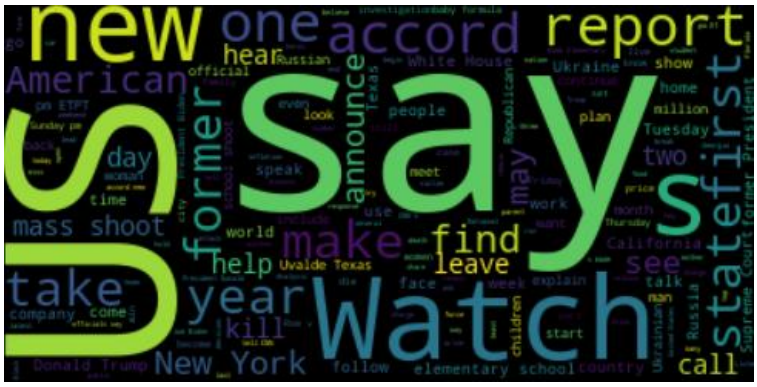
Do którego celebryty należy tweet?

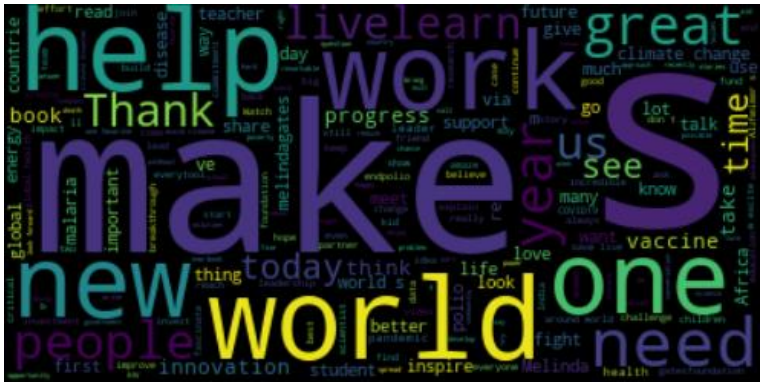
W celu zrealizowania projektu posłużyłam się bazą danych o około 20 000 rekordów, zawierających tweety z kont takich osób jak – Elon Musk, Oprah Winfrey czy Donald Trump. W sumie zebrałam treści z sześciu różnych kont.

Na początek, po wstępnej obróbce tekstu użyłam wordcloud, aby przedstawić najbardziej charakterystyczne słowa dla każdego z kont.

