TOWARDS SOCIAL MEDIA MINING: TWITTEROBSERVATORY

Inna Novalija, Miha Papler, Dunja Mladenić

Artificial Intelligence Laboratory

Jožef Stefan Institute

Introduction

- **Goal** of the **Social Media Mining** is data mining of **content** streams **produced by people** through **interaction** via Internet based applications.

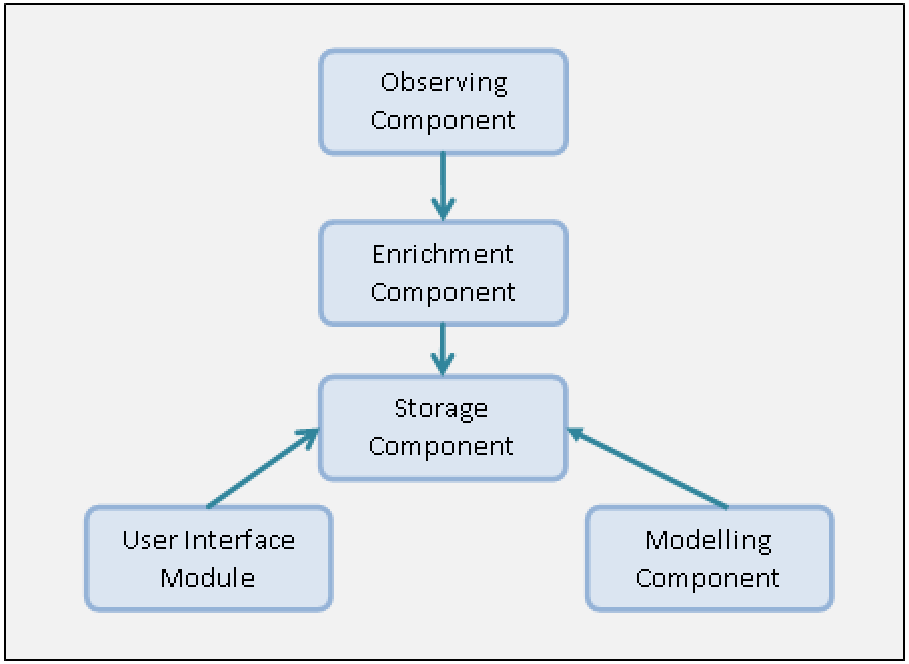
- **Social media mining** is usually associated with **noisy, distributed, unstructured and dynamic data**, as well as with informal text processing.

- In this research we introduce a novel **Social Media Mining Pipeline** and **TwitterObservatory** tool for:

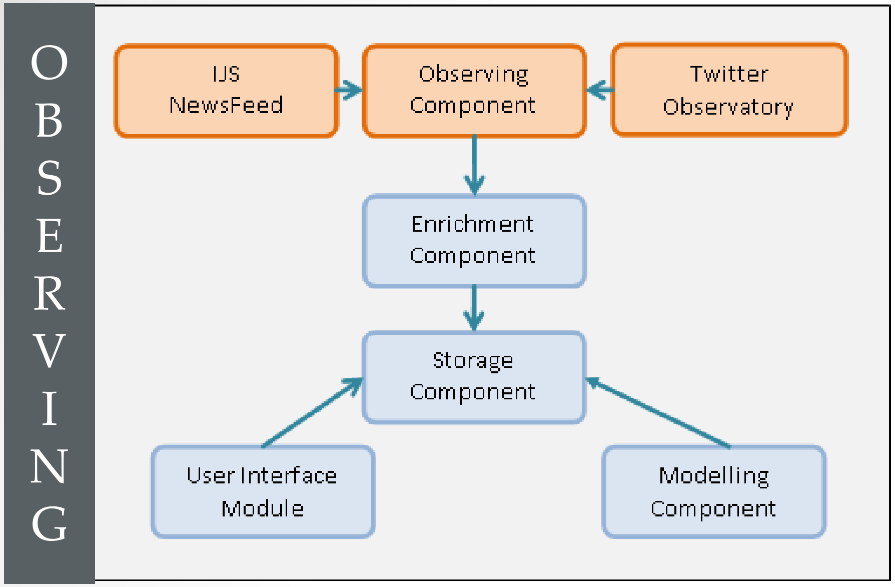
  observing,
  enriching,
  storing,
  analyzing and
  presenting information

obtained from social media and in particular, from Twitter.

