Analiza emocji tweetów na temat partii politycznych w roku wyborczym 2015

Piotr Lewandowski

26.01.2016

Wstęp

Projekt ma na celu zbadanie opinii wywoływanych przez dane partie polityczne na przestrzeni wyborów parlamentarnych 2015 i po wyborach.

Badanie może zostać wykorzystane przez agencje marketingowe obsługujące profile partii podczas kampanii do wyciągnięcia wniosków bądź przez wolnych czytaczy jako ciekawostka obrazująca obecną scenę polityczną.

Zbadane partie polityczne:

Nazwa partii	**@username**	**URL**
KORWIN	@partia_korwin	https://twitter.com/partia_korwin
Prawo i Sprawiedliwość	@pisorgpl	https://twitter.com/pisorgpl
Platforma Obywatelska	@Platforma_org	https://twitter.com/Platforma_org
.Nowoczesna	@_Nowoczesna	https://twitter.com/_Nowoczesna
Kukiz ’15	@Kukiz15	https://twitter.com/Kukiz15

Narzędzia

- **FluentLenium**
  Web scrapper do agregacji danych. API tweetera nie udostępnia metod do pobierania tweetó z odległej historii.. Dodany do paczki. Nie ma potrzeby osobnej instalacji.
  - Strona domowa: http://www.fluentlenium.org/
  - Dokumentacja: https://github.com/FluentLenium/FluentLenium/wiki

- **Java 8**
  Do uruchomienia projektu potrzebna jest JRE / JDK minimum w wersji 8

- **Gradle**
  Wykorzystywany do uruchamiania procedur projektu (pobieranie danych, analiza). Dodany do paczki. Nie ma potrzeby osobnej instalacji.
    - Strona domowa: http://gradle.org/

- **Libre Office**
  Tworzenie wykresów na podstawie wygenerowanych danych.

Dane

Sposób agregacji danych

- Okres: 6 miesięcy. Od 21 lipca 2015 do 17 stycznia 2016
- Maksymalnie 40 losowych tweetów dziennie dla każdej z partii
- Około 6 tysięcy tweetów na partie w ciągu całego badanego okresu
- Brane pod uwagę dni z co najmniej jednym tweetem
- Tweety brane pod uwagę są tylko te, które bezpośrednio odpowiadają na tweeta partii politycznej

Analiza emocji

Polski Instytutu Biologii Doświadczalnej im. M. Nenckiego w Warszawie opublikował w lipcu 2015 roku listę polskich słów emocjonalnych, bazując na Niemieckiej liście - Berlin Affective Word List (BAWL).

Lista zawiera 2,902 polskich słów skategoryzowanych na poszczególne emocje:
- Szczęście: 147 słów
- Gniew: 98 słów
- Smutek: 64 słowa
- Strach: 163 słów
- Wstręt: 48 słów
- Neutralne: 219 słów
- Słowa niesklasyfikowane: 2163

W badaniu założyłem, słowa neutralne i niesklasyfikowane jako jedną kategorię.
Każde słowo ma przypisany współczynnik dla każdej emocji.

Źródło badania wraz z bazą słów do pobrania: http://lobi.nencki.gov.pl/research/18/
Interaktywny wykres dla każdego słowa wraz z intensywnością emocji (współczynnik Arousal): http://exp.lobi.nencki.gov.pl/nawl-analysis

Użyta baza danych jest dołączona do projektu.

Teoria

Porównywanie podobności słów

Baza słów zawiera bezokoliczniki, natomiast tweety zawierają słowa odmienione przez przypadki.

Porównując słowa tweetów do oryginalnej bazy danych daje bardzo niską skuteczność klasyfikacji (poniżej 30%). Z użyciem algorytmu do obliczania podobności słów, zwiększyłem ten współczynnik do 66-70%.

Levenshtein distance
[EN] https://en.wikipedia.org/wiki/Levenshtein_distance
[PL] https://pl.wikipedia.org/wiki/Odleg%C5%82o%C5%9B%C4%87_Levenshteina

Algorytm rozszerzony o znormalizowanie wyniku do skali od 0 do 1.

Wzór
1 - (WSPÓŁCZYNNIK / DŁUGOŚĆ_DŁUŻSZEGO_NAPISU)
WSPÓŁCZYNNIK - ilość zmian jakie trzeba wykonać, by słowa były takie same.

Przykład
- orczyk
- oracz

Współczynnik: 3
Długość dłuższego wyrazu: 6
Wynik: 1 - (3 / 6) = 0.5
Wniosek: Słowa nie są podobne.

Słowa uznawane są za przypasowane jeżeli są w co najmniej w 76% podobne do jednego ze słów z bazy danych.
Im niższy wynik stwierdza podobieństwo słowa, tym większa ilość tweetów jest oznaczana jako sklasyfikowane.
Jeżeli 70% podobności słowa je klasyfikacje, ilość poprawnie sklasyfikowanych tweetów zwiększona jest do 90%. Jednak istnieje zbyt duże ryzyko błędnej klasyfikacji.