CNN



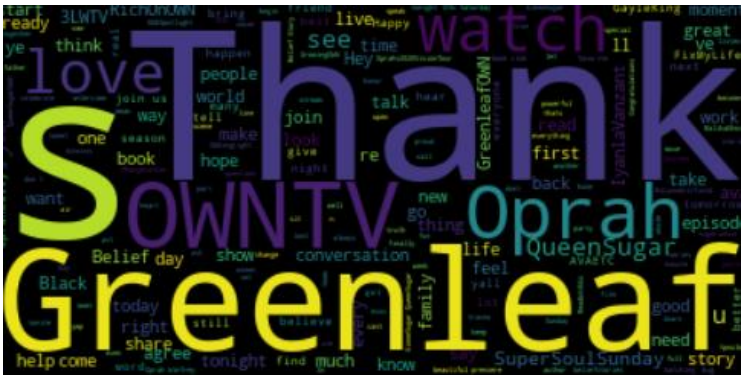
Gates



Musk

NASA

Oprah



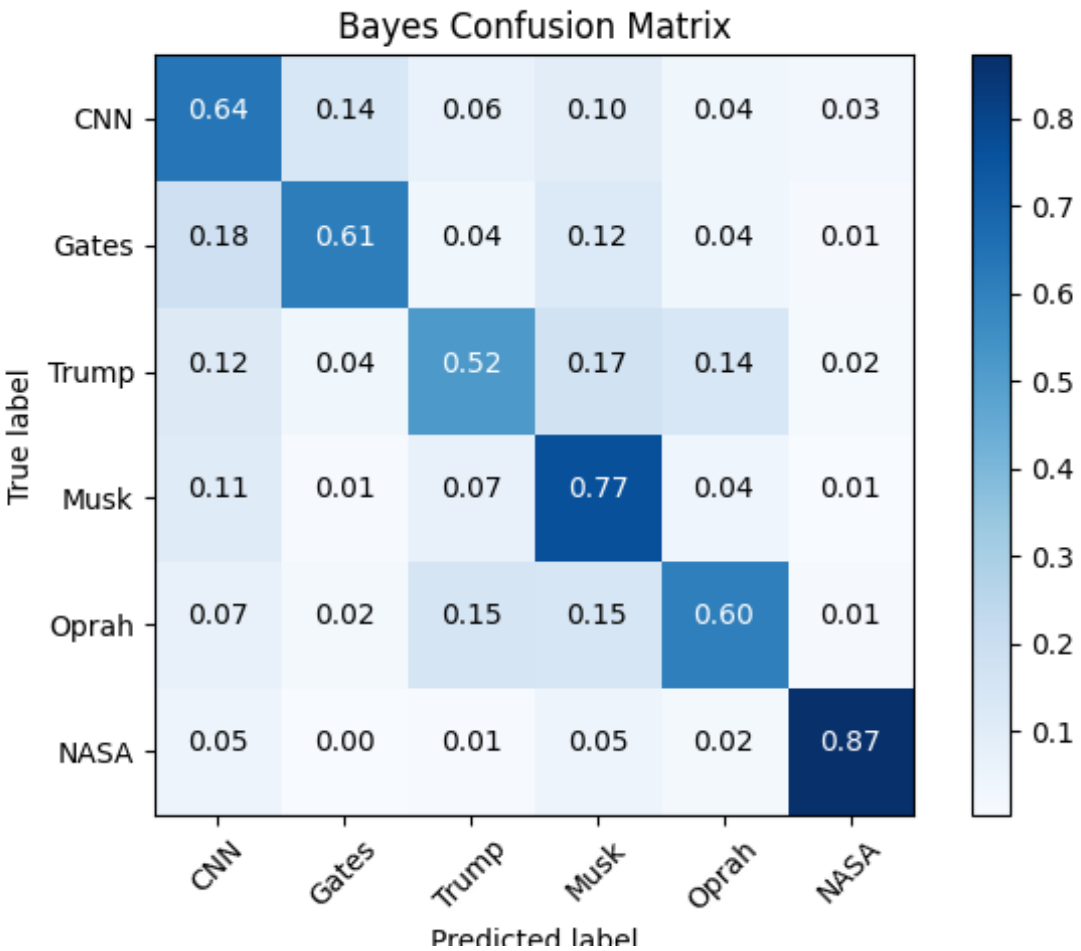
Trump

Jak widać, wypowiedzi są charakterystyczne dla każdego z autorów, chmury słów są bardzo różne, jednak na pierwszy rzut oka, zdecydowanie trudniej powinno być rozróżnić między CNN, Oprah czy Billem Gates'em ze względu na to, że największe słowa, nie są tak unikatowe jak np. President dla Donalda Trumpa lub Tesla dla Elona Muska.

Naive Bayes

Pierwszym klasyfikatorem, którego użyłam jest Naive Bayes, czyli klasyfikator probabilistyczny. Spodziewałam się po nim słabych wyników, jednak jak widać w przypadku NASA, którego tweety są bardziej charakterystyczne poradził sobie całkiem nieźle.

