Streszczenie

W rozprawie przedstawiony został proces budowy i analizy korpusu złożonego z 9 milionów komentarzy w języku polskim opublikowanych na portalu YouTube. Rozprawa składa się ze wstępu, stanu badań, czterech rozdziałów dotyczących części empirycznej, podsumowania, bibliografii oraz aneksu. W pierwszym rozdziale określony został przedmiot i cel rozprawy, wyjaśnione zostały także kluczowe terminy. W drugim rozdziale opisany został stan badań. Nawiązano do dotychczas dostępnej literatury dotyczącej komentarzy internetowych. Opisane zostały przykłady zastosowania różnych podejść i narzędzi badawczych, między innymi badania jakościowe, ilościowe i korpusowe. W rozdziale trzecim, zatytułowanym „Projekt i budowa korpusu" rozpoczyna się opis empirycznej części badania. Przedstawiona została procedura doboru odpowiedniego portalu internetowego, który posłuży jako źródło danych do budowy korpusu. Opisane zostały funkcje i mechanizmy związane z publikowaniem komentarzy internetowych na popularnych w polskim internecie portalach: wp.pl, gazeta.pl, youtube.com, facebook.com. Ustalono, że zakładka „Na czasie" polskiej podstrony portalu youtube.com najlepiej spełnia określone kryteria i przystąpiono do budowy korpusu komentarzy internetowych. Pobrane zostało ponad 9 milionów komentarzy pod 4323 filmami, opublikowanymi między wrześniem 2018 a wrześniem 2019 roku. Udokumentowana i opisana została techniczna procedura pobierania i przetwarzania danych.

W rozdziale czwartym przedstawiona została ogólna charakterystyka zbudowanego korpusu. Przedstawione zostały podstawowe statystyki dotyczące liczby słów w korpusie i użytkowników piszących komentarze. Wyodrębnione zostały słowa kluczowe dla całego korpusu. Zauważono obecność słów wygrywających plebiscyty na „Młodzieżowe słowo roku", takie jak „xD" czy „sztos", skrótów jak „nwm", „wgl", anglojęzyczne akronimy, liczne wulgaryzmy oraz słowa związane z samym portalem YouTube, takie jak „like", „sub", czy pseudonimy autorów filmów. Aby uzyskać bardziej precyzyjne informacje, korpus został podzielony na sub-korpusy utworzone według podziału kategorii filmów określonego przez portal YouTube: rozrywka, muzyka, ludzie i blogi, poradniki i styl, śmieszne, sport, motoryzacja, film i animacja, edukacja, gry, nauka i technika, podróże i wydarzenia, wiadomości i polityka, społeczne i non-profit, zwierzęta. Określono słowa kluczowe dla każdej z kategorii, opisano podobieństwa między niektórymi grupami kategorii i przeprowadzono analizę jakościową wybranych komentarzy.

W rozdziale piątym przeprowadzone zostały analizy o wyższym poziomie szczegółowości. Na podstawie metadanych wyodrębniono sub-korpus najpopularniejszych komentarzy z każdej kategorii i przedstawiono przykładowe treści. Zauważono, że komentarze mogą być wyróżniane zarówno przez widzów, jak i przez autorów filmów. Opracowano metody pozwalające na rozróżnienie tych grup komentarzy przy użyciu metadanych. Zauważono również liczne przykłady komentarzy, których autorzy w otwarty sposób zabiegają o uwagę innych. Wyodrębniono ponad 60 tysięcy wystąpień zwrotów w których komentujący proszą o „lajki", „suby", czy wejścia na ich kanał, niekiedy obiecując różnego rodzaju rewanż za spełnienie prośby. W kolejnym podrozdziale opisane zostały również komentarze charakteryzujące się nietypową kompozycją wizualną. Wyodrębniono różne zabiegi graficzne i typograficzne, takie jak budowa grafik z użyciem emoji, tworzenie gier i rebusów, czy nietypowe formatowanie treści. Opracowano również statystyki dotyczące wykorzystania znaków Emoji w poszczególnych sub-korpusach. Następny podrozdział dotyczy sub-korpusu kontrowersyjnych dyskusji, w którym analizie poddane zostały komentarze mające więcej odpowiedzi niż „polubień". Wyodrębniono charakterystyczne ciągi wielowyrazowe. Opracowano również statystyki dotyczące części mowy i zauważono wysoki udział wykrzykników, rozkaźników i zaimków nietrzecioosobowych. Analiza ich konkordancji pozwoliła na wyodrębnienie przykładów wulgarnych dyskusji i inwektyw. W kolejnej sekcji uwagę poświęcono statystykom dotyczącym wulgaryzmów. Zauważono wielokrotnie wyższą frekwencję wulgaryzmów w komentarzach internetowych niż w innych korpusach referencyjnych. Jednocześnie stwierdzone zostało, że wulgaryzmy często są wykorzystywane do wyrażania pozytywnych emocji, co pokazały w szczególności przykładowe komentarze z kategorii „Muzyka", która cechowała się najwyższą wulgarnością.

W rozdziale szóstym opisane zostały kwestie dotyczące zjawisk słowotwórczych, nowych słów, zapożyczeń i poprawności językowej. Zauważono, że ponad 323 tysiące razy powtórzone zostały w komentarzach wybrane przykłady słów uznanych przez Obserwatorium Językowe UW za „nowe słowa", a przodowało słowo „wow". Sprawdzono również udział wyrazów niesłownikowych w korpusie, czyli takich, które z dowolnego powodu nie

zostały uznane przez komputer za poprawne słowo w języku polskim. Opisane również nowe słowa tworzone za pomocą różnych sufiksów. Omówione zostały także przykłady modyfikacji wyrazów poprzez reduplikacje liter.

W siódmym rozdziale zostało przedstawione podsumowanie i wnioski dotyczące komentarzy internetowych oraz wnioski metodologiczne. Dla wszystkich tematycznych kategorii komentarzy internetowych zostały zbiorczo przedstawione kluczowe kategorie treści, długość komentarzy, liczba dyskusji, niesłownikowość, udział wulgaryzmów i liczba reduplikacji. Zauważono różnice tych cech między komentarzami z kategorii prymarnie rozrywkowych a komentarzami z kategorii społecznych, politycznych czy edukacyjnych. Jednocześnie cechą wspólną komentarzy z większości kategorii było nawiązywanie do twórców filmów lub osób w tych filmach występujących. Za ogólne cechy komentarzy uznano ekspresywność, potoczność i w pewnym wymiarze wulgarność, choć niekoniecznie łączącą się z agresją. Wyodrębnione zostały również cztery, charakterystyczne typy komentarzy internetowych: ekspresywne, specjalistyczne, interdyskursywne i cyfrowonatywne. Ostatni typ okazał się szczególnie interesujący. Jego nazwa nawiązuje do obecnego w literaturze terminu „Digital Natives" i określa komentarze, w których szczególnie duży udział ma nowe słownictwo, słowa dotyczące przestrzeni cyfrowej, czy samego portalu YouTube.

Przedstawione zostały również wnioski metodologiczne i narzędziowe. Opisana została rola metadanych i technicznych narzędzi podczas pracy z korpusem. Przedstawiono niektóre problemy pojawiające się podczas analizy komentarzy i sposoby ich rozwiązania. Opisano również perspektywy dla dalszego wykorzystania zbudowanego korpusu i przykłady tych zastosowań. Opracowane pliki z korpusem i metadanymi zostały umieszczone na nośniku dołączonym do pracy.

Wojciech Jastrzębski

09.11.2021

Wojciech Jastrzębski