From Research to Applications: What can we Extract with Social Media Sensing?

Dr. Yiannis Kompatsiaris, ikom@iti.gr

CERTH – Information Technologies Institute, Director

Multimedia, Knowledge and Social Media Analytics Lab

Overview

- Introduction
    - Motivation – Challenges – Approaches
- Social Media mining for
    - Crisis management
    - Water management
    - Crime prediction, detection and prevention
    - Cultural and Architecture design applications
- Contributions – Support - Conclusions

http://petapixel.com/2013/03/14/a-starry-sea-of-cameras-at-the-unveiling-of-pope-francis/

Pope Benedict

2007: iPhone release

2008: Android release

2010: iPad release

Pope Francis

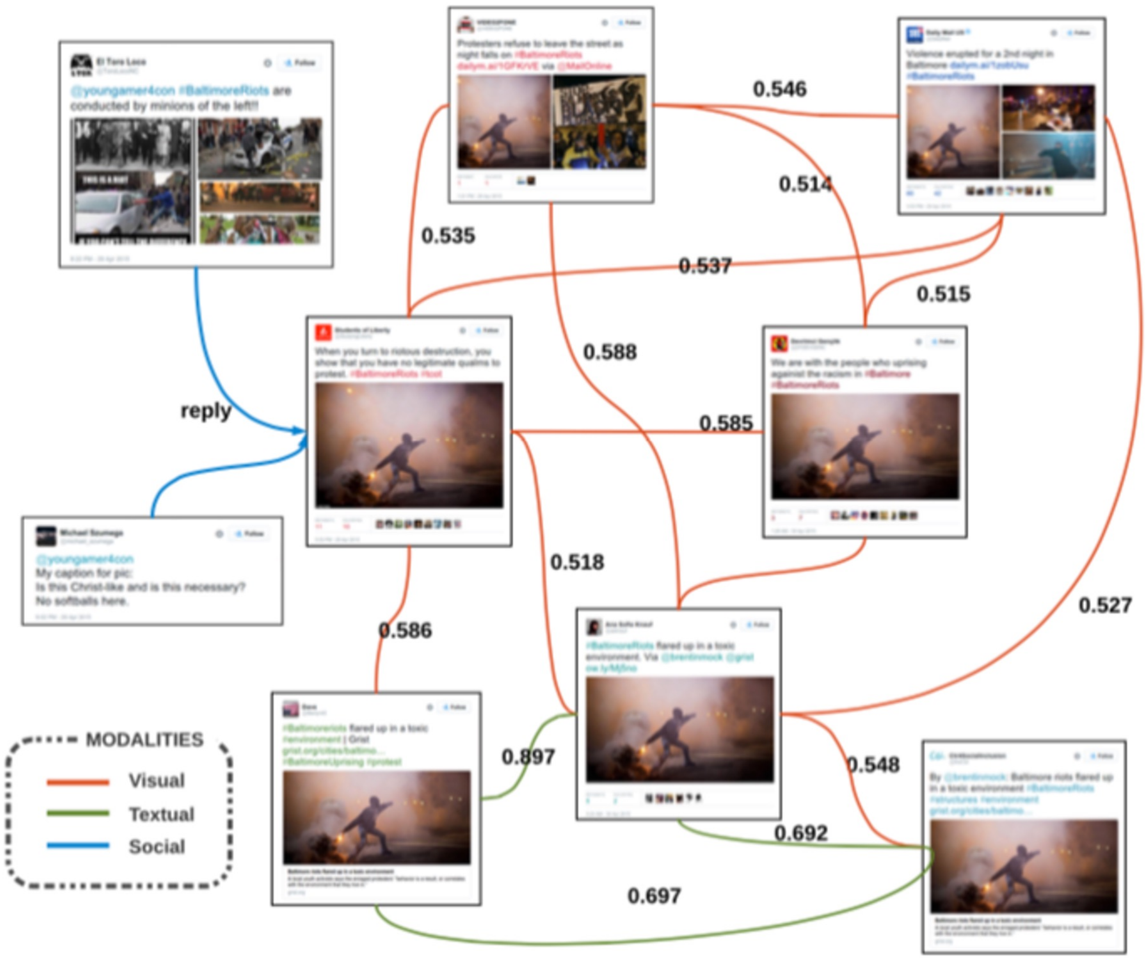
Hillary Clinton's Epic Group Selfie

Social Media as Real-Life Sensors

- Social Networks is a **data source** with an **extremely dynamic nature** that reflects events and the evolution of community focus (user's interests)
- Huge smartphones and mobile devices penetration provides **real-time and location-based** user feedback
- Transform **individually rare but collectively frequent** media to meaningful topics, events, points of interest, emotional states and social connections
- **Present** in an efficient way for a variety of applications (news, security (cyber and physical), marketing, science, health)

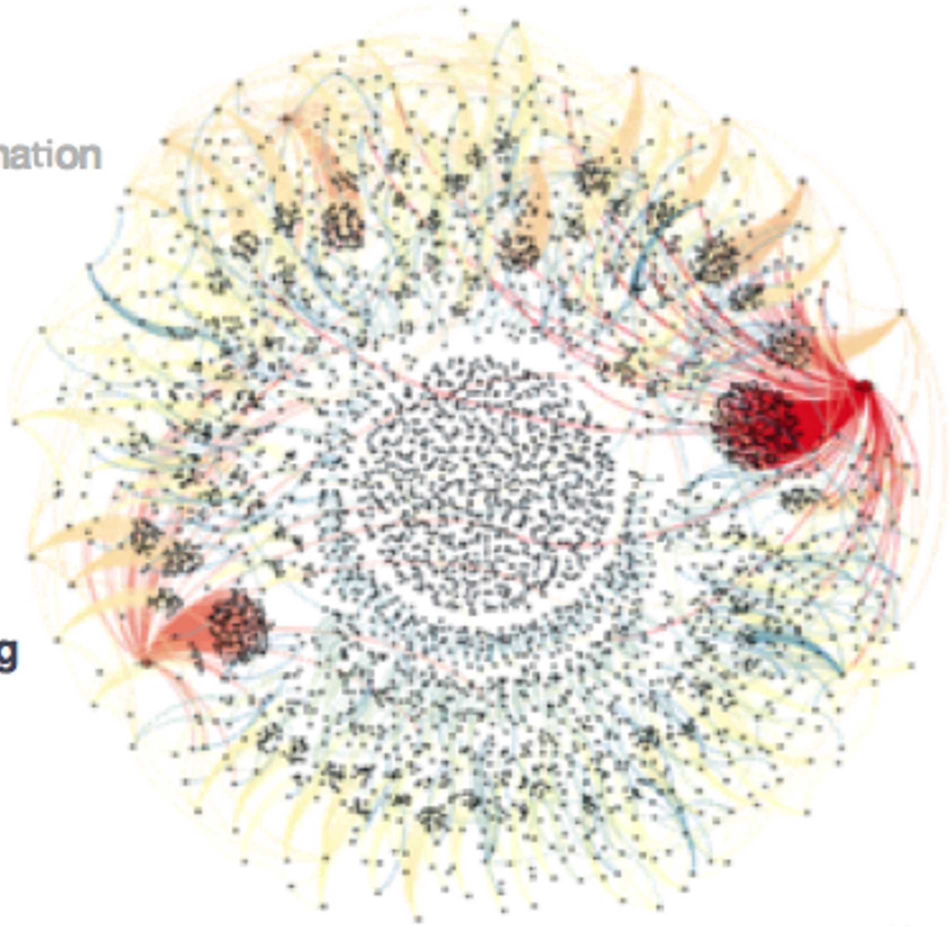
Social Media Aspects

Multi-Modal Social Media Graphs



Multi-Modal Social Media Graphs

announcement of Mubarak's resignation

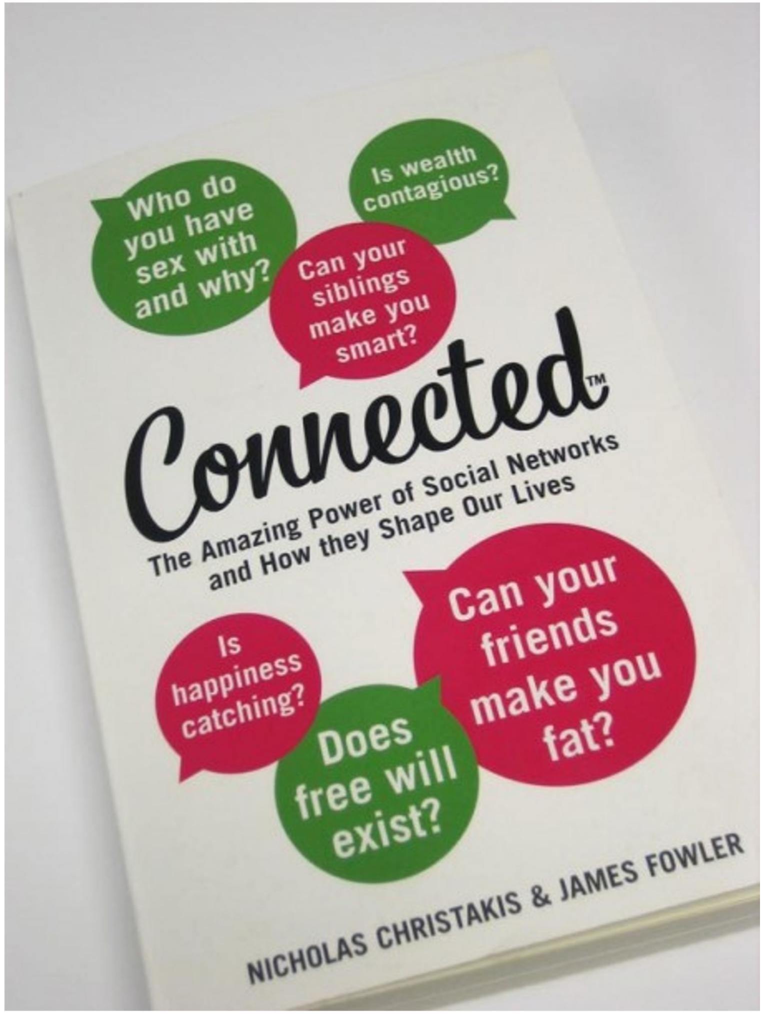


nodes=twitter users
edges=retweets on #jan25 hashtag

http://gephi.org/2011/the-egyptian-revolution-on-twitter/

Real-life Social Networks

- Social networks have **emergent properties**. Emergent properties are new attributes of a whole that arise from the interaction and interconnection of the parts
- Emotions, Health, Sexual relationships depend on our connections (e.g. number of them) and on our position - structure in the social graph
  - Central – Hub
  - Outlier
  - Transitivity (connections between friends)



Example – twitter and earthquakes

EPFL

idiap RESEARCH INSTITUTE

UNDERSTANDING EATING AND DRINKING IN CONTEXT FROM CROWDSOURCED DATA

Thanh-Trung Phan

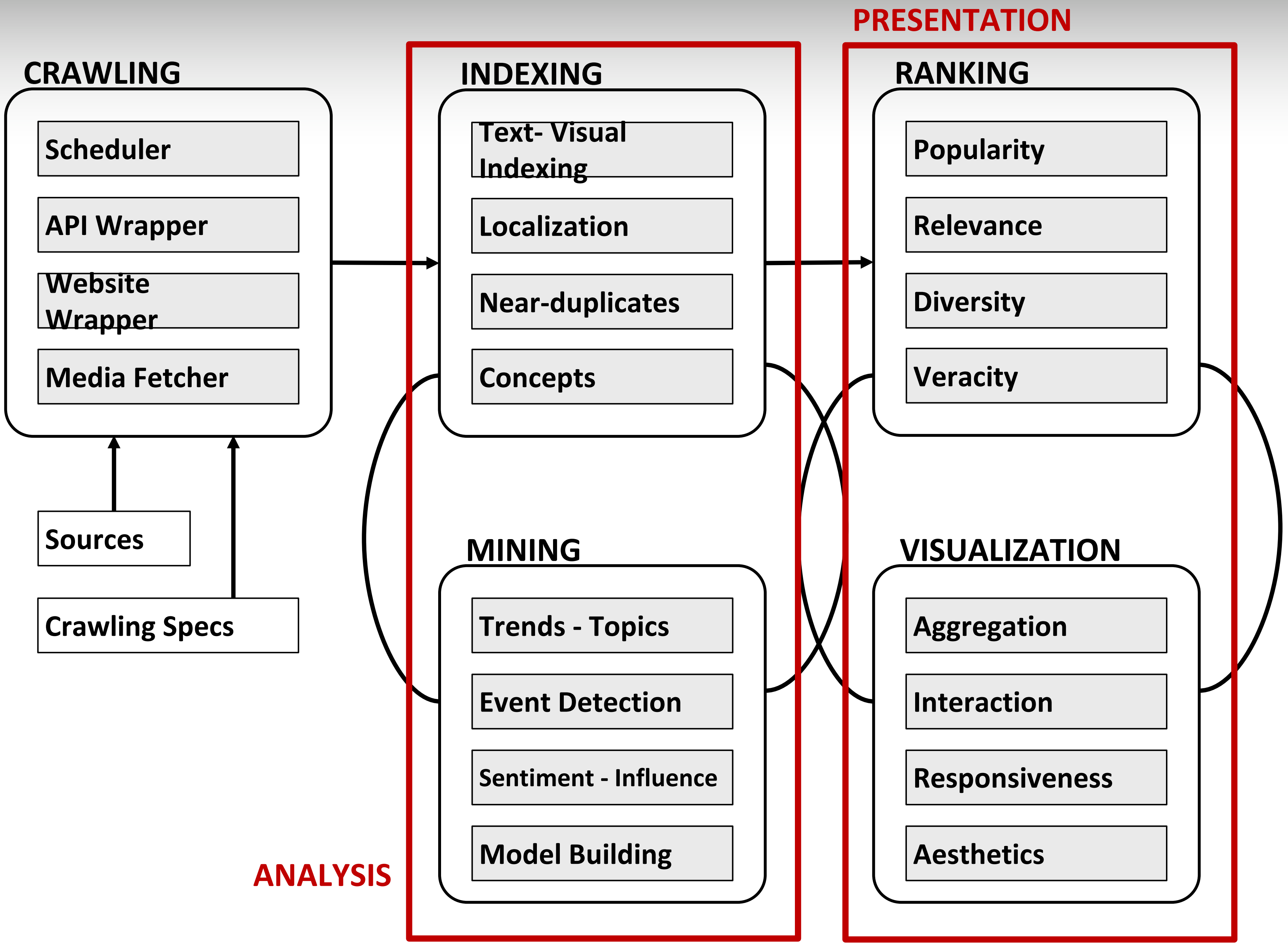
May 2020

École polytechnique fédérale de Lausanne

Technical Challenges (and opportunities)

