Politechnika Warszawska
Wydział Elektryczny

Statut Projektu v4.0

Autorzy:
Aleksei Haidukevich, nr albumu 295233
Jakub Korczakowski, nr albumu 291079
Marharyta Kruk, nr albumu 295235
Maciej Leszczyński, nr albumu 291085
Piotr Rosa, nr albumu 291112

4 listopada 2019

Spis treści

1 Opis projektu

1.1 Temat projektu

Stworzenie systemu do analizy sentymentów spółki na podstawie wpisów w mediach społecznościowych.

1.2 Opis problemu

Analiza sentymentów wiadomości jest stosunkowo nowym zagadnieniem, które staje się coraz bardziej popularne. Poprzez trafne określenie nacechowania emocjonalnego wiadomości klientów, spółki są w stanie zminimalizować ewentualne straty i zmaksymalizować zyski. Dzięki dokładnym badaniom firma wie, czego rzeczywiście chcą odbiorcy jej produktów/usług. Analiza sentymentów może być wykorzystywana w wielu rodzajach biznesów, a nawet w polityce (tworzenie sondaży). Jest ona obecnie stosowana przez przedsiębiorstwa takie jak Facebook (Facebook Insights), Google (Google Insights, Google Alerts), czy Hootsuite. Dodatkowym atutem przemawiającym za wykorzystywaniem jej jest dostępność dużej ilości darmowych danych w internecie, które po odpowiednim przetworzeniu są w stanie dostarczyć cenne informacje.

1.3 Cel projektu

Głównym celem projektu jest budowa systemu umożliwiającego analizę poziomu hejtu odnoszącego się do wybranych organizacji. Zbiór danych zostanie przez nas otrzymany z wpisów znajdujących się w mediach społecznościowych (Twitter, Reddit). Nasz system pozwoli badać, w jaki sposób formuje się sentyment wiadomości danej spółki. W przypadku nieplanowanego nagromadzenia się wiadomości o negatywnym sentymencie, będzie wysyłane ostrzeżenie, które może pozwolić na odpowiednio wczesną reakcję firmy.

2 Etap 1 - Przygotowanie środowiska i danych

czas: 4.11.2019 - 15.12.2019

2.1 Opis

W pierwszym etapie projektu najważniejszym celem będzie utworzenie narzędzi do zbierania danych z Twittera i Reddita. Zostaną one użyte w celu stworzenia zbioru danych pozwalającego na naukę i testowanie algorytmów. Pobrane wiadomości i nagłówki będą składowane w bazie danych. Ten etap prac zawiera również przygotowanie infrastruktury zdolnej analizować i przechowywać zebrane dane.

2.2 Pobieranie Danych

Opis zadania

Budowa programów pobierających dane z wybranych mediów społecznościowych (Twitter i Reddit).

Lista zadań pobocznych

1. Analiza dostępności API Twittera.
2. Analiza dostępności API Reddita.
3. Stworzenie programu pobierającego dane z Twittera.
4. Stworzenie programu pobierającego dane z Reddita.
5. Umożliwienie zapisu pobranych z Twittera danych do bazy danych.
6. Umożliwienie zapisu pobranych z Reddita danych do bazy danych.
7. Analiza dostępnych zbiorów opisanych klasami, pozwalających na testowanie algorytmu.

Kamień milowy

Utworzenie wyselekcjonowanego zbioru danych pozwalającego na naukę i testowanie algorytmów.

Parametry

Zbiór danych musi zawierać 250 tweetów, 150 nagłówków z Reddita zbieranych codziennie przez 30 dni dla 10 organizacji(razem 75000 tweetów i 45000 nagłówków).

Zbiór testowy musi zawierać 5000 tweetów i 3000 nagłówków, opisanych poprzez klasy pozwalające na sprawdzenie algorytmu.

Ryzyka

Ograniczona dostępność API serwisów społecznościowych.
Mitygacja: Automatyzacja zakładania kont dewloperskich, w celu ominięcia ograniczeń.

Brak zbioru pozwalającego na testowanie algorytmów analizujących sentyment.
Mitygacja: Ręczne opisanie zbioru testowego.

2.3 Budowa architektury

Opis zadania

Przygotowanie architektury systemu zdolnego składować i analizować pobrane wcześniej dane z mediów społecznościowych.

Lista zadań pobocznych

1. Analiza i wybór sposobu wdrożenia aplikacji (rozwiązania chmurowe).
2. Analiza dostępnej infrasturktury do przetwarzania danych.
3. Wybór odpowiedniej bazy do składowanych danych (porównanie SQL i noSQL).
4. Instalacja wybranej bazy.
5. Połączenie bazy danych z programami pobierającymi dane.
6. Pobranie danych do bazy danych.

Kamień milowy
Utworzenie systemu pozwalającego na przetwarzanie i analizowanie pobranych danych.

Ryzyka
Ograniczona dostępność do chmury (Azure).
Mitygacja: Utworznie kolejnych kont w usłudze.

