Integrating Social Network Structure into Online Feature Selection

ANTONELA TOMMASEL

ADVISOR: DANIELA GODOY

ISISTAN
Instituto Superior de Ingenieria de Software Tandil

Motivation

- Short-texts **accentuate** the **challenges** posed by the **high feature space dimensionality** of text learning tasks.

- The **linked** nature of **social data** causes <u>**new dimensions**</u> to be added to the feature space, which, also becomes **sparser**.

Efficient and scalable online feature selection becomes a crucial requirement of numerous large-scale social applications.

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.

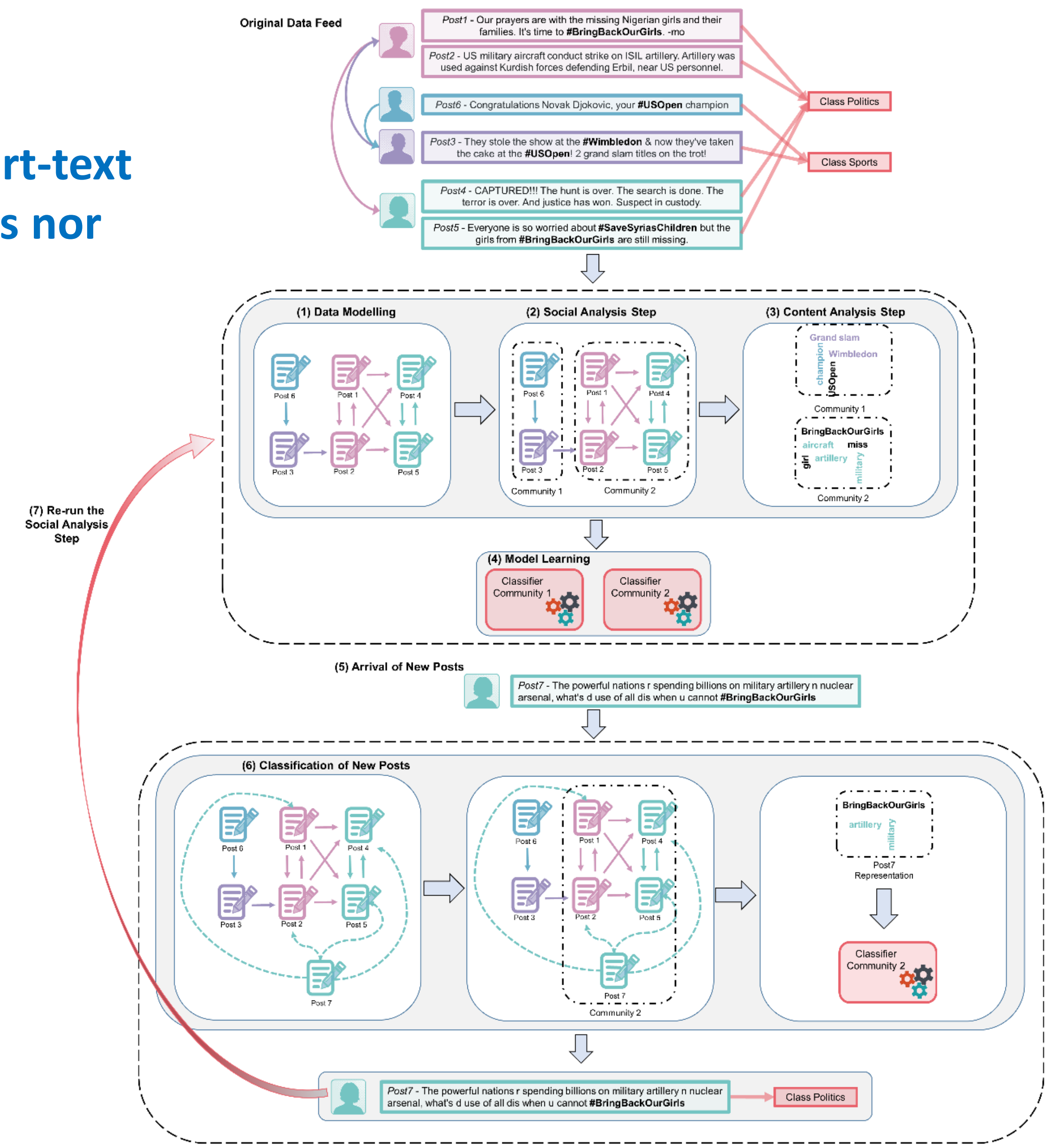
- *Objectives*?
  - Enhancing the process of knowledge discovery in social-media.
  - Helping in the development of new and more effective models for personalisation and recommendation of content in social environments.

