SYSTEAM Presentation

Nathalia, Fabien, Yongchao, Josep

Stratégie générale

- **Enjeux principaux**
  - Approche automatique (presque)
  - Equilibre entre précision et rappel
  - Croiser les sources d'information en différentes langues
  - Combiner différentes approches pour meilleur débruitage

- **Plan d'action**
  - Preprocessing
  - Detection d'évènements
  - Extraction des faits et validation

Workflow: 1ère étape

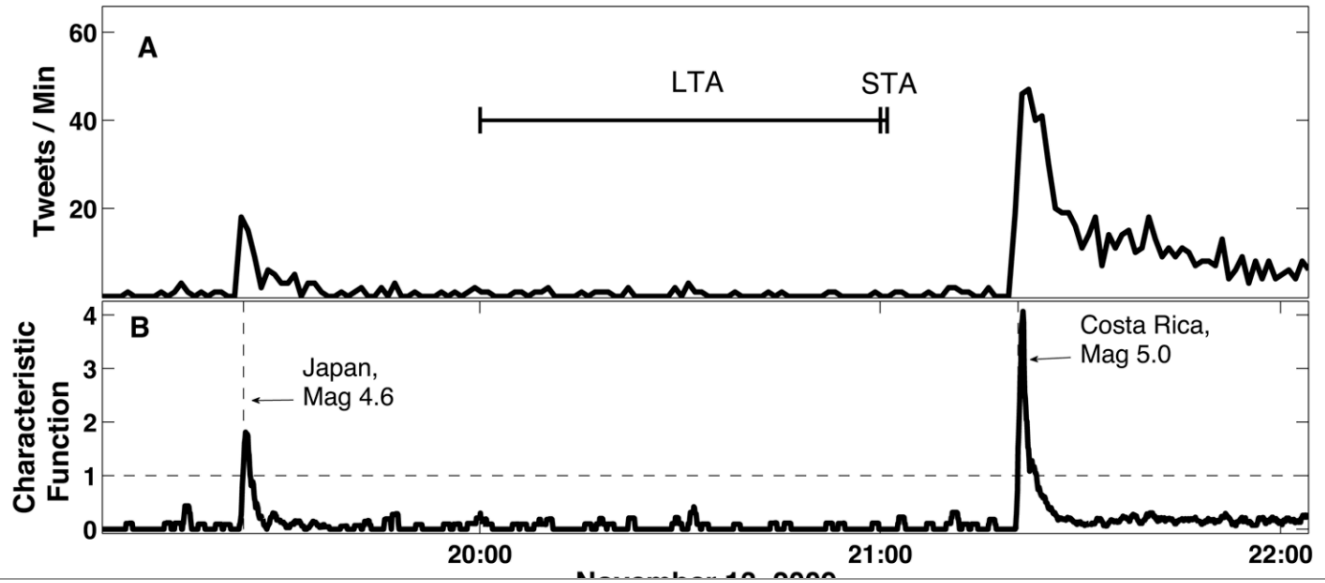
- **Preprocessing**
  - Tokenization + Normalisation (LDK pour en, fr, ar)
  - Filtrage des tweets “on domain” (clusters de #Hashtags)

Workflow: 2ème étape

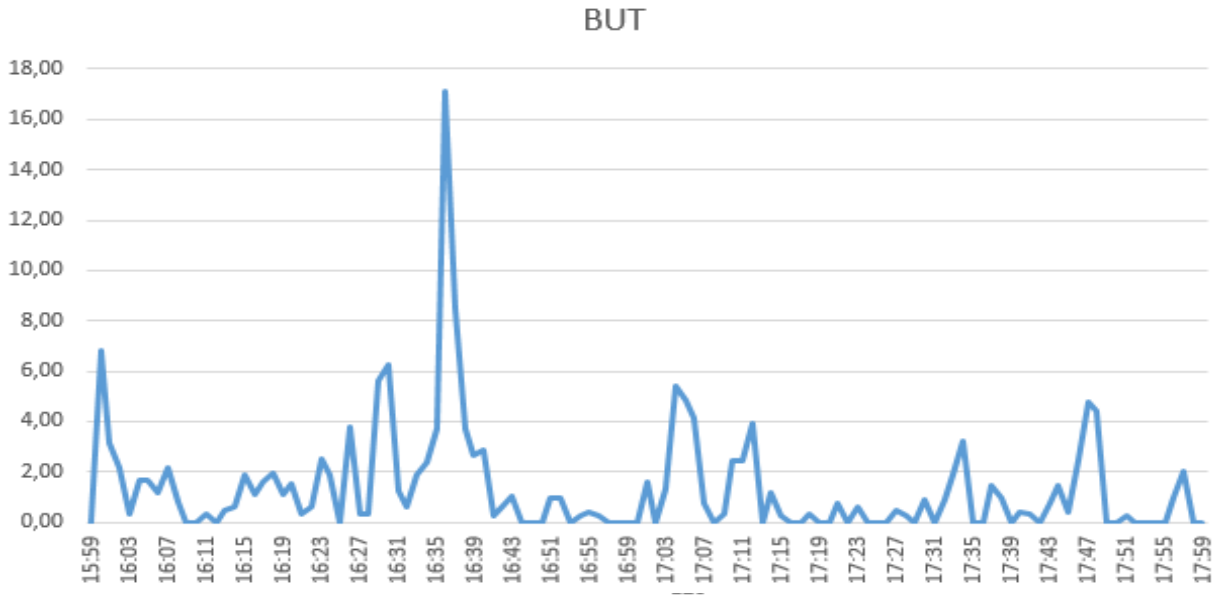
- **Détection d'évènement par minute**
  - **3 approches:**
    - ✓ Word2vec
    - ✓ Tfidf
    - ✓ Ngrams
  - **Training Data:**
    - ✓ Filtrage par minute utilisant annotations
    - ✓ Filtrage supplémentaire par mots-clés
    - ✓ Filtrage minimal pour word2vec (#EURO2016)

