Analiza opinii Tweetów dotyczących przeprowadzenia wyborów prezydenckich w maju

Magda Boruch, Filip Dratwiński, Maksym Telepchuk

Business/Scientific/Social Problem/Question

- Jak społeczeństwo polskie jest nastawione na pomysł wybór w maju?
- Jak można badać opinię publiczną związaną z wyborami w lepszy sposób niż na podstawie sondaży?
- Czy parlament polski podejmuje decyzje, które są zgodne z opinią publiczną?

Business/Scientific/Social Value

Analiza opinii związanych z wyborami pozwoli lepiej ocenić opinię społeczeństwa polskiego, co może pomóc podejmować decyzję przez parlament polski.

Team/Collaborations

- Osoba z doświadczeniem zbierania danych z mediów społecznościowych
- Osoba z wiedzą w zakresie przetwarzania języka naturalnego
- Osoba mająca wiedzę na temat uczenia nadzorowanego dla problemu klasyfikacji
- Osoba z doświadczeniem tworzenia aplikacji internetowych

Partners/Collaborations

- Osoby, które będą przypisywać tweety do jednej z klas (nie dla prototypu)
- Parlament polski
- Media informacyjne

Data

Dane pobierane z twittera po tagach - #wybory

Oznaczenie zebranych danych - czy tweet wyraża opinie za czy przeciw przeniesieniu wyborów lub nie wyraża żadnej.

Ocena współczynnika Kappa Cohena na podstawie oznaczanie części danych wspólnie

Model

FastText - wieloklasowa klasyfikacja, gdzie są trzy możliwe klasy - tweet nastawiony pozytywnie na wybory w maju, tweet nastawiony neutralnie (tweety informacyjne) i tweet nastawiony negatywnie

Evaluation

Badanie metryk odpowiednich dla zadania klasyfikacji w celu sprawdzenia dokładności modelu - F1 score, Accuracy

Deployment/UX

Strona internetowa połączona z bazą danych zawierającą tweety z opiniami. Strona powinna dawać możliwość dodawania nowych opinii - które będą odpowiednio klasyfikowane.

Users

- osoby odpowiedzialne za podejmowanie decyzji w kraju
- wymagają precyzyjnych i szczegółowych danych, muszą wiedzieć dlaczego podjęta została taka decyzja a nie inna
- społeczeństwo polskie

Expected Costs

- Hosting strony internetowej
- Sprzęt do wytrenowania modelu
- Baza danych, na której są przechowywane tweety

Expected Benefits

Narzędzie, które pozwala na ocenienie aktualnych emocji Polaków związanych z przeprowadzeniem wyborów w maju

Zdobycie zaufania w temacie analizy emocjonalnej mediów społecznościowych

Wykonane zadania w ramach utworzenia prototypu

Zebranie danych

Do zebrania danych użyto biblioteki **tweepy**.

Dzięki kontu developerskiemu na twitterze, mogliśmy zebrać dane z ostatnich **7 dni**.

Zebrano tweety, które zawierały przynajmniej jeden z hashtagów :

#wybory2020 #wybory #WyboryKorespondencyjne #WyboryKopertowe #pseudowybory #WyboryPrezydenckie2020

Przed przejściem do etykietowania usunięto z tweetów jedynie URL aby zachować jak najwięcej treści nacechowanej pozytywnie lub negatywnie względem tematu.

Oznaczanie danych

- Tweet'y były oznaczane na 4 różne kategorie:
  - 0 - przeciw przeprowadzeniu wyborów w maju, przykład : "Nie będę głosować. Niech się zdecydują, czy jest koronawirus, czy nie ma. Skoro mogą być otwarte lokalne wyborcze, to dlaczego nie restauracja?" #wybory2020 #ciechocinek"
  - 1 - za przeprowadzeniem wyborów w maju, przykład: "Uważam, że wybory powinny się odbyć jaknajszybciej. Nie można doprowadzić do kryzysu konstytucyjnego - @BeataKempa_MEP do @Marek_Pyza #wybory"
  - 2 - tweet informacyjny, przykład: "#Wybory 2020. Reakcje polityków opozycji na rekomendację ministra zdrowia #koronawirus"
  - N/D - Nie dotyczy tematu terminu wyborów lub ocena opinii jest niemożliwa na podstawie aktualnej wiedzy, przykład: "Oficjalnie: Przedłużenie kadencji władz związków sportowych.Głosowania elektroniczne w stowarzyszeniach #koronawirus #tarczaantykryzysowa #COVID19 #polskizwiązeksportowy #związeksportowy #klubsportowy #stowarzyszenie #kadencja #wybory #walne #prawosportowe"

