Przetwarzanie Języka Naturalnego - Projekt

Analiza tekstów piosenek w odkrywaniu emocji, gatunku i nastroju w klasyfikacji utworów muzycznych z wykorzystaniem technik przetwarzania języka naturalnego

Julia Różycka[254756] and Patrycja Sekuła[254716]

Politechnika Wrocławska

Spis treści

1 Cel projektu

Celem tego projektu jest zapoznanie się z dostępnymi rozwiązaniami oraz różnorodnością zastosowanych metod do klasyfikacji utworów muzycznych. Projekt skupia się wokół możliwości wyciągania informacji czy też predykcji emocji związanych z danym utworem, jego nastroju czy klasyfikacji gatunku. Analizowanymi danymi są teksty utworów muzycznych.

Do postawionych w tej pracy pytań należą:

- W jaki sposób techniki przetwarzania języka naturalnego mogą być wykorzystane do ekstrakcji cech z tekstów piosenek?
- Jak wybrana technika wstepnego przetwarzania wpływa na wyniki klasyfikacji czy też predykcji nastroju czy gatunku utworów?
- W jakim stopniu wybrane modele mają zdolność do generalizacji problemu klasyfikacji gatunku i nastroju utworu muzycznego na podstawie samego tekstu piosenki?

2 Opracowanie literaturowe

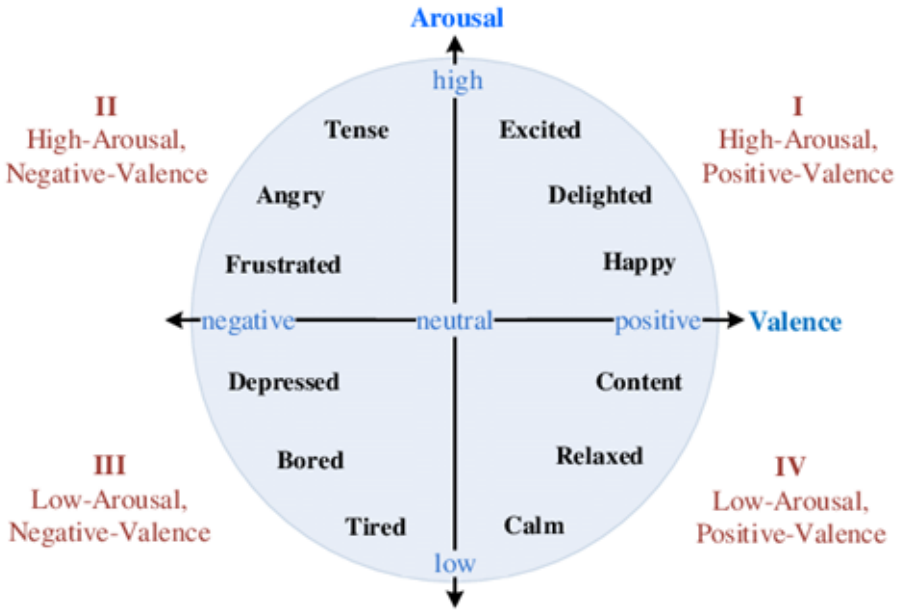
Przed podjęciem się postawionego zagadnienia projektowego, przygotowano opracowanie literatury podejmującej się zagadnienia wykorzystania danych audio oraz tekstów muzycznych w identyfikacji nastroju piosenek, czy też detekcji emocji. Celem opracowania literaturowego było pozyskanie informacji o preprocesingu danych, dostępnych zbiorach danych, wykorzystanych modelach oraz wyekstrahowanych cech, rodzajach architektury sieci i algorytmów uczenia maszynowego w tym temacie.

Wstępnie wyróżnić można dwa główne podejścia w opracowaniu technik klasyfikujących utwory muzyczne: unimodalne i wielomodalne. W analizie unimodalnej wykorzystuje się tylko jeden rodzaj danych. Z kolei podejścia wielomodalne bazują na różnych źródłach danych. Mogą nimi być przykładowo dane tekstowe, nagrania audio, czy obrazy. Wykorzystanie kilku rodzajów danych pozwala na dostarczenie dodatkowych informacji, które mogą poprawić jakość klasyfikacji. W przypadku detekcji emocji, nastroju czy gatunku w podejściu wielomodalnym korzysta się zazwyczaj z tekstów piosenek oraz nagrania audio.

Często poruszanym w przeczytanych przez nas pracach problemem był także problem 'ground truth' w identyfikacji emocji. Odnosi się on do trudności w dokładnym zdefiniowaniu emocjonalnej zawartości utworu muzycznego. Emocje są złożonymi i subiektywnymi doświadczeniami, które mogą się znacznie różnić w zależności od osoby, kultury i kontekstu. W rezultacie określenie dokładnej treści emocjonalnej utworu może być wysoce subiektywne i trudne do obiektywnego określenia.

W dziedzinie rozpoznawania emocji w muzyce, dwa powszechne modele używane do opisywania emocji to model dyskretny i model *Valence–Arousal* [1].

Model dyskretny zakłada, że wszystkie emocje mogą się wywodzić z podstawowych emocji takich jak radość gniew czy strach. Natomiast model wymiarowy opisuje emocje za pomocą przestrzeni afektywnej, gdzie osiami ortogonalnymi są pobudzenie i wartościowość.



Rysunek 1: Model Valence-Arousal (walencja-pobudzenie).

W ostatnim czasie modele dwuwymiarowe zyskują na popularności. Trzy różne modele emocji omawiane w tym artykule to: Model Henvera, Model Russela oraz Model Thayera. Każdy z tych modeli opisuje emocje w kontekście muzyki z różnych perspektyw, co pozwala na bardziej różnorodną analizę i zrozumienie związku między muzyką a emocjami. Autorzy wspominają także o czymś takim jak leksykon afektywny (z ang. *affective lexicon*) - odnosi się to do wybranych podzbiorów słów w języku dotyczących stanów afektywnych (pozytywnych lub negatywnych), a większość z tych słów odnosi się do emocji.

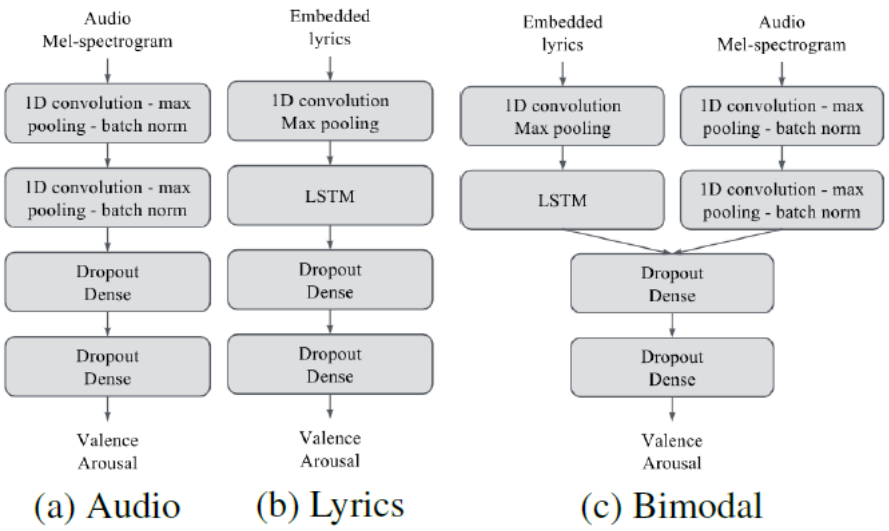
W dalszej części podjęto się opracowania 10 artykułów bazujacych zarówno na modelu dykretnym jak i dwuwymiarowym (a także jego wariacjach) w podejsciu unimodalnym i multimodalnym.

Automatyczna detekcja nastroju, czy też gatunku utworów muzycznych jest wysoce pożądana, co wynika z potrzeb przemysłu muzycznego dla serwisów streamingowych. Nacisk na stworzenie szybkiej i automatycznej metody klasyfikowania utworów muzycznych stał się jedną z powódek do zbadania tego problemu. Najściślej z tym zagadnieniem wiąże się dziedzina przetwarzania języka naturalnego oraz uczenia głębokiego.

W pierwszym z analizowanych artykułów [2] autorzy podjęli się klasyfikacji multimodalnej nastrojów piosenek. Do tego celu wykorzystali techniki uczenia głębokiego, a otrzymane wyniki porównali z klasycznymi metodami ekstrakcji cech. Badany zbiór danych zawierał nagrania audio oraz teksty piosenek.

W literaturze znaleźć można dwa sposoby reprezentacji nastroju utworów muzycznych. Pierwszy z nich opisuje utwór za pomocą etykiet. Drugi sposób to reprezentacja ciągła w postaci wykresu walencja-pobudzenie (z ang. *Valence-Arousal*). Ta reprezentacja to punkt w przestrzeni dwuwymiarowej cech walencji (wartościowości) oraz pobudzenia (energii).

