Agregacja i automatyzacja publikacji treści prasowych

Adrian Orłów
adrian@orlow.me

Definicje

- *Źródło pierwotne* – osoba lub organizacja, która posiada pełną lub niemal pełną wiarygodność w podawanych przez nię informacjach. Są to np. politycy czy organizacje polityczne, będące same w sobie źródłem oryginalnych informacji o działaniach i postawach reprezentowanego podmiotu.
- *Zaufane źródło wtórne* – osoba lub organizacja, która posiada wysoką wiarygodność w podawanych informacjach, co wynika z pełnionej przez nią roli i wysokiej pewności poprawności informacji w przeszłości. Są to agencje prasowe i ich korespondenci.
- *Kondensacja informacji* – skrócenie informacji do maksymalnie minimalnego poziomu, zachowując poprawność finalnego przekazu
- *Czynnik ludzki* – wszelkie działania podejmowane przez człowieka w ramach określonej sytuacji.
- *Treść długa i krótka* – w kontekście komunikatu prasowego, treścią krótką jest tekst mający mniej niż 280 znaków. Mający więcej – treść długa.

Wstęp

Zasadniczym problem redakcji internetowych, których fundamentem jest szybkość przekazywania wartościowych dla odbiorcy treści prasowych, stanowi szybkość reakcji i weryfikacja istotności informacji podawanych w Źródłach pierwotnych i Zaufanych źródłach wtórnych (dalej: Źródłach).

Poprzez minimalizację pracy potrzebnej do pozyskiwania informacji ze Źródeł (agregacja treści), ustalania ich istotności (obserwacja odbioru), a także tłumaczenia i kondensacji finalnej informacji (sztuczna inteligencja), można w efekcie doprowadzić do minimum – *a nawet do zera* – udział czynnika ludzkiego w skutecznej agregacji krótkich treści prasowych.

To samo stosuje się dla treści długich, z wyłączeniem pełnej kondensacji informacji, a włączeniem udziału czynnika ludzkiego w tworzeniu finalnego komunikatu prasowego, poprzez jego rozszerzenie o niezbędne elementy dodatkowe.

Model realizacji

Realizacja założeń niniejszego dokumentu wymaga oparcia się na trzech głównych filarach, które należy dokładnie rozpatrzyć w procesie realizacji. Są to:

1. **Agregacja**
   pozyskiwanie informacji ze Źródeł poprzez kanały komunikacji w postaci mediów społecznościowych, kanałów RSS, Atom, bądź z wykorzystaniem innych sposobów ich aktywnego pozyskiwania w czasie rzeczywistym
2. **Priorytetyzacja**
   obserwacja zaangażowania odbiorców w treść w czasie
3. **Transformacja**
   doszlifowanie finalnego komunikatu przez jego tłumaczenie czy kondensację

Przedstawione filary powinny być realizowane w sposób maksymalnie minimalizujący udział czynnika ludzkiego.

Agregacja treści powinna następować w skutek ich pozyskiwania poprzez interfejsy programistyczne API oraz obserwację kanałów RSS i innych kanałów teleinformatyczych w czasie rzeczywistym, co pozwoli na ich aktywne pozyskiwanie i obserwację.

Priorytetyzacja powinna być kluczowym elementem w procesie automatyzacji publikacji, co wynika z założenia efektywności opiniotwórczej bezpośrednio przekładającej się z aktywności odbiorców treści np. w mediach społecznościowych. Obserwację odbioru treści należy dokonać przez interfejsy programistyczne API.

Transformacja powinna odbywać się w sposób automatyzowany dla tłumaczeń, z wykorzystaniem narzędzi tłumaczących opartych na sieciach neuronowych. Kondensacja treści powinna być dokonywana przez człowieka lub złożone modele sieci neuronowych (np. GPT-3) pozwalające na poprawną tego działania realizację.

Finalna publikacja powinna odbywać się na podstawie decyzji człowieka lub predefiniowanej wiarygodności Źródła połączonej z grupowaniem informacji.

Proof-of-concept

Fig. 1 (po prawej) przedstawia proces realizacji na podstawie modelu w dwóch etapach, podzielonych według stopnia oczekiwanej automatyzacji.

W akcji *Agregacja informacji*, wobec popularności mediów społecznościowych w publikacji krótkich treści prasowych, skorzystano ze źródła wpisów stron Źródeł na Twitterze pobieranych poprzez interfejs programistyczny platformy.

Poprzez kanał RSS i podobne pobierane są informacje w czasie rzeczywistym poprzez metodę pollingu (pobieranie danych w równych odstępach czasu z pomocą narzęcia cron) lub komunikację ze źródłem informacji poprzez protokół komuniacyjny WebSocket.

