Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu

D2.3 Initial Release of Segmentation, Summarization, and News Classification Capabilities

Work Package	2
Responsible Partner	IMCS
Author(s)	Tugtekin Turan, Sebastião Miranda, Afonso Mendes
Contributors	Guntis Barzdins
Reviewer	Salima Mdhaffar
Version	V1.0
Contractual Date	31 March 2022
Delivery Date	31 March 2022 (Resubmitted 25 May 2022)
Dissemination Level	Public

Version History

Version	Date	Description
0.1	13/01/2022	Introduction and Initial Table of Contents (ToC)
0.2	11/03/2022	Section on PiniTree added
0.3	29/03/2022	Priberam input added
0.4	30/03/2022	Internal Review version
0.5	30/03/2022	Finalization
1.0	31/03/2022	Publishable version
1.5	17/05/2022	Unmerged deliverables
2.0	25/05/2022	Publishable version

Executive Summary

This initial deliverable describes the first release of software components developed within the WP2 and integrated with the SELMA orchestration platform. This first batch of components is mainly applicable to use case 1 (UC1), the “Media Monitoring Platform”.

SELMA targets to create stream-based models that can leverage updated news information to improve topic classification. Using SELMA’s monitoring platform, users can keep track of current trending news stories across multiple languages.

This document provides an overview of the initial release of the “Multilingual Topic Classification” and “Online Multilingual News Clustering” capabilities to follow the interim and final releases later in the project.

Table of Contents

Table of Figures

1. Introduction

This initial deliverable describes the first release of software components developed within WP2 and integrated with the SELMA orchestration platform. This first batch of components is mainly applicable to use case 1 (UC1), the "Media Monitoring Platform". We address two components of the Natural Language Processing (NLP) document enrichment pipeline, namely, "Multilingual Topic Classification" and "Online Multilingual New Clustering". This report aims to describe those components from the point of view of the release of software components. This deliverable should be read in conjunction with D2.1, where a complete technical description is reported.

All components are deployed as Docker containers (www.docker.com) and will be made available at https://hub.docker.com/orgs/selmaproject, expose REST APIs and provide swagger documentation pages (https://swagger.io). These components are integrated with the SELMA orchestration platform and are already being used by UC1 (https://app.monitio.com).

2. Released Components

2.1 Online Multilingual News Clustering

Multilingual News Clustering is a core piece of the Media Monitoring use case as it allows the user to focus on stories instead of being overwhelmed by a huge amount of scattered news articles. Our work for the SELMA platform was presented in D2.1 and was later accepted at the Text2Story Workshop to be held at ECIR 2022. Unlike other enrichment tasks like Topic Classification or NER/NEL, news clustering is not easily scalable to handle a very big incoming stream of documents and thus can be a bottleneck in the processing pipeline. The reason for this is that the clustering of a document is dependent on the status of the clustering pool at a certain point in time, making it impossible to scale by simply adding additional workers. The performance of the component is thus a key aspect when integrating into the pipeline.

The currently deployed version is capable of processing 3.2 documents per second on average while maintaining a cluster pool of 25,000 clusters. Given that the platform is currently ingesting about 150,000 documents per day, this seems an acceptable performance, nevertheless, and since the flow is not uniform during the day, we already see times when there is a perceivable delay between the ingestion and the clustering. Research efforts are being carried in WP2 in order to obviate this problem. The current model handles 50 languages *(en, ar, bg, ca, cs, da, de, el, es, et, fa, fi, fr, fr-ca, gl, gu, he, hi, hr, hu, hy, id, it, ja, ka, ko, ku, lt, lv, mk, mn, mr, ms, my, nb, nl, pl, pt, pt-br, ro, ru, sk, sl, sq, sr, sv, th, tr, uk, ur, vi, zh-cn, zh-tw),* as described in D2.1. This model is *state-of-the-art* in the task of online multilingual news clustering as reported in D2.1 and in our publication at the Text2Story workshop "Simplifying News Clustering Through Projection from a Shared Multilingual Space" by João Santos, Afonso Mendes, and Sebastiao Miranda.