- **Multi-modality**: e.g. image + tags, video, audio
- **Rich social context**: spatio-temporal, social connections, relations and social graph
- **Specific messages**: short, conversations, errors, no context, emoticons, abbreviations (OMG!)
- **Inconsistent quality**: noise, spam, fake, propaganda
- **Huge volume**: Massively produced and disseminated
- **Multi-source**: may be generated by different applications and user communities
- **Dynamic**: Fast updates, real-time

Overall Conceptual Architecture



NLP approaches

- Word2Vec & GloVe
  - **Word2Vec**: uses neural networks to train a predictive model
  - **FastText:Extension** of the continuous skip-gram model (Word2Vec) which takes into account subword information.

    Learns representations for character n-grams, and represents words as the sum of the n-grams vectors, thus taking into account morphology

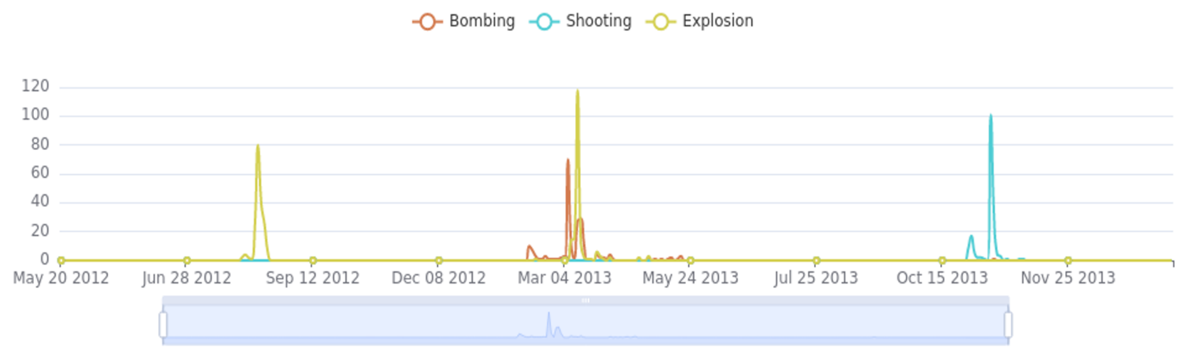
    Capable of computing word representations for words that did not appear in the training data as opposed to Word2Vec.
  - **GloVe**: considers statistical information for each word using a global co-occurrence matrix
  - Decent results in many NLP tasks
  - **Non-contextual word embeddings** => cannot distinguish between different meanings of the same word in a sentence
- ELMo & BERT
  - **ELMo**: the representation of each word depends on the **surrounding context**
  - **BERT**: use of transformers, i.e. an attention mechanism that learns contextual relationship between words in a text, also enables **parallelization**
  - **Contextualised word representations** => syntactic and semantic understanding of a text

Event Detection in Social Media

- Collection and analysis of content produced in social media platforms can play a vital role in almost real-time incident detection
- Highlighting and locating timely and valuable information and knowledge about **events** (e.g. floods, fire, bombings, etc) is more than necessary

Event Detection

Automatic identification of significant incidents through the analysis of social media data



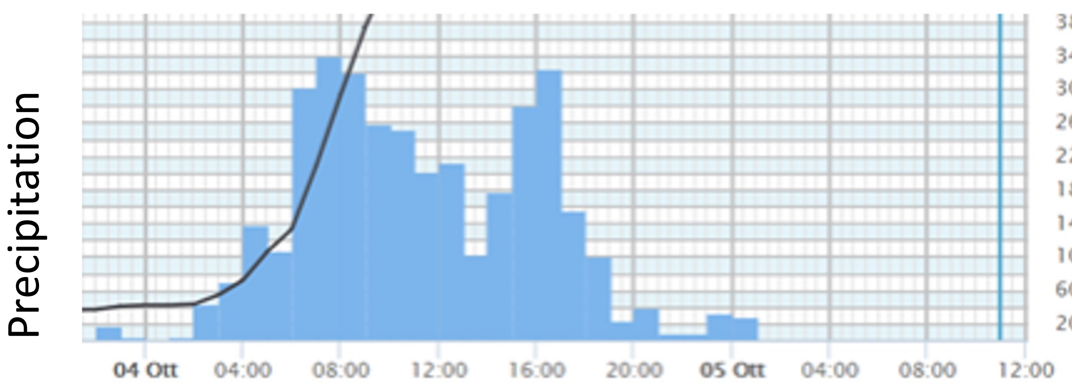
Event Detection

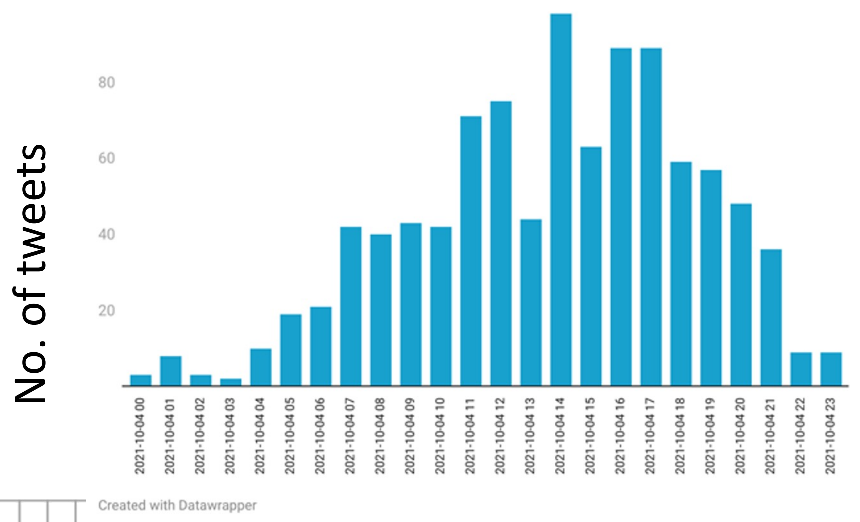
- **Statistical approaches:** STA/LTA (parametric), Z-Score (parametric), KDE (non-parametric)
  - Focus on number of posts in relation to time
  - No need for training data
  - Easy implementation
  - Need for thresholds
- **Graph-Based**: Community Detection
  - Focus on social connections - user behaviour (follow, mention, etc.)
  - No need for training data
  - Need for threshold (modularity)
- **Supervised-based**: Deep Neural Networks & Self-attention encoder
  - Need for training data
  - Able to capture complex events
    - Modeling of long-term dependencies in a sequence
    - Give more importance to some of the words in a sentence

Representativeness and coverage

- Specific categories might be **under-repesented** leading to **bias**
- E.g temporal, age representation

- Visual correlation between precipitation measurements & number of tweets posted on Oct 04, 2021 in the region of Liguria, Italy:

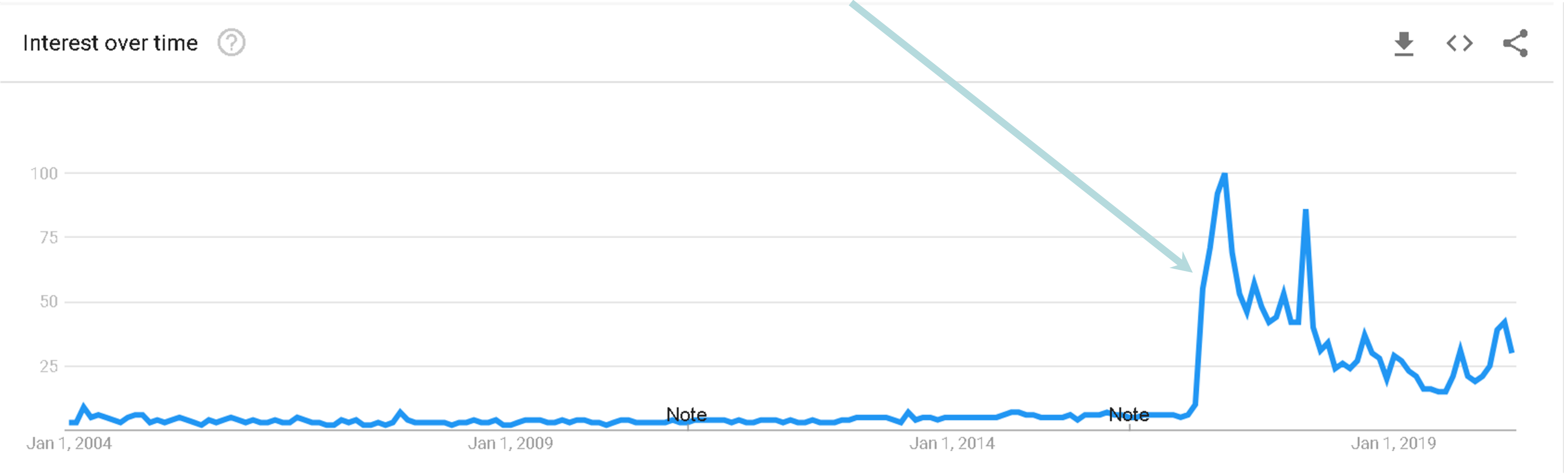




The Rise of Fake News

Volume for query “fake news” over time: A key milestone has been the US Elections in 2016, which marked the beginning of large-scale coordinated disinformation campaigns.

US Elections 2016

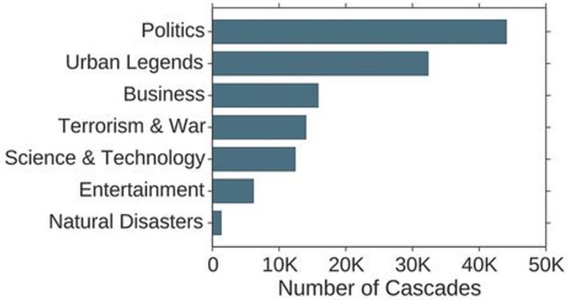


https://trends.google.com/trends/explore?date=all&geo=US&q=fake%20news

Misleading posts tend to spread faster and wider compared to accurate ones.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.

