Uniwersytet Gdański

Majkowska Agata
Migdał-Najman Kamila
Najman Krzysztof
Raca Katarzyna

Identyfikacja grup wiekowych użytkowników portalu X

www.ug.edu.pl

Plan prezentacji

Problem i cel badawczy

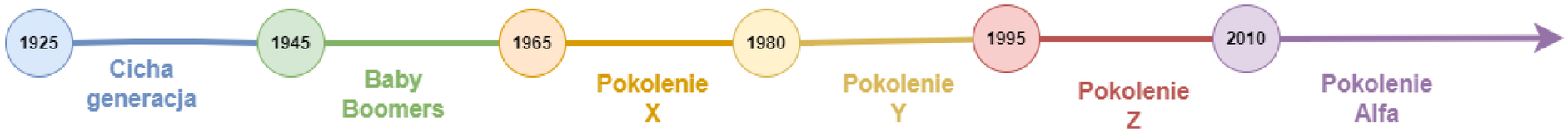
Z badań językowych wynika, że przedstawiciele różnych grup wiekowych posługują się **odmiennym słownictwem** i formami gramatycznymi. Wydaje się, że różnicują je także powszechnie stosowane w wypowiedziach **znaki graficzne**, takie jak emotikony, emoji, piktogramy i inne znaki graficzne.

Celem prezentowanych badań jest próba identyfikacji wieku użytkownika portalu X na podstawie **charakterystycznych elementów w tekście takich jak słowa czy emoji**. Do realizacji badania wykorzystano narzędzia Text Mining oraz naiwny klasyfikator Bayesa.

Pokolenie

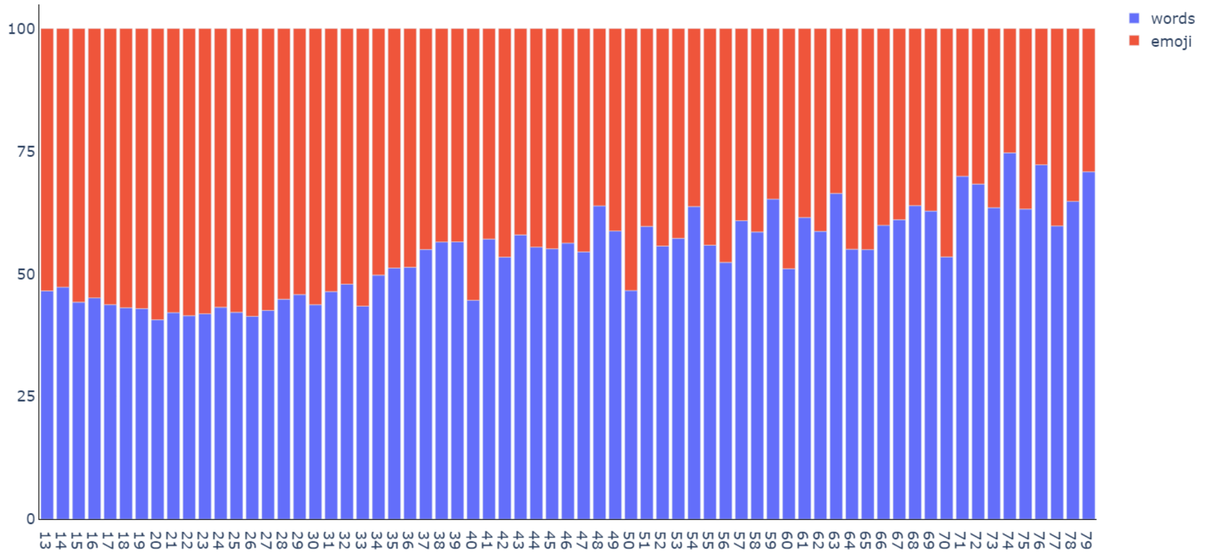
Grupa, która ma wspólne dzieciństwo, wspólny wiek młodzieńczy i u których etap dojrzałości przypada na ten sam okres. Generację tworzą jednostki, które uczestniczyły w tych samych przemianach i wydarzeniach, jakie miały miejsce w okresie ich pobudliwości. Źródłem tej odmienności są różne doświadczenia i perspektywy życiowe. [Dilthey, 1924]

