Paweł Zawistowski 10.06.2024

Inteligencja Obliczeniowa
Analiza tekstu z mediów społecznościowych

1 Wstęp

Zbiór danych użyty do analizy zawiera ponad 500 tysięcy postów z platformy **X** (dawniej Twitter) dotyczących wydarzenia *FIFA World Cup 2018*. Posty obejmują okres od 29 czerwca 2018 roku (1/8 finału) do 15 lipca 2018 roku, kiedy odbył się finał wygrany przez Francję. Dane zostały pobrane z serwisu **Kaggle**.

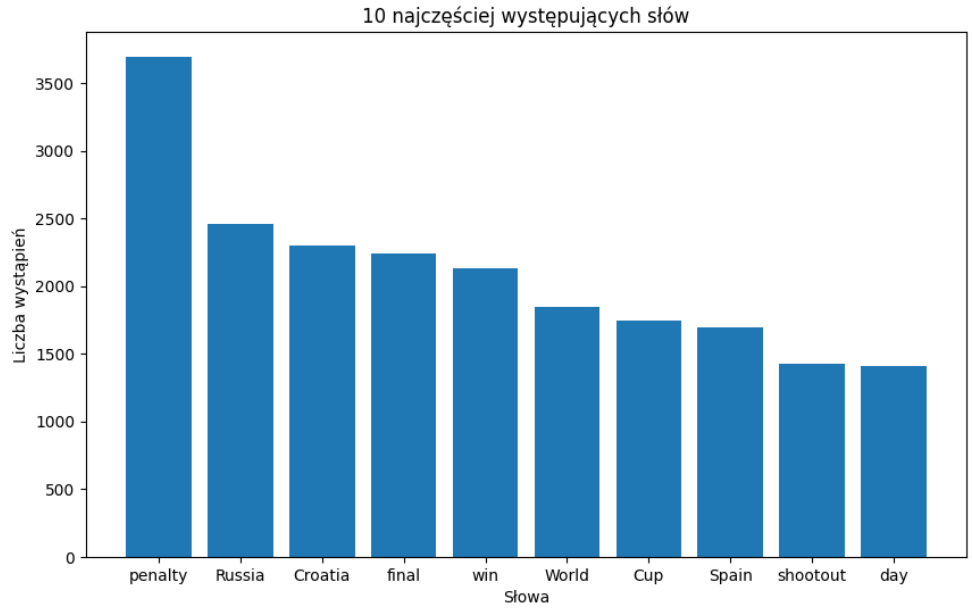
2 Preprocessing

Preprocessing danych tekstowych z mediów społecznościowych obejmował czyszczenie danych, tokenizację, usuwanie stop słów oraz lematyzację. W implementacji w Pythonie użyto biblioteki **nltk** do załadowania stop słów, tokenizacji tekstu oraz lematyzacji tokenów. Dodatkowe stop słowa specyficzne dla analizowanej treści zostały również uwzględnione. Teksty zostały przekształcone do formy bardziej odpowiedniej do dalszej analizy.