Autorzy pracy wykorzystali drugi sposób reprezentacji i zaproponowali autorski model bimodalny, porównując go do dwóch podejść unimodalnych. Uproszczoną architekturę przedstawiono na rysunku **??**. Testom zostały poddane unimodalne modele osobno dla danych audio oraz dla tekstów piosenek. Trzecim było połączenie rozwiązań we wspólną multimodalną sieć. w każdym z przypadków model przewidywał jednocześnie walencję oraz pobudzenie. Wejścia zostały podzielone na kilka segmentów treningowych, których ostateczny wynik był średnią z uzyskanych wyników dla każdego z wejść. Modele dla danych audio opierały się o konwolucyjnej sieci neuronowe takie jak ConvNet. Dla danych tekstowych, sieć łączyła architekturę sieci rekurencyjnych z warstwami konwolucyjnymi. Model bimodalny stanowił fuzję tych rozwiązań poprzez połączenie wyjść z unimodalnych modeli pozbawionych ostatnich warstw gęstych. Wartości uzyskane z połączenia tych wyjść zostały następnie skierowane na dwie warstwy w pełni połączone. Uzyskano podobne wyniki dla



Rysunek 2: Architektura modeli jedno- i wielomodalnych [2].

podejść unimodalnych, czy to z danymi audio, czy tekstowymi, zarówno dla klasycznych algorytmów uczenia maszynowego, jak i metod uczenia głębokiego. Podkreślono, że najlepsze wyniki uzyskano poprzez połączenie warstw konwolucyjnych i rekurencyjnych. w przypadku predykcji pobudzenia/energii dane dźwiękowe dają lepsze rezultaty. Natomiast dla walencji predykcja z obu rodzaju danych daje bardzo zbliżo-

ne rezultaty. znacznie jednak przewyższające okazały się wyniki dla podejścia z wykorzystaniem bimodalnym. Poprawę tą przypisano zdolności modeli do odkrywania i wykorzystania korelacji między dźwiękiem, a tekstem, szczególnie w przypadku przewidywania walencji (wartościowości).

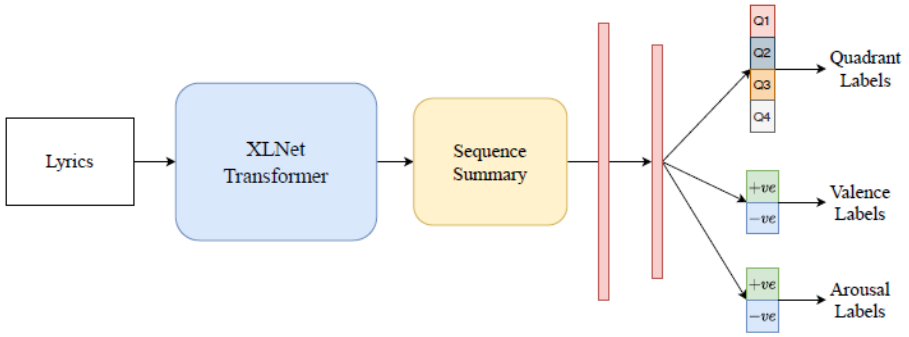
Nowoczesną strategię w przewidywaniu emocji utworów muzycznych opisano w publikacji [3], której autorską metodę oparto na transformerach. Zaznaczono, że analiza tekstów muzycznych odnosi w tej dziedzinie najlepsze wyniki. Dotychczas do wyciągania informacji o emocjach utworu muzycznego stosowane były tradycyjne techniki przetwarzania języka naturalnego ograniczone do reprezentacji słów na poziomie słów. Nowsze techniki NLP wykorzystują kontekst oraz wspólne zależności między słowami. Celem efektywnej identyfikacji emocjonalnych konotacji z tekstów utworów muzycznych, transformery stają się oczywistym wyborem.

W pracy wykorzystano dwa zbiory danych: MoodyLyrics oraz MER Dataset. MoodyMusic to zbiór zawierający 2595 piosenek opisanych przez punkty w przestrzeni dwuwymiarowej wykresu zależności walencji i pobudzenia (*Valence-Arousal* dalej jako: VA). Zbiór charakteryzuje równomierne rozłożenie punktów między ćwiartkami. Walencja charakteryzuje przyjemność utworu (w zakresie od pozytywnego do negatywnego), a pobudzenie określa energię utworu (zakres od spokojnego do energetycznego). zbiór MER również zawiera dane opisujące 180 piosenek na wykresie VA. Również rozkład punktów między ćwiartkami jest równomierny.

Architektura sieci składa się z dwukierunkowego transformera XLNet, którego podstawę stanowi BERT. Głębsze zrozumienie informacji kontekstowej zostało umożliwione poprzez dodatkowe warstwy rekurencyjne. Następnym blokiem przetwarzania informacji jest *Sequence Summary*, który przetwarza sekwencje stanów ukrytych. Model zakończony jest dwiema warstwami gęstymi. Wynikowa informacja zostaje podana jako wektor ośmiu wartości, który zostaje rozdzielony osobno na klasyfikację według 4 ćwiartek wykresu VA w sposób binarny oraz etykiety Valence oraz Arousal. Funkcja straty użyta w trakcie uczenia bierze pod uwagę reprezentację osobno dla predykcji ćwiartek, etykiet walencji i pobudzenia. Na rysunku **??** zaprezentowano graficzną postać modelu.

Porównano otrzymane wyniki do dwóch innych metod jako zadanie multi-task odpowiednio dla każdego zbioru danych. Podsumowując wyniki zaprezentowane w pracy, wykorzystanie metody opartej na transformerach wykazuje znaczną poprawę klasyfikacji. Najlepsze wartości obliczonych klasycznych metryk w detekcji emocji z tekstów w porównaniu do klasyfikatora naiwnego oraz metody wykorzystującej architekturę LSTM otrzymano właśnie dla autorskiej metody z transformerami.

W literaturze znaleźć można publikacje prezentujące wykorzystanie klasycznych metod uczenia maszynowego do zagadnień związanych z wyciąganiem informacji o nastroju utworu muzycznego. Autor [4] publikacji zaprezentował wykorzystanie algorytmu klasyfikatora naiwnego (z ang. *Naïve Bayes*) do klasyfikacji muzyki na tą z pozytywnym odbiorem versus z negatywnym odbiorem. Wkładem tej pracy jest autorski zbiór danych oraz naiwny klasyfikator Bayesowski do predykcji nastroju utworu na podstawie tekstu piosenki. Wybór metody został uzasadniony przez rodzaj zadania jako prostej binarnej klasyfikacji tekstu, dla którego algorytm ten wykazał swoją skuteczność w podobnych zadaniach.



Rysunek 3: Schemat architektury modelu [3].

Zbiór danych został pozyskany jako część zbioru Million Song Dataset z którego pobrano 10 tys. piosenek w postaci audio, jednak kluczową informacją były dane o autorze i tytule utworów. Następnie korzystając ze strony internetowej LyricWikia teksty tych utworów zostały pobrane. Etykiety piosenek zostały przypisane w dwojaki sposób. Większość została zebrana automatycznie ze zbioru LastFm, a pozostałe etykiety zostały nadane ręcznie.

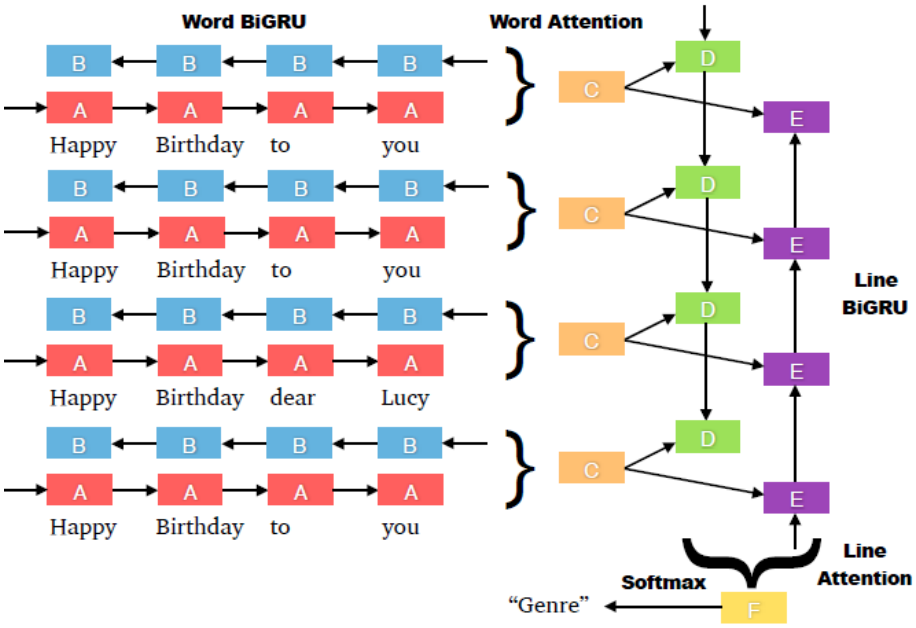
Obróbka danych obejmowała kolejno zmianę tekstów piosenek na reprezentację w postaci wektorów słów, wykorzystanie sekwencji n-gramów, czy obcinanie końcówek. Proces selekcji najlepszego modelu klasyfikatora odbył się z przeprowadzeniem selekcji hiperparametrów, wykorzystaniem metody przeszukiwania siatki i ewaluację poprzez 10-foldową walidację krzyżową. Wyznaczone metryki wskazały, że uzyskano model, który klasyfikował utwory o pozytywnym nastroju z wysoką precyzją.

Artykuł pokazał, że klasyczne algorytmy uczenia maszynowego w detekcji nastroju utworu muzycznego dają względnie zadowalające wyniki. Jako że w pracy autor podjął się względnie prostego problemu w tej dziedzinie, takie metody są możliwe do implementacji i oferują względnie dobre wyniki.

