ZZSN Dokumentacja Końcowa
Projekt 12. Analiza wydźwięku depesz giełdowych za pomocą modelu typu BERT.

Dawid Brzozowski
Robert Ostoja-Lniski

8 czerwca 2021

Spis treści

1 Opis zadania

Celem projektu jest wykorzystanie polskiego modelu typu BERT (np. Herbert albo Polbert) do oceny wydźwięku (negatywny/neutralny/pozytywny) depesz na polskiej Giełdzie Papierów Wartościowych.

2 Pobieranie danych finansowych

Pierwszym etapem rozwiązania problemu jest pobranie danych finansowych z systemu ESPI, na którym znajdują się treści depesz giełdowych oraz z platformy Quandl, która udostępnia API do pobierania kursów akcji spółek. Stworzyliśmy skrypty, które zostały uruchomione na platformie Google Cloud Platform w celu pozyskania danych.

2.1 Pobieranie raportów ESPI

Na podstawie udostępnionych depesz na platformie ESPI zaimplementowaliśmy CLI służące do pobierania raportów dla dowolnej spółki z zadanego zbioru (zbiór spółek notowanych na Warszawskiej Giełdzie Papierów Wartościowych) z zadanego okresu. Na wejściu podana jest nazwa spółki i przedział czasowy, a zwrócony plik zawierający treści depesz dla tej spółki dla każdego dnia z zadanego przedziału (o ile w tym dniu opublikowano raport). Plik jest w formacie JSON. Wykorzystaliśmy w celu pobierania udostępnionych raportów moduł requests oraz bs4. Aby pobrać depesze, należało przeparsować strony o różnym formacie w celu wydobycia istotnych dla nas informacji (nazwa spółki, treść depeszy, data depeszy).

2.2 Pobieranie raportów Quandl

Dla poszczególnych wierszy uzyskanych z systemu ESPI pobrane zostaną ceny kursów akcji spółek z dnia roboczego poprzedzającego oraz następującego po dniu udostępnienia depeszy. Następnie zmiana kursu akcji została porównana do zmiany kursu indeksu WIG aby ustalić wydźwięk dla depeszy względem szerokiego rynku. Porównanie to zostało szczegółowo opisane w podpunkcie 4.1.

3 Przetwarzanie pobranych danych

3.1 Uzyskiwanie informacji z raportu ESPI

Otrzymany raport jest w formacie html. Większość danych w tym raporcie to informacje nadmiarowe, nieistnone względem rozwiązywanego problemu. Z raportu wyznaczamy jedynie datę utworzenia, nazwę spółki oraz treśc depeszy, a wyniki zapisujemy do pliku. Z treści depesz przechowywanych w zbiorze usunięte są nazwy spółek. Raporty mają jednakową budowę dzięki czemu te dane

znajdują się w miejscach oznaczonych przez odpowiednie tagi. Podsumowując, nasze API do przetwarzania raportów na wejściu otrzymuje raport, a zwraca plik z kluczowyi informacjami.

3.2 Wyzanczenie sentymentu za pomocą Quandl

Dane zdobyte przez biliotekę Quandl (cena akcji spółki oraz WIG) służą do określenia liczbowej wartości sentymentu. W punkcie 4.1 przedstawiony jest stworzony w tym celu wzór. Wyliczony sentyment dołączamy do każdego przetworzonego raportu.

4 Ustalanie wydźwięku dla pobranych danych

Wydźwięk danej depeszy jest określany początkowo jako liczba rzeczywista z przedziału od -1 do 1, a nastepnie przyporządkowywana jest jemu jedna z trzech etykiet zgodnie z parametrami wejściowymi do do trenowania modelu.

