Analiza polskiej bazy językowej

Konkurs ParlaMint

Jakub Chojnacki, Bartłomiej Kruczek

Kraków, 15.11.203

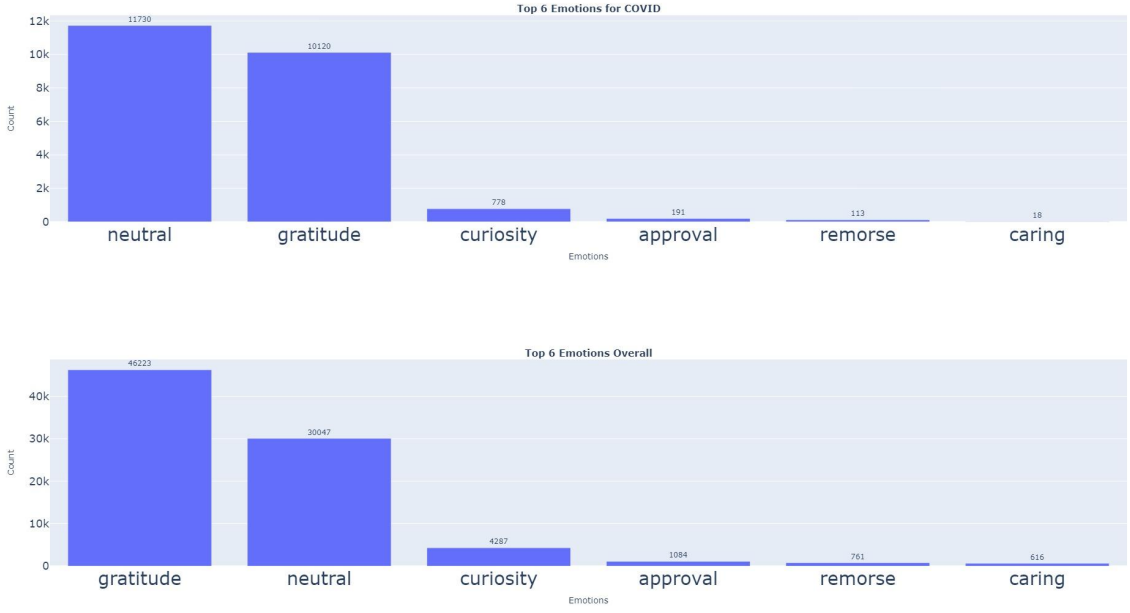
WSTĘP

W Europie znajduje się wiele państw parlamentarnych, m. in. Polska, Niemcy czy Francja. Podczas każdego zebrania głów państw, generowane są rozmaite dane; począwszy od ilości zebranych osób na sali, po każdą ich wypowiedź. Stenograf ma za zadanie notować usłyszane zdania, słowa, reakcje publiczności. Baza danych, zawierająca lata dokumentacji stenograficznych, może posiadać nawet setki tysięcy jak nie milionów słów. Każde wypowiedziane przez nas sformułowanie niesie ze sobą jakąś emocję. Żywsza dyskusja może w skrajnych przypadkach doprowadzić nas do granic kontroli emocjonalnej, co może mieć realny wpływ na rozumienie i skuteczną ocenę otaczającej nas rzeczywistości. Dodatkowo, żyjemy w czasach algorytmów uczenia maszynowego, które są w stanie scharakteryzować oraz sparametryzować pożądane przez nas dane lepiej od człowieka. W poniższym, krótkim raporcie, postaramy się Państwu pokazać wybrane zależności statystyczne w oparciu o polską bazę danych parlamentarnych-ParlaMint.