Klasyfikacja gatunku muzycznego może zostać rozszerzona o wykorzystanie struktury hierarchicznej w danych tekstowych. w tym wypadku autorzy publikacji [5] zaprezentowali metodę z wykorzystaniem hierarchicznej sieci atencji. Hierarchia w zbiorze danych może zostać opisano w postaci kolejnych łączeń. Pojedyncze słowa najpierw są połączone w wiersza. Wiersze w całe segmenty, a segmenty składają się w cały tekst utworu muzycznego.

Problem ten również przynależy do dziedziny przetwarzania języka naturalnego, jednak podejście hierarchiczne próbuje wykorzystać pewien rodzaj ustrukturyzowania danych do poprawy jakości modeli. W artykule zwrócono uwagę na problematykę związaną z prawami autorskimi. Przedstawionym rozwiązaniem było nawiązanie współpracy z stroną LyricFInd, która dostarczyła zbiorów tekstów piosenek za podpisaną zgodą. Natomiast wydobycie informacji o gatunkach muzycznych zostało dokonane poprzez wykorzystanie API serwisu iTunes. z danych, po wstępnej filtracji wyróżniono 20 kategorii gatunków muzycznych.

Strukturę architektury sieci hierarchicznej atencji (z ang. *Hierarchical Attention Network*) zaprezentowano na rysunku **??**. Każda warstwa posiada dwukierunkową jednostkę rekurencyjną.



Rysunek 4: Reprezentacja architektury HAN [5].

Część eksperymentalna obejmowała oprócz sieci HAN, klasyczne podejścia celem dokonania analizy porównawczej. Do nich wliczały się kolejno regresja logistyczna (MLR), LSTM (z ang. *Long Short-Term Memory*) i sieć hierarchiczna bez atencji.

Końcowe wnioski pracy wskazały, że metody oparte na neuronach, takie jak hierarchiczna sieć atencji i LSTM, mogą znacznie poprawić dokładność klasyfikacji gatunków. HAN wykazuje lepszą wydajność z warstwami na poziomie słowa, wiersza i piosenki w porównaniu do konfiguracji z segmentami, podkreślając znaczenie hierarchicznych mechanizmów uwagi. W pracy zwrócono uwagę, na kwestię która mogła stanowić spore ograniczenie, którym było przypisanie gatunku do artysty, niż do danego utworu przez serwis iTunes. Mogło to stanowić powód pogorszenia klasyfikacji, gdyż dany artysta niekoniecznie może być autorem muzyki tylko w obrębie danego gatunku.

Celem tego artykułu [6] było odpowiedzenie na pytanie, czy podejście multimodalne w detekcji emocji z utworów muzycznych skutkuje poprawą wyników w po-

równaniu do podejść unimodalnych. W projekcie wykorzystano 11 cech piosenek pozyskanych z API serwisu Spotify. Wykorzystano zbiór danych Deezer Moodd Detection Dataset przedstawiający piosenki za pomocą wartości Valence-Arousal (VA) wspomnianej już w poprzednich artykułach. Ponawiając wyjaśnienie jest to punkt w przestrzeni dwuwymiarowej wartości walencji oraz pobudzenia (energii piosenki). Zbiór ten jest oparty o zbiór danych Million Song Dataset. Etykiety nastroju zostały natomiast pozyskane z LastFM.

Celem przedstawienia informacji tekstowej stworzono 3 typy cech:

- sentiment information,
- TF-IDF features,
- extended ANEW features.

Cechy te zostały poddane wnikliwej selekcji celem dopasowania podkategorii cech dla obu modalności. Oznaczało to scharakteryzowanie osobnych cech dla podejścia multimodalnego oraz unimodalnego.

Ewaluację obu podejść dokonano za pomocą kilku algorytmów którymi były: Regresja Liniowa, Regresja Lasu Losowego, maszyna wektorów nośnych SVM oraz Wielowarstwowy Perceptron. Do optymalizacji hiperparametrów wykorzystano metodę przeszukiwania siatki.

Analiza wyników wskazała, że połączenie modalności danych dźwiękowych i tekstowych pozwala na skuteczne przewidywanie walencji i pobudzenia. Połączone podejście znaczenie poprawia przewidywanie wartości walencji w porównaniu do podejść unimodalnych. Jednak w przypadku przewidywania pobudzenia, modele z danymi audio przewyższały podejścia multimodalne. Końcowo można stwierdzić, że podejścia multimodalne nie zawsze kierują w stronę lepszych rezultatów.

Kolejna praca [7] na temat multimodalnego podejścia do klasyfikacji gatunku muzycznego, trenowanego na danych tekstowych, danych audio oraz obrazie (zdjęcia okładki). Informacje z każdej modalności zostały zintegrowane, aby poprawić skuteczność modelu i poprawić jego zrozumienie. Eksperymenty zostały przeprowadzone dla jedno- i wieloetykietowych danych. Dodatkowo został zaproponowany także sposób redukcji wymiarowości, który daje znaczną poprawę klasyfikacji nie tylko pod względem dokładności, ale także różnorodności modelu.

Autorzy przybliżają powszechny problem z etykietowaniem utworów - mówiąc dokładniej z anotowaniem gatunku muzycznego. Gatunki muzyczną są niezwykle przydatną informacją dla różnego rodzaju platform streamingowych i nowych aplikacji takich jak Spotify AI DJ, czy Smart Swap. Większość dostępnych prac skupia się na klasyfikowaniu do bardzo ogólnych gatunków: Pop, Rock etc. Jest to problematyczne pod wieloma względami. po pierwsze istnieją setki bardziej specyficznych gatunków muzycznych, które niekoniecznie muszą się wzajemnie wykluczać, na przykład piosenka może być popowa, ale zawierać elementy Deep House i Reggae. Po drugie duże zbiory muzyczne zawierają dane z różnych modalności tj. dźwięk, obrazy i tekst, więc czemu tego nie wykorzystać?

Jako że głównym przedmiotem tego projektu jest ocena gatunku na podstawie tekstu, postanowiono skupić się na tej części artykułu opisującej wykorzystanie in-

formacji tekstowej w modelu. Autorzy nie ograniczyli się tylko do tekstu piosenki, ale także do biografii artysty czy recenzji albumu.

W wykorzystanym zbiorze danych każdy album posiadał zmienną liczbę recenzji, dlatego wszystkie recenzje z tego samego albumu są agregowane w jeden tekst. zagregowany tekst jest obcinany do około 1500 znaków (niekompletne zdania są usuwane z końca obciętego tekstu), równoważąc w ten sposób ilość tekstu na album. Ponieważ recenzje są uporządkowane chronologicznie w zbiorze danych, starsze recenzje są faworyzowane w tym procesie.

Autorzy zastosowali proces semantycznego wzbogacenia tekstu, wykorzystując Babelyfy i ELVIS. Babelyfy i ELVIS to często wykorzystywane narzędzia podczas łączenia encji.

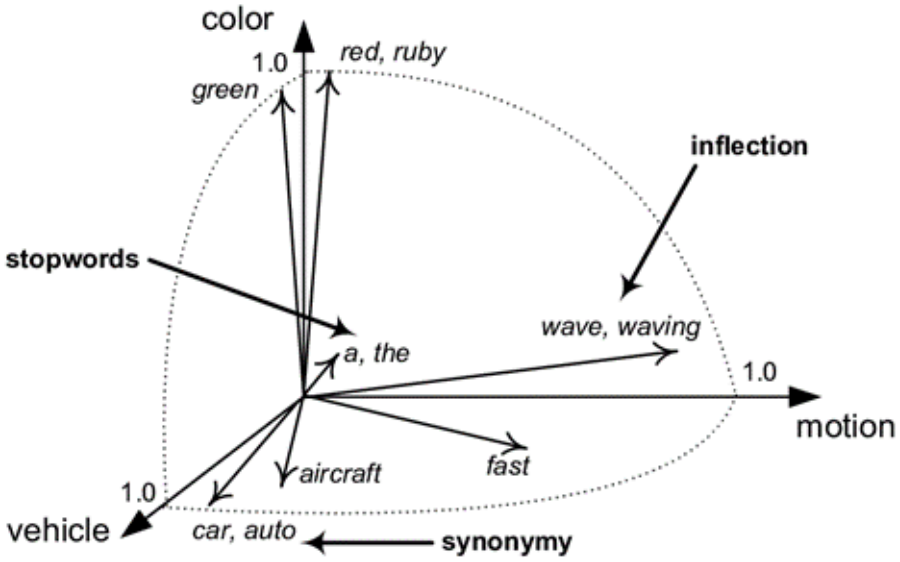
Rysunek 5: Działanie Babelfy w łączeniu encji (z ang. entity linking). EL (z ang. Entity Linking - łączenie encji to zadanie rozpoznawania (z ang. Named Entity Recognition) i identyfikowania czy mapowania (z ang. Named Entity Disambiguation) encji w bazie wiedzy (np. Wikidata, DBpedia lub YAGO).

Babelyfy służy do łączenia encji, natomiast ELVIS to system do homogenizacji i łączenia wyników różnych narzędzi służących do łączenia encji [8].

Autorzy wyodrębnili encje z tekstów za pomocą Babelyfy, a następnie za pomocą ELVIS pobrali odpowiednie adresy URL i kategorie z Wikipedii. W Wikipedii kategorie są używane do organizowania zasobów i grupowania artykuły na ten sam temat. Kategorie te są dodawane na końcu tekstu jako nowe słowa - co wzbogaca semantycznie tekst.