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.

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.


Original Data Feed
Post1 - Our prayers are with the missing Nigerian girls and their families. It's time to #BringBackOurGirls. -mo
Post2 - US military aircraft conduct strike on ISIL artillery. Artillery was used against Kurdish forces defending Erbil, near US personnel.
Post6 - Congratulations Novak Djokovic, your #USOpen champion
Post3 - They stole the show at the #Wimbledon & now they've taken the cake at the #USOpen! 2 grand slam titles on the trot!
Post4 - CAPTURED!!! The hunt is over. The search is done. The terror is over. And justice has won. Suspect in custody.
Post5 - Everyone is so worried about #SaveSyriasChildren but the girls from #BringBackOurGirls are still missing.
Class Politics
Class Sports
(1) Data Modelling
Post 6
Post 1
Post 4
Post 3
Post 2
Post 5
(2) Social Analysis Step
Community 1
Community 2
(3) Content Analysis Step
Grand slam
Wimbledon
champion
USOpen
Community 1
BringBackOurGirls
aircraft
miss
girl
artillery
military
Community 2
(4) Model Learning
Classifier Community 1
Classifier Community 2
(5) Arrival of New Posts
Post7 - The powerful nations r spending billions on military artillery n nuclear arsenal, what's d use of all dis when u cannot #BringBackOurGirls
(6) Classification of New Posts
Post 7
Community 2
BringBackOurGirls
artillery
military
Post7 Representation
Classifier Community 2
Post7 - The powerful nations r spending billions on military artillery n nuclear arsenal, what's d use of all dis when u cannot #BringBackOurGirls
Class Politics
(7) Re-run the Social Analysis Step

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.

2. ***Social Analysis Step.*** Social relationships between posts a analysed to find groups of socially related posts.