Obliczanie decydującej emocji tweeta

Każde słowo w bazie danych ma przypisane współczynnik (od 1 do 7) jak bardzo kwalifikuje się pod daną emocję. Dla jednego tweeta sumujemy wszystkie współczynniki i wybieramy emocję z największym wynikiem po zsumowaniu.

Przetasowanie bazy słów

Przy zastosowaniu powyższego algorytmu rozpoznawania słów, istnieje ryzyko, że słowa będą błędnie rozpoznane jako podobne i zakwalifikowane do początkowej kategorii. W celu zminimalizowania ryzyka, należy przelosować kolejność słów w bazie danych przed porównaniem.

Omijanie mało istotnych słów

Bierzemy pod uwagę tylko słowa dłuższe niż 3. Broni to przed błędnie zakwalifikowaniem spójników do emocjonalnych słów. Minimalnie zmienia to wyniki.

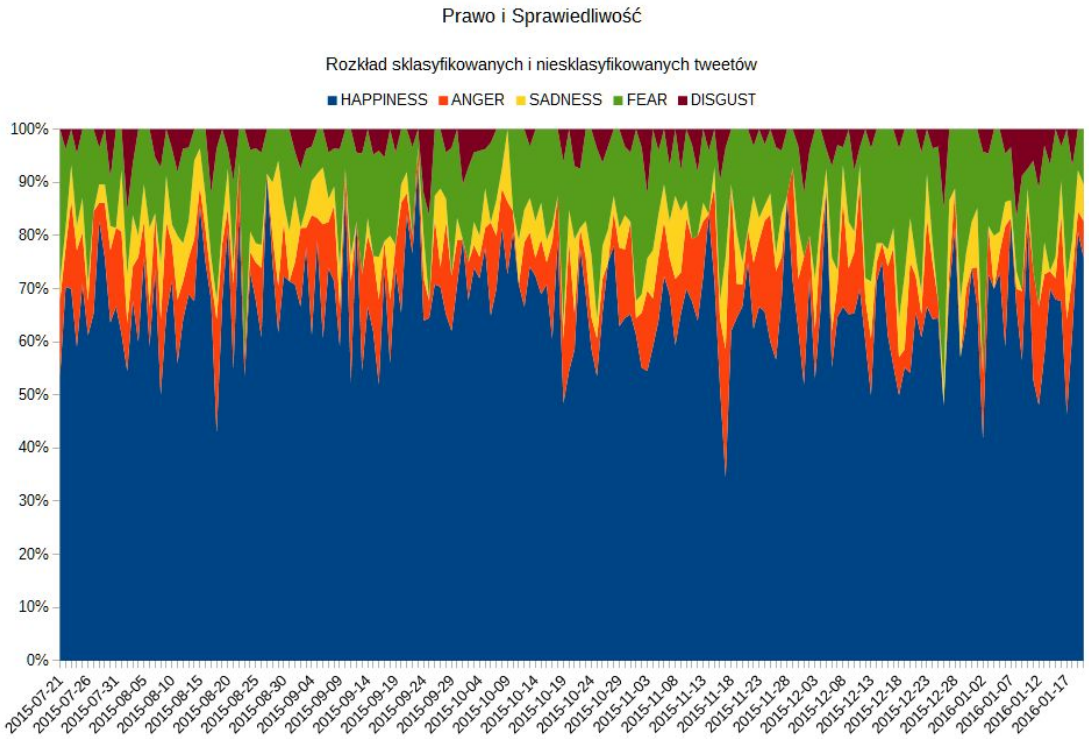
Po zastosowaniu zmiany: Rozpoznawalność tweetów spadła o 3%.