Wzbogacone teksty są dalej wykorzystywane dla modelu w przestrzeni wektorowej (VSM) z ważeniem TF-IDF. Rozmiar VSM został ograniczony do 10 tys. słów, ponieważ zapewnia to dobrą równowagę między złożonością sieci a otrzymywaną dokładnością. Dzięki ważeniu TF-IDF, słowa występujące często w danym dokumencie, ale rzadko w całym zbiorze dokumentów, otrzymują wyższe wagi. w efekcie, VSM z wykorzystaniem TF-IDF umożliwia przekształcenie tekstowych danych wejściowych na numeryczne wektory cech, które można wykorzystać do analizy i klasyfi-

kacji tekstu. Następnie wytrenowano przy użyciu tej reprezentacji sieć neuronową typu *feed forward* z dwiema warstwami gęstymi o liczbie 2048 neuronów oraz ReLU po każdej warstwie. Aby zapobiec przetrenowaniu zastosowano dropout o wartości 0.5. Sieć trenowana była mini-batchami po 32 elementy z zastosowaniem optymalizatora Adam.



Rysunek 6: Przykładowa wizualizacja VSM [7].

Do klasyfikacji tekstu użyto dwóch wektorów cech: jeden oryginalny (VSM), a drugi zbudowany z semantycznie wzbogaconych tekstów. Oba wektory cech zostały wykorzystane w zadaniu klasyfikacji gatunku. Wyniki klasyfikacji na podstawie tekstów pokazują, że semantyczne wzbogacenie wyraźnie daje lepsze wyniki oraz wyniki oparte na tekście są nieco lepsze od wyników opartych na samym dźwięku oraz co ciekawe podczas gdy klasyfikacja oparta na obrazie dała najniższe wyniki, pomogła poprawić ogólną wydajność modelu multimodalnego.

W [1] Afreen Ara i in. skupiają się głównie na generalnym problemie klasyfikacji emocji wspomnianym na początku tego rozdziału. Następnie oba artykuły wymieniają cechy ekstrahowane z tekstu pomagające w ocenie emocji z tekstu. [9] koncentruje się przede wszystkim na tradycyjnych cechach tekstowych, takich jak BOW i n-gramy, podczas gdy [1] wprowadza szerszy zakres cech obejmujących słownictwo, styl, semantykę i strukturę utworu. W [9] wykorzystano 2 rodzaje cech: cechy oparte na strukturze tekstu, takich jak Jlyrics3 [10], Synesketch4 [11] i ConceptNet5 [12] oraz cechy BOW (z ang. *bag-of-words*). Rozważano także cechy BOW z kilkoma transformacjami: *stemmingiem* - procesem redukowania odmienianych (lub czasami pochodnych) słów do ich rdzenia, podstawy lub formy źródłowej, usuwaniem sto-

pwords, z żadną lub z obiema operacjami. Dla każdej operacji porównano dwa rodzaje reprezentacji dla cech: Boolean i TF-IDF. Dla każdej z poprzednich kombinacji obliczono unigramy, bigramy i trygramy, tworząc łącznie 24 zestawy cech.

Co ciekawe w [1] autorzy stwierdzają, że BOW czy TF-IDF mimo, że są bardzo popularnymi cechami w NLP to ich skuteczność w identyfikacji emocji jest niewielka. Jednak podobnie jak w [9] używają modeli n-gramów, które uszeregowane są zgodnie z TF-IDF dla klasy.

Co do datasetów w [9] użyto Allmusic a w [1] rozszerzono to o Genius i LastFm.

W [1] przywołano 5 modeli, jednak bez ich praktycznej oceny, a jedynie opierając się na wnioskach artykułów, w których zostały wykorzystane. Te modele to nawiny klasyfikator Bayesa, SVM, DNN, CNN-LST model i MLP. w [9] wykorzystano SVM, kNN, C4.5 i NB (z ang. *Naïve Bayes*).

Najlepsze wyniki uzyskano przy użyciu klasyfikatora SVM. Jeśli chodzi o cechy, to cechy oparte na treści (BOW) osiągnęły lepszy wynik niż cechy oparte głównie na strukturze tekstu. Dodatkowo połączono także cechy audio i liryczne, w wyniku czego także poprawiono efektywność modelu. Najlepszy wynik (63.9% F-Score) uzyskano na zestawie 12 cech (11 z audio, 1 z tekstu) z wykorzystaniem *stemmingu*.

Klasyfikacja gatunku muzycznego jest kwestią wstępnie problematyczną, ze względu na nieustalone granice między gatunkami, czy brak dokładnych wytycznych co czyni dany utwór reprezentantem danego gatunku muzycznego. Próby wyznaczenia tych granic podjęli się badacze dla 6 gatunków muzycznych z wykorzystaniem technik regresji oraz LSTM. Wyniki zostały opublikowane w pracy [13]. Wykorzystany zbiór danych został pozyskany z strony kaggle i zawierał 160 tys. Utworów w różnych językach.

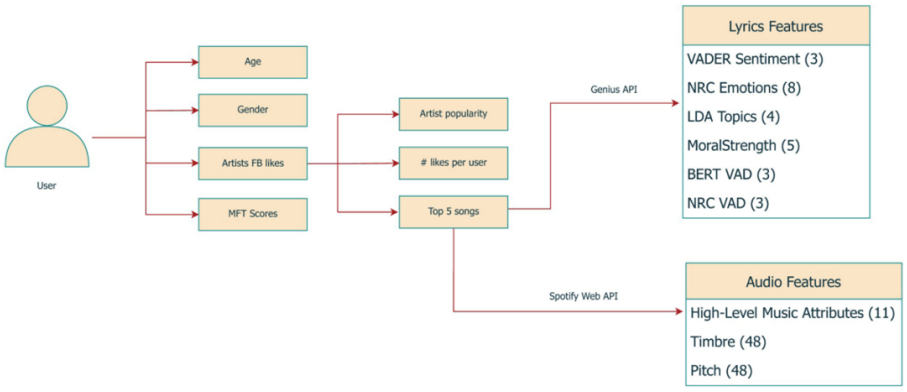
Wykorzystany zbiór danych zawierał 160 tys. utworów w różnych językach dla 6 gatunków muzycznych. Dane zawierają treść piosenki oraz odpowiadający jej gatunek. Dodatkowo zbiór posiada informację o autorze utworu. zmiana reprezentacji słów na wektory została dokonana za pomocą modelu GloVe celem stworzenia przestrzeni embeddingów. w celach zrozumienia trudności zadania i analizy danych, autorzy wytrenowali dwa modele regresji logistycznej. Wejściem dla tych modeli była reprezentacja wektorowa słów otrzymana z metody GloVe. Otrzymane wyniki przeanalizowano celem zbalansowania zbioru danych, aby zawierał równą liczbę utworów dla każdego z 6 gatunków muzycznych. Finalnie wytrenowano dwa modele: LSTM oraz dwukierunkowy LSTM. Przeprowadzono eksperymenty zarówno dla zbioru niezbalansowanego i zbalansowanego. Ponadto uwzględniono klasyfikację z przypisaniem wielu gatunków do danego utworu. w analizie końcowej porównano wyniki między modelami regresji logistycznej a LSTM z uwzględnieniem balansowania zbioru danych. Model LSTM z zbalansowanym zbiorem oraz możliwością predykcji wielu etykiet wyznaczył najwyższy poziom dokładności na poziomie 68%.

Ten artykuł prezentuje wyniki badania związku muzyki z cechami psychologicznymi jednostki. Zespół przeprowadził eksperymenty, które badały stopień przewidywalności wartości moralnych na podstawie tekstów piosenek i cech audio.

Vjosa Preniqi i in. [14] porównali ze sobą wyniki psychometryczne 1480 uczestników badania z piosenkami pięciu najbardziej preferowanymi przez nich artystów muzycznych, wyłonionych na podstawie polubień stron na Facebooku.

Jako bazy danych używano danych z LikeYouth [15] zawierająca dane 64 000 osób, w tym MFT (Moral Foundations Theory), kwestionariusze psychometryczne, informacje demograficzne i dostęp do polubień na Facebooku. z bazy pobrano dane 3880 użytkowników, którzy poprawnie wypełnili kwestionariusz MFT. Dalej dane przefiltrowano zatrzymując uczestników, którzy polubili więcej niż 10 stron artystów - otrzymując finalnie liczbę 1480 osób. Teksty piosenek uzyskiwano za pomocą API Genius, audio za pomocą API Spotify. w rezultacie otrzymano 47 580 utworów. Następnie użyto biblioteki spaCy, aby zidentyfikować i wybrać tylko utwory z angielskimi tekstami, ponieważ cechy liryczne zostały wyodrębnione za pomocą narzędzi opracowanych dla tego języka. W ten sposób otrzymano 36 902 utworów. Ostateczny zestaw obejmował 5464 artystów, 27320 piosenek. Autorzy opracowali macierz rzadką polubień stron na użytkownika i zastosowali SVD w celu zmniejszenia wymiarowości. Z źródła, na które powołali się w tym miejscu autorzy [16] wynika, że w powstałej macierzy L każdy rząd r reprezentował uczestnika, a każda kolumna c reprezentowała artystę, tak że $L(r,c)$ równała się 1 jeśli uczestnik r polubił stronę artysty c, 0 w przeciwnym wypadku. Macierz miała 5464 (artystów) x 1480 (uczestników). Wykorzystano także algorytm XGBoost do przewidywania brakujących wartości wieku.

