



Open Archive Toulouse Archive Ouverte

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible

This is an author's version published in: <http://oatao.univ-toulouse.fr/22354>

To cite this version:

Mothe, Josiane *Learning from Social Media and Contextualisation*. (2018) In: Dagstuhl Seminar 17301 (2017), 24 July 2017 - 28 July 2017 (Dagstuhl, Germany).

Any correspondence concerning this service should be sent to the repository administrator:

tech-oatao@listes-diff.inp-toulouse.fr



LEARNING FROM SOCIAL MEDIA & CONTEXTUALISATION

Josiane.Mothe@irit.fr

1

**User-Generated Content in Social Media
Dagstuhl Seminar July 2017**

RESEARCH OBJECTIVES

○ Learning from Social Media: main trends vs peculiarities

- Information Mining to extract main trends and behaviour
 - Main topics
 - Main users (e.g. influential users on a topic)
 - Following dissemination (e.g. flu, typhon)
 - Link between information (e.g. location information extraction)
- Information Mining to detect peculiarities and weak signals
 - Used in Business Intelligence
 - Knowledge of the environment / contexts
 - Detect opportunities / risks
 - Detect changes

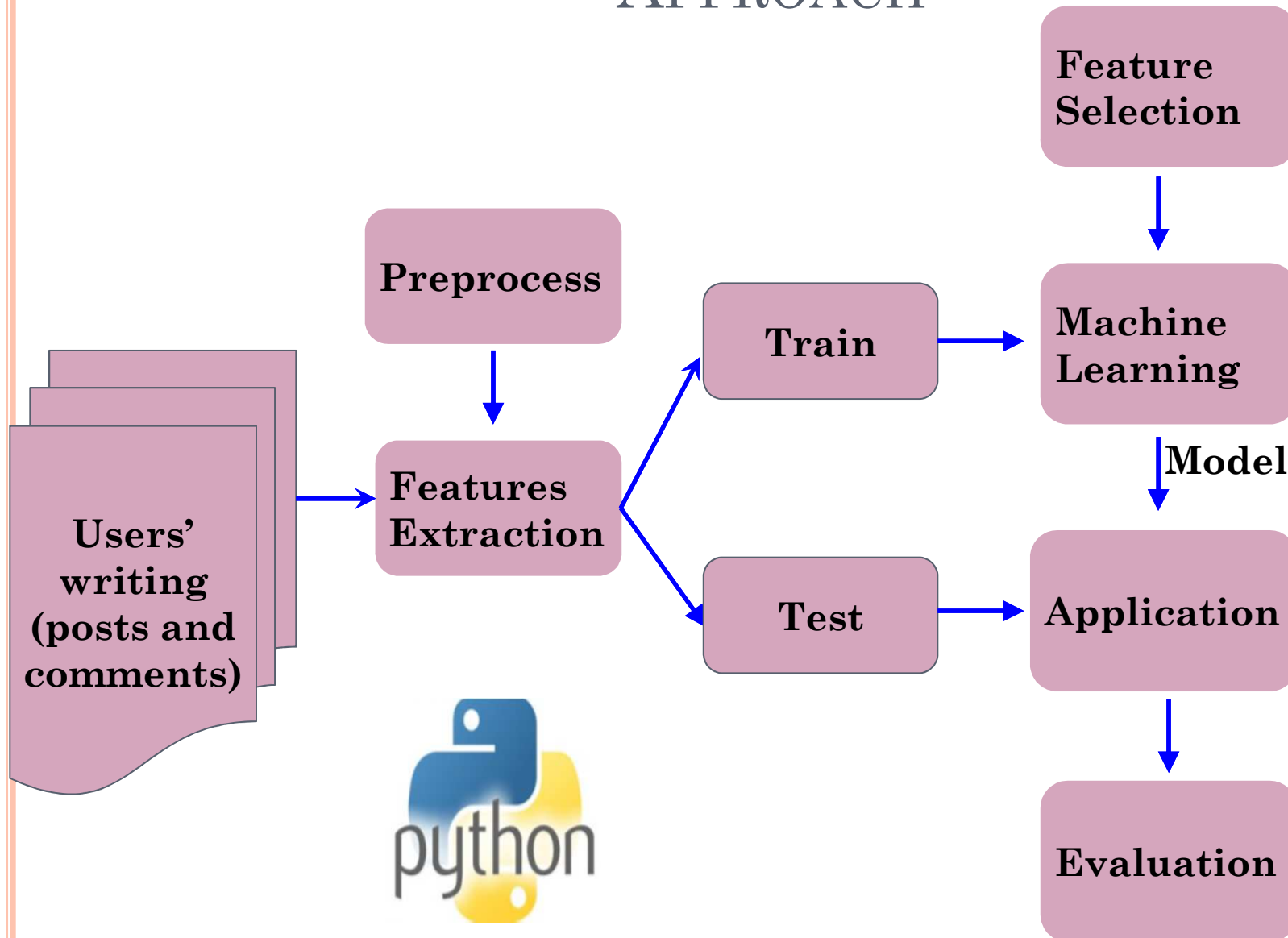
Detecting changes in behaviour : early detection of depression