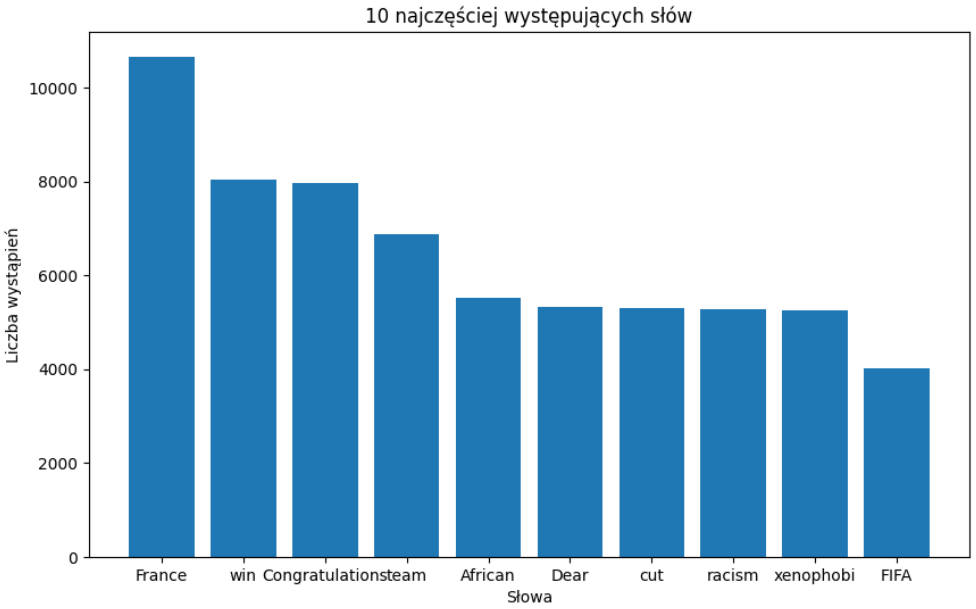
```
78  stop_words = set(stopwords.words('english'))
79  lemmatizer = WordNetLemmatizer()
80  additional_stop_words = ["also", '``', '\"', '"', '"', "v", "vs"]
81  stop_words.update(additional_stop_words)
    1 usage   pz090922*
82  def preprocess(text):
83      tokens = word_tokenize(text)
84      tokens = [lemmatizer.lemmatize(token.lower()) for token in tokens if token.isalpha() and token.lower() not in stop_words]
85      return tokens
```

3 Word Frequency i Word Cloud

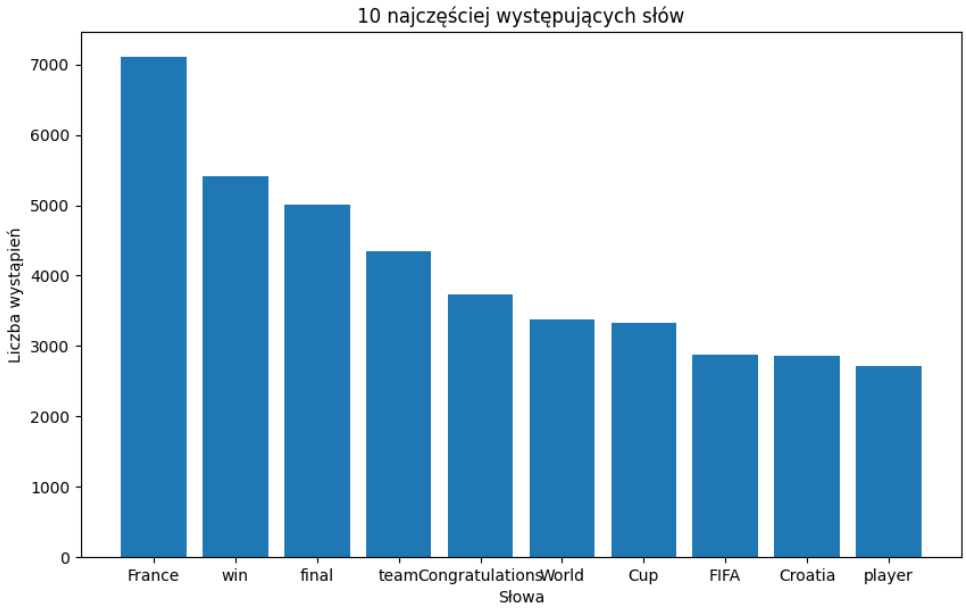
W ramach analizy danych dotyczących Mistrzostw Świata w Piłce Nożnej 2018, przeprowadzono badanie częstotliwości używanych słów. Analiza ta miała na celu zidentyfikowanie, które słowa były najczęściej używane w różnych fazach turnieju: 1/8 finału, podczas finału oraz w całym okresie mistrzostw. Oto jak prezentują się wyniki:



