Profiling Fake News Spreaders on Twitter based on TFIDF Features and Morphological Process

Notebook for PAN at CLEF 2020

M. Lichouri* **M. Abbas*** *B. Benaziz**

CLEF 2020, 24-25 September 2020 - Thessaloniki, Greece.

Contact: **m.lichouri@crstdla.dz**; **m.abbas@crstdla.dz**

*: Computational Linguistics Department, CRSTDLA, Algeria

Summary

1 Task information
- Multilingual Fake News Detection

2 Characteristics, problems and solutions
- Characteristics
- Problems
- Solution

3 Experimental Setup and Evaluation
- Experimental Setup
- Evaluation

4 Conclusion

Author Profiling for Fake News Detection

Is-it possible to detect Fake News Spreaders based on their profile?

Fake news is a threat to the

- Presidential Election like the US's 2016.
- Public health like in the time of Coronavirus.
- Security where it can lead to war.
- Commerce like in the Bourse Market.

Problems

How to and who decide which is real and which is fake?

Solution

Naive assumption

We can detect and decide if a news is fake or not by either:

1. Checking facts.
2. Profiling the users who reported the news.

Corpus Statistics

	English	Spanish
# authors (XML files)	300	300
# sentences per author (XML file)	30,000	30,000
# words per author (XML file)	717,596	786,965
Max # word per author (XML file)	3,636	5,373
Min # word per author (XML file)	1,524	1,603
Max # char per author (XML file)	12,962	23,588
Min # char per author (XML file)	5,238	5,799

Table: PAN Train set statistics for both English and Spanish

System Architecture

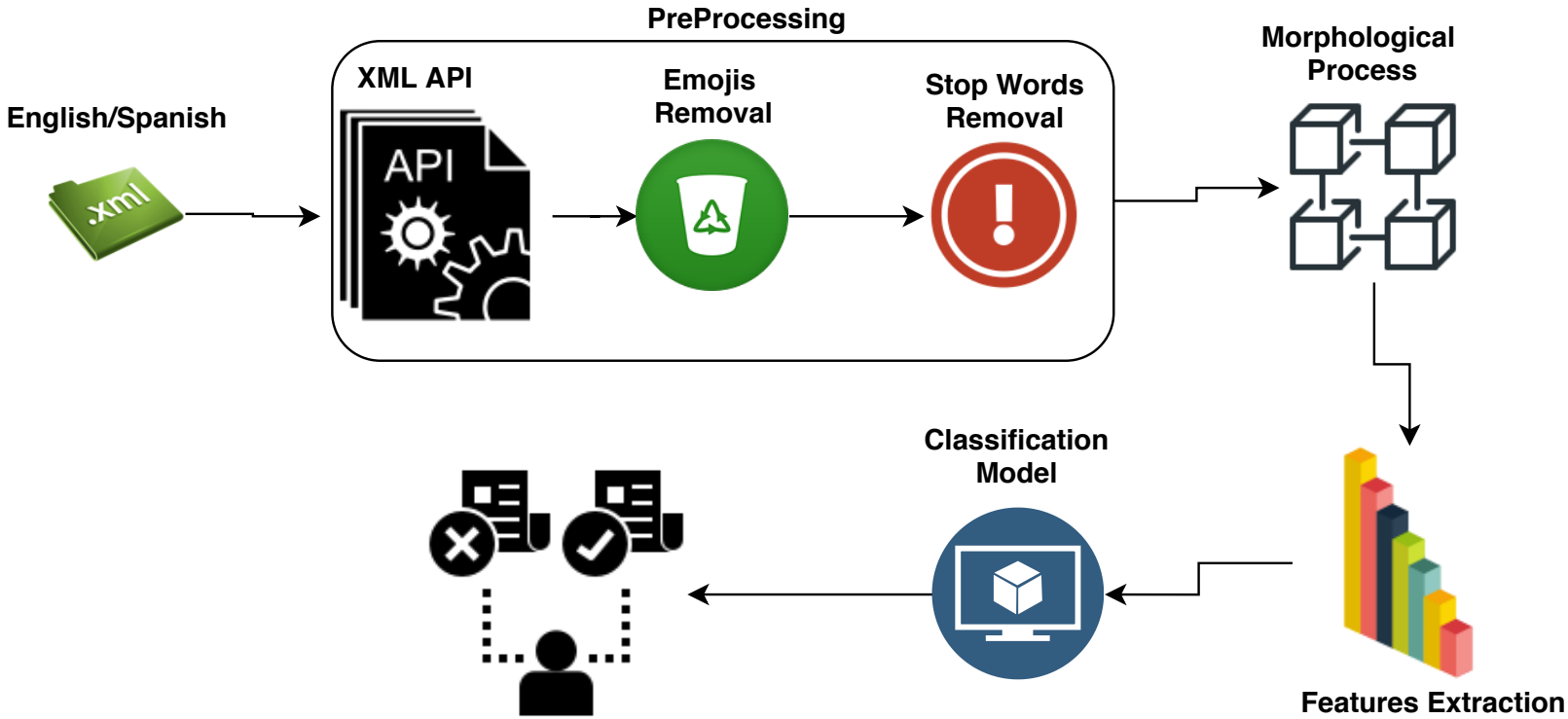
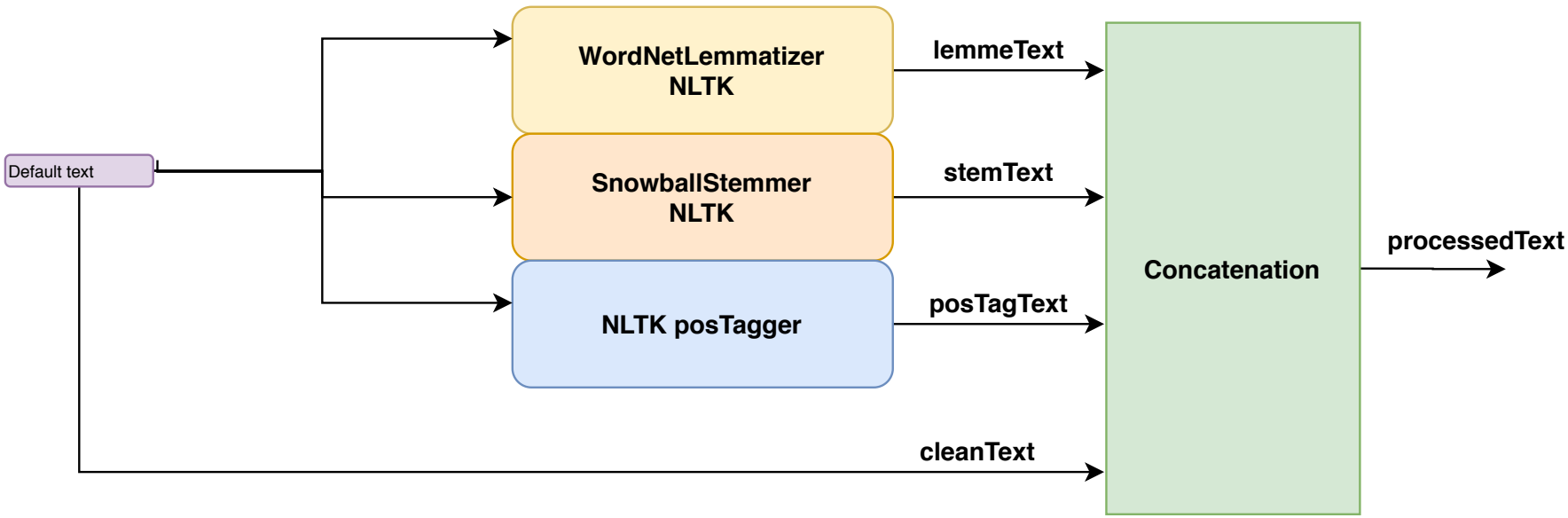