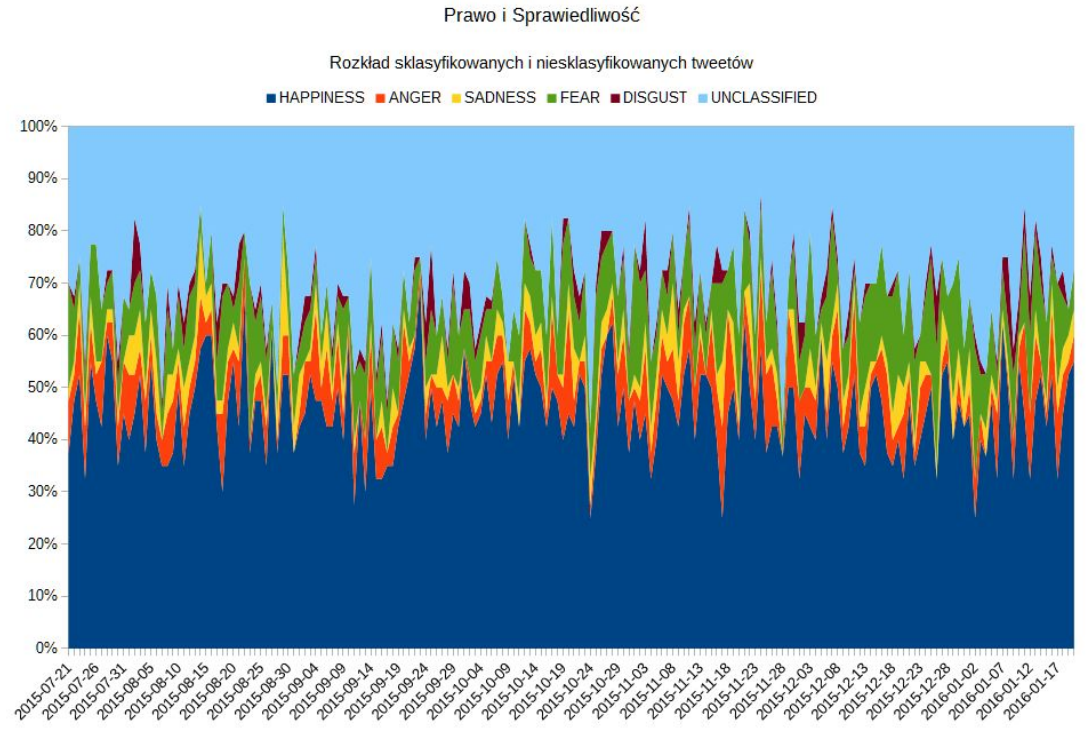
Opis eksperymentów

Dla każdej partii:

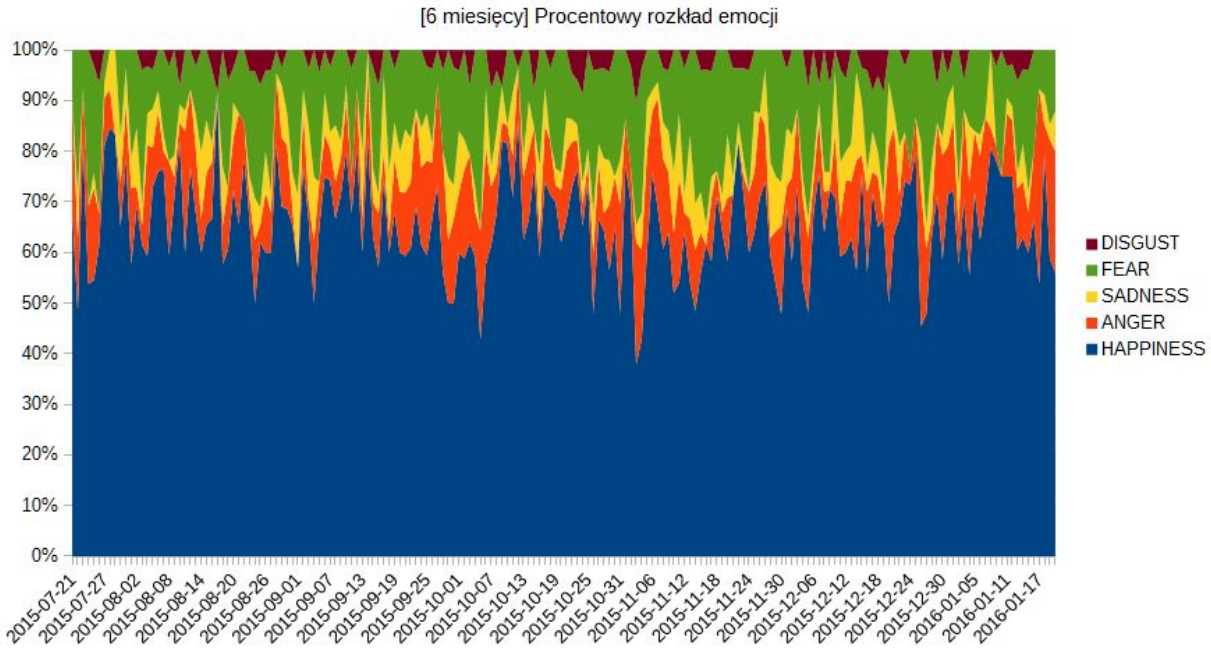
1. Zagregować dane z ostatnich 6 miesięcy.
2. Wygenerować CSV z zagregowanych tweetów
3. Wczytać bazę słów do pamięci
4. Zanalizować emocje każdego tweeta
5. Zsumować dane z wszystkich tweetów
6. Wygenerować CSV zawierający ilość tweetów z ilością każdej z emocji danego dnia
7. Wygenerować wykresy