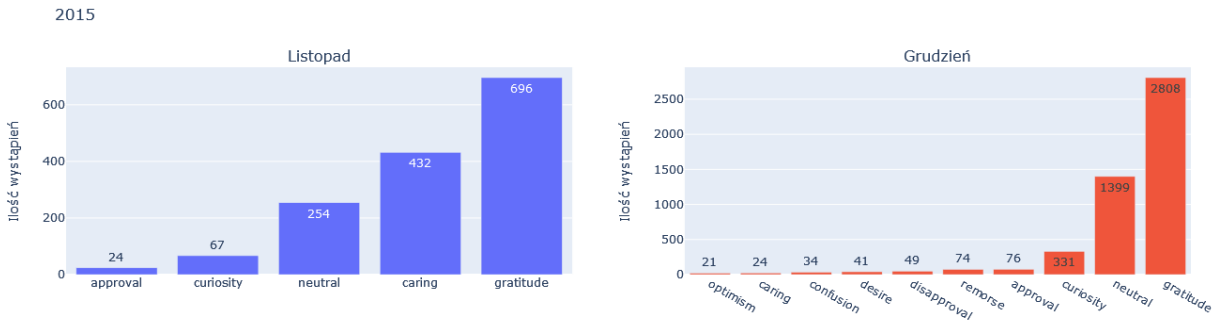
METODYKA I WYNIKI

Ogólnym przyjętym przez nas założeniem była próba pokazania liczebności występowania emocji na wybranych latach ze zbioru danych. Jest to bardzo prosty problem klasyfikacyjny, do którego został wykorzystany pre-trenowany model oparty na algorytmach uczenia maszynowego. Przyjmuje on słowa, zdania lub całe wypowiedzi w formie pisanej jako dane wejściowe i z pewną dozą prawdopodobieństwa zwraca sklasyfikowaną emocje. Zebrane w ten sposób dane można wykorzystać na wiele interesujących sposobów. Jednym z nich jest próba zestawienia ich częstotliwości występowania z podziałem na miesiąc, czy pokazanie jak (oraz czy) zmieniały się one podczas poważniejszych obrad.

Rys. 1. Zestawienie częstotliwości występowania emocji dla 2022 roku, posiedzenia parlamentarne związane z wojną na Ukrainie