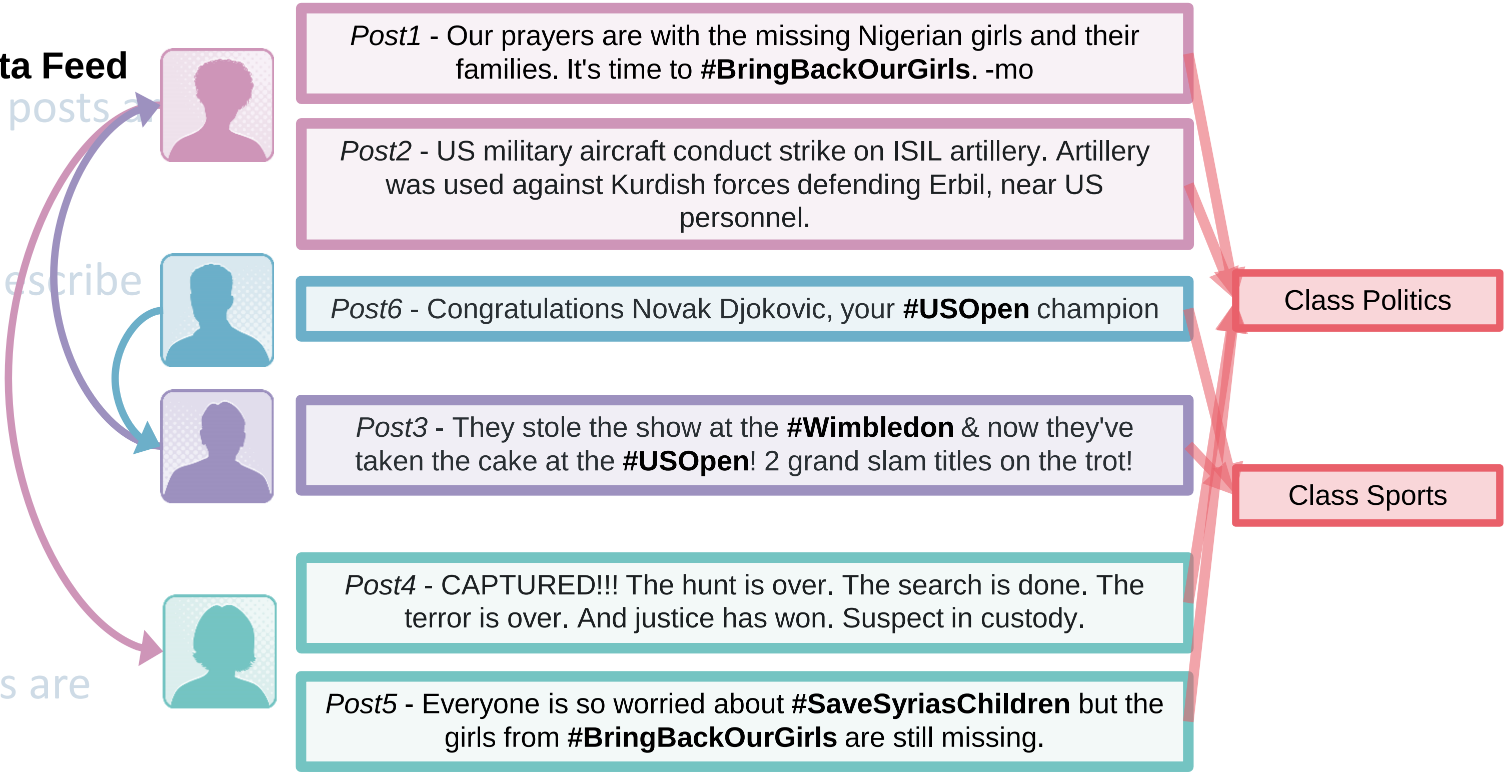
3. ***Content Analysis Step.*** An optimal feature set to describe each group of socially related posts is found.

4. ***Model Learning.***

5. ***Arrival*** and classification of new posts.

6. ***Re-run of the Social Analysis Step****.* After new posts are classified, the feature space is updated.

Original Data Feed



Post1 - Our prayers are with the missing Nigerian girls and their families. It's time to **#BringBackOurGirls**. -mo

Post2 - US military aircraft conduct strike on ISIL artillery. Artillery was used against Kurdish forces defending Erbil, near US personnel.

Post6 - Congratulations Novak Djokovic, your **#USOpen** champion

Post3 - They stole the show at the **#Wimbledon** & now they've taken the cake at the **#USOpen**! 2 grand slam titles on the trot!

Post4 - CAPTURED!!! The hunt is over. The search is done. The terror is over. And justice has won. Suspect in custody.

Post5 - Everyone is so worried about **#SaveSyriasChildren** but the girls from **#BringBackOurGirls** are still missing.

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.