W sumie oszacowano 26 cech dla tekstów i 107 dla audio. Cały proces w celu lepszego zrozumienia przedstawiono na rysunku **??**.



Rysunek 7: Zastosowany proces od ekstrakcji danych z bazy LikeYouth do ekstrakcji cech.

Dla każdej piosenki wyekstrahowano cechy tekstowe takie jak temat, wartość moralna, nastrój i emocje. Następnie wykorzystano różne podejścia preprocessingu tekstu – tradycyjne podejście leksykalne oraz bardziej zaawansowane podejście NLP

takie jak pretrenowany model BERT. zastosowano także oczyszczenie i lemmatyzację tekstu. zastosowano podejście modelowania tematycznego (z ang. *topic modelling*) opartego na LDA (z ang. *Latent Dirichlet Allocation*), aby zidentyfikować wspólne wzorce w narracjach tekstów. Dla optymalnej ilości tematów, postanowiono zaaplikować koherencję tematów (metryka Cv). Ponieważ utwory z dużą liczbą powtórzeń mają tendencję do bycia dominującymi, ręcznie wybrano utwory o wysokiej trafności tematycznej oraz bogatej i zróżnicowanej treści.

Aby uzyskać informację o nastroju i emocjach piosenki użyto modelu VADER oraz NRC Word-Emotion Association Lexicon. w tym przypadku zastosowano trójwymiarowy model: Valence-Arousal-Dominance (model 2-wymiarowy został wspomniany na początku opracowania, w tym wypadku dodawany jest 3 wymiar – *Dominance*). zastosowano 2 podejścia w przewidywaniu emocji - jedno oparte na pretrenowanym modelu BERT, drugie na VADER oraz NRC Lexicon. Cechy audio pobrano w całości z Spotify API.

Wykorzystano cztery nadzorowane modele regresji, Support Vector Regressor, Random Forest, XGBoost i ElasticNet, aby przewidzieć wartości moralne przy użyciu podejścia wielowymiarowego w 10-krotnej walidacji krzyżowej. Random forest uzyskał najlepsze wyniki. Jako miernik oceny wydajności wykorzystano współczynniki korelacji Pearsona. zbudowano modele predykcyjne przy użyciu każdego zestawu cech tekstowych i dźwiękowych. Następnie połączono i przeanalizowano wpływ tych cech na proponowane modele. w tym celu oceniono wartości SHAP, aby zrozumieć ogólne zachowanie modeli i znaczenie każdej cechy. SHAP (z ang. *SHapley Additive exPlanations*) to podejście oparte na teorii gier, opracowane w celu zilustrowania wkładu cech w ostateczny wynik dowolnego modelu uczenia maszynowego. Na koniec użyto testu ANOVA, do zbadania istotności różnic w otrzymanych wynikach.

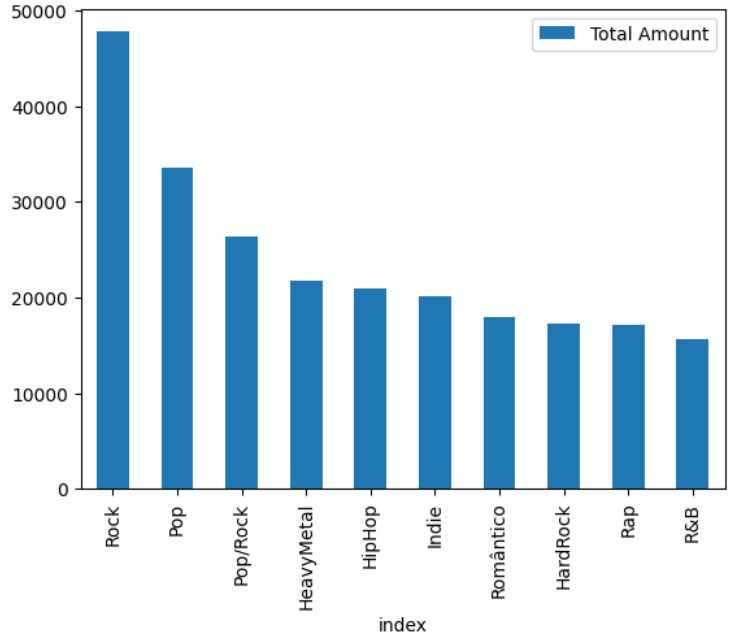
Autorzy zastosowali różne techniki, aby zidentyfikować wartości moralne, postawy i emocje przekazywane w tekstach piosenek. zbadali zarówno cechy niskojak i wysokopoziomowe dostarczane za pomocą interfejsu Spotify API. Na podstawie wyników stwierdzono, że cechy audio są lepsze niż te liryczne we wnioskowaniu o wartościach takich jak empatia i równość, ale liryczne są lepsze w przypadku wartości takich jak tradycja i hierarchia. Modele multimodalne uzyskały najwyższą skuteczność w przewidywaniu wartości moralnych. Udowadnia to, że preferencje muzyczne odgrywają rolę w więzi społecznych, ponieważ wspólne gusta muzyczne mogą wskazywać na podobne wartości.

3 Etap II - Implementacja środowiska badawczego

3.1 Zbiory danych

Wybrane zbiory danych to Genius Song Lyrics [17] oraz Song lyrics from 79 musical genres [18]. Dokonaliśmy już wstępnego preprocessingu danych. Baza [18] składała się z 2 plików, artist_data.csv oraz lyric_data.csv. W pierwszej znajdowały się: link (unikalne ID), nazwa artysty, gatunki muzyczne, natomiast w drugiej link, nazwa piosenki, tekst, język. Najpierw odflitrowano z bazy wszystkie nie anglojęzyczne teksty piosenek, następnie połączono ze sobą pliki tak aby uzyskać pola: artysta, tytuł, tekst

i gatunki muzyczne (w postaci stringa). Następnie przeprowadzono one-hot encoding a następnie wyekstrahowano 10 najbardziej popularnych gatunków. Po takim wstępnym preprocessingu w bazie zostało około 100 tysięcy utworów. Dataset [17]



Rysunek 8: Rozkład danych w datasecie 79 musical genres po preprocessingu.

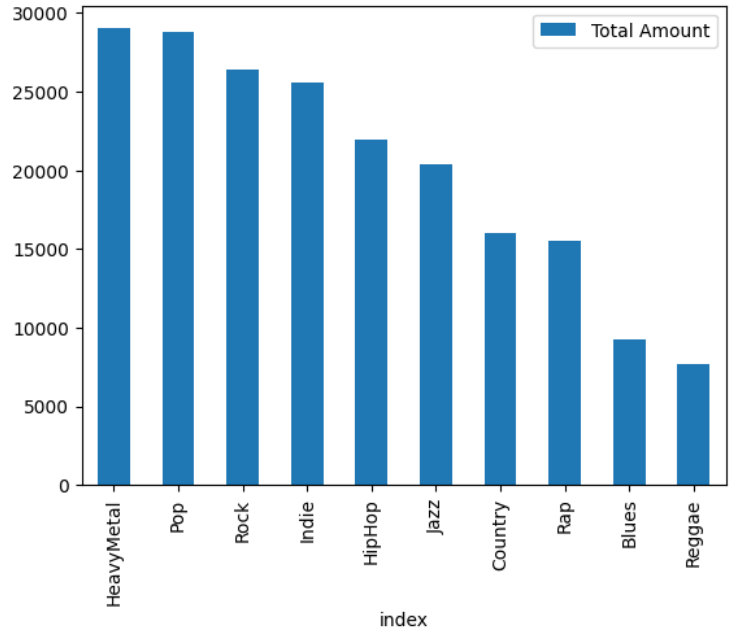
także wykorzystano aby zbalansować nieco poprzedni zbiór, ponieważ jak pokazano na wykresie 8 zbiór jest dosyć niezbalansowany, a jako, że nie jest to problem, który chcieliśmy poruszać w naszym projekcie, postanowiłyśmy zbalansować dataset, dodając więcej tekstów piosenek z gatunków innych niż Rock.

Do tak przygotowanych danych dodaliśmy jeszcze trzeci dataset w celu zbalansowania nieco danych. Na rysunku 9 prezentuje sie finalny dataset po wybraniu tylko angielskich piosenek i usunieciu duplikatów. Ma on ponad 150 tysięcy instancji.

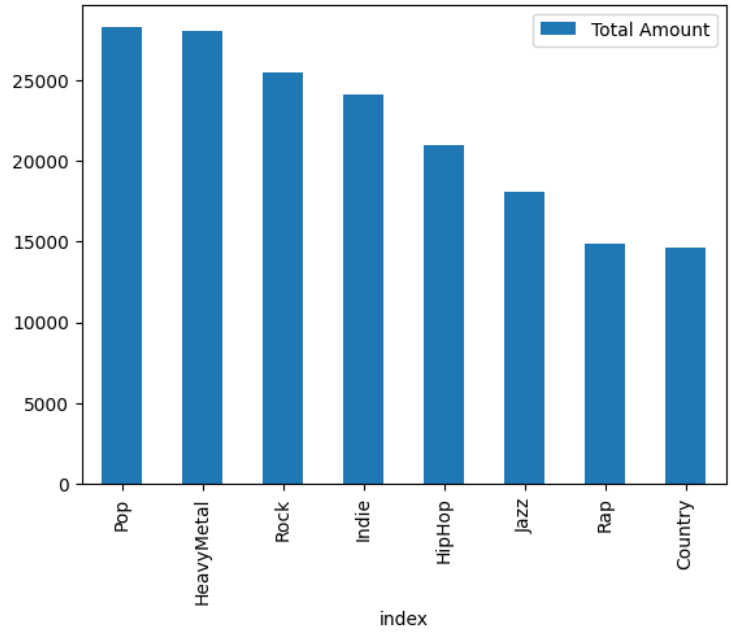
Niestety zauważyłyśmy, że na ten moment nie jesteśmy w stanie uzupełnić 2 klas, które najbardziej odstają pod względem ilości próbek: blues oraz reggae, dlatego tez postanowiłyśmy usunąć te dwie klasy 10.