Social Media Mining Pipeline



Social Media Mining Pipeline



OBSERVING

IJS NewsFeed

Observing Component

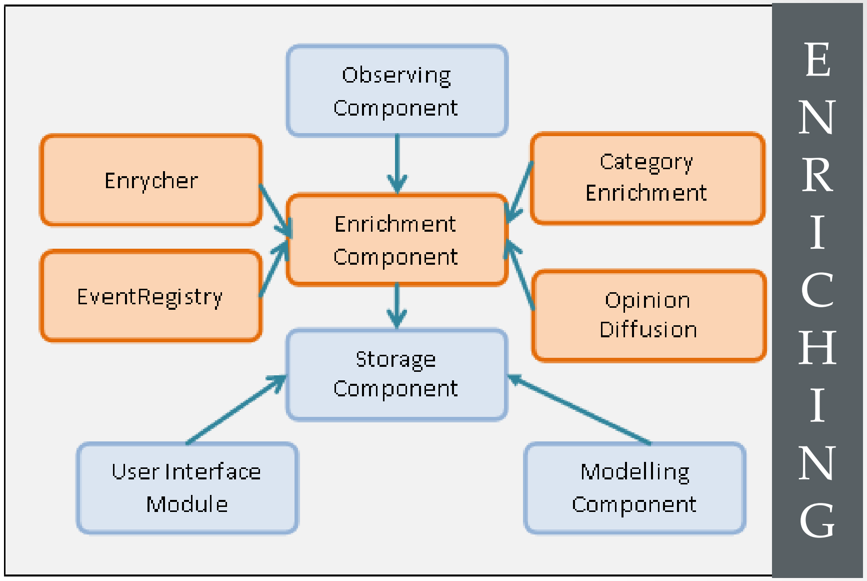
Twitter Observatory

Enrichment Component

Storage Component

User Interface Module

Modelling Component



ENRICHING

Observing Component

Enrycher

EventRegistry

Enrichment Component

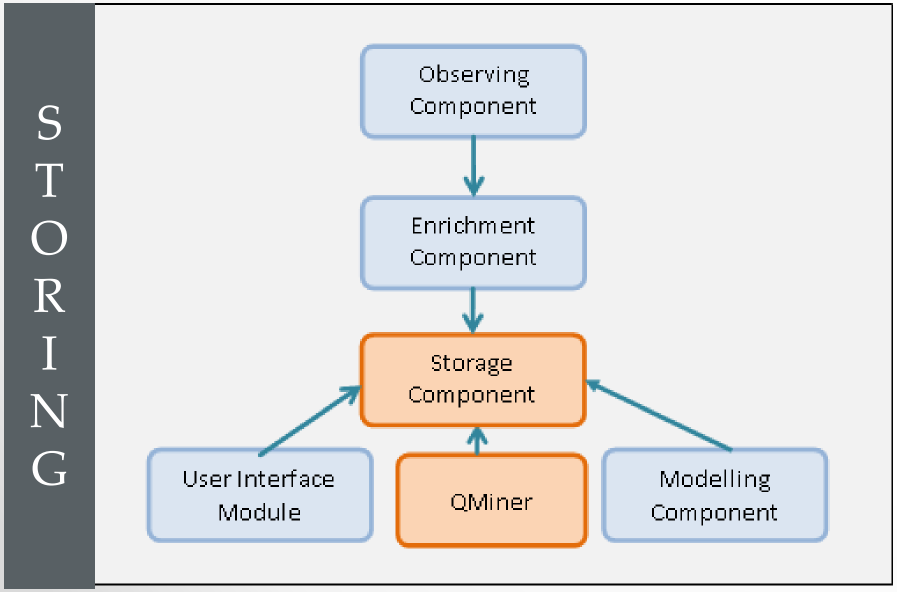
Category Enrichment

Opinion Diffusion

Storage Component

User Interface Module

Modelling Component



STORING

Observing Component

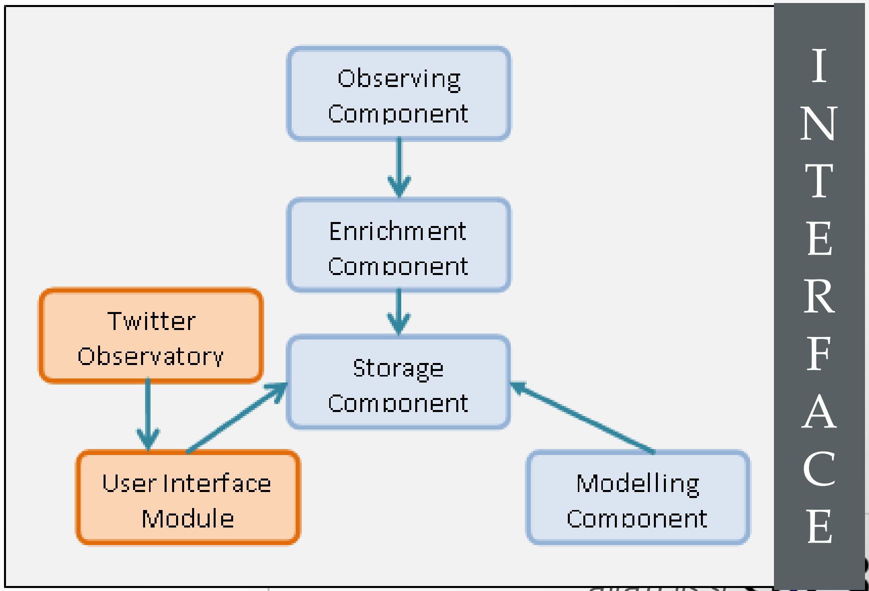
Enrichment Component

Storage Component