Workflow: 2ème étape

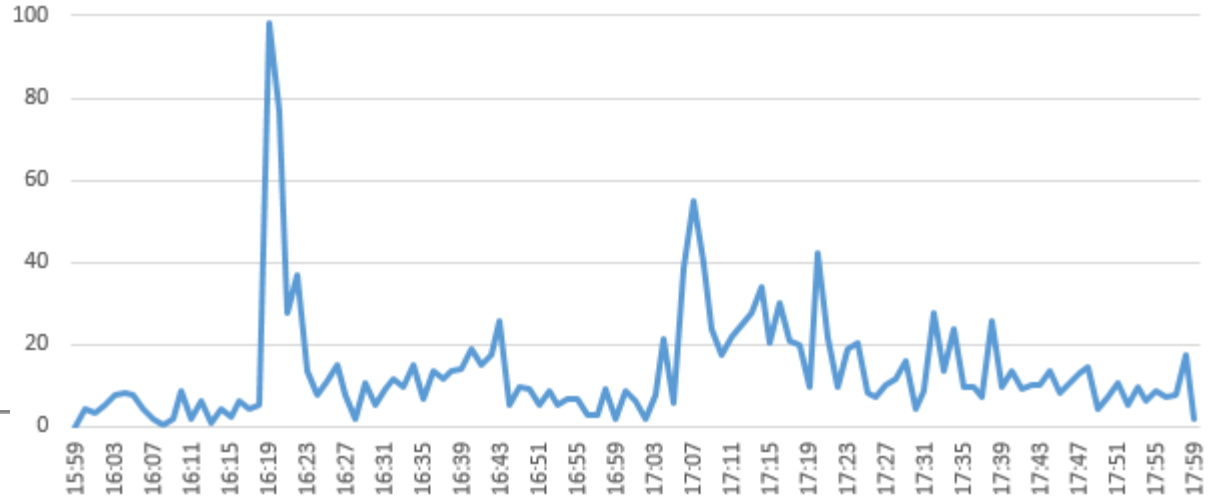
- **Détection d'évènement au niveau du match**
  - Seuillage par fênetre glissante:
    - ✓ C(t) = STA / (m * LTA + b)
      - STA: Short-Term Average
      - LTA: Long-Term Average (3min before)
      - b: Standard Deviation of LTA



Workflow: 2ème étape



TIR



SYSTRAN

Workflow: 3ème étape

- **Extraction des faits**
  - Identification des ngrams autour des faits dans les training data
  - NER (Named Entities Recognition)
  - Exemple: TIR

tête de
frappe de
la tête de
la frappe de
reprise de
poteau de
occasion de
la reprise de
coup franc de

Aperçu des résultats

- **Espagne_Croatie (TIR & BUT)**

21:04	TIR		
21:06	TIR	morata	
21:07	BUT	morata	
21:09	BUT	morata	But de Yilmaz
21:09	TIR	morata	
21:11	TIR	morata	
21:15	BUT	morata	False Positive
21:15	TIR	morata	
21:17	TIR	ivan	
21:19	TIR	rakitic	
21:24	TIR	morata	
21:29	TIR	morata	
21:33	TIR	morata	
21:45	BUT	kalinic	
21:47	BUT	kalinic	
21:51	TIR	morata	
21:56	TIR	burak	
22:04	TIR	burak	
22:11	TIR	kalinic	
22:14	TIR	burak	
22:22	TIR	modric	
22:24	TIR	ozan	
22:28	TIR	tufan	
22:30	TIR	payet	
22:32	TIR	kalinic	
22:34	TIR	kalinic	
22:38	TIR	will	
22:42	TIR	burak	
22:44	BUT	perisic	

Axes d'amélioration

- **Séparation des tweets par match:**
  - Filtrage par tags en amont
  - Validation par entités détectés en aval

- **Utiliser un(des) classifieur(s) à la place du tuning des paramètres**
  - Utiliser les scores comme features
  - Combiner plusieurs classifieurs simples

- THANK YOU!