W akcji *Zagreguj i obserwuj* wykonywanej wskutek zdarzenia *Nowa informacja* pobrane dane zapisywane są do nierelacyjnej bazy danych, a wpisy w mediach społecznościowych powiązane z informacją – śledzone poprzez API odpowiednich platform. Jeśli aktywność w krótkim czasie jest znacząca, następuje zdarzenie *Znacząca aktywność*.

W drugim etapie istnieje możliwość pełnej automatyzacji publikacji, co wymaga grupowania informacji umożliwiające uniknięcie duplikatów, czy częściowej – przez delegowanie człowiekowi finalnej akcji w postaci publikacji komunikatu prasowego.

W celu wykrycia podobieństw, tranformacji treści do oczekiwanego formatu wypowiedzi i jej ewentualnego tłumaczenia z języka obecego, zastosowano rozwiązania typu NLP (Natural Language Processing) oparte na sieciach neuronowych – model GPT-3 oraz narzędzie DeepL (do tłumaczenia).

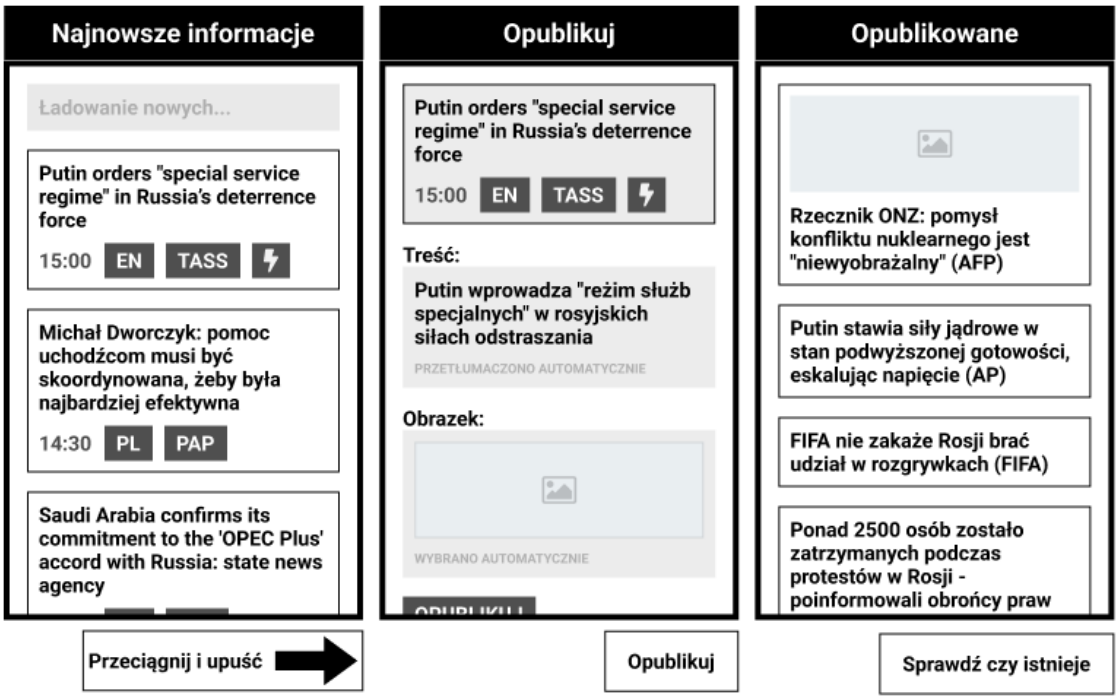

Fig. 2. (powyżej) przedstawia przykład interfejsu aplikacji redaktora wspomaganego procesami zautomatyzowanymi. Zawarty się w nim zbiór zagregowanych informacji prasowych (oznaczenie wysokiego zaangażowania – piorun) wraz ze zbiorem już opublikowanych treści, a także polem do publikacji nowej.

Poprzez metodę “przeciągnij i upuść” ładowane są wszystkie niezbędne dane z wybranego komunikatu, tłumaczona treść (jeśli potrzebne) i ładowany obrazek. Jedynym pozostającym w procesie czynnikiem ludzkim jest finalna weryfikacja poprawności treści i jej ewentualne poprawienie.

Dla treści długich narzędzie może zostać rozszerzone o funkcjonalności tworzenia dłuższego tekstu komunikatu prasowego czy też grupowanie tych treści.

Podsumowanie

Niniejszy dokument przedstawia skuteczne i realizowalne technicznie rozwiązanie dla rozwijającego się rynku medialnego, w którym minimalizacja długości treści i maksymalizacja ich istotności staje się priorytetem.

Odpowiada także na potrzebę minimalizacji czynnika ludzkiego w rutynowych procesach działania redakcji internetowych, co umożliwia priorytetyzowanie działań kreatywnych i jednoznacznie wymagających udziału człowieka.