Wyniki

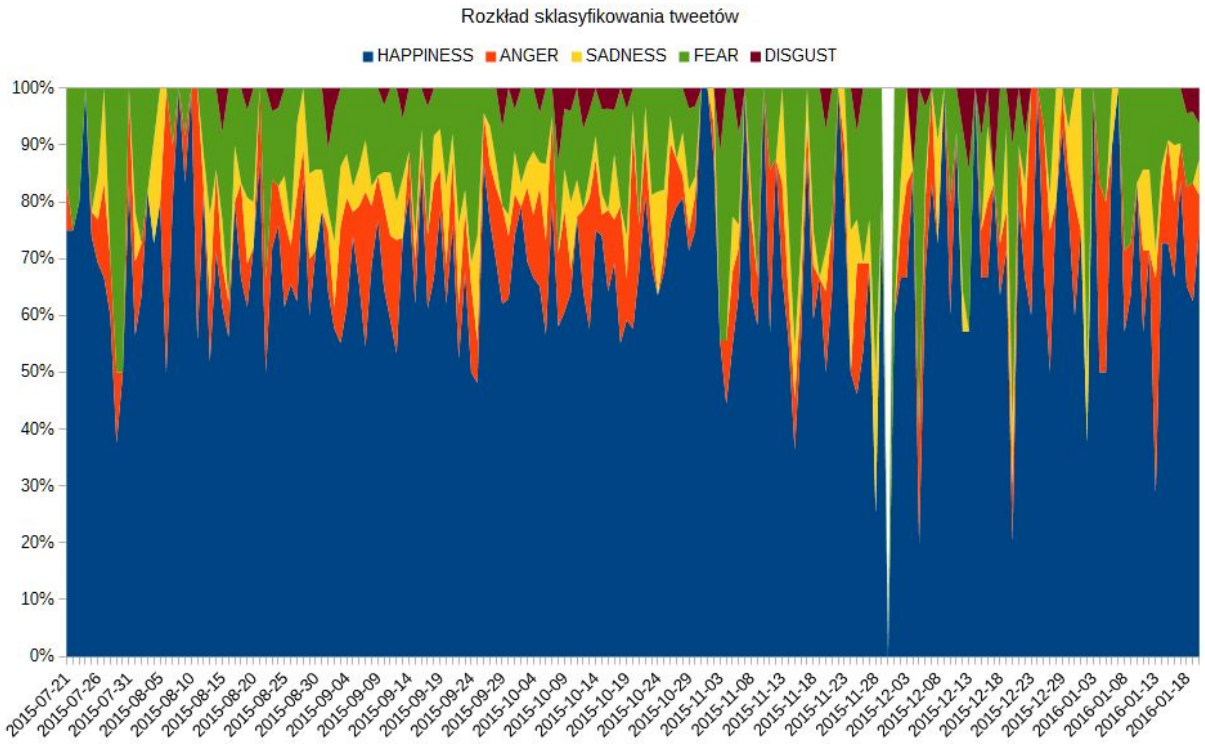

Prawo i Sprawiedliwość
Rozkład sklasyfikowanych i niesklasyfikowanych tweetów
HAPPINESS ANGER SADNESS FEAR DISGUST
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2015-07-21
2015-07-26
2015-07-31
2015-08-05
2015-08-10
2015-08-15
2015-08-20
2015-08-25
2015-08-30
2015-09-04
2015-09-09
2015-09-14
2015-09-19
2015-09-24
2015-09-29
2015-10-04
2015-10-09
2015-10-14
2015-10-19
2015-10-24
2015-10-29
2015-11-03
2015-11-08
2015-11-13
2015-11-18
2015-11-23
2015-11-28
2015-12-03
2015-12-08
2015-12-13
2015-12-18
2015-12-23
2015-12-28
2016-01-02
2016-01-07
2016-01-12
2016-01-17