Topic frequency



Disaster management

Social media in disaster management

- Social media platforms have been proven to be a valuable source of information for early warning tools during a disaster

- Real-time collection of tweets about fires/earthquakes/flooding in order to detect events in time

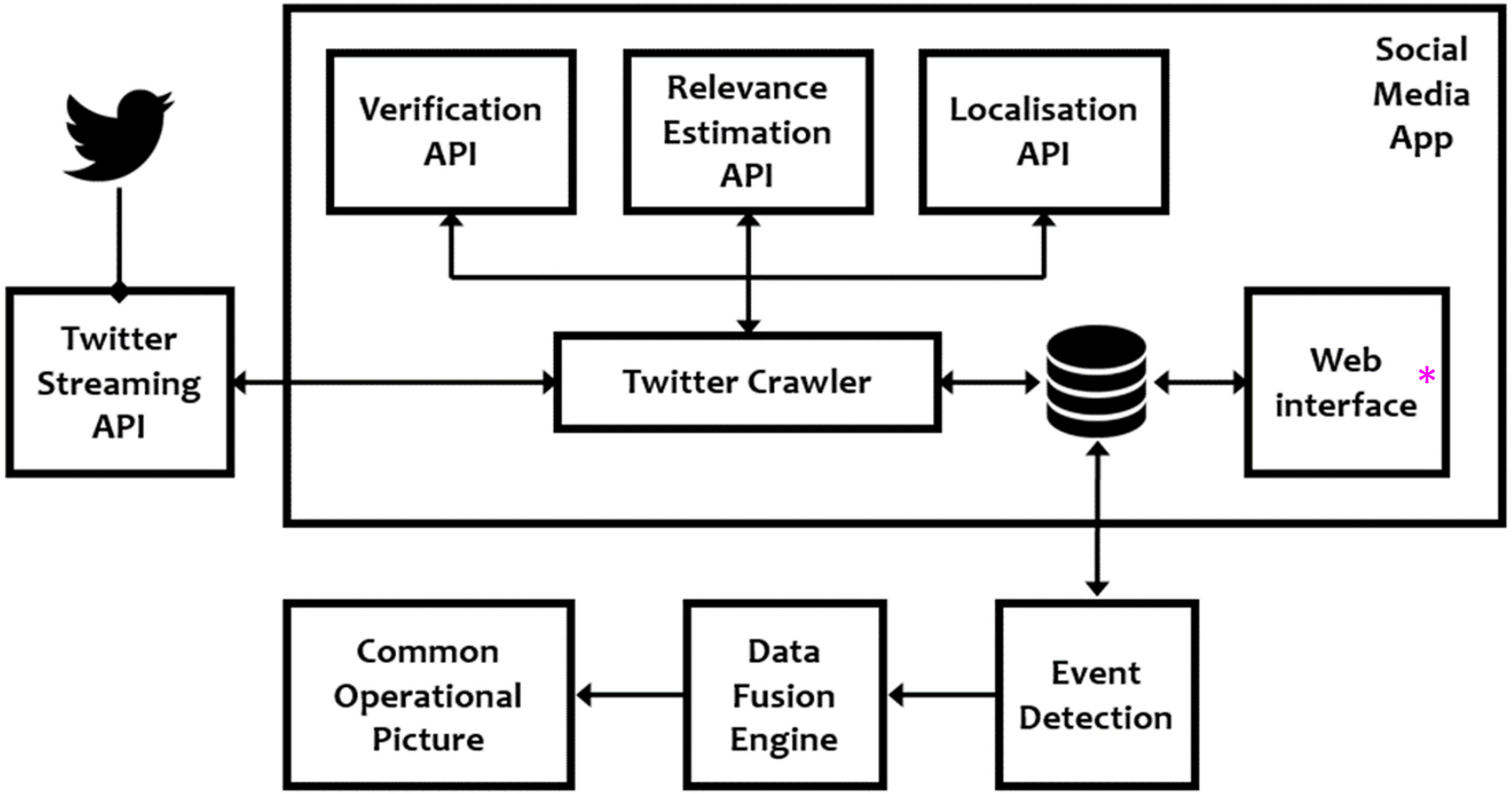
- Challenges:
  - Too much noise in Twitter (e.g. metaphorical use of incident-related words) → possible false warnings
  - Huge stream of single posts → more compact information is needed (**events**)

Posts during emergencies

fake

irrelevant

Overall specific framework

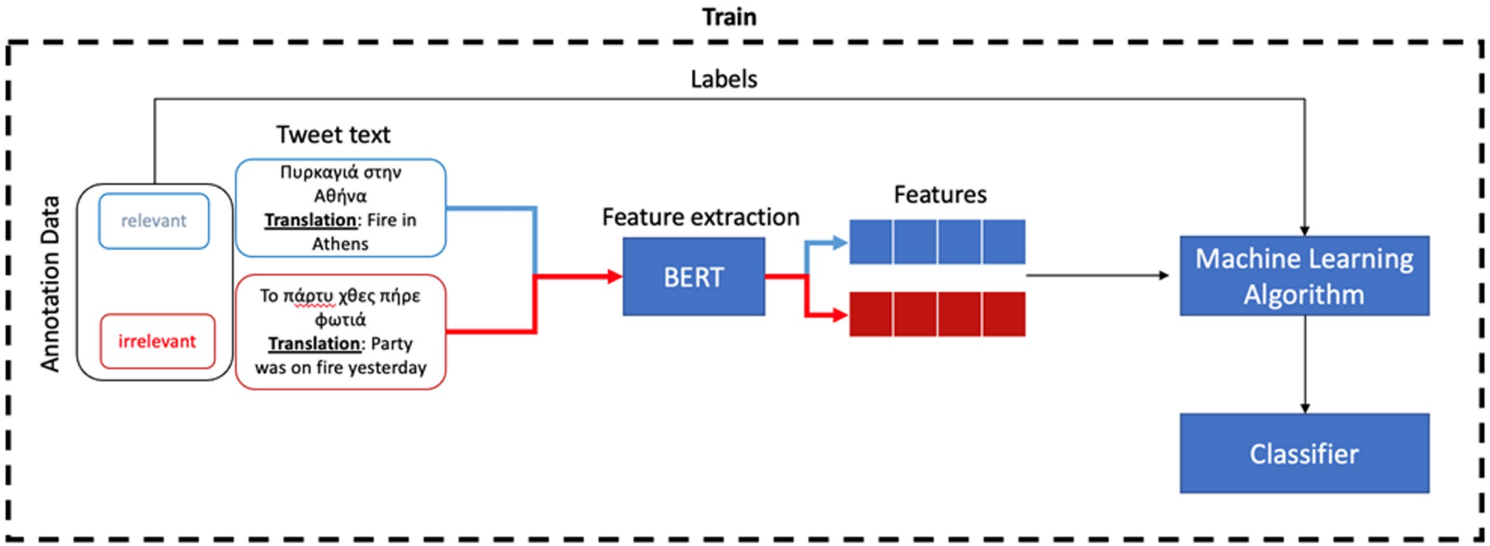


- **Twitter Crawler**: real-time tweet retrieval with keyword- or account-based search
- **Verification**: estimation of reliability score (real/fake)
- **Relevance Estimation**: filtering out the irrelevant tweets
- **Localisation**: detecting the locations mentioned in the text
- **Event detection**: producing warnings for potential events

* https://socialmedia-server-m4d.iti.gr/ingenious/relevancy-annotation.html

Relevance estimation

- Aim: train a model that will classify a new tweet as relevant or not to a disaster
- Training dataset: tweets about fires in Greek

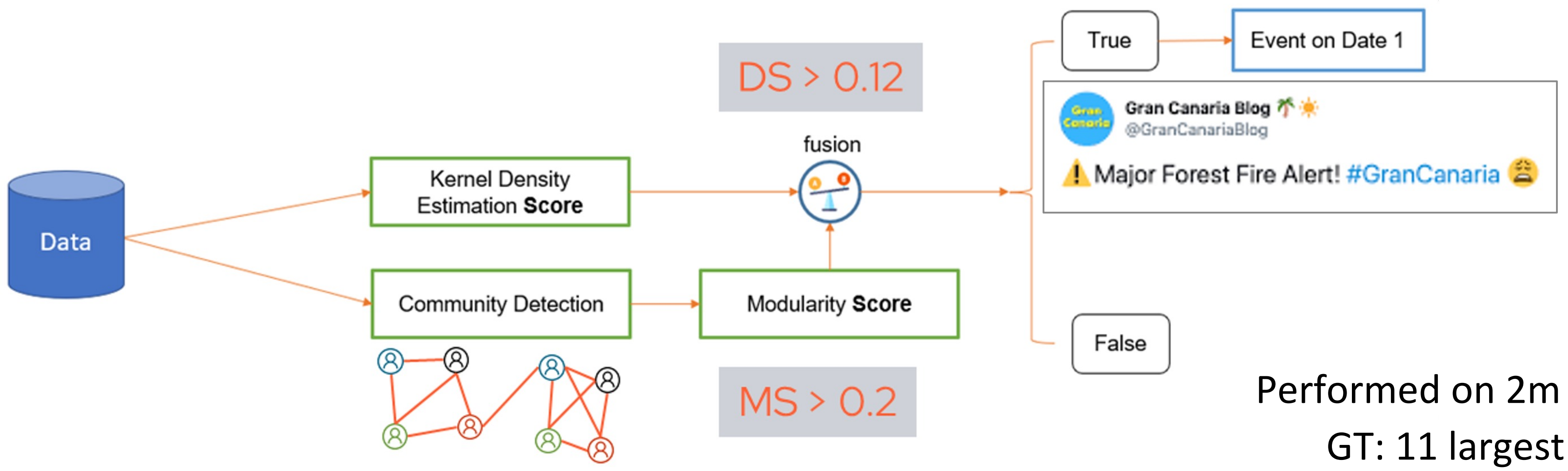


- Results for Binary Logistic Regression:

Accuracy	Recall	Precision	F1-score
0.8481	0.8036	0.8818	**0.8364**

Event detection

- Kernel Density Estimation (KDE): considers not only no. of tweets, but also sparsity & density when posted → Density score (DS)
- Community Detection (CD): discovers communities of Twitter users (graph representation) → Modularity score (MS)



Method	Accuracy
STA/LTA	0.7726
Z-score	0.8301
KDE	0.8986
KDE+CD	**0.9589**

Performed on 2m tweets about fires in Spain (2019)
GT: 11 largest fires reported by Copernicus EMS

T. Papadimos, N. Pantelidis, S. Andreadis, A. Bozas, I. Gialampoukidis, S. Vrochidis, and I. Kompatsiaris, "Real-time Alert Framework for Fire Incidents Using Multimodal Event Detection on Social Media Streams", 19th International Conference on Information Systems for Crisis Response and Management, 22-25 May 2022, Tarbes, France (accepted for publication).

Social media in creeping crisis

- *Sudden crisis*: natural or human-caused disasters that occur without warning
  - E.g. fires, earthquakes or terror attacks

- *Creeping crisis*: a threat to life-sustaining systems that evolves over time and space and is foreshadowed by precursor events
  - E.g. air quality or water quality, safety and security

Detection of water quality incidents

- Dataset: 212k English tweets that combine a water source (List A) & an issue (List B), collected during one year (Aug 1, 2020 - Jul 31, 2021)
    - 51k geotagged
- Examined methodologies: Z-score, STA/LTA, DBSCAN
- No ground truth (all water incidents are not known)

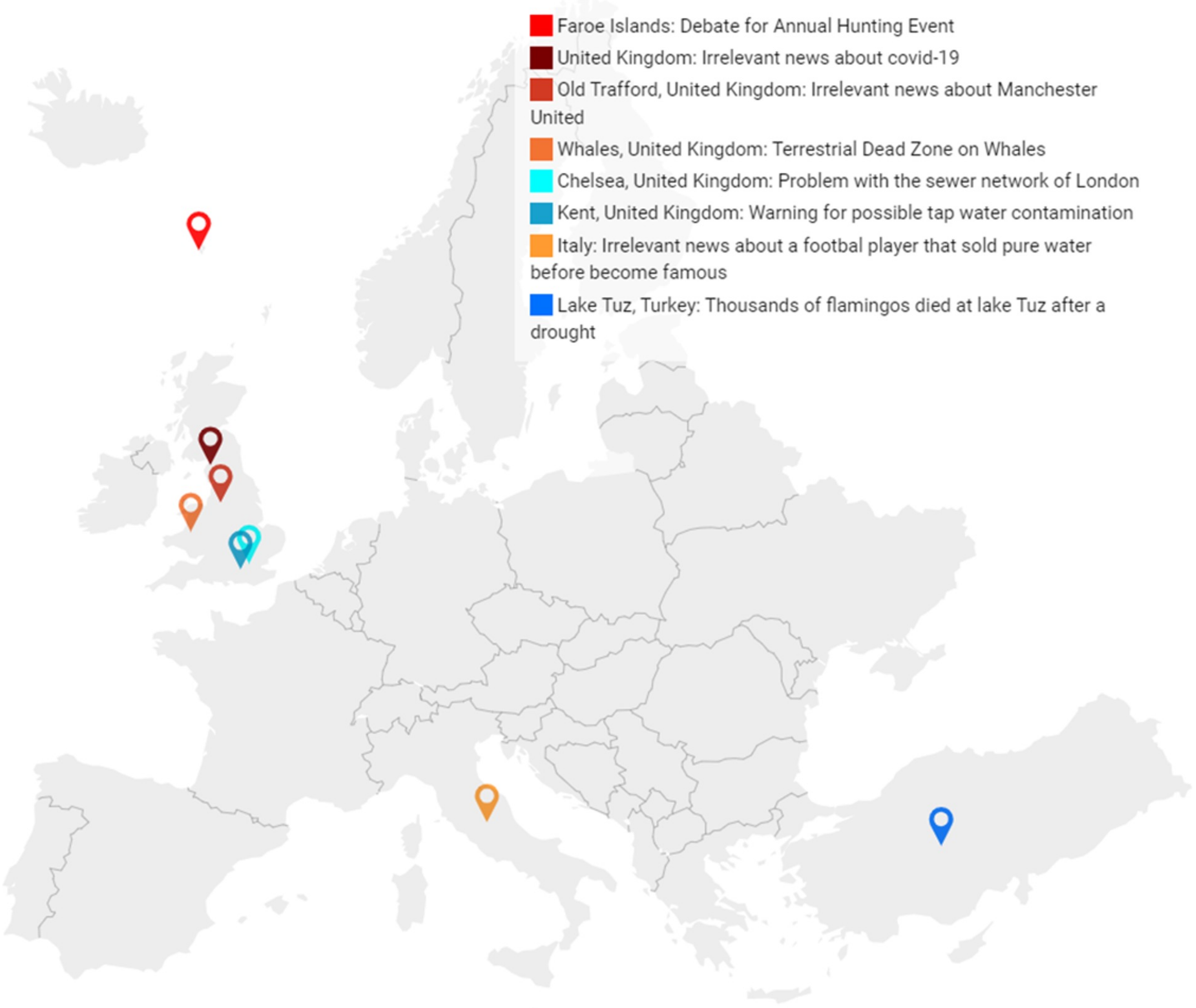
List A	List B
tap water	chlorine taste
water bottle	strange colour
aqueduct	bad odour
river	muddy
etc.	etc.