User Interface Module

QMiner

Modelling Component



INTERFACE

Observing Component

Enrichment Component

Twitter Observatory

Storage Component

User Interface Module

Modelling Component

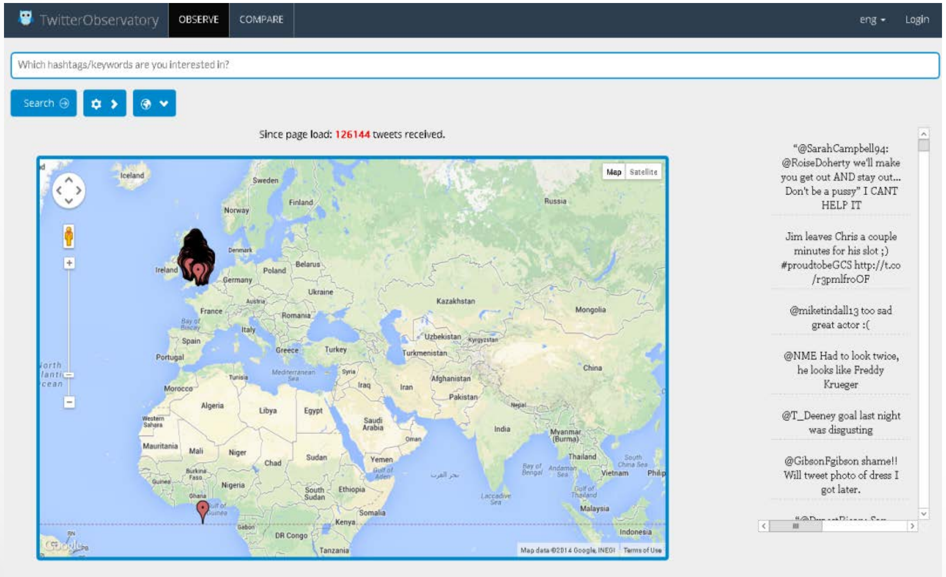
Observing

- **OBSERVING SOCIAL MEDIA by LOCATION:**
  - Geo coordinates from United Kingdom
  - 10 largest cities (by population)

Location	Number of Tweets (archived)
United Kingdom	31 GB

- **OBSERVING SOCIAL MEDIA by KEYWORDS:**
  - 400 most common words from Wikipedia,

Keywords	Number of Tweets (archived)
Common words	500 GB



Since page load: 126144 tweets received.