RELATED WORK IN DEPRESSION DETECTION

Supervised learning based on 6 main groups of features:

- **N-grams:** unigram, bigram and trigram [*Yang et al., 2015*];
- **Relevant Lexicons:** depression symptoms, lexicon of drug names etc.(wikipedia);
- **Linguistic Style:** frequency of negative word, quantifiers, quantity of first personal pronoun, quantity of emoticons, numbers of exclamation and question marks etc. [*Coppersmith et al., 2014*];
- **Users Behaviors:** number of words/posts, proportion of reply posts from a user per day etc., user's active period or posting time span [*Choudhury, 2016*];
- **Sentiment analysis:** sentence polarity, sentiment words - positive and negative sentiment [*Mowery et al., 2016*];
- **Emotion analysis:** categories of emotions - emotional degree [*Choudhury, 2016*].

OUR NATURAL LANGUAGE PROCESSING APPROACH



5 MAIN GROUPS OF EXTRACTED FEATURES

- **Bag of Words:** 18 most frequent unigrams (depressive) - our baseline
- **Group 1: Language Style** 13 features
 - Negation frequency
- **Group 2: Self-Preoccupation** 9 features
 - Frequency of Personal Pronouns
- **Group 3: Reminiscence and Relevant Words** 5+5 f
 - Past tense
 - Reference to related drugs
- **Group 4: Sentiment and Emotion** 3+8 features
 - Positive sentiment
 - Sadness

EXPERIMENTAL FRAMEWORK: ERISK DATA AND WEKA

CLEF eRisk 2017: early risk prediction of depression from data extracted from **Reddit** (american social news aggregation, web content rating, and discussion website.)