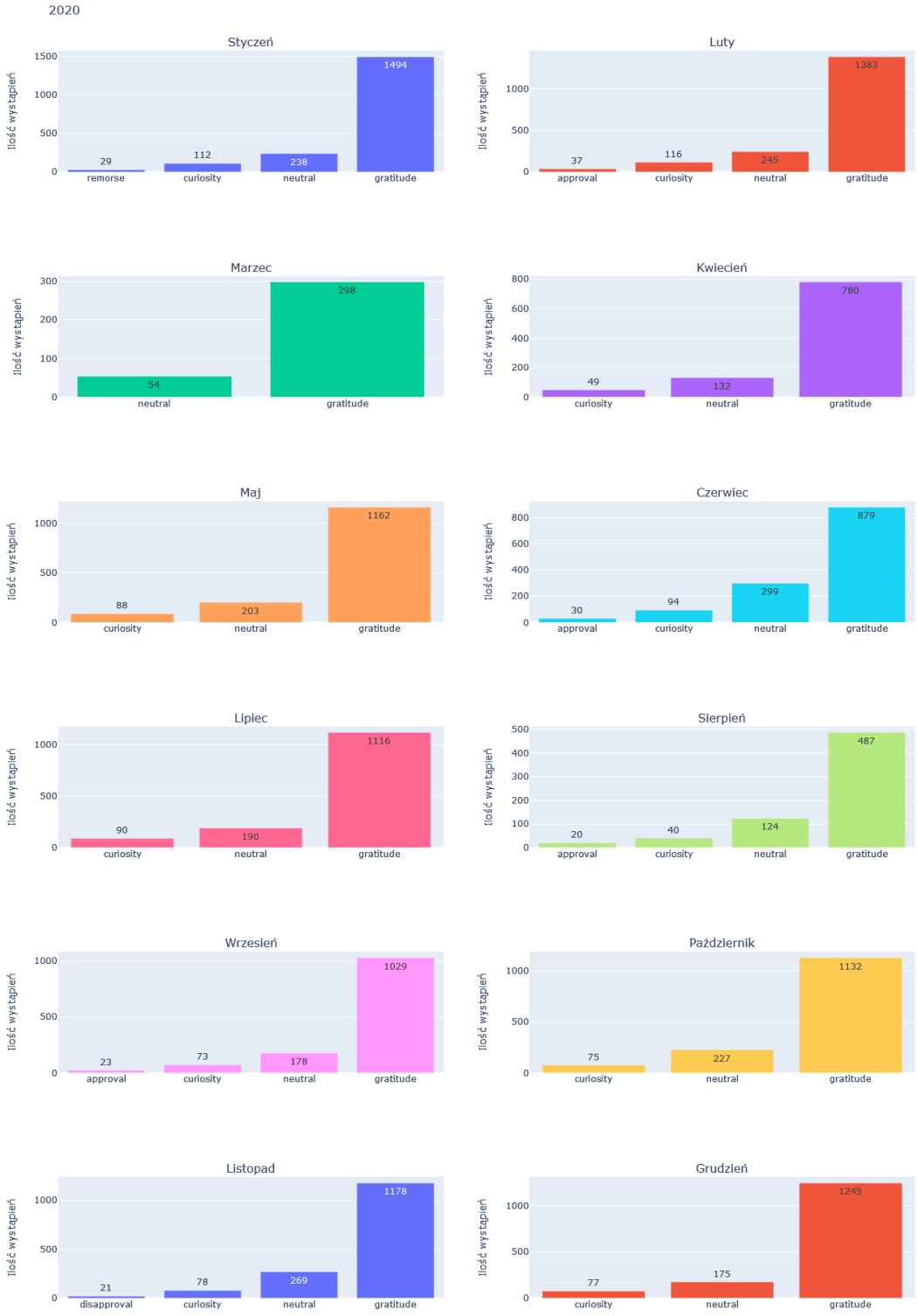


Rys. 2. Zestawienie częstotliwości występowania emocji dla 2020-2021 roku, posiedzenia parlamentarne związane pandemią Covid

