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 Twitter Observatory tool for:
  - observing,
  - enriching,
  - storing,
  - analyzing and
  - presenting information obtained from social media and in particular, from Twitter.
Social Media Mining Pipeline

- Observing Component
- Enrichment Component
- Storage Component
  - User Interface Module
  - Modelling Component
Social Media Mining Pipeline

**Observing**
- IJS NewsFeed
- Observing Component
- Enrichment Component
- Storage Component
- User Interface Module
- Modelling Component

**Enriching**
- Observing Component
- Enrichment Component
- Storage Component
- User Interface Module
- Modelling Component
- EventRegistry
- Category Enrichment
- Opinion Diffusion

**Storing**
- Observing Component
- Enrichment Component
- Storage Component
- User Interface Module
- Modelling Component
- QMiner

**Interface**
- Observing Component
- Enrichment Component
- Storage Component
- User Interface Module
- Modelling Component
- Twitter Observatory
Observing

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

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

- **DATA ENRICHMENT with ENRYCHER for Twitter Observatory:**
  - **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 Twitter Observatory:**
  - **Cross-lingual** DMOZ categorization
Storing

- QMiner as a storing and analytics platform for Twitter Observatory.

- 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.

```json
{
  "name": "SocialMediaInput",
  "fields": [
    {
      "name": "URL", "type": "string", "primary": true},
    {
      "name": "Language", "type": "string", "codebook": true},
    {
      "name": "DateTime", "type": "datetime"},
    {
      "name": "Title", "type": "string", "store": "cache"},
    {
      "name": "Body", "type": "string", "score": "cache"},
    {
      "name": "Sentiment", "type": "float", "null": true},
    {
      "name": "ExtractedDates", "type": "string", "null": true},
    {
      "name": "Date", "type": "string", "default": ""},
    {
      "name": "ContentMentions", "type": "bool", "default": 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": "text"},
    {
      "field": "Date", "type": "value"}
  ],
  "timeWindow": {
    "duration": 365,
    "unit": "day",
    "field": "DateTime"}
}
```
User Interface

- **TWITTER OBSERVATORY** provides a suitable user interface that allows the 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

Observed Tweets with Details (Filter: “job”)

@Raffael26: Emptilytridge: No one with a job job wants to hear us begging up our jobs. Or jobless as I call them. But a private job off?

New job new job new job new

Finally after 8 months of applying for job after job I have got my first job interview... Gaining more

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

@JasonRN74: Many about to be booted out. Get ur CV ready to hand in for the job. Monkey wrong way for job and now egg on SAF - bye job
User Interface: Tag Cloud

Tag Cloud for Tweets (Filter: "job")
User Interface: Sentiment

Sentiment for Tweets (Filter: "job")
User Interface: Timeline

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?