"@SarahCampbell94: @RoiseDoherty we'll make you get out AND stay out... Don't be a pussy" I CANT HELP IT

Jim leaves Chris a couple minutes for his slot ;) #proudtobeGCS http://t.co/r3pmlfroOF

@miketindall13 too sad great actor :(

@NME Had to look twice, he looks like Freddy Krueger

@T_Deeney goal last night was disgusting

@GibsonFgibson shame!! Will tweet photo of dress I got later.

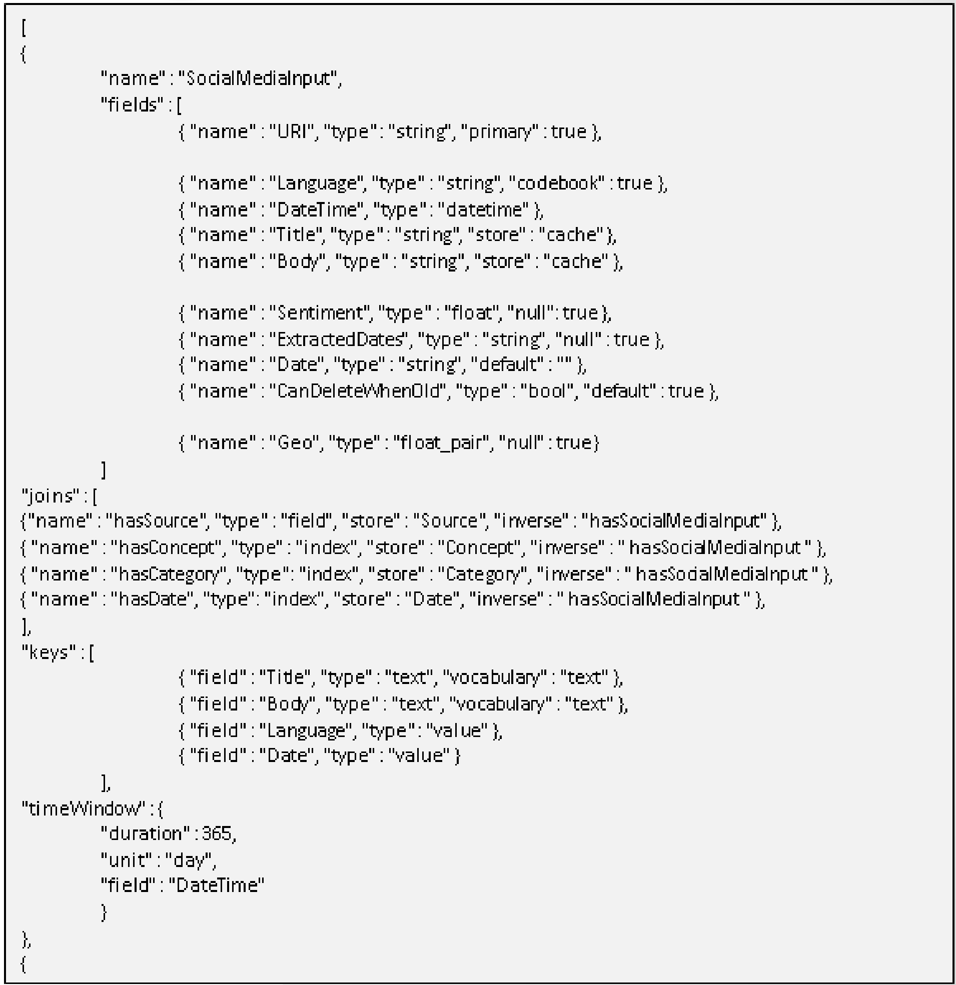
Enriching

- **DATA ENRICHMENT with ENRYCHER for TwitterObservatory:**
  - **Enrycher** is a service-oriented system that aims to **shallow and deep text processing** functionalities.
  - **Enrycher** processing **functionalities** include:
    - **Topic and keyword detection**
    - **named entity extraction**: names of people, locations and organizations, dates, percentages and money amounts
    - **sentiment** enrichment (English language)
    - Etc.
- **CATEGORY ENRICHMENT with XLING for TwitterObservatory:**
    - **Cross-lingual** DMOZ categorization

Storing

- **QMINER** as storing and analytics platform **for TwitterObservatory**.

- **QMiner** is a **data analytics platform** for **streams of structured and unstructured** data that at the same time contains a number of techniques for **supervised, unsupervised and active learning** on streams of data.

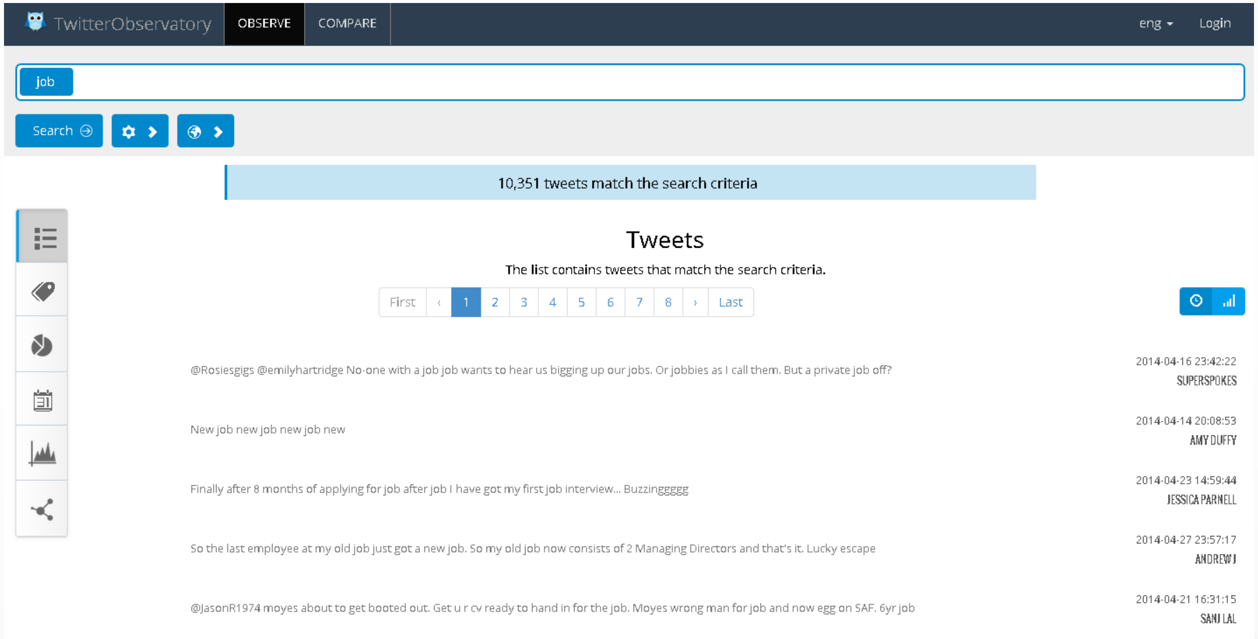


```
[
{
        "name" : "SocialMediaInput",
        "fields" : [
                { "name" : "URI", "type": "string", "primary" : true },