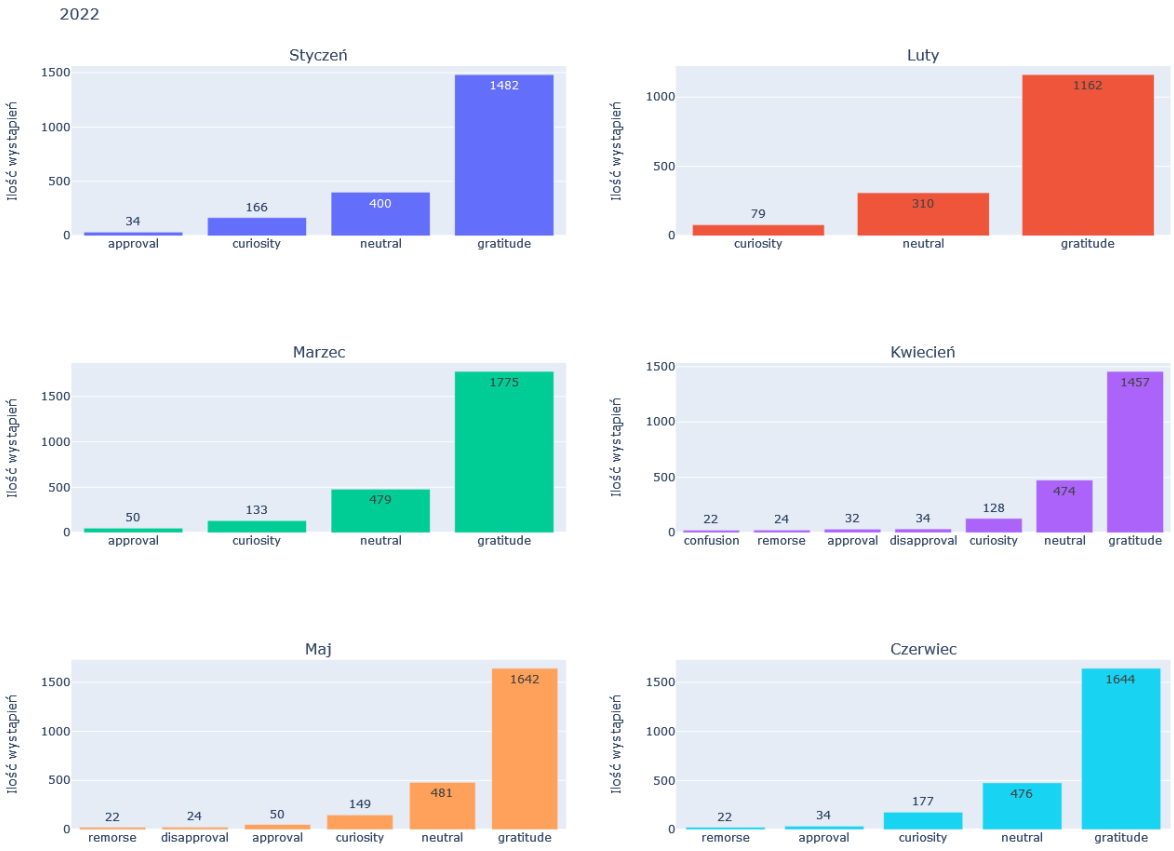


Rys. 3. Miesięczne zestawienie występujących w wypowiedziach - 2015

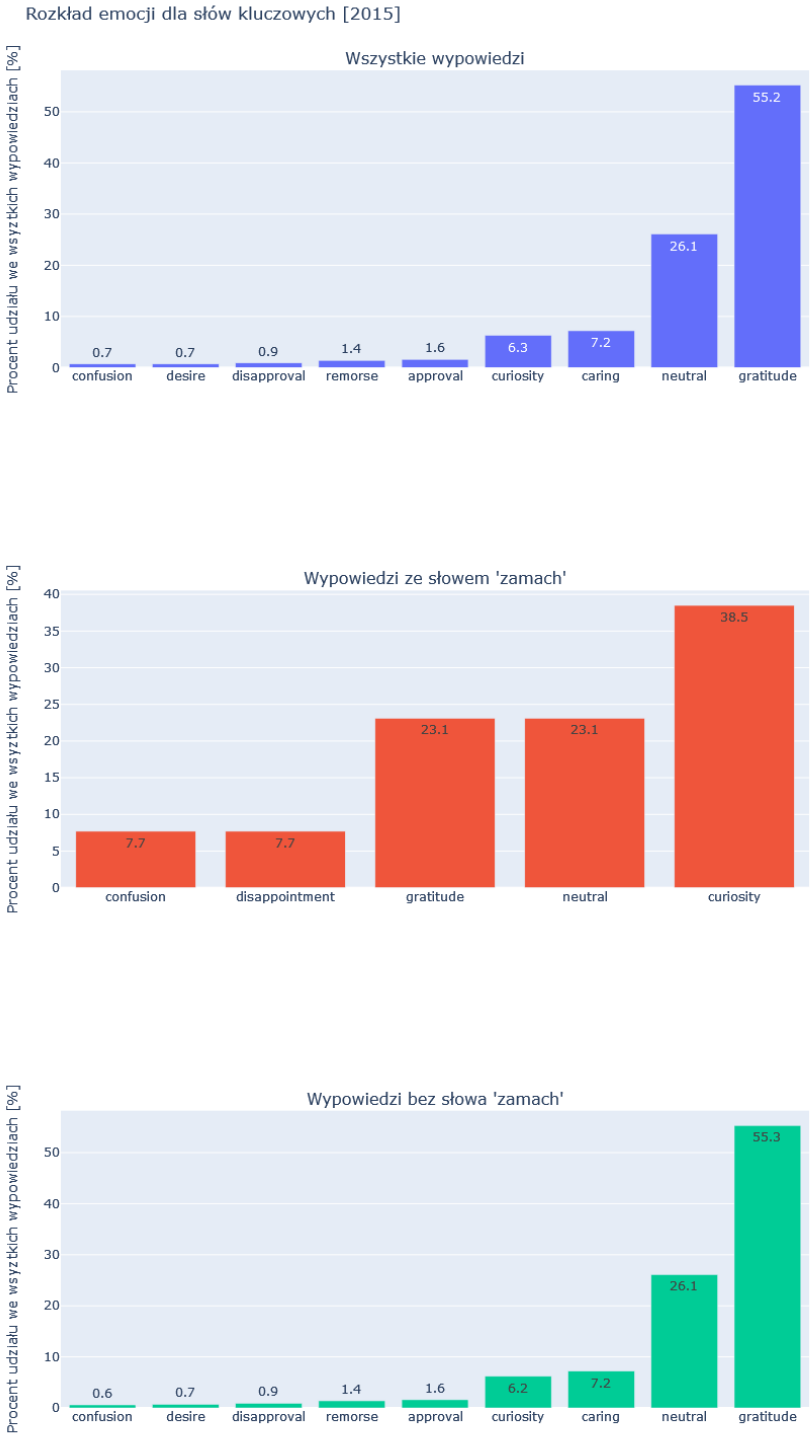
Rys. 4. Miesięczne zestawienie występujących w wypowiedziach - 2016



