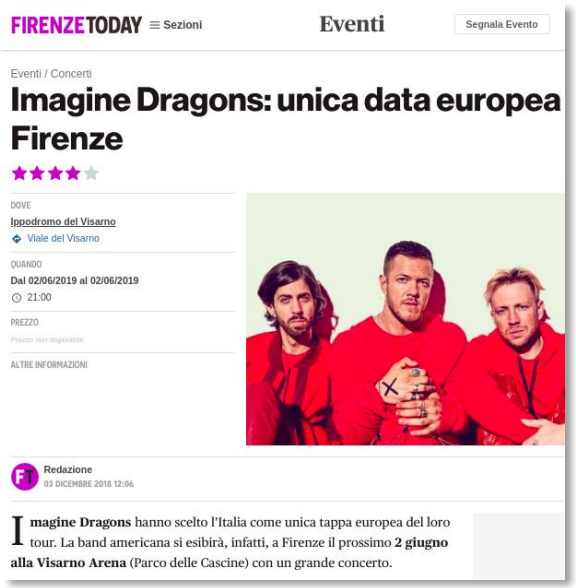
RSS feed - Extraction of
- Text
- Data
- Metadata

Step 3

Text mining

Identification of
Event Object Attributes
- Named Entity Recognition
- Regular Expression
- Keyword Extraction

Event Detection Sub-Module
Knowledge Extraction from Web Data



FIRENZETODAY ≡ Sezioni Eventi Segnala Evento

Eventi / Concerti

Imagine Dragons: unica data europea Firenze

★★★★☆

DOVE
Ippodromo del Visarno
Viale del Visarno

QUANDO
Dal 02/06/2019 al 02/06/2019
21:00

PREZZO
Prezzo non disponibile

ALTRE INFORMAZIONI

Redazione
03 DICEMBRE 2018 12:06

Imagine Dragons hanno scelto l'Italia come unica tappa europea del loro tour. La band americana si esibirà, infatti, a Firenze il prossimo **2 giugno alla Visarno Arena** (Parco delle Cascine) con un grande concerto.

« Imagine Dragons hanno scelto l'Italia come unica tappa europea del loro tour. La band americana si esibirà, infatti, a Firenze il prossimo 2 giugno alla Visarno Arena (Parco delle Cascine) con un grande concerto. Il loro ultimo album è uscito il 9 novembre con l'etichetta Universal Music il quarto album degli Imagine Dragons, intitolato "Origins" [...] »

- Location (based on NER)
- Date (based on Regex)
- Kind of Event (based on Keywords search)
- Miscellaneous entities (based on NER)

Topic Monitoring Sub-Module

Sentiment Analysis and Opinion Mining for Brand Reputation

Step 1
Configuration

Query parameters

Step 2
Data Collection

Twitter Streaming API

Step 3
Text Mining

Classification Task:
- Lexicon
- Supervised ML

Step 4
Alert mechanism

Negative Polarity?

Experimental Setup: Test on field

Task definition

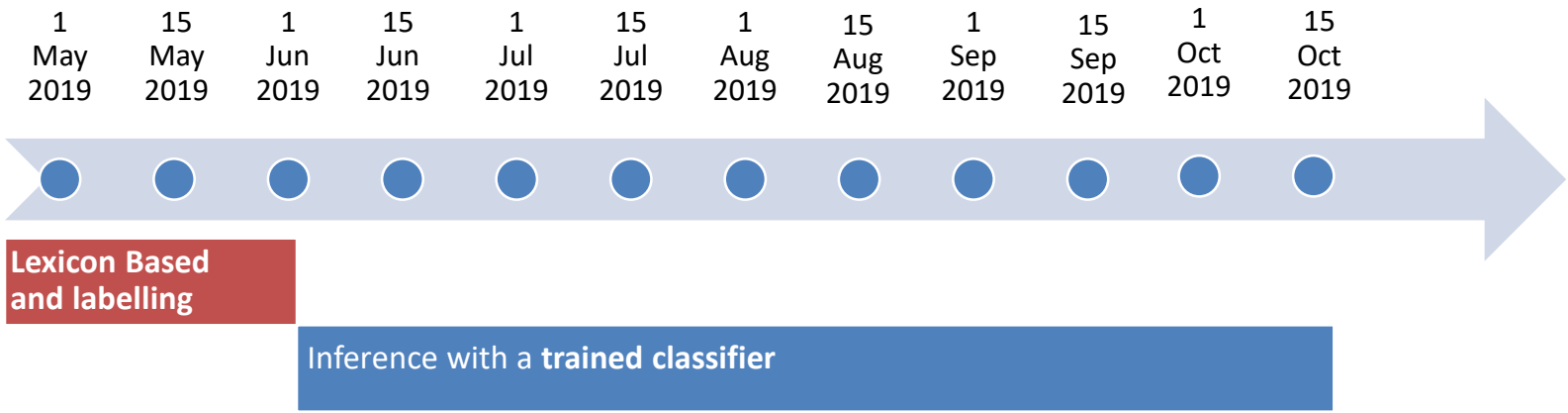
Final goal:

> *uncovering the Italian Twitter user's perception about Gucci company and, explicitly, about the* ***advertising campaign Gucci Tailoring Pre-Fall 2019****, with singer-songwriter* ***Harry Styles*** *in the role of Testimonial.*

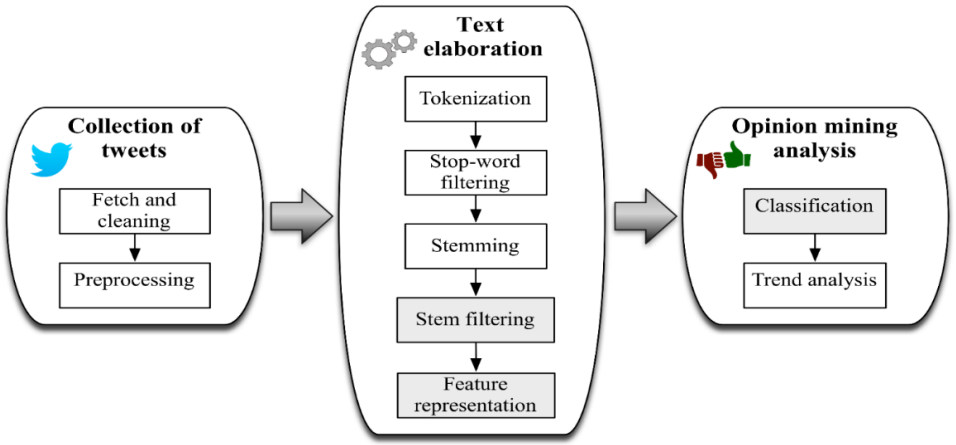
- Time window of interest: May 2019 – October 2019
- Query keywords: «harry gucci»

Experimental Setup: Test on field

Task definition



- *Cold start:* lexicon-based approach and data annotation
- Supervised Machine Learning approach: standard classification pipeline