Nastepnie zajęłyśmy się czyszczeniem tekstów piosenek i przygotowywaniem ich do dalszego preprocessingu. Po usunieciu zbędnych znaków, liczb, apostrofów, skrótów (np. you're) taki tekst:

> Yeah, yeah, yeah, yeah [4x] We started out good friends Said you're the kind of man that Takes every girl for a fool. Shoot, shoot, shoot A fantastic into romantic [say what?] Romantic into fantastic. We came here to rock the microphone. We came here



Rysunek 9: Finalny dataset stworzony z połączonych 3 zbiorów.



Rysunek 10: Finalny dataset, po usunieciu dwóch najmniej licznych klas.

> to rock the microphone. Our aim is to break you down to the bone. Our aim is to break you down to the bone. Lets talk about [4x] We started out good friends Said you're the kind of man that Takes every girl for a fool. We came here to rock the microphone. We came here to rock the microphone. Our aim is to break you down to the bone. Our aim is to break you down to the bone [2x] Yeah, yeah, yeah, yeah [4x]

Zamienił się w taki tekst:

> yeah yeah yeah yeah we started out good friends said you are the kind of man that takes every girl for a fool shoot shoot shoot a fantastic into romantic say what romantic into fantastic we came here to rock the microphone we came here to rock the microphone our aim is to break you down to the bone our aim is to break you down to the bone let us talk about we started out good friends said you are the kind of man that takes every girl for a fool we came here to rock the microphone we came here to rock the microphone our aim is to break you down to the bone our aim is to break you down to the bone yeah yeah yeah yeah

Następnie finalny dataset chcemy wzbogacić o *valence* i *energy* (*arousal*) w celu badania nastroju utworu, wykorzystując do tego Spotify API - w tym momencie jednak natrafiłyśmy na pewne niedogodności - około 150 tysięcy zapytań czy słabej jakości internetu skończyło się niestety zbyt długim czasem wykonywania komórki, przez co dane nie zostały w pełni zebrane. Planujemy zrównoleglić proces wysyłania zapytań i zobaczyć, czy finalnie uda nam się pobrać dane. Edit: niestety nie udało się, SpotifyAPI się jakoś zacieło dla naszego konta i zwraca cały czas 429.

3.2 Plan eksperymentów

1. **Przygotowanie danych** — przygotowanie danych tekstowych do analizy: tokenizacja, usuwanie stopwords, stemming/lemmatyzacja. Prezentacja struktury danych i cech, w tym dodatkowych danych z Spotify API (*valence*, *energy*). Wstępna analiza danych: podział na zbiór treningowy i testowy, sprawdzenie balansu klas (gatunków, nastrojów).
2. **Eksploracja technik wyciągania cech z tekstów piosenek** — bag-of-words, TF-IDF, word embeddings.
3. Eksperymenty z z wykorzystaniem **Word Embeddings i Deep Learningu** — wybór modelu (np. prosty klasyfikator oparty o sieci nauronowe), trening, strojenie parametrów modelu (np. liczba warstw, rozmiar embeddingów, learning rate) i ocena modelu w klasyfikacji gatunków i nastroju utworów (np. accuracy, precision, recall).
4. Eksperymenty z wykorzystaniem **modeli opartych na transformerach (BERT)** — implementacja gotowego modelu BERT, fine-tuning modelu na zbiorze danych, porównanie wyników uzyskanych z różnych modeli i metod przetwarzania danych.
5. **Analiza wyników i interpretacja** — prezentacja wyników klasyfikacji i predykcji nastroju i gatunku utworów, interpretacja zdolności modeli do generalizacji problemu klasyfikacji gatunku i nastroju utworów muzycznych

4 Preprocessing

Po wyczyszczeniu danych, przystąpiono do preprocessingu tekstu pod dalsze eksperymenty, w celu odpowiedzi na pierwsze pytanie badawcze i przeanalizowania, jak dana technika preprocessingu wpływa na uzyskane wyniki. W tym celu za pomocą pakietu NLTK przeprowadzono:

- lemmatyzację
- stemming
- lemmatyzację wraz z usunięciem stopwords

Lemmatyzacja to proces redukcji różnych form słów do ich podstawowej formy słownikowej. Na przykład, w zdaniu "The cats are playing with the mice," słowa takie jak "cats" zostają przekształcone w "cat," "playing" w "play," a "mice" w "mouse." Ten proces zapewnia spójność w reprezentacji słów, ułatwiając skuteczne analizowanie tekstu przez systemy przetwarzania języka naturalnego.

Stemming polega na redukcji słowa do jego podstawowej formy poprzez odcięcie końcówek fleksyjnych. Na przykład słowo "running" zostanie zstemowane do "run." Stemming pomaga w redukcji złożoności słów, co jest przydatne w zadaniach takich jak klasyfikacja tekstu czy wyszukiwanie informacji, gdzie różne formy słów wymagają traktowania w podobny sposób.

W zdaniu "The weather forecast predicts sunny skies tomorrow," stopwords takie jak "the," "is," "and," czy "it" są usuwane. Proces ten pozostawia tylko istotne słowa, takie jak "weather," "forecast," "predicts," "sunny," "skies," i "tomorrow," które są kluczowe dla zrozumienia sensu zdania.

5 Analiza danych

W naszym projekcie przeprowadziliśmy analizę danych z kilku kluczowych powodów. Przede wszystkim konieczne było zrozumienie struktury i charakterystyki naszych danych. Analiza ta pozwoliła nam lepiej zapoznać się z różnorodnością tekstów, ich długością oraz wykryć potencjalne problemy jakościowe, takie jak błędy w danych czy brakujące wartości. Eksploracyjna analiza danych była również kluczowa, pozwalając nam na zidentyfikowanie częstości występowania słów, analizę unikalnych słów oraz dystrybucję długości zdań, co przyczyniło się do lepszego zrozumienia charakterystyki naszego zbioru danych. Najpierw wyrysowałyśmy sobie dla każdej klasy tzw. *word cloud*. Wyniki zaprezentowałyśmy na rysunku 11.

Następnie przenalizowałyśmy także średnią długość sekwencji i średnią ilość unilanych słów dla każdego gatunku. Najbardziej wyróżniają się pod tym względem rap oraz hip–hop co pokazano na rysunku 12 oraz 13.

6 Eksperymenty

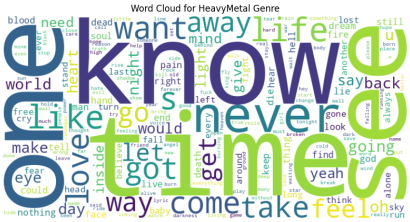
6.1 BiLSTM + CNN + GloVe

Najbardziej podobną pracą z jaką mozemy porównać nasze wyniki jest wcześniej przywołana [13], dlatego też często się będziemy odwoływać w tej części. Dla tej czę-

(a) Rock.

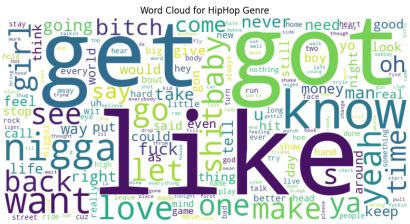
(b) Pop.

(c) Jazz.

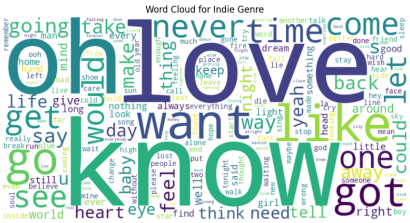


(d) HeavyMetal

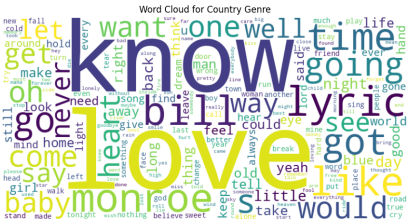
(e) Rap



(f) HipHop

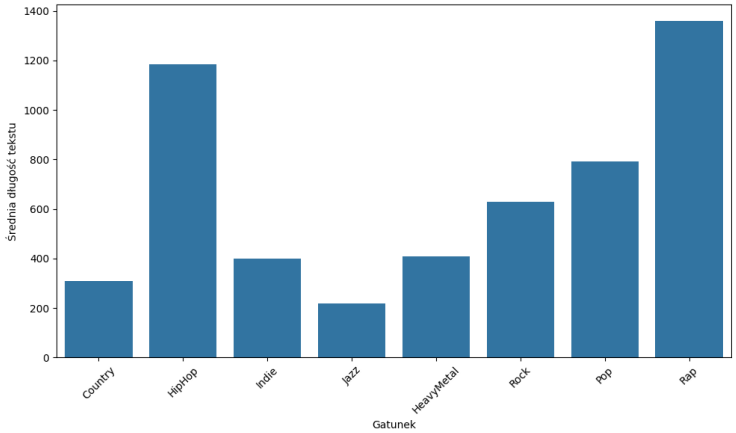


(g) Indie.

