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.

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.

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.

---

**Dictionary Creating Algorithm**

```
START
Create list of general expressions

1. By one-to-one analysis of the general expressions
2. List-based
3. Review of the list of general expressions
4. Final list of general expressions

IF YES
Create list of specific expressions
ELSE
NO
```

---

**Extra Cleaning Phase**

- **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.**

---

**Performance Evaluation of the Method**

<table>
<thead>
<tr>
<th>Performance Index (% values)</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>98.42</td>
<td>99.32</td>
<td>97.55</td>
<td>97.50</td>
<td>98.40</td>
</tr>
</tbody>
</table>

---

**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.

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.

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.