Figure: The architecture of the proposed system.

Surface Pre-processing

- Read XML files by the XML API.
- Remove Emojis using specified rules.
- Remove stop words (English/Spanish) using the NLTK stop words lists.
- Save the output as cleanText.

Morphological Processing



Features Extraction using TF-IDF

1. Setup 1: TF-IDF with word 5-grams tokenizer.
2. Setup 2: TF-IDF with character 5-grams tokenizer.
3. Setup 3: TF-IDF with character with-boundary 5-grams tokenizer.
4. Setup 4: Union of the three above.

Training Phase

Dataset	Model	F1-score (%)
English	LSVC	100
	RDG	99.58
	SGD	100
Spanish	LSVC	100
	RDG	99.16
	SGD	100

Table: Results using the development set.

Test Evaluation

In the final submission, the model is trained on the whole training set.

Dataset	Model	F1-score (%)
English	LSVC	58.50
	RDG	**61.50**
	SGD	52.00
Spanish	LSVC	**76.00**
	RDG	74.50
	SGD	54.50

Table: Results of the final submission using LSVC (ranked system), RDG and SGD (Test set).

Discussion

Dataset	Model	Dev	Test
English	LSVC	100	58.50
	RDG	99.58	**61.50**
	SGD	100	52.00
Spanish	LSVC	100	**76.00**
	RDG	99.16	74.50
	SGD	100	54.50

Table: Comparison of obtained results (Dev and Test) in term of F1-score(%).

Conclusion

- We proposed a simple, yet competitive system (ranked 45/66).
- Best features: a union of three TF-IDF features (word 5-grams, char 5-grams and char_wb 5-grams), in addition to three important morphological features: stemming, lemmatization and part of speech tagging.
- Best results: Our system achieved an F1-score of 76% for Spanish and 58.50% for English.

Danger: Believe Fake News

Ref:How our brains trick us into believing fake news. https://www.yourlifechoices.com.au/the_meeting_place/post/how-our-brains-trick-us-into-believing-fake-news