Multilingual Fake News Detection with Satire

Gaël Guibon[1,2], Liana Ermakova[3], Hosni Seffih[4,5], Anton Firsov[6], Guillaume Le Noé-Bienvenu[7]

(1) **Télécom Paris**, Palaiseau, France
(2) **SNCF**, **Innovation & Recherche**, Saint-Denis, France
(3) **HCTI – EA 4249, Université de Bretagne Occidentale**, Brest, France
(4) **IUT de Montreuil, Université Paris 8**, **LIASD-EA4383**, France
(5) **GeolSemantics**, Gentilly, France
(6) **Knoema Corporation Perm**, Russia
(7) **CNRS, IRISA**, Rennes, France

Context

Related Work: Some key points

- Satirical Tweet detection often **rely on account-level labels** (Burfoot et al. 2009; Barbieri et al. 2015; Volkova et al. 2017; etc.)
- Satire datasets are often **unbalanced** (Yang et al. 2017)
- Fd
- Fdfdf
- Fdfdf
-

Hackathon on NLP: HackaTAL 2018

Task

Identify and **Categorize** Fake News

Participants

6 teams ; 40 participants

Data Provider

Storyzy company "storyzy
Specialized in **disinformation** and
websites classification

Data Topic

Vaccination Fake News Dataset

Data Size

6358 articles

**Data Information**

EN & FR: mainly English

Dataset

Fake News Dataset

- Train | Test distribution = 4767 | 1591

- Restricted information on Test

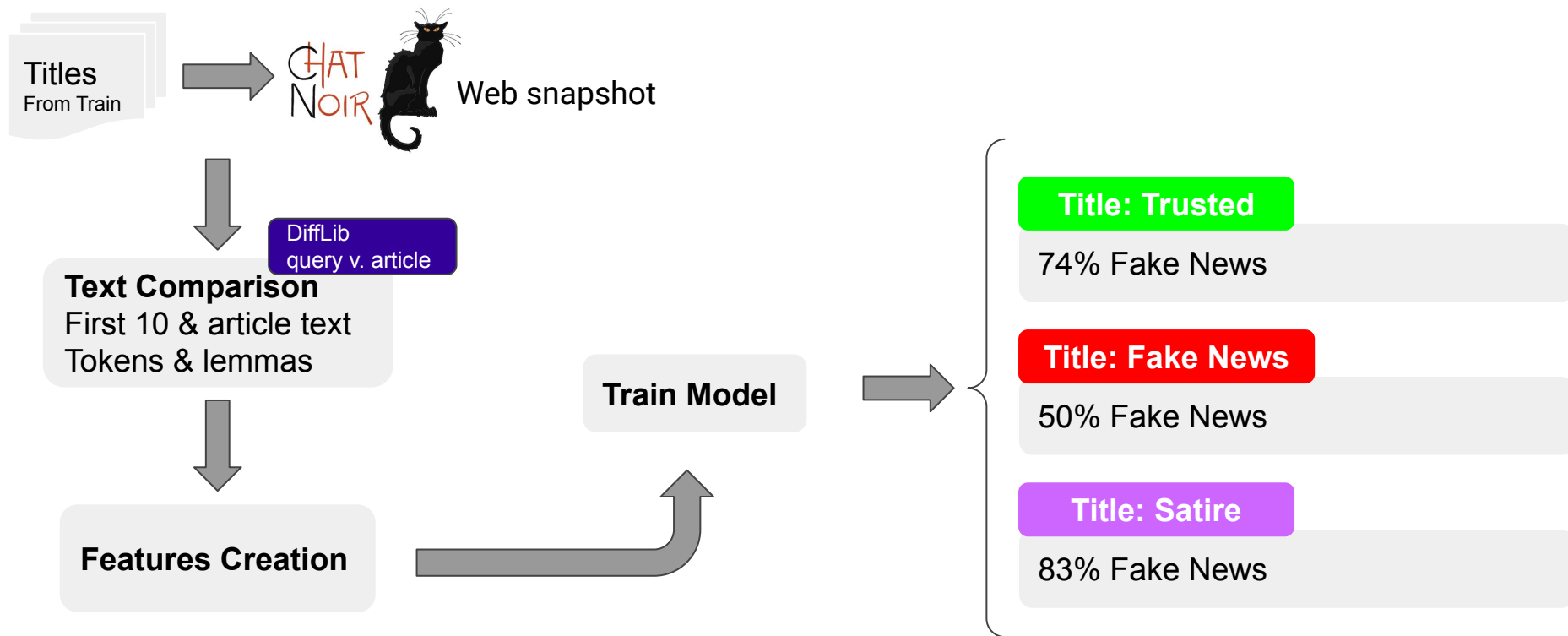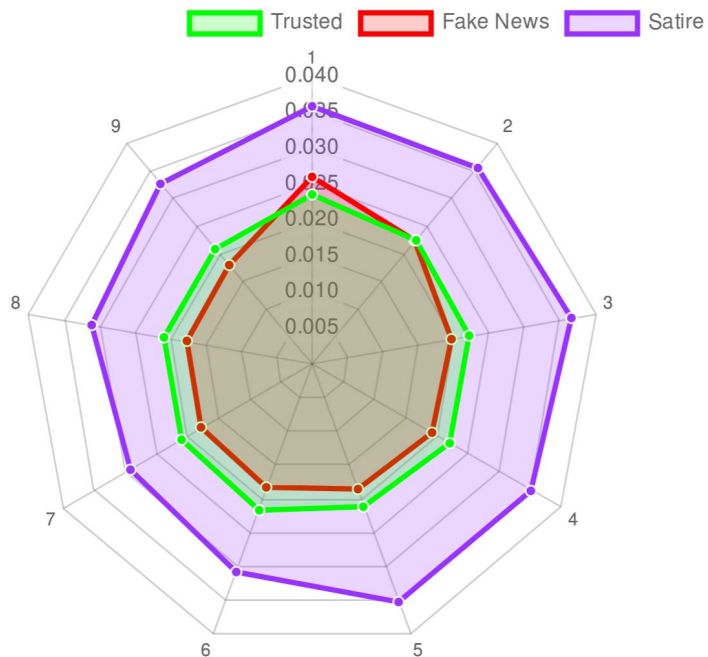| Language	Train	Test	Train format	Test format
English	3828	1277	id, domain, type, uri, author, language, title, text, date, external_uris	id, title, text
French	705	236	id, domain, type, uri, author, language, title, text, date, external_uris	id, title, text
YouTube	234	78	video-id, channel-id, video-title, video-view-count, lang, type, channel-title, text, id	id, video-title, text