                { "name" : "Language", "type" : "string", "codebook" : true },
                { "name" : "DateTime", "type" : "datetime" },
                { "name" : "Title", "type": "string", "store" : "cache"},
                { "name" : "Body", "type" : "string", "store" : "cache" },

                { "name" : "Sentiment", "type" : "float", "null": true },
                { "name" : "ExtractedDates", "type" : "string", "null" : true },
                { "name" : "Date", "type" : "string", "default" : "" },
                { "name" : "CanDeleteWhenOld", "type" : "bool", "default" : true },

                { "name" : "Geo", "type" : "float_pair", "null" : true}
        ]
"joins" : [
{"name" : "hasSource", "type" : "field", "store" : "Source", "inverse" : "hasSocialMediaInput" },
{ "name" : "hasConcept", "type" : "index", "store" : "Concept", "inverse" : " hasSocialMediaInput " },
{ "name" : "hasCategory", "type": "index", "store" : "Category", "inverse" : " hasSocialMediaInput " },
{ "name" : "hasDate", "type": "index", "store" : "Date", "inverse" : " hasSocialMediaInput " },
],
"keys" : [
                { "field" : "Title", "type" : "text", "vocabulary" : "text" },
                { "field" : "Body", "type" : "text", "vocabulary" : "text" },
                { "field" : "Language", "type" : "value" },
                { "field" : "Date", "type" : "value" }
        ],
"timeWindow" : {
        "duration" : 365,
        "unit" : "day",
        "field" : "DateTime"
        }
},
{
```