Oznaczanie danych - ustalenia i ocena współczynnika Kappa Cohen’a

- Przed etykietowaniem niezależnym (każda osoba etykietuje inne Tweet’y) trzeba było przeprowadzić dwie wspólne “sesje”
- Pierwsza - każdy etykietuje te same 100 Tweet’ów
- Po etykietowaniu dyskusja na temat każdego Tweet’a i dyskutowanie, dlaczego tak zaetykietowaliśmy, a nie inaczej. Uporządkowanie wiedzy na temat wyborów
- Druga - każdy etykietuje te same 200 innych Tweet’ów
- Brak dyskusji po tej “sesji”. Celem tej sesji jest zbadanie zmiany współczynnika Kappa Cohen’a po ustaleniach
- Zmierzone zostały współczynniki Kappa Cohen’a na podstawie pierwszej i drugiej “sesji”

- 0 = agreement equivalent to chance.
- 0.1 – 0.20 = slight agreement.
- 0.21 – 0.40 = fair agreement.
- 0.41 – 0.60 = moderate agreement.
- 0.61 – 0.80 = substantial agreement.
- 0.81 – 0.99 = near perfect agreement
- 1 = perfect agreement.

Współczynnik Kappa Cohen’a - pierwsze etykietowanie	Współczynik Kappa Cohen’a - drugie etykietowanie
0.3426378785929643	0.5859641053876142

Przygotowanie danych do uczenia

W modelu zostały użyte tylko tweety oryginalne

Z każdego tweeta zostały usunięte

- znaki specjalne
- URL
- Nazwy użytkowników
- Emoji

Każdy tweet oraz etykieta do niego eksportowano do formatu, którego wymaga fastText

Proces uczenia modelu FastText

- Na podstawie oetykietowanych danych udało się uzyskać 1240 Tweet'y (po odrzuceniu Tweet'ów oznaczonych jako N/D) do uczenia modelu
- 369 z nich to Tweet'y pozytywne (za przeprowadzeniem wyborów w maju)
- 507 to Tweet'y negatywne (przeciwko przeprowadzeniu wyborów w maju)
- 364 to Tweet'y informacyjne
- Tweet'y zostały podzielone na zbiór treningowy i testowy z proporcją 80%/20%
- Wykorzystany został klasyfikator FastText do uczenia
- Przetestowane zostały maksymalne długości Word N-grams dla N=1,2,3 razem z różnymi początkowymi krokami uczenia - 1e-4, 1e-3, 1e-2, 1e-1
- Uczenie zostało przeprowadzone na 500 epokach i z wykorzystaniem wyuczonego word vector'a dla języka polskiego dostępnego tutaj: https://fasttext.cc/docs/en/crawl-vectors.html

Proces uczenia modelu FastText - wyniki

lr	wordNgrams	precision	recall	f-score
0.0001	1	0.494032258	0.494032258	0.494032
0.0001	2	0.4838709677419355	0.4838709677419355	0.483871
0.0001	3	0.47580645161290325	0.47580645161290325	0.475806
0.001	1	0.4959677419354839	0.4959677419354839	0.495968
0.001	**2**	**0.5**	**0.5**	**0.5**
0.001	3	0.4879032258064516	0.4879032258064516	0.487903
0.01	1	0.4838709677419355	0.4838709677419355	0.483871
0.01	**2**	**0.5**	**0.5**	**0.5**
0.01	3	0.4959677419354839	0.4959677419354839	0.495968
0.1	1	0.47580645161290325	0.47580645161290325	0.475806
0.1	2	0.47580645161290325	0.47580645161290325	0.475806

Wykorzystanie modelu do predykcji

- Najlepszy model został wyuczony na 1000 epokach
- 29.04.2020 zostało ściągniętych 1442 nowych Tweet'ów, które zostały oetykietowane przez wyuczony model
- Statystyki predykcji zostały ukazane na prototypie interfejsu

Aneta ObserwatorXY
@ObserwatorXY

Rząd z dnia na dzień zaskakuje branżę CH i jej najemców otwarciem na maja zaraz po wkndzie majowym mając w nosie czy to logistycznie i organizac możliwe bo Sasin pierwszy pocztowiec RP musi zrobić wybory Dudzie WyboryPrezydenckie2020 WyboryKopertowe WYBORY2020

7:20 pm · 29 Apr 2020 · Twitter Web App

20 Retweets **219** Likes

Uncle_Jun
@_uncle_jun

Debata Dnia Polsat News Kwietnia 5 NaprzódPolsko Bosak2020 PilnujmyPolski WYBORY2020 WyboryPrezydenckie2020 WyboryKorespondencyjne Konfederacja

6:29 pm · 29 Apr 2020 · Twitter Web App

30 Retweets **350** Likes

Lud w majtkach
@marcin_kulinski

Dr Artur Wróblewski w SygnałyDnia w Prawica ma wizję przyszłości Polski opozycja jej nie ma i z tego wynika jej słabe notowanie WyboryPrezydenckie2020

7:20 pm · 29 Apr 2020 · Twitter Web App

14 Retweets **100** Likes

Prototyp interfejsu

Podsumowanie

- Dyskusja na temat oznaczania była bardzo potrzebna
- Przy oznaczaniu przydatne by była pomoc politologów lub osób, które mają więcej wiedzy w kwestii oceny opinii
- Możliwą perspektywą jest przede wszystkim większy zbiór danych do uczenia