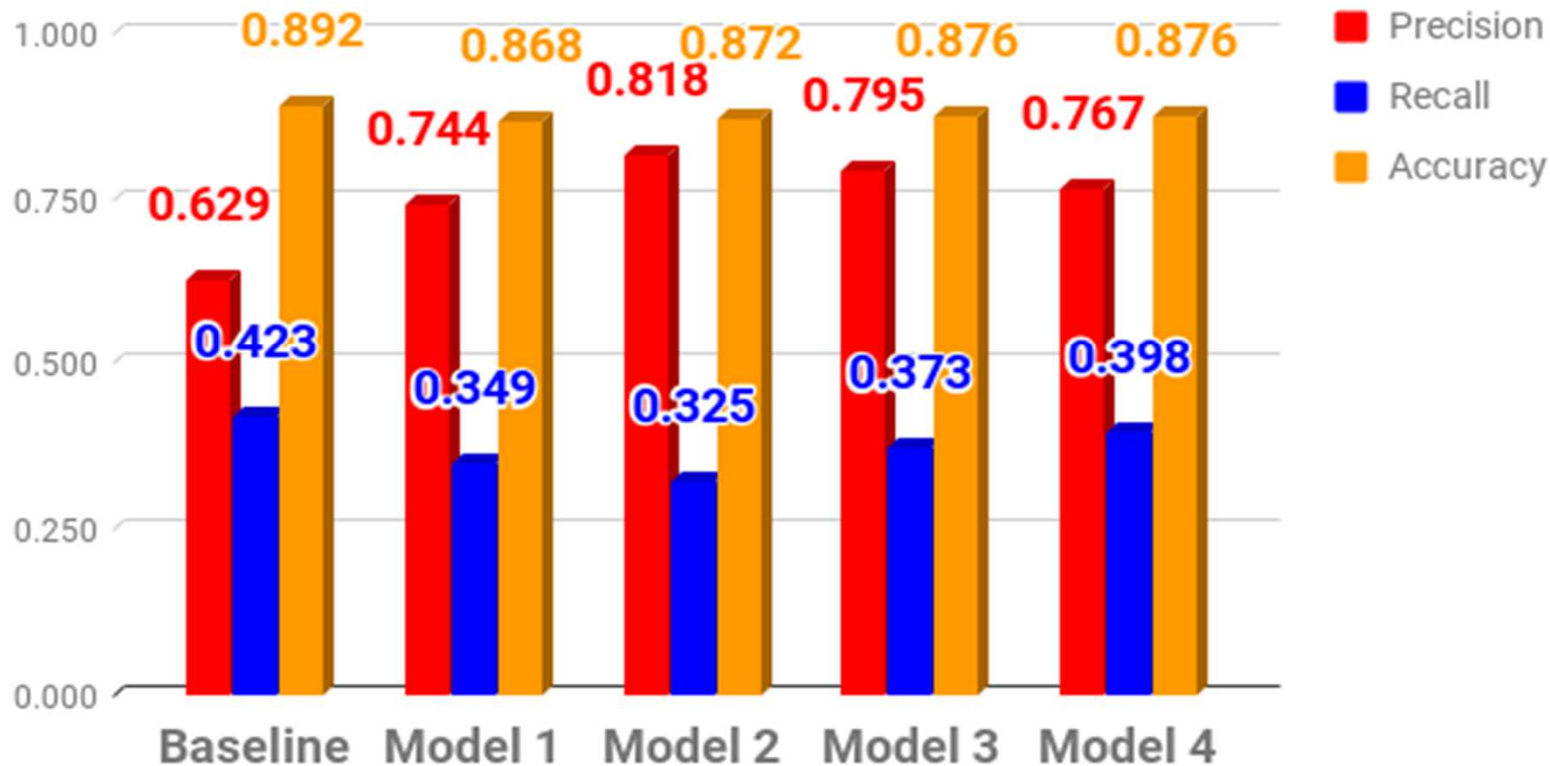


Training: 403 Non-depressive + 83 Depressive
Testing: 349 Non-depressive + 52 Depressive

D. Losada, F. Crestani. *A Test Collection for Research on Depression and Language Use*. Conference and Labs of the Evaluation Forum (CLEF), 2016

RESULTS

Precision/Recall for "Positive" and Accuracy (Random Forest)



Model 1: Baseline + Group 1: Language Style;

Model 2: Features of Model 1 + Group 2: Self-Preoccupation;

Model 3: Features of Model 2 + Group 3: Reminiscence and Relevant Words;

Model 4: Features of Model 3 + Group 4: Sentiment and Emotion.

ML FOR DETECTING LOCATION



Jamie Fox tonight
in **Budapest**
waiting for
someone to jump
off the 4th floor of
a building.