User Interface

- **TWITTEROBSERVATORY** provides a suitable user interface that allows user to:
  - **view** upcoming social media data (tweets),
  - **search** tweets by different queries and
  - **analyze** the search results within different dimensions.

User Interface: Observed Tweets



10,351 tweets match the search criteria

Tweets

The list contains tweets that match the search criteria.

@Rosiesgigs @emilyhartridge No-one with a job job wants to hear us bigging up our jobs. Or jobbies as I call them. But a private job off?
2014-04-16 23:42:22
SUPERSPOKES

New job new job new job new
2014-04-14 20:08:53
AMY DUFFY

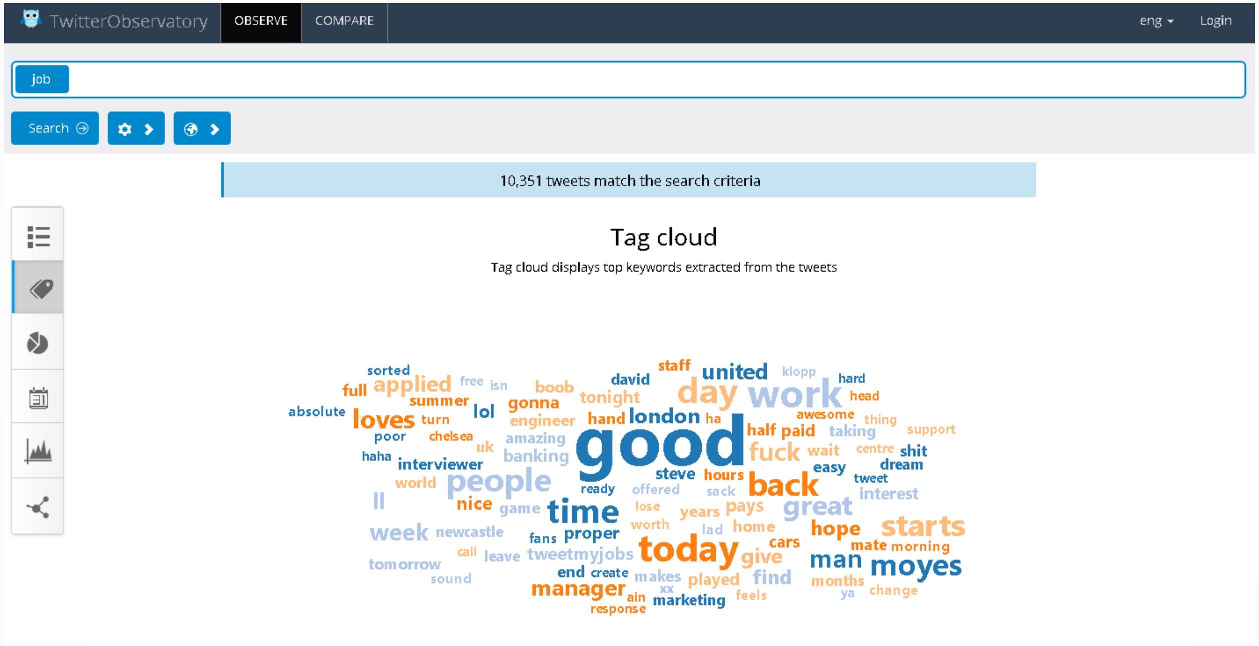
Finally after 8 months of applying for job after job I have got my first job interview... Buzzingggggg
2014-04-23 14:59:44
JESSICA PARNELL