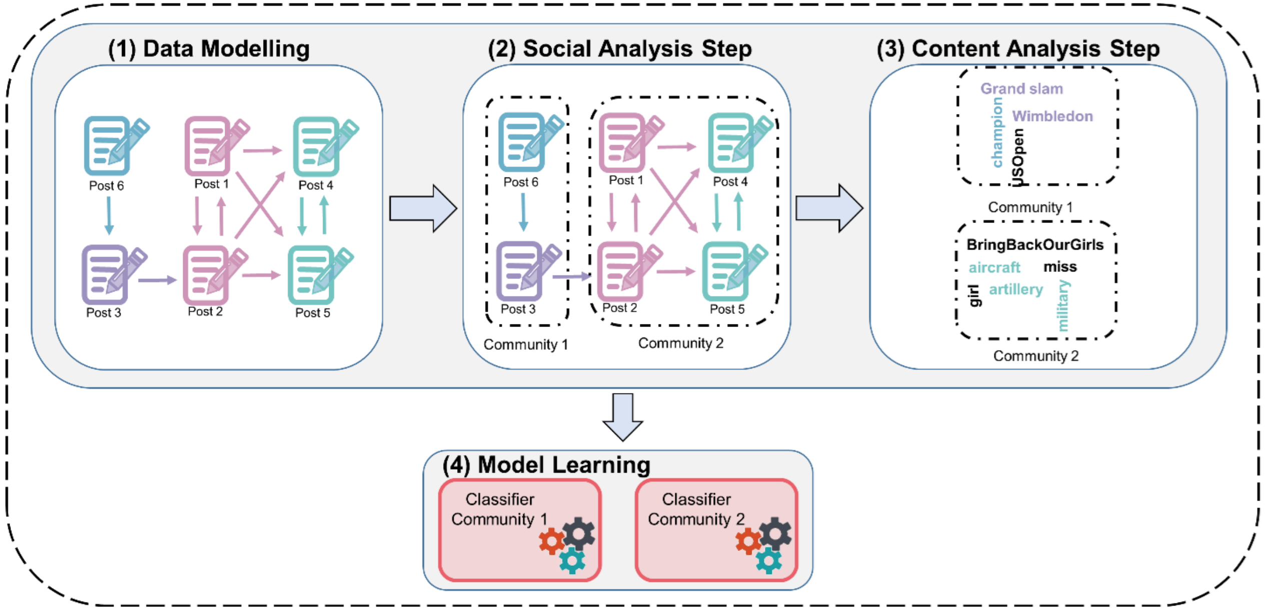
2. ***Social Analysis Step.*** Social relationships between posts are analysed to find groups of socially related posts.

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

4. ***Model Learning.***

5. ***Arrival*** and classification of new posts.

6. ***Re-run of the Social Analysis Step.*** After new posts are classified, the feature space is updated.



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.

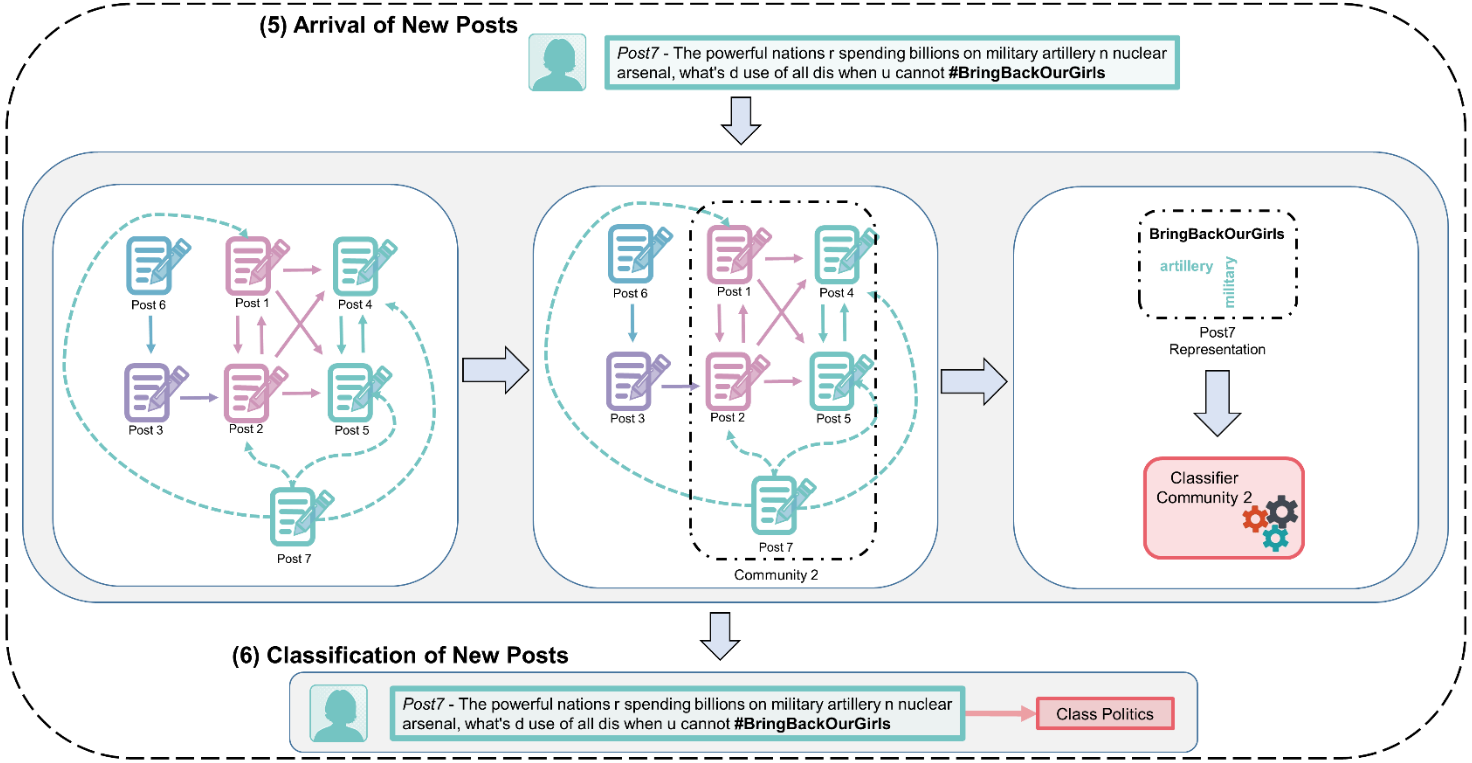
2. ***Social Analysis Step.*** Social relationships betwee analysed to find groups of socially related posts.