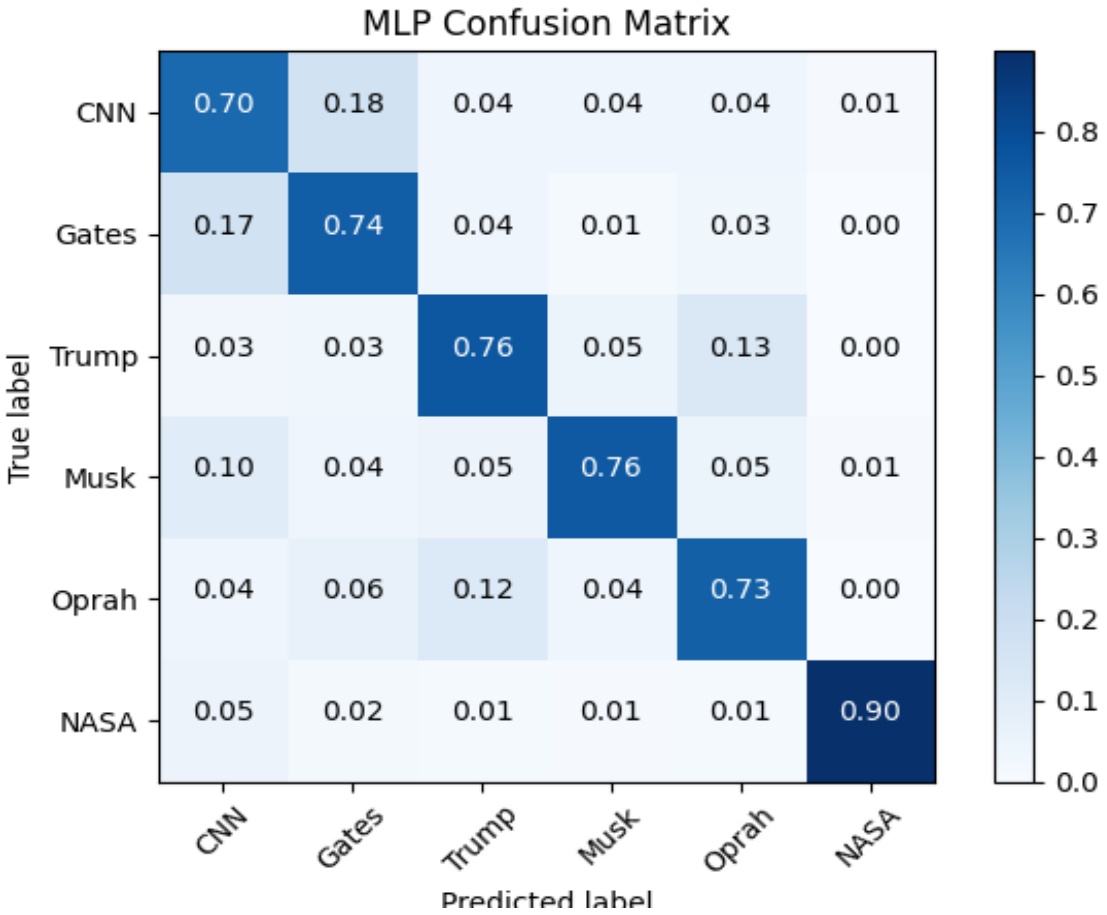
Uśredniając wyniki Naive Bayes zakończył z wynikiem 64% skuteczności.



Multilayer Perceptron

Kolejnym klasyfikatorem, zostało MLP, w którym pokładałam największe nadzieje z uwagi na to, że jest to klasyfikator wykorzystujący sztuczną sieć neuronową i jako jedyny podchodzi do problemu w bardziej złożony sposób. Podobnie jak przy użyciu klasyfikatora Bayesa, rozpoznanie NASA było najprostsze i w 90% przypadków udało się to bez problemu, znaczną przewagę MLP widać dla wszystkich innych autorów, przy których skuteczność sieci utrzymuje się na wysokim około 70% poziomie. Co warto zaznaczyć MLP zajmowało zdecydowanie najwięcej czasu na obliczenia, ze wszystkich użytych przeze mnie algorytmów.