	TP	FP	Precision
Z-score	6	1	**0.86**
STA/LTA	2	3	0.4
DBSCAN	**8**	4	0.66

S. Andreadis, N. Pantelidis, I. Gialampoukidis, S. Vrochidis, and I. Kompatsiaris, "Water quality issues: Can we detect a creeping crisis with social media data?", 2022 IEEE Symposium on Computers and Communications (ISCC), 30 June - 3 July 2022, Rhodes, Greece (accepted for publication).

Detection of water quality incidents

- Some detected events
  - Example of relevant event: Possible tap water contamination in Kent
  - Example of not relevant event: Football player selling water



Detected events
Z-score

Incident	Date
✓ Gas and groundwater bubbling up on farmland near Chinchilla	2020-08-31
✓ Need for Biden's EPA to act quickly to undo the damage Trump caused	2020-09-01
✓ Water outages in Selangor	2020-09-04
❌ Miscellaneous noisy tweets	2021-02-09
✓ Water outages in Texas	2021-02-20
✓ Japan's approval of plan to release wastewater into ocean	2021-04-13
❌ Protest for "Global Myanmanr Spring Revolution"	2021-05-02
✓ Contamination of Okhchuchay River	2021-07-09

Detected events STA/LTA

Incident	Date
✓ Need for Biden's EPA to act quickly to undo the damage Trump caused	2020-09-01
✓ Water outages in Selangor	2020-09-04
✗ Miscellaneous noisy tweets	2020-09-05
✗ Miscellaneous noisy tweets	2020-09-06
✗ Miscellaneous noisy tweets	2020-09-07

Gwenash
@Kledangrange

Cops nab four brothers over pollution that led to Selangor water disruption”👍👍👍👍

malaymail.com
Cops nab four brothers over pollution that led to Selangor water disruption
KUALA LUMPUR, Sept 5 — Four men have been arrested in connection with the odour pollution incident in the Sungai Gong Industrial Area in Rawang. Selango...

5:37 AM · Sep 5, 2020 · Twitter for iPhone

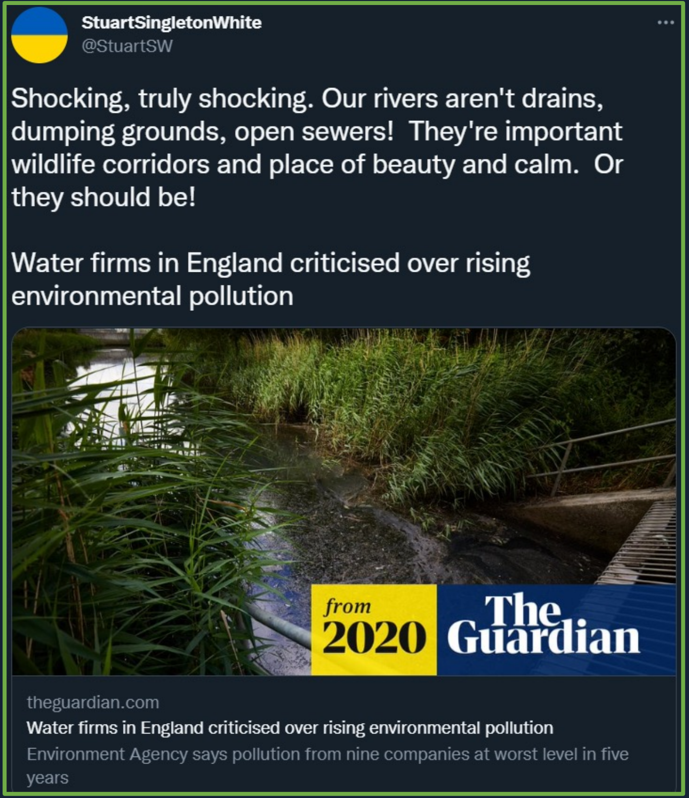
asyraf
@Asyrfmlek

4 days task for Air Selangor. Will they get to clear that 1 ton pollution at the raw water supply ? Hope so.

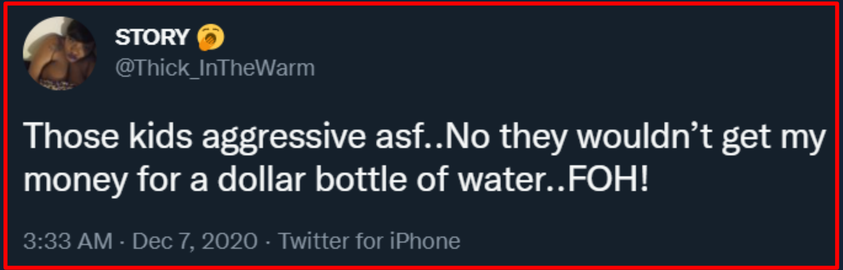
8:07 PM · Sep 4, 2020 · Twitter for Android

Detected events DBSCAN (1)

Incident	Date
✓ Criticism of water firms in England over rising environmental pollution	2020-10-02
✗ Pushing for a citywide ban on water boys in Atlanta	2020-12-07
✗ Promotional material for scientific journal "Water, Air, and Soil Pollution"	2020-12-11
✗ Viral article about Mississippi river	2021-02-08
✗ 6-year-old girl shot over spilled water	2021-03-21



→

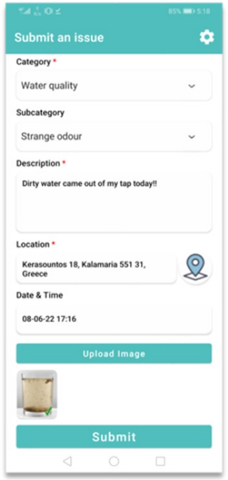


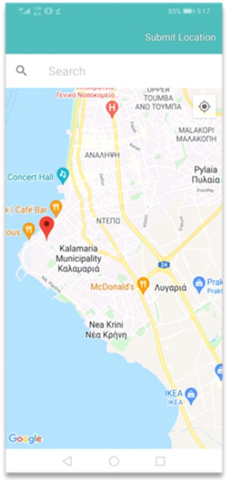
Detected events DBSCAN (2)

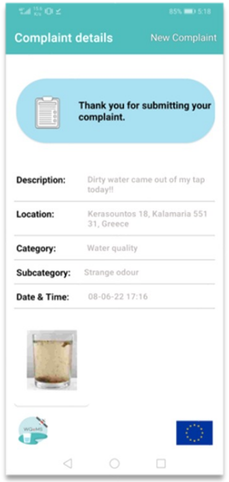
Incident	Date
. . .	. . .
✓ Japan's approval of plan to release wastewater into ocean	2021-04-13
✓ Japan's approval of plan to release wastewater into ocean	2021-04-14
✓ Lack of water in refugees camp in Myanmar	2021-06-02
✓ Environment Agency's discovery on water industry's failure to stop pollution with raw sewage in England	2021-07-13
✓ Double environmental protection budgets in England & Wales to fight river pollution	2021-07-14
✓ Double environmental protection budgets in England & Wales to fight river pollution	2021-07-15

More sources of social data

- WQeMS Crowdsourcing Mobile Application
  - A dedicated Android app enabling users to easily report water issues
  - Description of issue, location, date/time, attached photo



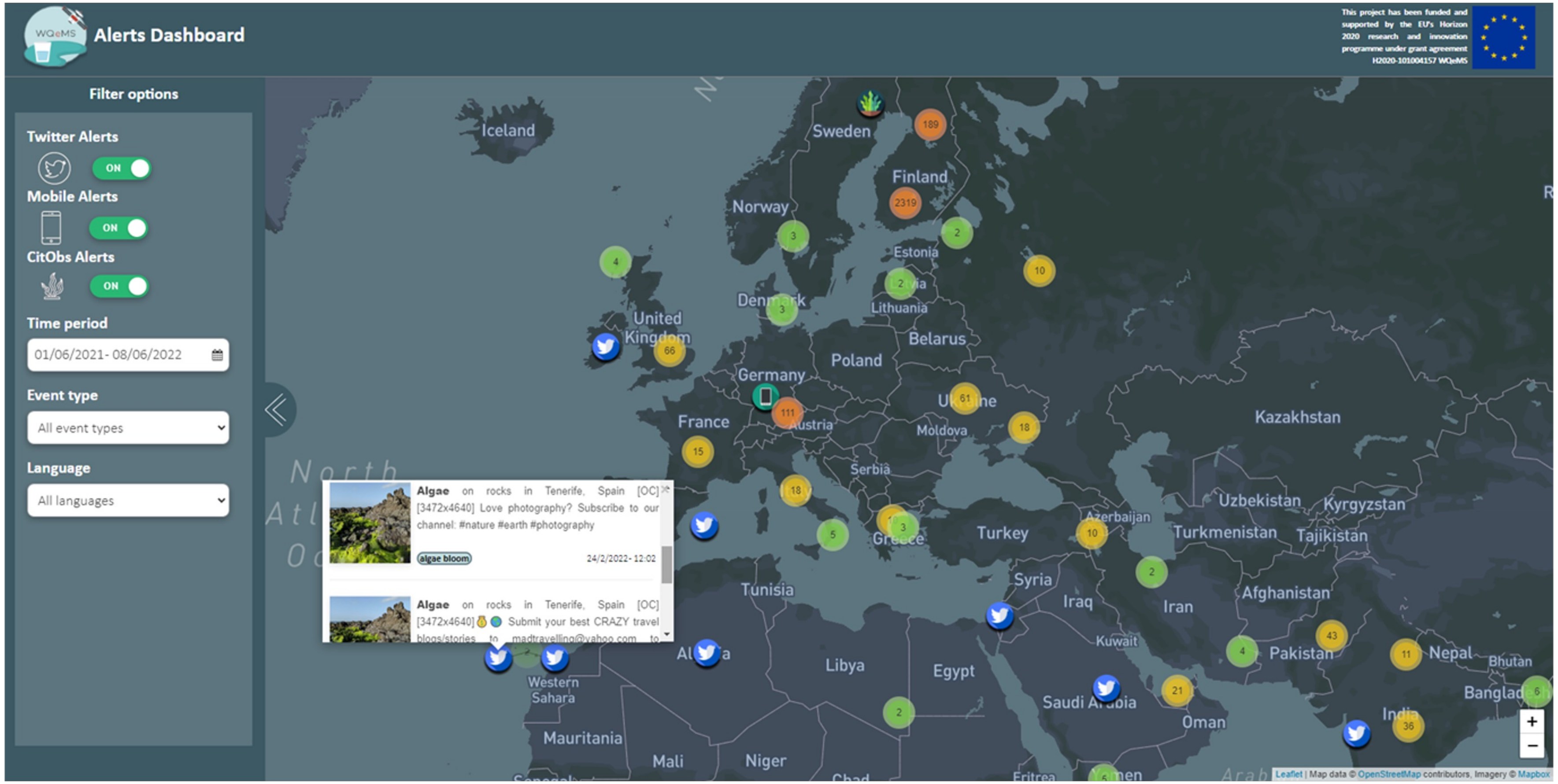




- Custom parsers for existing services
  - E.g. a parser for SYKE's CitObs API that serves citizen observations about algae blooms in Finland

More sources of social data

- Different types of crowdsourcing can be combined in a dashboard



http://m4d-apps.iti.gr:8007/WQeMS_Alerts_Dashboard

Multimodal Data Fusion for snow depth estimation

- Various modalities snow related analysis
  - **Twitter text**: BERT representation model for text to then classify each tweet as relevant or irrelevant to snow
  - **Twitter image**: GoogLeNet neural network for the extraction of the concept "snow"
  - **Satellite**: Backscatter measurements VV and VH are combined with snow cover

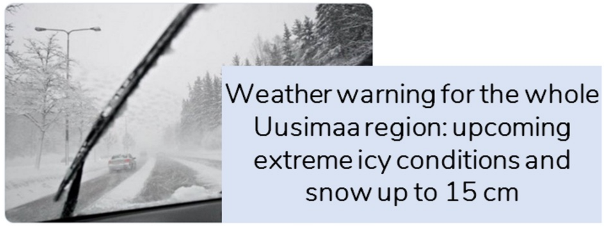
Collected tweets

Ilmastonmuutos on totta. Äkäslompolo 20.11.2018. Lunta 0 cm. #ilmastonmuutos #Lappi #lumetontalvi