Rysunek 1: Najpopularniejsze słowa w dniu 1/8 finału



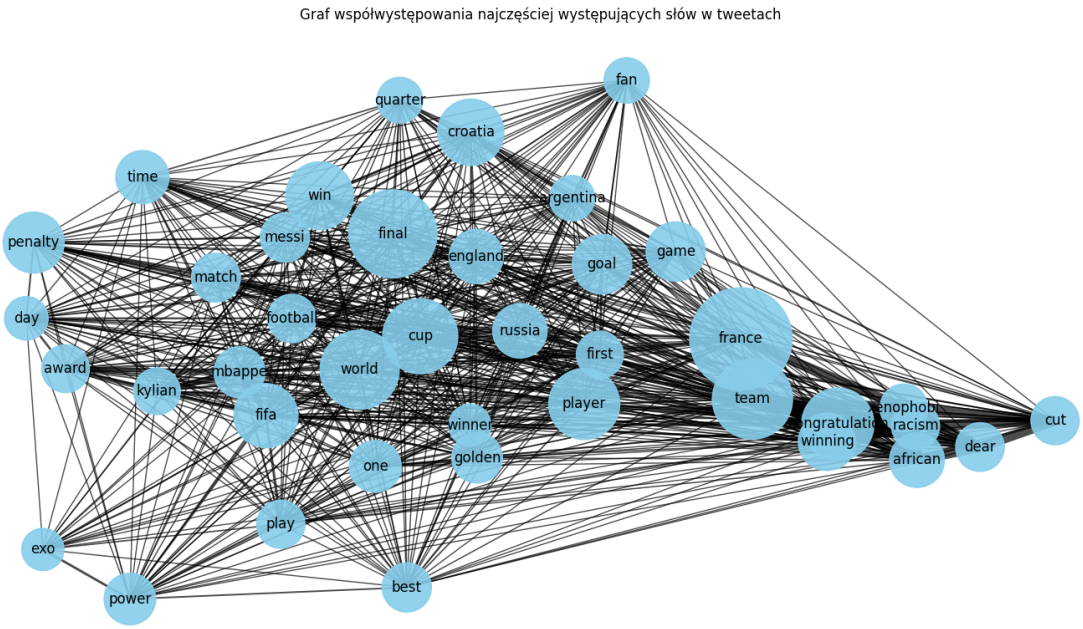
Rysunek 2: Najpopularniejsze słowa w dniu finału mistrzostw



Rysunek 3: Najpopularniejsze słowa podczas całych mistrzostw

Rysunek 4: Chmura tagów

Jak widać na wykresie dotyczącym 1/8 finału, najczęściej występującymi słowami są: **Penalty**, **Russia** i **Coratia**. Jest to niewątpliwie związane z rozstrzygającym konkursem rzutów karnych między reprezentacją Chorwacji a reprezentacją Rosji. Natomiast w dniu wielkiego finału, najczęściej używanymi słowami są: **France**, **Win** oraz **Congratulations**. Ten wynik jest zrozumiały, biorąc pod uwagę fakt, że to Francji udało się zwyciężyć cały mundial. Analizując najpopularniejsze słowa w całym okresie turnieju, zauważamy, że niewiele różnią się one od tych używanych podczas finału. Jest to prawdopodobnie związane z tym, że największa liczba postów została opublikowana już po zakończeniu turnieju, kiedy emocje były najintensywniejsze.