4.1 Wartość liczbowa sentymentu

Do ustalenia wartości wydźwięku posłużył nam wzór.

$$m = \frac{a2 - a1}{a1} - \frac{w2 - w1}{w1}$$

- a1 jest ceną akcji spółki podczas poprzedniego zamknięcia
- a2 jest ceną akcji spółki podczas zamknięcia kolejnego notowania po wydaniu depeszy
- w1 jest ceną indeksu WIG podczas poprzedniego zamknięcia
- w2 jest ceną indeksu WIG podczas zamknięcia kolejnego notowania po wydaniu depeszy

4.2 Przyporządkowanie etykiety dla sentymentu

Podczas eksperymentów wyznaczaliśmy dokładne wartości parametrów p i n (p $>$ n), takich, że

- dla wydźwięku pozytywnego (etykieta positive)

$$m \geq p$$

- dla wydźwięku neutralnego (etykieta negative)

$$n \leq m < p$$

- dla wydźwięku negatywnego (etykieta neutral)

$$m < n$$

Eksperymentalnie wyznaczyliśmy wartości wartości:

- p = 0.03
- n = -0.07

Wartości te wynikają z naszych własnych badań przeprowadzonych na 200 próbkach. Celem badania było wyznaczenie optymalnej wartości progów. Wyniki są przybliżone do części setnych. Początkowo przyjmowaliśmy stałą wartość parametru p, a modyfikowaliśmy wartość n. W kolejnym kroku dla ustalonego n optymalizowaliśmy p.

Chcielibyśmy także zwrócić uwagę na wyraźną różnicę między wartościami bezwzględnymi obydwu parametrów. Naszym zdaniem, oddaje to sytuację, w której inwestorzy są bardziej skłonni wycofać kapitał, jeśli spółka odnotowuje stratę niż zakupić akcje w przypadku pozytywnych informacji.

5 Model BERT

5.1 Typ modelu

Za pomocą biblioteki transformers wykorzystaliśmy model HerBERT (base) ze względu na wysokie wyniki uzyskane w benchmarkach Klej [1]. Wybrana została wersja bazowa ze względu na krótszy czas treningu w porównaniu do modelu w wersji large.

5.2 Trenowanie na zbiorze PolElmo

Model ten został wytrenowany na zbiorze PolElmo oraz zmierzona została jego jakość na zbiorze testowym z tej samej dziedziny jak na zmienionej (odpowiednio PolElmo2.0-IN, PolElmo2.0-OUT). Z uwagi na stosunkowo dużą złożoność trenowania modelu posłużyliśmy się platformą Google Colab, na której trenowaliśmy model przy użyciu GPU. Do wytrenowania modeli użyliśmy biblioteki PyTorch wraz z Trainer API pochodzącym z transfomers. Należy zauważyć, że ze względu na chęć uzyskania tej samej liczby klas co dla zbioru danych finansowych, zdecydowaliśmy się odrzucić dane z kategorii AMB. W ten sposób, uzyskaliśmy zbiory 3-klasowe dla klas: "positive", ńegative", ńeutral".

5.3 Trenowanie na zbiorze depesz

W kolejnym etapie projektu wytrenowaliśmy model na danych zdobytych z platformy ESPI w celu porównania wyników z referencyjnym modelem trenowanym na PolElmo2.0. Wielkośc zbioru ESPI liczyła kolejno:

- ok. 1800 dla zbioru trenującego.
- ok 250 dla zbioru walidacyjnego.
- ok 500 dla zbioru testowego.

. Podzieliliśmy ten zbiór na trenujący, testowy oraz walidacyjny w stosunku 7:2:1. Trenowanie również miało miejsce na platformie Google Colab.

5.3.1 Podział danych na zbiory: testowy, treningowy i walidacyjny

Dzielając zbiór depesz giełdowych na zbiory: testowy, treningowy i walidacyjny mogło dojść do sytuacji, w której ta sama spółka znalazła się w zbiorze testowym i treningowych. Jeśli występowałby pewien związek między wszystkimi depeszami w ramach jednej spółki, to mogłoby to zaburzyć wyniki modelu. Aby rozwiązań ten potencjalny problem wprowdziliśmy parametr *mixed*. Jest to flaga, która tylko po zapaleniu umożliwia dodanie danych z tej samej spółki do różnych zbiorów. Przeprowadziliśmy eksperymenty, które porównują działanie modelu względem tego parametru. Wyniki badań dostępne są w sekcji 8.

