Integrating Social Network Structure into Online Feature Selection

Antonela Tommasel
PhD. Advisor: Daniela Godoy

Why Online Feature Selection?

Efficient and scalable Online Feature Selection (OFS) becomes a crucial requirement of numerous large-scale social applications.

- Short-texts **accentuate** the **challenges** posed by the **high feature space dimensionality** of text learning tasks.
- The **linked** nature of **social data** causes **new dimensions** to be added to the feature space, which, also becomes **sparser**.

However…

- Most studies have focused on developing batch techniques.
- Existing OFS techniques are not always accurate enough and are not sufficiently efficient when handling massive-scale data with high-dimensionality.
- Most techniques do not leverage on the linked nature of social media data.

Proposal!

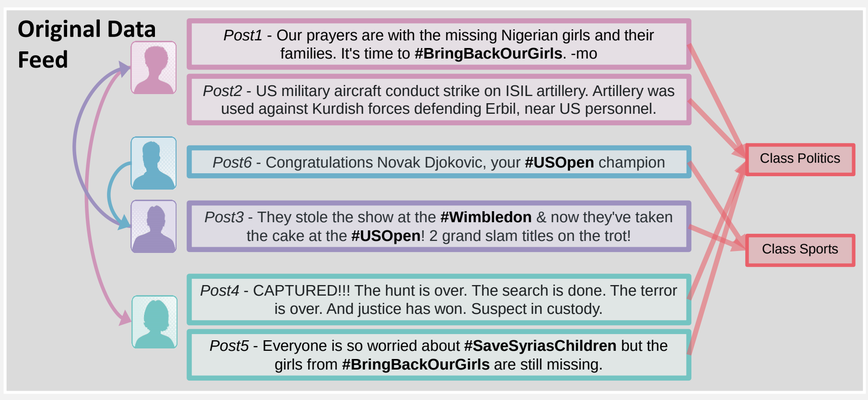
An **Online Feature Selection** technique for **high-dimensional** data based on both **social** and **content-based information** for the **real-time** classification of **short-text** streams coming from social media.

Objectives?

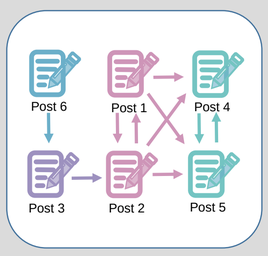
- Enhancing the process of knowledge discovery in social-media.
- Help in the development of new and more effective models for personalisation and recommendation in social environments.

Social-based OFS

Addresses the massive-scale OFS task for high-dimensional short-text data arriving in a continuous stream, in which neither features nor data instances are fully known in advance.

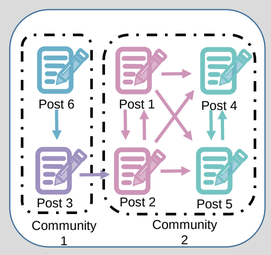


1. ***Data Modelling*** as a graph representing the social posts and their relations.



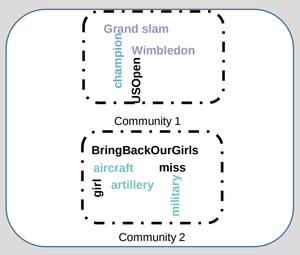
Relations:
- Social
  - Follower/Followee
  - Friends
- Content-based
  - Share tags
  - Similar content
  - Same class

2. ***Social Analysis Step.*** Social relationships between posts are analysed to find groups of socially related posts. Community detection techniques are explored.



- Content-based relations can use to weight social relations.
- Content-based relations can create new relationships between nodes.
- Clustering Consensus techniques can be applied.

3. ***Content Analysis Step.*** An optimal feature set to describe each group of socially related posts is found.