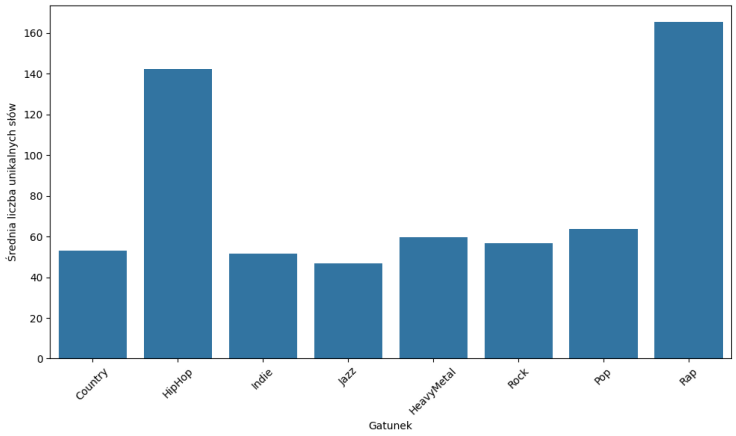


(h) Country.

Rysunek 11: Word clouds dla poszczególnych gatunków.



Rysunek 12: Średnia długość sekwencji dla każdego gatunku.



Rysunek 13: Średnia liczba unikalnych słów dla każdego gatunku.

ści przeprowadziłyśmy wiele eksperymentów - wykorzystując LSTM, BiLSTM, stosując warstwę atencyjną oraz bez zastosowania takowej warstwy. Stosując konwolucję w postaci bloku, czy też jednej warstwy. Jednak technika będąca tytułem tej sekcji okazała się najbardziej skuteczna. Poniżej znajdują się opisane warstwy modelu: Dobierano także hiperparametry, nie wybrano jednak tych, które dawały najlepsze

Layer (type)	Output Shape	Param #
embedding_59 (Embedding)	(None, 512, 300)	65152800
conv1d_57 (Conv1D)	(None, 510, 64)	57664
max_pooling1d_37 (MaxPooling1D)	(None, 255, 64)	0
bidirectional_77 (Bidirectional)	(None, 255, 128)	66048
dropout_114 (Dropout)	(None, 255, 128)	0
bidirectional_78 (Bidirectional)	(None, 64)	41216
dropout_115 (Dropout)	(None, 64)	0
dense_92 (Dense)	(None, 64)	4160
dropout_116 (Dropout)	(None, 64)	0
dense_93 (Dense)	(None, 8)	520

Total params: 65322408 (249.19 MB)
Trainable params: 169608 (662.53 KB)
Non-trainable params: 65152800 (248.54 MB)

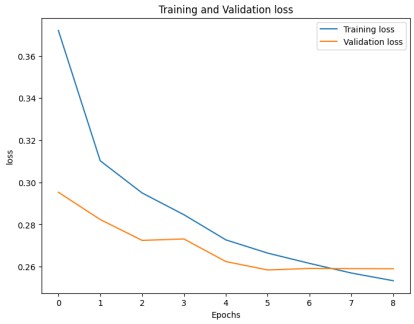
Rysunek 14: Parametry modelu.

wyniki ze względu na czas trwania jednej epoki. Rozmiar batcha ustawiono na 64 próbki, liczba epok to 10. Maksymalną długość sekwencji ustawiono na 400. Otrzymano następujace wyniki dla accuracy i loss 15.

Wykonano także macierze pomyłek — jako, że jest to problem multiklasowy, jedna macierz dla każdego gatunku 16.

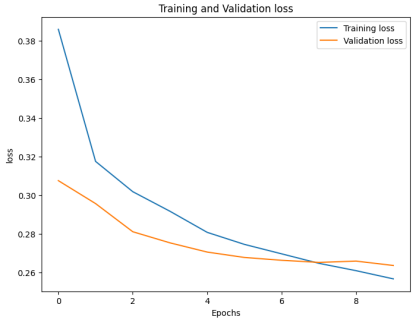
W naszej analizie eksperymentalnej zastosowaliśmy podejście wykorzystujące połączenie modeli BiLSTM, CNN oraz wektorów GloVe. Takie podejście wynika z chęci wykorzystania zalet każdego z tych elementów: BiLSTM (dwukierunkowe długie krótkoterminowe pamięci) są skuteczne w modelowaniu zależności czasowych w danych tekstowych, CNN (splotowe sieci neuronowe) są znane z efektywnej ekstrakcji cech lokalnych, a GloVe (Globalne Wektory Kontekstowe) dostarczają wysokiej jakości reprezentacji słów na podstawie ich współwystępowania w dużych korpusach tekstowych.

(a) Accuracy z lematyzacją.



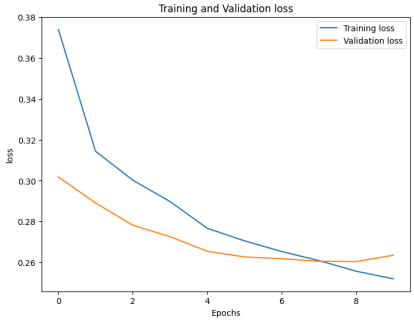
(b) Loss z lematyzacją.

(c) Accuracy ze stemmingiem.



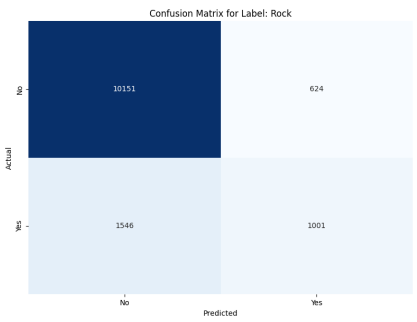
(d) Loss ze stemmingiem.

(e) Accuracy — bez stopwords.

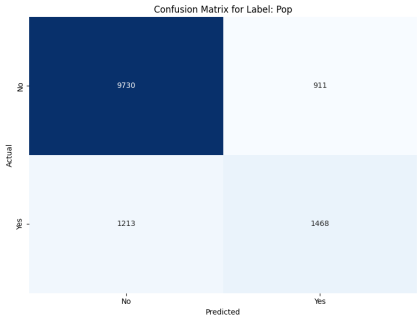


(f) Loss — bez stopwords.

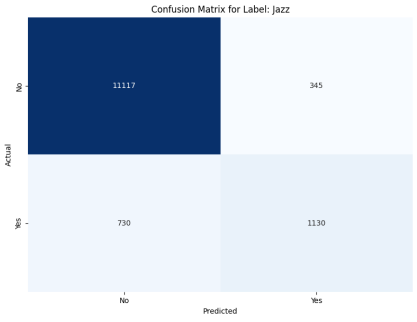
Rysunek 15: Accuracy i loss dla różnych technik preprocessingu.



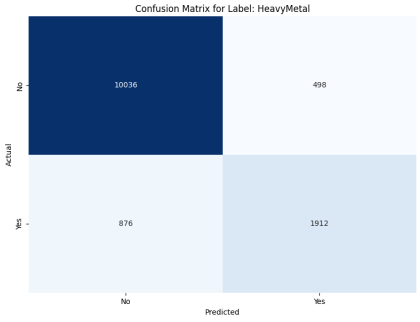
(a) Rock.



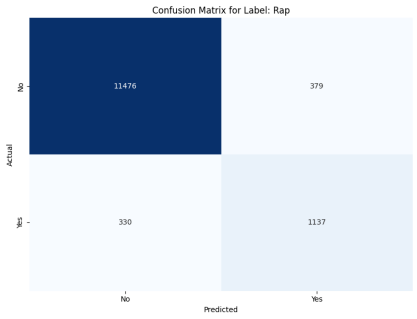
(b) Pop.



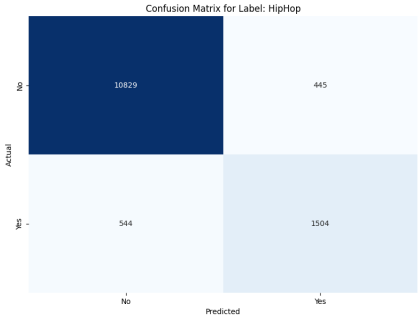
(c) Jazz.



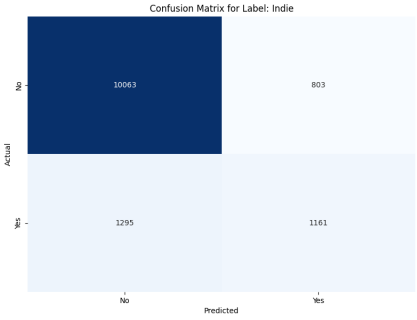
(d) HeavyMetal



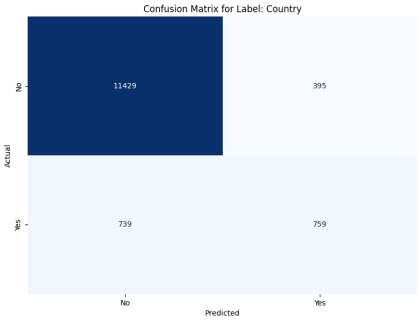
(e) Rap



(f) HipHop



(g) Indie.



(h) Country.

Rysunek 16: Macierze pomyłek dla poszczególnych gatunków.