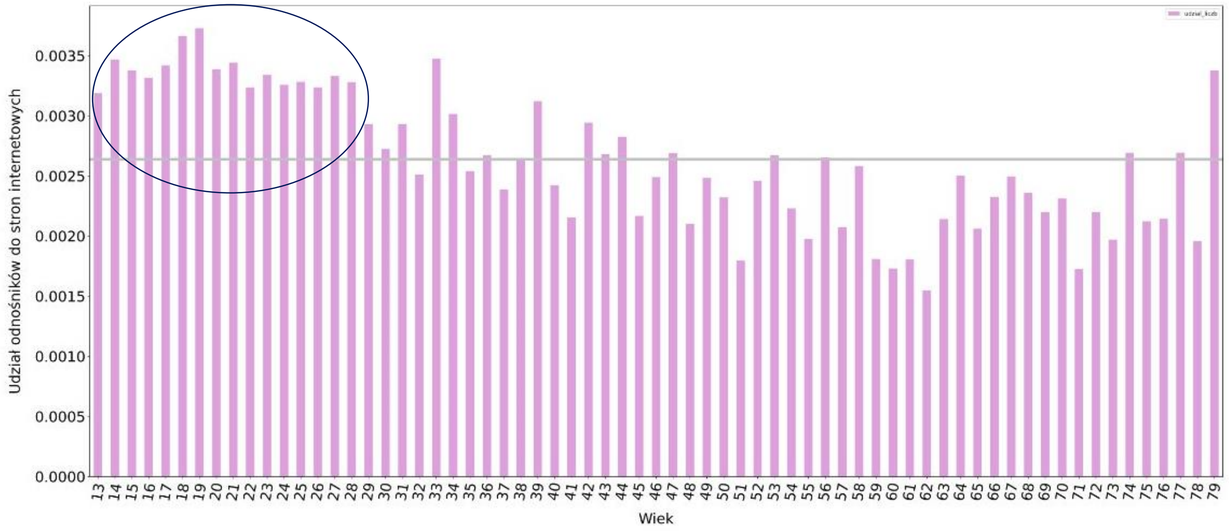
Grupa społeczna, która nawiązuje ze sobą wspólną więź i wspólnie działa. Charakteryzuje się specyficznymi wzorcami zachowań, widocznie różnymi od innych grup postawami, poglądami, uznawanymi wartościami, aspiracjami, sposobem życia czy na swój sposób odmiennym pojmowaniem rzeczywistości. W definicji tej zwraca się uwagę na aspekt czasowy i jakościowy. [Sztompka, 2003]

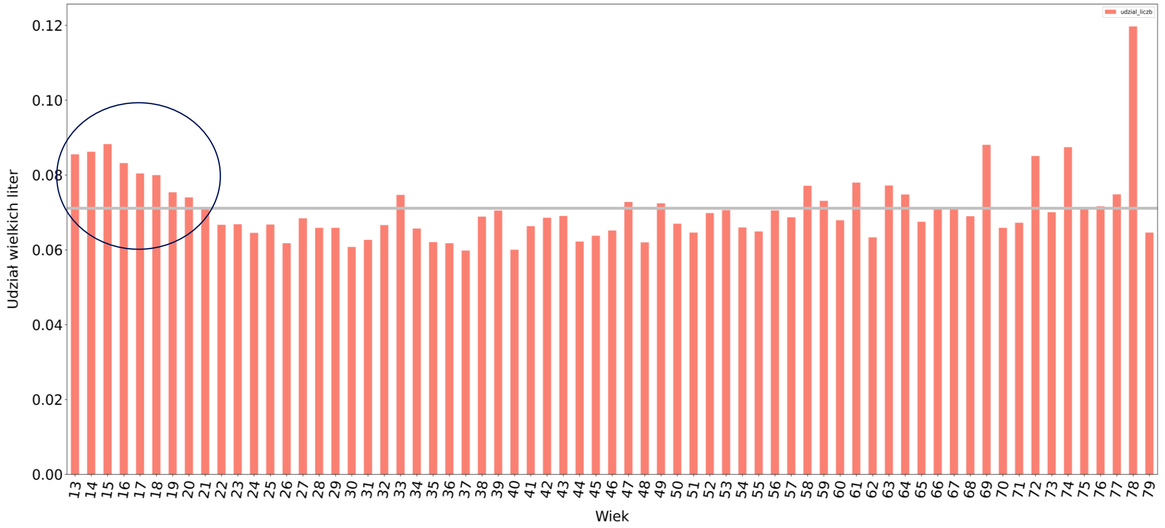


Uzyskane grupowania we wcześniejszych badaniach

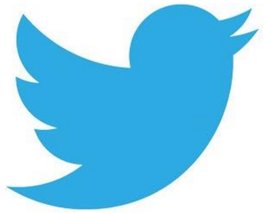
Metoda grupowania	Liczba powstałych grup	Cechy wyodrębnione do analizy
grupowanie aglomeracyjne	4	słowa
sieć neuronowa	5	emoji
grupowanie aglomeracyjne	4	emotikony, stopwords, interpunkcja, liczba wzmianek o użytkownikach (@użytkownik), udostępnianie linków do stron, liczba wielkich liter
grupowanie aglomeracyjne	4	słowa i emoji