6 Przeprowadzone eksperymenty

6.1 Eksperymenty dla modelu trenowanego na zbiorze Klej In

Podczas przeprowadzonych badań zauważono, że model wytrenowany na zbiorze Klej, choć działa bardzo dobrze zarówno na zbiorze testowym Klej In jaki i Klej Out - zupełnie nie radzi sobie z analizą sentymentu depesz giełdowych. Dla tego modelu zawsze zwracana jest klasa neutralna, co jest zrozumiałe, ponieważ nie występują w depeszach słowa typowo związane z negatywnym lub pozytywnym sentymentem w kontekście zbioru Klej. Wyniki dla tego modelu zostały umieszczone w sekcji 8.

6.2 Eksperymenty dla modeli trenowanych na danych finansowych

Wyniki dla modeli trenowanych na danych finansowych również nie są na poziomie zadowalającym. Dane finansowe układają się w ten sposób, że zdecydowanie przeważającą klasą jest ńeutral". Wynika to z faktu, że większość depesz nie ma większego wpływu na zmianę kursu akcji spółek. Z tego powodu, model trenowany na oryginalnym podziale nauczył się przewidywać tylko i wyłącznie klasę neutralną. Chodź accuracy było w tym momencie wysokie, jest to działanie bardzo nieporządane. Z tego powodu, zdecydowaliśmy się na modyfikację architektury modelu oraz próby manipulowania zbiorem treningowym w celu zmniejszenia różnic między liczbą próbek z danych klas.

- Sprawdzono jak model zachowa się, kiedy zastosujemy undersampling na zbiorze treningowym w celu wyrównania liczby próbek z danych klas. Wyniki dla takiego eksperymentu są dostępne w sekcji 8.
- Sprawdzono jak model zachowa się, kiedy zasosujemy undersampling na zbiorze treningowym, kiedy ograniczymy najliczniejszą klasę do dwukrotności najmniej licznej klasy. Wyniki dla tego eksperymentu nie były zadowalające i nie zdecydowaliśmy się ich umieścić w sprawozdaniu.
- Sprawdzono jak model zachowa się, kiedy funkcja straty dodatkowo weźmie pod uwagę liczby próbek dla danych klas. Wykorzystano do tego celu funkcję compute_class_weight z pakietu scikit-learn. Wyniki również nie były zadowalające i nie zostały umieszczone w sprawozdaniu.

7 Wybrane zbiory danych

W projekcie skorzystaliśmy ze zbioru depesz opublikowanych na portalu Infosfera [2]. Zawiera on raporty dla wybranych spółek (od października 2004 roku do bieżącego kwietnia 2021 roku). Wartość cen akcji spółki w danym dniu, oraz wysokość notowania WIG (w celu zbadania korelacji między szerokim rynkiem) została pobrana na dzięki API udostępnionemu przez platformę Quandl. Zbiór PolElmo2.0 (zarówno IN jak i OUT) został pobrany za pomocą linku w treści zadania.

8 Zrealizowana funkcjonalność

Porównano działanie modelu wytrenowanego na danych ze zbioru Klej IN dla zadania analizy sentymentu depesz giełdowych. Okazuje się, że zadanie to jest bardzo trudne, a ocena sentymentu dla depesz wymaga dużej wiedzy dziedzinowej, która nie jest zawarta w tekście. Okazuje się, że modele nauczone na klasycznych danych do analizy sentymentu zawsze wskazują klasę neutralną. Model po procesie uczenia na danych finansowych potrafi wskazać sentyment (pozytywny/neutralny/negatywny) dla treści depeszy giełdowej.

9 Wyniki

9.1 Ewaluacja na zbiorach PolElmo

9.1.1 Ewaluacja na PolElmo2.0-In

W Tablicy 1 przedstawione są wyniki ewaluacji na zbiorze testowym PolElmo2.0-In dla modelu trenowanego na PolElmo2.0-In. Każdy rodzaj etykiety został przyporządkowany poprawnie dla praktycznie wszystkich danych.

Label	Precision	Recall	F1-score	support
positive	0.98	0.98	0.98	209
negative	0.98	0.98	0.98	271
neutral	1.00	1.00	1.00	127
macro avg	0.99	0.99	0.99	330
weighted avg	0.99	0.99	0.99	330

Tablica 1: Na zbiorze PolElmo uzyskaliśmy bardzo wysokie wyniki. Wszystkie parametry, w szczególności miara F, są bliskie 1.00.