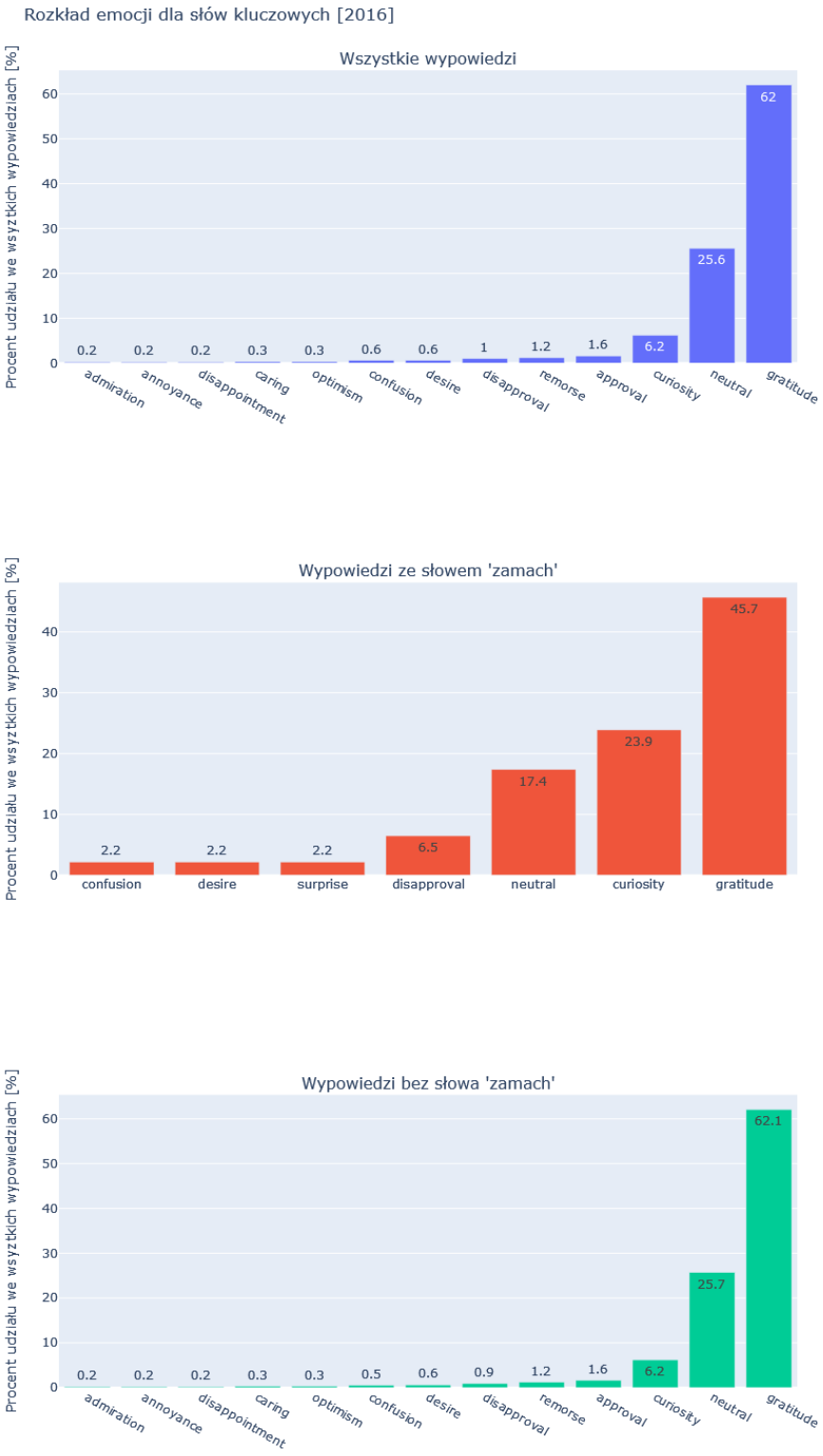
Rys. 5. Miesięczne zestawienie występujących w wypowiedziach - 2020