Experimental Setup: Test on field

Experimental results

Lexicon-based method

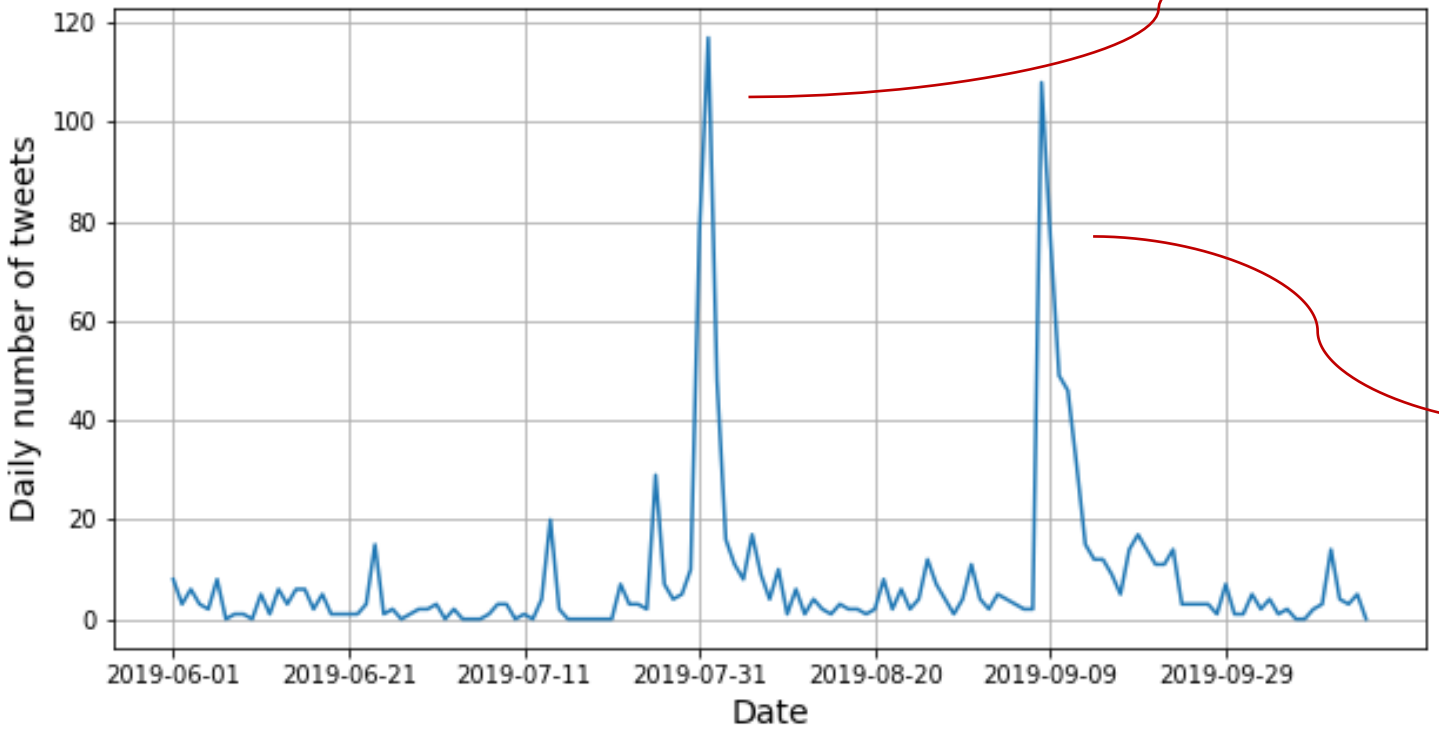
accuracy 0.535				
	precision	recall	f1-score	support
Neutral	0.60	0.43	0.50	121
Positive	0.76	0.57	0.66	242
Negative	0.17	0.62	0.27	37

Machine Learning method: Linear Support Vector Machine (LinearSVM)

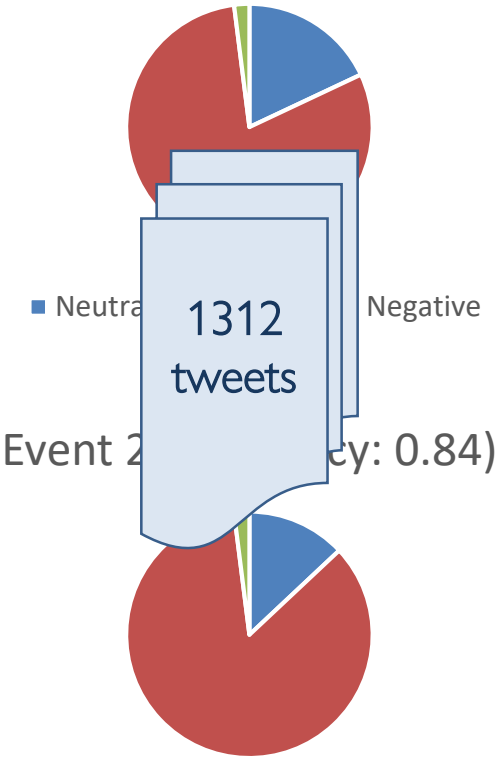
accuracy 0.677				
	precision	recall	f1-score	support
Neutral	0.72	0.77	0.74	121
Positive	0.67	0.60	0.61	242
Negative	0.21	0.11	0.14	37

Experimental Setup: Test on field

Inference with trained classifier



Event 1: (Accuracy: 0.7)



Event 2 ... cy: 0.84)

Event 1: piece of news related to advertising campaign

Event 2: broadcasting of the advertising on TV

Conclusions

- Participation in SIBILLA regional project:
  Design and development of a software module
  for the **enhancement of Business Intelligence tools functionalities**

- **Web Crawling and Data Mining Module:**
  - **Event Detection** functionality
  - **Topic Monitoring** functionality
  - Seamless integration with existing Business Intelligent system

- Experimental Campaign: monitoring Italian luxury fashion **brand reputation**

Thank you for your attention

Alessandro Renda

alessandro.renda@unifi.it

Ph.D. Student – Smart Computing

University of Pisa, Dept. Information Engineering

Computational Intelligence Group