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

4. ***Model Learning.***

5. ***Arrival*** and classification of new posts.

6. ***Re-run of the Social Analysis Step.*** After new pos classified, the feature space is updated.



(5) Arrival of New Posts

Post7 - The powerful nations r spending billions on military artillery n nuclear arsenal, what's d use of all dis when u cannot **#BringBackOurGirls**

(6) Classification of New Posts

Post7 - The powerful nations r spending billions on military artillery n nuclear arsenal, what's d use of all dis when u cannot **#BringBackOurGirls**

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.

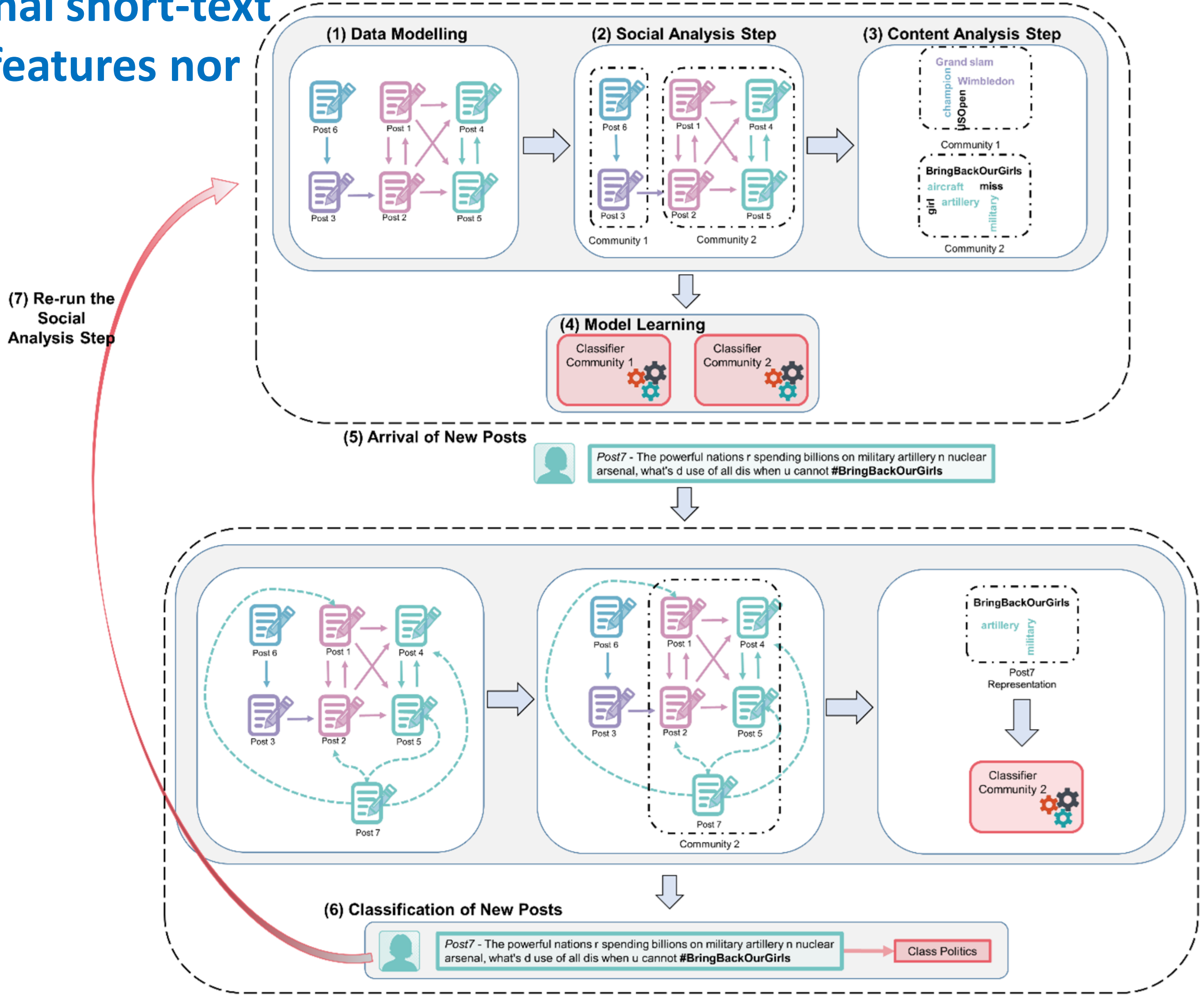
2. ***Social Analysis Step.*** Social relationships between posts are analysed to find groups of socially related posts.

