

Semiautomatic dictionary-based classification of environment tweets by topic

Michela Cameletti – Stephan Schlosser – Daniele Toninelli – Silvia Fabris

Digital data are characterized by **advantages** such as reduced collection costs, short retrieval times and almost real-time outputs. New approaches are required for two **challenging** tasks:

- the **selection** of posts related to one or more specific topics;
- **retrieving information** of interest inside Twitter posts.

Tweets of **Official Social Accounts** linked to the subject of interest

A **list of keywords** is identified to set a topic-oriented **dictionary**

Method **evaluation** by applying the dictionary to more than 54 million geolocated tweets

In order to **classify** and **filter** tweets by their **content**, two main approaches have been proposed in the literature when categories are known: 1) **dictionary-based** and 2) **supervised** or **unsupervised** methods. We propose an unsupervised and dictionary-based method.

The **first** dataset is composed by tweets posted by **Official Social Accounts (OSA)** related to the analyzed topic, environment. Starting from these data, the algorithm sets up the **dictionary**. That latter is applied to the **second** dataset, composed by the tweets posted in Great Britain between January and May 2019.

~ 40,000 Tweets
of 12 OSA's

Tweets
Cleaning Corpus

**Bi- Trigrams
Hashtags**

**Dictionary Creating
Algorithm**

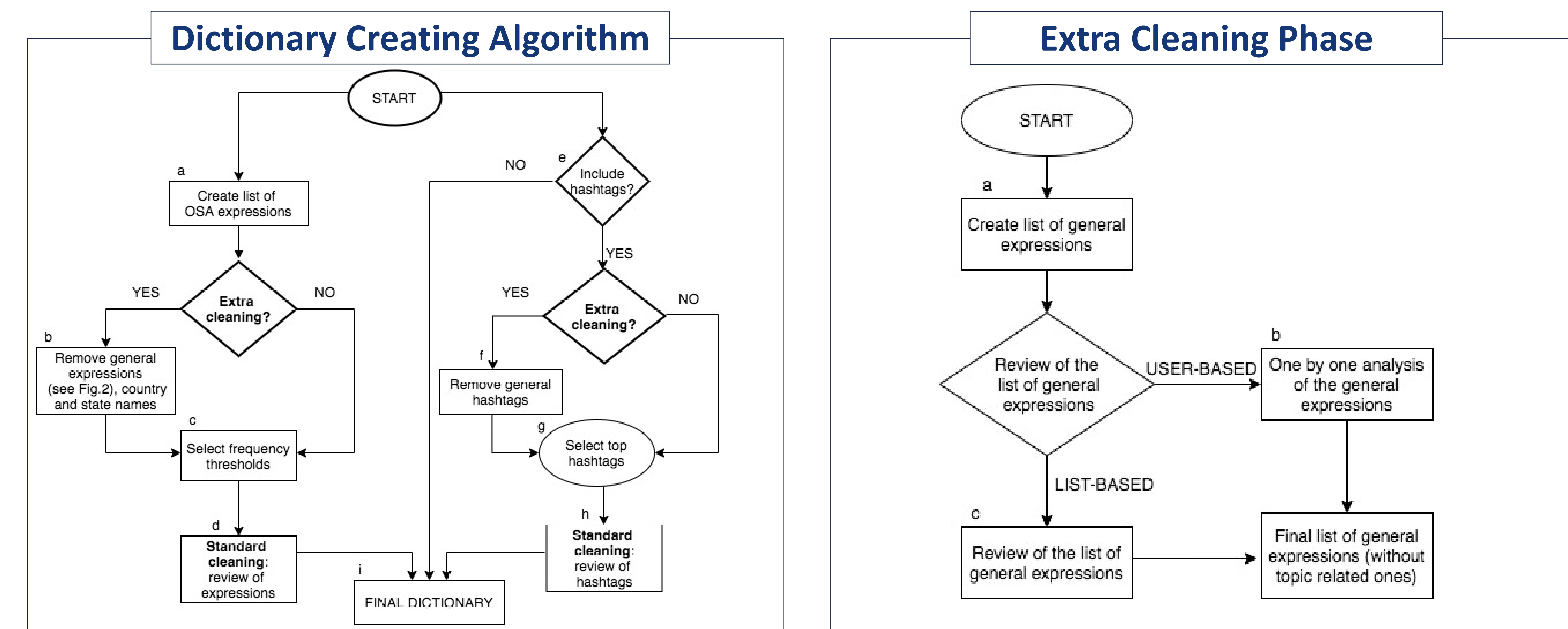
**Dictionary
"Environment"**

35 Expressions
52 Hashtags

Selection of
~ 100,000 Tweets
out of 54 million

Performance
Evaluation of
the Method

Given the tweets collected from the selected OSA and preprocessed, we produce the list of all bigrams and trigrams with the corresponding frequencies. This represents the starting point of the dictionary creation. Expressions which do not appear frequently are usually not related to the topic or are too general to be included in the final dictionary. For this reason, in order to select the most pertaining bigrams and trigrams, some additional steps are required.



In order to evaluate the performance of the dictionary-based filtering, we randomly choose 600 tweets selected and 600 not selected by the algorithm. Then, we manually classify these posts into two categories: "related" and "non-related" to environment. Our method is able to **correctly classify 98.42%** of the total number of tweets.

	Accuracy	Sensitivity	Specificity	Precision	F ₁ Score
Performance Index (% values)	98.42	99.32	97.55	97.50	98.40

- This method can be **easily applied to any topic of interest**.
- Instead of using pre-set and already-available dictionaries, the user can create an **ad-hoc and personalized dictionary**.
- The method does not rely on a single or on a short list of predefined keywords, the **list can be expanded as necessary to be updated and renewed any time it is needed**.
- The method relies on a dictionary that is not based on just single words, but on **combinations of words** (i.e. on bigrams and trigrams), thus **reducing the inclusion of non-pertinent tweets**.