Aggregation and automation of press content publishing

Adrian Orłów
adrian@orlow.me

Definitions

- *Primary Source* – a person or organization that has full or near full credibility in the information they provide. These include, e.g., politicians or political organizations, which are by themselves a source of primary information about the actions and attitudes of the represented entity.
- *Trusted Secondary Source* – an individual or organization that has a high degree of credibility in the information it provides, as a result of its role and high reliability of the correctness of the information in the past. These include news agencies and their correspondents.
- *Information condensation* – shortening information to the minimal possible form, maintaining the correctness of the final message
- *Human Factor* – any action taken by a human being within a specific situation.
- *Long and short content* – in the context of press content, short content is text with less than 280 characters. If the text is longer, it is a long content type.

Introduction

A fundamental challenge for online editors, who rely on the speed of delivering valuable press content to their audience, is the speed of response and validation of the relevance of the information provided in Primary Sources and Trusted Secondary Sources (hereinafter: Sources).

By minimizing the work needed to extract information from Sources (content aggregation), determine its relevance (audience response observation), as well as translate and condense the final information (artificial intelligence), it is possible, in effect, to reduce to a minimum – or even to zero – the contribution of the human factor in the process of effective aggregation of short press content.

The same applies for long content, excluding the full condensation of information, but including the participation of the human factor in the creation of the final press message, by expanding it with the necessary additional elements.

Implementation model

The implementation of the assumptions of this paper requires reliance on three main phases that must be carefully considered in the implementation process. These are:

1. **Aggregation:**
   Obtaining information from the Sources through communication channels such as social media, RSS feeds, Atom, or other ways of actively obtaining it in real-time
2. **Prioritization:**
   Observation of audience content engagement over time
3. **Transformacja**
   Fine-tuning the final message through translation or condensation

The phases outlined above should be implemented in a way that minimizes the Human Factor as much as possible.

Aggregation of content should take place as a result of their obtainment through API programming interfaces and observation of RSS feeds and other ICT channels in real-time, which will ensure their active obtainment and observation.

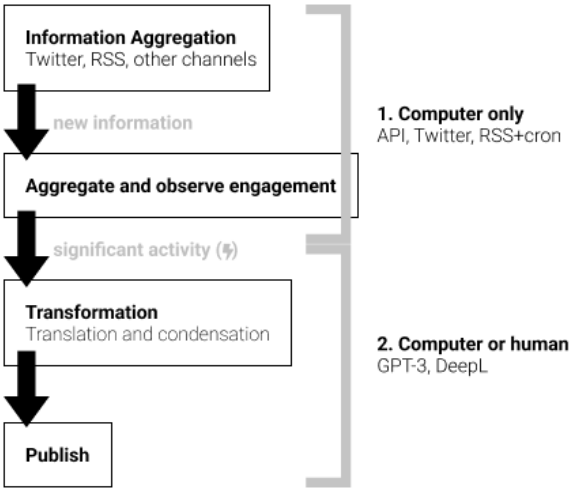
Prioritization should be a key element in the publication automation process, which is due to the assumption of opinion-forming effectiveness being directly expressed in the engagement of content recipients e.g. in social media. Observation of content reception should be done through software APIs.

Transformation should be done in an automated way for translations,
using translation tools based on neural networks. Content condensation should be performed by human or complex neural network models (e.g. GPT-3) allowing for correct implementation of this action.

The final publication should be based on human decision
or predefined source credibility combined with information clustering.

Proof-of-concept

Fig. 1 shows the implementation process based on the model in two phases, split according to the degree of automation expected.



In the Information Aggregation action, in view of the popularity of social media in the publication of short press content, the source of Source pages' posts on Twitter fetched through the platform's programming interface was used.

Through RSS feeds and similar, information is collected in real-time by means of the polling method (collecting data in equal time intervals with the help of e.g. the cron tool) or communication with the information source through the WebSocket communication protocol.