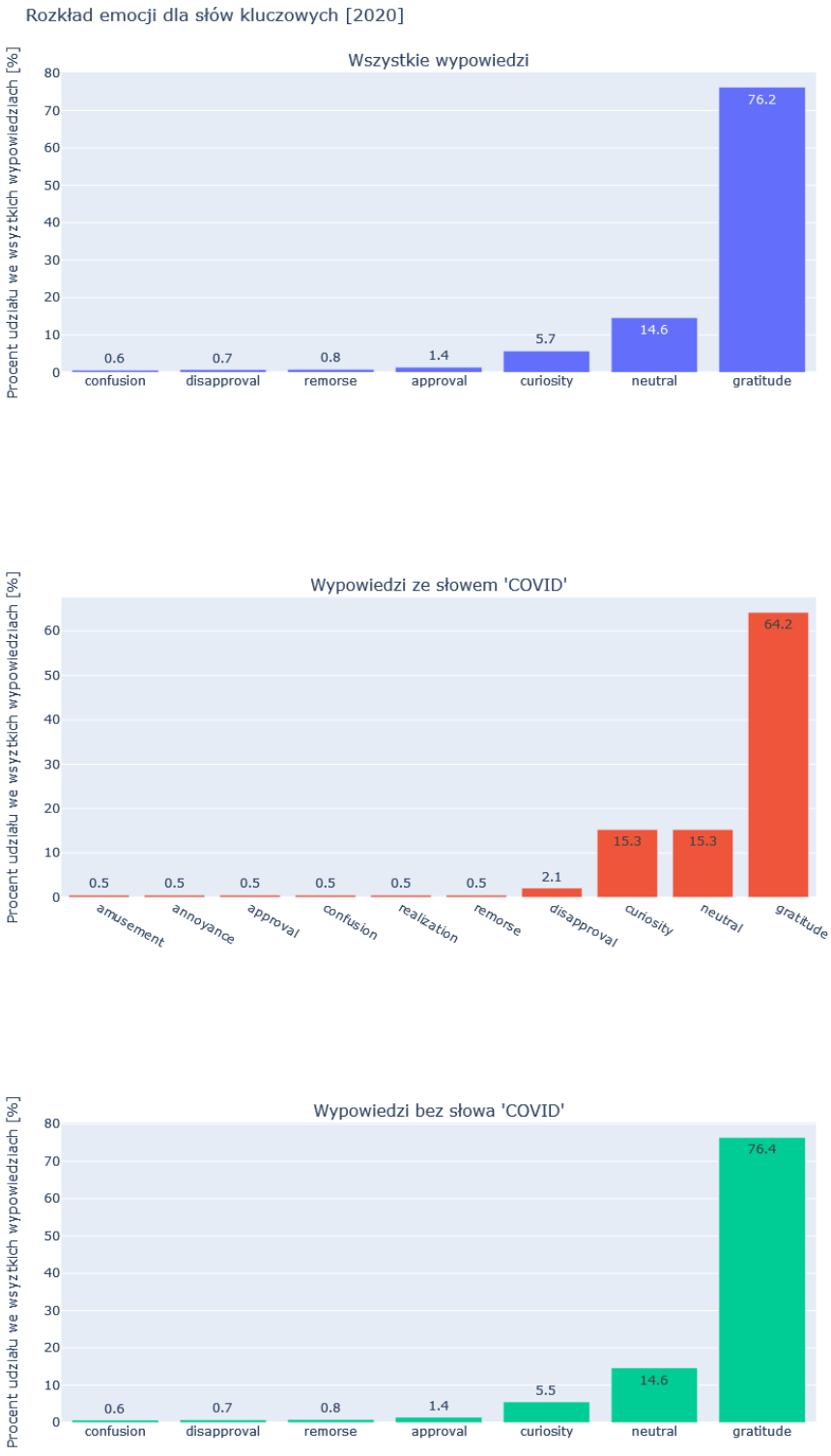
Rys. 6. Miesięczne zestawienie występujących w wypowiedziach - 2022