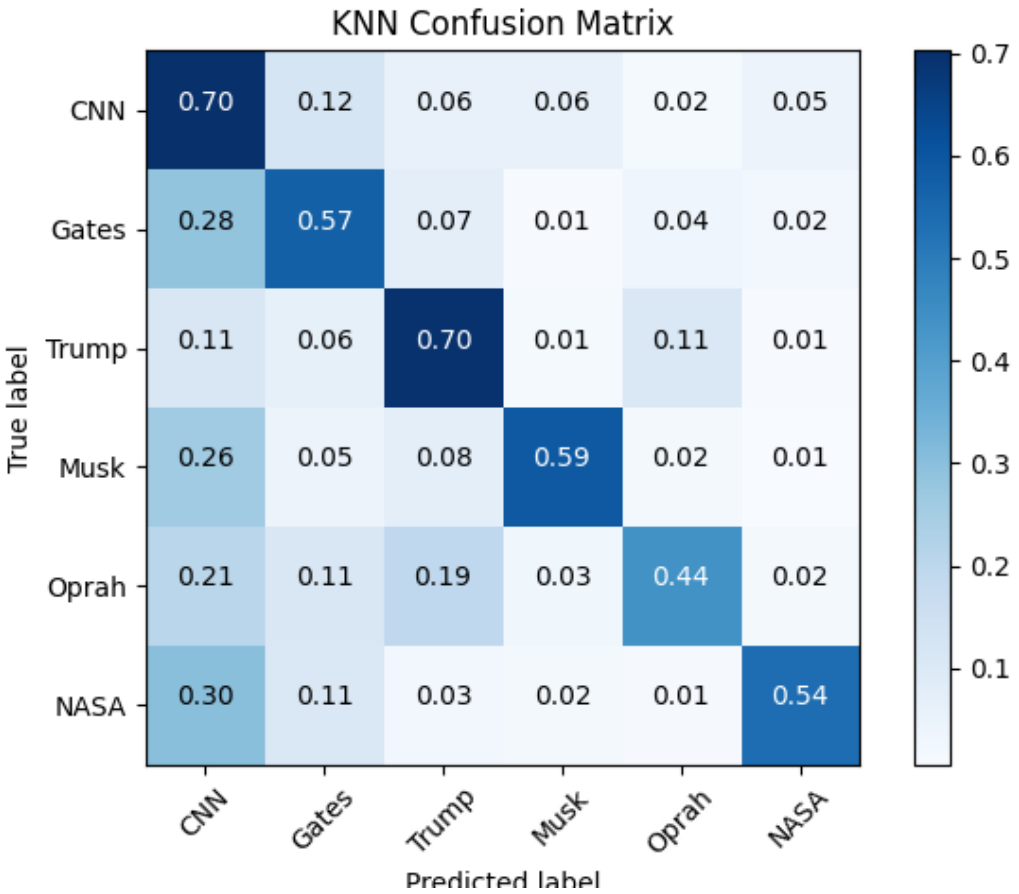
Multilayer Perceptron osiągnął średni wynik w wysokości 75%.



K- nearest neighbors

Czyli k- najbliższych sąsiadów, najprostszy z algorytmów, który nie nadaje się do rozwiązywanie bardziej złożonych problemów. Jak widać poniżej, macierz KNN najbardziej różni się wynikami od MLP i Bayesa. NASA, które jak udało nam się ustalić, najłatwiej rozpoznać spośród innych autorów, uzyskało wynik tylko 54%, a za to CNN, które było jednym z najtrudniejszych do rozpoznania, dla KNN było najłatwiejsze.

Klasyfikator KNN ostatecznie osiągnął 69% skuteczności, co o dziwo nie odstaje znacznie od dwóch pozostałych.



Wnioski

Niestety żaden z algortymów nie zbliżył się do 90% skuteczności. Jednak moim zdaniem, mimo kosztów czasowych jakie idą z zastosowaniem klasyfikatora MLP, to on sprawdza się najlepiej dla uczenia maszynowego z analizą tekstów, być może przy użyciu większej bazy danych i dobraniu najlepszych parametrów, osiągnąłby porządane wyniki, bez dwóch zdań to on ma na to największy potencjał.

Autor: Natalia Turska