In the Aggregate and Observe action performed as a result of the New Information event, the retrieved data are stored in a non-relational database and the social media posts associated with the information are tracked via API of the respective platforms. If the engagement in a short period of time is significant, a Significant Activity event occurs.

In the second phase, it is possible to fully automate the publication, which requires grouping the information to avoid duplicates, or partial - by delegating the final action in the form of publishing press content to a human.

In order to detect duplicates, transformation of the content to the expected format of the statement and its possible translation from the current language, solutions of NLP (Natural Language Processing) type based on neural networks - GPT-3 model and DeepL tool (for translation) were used.

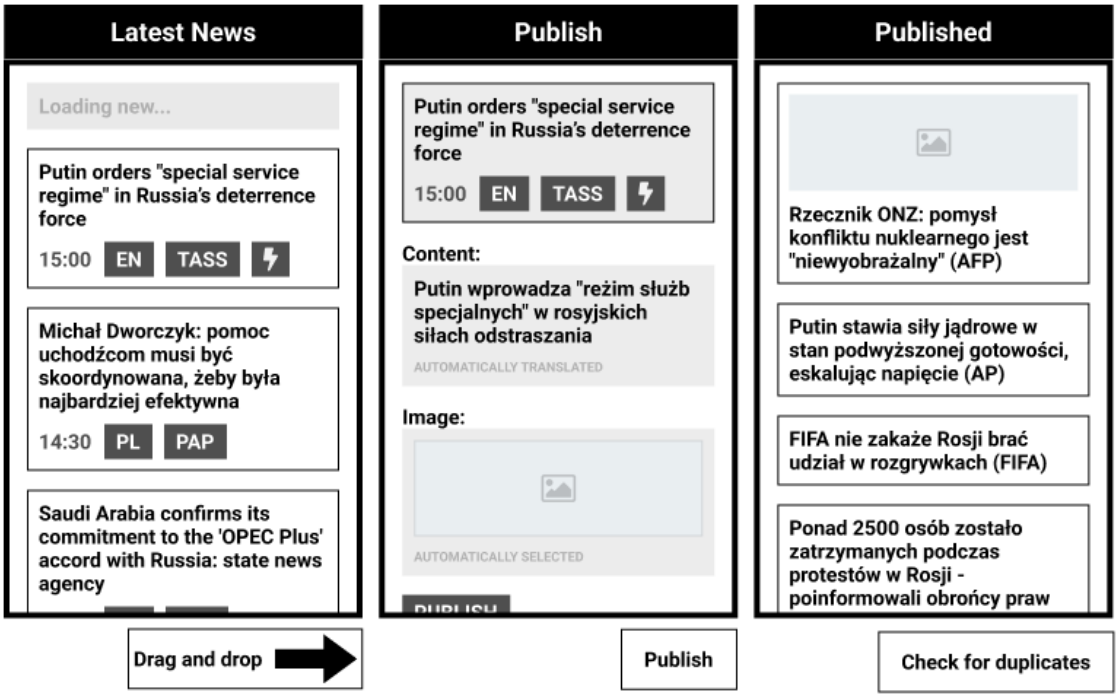


Fig. 2. (above) shows an example of an editor application interface supported by automated processes. It contains a collection of aggregated press messages (high engagement mark - lightning bolt) along with a collection of already published content, as well as a field for publishing new content.

Through the "drag and drop" method, all the necessary data is loaded from the selected message, the content is translated (if necessary) and an image is selected and uploaded. The only remaining human factor in the process is the final verification of the correctness of the content and its possible improvement.

For long content, the tool can be extended with the functionality of creating longer press message text or grouping the content.

Summary

This paper presents an effective and technically feasible solution for the evolving media marketplace, where minimizing content length and maximizing its relevance becomes a priority.

It also addresses the need to minimize the human factor in the routine processes of online editorial operations, enabling the prioritization of creative and explicitly human-driven activities.

adrian@orlow.me | Adrian Orłów