So the last employee at my old job just got a new job. So my old job now consists of 2 Managing Directors and that's it. Lucky escape
2014-04-27 23:57:17
ANDREW J

@JasonR1974 moyes about to get booted out. Get u r cv ready to hand in for the job. Moyes wrong man for job and now egg on SAF. 6yr job
2014-04-21 16:31:15
SANJ LAL

Observed Tweets with Details (Filter: “job”)

User Interface: Tag Cloud



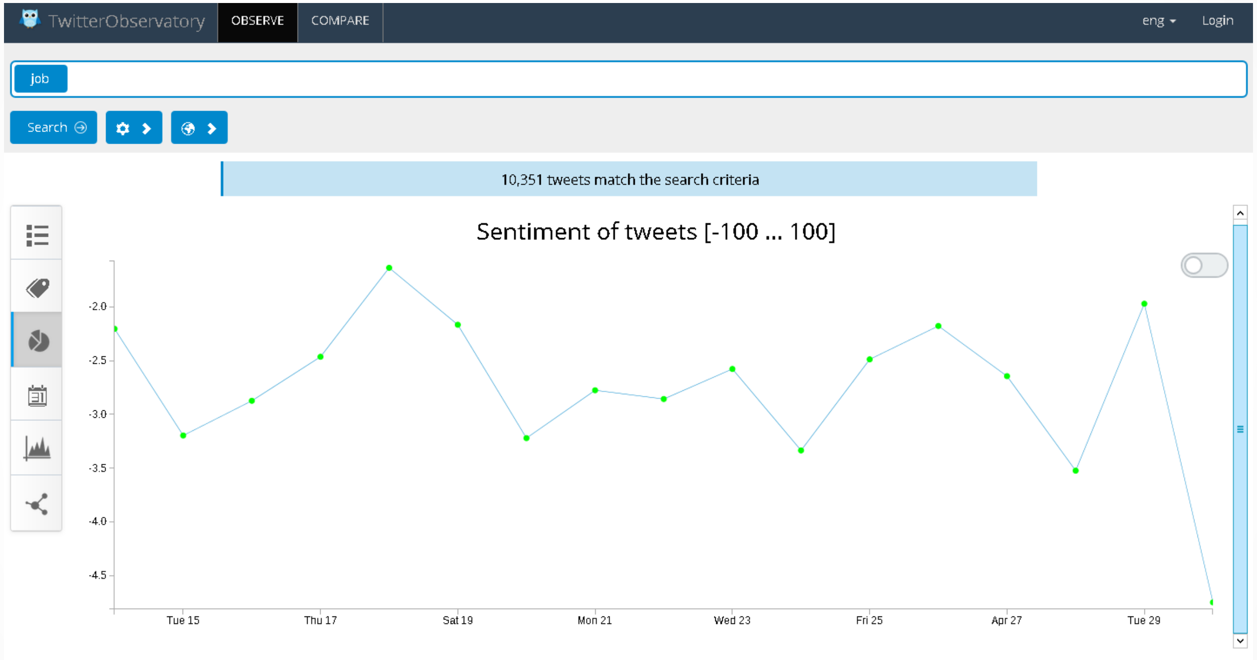
10,351 tweets match the search criteria

Tag cloud

Tag cloud displays top keywords extracted from the tweets

Tag Cloud for Tweets (Filter: "job")