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

4. ***Model Learning.***

5. ***Arrival*** and classification of new posts.

6. ***Re-run of the Social Analysis Step.*** After new posts are classified, the feature space is updated.



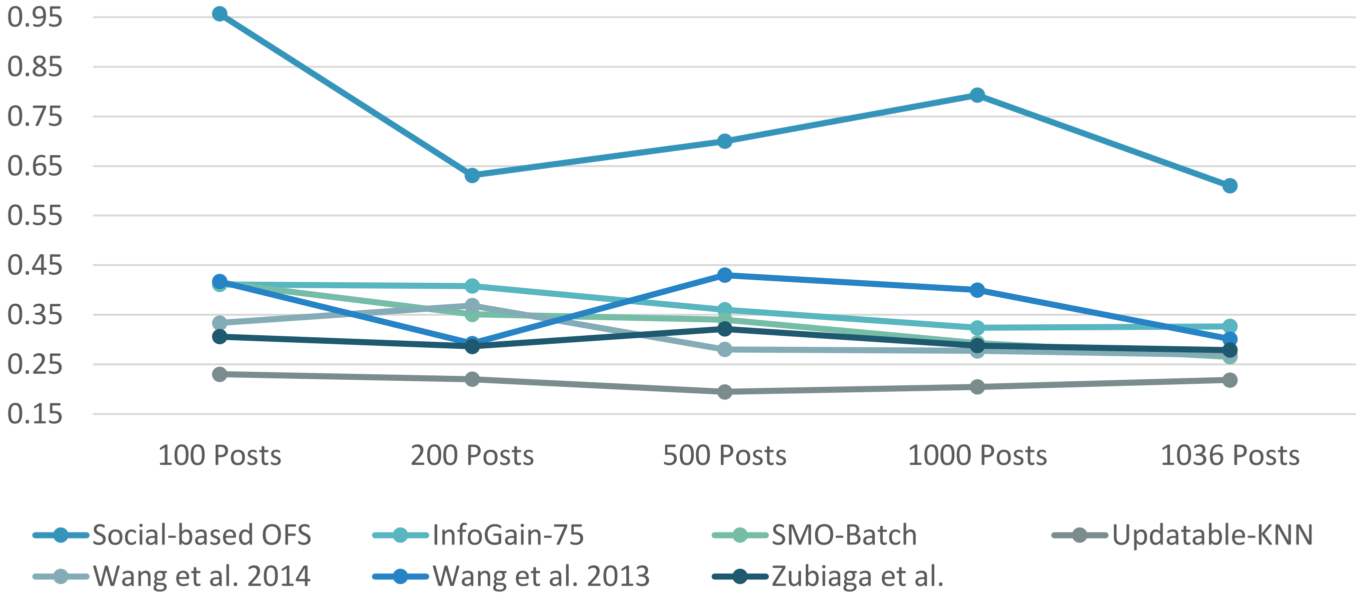
Post7 - The powerful nations r spending billions on military artillery n nuclear arsenal, what's d use of all dis when u cannot **#BringBackOurGirls**