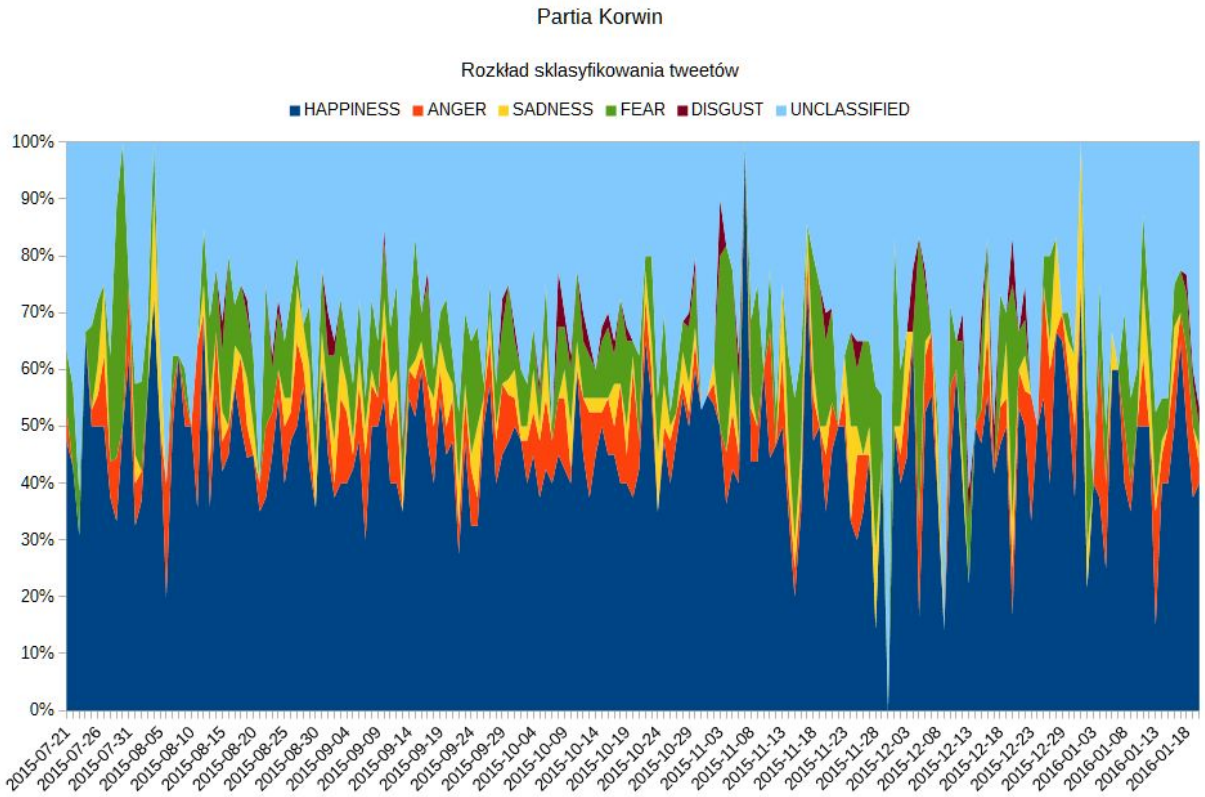

Prawo i Sprawiedliwość
Rozkład sklasyfikowanych i niesklasyfikowanych tweetów
HAPPINESS ANGER SADNESS FEAR DISGUST UNCLASSIFIED
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2015-07-21
2015-07-26
2015-07-31
2015-08-05
2015-08-10
2015-08-15
2015-08-20
2015-08-25
2015-08-30
2015-09-04
2015-09-09
2015-09-14
2015-09-19
2015-09-24
2015-09-29
2015-10-04
2015-10-09
2015-10-14
2015-10-19
2015-10-24
2015-10-29
2015-11-03
2015-11-08
2015-11-13
2015-11-18
2015-11-23
2015-11-28
2015-12-03
2015-12-08
2015-12-13
2015-12-18
2015-12-23
2015-12-28
2016-01-02
2016-01-07
2016-01-12
2016-01-17