Multilingual Dataset
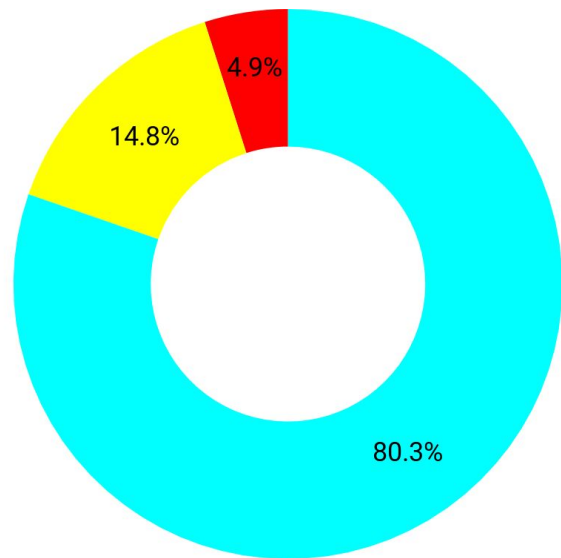
English
French
YouTube (FR)

80.3%
14.8%
4.9%

3 classes

Trusted

Fake News

Satire

Textual Features and Exploration

Domain Type Detection

Titles
From Train

Web snapshot

DiffLib
query v. article

Text Comparison
First 10 & article text
Tokens & lemmas

Features Creation

Train Model

Title: Trusted

74% Fake News

Title: Fake News

50% Fake News