2.4 Przygotowanie do implementacji algorytmów

Opis zadania
Analiza dostępnych algorytmów do analizy sentymentu oraz wykrywania anomalii i ocena ich przydatności w naszym projekcie.

Lista zadań pobocznych

1. Analiza i porównanie dostępnych algorytmów uczenia nienadzorowanego do analizy sentymentu.
2. Analiza i porównanie dostępnych algorytmów uczenia nadzorowanego do analizy sentymentu.
3. Analiza i porównanie dostępnych algorytmów wykrywania anomalii.
4. Analiza wymagań, jakie musi spełniać zbiór danych przeznaczony do uczenia i testowania algorytmu.
5. Analiza metod przetwarzania tekstu.

Kamień milowy
Zdobycie wiedzy i opis dostępnych algorytmów.

3 Etap 2 - Budowa algorytmów

czas: 15.12.2019 - 30.03.2020

3.1 Opis

W drugim etapie projektu najważniejszym celem będzie stworzenie oraz uczenie modelów do analizy sentymentów. Modele dla Twittera i Reddita będą w stanie określić sentyment wiadomości (dobry - zły - neutralny). Po stworzeniu tych modelów głównym celem będzie dokonanie analizy zebranych wiadomości dla kilku określonych haseł. W tym celu będzie stworzony jeszcze jeden model, przeznaczony do analizy rozkładów sentymentów i wykrywania anomalii. Ten etap jest również przeznaczony do stworzenia infrastruktury, która będzie w stanie ciągle analizować nowe dane.

3.2 Zadania

1. Stworzenie oraz uczenie modelu na podstawie już znalezionych zbiorów do analizy sentymentów Twittera.
2. Stworzenie oraz uczenie modelu na podstawie już znalezionych zbiorów do analizy sentymentów Reddita.
3. Przeprowadzenie testów modelów na samodzielnie stworzonych zbiorach testowych.
4. Dokonanie analizy danych, zebranych w wyniku działania programów, stworzonych w poprzednim etapie, za pomocą modelu do analizy sentymentów (p.3-4)
5. Dokonanie analizy występowania anomalii. Stworzenie zbioru testowego z występującą anomalią.
6. Stworzenie modelu, który będzie w stanie na podstawie dokonanej wcześnie analizy (p.6) wykrywać anomalie w rozkładzie negatywnych wiadomości. Testowanie tego modelu na utworzonym wcześniej (p.7) zbiorze.

3.3 Kamienie milowe

1. Stworzenie modelów do określenia sentymentów wiadomości Twittera oraz Reddita
2. Stworzenie modelów do analizy określonych sentymentów w celu wykrywania anomalii

3.4 Parametry

1. Modele muszą określać sentymenty wiadomości z precyzją co najmniej 85%.

2. Modele do wykrywania anomalii muszą mieć precyzję co najmniej 70% przy sprawdzeniu zbiorów testowych

3. Zbiory testowe, zawierające w sobie wiadomości oraz sentymenty, stworzone sztucznie w taki sposób, aby zawierały w sobie anomalie — przeznaczony do testowania modelu wykrywania anomalii.

3.5 Ryzyka

1. Brak wystarczającej ilości danych do nauczenia modelu wykrywania anomalii.

2. Brak możliwości dokładnej analizy wiadomości pobranej z Reddita.

4 Etap 3 - Budowa strony internetowej i wizualizacja

czas: 30.03.2020 - 27.04.2020

4.1 Opis

W trzecim etapie projektu celem jest stworzenie medium pomiędzy wynikiem działania algorytmów i użytkownikiem. Medium ma postać serwisu internetowego i służy do wizualizacji zebranych danych oraz wyników ich analizy. Serwis jest połączony z bazą danych, wytworzoną w poprzednich etapach.

4.2 Zadania

1. Wybranie technologii najbardziej pasującej do tworzenia stron webowych (np. Django, Flask).

2. Budowa struktury serwisu internetowego.

3. Stworzenie modułu wizualizacji na podstawie istniejących rozwiązań (np. Prophet library).

4. Integracja modułu wizualizacji do serwisu internetowego.

5. Połączenie modułu wizualizacji z bazą danych.

6. Automatyzacja odświeżania wyników wizualizacji na podstawie aktualnych danych z bazy.

4.3 Kamienie milowe

1. Stworzenie szkieletu serwisu internetowego
2. Reprezentacja danych w postaci wykresów na stronach serwisu internetowego. Wykresy powinny w sposób jawny wyróżniać anomalie w danych.

4.4 Parametry

1. Serwis internetowy składa się z przynajmniej 3 stron.
2. Wykresy przedstawiają dane z okresu 30 dni.

4.5 Ryzyka

1. Problem połączenia serwisu internetowego z bazą danych.
   Mitygacja: Użycie rozwiązań istniejących, w których back-end i baza danych są połączone w sposób automatyczny.