Zbieranie danych



API TWITTERA

„happy birthday to me”

„I am” + x + „years old”

14 224 użytkowników
około 4 miliony tweetów

R Studio

1.

l.p.	user_name	year
0	Surendr37595033	17
1	seulgigomie	20
2	monumental2020	22
3	iamjuwijichu	22
4	manideeraj	23
5	RealLilbaddie	23
6	poserman554	25
7	Killer88Munir	28
8	Adam_GSmith	30
9	BlainToni	44
10	Rb2LzB4JZZgauqv	45

2.

	user_name	year	tweets
0	Surendr37595033	17	@sierrabensen122 @cr3wdawg @TheEllenShow Who t...
1	seulgigomie	20	@RVsmtown Best girls https://t.co/B2WtzktvUO @...
2	monumental2020	22	This move is long overdue https://t.co/TjjwkET...
3	iamjuwijichu	22	@nvsmlnd Serem poll kak icaa :(@nvsmlnd Iya k...
4	manideeraj	23	@PuneethRajkumar Sir your suddenly looking lik...

3.


```
{'id': 1288530861046956032,
 'id_str': '1288530861046956032',
 'name': 'Surendra',
 'screen_name': 'Surendr37595033',
 'location': '',
 'profile_location': None,
 'description': 'my dream want to live and die in america...an
Katy perry',
 'url': None,
 'entities': {'description': {'urls': []}},
 'protected': False,
 'followers_count': 3,
 'friends_count': 5,
 'listed_count': 0,
 'created_at': 'Wed Jul 29 17:44:55 +0000 2020',
 'favourites_count': 38,
 'utc_offset': None,
 'time_zone': None,
 'geo_enabled': False,
 'verified': False,
 'statuses_count': 112,
 'lang': None,
 'status': {'created_at': 'Mon Aug 24 09:16:14 +0000 2020',
  'id': 1297825030118629376,
  'id_str': '1297825030118629376',
  'text': '@gableigh39 @TheEllenShow Who stalker...??lol',
  'truncated': False,
  'entities': {'hashtags': [],
   'symbols': [],
   'user_mentions': [{'screen_name': 'gableigh39',
```


year	tweet	emoji_count	word_count	emoji/words	emoji	sorted_emoji
13	@antmccrossan what does B mean in the Tour de ...	8915	210641	0.042323	{'😎': 30, '😑': 71, '💃': 7, '😕': 324, '♡': 25, ...	{'😭': 797, '😂': 606, '❤': 446, '😕': 324, '💜': ...
14	i'm easy apparently 😟💔he said he has big hands...	8332	226085	0.036853	{'😟': 10, '💔': 48, '😰': 4, '🍆': 1, '😁': 22, '‼...	{'😭': 1008, '😂': 530, '❤': 498, '😕': 318, '🙂':...
15	@byuwo2 Right???? Right?? It's fucking infuria...	15120	267620	0.056498	{'💔': 180, '💛': 59, '✨': 249, '❤': 758, '😳': 5...	{'😭': 1699, '😂': 908, '❤': 758, '😕': 540, '😍':...
16	@xoitsmia Can I volunteer?@Reasonsrr U to. R u...	12788	277268	0.046121	{'😟': 34, '😰': 3, '😳': 5, '❓': 19, '😢': 7, '😒'...	{'😭': 1490, '😂': 865, '❤': 580, '💜': 578, '🙏':...
17	Go and gownlaod this link\n\nYou will get mone...	16257	333465	0.048752	{'👌': 11, '🙏': 105, '♥': 69, '🤣': 340, '😢': 22...	{'😭': 1550, '😂': 1324, '❤': 976, '😕': 591, '🍖'...
...	...	...	...	...	...	...

Przygotowanie danych i zastosowana metoda klasyfikacji

1.

Podział bazy użytkowników

Pokolenie Z (**7045** użytkowników)

Reszta grup wiekowych (**7179** użytkowników)

2.

Macierz danych

119 najczęściej występujących słów

120 najczęściej występujących emoji

liczba wielkich liter, liczba wzmianek innych użytkowników, liczba linków do stron

38 asocjacji słów - "emoji-słowo" i "słowo-emoji"

91 emotikon

3.