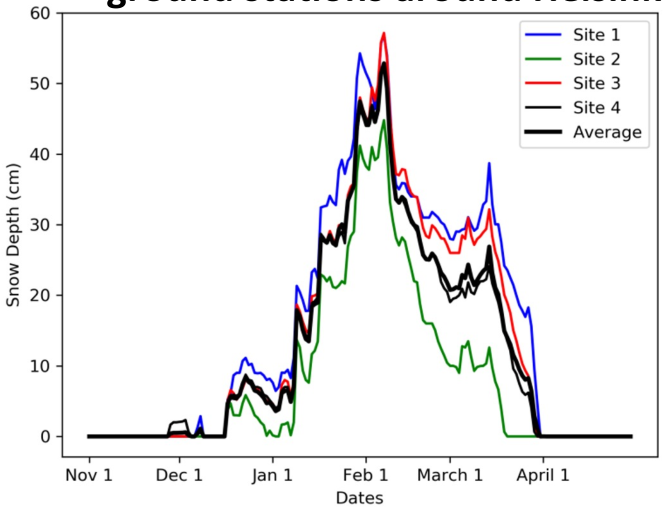
Syksyn viimeiset lumet Leivomäen #kansallispuisto tänään 12.11.2018

On hanget korkeat nietokset #talvi #Lappi #marraskuu #lumi #ilmastonmuutos

Säävaroitus koko Uudellemaalle: Nyt tulevat pääkallokelit ja lunta jopa 15 senttiä dlvr.it/QFFDNK



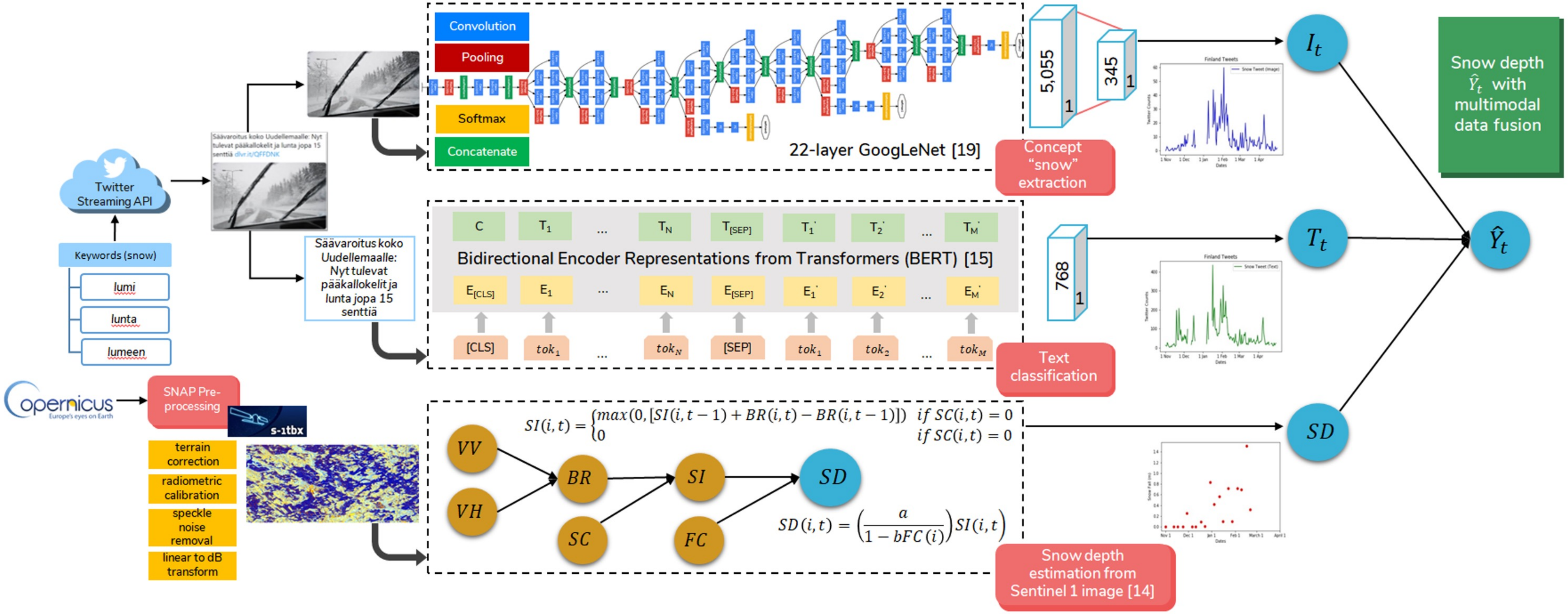
Observed snow depth in 4 ground stations around Helsinki



Multimodal Data Fusion framework

- Snow depth with multimodal data fusion combines the number of relevant-to-snow images I_t, tweets T_t and remotely sensed snow depth SD so as to provide a more accurate snow depth estimation.

$$\hat{Y}_t = \alpha_1 SD(t) + \alpha_2 I_t + \alpha_3 T_t + \beta$$



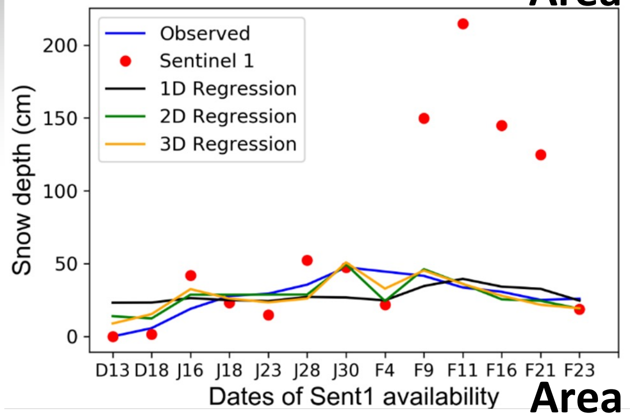
D. Mantsis, M. Bakratsas, S. Adreadis, P. Karsisto, A. Moumtzidou, I. Gialampoukidis, A. Karppinen, S. Vrochidis, I. Kompatsiaris: Multimodal Fusion of Sentinel-1 images and Social media Data for Snow Depth estimation. *IEEE Geoscience and Remote Sensing Letters,* 2020.

Snow depth model validation

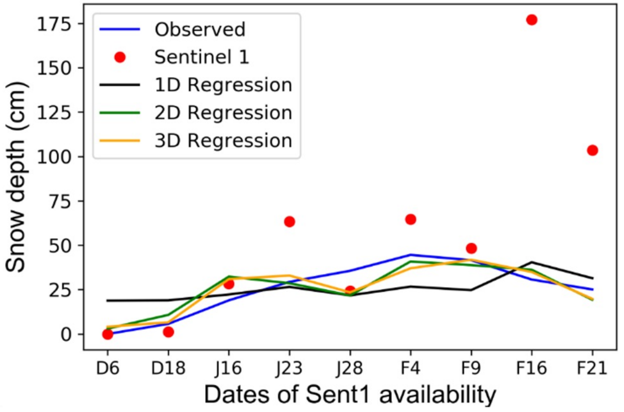
- 11,024 tweets were collected, covering a period of 151 days, i.e. from November 2018 till March 2019
- Two areas have been considered, 70Km distance far from the city center of Helsinki
- Ground truth measurements provided by the Finnish Meteorological Institute
- Evaluation metric: Mean Squared Error (MSE), with the objective to be minimised

Modalities used	MSE (Area B)	MSE (Area A)	MSE (average)
Satellite image	164.23	152.68	158.46
Satellite image, Twitter text	60.81	68.81	64.81
Satellite image, Twitter image	54.64	54.99	54.82
Satellite image, Twitter text, Twitter image	**47.11**	**54.98**	**51.05**

Area A



Area B



Social media monitoring for **crime prediction, detection** and **prevention**

Social Media Data for Crime Prediction, Detection and Prevention

- Despite the multitude of positive effects of social media, they are also used for nefarious reasons
- Social media have been exploited for recruiting terrorist and criminals online
- Common crimes can be facilitated through social media, such as trafficking of human beings
- Can be considered as a valuable source of information
  - ⇒ E.g. In a study of a violent incident (shooting of four police officers) in the Seattle-Tacoma, Washington, it was demonstrated that the majority of the messages posted on Twitter, regarding such incident, contained useful information

Crisis Event Detection
Architecture – Key components

Language Modelling

Self Attention Encoder

Multi-head Attention

CNN

Crisis Event Detection
Architecture – Key components

Language Modelling

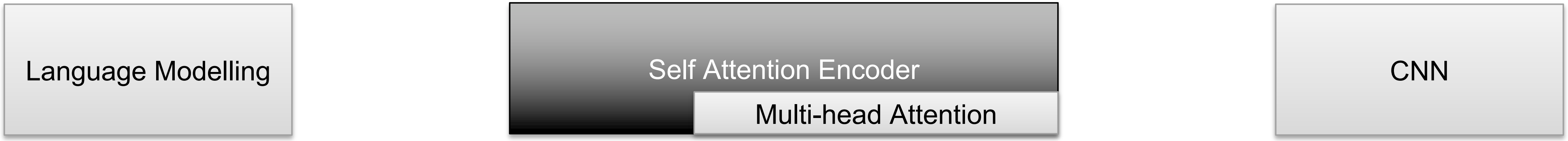
Self Attention Encoder

Multi-head Attention

CNN

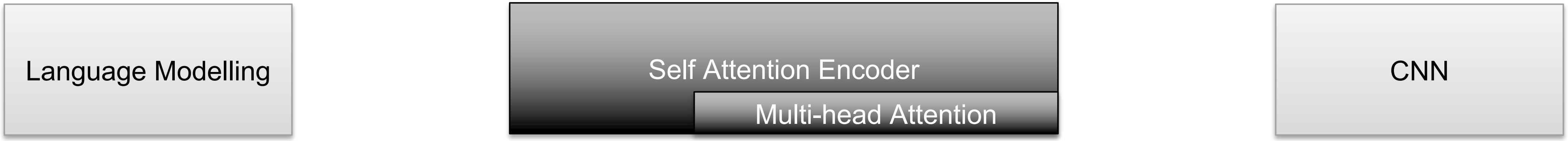
- Translation of the human readable characters and words to a mathematical representation
  $\Rightarrow$ Modelling the semantic meaning of each word
- Word2Vec model: 300 dimensions per word
- Pretrained on Google news

Crisis Event Detection
Architecture – Key components



- Decide which parts of a sequence are more important
  ⇒ I.e. in which parts more attention should be placed to
- The SOTA attention encoding method from Transformers is employed as a feature extractor

Crisis Event Detection
Architecture – Key components



- Decide which parts of a sequence are more important
  ⇒ More attention should be placed to
- The SOTA attention encoding method from Transformers is employed as a feature extractor

Multi-head Attention

- The layer where the attentions are calculated

Crisis Event Detection
Architecture – Key components

Language Modelling

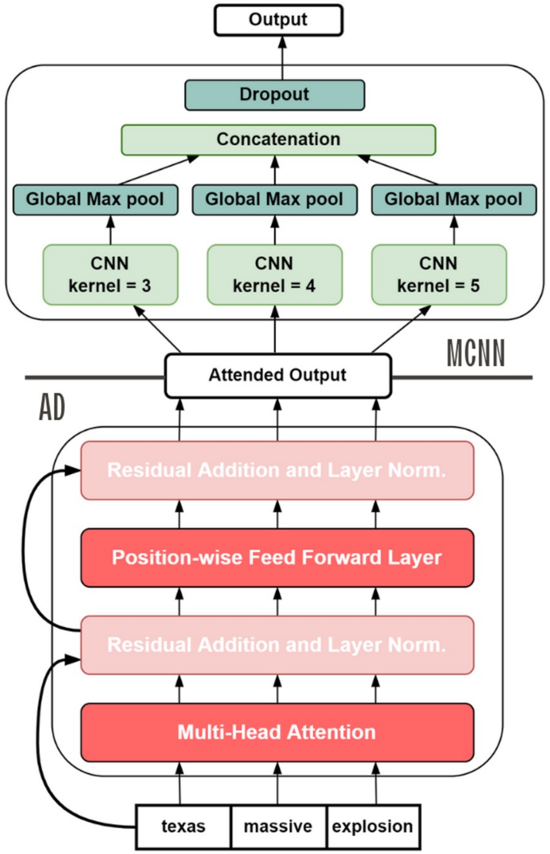
Self Attention Encoder

Multi-head Attention

CNN

- Convolutional Neural Networks (CNN)
  - Have proven to be invaluable for NLP tasks
  - Capture of salient information from n-gram word combinations
  - Capable to acquire local information from the input
  - Often used in event detection tasks

Crisis Event Detection
Architecture



Attention Denoised Multi-channel CNN (AD-MCNN)

- One self-attention encoder as feature extraction mechanism
- Use of three parallel CNN layers operating under different kernel sizes
  - ⇒ Capture different *n*-gram combinations from the text
- Max-over-time pooling operation

Crisis Event Detection
Ground Truth & Baseline

Ground Truth

CrisisLexT26

~28k Twitter posts

26 Crisis events 2012-2013

~1k posts per crisis event

Baseline Model

Multi-channel CNN (MCNN)