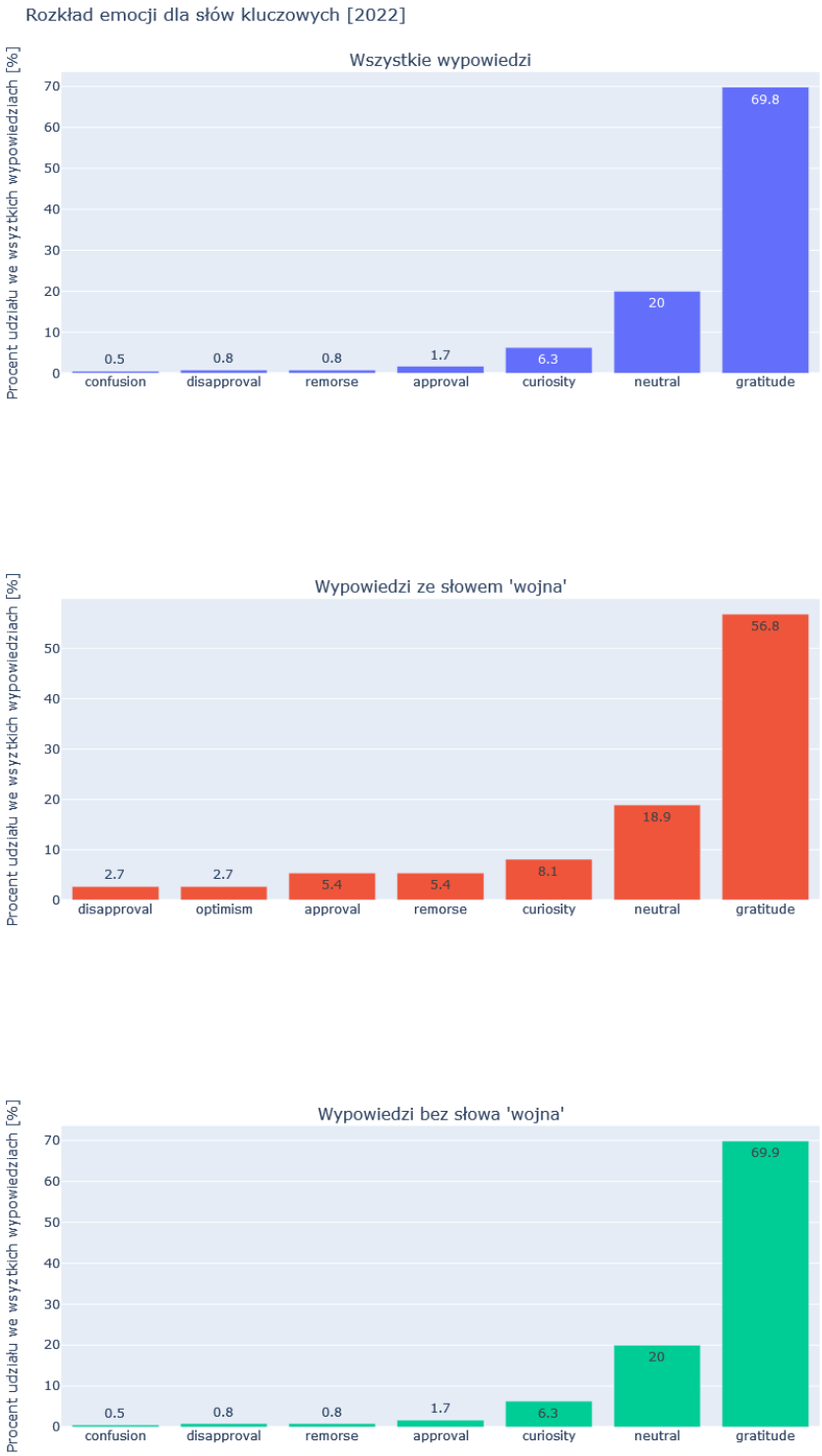
Rys. 7. Porównanie rozkładu emocji dla słowa kluczowego "zamach" – 2015