	😂	😭	❤️	😍	🤣	😊	💕	🙏	😘	💜
0	0	0	0	0	0	0	0	0	0	0
1	3	5	0	0	0	0	2	0	0	0
2	41	0	0	0	11	0	0	0	0	0
3	11	156	1	0	0	0	6	0	0	0
4	5	0	4	0	0	9	0	0	1	0
5	17	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	2	0	59	0	0	0	0	0	0	0
8	12	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0
11	15	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0

	get	every	fucking	good	done	person	new	would	thank	next
0	0	0	1	0	2	4	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	3	3	0	3	2	0	1	2	0	1
3	16	0	0	1	1	0	0	0	0	2
4	2	0	0	4	1	0	1	0	0	0
5	1	10	0	3	2	10	1	12	1	0
6	7	0	1	6	0	0	1	3	1	0
7	1	0	1	1	0	0	1	0	1	0
8	12	8	2	8	3	1	9	8	0	7
9	1	0	0	0	0	0	1	0	0	0
10	0	0	0	0	0	0	1	0	0	0
11	3	2	0	1	0	0	0	0	0	2
12	4	1	8	1	0	1	3	0	11	0
13	8	1	0	2	7	8	8	12	0	0
14	6	3	1	4	0	1	2	0	0	5
15	0	1	0	0	0	0	0	0	0	0

4.

```
[ "😂 love", "😭 omg", "❤️ love", "😊 thank","🤣 lmao","😍 love", "💜 bts","😇 thank",
"💕 thank", "✨ brunette", "😂 lol", "🤣 lol", "😭 thank", "❤️ love", "😍 love",
 "💜 thank", "🔥 fire", "😇 thank", "🙏 thank", "love 😂", "omg 😭","love ❤️",
 "thank 😊", "lmao 🤣", "love 😍", "bts 💜","thank 😇", "thank 💕", "brunette ✨",
 "lol 😂" ,"lol 🤣","thank 😭", "love ❤️", "love 😍", "thank 💜", "fire 🔥","thank 😇","thank 🙏"]
```

NAIWNY KLASYFIKATOR BAYESA

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Ocena podobieństwa uzyskanych klasyfikacji – macierz błędów

FAKTYCZNE DANE

PROGNOZA

	Pozytywna	**Negatywna**
Pozytywna	TP	FP
Negatywna	FN	TN

Precyzja – wskazuje udział liczby prawidłowo pozytywnie przydzielonych względem zaprognozowanych pozytywnie (TP/ TP+FP))

Czułość – wskazuje jaka część obserwacji jest prawidłowo pozytywnie przydzielona względem wszystkich pozytywnych tj. (TP/(TP+FN))

Dokładność – wskazuje jaka część obserwacji została zaklasyfikowana poprawnie (TP+TN)/N

F1-score – średnia harmoniczna z precyzji oraz czułości.(2TP/2TP+FP+FN)

Porównanie wyników klasyfikacji dla różnych zbiorów danych
Odsetek poprawnie zaklasyfikowanych użytkowników

Model	Klasa	Słowa	Emoji	Asocjacje	Emotikony	Wszystko
Naiwny klasyfikator Bayesa	Pokolenie Z	91%	77%	16%	39%	70%
	Reszta grup wiekowych	59%	63%	80%	72%	65%

Porównanie wyników klasyfikacji dla różnych zbiorów danych

Model	Słowa	Emoji	Asocjacje	Emotikony	Wszystko
Dokładność	72%	69%	54%	58%	67%
Precyzja	74%	67%	52%	56%	65%
Czułość	70%	74%	94%	85%	76%
F1-score	72%	71%	67%	67%	70%

Wnioski końcowe

- W identyfikacji pokolenia Z najwyższą precyzją (74%) i dokładnością (72%) charakteryzuje się model wykorzystujący jedynie słowa.
- Uwzględnienie w zbiorze danych innych elementów tekstu nie daje poprawy klasyfikacji.
- W identyfikacji pozostałych pokoleń najskuteczniejszy (80%) jest zbiór danych obejmujący asocjacje emoji - słowo.

Dalsze kierunki badań

- Zastosowanie innych modeli klasyfikacyjnych w tym regresji logistycznej, drzew klasyfikacyjnych i modeli AI.
- Walidacja modeli na nowych danych.
- Próba stworzenia modelu pozwalającego na klasyfikację użytkownika do pokolenia Z na podstawie pojedynczego wpisu.
- Próba stworzenia modelu pozwalającego na identyfikację wieku autora wpisu na podstawie wiadomości tekstowej napisanej w innym języku niż angielski.

Uniwersytet Gdański

Majkowska Agata
Migdał-Najman Kamila
Najman Krzysztof
Raca Katarzyna

Identyfikacja grup wiekowych użytkowników portalu X

www.ug.edu.pl