The clustering engine is deployed as a docker container that exposes a REST API which exposes the following methods:

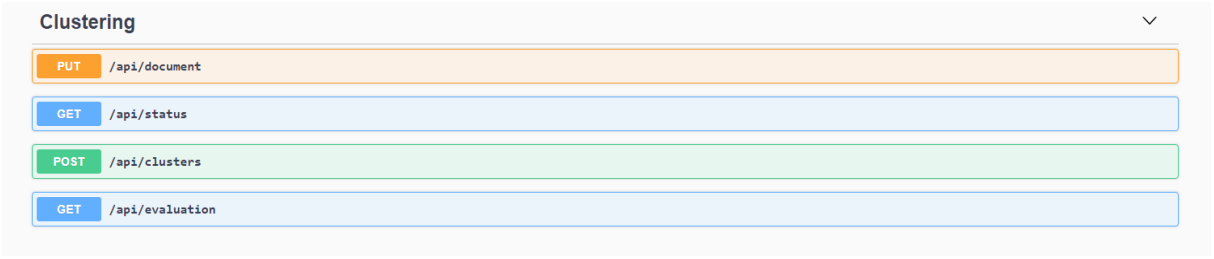

Figure 1: *NER Annotation Tool*

The first method /api/document receives a document and returns the clustering ID to associate the document with, the method also returns pairs of (doc id, cluster id) when previously clustered documents, due to a cluster merge, are reassigned after the current operation. The /api//status and /api/clusters are for state monitoring purposes and the /api/evaluation performs an evaluation of the algorithm as a sanity check before deployment.

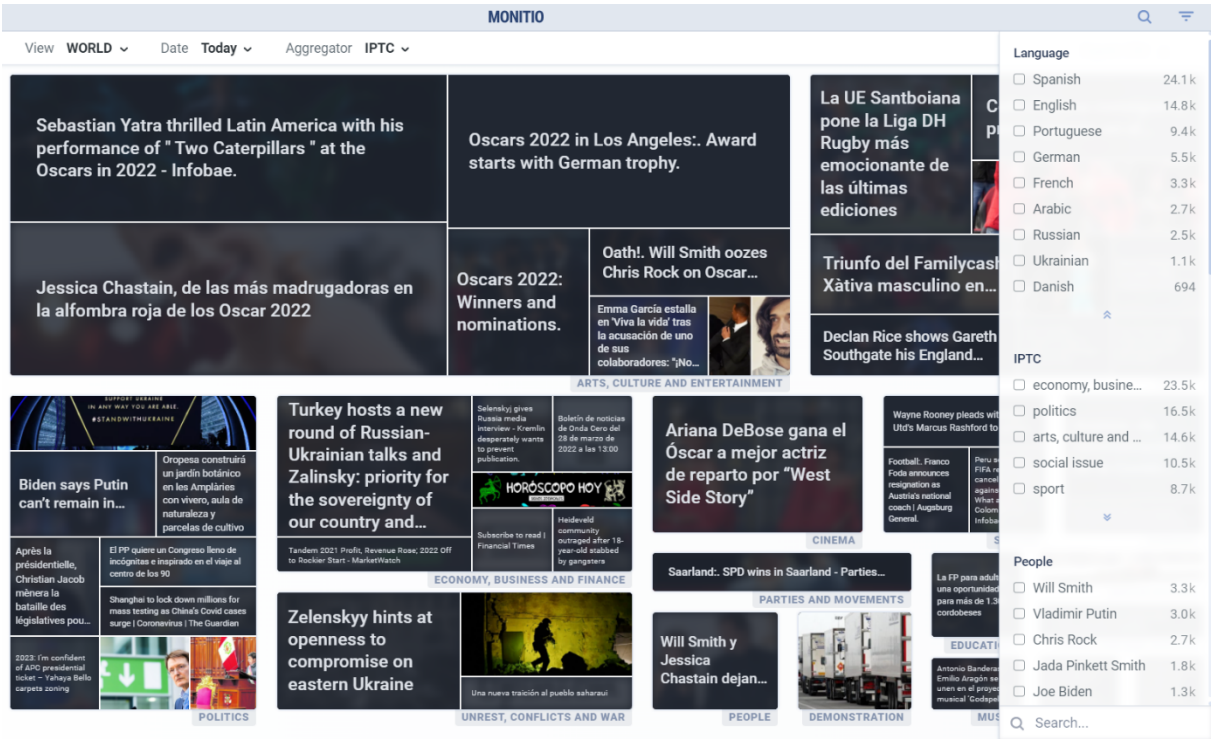

Figure 2: *"Storylines" Dashboard from Monitio Platform Showing Multilingual Clustering Model*