It has shown the best performance so far in the CrisisLexT26 dataset

Crisis Event Detection
Experimental Setup

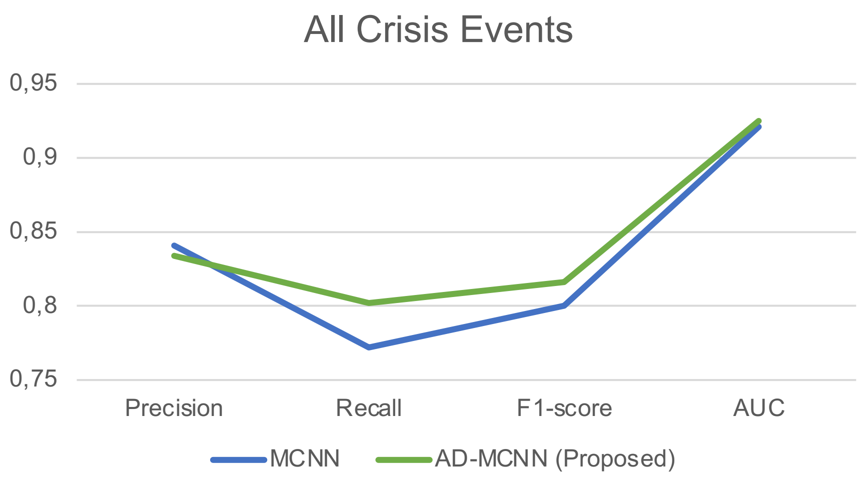
All Crisis

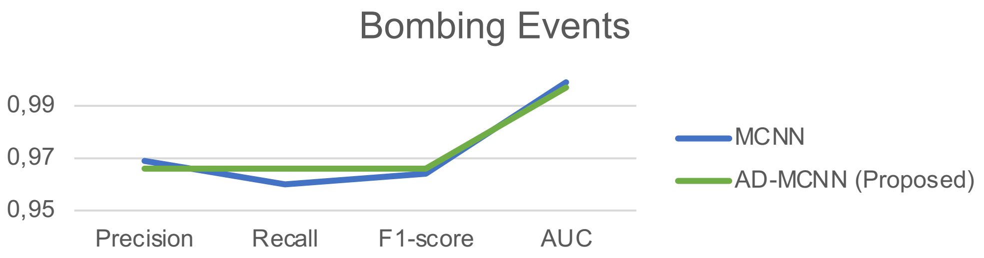
- Binary classification setup: related to an event or not
  ⇒ all crisis events are considered as one classification category

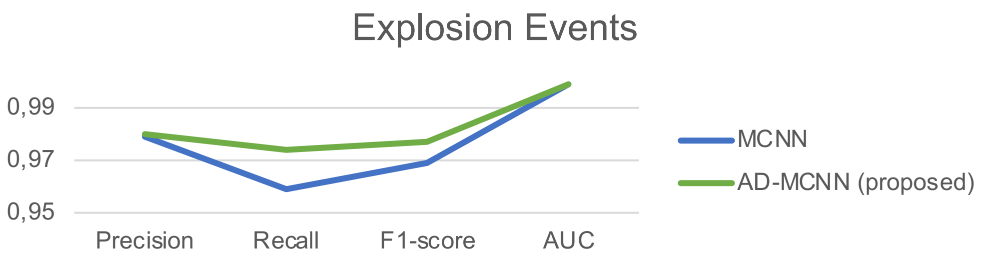
Specific Crisis

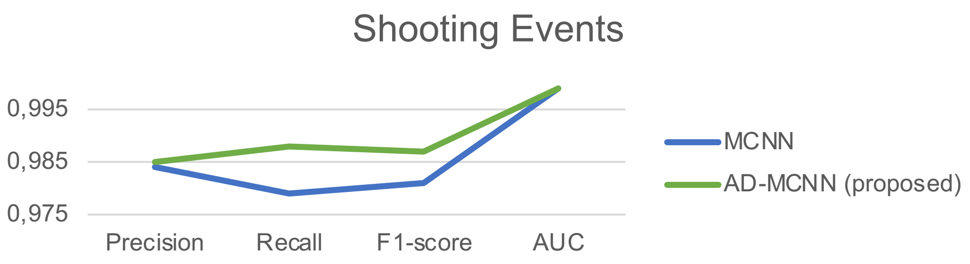
- Unique models for detecting crime-related crisis events
- Three classification models able to inference **explosion**, **shooting** and **bombing** types of events
- Binary classification setup:
  ⇒ 1^{st} class: events of a specific type
  ⇒ 2^{nd} class: all the other types of events from the CrisisLexT26 dataset

Crisis Event Detection
Experimental Results







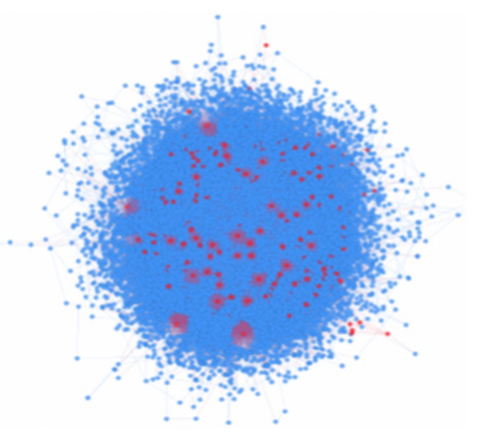


Multiple Identities Detection in Social Media

- Users often hold several accounts in their effort to multiply the spread of their thoughts, ideas, and viewpoints
- Illegal activities: creation of multiple accounts to bypass the combating measures enforced by social media platforms

User Identity Linkage

Detect accounts likely to belong to the same natural person ("linked accounts")

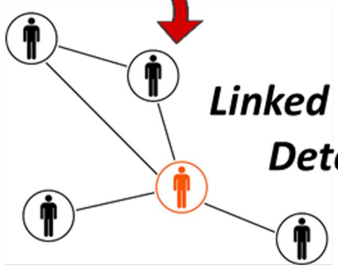


User Identity Linkage Framework

Feature extraction
- Stylometric
- Semantic

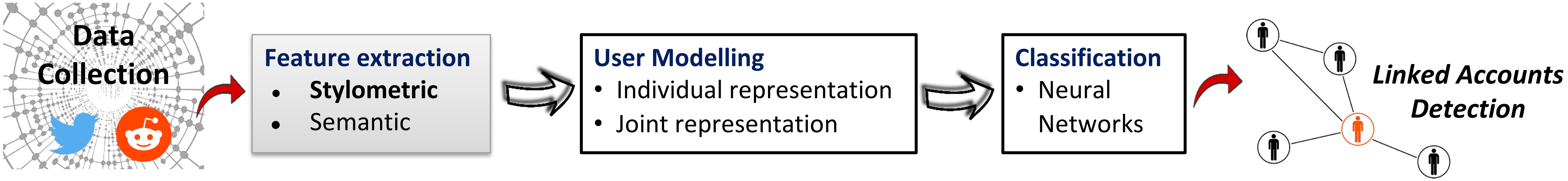
User Modelling
- Individual representation
- Joint representation

Classification
- Neural Networks



Linked Accounts Detection

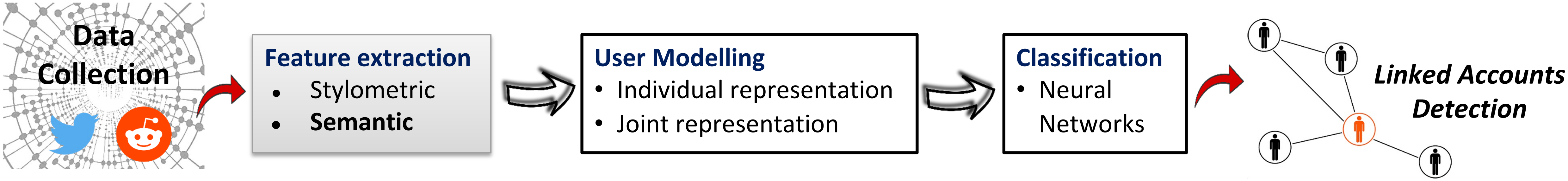
User Identity Linkage
Feature Extraction – Stylometric Features



- Characterise at several different levels the inherent writing style of users
- Users can easily change the words they use to mask their identity, but cannot easily change the small stylometric characteristics that they are not even aware of
- **Features:**
  (i) Character-based, (ii) Word-based, (iii) Sentence-based, (iv) Dictionary-based, (v) Syntactic-based

Around 200 stylometric features are extracted

User Identity Linkage
Feature Extraction – Semantic Features



- **Word2Vec**: uses neural networks to train a predictive model
- **GloVe**: considers statistical information for each word using a global co-occurrence matrix
- **ELMo**: the representation of each word depends on the surrounding context

Pre-trained word embeddings are used to encode the input texts:

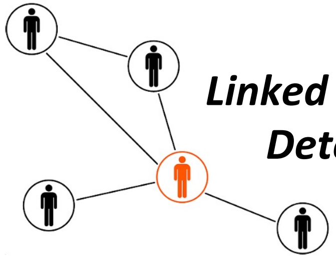
- Word2Vec (300d) & GloVe (100d): Twitter pre-trained embeddings
- ELMo: pre-trained embeddings from a language model trained on 1B word benchmark

User Identity Linkage
User Modelling – Individual representation

Feature extraction
- Stylometric
- Semantic

User Modelling
- **Individual representation**
- Joint representation

Classification
- Neural Networks



Linked Accounts Detection

Based on Stylometric Features

$$u_i: V_{u_i} = < f_{i_1}, f_{i_2}, \dots, \boxed{f_{i_j}}, \dots, f_{i_{\boxed{m}}} >,$$

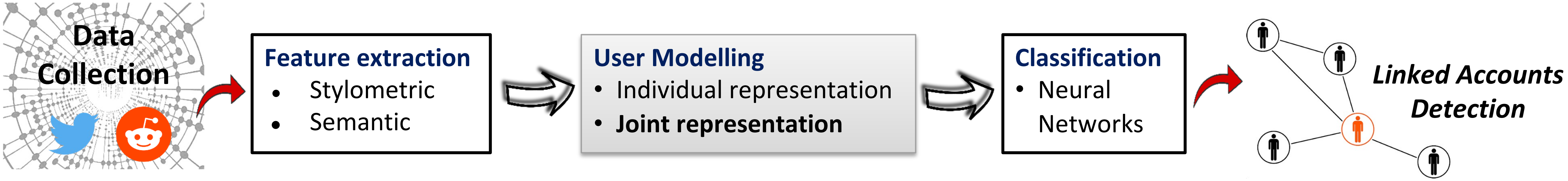
j_{th} feature for user u_i

Total number of stylometric features

Example

$$V_{u_i} = < chars_i, acronyms_i, \dots, pronouns_i >$$

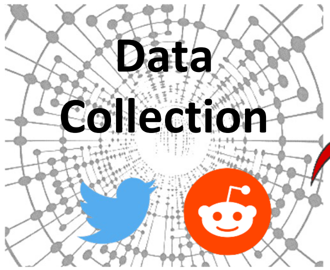
User Identity Linkage
User Modelling – Joint representation



Joint representation of each pair of users

- Identify their potential relationship
- Use that as input to the classifier
- Absolute difference of feature vectors of ui, uj

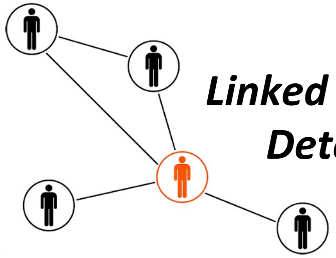
User Identity Linkage
Classification



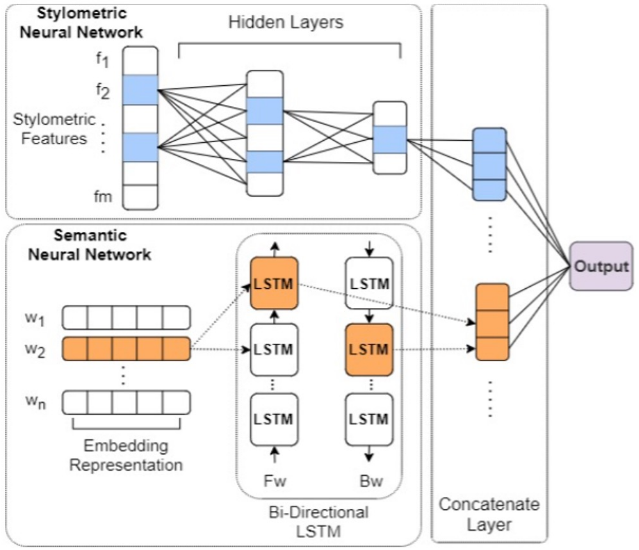
Feature extraction
- Stylometric
- Semantic

User Modelling
- Individual representation
- Joint representation

Classification
- **Neural Networks**



Linked Accounts Detection



Stylometric neural network: stylometric features that have been previously jointly represented as a unique feature vector

- Regularized fully connected (dense) layers

Semantic neural network: includes all the posts of a user pair

- Bidirectional LSTM

Combined neural network

- Semantic and Stylometric representations are concatenated and used to determine if two texts are written by the same author or not.