Platforma obywatelska

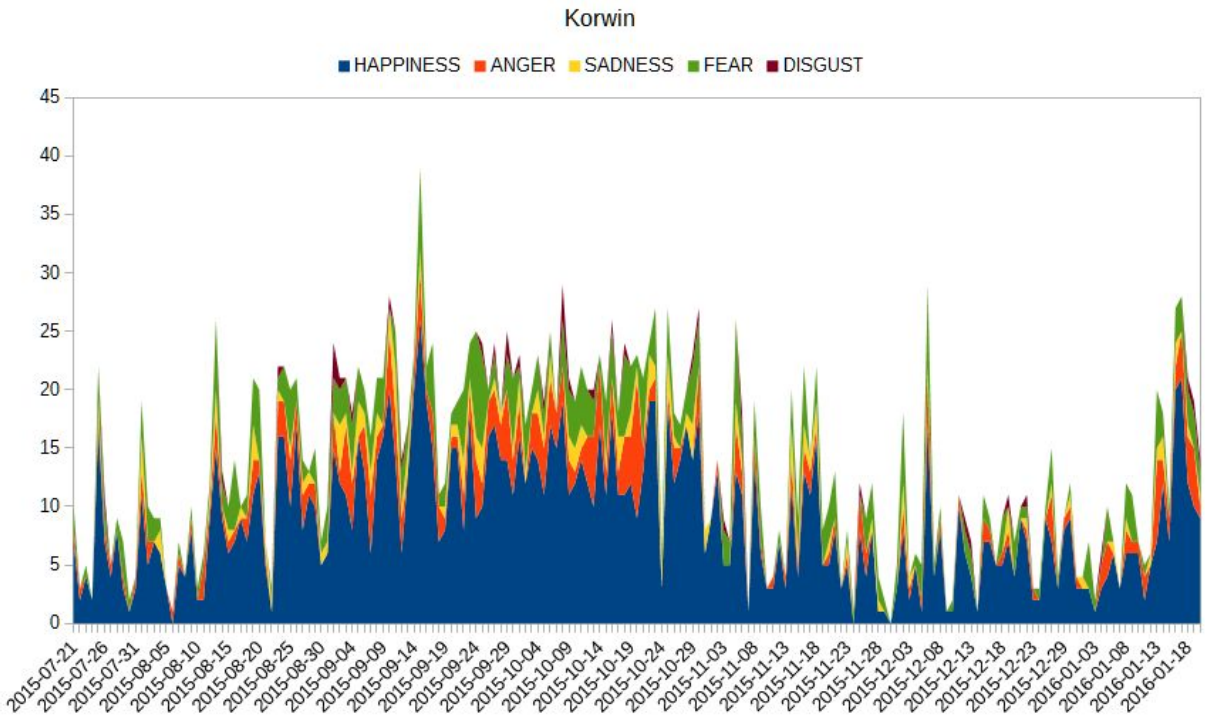


Partia Korwin

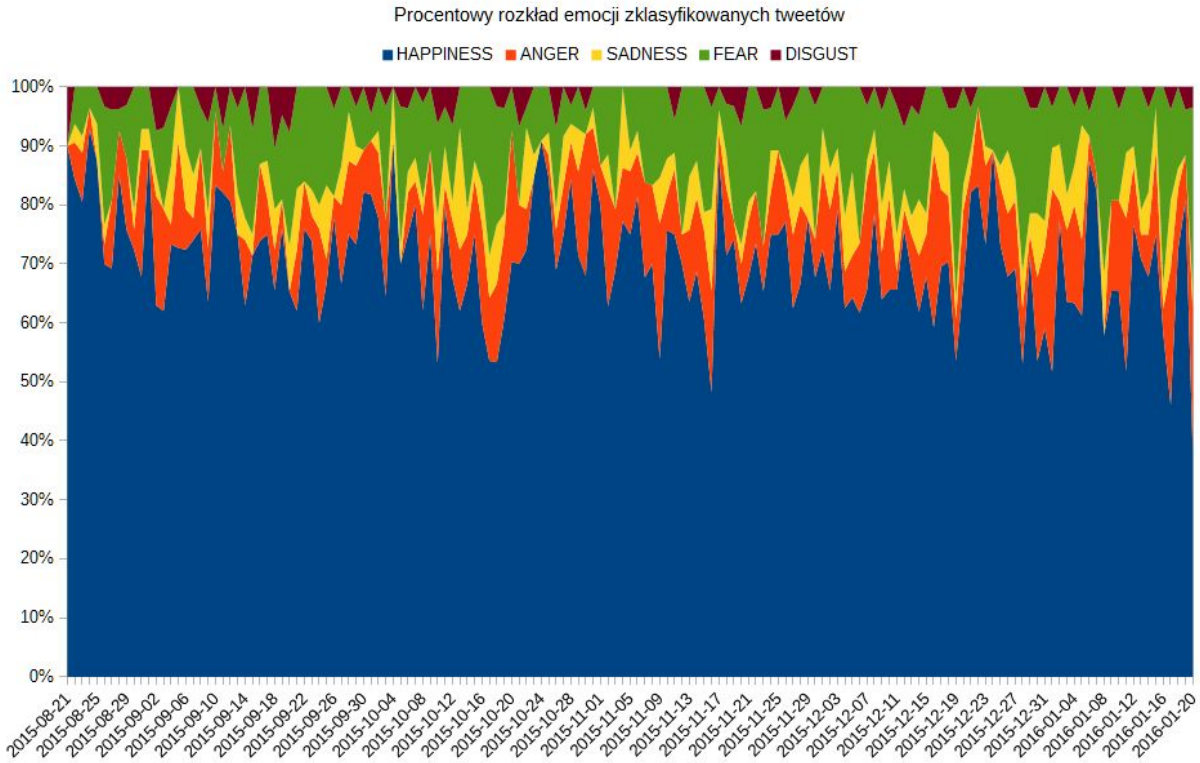