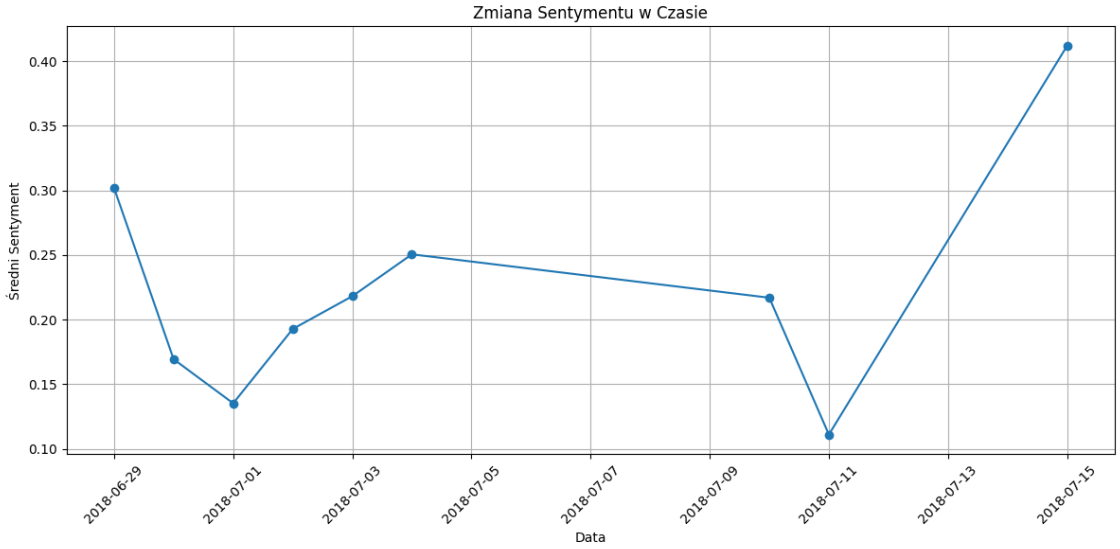


Rysunek 5: Graf współwystępowania najczęsciej wystepujących słow w tweetach

Powyższy graf przedstawia współwystępowanie słów w postach na platformie **X**. Rozmiar każdego węzła odwierciedla częstotliwośc występowania słowa. Krawędzie łączące węzły reprezentują współwystępowanie słów w tych samych postach. Grubość krawędzi wskazują na siłe współwystępowania. Na naszym grafie widzimy, że zcentralizowanymi węzłami są **Final**, **Cup**, **World** i to one mają najwięcej połączeń, co jest spodziewanym rezultatem. Z drugiej strony widzimy również słowo **France** wraz z takimi słowami jak: **Team**, **Congratulations**, ale też **Racism**, **African** czy **Xenophobic**, co za pewne niestety ma związek z aktami rasistowskimi ze strony innych kibiców, ponieważ skład reprezentacji Francji składa się w większości z osób ciemnoskórych.

4 Sentyment

W tej sekcji, przy użyciu narzędzia *Vader*, przeanalizowano zmiany sentymentu w trakcie trwania turnieju. Poniżej przedstawione są wyniki tej analizy:

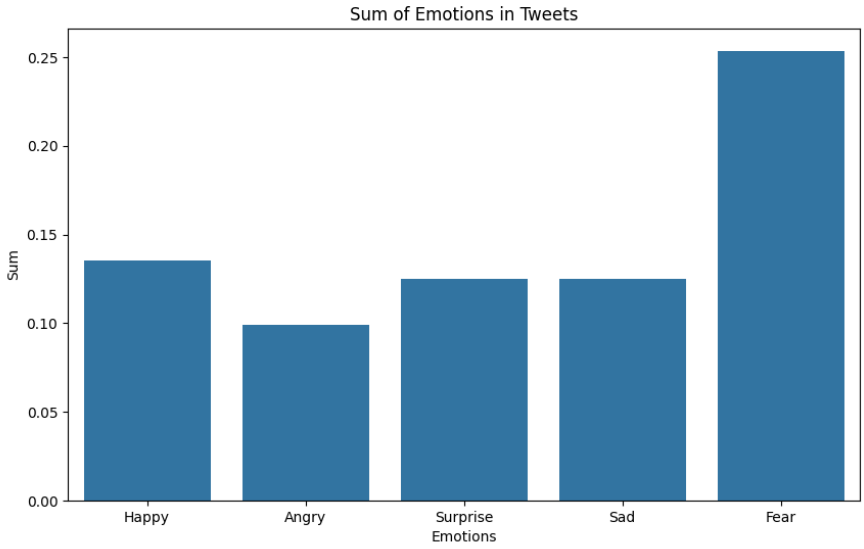


Rysunek 6: Wykres zmiany sentymentu w czasie

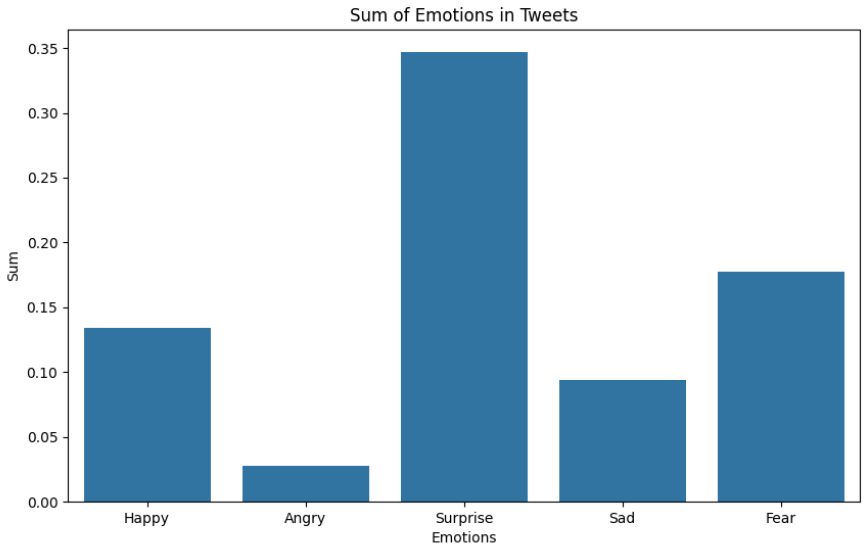
Analiza wykazała, że ogólny sentyment w postach był przeważnie pozytywny. Najniższy poziom sentymentu odnotowano 11 lipca, co prawdopodobnie jest związane z meczem pomiędzy reprezentacją Anglii a reprezentacją Chorwacji, zakończonym wynikiem 2:1 dla Chorwacji. Angielscy kibice, znani ze swojej ekspresyjności, mogli w tym czasie wyrażać swoje niezadowolenie, co mogło wpłynąć na nagły spadek sentymentu. Z kolei najwyższy poziom sentymentu odnotowano w dniu finału, co jest spodziewanym wynikiem, zwłaszcza że finał zakończył się bez większych kontrowersji.

5 Badanie emocji

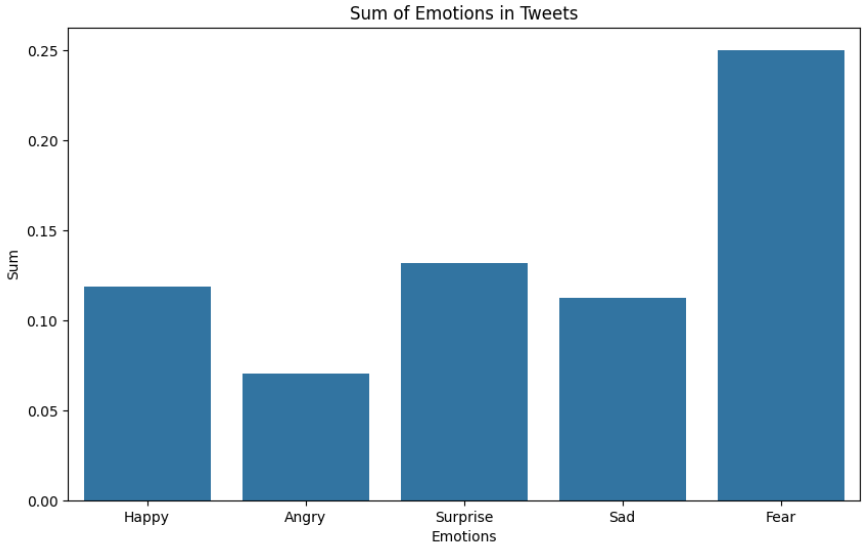
W niniejszej sekcji wykorzystano bibliotekę *text2Emotions* w celu analizy emocji związanych z 1/8 finału, finałem oraz całym wydarzeniem. Poniżej przedstawione są wyniki analizy:



Rysunek 7: Emocje podczas 1/8 finału



Rysunek 8: Emocje podczas finału



Rysunek 9: Emocje podczas całego turnieju

Po analizie wykresów zauważyliśmy ciekawy trend: w 1/8 finału dominuje emocja strachu, co budzi pewne zaskoczenie, biorąc pod uwagę brak dramatycznych wydarzeń w tej fazie turnieju. Istnieje podejrzenie, że może to być wynikiem niedoskonałości użytej biblioteki, bowiem przecież nie odnotowano żadnych szczególnie stresujących momentów w tym czasie. Natomiast w przypadku finału obserwujemy dominację emocji zaskoczenia, co prawdopodobnie związane jest z historycznym osiągnięciem Chorwacji, po raz pierwszy w historii docierając do finału Mistrzostw Świata. Przyglądając się całemu przebiegowi mistrzostw, zauważamy ponownie wyraźne występowanie emocji strachu. Niemniej jednak, pozostałe emocje utrzymują się na stabilnym poziomie, co jest spodziewanym rezultatem – zwycięstwo wywołuje radość, a porażka smutek. To pokazuje, że turniej był pełen napięcia, dając fanom piłki nożnej szeroki wachlarz emocji, które towarzyszą sportowej rywalizacji.

6 Klasteryzacja

W badaniu użyto klasteryzacji tekstu, konkretniej algorytmu *Latent Dirichlet Allocation (LDA)*, w celu analizy dyskusji na platformie **X** związanych z Mistrzostwami Świata FIFA 2018. Na początku przetworzono teksty tweetów, używając techniki TF-IDF (Term Frequency-Inverse Document Frequency). Następnie zastosowano wspomniany już algorytm *LDA*, który jest popularnym narzędziem w analizie tematycznej tekstów. Działa on poprzez grupowanie słów wzdłuż tematów, które są ukryte w zbiorze danych. O to jak się prezentują najciekawsze wygenerowane tematy:

Rysunek 10: Tematy wygenerowane za pomocą LDA

Analizując wyodrębnione tematy, zauważamy, że algorytm LDA wykonał imponującą pracę. W szczególności temat 7 wydaje się być niezwykle trafny. Odwołuje się on do Kyliana Mbappe, który w tamtym okresie stał się rzeczywiście gwiazdą wschodzącą i zagrał genialnie podczas turnieju. Podobnie można zauważyć odniesienie do Thibauta Courtoisa, w którym uzyskał on nagrodę Golden Glove, co dodatkowo podkreśla trafność tego tematu.

7 Podsumowanie

Podsumowując, uważam, że cały projekt zakończył się sukcesem. Wyniki analizy były interesujące i w dużej mierze zgodne z oczekiwaniami. Odkryto wiele pozytywnych aspektów, które odzwierciedlały przewidywane tendencje. Niemniej jednak, zauważono również pewne negatywne aspekty, w szczególności wysoką intensywność emocji strachu wśród tweetów, co jednak przypisuję bardziej specyfice użytej biblioteki niż samym tweetom. Pomimo tych zastrzeżeń, ogólny obraz był zgodny z moimi przewidywaniami i stanowi wartościowy wkład w zrozumienie sentymentu i emocji wyrażanych przez użytkowników mediów społecznościowych podczas turnieju.

8 Źródła informacji

Baza Danych: **Link**
Chat-GPT: **Link**