User Identity Linkage
Datasets

June to August 2016

Relevant to Gamergate controversy

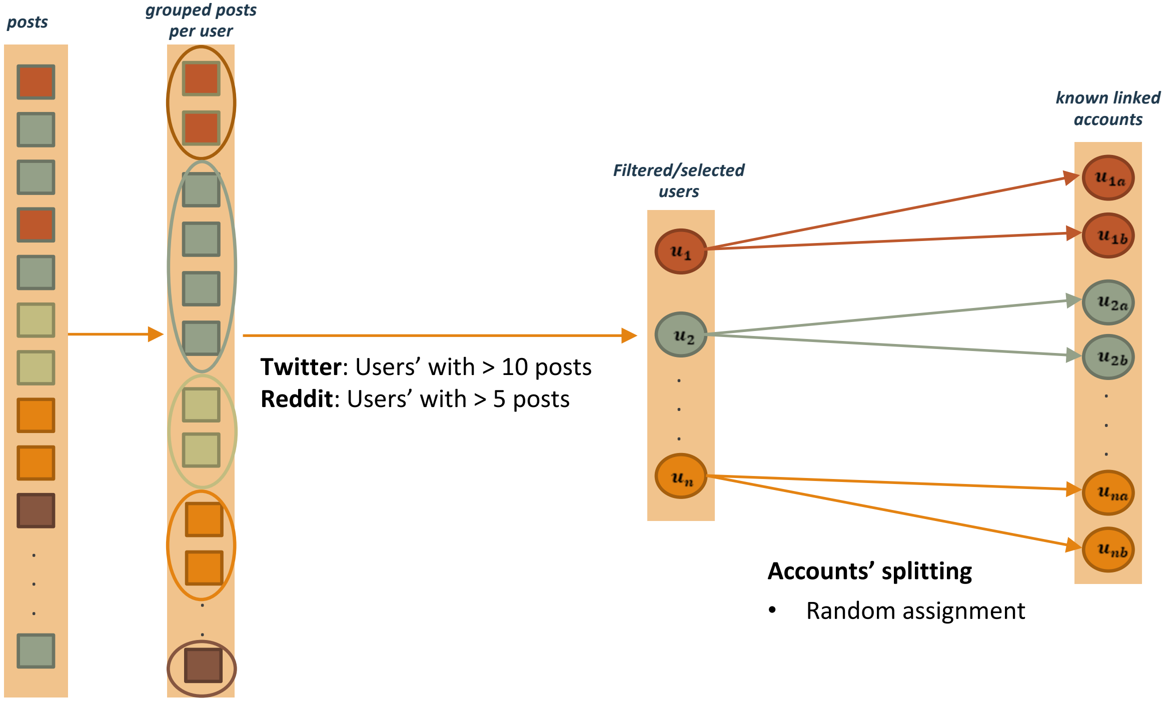
Abusive-related English hashtags

650K tweets and 312K users

Extraction of Twitter usernames in the GamerGate data and search for them in Reddit

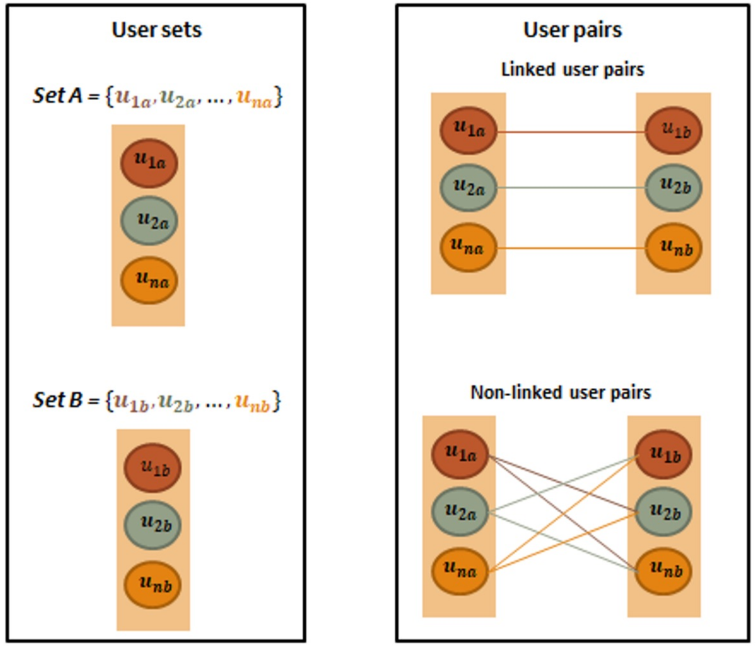
9,615 posts and 324 users

User Identity Linkage
Ground Truth – Data sampling



- Absence of well established ground truth that indicates which user accounts belong to the same person
- We follow an approach commonly used for such a task

User Identity Linkage
Ground Truth – Generation



- **Overall number of non-linked user pairs**
    - **Twitter**: *2,958* linked and *26,622* non-linked accounts
    - **Reddit**: *215* linked and *1,935* non-linked accounts

- **Ground truth**
    - 10% linked accounts
    - 90% non-linked accounts

User Identity Linkage
Experimental Setup

Single platform setup

- Separately, on the Twitter and Reddit datasets
  - Using the semantic neural network
  - Using the stylometric neural network
  - Combining both

Cross platform setup

- Train with Twitter data, test with Reddit data and vice versa
- What is the effect of the written language when training in a particular platform and testing to another?

For all experiments: 90% training, 10% testing and 10% of training as development set

User Identity Linkage
Experimental Results – Single Platform Setup

	Prec	Rec	Acc	AUC
Semantic neural network				
Word2Vec	0.8098	0.8999	0.8999	0.4749
GloVe	0.8534	0.8995	0.8995	0.6599
ELMo	0.8614	0.8982	0.8982	0.8059
Stylometric neural network				
	0.9356	0.9405	0.9405	0.9307
Combined neural network				
	0.9424	**0.9428**	**0.9428**	**0.9576**

	Prec	Rec	Acc	AUC
Semantic neural network				
Word2Vec	0.8058	0.8976	0.8976	0.4551
GloVe	0.8058	0.8976	0.8976	0.5442
ELMo	0.8053	0.8930	0.8930	0.5737
Stylometric neural network				
	0.9099	0.9162	0.9162	0.8316
Combined neural network				
	0.9433	**0.9395**	**0.9395**	**0.9126**

- **Semantic neural network:** best performance with the ELMo embeddings
  - ELMo generates an embedding for each word based on its context; instead of using fixed embedding for each word
- **Stylometric neural network:** outperforms the semantic neural network by a large margin
  - Clear stylometric patterns are extracted with the proposed feature set: facilitate a clear distinction between users in the two different scenarios (Twitter and Reddit)
- **Combined neural network (with ELMo):** better performance compared to only using either the semantic or stylometric neural network
  - Each configuration analyses the content at different levels and thus their combination would be expected to yield better performance

User Identity Linkage
Experimental Results – Cross Platform Setup

Train → Test: Twitter → Reddit

	Prec	Rec	Acc	AUC
Semantic neural network	0.8383	0.7320	0.7320	0.5849
Stylometric neural network	**0.9113**	**0.9102**	**0.9102**	0.7101
Combined neural network	0.8888	0.9083	0.9083	**0.7404**

Train → Test: Reddit → Twitter

	Prec	Rec	Acc	AUC
Semantic neural network	0.8099	0.9000	0.9000	0.5586
Stylometric neural network	0.8556	0.8159	0.8159	0.6889
Combined neural network	**0.8801**	**0.8984**	**0.8984**	**0.8039**

The same patterns hold as in the single platform setup

- Stylometric neural network outperforms the semantic one (in terms of AUC)
- The combined network yields better performance compared to the individual ones

General observations

- The basic patterns found on the one platform are generalisable enough and can be found on the other platform as well
- The stylometric feature set succeeds in extracting several patterns that are source-independent

The proposed approach is generalisable obtaining competitive performance in both cross-platform experiments

Related Publications

- Kyriakidis, P., Chatzakou, D., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2022, April). Leveraging Transformer Self Attention Encoder for Crisis Event Detection in Short Texts. In *European Conference on Information Retrieval* (pp. 163-171). Springer, Cham.
- Theodosiadou, O., Pantelidou, K., Bastas, N., Chatzakou, D., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2021). Change point detection in terrorism-related online content using deep learning derived indicators. *Information*, *12*(7), 274.
- Chatzakou, D., Soler-Company, J., Tsikrika, T., Wanner, L., Vrochidis, S., & Kompatsiaris, I. (2020, July). User Identity Linkage in Social Media Using Linguistic and Social Interaction Features. In *12th ACM Conference on Web Science* (pp. 295-304).

Cultural Applications

Use Cases

- Use of social media networks in order to create (training) datasets or **relevant** items for various applications
- Design and implementation of interactive virtual reality platforms that will gather elements of **intangible cultural heritage (e.g. traditional dances)**
- Develop tools to enable sharing of cultural heritage and co-creation of new cultural materials with and for **refugees**
  - Storytelling based on maps, with interactive visual elements and textual resources are created to present and link the data geographically
- Provide repurposed content to targeted creative industries (e.g. **architects**)
  - Visual tags (architectural style, scene description, localized objects and buildings)
  - Textual tags, sentiment analysis metadata and natural language descriptions
- Key requirement: **Copyrights / Distribution licenses** that restrict using and sharing social media content

Traditional dances dataset creation

- Total miscellaneous videos retrieved from social media platforms: 513

Dance	No. of Videos	Duration of Videos	GBs in Storage
Gikna	13	~ 28 minutes of footage	0.57
Mpaintouska	15	~ 28 minutes of footage	0.85
Karsilamas	16	~ 40 minutes of footage	1.89
Hasapikos	15	~ 37 minutes of footage	1.40

Thracian Folklore Dance Dataset

Main topics of Collection

1. Folklore customs
2. Actuators of customs
3. People participating in folklore customs
4. Places of interest
5. Songs and odes
6. Dances

Keywords/Keyphrases (Greek Words with English Letters):

Anastenaria, Anastenarides, Agia Eleni, Pirovasia, Konaki, O Konstantinos o mikros, Sta prasina livadia, Skopos tou dromou, Arapides, Monastiraki Dramas, Tseta, Gkiligkes, Flampouro, Kodonoforoi, Koudounia, Mpatalia, Trakarntaki, Podopania, Mpampougera, Mpampougeros, Lira, Kasnaki

Keywords/Keyphrases (Native Greek letters):

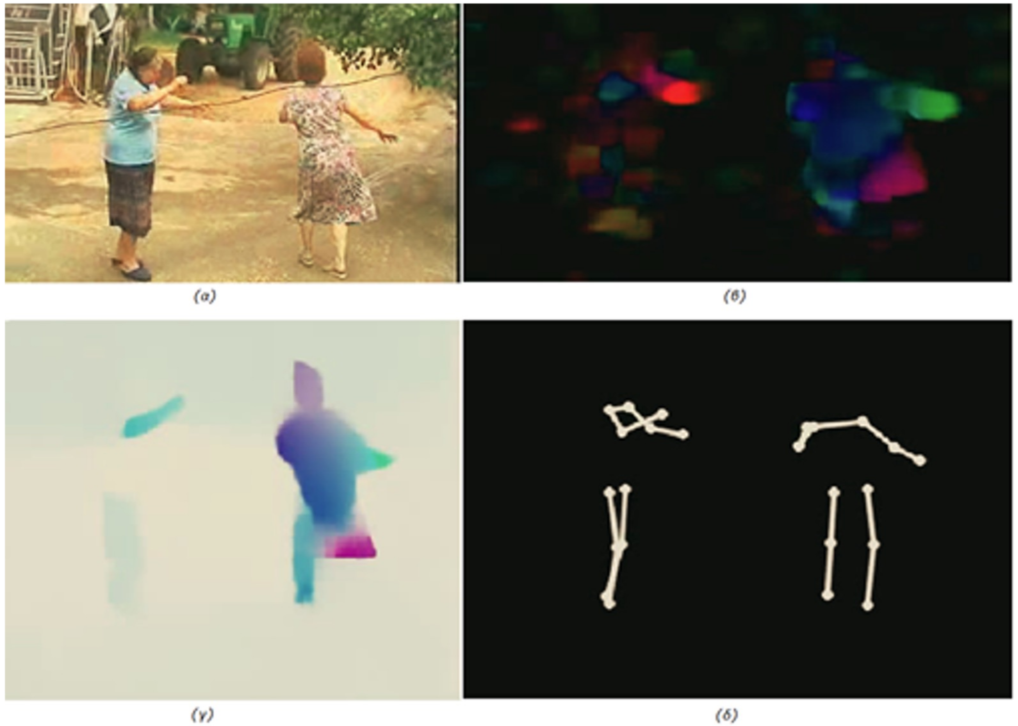
Αναστενάρια, Αναστενάρηδες, Αγία Ελένη, Πυροβασία, Κονάκι, Ο Κωνσταντίνος ο μικρός, Στα πράσινα λιβάδια, Σκοπός του δρόμου, Αράπηδες, Μοναστηράκι Δράμας, Τσέτα, Γκιλίγκες, Φλάμπουρο, Κωδωνοφόροι, Κουδούνια, Μπατάλια, Τρακαρντάκι, Ποδοπάνια, Μπαμπούγερα, Μπαμπούγερος, Λύρα, Κασνάκι

Example content

The collected content is used for training algorithms that aim to extract useful information regarding the human activities conducted in a cultural framework.

More precisely, the content was used for

- 3D Pose Estimation
- Dance Recognition
- Sentiment Analysis
- Laban Generation



Various depictions of the training datasets used for dance recognition. (α) Initial content from social media, (β) Farneback optical flow, (γ) RAFT optical flow and (δ) 2D pose.