Title: Satire

83% Fake News

Text Resemblance

Average Comparison

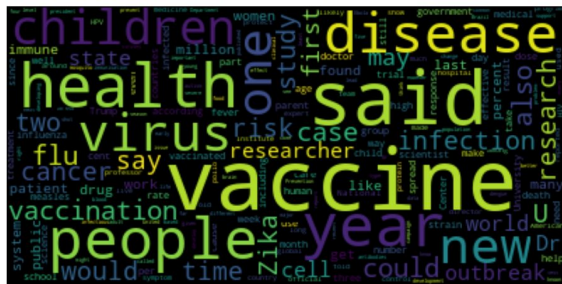**Train text** + **ChatNoir** results

Irregular Ratios

Irregular ratios for Fake and Satire
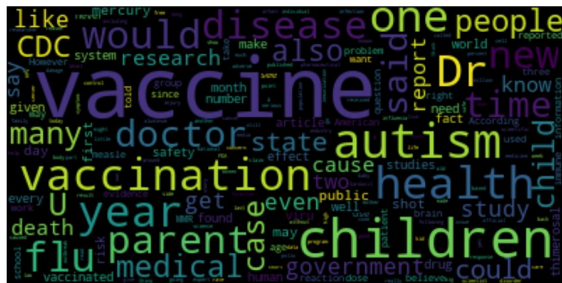
Satire Peculiar Subjects

Facebook, Pharma, Science, Monsanto, etc.
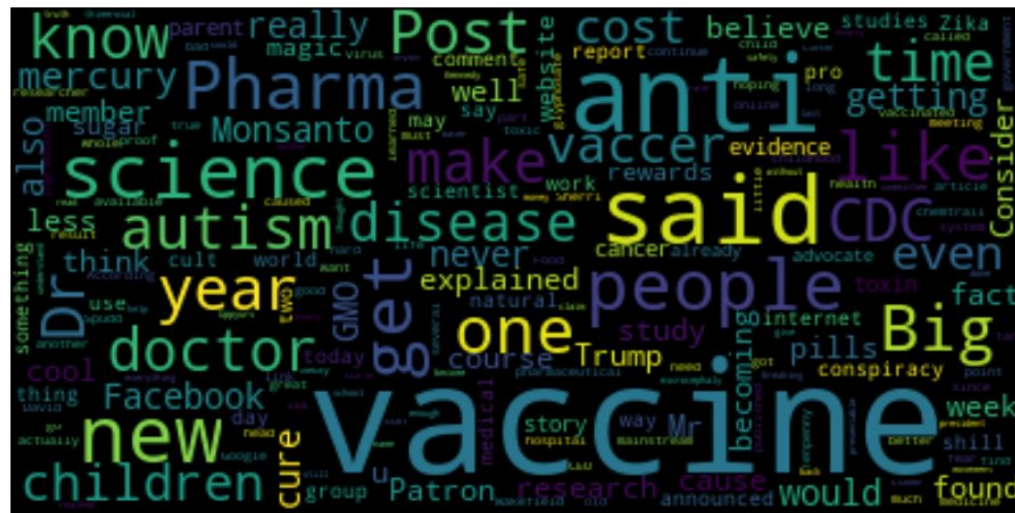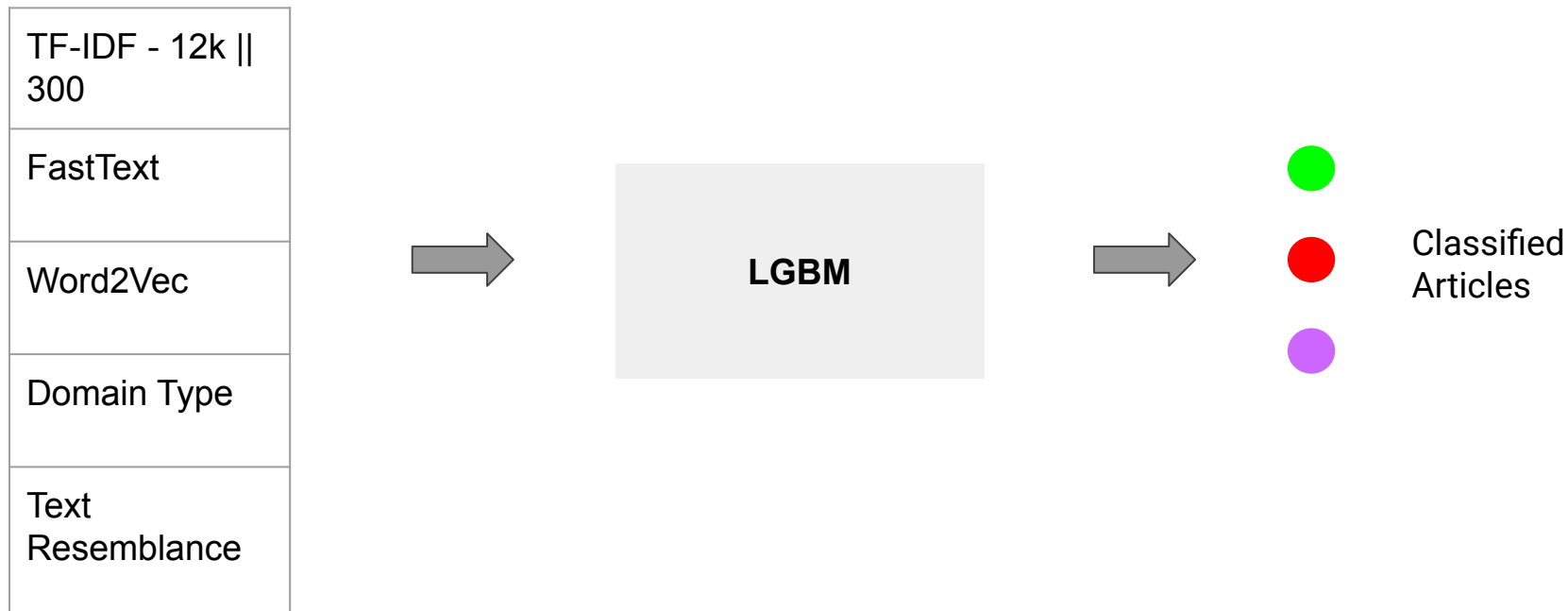
Recurrent Topics per Class in English