**CONTAIN a
LOCATION**

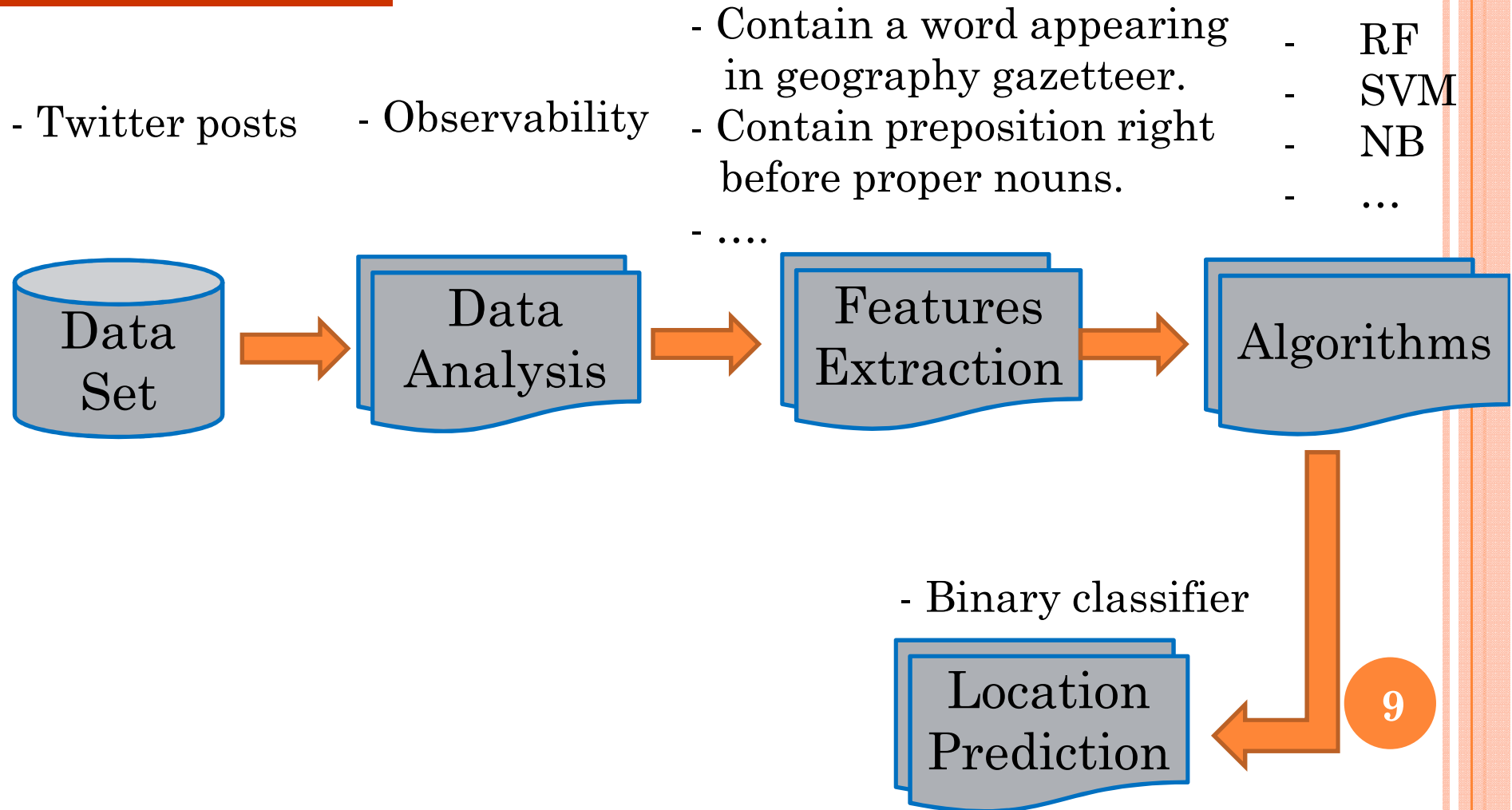
If you love **NY**,
watch this video
and take action
against



Predicting and recognizing
location names in tweets

ML FOR DETECTING LOCATION

Predictive model



ML FOR DETECTING LOCATION

- **Features**
- **Collections**
 - **Existing** (Ritter's; MSM2013)
 - **New** (1% tweet)
- **ML: Naïve Bayes, SVM, Random Forest**
- **Results**
 - **Accuracy: from 84 % to 94% (prediction of occurrence)**
 - **Increases accuracy of location extraction**
- **Current: predict diffusion**

Thi Bich Ngoc Hoang, Josiane Mothe,
Location Extraction from Tweets
IPM (submitted) 2017

RESEARCH TOPICS

- Learning from Social Media
- Contextualization
 - Developed a CLEF task 2011-> now
 - Aim: provide context to help short text understanding
 - Collection: 1,000 tweets + Wikipedia
 - Mean: multi-document summarization
 - Evaluation: informativeness & readability
 - Outcome and main results