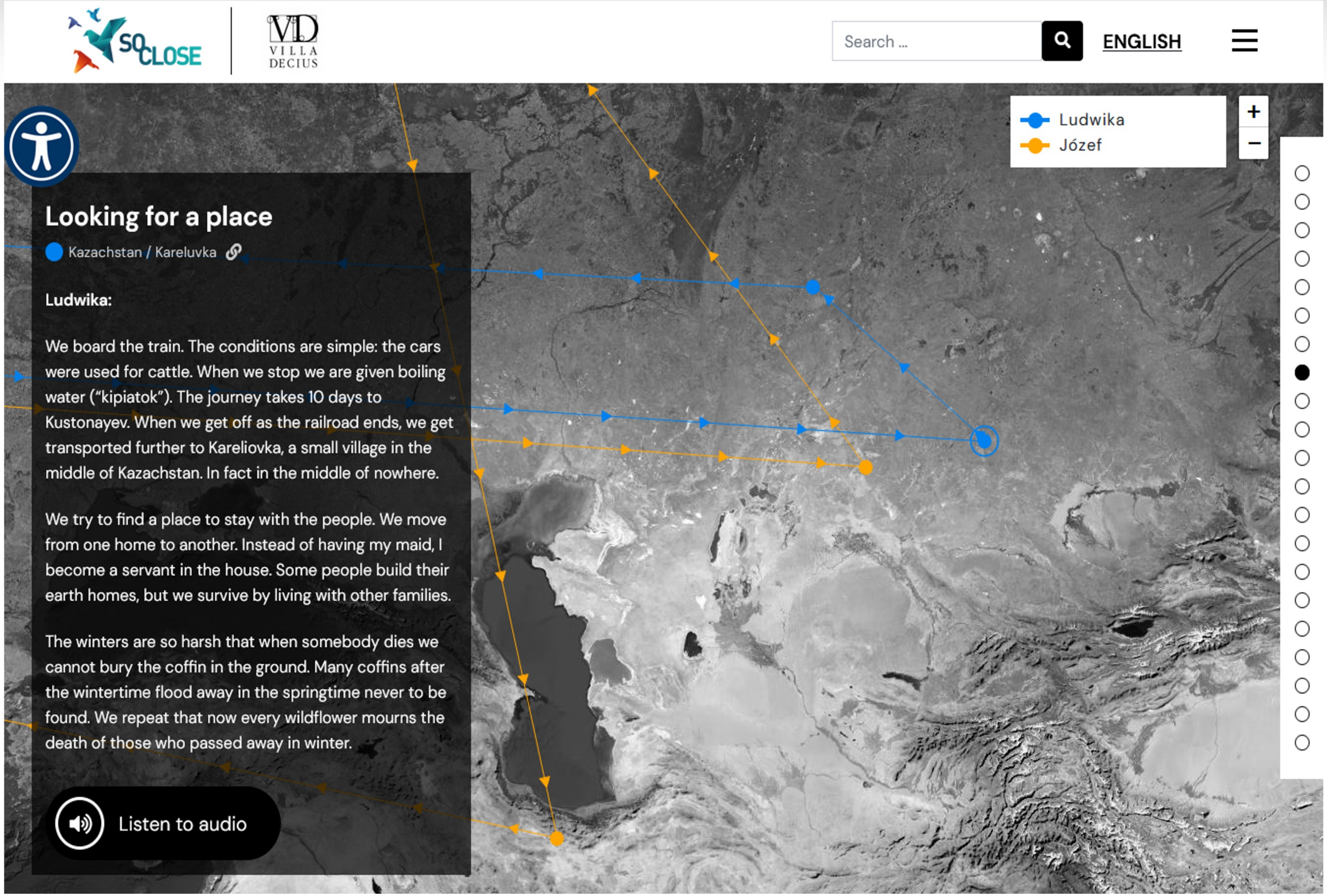
Refugee related collection of keywords examples

- Public twitter posts were gathered for creating collections in the database for textual analysis tasks.
- Example of two collections based on the following keywords:
  - Needs and rights of refugees
    - refugee needs, refugee rights, right to citizenship, refugee documents, right to work, right to food, access to health, access to education, access to housing
  - Geography
    - refugee Poland, refugee Italy, refugee Greece, refugee Spain, refugee Catalunya, crossing borders

Dataset name	Tweets Found	Tweets Added to database	Time to Fetch (min)
Needs and rights of refugees	279916	105338	226.480
Geography	4025	1386	6.774

Example datasets with social media sources

Story map example

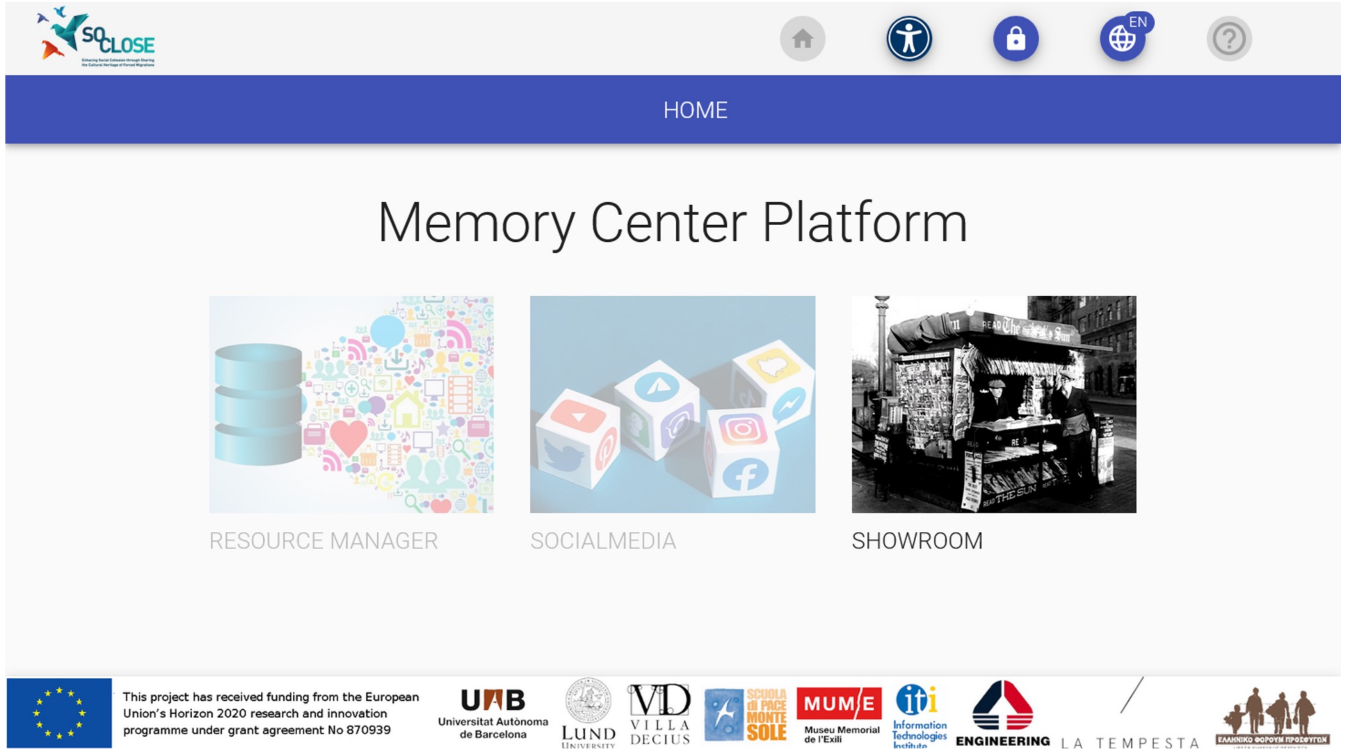


Tools

- Interactive story map:
  - https://so-closetools.eu/storymap/scuoladipacemontesole/
  - https://so-closetools.eu/storymap/willa-decjusza/
  - https://so-closetools.eu/storymap/museuexili
- Participatory virtual exhibition:
  - https://so-closetools.eu/virtualexhibition/displaced-voices/
  - https://so-closetools.eu/virtualexhibition/scuoladipacemontesole
  - https://so-closetools.eu/storymap/museuexili
- Immersive web documentary:
  - https://so-closetools.eu/webdoc/museuexili
  - https://so-closetools.eu/webdoc/displaced-voices-in-fences/

Tools

- Memory Center Platform (MCP):
    - https://mcpwebstart.net/



Memory Center Platform (MCP) main page

Demos

- V4Design: 3D reconstruction from a Youtube video input
  - https://www.youtube.com/watch?v=w9G-FgyjyHg

Related publications

- S. Mille, S. Symeonidis, M. Rousi, M. Marimon Felipe, K. Stavrothanasopoulos, P. Alvanitopoulos, R. Carlini Salguero, J. Grivolla, G. Meditskos, S. Vrochidis and L. Wanner, "A Case Study of NLG from Multimedia Data Sources: Generating Architectural Landmark Descriptions", in WebNLG+: 3rd Workshop on Natural Language Generation from the Semantic Web, (INLG 2020), 15-18 December 2020.
- E.A. Stathopoulos, A. Shvets, R. Carlini, S. Diplaris, S. Vrochidis, L. Wanner and I. Kompatsiaris, “Social Media and Web Sensing on Interior and Urban Design”, Fourth International IEEE Workshop on Social (Media) Sensing, 30 June – 3 July 2022 (accepted for publication)

Overall Closing

Policy – Licensing – Legal challenges

- Fragmented access to data
  - Separate wrappers/APIs for each source (Twitter, Facebook, etc.)
  - Different data collection/crawling policies
- Limitations imposed by API providers ("Walled Gardens")
  - Full access to data impossible or extremely expensive (e.g. see data licensing plans for GNIP and DataSift)
  - Non-transparent data access practices (e.g. access is provided to an organization/person if they have a contact in Twitter)
- Constant change of model and ToS of social APIs
  - No backwards compatibility, additional development costs
- Ephemeral nature of content
  - Social search results often lead to removed content, inconsistent and unreliable referencing
- User Privacy & Purpose of use
- Fuzzy regulatory framework regarding mining user-contributed data

Conclusions

- Social media data useful in many applications: from confirming existing and known correlations to prediction and decision-making
- Many challenges exist
  - Data availability and representativeness (of society, real-event)
  - Coverage, robustness and reproducibility
  - Real-time and scalable approaches
  - Selection and fusion of various modalities (content, network-social, temporal, location) and combination with external sources
- Required contribution from various disciplines
  - Content Analytics
  - Machine Learning
  - Network Analysis
  - Big Data Architectutre, Cloud
  - Psychology – Social Sciences (patterns of presentation, sharing)
  - Visualization
- Currently mostly an auxiliary means for real-events assessment and decision-making, which can generate additional insights

Contributions

Spyridon Symeonidis,
Architecture Design

Alexandros Kokkalas,
Cultural Applications

Sotiris Diplaris,
Cultural Applications,
Architecture Design

Despoina Chatzakou,
Data mining, Deep learning,
Natural Language Processing,
Behaviour analysis

Ourania Theodosiadou,
Stochastic modelling,
Computational statistics,
Time series analysis,
Predictive analytics

Pantelis Kyriakidis,
Deep learning,
Reinforcement learning,
Natural language processing

Theodora Tsikrika,
Web and social media search and
mining, Multimedia indexing and
retrieval, AI-based multimodal analytics

Stelios Andreadis,
Social media monitoring
& analytics, Web design

Ilias Gialampoukidis
Multimodal data fusion, Web and
social media mining, Multimedia
analysis and retrieval

Stefanos Vrochidis,
Multimodal data fusion, Web and social
media mining, Multimedia analysis and
retrieval, Multimodal analytics

Support

V4Design

Visual and textual content re-purposing FOR(4) architecture, Design and video virtual reality games

S+T+ARTS LIGHTHOUSE MINDSPACES

Art-driven adaptive outdoors and indoors design

CONNEXIONs

InterCONNected NEXt-Generation Immersive IoT Platform of Crime and Terrorism DetectiON, PredictiON, InvestigatiON, and PreventiON Services

STARLIGHT

Sustainable Autonomy and Resilience for LEAs using AI against High priority Threats

INFINITY

IMMERSE. INTERACT. INVESTIGATE.

Investigative, Immersive, and Interactive Collaboration Environment

SO CLOSE

Enhancing Social Cohesion through Sharing the Cultural Heritage of Forced Migrations

Support

CHROMATA

Development of Technologies for Intagible Heritage

Dances, Songs, Myths and Customs for the Development of Technologies for Intangible Cultural Heritage

aqua3S

Enhancing Standardisation strategies to integrate innovative technologies for Safety and Security in existing water networks

PathoCERT

Pathogen Contamination Emergency Response Technologies

WQeMS

Copernicus Assisted Lake Water Quality Emergency Monitoring Service

CALLISTO

Copernicus Artificial Intelligence Services and data fusion with other distributed data sources and processing at the edge to support DIAS and HPC infrastructures

INGENIOUS

Enhancing Situational Awareness, Safety and Collaboration for the First Responder of the future

The First Responder (FR) of the Future: a Next Generation Integrated Toolkit (NGIT) for Collaborative Response, increasing protection and augmenting operational capacity

Thank you for your attention!

ikom@iti.gr

http://mklab.iti.gr