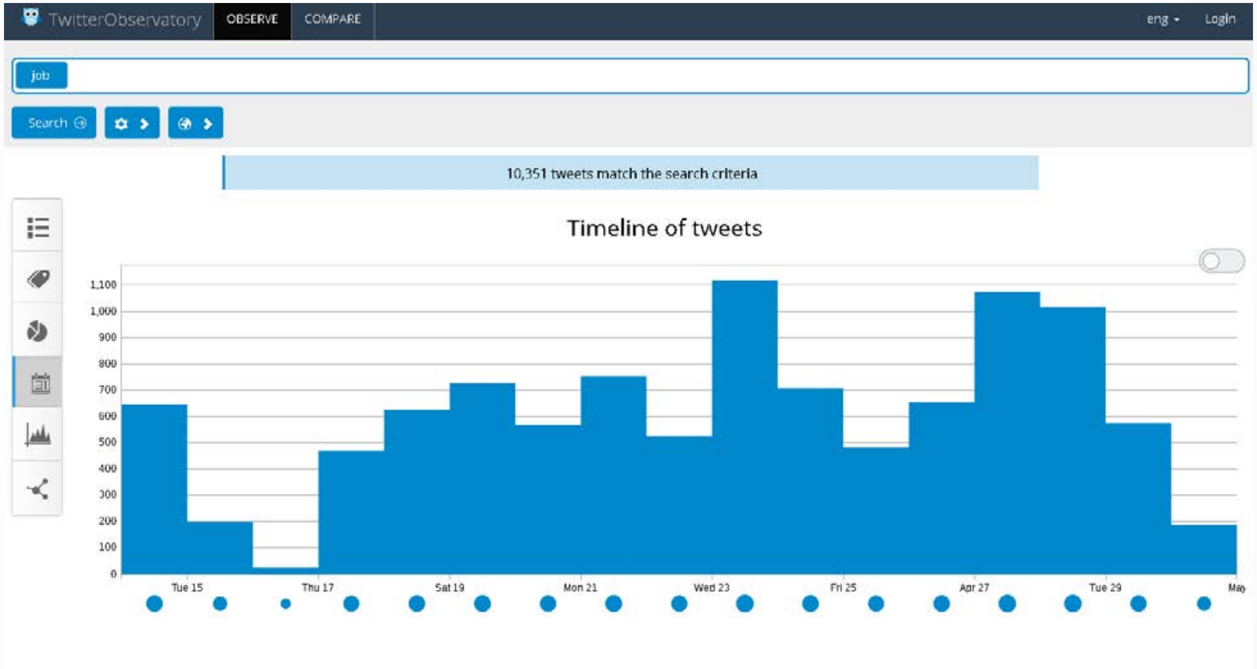
User Interface: Sentiment



10,351 tweets match the search criteria

Sentiment for Tweets (Filter: "job")

User Interface: Timeline



10,351 tweets match the search criteria

Tweets Timeline (Filter: "job")

Introduction to Data Modeling

- **MODELING AND NOWCASTING** functionalities are intended to connect social media with external datasets, such as macroeconomic data.

- **Goal** of modeling and nowcasting is to **relate micro-signals** coming from social media (such as micro-signals related to stocks, micro-signals related to labor, micro-signals related to consumers, micro-signals related to real estate and credit, micro-signals related to energy) with **macro-economic variables**.

- **First test** - on data such as NTSF indices and other stock indices relevant to regional based crawling of tweets.

- **Combined features** from social media - **correlated** with macroeconomic time series, with a number of operators for time series analysis used (**moving average** (MA), **exponential moving average** (EMA), etc.

Conclusion

- We presented an **approach for social media mining** based on a pipeline that implements observing, enriching, storing, modeling and presentation techniques.

- A novel tool **TwitterObservatory** that allows observing, searching, analyzing and presenting social media has been introduced.

- The developed **software components** enable **monitoring of social media stream including enrichment and storing of the data**.

- The **future work** will be based on implementing **additional functionalities for social media mining pipeline and on developing extensive modeling and nowcasting functionalities** for social media and external datasets.

Questions?