GloVe (Global Vectors for Word Representation) to metoda osadzania słów, która tworzy wektory słów poprzez analizę globalnych statystyk współwystępowania słów w korpusie. BiLSTM (Bidirectional Long Short-Term Memory) to rodzaj sieci neuronowej, która potrafi przetwarzać dane sekwencyjne w obu kierunkach, co jest szczególnie przydatne w NLP, gdzie kontekst z obu stron słowa może być istotny.

W naszych eksperymentach osiągnęliśmy następujące wyniki na zbiorze testowym:

- dane z lematyzacją: dokładność – 0.602
- dane z stemmingiem: dokładność – 0.592
- dane z lematyzacją i usunięciem stop words: dokładność – 0.590

W pracy, do której się odnosimy, osiągnięto dokładność na poziomie 65% na zbalansowanym zbiorze danych. Nasze gorsze wyniki mogą wynikać z kilku czynników. Po pierwsze, użycie niezbalansowanego zbioru danych mogło negatywnie wpłynąć na wyniki naszego modelu, ponieważ model mógł być bardziej skłonny do przewidywania bardziej reprezentowanych klas. Po drugie, różnice w preprocessing danych, takie jak lematyzacja, stemming oraz usunięcie stop words, mogły wpłynąć na efektywność modelu w różny sposób.

Lematyzacja i stemming mają na celu redukcję słów do ich podstawowej formy, co może zmniejszyć różnorodność danych wejściowych i pomóc modelowi skupić się na istotnych cechach. Usunięcie stop words natomiast może pomóc w eliminacji mniej istotnych słów, ale jednocześnie może usunąć istotne informacje kontekstowe, co może być powodem nieznacznie niższej dokładności w naszym przypadku.

Wyniki sugerują, że różne techniki przetwarzania wstępnego mogą mieć znaczący wpływ na wydajność modelu. Lematizacja bez usuwania stop words dała najlepsze wyniki, co sugeruje, że zachowanie pełnego kontekstu może być kluczowe dla naszego modelu.

Dodatkowo co mozemy zauważyć, to macierze pomyłek są najlepsze dla takich gatunków jak metal czy hiphop, a w drugiej kolejności rap i jazz. Może to wynikać z charakterystycznych cech tekstowych tych gatunków muzycznych, które są łatwiejsze do uchwycenia przez nasz model. Teksty w gatunkach takich jak metal czy hiphop mogą zawierać bardziej specyficzne słownictwo, powtarzające się frazy i unikalne struktury, które są bardziej rozpoznawalne dla naszego modelu, co mogliśmy zauważyć w rozdziale z analizą danych - rap i hiphop miały najdłuzsze sekwencje i najwięcej unikalnych słów. Hiphop i rap to bardzo podobne warunki, ale w naszym datasecie hip hop posiadał większą ilość próbek, dlatego pewnie wypadł lepiej podczas treningu. Najgorzej wypadly najmniej charakterystyczne gatunki - pop, rock, country oraz indie - te cztery gatunki same w sobie są do siebie bardzo podobne pod względem tekstu — czymś co mogłoby poprawić te wyniki jest uczenie multimodalne i wydobycie informacji o muzyce.

Odpowiadając na pytanie o genralizację tego modelu — niestety zauważylismy, że pik dokładności na zbiorze walidacyjnym osiągamy w 8-10 powtórzeniu, natomiast dokładność na zbiorze treningowym dalej rośnie - model zaczyna zbytnio dopasowywać się do danych. Sama klasyfikacja na zbiorze tetsowym wyniosła około 45%, czyli niestety nie zbyt dobrze.

6.2 Podejście oparte na transformerach z wykorzystaniem pretrenowanego modelu BERT

W drugim podejściu zastosowaliśmy transformery, a dokładniej pretrenowany model BERT (Bidirectional Encoder Representations from Transformers).

BERT to model oparty na architekturze transformerów. Jedną z głównych cech BERT-a jest jego dwukierunkowe przetwarzanie tekstu, co oznacza, że analizuje on kontekst słów zarówno z lewej, jak i z prawej strony. To odróżnia go od tradycyjnych modeli przetwarzających tekst tylko w jednym kierunku. BERT został wstępnie wytrenowany na ogromnych zbiorach danych tekstowych (Wikipedia i BookCorpus) przy użyciu dwóch zadań: maskowania słów (Masked Language Modeling, MLM) oraz przewidywania następnego zdania (Next Sentence Prediction, NSP). W zadaniu MLM losowo maskuje się niektóre słowa w tekście, a model uczy się je przewidywać na podstawie otaczającego kontekstu. W zadaniu NSP model uczy się przewidywać, czy jedno zdanie następuje po drugim, co pomaga w zrozumieniu relacji między zdaniami.

BERT może być dostosowany do konkretnych zadań (np. klasyfikacja tekstu, analiza sentymentu) poprzez proces fine-tuningu, gdzie model jest dalej trenowany na mniejszych, specyficznych zbiorach danych. W naszym projekcie zastosowaliśmy pretrenowany model BERT do klasyfikacji tekstów piosenek do 8 gatunków.

Ze względu na duży koszt GPU i czasowy wybraliśmy dane które zostały poddane lematyzacji i je wykorzystaliśmy tutaj dalej. Wykorzystaliśmy pretrenowany model BERT i dostosowaliśmy go do naszego zadania klasyfikacji. Proces fine-tuningu obejmował dalsze trenowanie modelu na naszym specyficznym zbiorze danych.

Zastosowanie BERT-a w naszym projekcie ma kilka potencjalnych korzyści. Dzięki dwukierunkowemu przetwarzaniu tekstu, BERT lepiej rozumie kontekst słów, co powinno prowadzić do poprawy wyników klasyfikacji w porównaniu do tradycyjnych metod. Wykorzystanie pretrenowanego modelu pozwala na korzystanie z wiedzy zdobytej na ogromnych zbiorach danych, co może być szczególnie korzystne, gdy mamy ograniczoną ilość danych specyficznych dla naszego zadania.

Podsumowując, drugie podejście oparte na transformerach z wykorzystaniem pretrenowanego modelu BERT oferuje zaawansowane możliwości przetwarzania tekstu, które mogą znacząco poprawić wyniki klasyfikacji tekstów piosenek. Mamy nadzieję osiągnąć wyższą dokładność i lepsze wskaźniki ewaluacyjne w porównaniu do poprzednich metod opartych na BiLSTM, CNN i GloVe jednak ze względu na bardzo duże napotkane problemy z GPU, koniecznością wykupienia Colaba Pro i bardzo wolnym treningiem nie uzyskałysmy dla niego jeszcze pełnych wyników. Po 1 epoce model skończył na dokładności na poziomie 45% co jest porównywalnym wynikiem do pierwszej techniki.

Literatura

1. Affreen Ara and Raju Gopalakrishna. *A Study on Emotion Identification from Music Lyrics*, pages 396–406. 05 2021.
2. Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. Music mood detection based on audio and lyrics with deep neural net. *ArXiv*, abs/1809.07276, 2018.

3. Yudhik Agrawal, Ramaguru Guru Ravi Shanker, and Vinoo Alluri. Transformer-based approach towards music emotion recognition from lyrics. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, pages 167–175, Cham, 2021. Springer International Publishing.
4. Sebastian Raschka. MusicMood: Predicting the mood of music from song lyrics using machine learning. *ArXiv*, 1611.00138v1, 2016.
5. Alexandros Tsaptsinos. Lyrics-Based Music Genre Classification Using A Hierarchical Attention Network. *ArXiv*, 1707.04678v1, 2017.
6. Krols Tibor, Nikolova Yana, and Oldenburg Ninell. Multi-Modality in Music: Predicting Emotion in Music from High-Level Audio Features and Lyrics. *ArXiv*, 2302.13321v1, 2023.
7. Nave Gideon, Minxha Juri, M Greenberg David, Kosinski Michal, Stillwell David, and Rentfrow Jason. Musical Preferences Predict Personality: Evidence From Active Listening and Facebook Likes. *Psychological Science*, 29:1145–1158, 2018.
8. https://github.com/sergiooramas/elvis. Accessed: 2024-04-03.
9. Ricardo Malheiro, Renato Panda, Paulo Gomes, and Rui Pedro Paiva. Music emotion recognition from lyrics: A comparative study. 09 2013.
10. https://jmir.sourceforge.net/jLyrics.html. Accessed: 2024-04-03.
11. https://github.com/parthenocissus/synesketch_v2.1/. Accessed: 2024-04-03.
12. https://conceptnet.io/. Accessed: 2024-04-03.
13. Stanford CS224N, Megan Leszczynski, Anna Boonyanit, and Andrea Dahl. Music genre classification using song lyrics. 2021.
14. Preniqi Vjosa, Kalimeri Kyriaki, and Saitis Charalampos. Soundscapes of morality: Linking music preferences and moral values through lyrics and audio. *PLoS ONE*, 18, 2023.
15. https://likeyouth.org. Accessed: 2024-04-03.
16. Gideon Nave, Juri Minxha, David M. Greenberg, Michal Kosinski, David Stillwell, and Jason Rentfrow. Musical preferences predict personality: Evidence from active listening and facebook likes. *Psychological Science*, 29(7):1145–1158, 2018. PMID: 29587129.
17. Nikhil Nayak. Genius song lyrics. https://www.kaggle.com/datasets/carlosgdcj/genius-song-lyrics-with-language-information.
18. Anderson Neisse. Song lyrics from 79 musical genres. https://www.kaggle.com/datasets/neisse/scrapped-lyrics-from-6-genres.