2.2 Topic Detection

In SELMA we have two objectives regarding topic detection and news classification. The first is to enhance the multilingual capabilities of the IPTC Subject Codes classification defined by the International Press Telecommunications Council. The second is the ability of the system to classify documents against user-defined topics given a minimum amount of user feedback.

During the reporting period, as reported in D2.1, we have developed and released a new multilingual model for the IPTC topic classification task based on the AttentionXML (You et al. 2019) that improves F1 performance in the test sets for Portuguese, English, and Spanish by 5.8%, 3,6%, and 11,8%, respectively, against our previous SUMMA model. Even though the model was not yet formally evaluated in other languages, initial user feedback on the accuracy for other languages like Russian and Arabic is very good, we expect a formal user evaluation to start soon in the scope of WP5. The model was trained to leverage the contextual embeddings of mBERT, which was trained in 102 languages.

The model was deployed as a docker container exposing the same REST API as the old SUMMA model for compatibility reasons and integrated into the SELMA orchestration platform.

While research efforts continue on the topic of few-shot classification and active learning, our second objective, we have created and deployed the UX for the “Smart tags” scenario as described in D1.2 together with a baseline model for few-shot classification so that we can collect user data for evaluation.

3. Future Plans

This first software release already covers the WP2 tasks of T2.3, T2.4, and T2.5. We successfully integrated a multilingual news classification module. We show that implementing new language transfer methods such as multilingual contextual embedding enables high-scale analytics with more precise predictions allowing better insights over news clustering and topic detection.

For the next releases of software components, our focus will be:

1) Further improve clustering performance to cope in real-time with the growing number of sources in the ingested stream,
2) Release of the first components for summarization and story segmentation,
3) Release of the first models for topic classification using user feedback.

Bibliography

Barzdins, G., Gosko, D., Cerans, K., Barzdins, O. F., Znotins, A., Barzdins, P. F., Gruzitis, N., Grasmanis, M., Barzdins, J., Lavrinovics, I., Mayer, S. K., Students, I., Celms, E., Sprogis, A., Nespore-Berzkalne, G., Paikens, P. (2020b). Pini Language and PiniTree Ontology Editor: Annotation and Verbalisation for Atomised Journalism. *In: ESWC 2020 Satellite Events. LNCS, Volume 12124, pp. 32-38.*

Peteris Paikens; Guntis Barzdins; Afonso Mendes; Daniel Ferreira; Samuel Broscheit; Mariana S. C. Almeida; Sebastiao Miranda; David Nogueira; Pedro Balage; Martins, Andre F. T. (2016a). SUMMA at TAC Knowledge Base Population Task 2016, *DOI: 10.5281/zenodo.827317*

Znotiņš, Artūrs & Barzdins, Guntis. (2020). LVBERT: Transformer-Based Model for Latvian Language Understanding. *Baltic HLT, IOS Press, pp. 111-115, DOI 10.3233/FAIA200610.*

Znotins A, Cirule E. (2018). NLP-PIPE: Latvian NLP Tool Pipeline. *In: Human Language Technologies - The Baltic Perspective. vol. 307. IOS Press; 2018. p. 183–189.*

Paikens P. (2016b). Deep Neural Learning Approaches for Latvian Morphological Tagging. *In: Baltic HLT; 2016. p. 160–166.*

João Santos, Afonso Mendes, and Sebastiao Miranda, “Simplifying News Clustering Through Projection from a Shared Multilingual Space” in Text2Story at ECIR, 2022.

You, R., Zhang, Z., Wang, Z., Dai, S., Mamitsuka, H., & Zhu, S. (2019). AttentionXML: Label Tree-based Attention-aware Deep Model for High-performance Extreme Multi-label Text Classification. *Advances in Neural Information Processing Systems*.