Trusted

Fake

Satire

Classification Task

Classification: Feature Stacking + LGBM

TF-IDF - 12k || 300
FastText
Word2Vec
Domain Type
Text Resemblance

LGBM
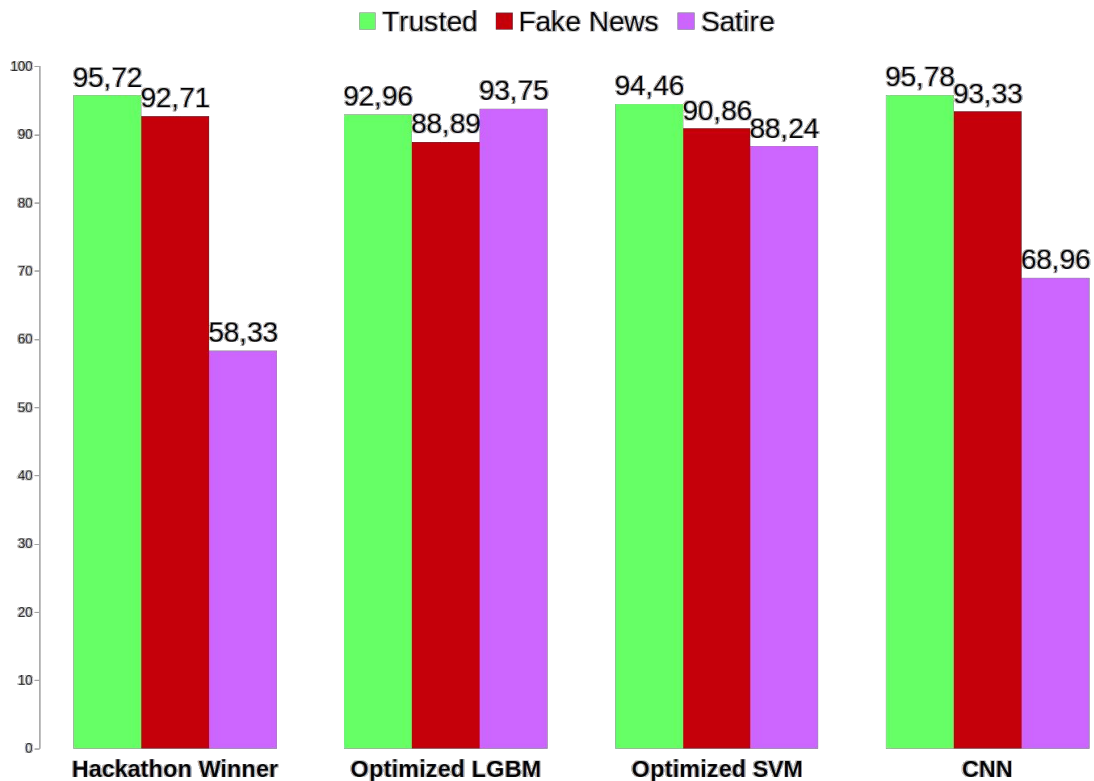
Classified Articles

Classification: CNN

- Features = Embedding layer

- Allows approach comparison
  - Standard embedding
  - Feature Stacking

InputLayer
ragged = false
sparse = false

Embedding
batch_input_shape = null, 10
input_dim = 1000
input_length = 10
output_dim = 64

Conv1D
activation = relu
filters = 512
groups = 1
kernel_size = 2

Dropout
rate = 0.4

GlobalMaxPooling1D
data_format = channels_last

Dense
activation = softmax
units = 3

dense_14

Classified Articles
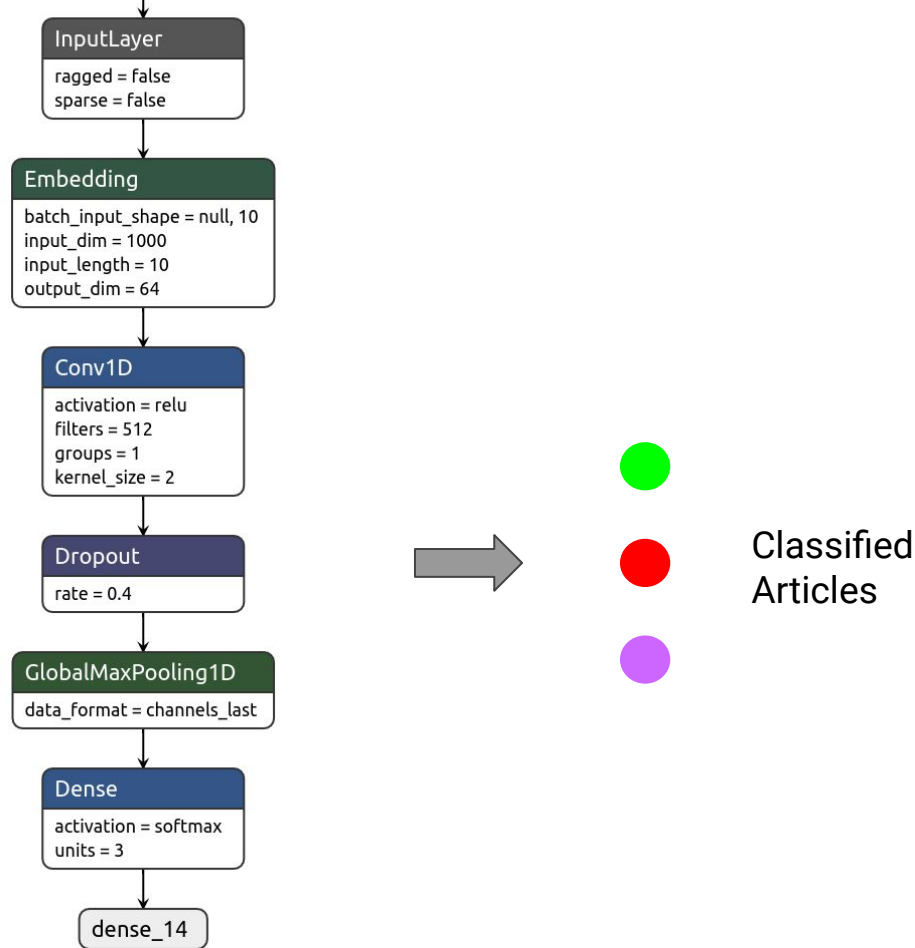
Classification: F1 scores

Some conclusions

- Stacking > scaling: opposite to (Burfoot et al. 2009) on Satire Detection
- Results indicates decision tree based algorithms are better suited for the task
  - Confirmed later by (Ozbay et al. 2020)
-