Post7 - The powerful nations r spending billions on military artillery n nuclear arsenal, what's d use of all dis when u cannot **#BringBackOurGirls**

Current State

- Preliminary evaluations conducted on two real-world short-texts datasets achieved **promising** results when compared to traditional and state-of-the-art in both batch and online settings!!

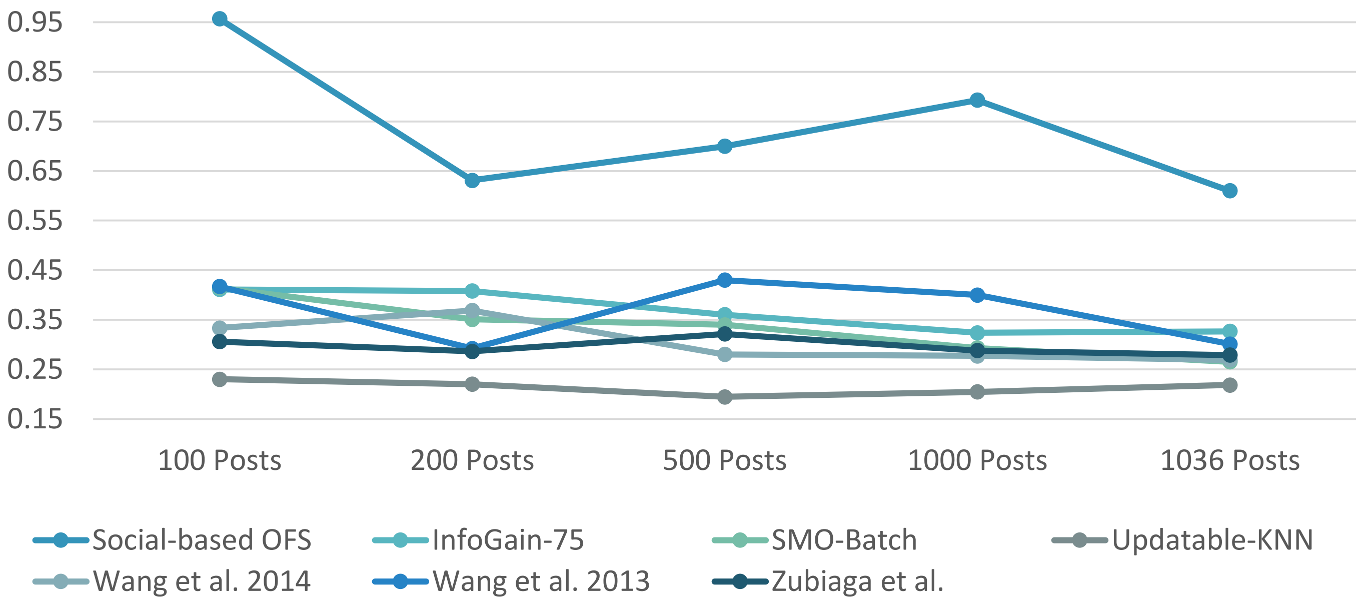
- The obtained results exposed the limitations of pure content-based techniques for classifying social media short-texts.



Current State

- Preliminary evaluations conducted on two real-world short-texts datasets 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.



Leveraging on social information becomes crucial for OFS.

Contributions

- This thesis 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 **leveraging on social relations**.
- **Combines social information** with **content** for effectively and efficiently performing feature selection.
- **Scalability**. Appropriate for real-time environments in which neither features nor instances are known in advance.

Questions?

Integrating Social Network Structure into Online Feature Selection

ANTONELA TOMMASEL

ADVISOR: DANIELA GODOY

ISISTAN
Instituto Superior de Ingenieria de Software Tandil