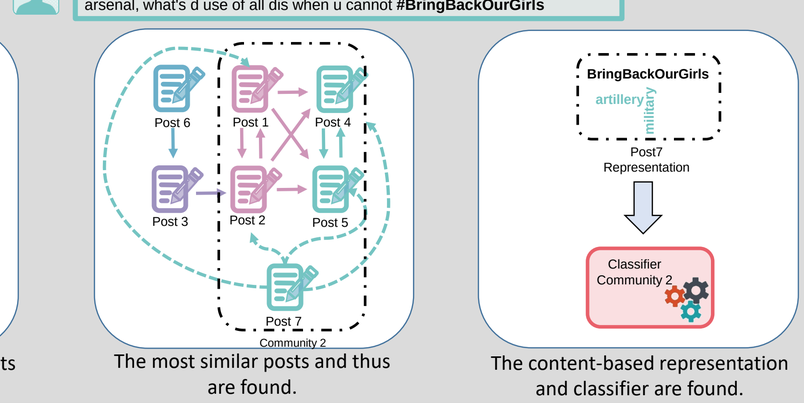
Text features must be relevant and non-redundant.

- Evaluates feature Redundancy.
  - Correlation.
  - Information Theory.
- Evaluates feature Relevance.
  - Ranking.

4. ***Model Learning.*** Content features are used for training learning models for classifying newly arriving posts.

Each community has its specialised learning model adapted to its particular textual features.

5. ***Arrival of new posts.***



The relations with the known posts are exploited.

The most similar posts and thus are found.

The content-based representation and classifier are found.

6. ***Classification of new posts.*** Once the most similar posts and are found, the post is assigned to a category.

Diverse alternatives are explored:

- *Single*. Selects the most similar community and assigns the post to the category with the highest probability.
- *Average*. Selects the N most similar communities and averages the probabilities of a post belonging to each category. The post is assigned to the category with the highest averaged probability.
- *Voting*. Selects the N most similar communities and applies a voting scheme. The post is assigned to the most voted category.

7. ***Re-run of the Social Analysis Step.*** After new posts are classified, the feature space is updated to cope with the continuous evolution of social data.

- New topics constantly emerge.
- Topics become obsolete and tend to disappear.

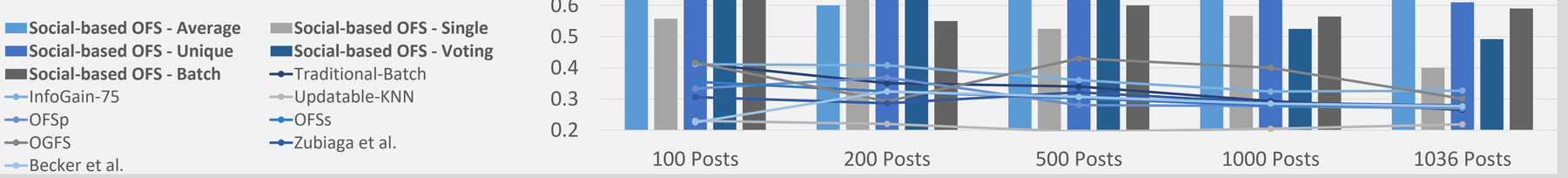
There is still work to do!

- Extensive experimental evaluation must be performed.
- New alternatives for further exploiting the social relations and communities are being considered.
- Several methods for assessing the redundancy and relevance of features will be analysed.

Current State

- Preliminary evaluations achieved **promising** results when compared to traditional and state-of-the-art in both batch and online settings!!
- The obtained results exposed the limitations of **pure content-based techniques** for classifying social media short-texts.
- Encouraging results were obtained for the community detection step when combining social and content-based information.

Leveraging on social information becomes *crucial* for OFS!



Contributions

- Tackles the **challenging** problem of **Online Feature Selection**.
- Addresses the problem of **how** to **exploit** the **linked nature** of social media data.
- Proposes a technique for exploiting the multiple relations among social data.
- **Combines** **social information** with **content** for effectively and efficiently performing feature selection.

antonela.tommasel@isistan.unicen.edu.ar