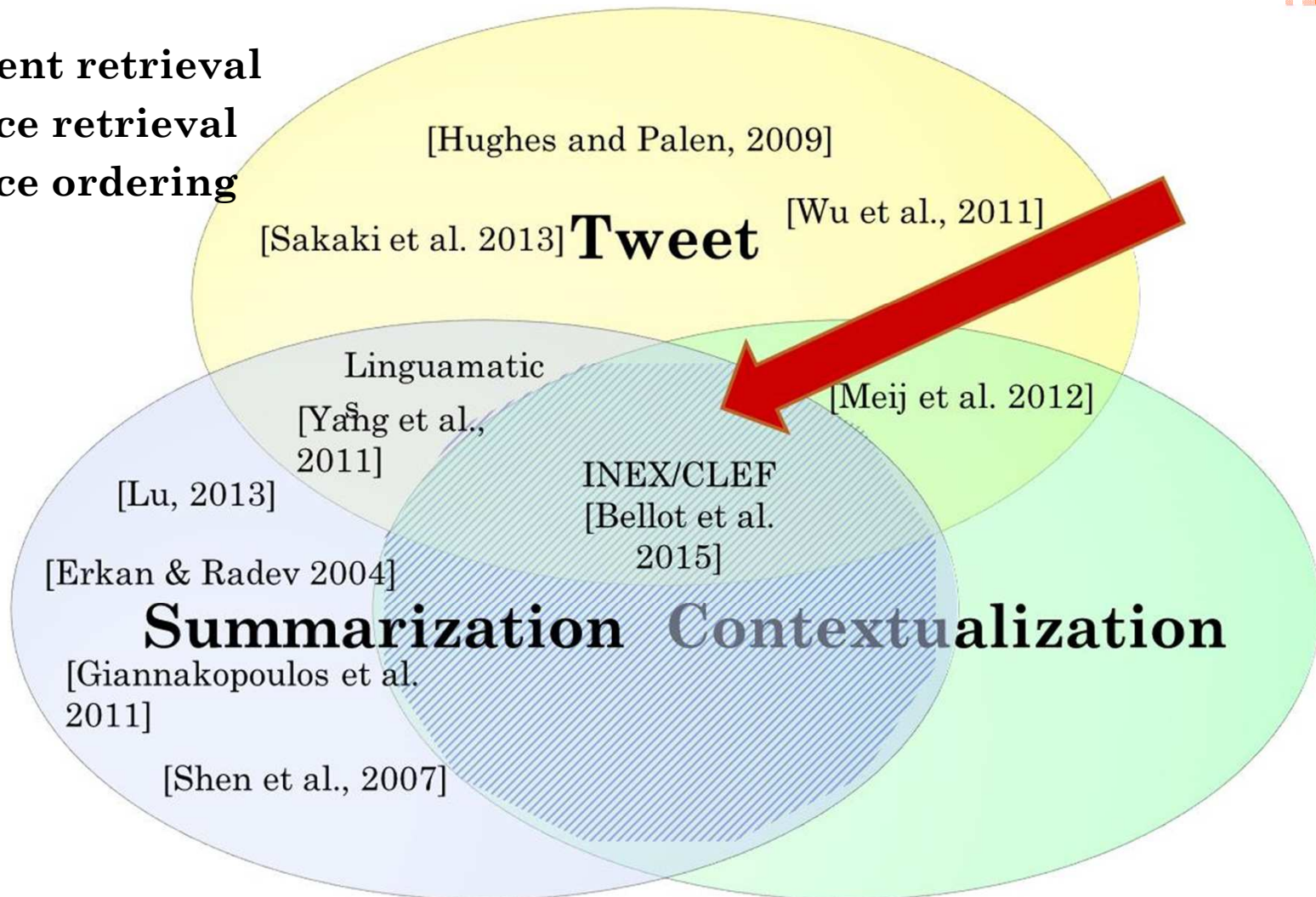
Patrice Bellot et al.

INEX Tweet Contextualization task: Evaluation, results and lesson learned.
IPM 52(5):801-819, 2016

CONTEXTUALIZATION

- **Method**

- Document retrieval
- Sentence retrieval
- Sentence ordering



CONTEXTUALIZATION

○ Contribution

- **Sentence weighting using many parameters**
 - Content
 - Syntax
 - Dependences
- **Graph theory to order sentences**
- **Topic/comment model**

Liana Ermakova.

A Method for Short Message Contextualization: Experiments at CLEF/INEX.
2015 (CLEF conference)

RESEARCH RESULTS

○ Trust in information

- How trust in Wikipedia evolves: a survey of students aged 11 to 25
Josiane Mothe, Gilles Sahut

○ Event dissemination

- [News Dissemination on Twitter and Conventional News Channels](#)
A Seth, S Nayak, J Mothe, S Jadhay

○ Detecting changes in behaviour

- Early detection of depression (CLEF e-risk 2017)
F. Benamara, Z. He, J. Mothe, V. Moriceau, F. Ramiandriosa,

○ Detection of locations in short messages

- Location extraction from tweeter
T. B. N. Hoang, J. Mothe IPM 2017 (submitted)

○ Short text (tweet) contextualization

- Task and collections (CLEF 2011 – 17)
P. Bellot et al.
INEX Tweet Contextualization task: Evaluation, results and lesson learned. IPM 52(5):801-819, 2016
- Propositions
Liana Ermakova (PhD), CLEF 2015