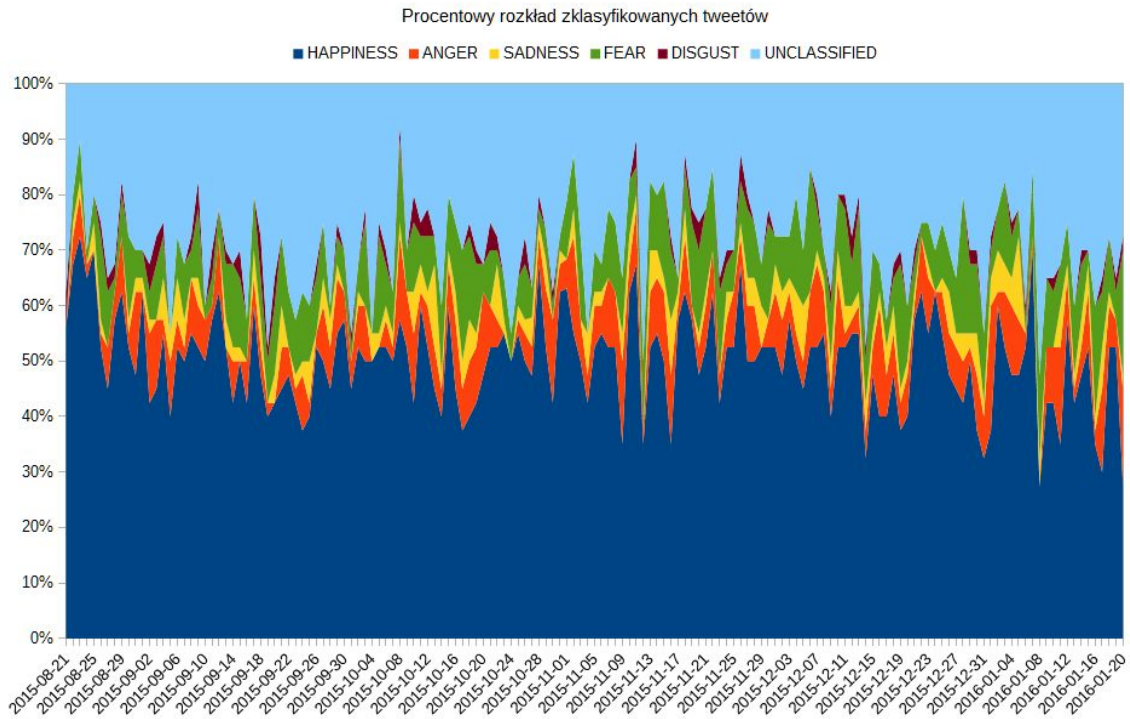
Przykład ilości zagregowanych tweetów w przeciągu okresu badania:



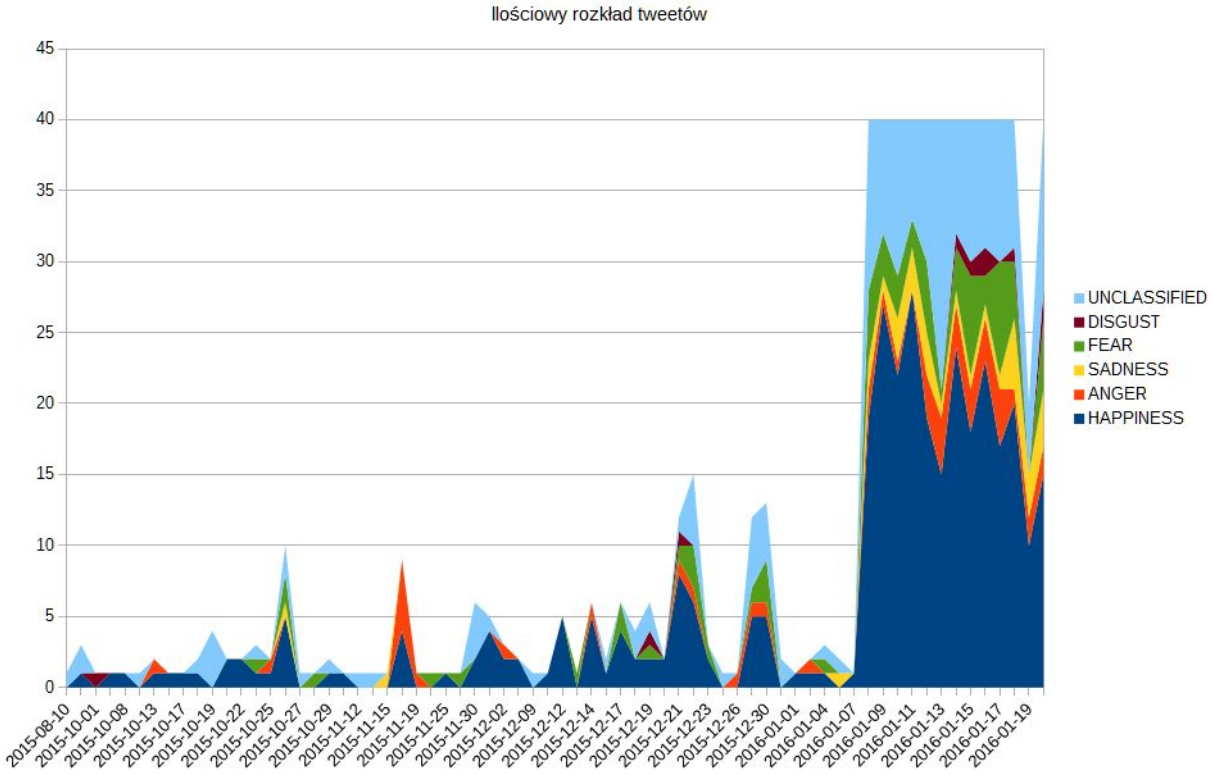
.Nowoczesna



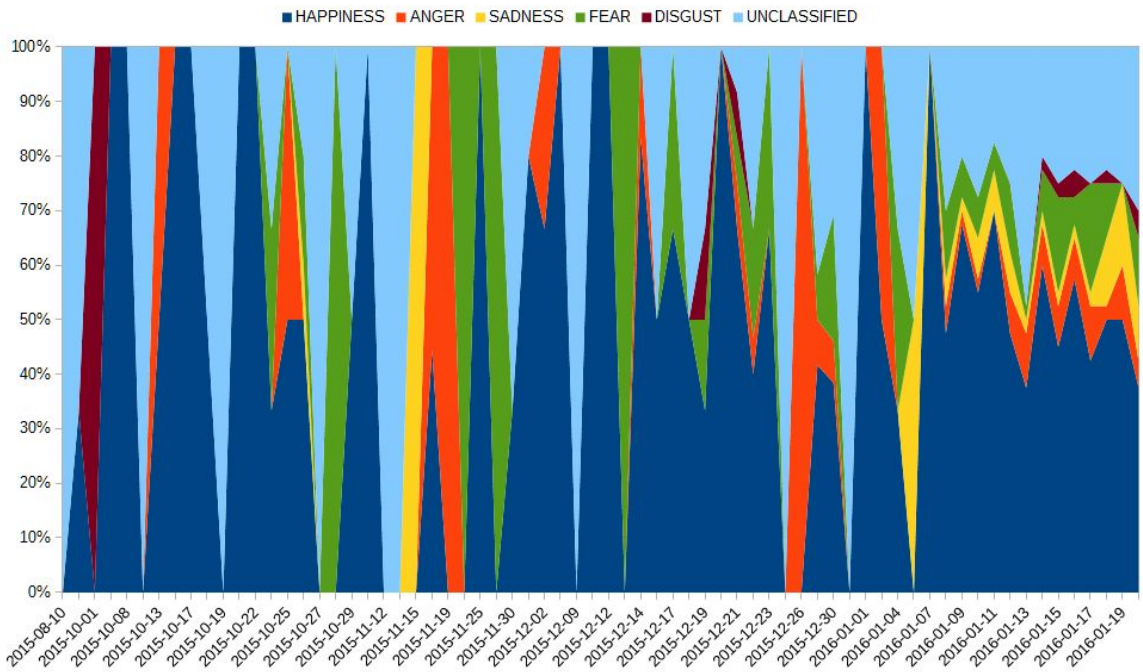
.Nowoczesna



Kukiz15



Kukiz '15



Interpretacja wyników

Partia	Tweetów	Sklasyfikowanych	Niesklasyfikowanych	% OK
Prawo i Sprawiedliwość	7295	4945	2350	67,8%
Platforma Obywatelska	6892	4647	2245	67,4%
.Nowoczesna	5933	4177	1756	70,4%
KORWIN	4601	3050	1551	66,3%
Kukiz ‘15	676	491	185	72,6%

Niższy wynik tweetów oznacza, że partia miała okresy niskiej aktywności użytkowników. np. Partia KORWIN miała bardzo niską aktywność w okresie wakacji. Natomiast partia .Nowoczesna założyła konto dopiero pod koniec sierpnia. Partia Kukiz ‘15 przez cały okres kampanii nie wzbudzała aktywności na twitterze.

Klasyfikacja emocji w tweetach w całym okresie badania:

Partia	Szczęście	Złość	Smutek	Strach	Wstręt	Neutral
Prawo i Sprawiedliwość	45,5%	6,5%	3,4%	10,2%	**2%**	32,2%
Platforma obywatelska	43,9%	**7,6%**	3,9%	**10,5%**	1,4%	32,6%
.Nowoczesna	**49,8%**	6,9%	3,2%	9,3%	1,3%	29,6%
KORWIN	44,4%	7%	3,9%	9,9%	0.9%	**33,7%**
Kukiz ‘15	**50,4%**	6,4%	**4,2%**	10%	**1,5%**	27,3%
Baza danych	*19,8%*	*13,2%*	*8,6%*	*22%*	*6,5%*	*29,6%*

Pogrubione zostały najwyższe wartości w kolumnie

Najczęstsze emocje:

1. Szczęście
2. Neutralny / Niesklasyfikowany
3. Strach
4. Złość
5. Smutek
6. Wstręt

Zaskakujące jest to, że blisko 50% tweetów każdej partii jest oznaczone jako pozytywne. Oraz wyjątkowo dużo jest tweetów sklasyfikowane jako strachliwe. Miesiąc wyborów nie odbiega znacząco od pozostałych miesięcy, nie można na podstawie tego wywnioskować zwycięzcy.

Uwagi

lość zagregowanych tweetów nie wpływała mocno na wyniki klasyfikacji.
Przykładem jest ilosć tweetów Kukiz ‘15.

Baza słów nie zawiera wulgaryzmów

Baza słów nie zawiera słów ze slangu

Algorytm nie jest odporny na sarkazm, ironię, szerszy kontekst wypowiedzi