Rys. 8. Porównanie rozkładu emocji dla słowa kluczowego "zamach" – 2016



Rys. 9. Porównanie rozkładu emocji dla słowa kluczowego "COVID" - 2020



Rys. 10. Porównanie rozkładu emocji dla słowa kluczowego "wojna" - 2022

WNIOSKI

Zależnością widoczną na wykresach jest dominacja niektórych emocji. W naszym przypadku można wyłonić stabilną trójkę najczęściej wskazywanych przez model ekspresji: “gratitude” - wdzięczność, “neutral” - neutralny oraz “curiosity” - ciekawość. Ze względu na profesjonalny charakter miejsca jakim jest sejm, naturalnym było zakładać że wypowiedzi będą jak najbardziej neutralne. Tymczasem występowanie neutralność zostało znacznie przewyższone, przez tym bardziej zaskakującą emocje jaką jest wdzięczność. Zauważyć również można, że żadna z trzech wymienionych emocji nie jest postrzegana jako negatywna.

Miesięczna analiza występowania emocji jak najbardziej pokazuje powyższy wniosek (Rys. 3-6). Wdzięczność występuje znacznie częściej w porównaniu do innych emocji, ale w większości zachowuje się również trójka najpopularniejszych emocji. Jedyną wartą wskazania anomalią jest spadek ilości danych w okolicy początku roku 2020. Oczywistym powodem jest nagły wybuch pandemii COVID-19 oraz związane z tym utrudnienia oraz spowolnienia odbywania obrad. Dane poprawiają się po około dwóch miesiącach choć do końca roku występują pewne spadki informacji.

Przeprowadzono również analizę rozkładu emocji uwzględniając rok oraz występowanie słów kluczowych. Słowa kluczowe zostały wybrane na podstawie występujących jednocześnie wydarzeń geopolitycznych. W tym przypadku zamiast czystego zliczania wystąpień, zdecydowano się na procentowy udział emocji we wszystkich wyznaczonych emocjach. Widzimy że na przykład w przypadku lat 2015 oraz 2016 (wzmożona aktywność terrorystów na terenach Europy) zauważamy znaczny spadek udziału “gratitude” w całości zbioru (choć dalej utrzymuje znaczącą przewagę) na rzecz występowania innych emocji, w tym bardziej negatywnych. Podobny efekt widzimy w przypadku słowa “COVID” dla roku 2020 (początek pandemii) oraz “wojna” dla 2022 (początek inwazji wojsk Rosyjskich na Ukrainę), choć w nieco mniejszej skali.

Powyższe wyniki należy przyjmować z dozą niepewności. Ze względu na ograniczone zasoby czasowe oraz obliczeniowe wprowadzono pewne kompromisy. Część wypowiedzi, po analizie zwróciło błędne wartości (NaN), jednak pełna, wymuszająca emocje analiza mogłaby zająć ponad 48 godzin na rok. Sam model sieci trenowany był oraz operuje na danych angielskich, przez co wykorzystana musiała być translacja. Krok ten również mógł wpłynąć na problemy z przypisaniem emocji.

BIBLIOGRAFIA

- https://huggingface.co/SamLowe/roberta-base-go_emotions