9.1.2 Ewaluacja na PolElmo2.0-Out

Tablica 2 przedstawia wyniki ewaluacji modelu trenowanego na PolElmo2.0-In na zbiorze testowym PolElmo2.0-Out. Etykieta neutral występowała tylko jednokrotnie w całym zbiorze. Dla tej etykiety precyzja na poziomie 0.06 oraz zupełność na poziomie 1.00 oznacza, że model poprawnie przyporządkował etykietę neutral, ale także oznaczył nią kilkanaście innych próbek.

Label	Precision	Recall	F1-score	support
positive	0.97	0.92	0.94	164
negative	0.96	0.90	0.93	165
neutral	0.06	1.00	0.11	1
macro avg	0.66	0.94	0.66	330
weighted avg	0.96	0.91	0.93	330

Tablica 2: Wyniki ewaluacji na zbiorze PolElmo2.0-Out. Model uzyskał dobre wyniki dla etykiety positive i negative, ale wykrywał neutral dla zbyt dużej liczby danych.

9.2 Ewaluacja na danych finansowych

9.2.1 Model trenowany na zbiorze PolElmo

Tablica 3 przedstawia wyniki ewaluacji modelu trenowanego na PolElmo2-0.In na zbiorze utworzonym z danych finansowych.

Rodzaj danych	Label	Precision	Recall	F1-score	support
finansowe *mixed*	positive	0.00	0.00	0.00	33
	negative	0.00	0.00	0.00	12
	neutral	0.83	1.00	0.91	218
	macro avg	0.26	0.33	0.29	263
	weighted avg	0.61	0.78	0.69	263
finansowe	positive	0.00	0.00	0.00	39
	negative	0.00	0.00	0.00	19
	neutral	0.78	1.00	0.88	210
	macro avg	0.26	0.33	0.29	268
	weighted avg	0.61	0.78	0.69	268

Tablica 3: Wyniki ewaluacji na zbiorze testowym danych finansowych. Model przewiduje tylko i wyłącznie klasę neutralną.

9.2.2 Model trenowany na zbiorze złożonym z danych finansowych

Tablica 4 przedstawia wyniki ewaluacji modelu trenowanego na danych finansowych.

Rodzaj danych	Label	Precision	Recall	F1-score	support
finansowe *mixed*	positive	0.00	0.00	0.00	25
	negative	0.00	0.00	0.00	23
	neutral	0.82	1.00	0.90	215
	macro avg	0.27	0.33	0.30	263
	weighted avg	0.67	0.82	0.74	263
finansowe	positive	0.00	0.00	0.00	39
	negative	0.00	0.00	0.00	19
	neutral	1.00	0.03	0.06	210
	macro avg	0.58	0.34	0.30	268
	weighted avg	0.69	0.75	0.65	268

Tablica 4: Wyniki ewaluacji na zbiorze testowym danych finansowych. Model przewiduje tylko i wyłącznie klasę neutralną.

Spróbowano również wytrenować model stosując undersampling na zbiorze treningowym uzyskując tym samym zbalansowany zbiór treningowy. Nie dało to jednak oczekiwanych efektów i wyniki nadal pozostają niskie.

Label	Precision	Recall	F1-score	support
positive	0.18	0.41	0.25	39
negative	0.06	0.58	0.11	19
neutral	1.00	0.00	0.01	210
macro avg	0.41	0.33	0.12	268
weighted avg	0.81	0.10	0.05	268

Tablica 5: Wyniki ewaluacji na zbiorze testowym danych finansowych. Model trenowany był na zbalansowanym zbiorze.

10 Wnioski

10.1 Sentyment w kontekście analizy fundamentalnej

Naszym zdaniem, próba predykcji wydźwięku depesz giełdowych jest zadaniem, do którego niezbędna jest odpowiednia wiedza dziedzinowa. Na podstawie samej treści depeszy bez wiedzy z zakresu analizy fundamentalnej praktycznie niemożliwe jest dokładne oszacowanie wydźwięku. Przykładowo jeśli w depeszy wykazono wzrost zarobków danej firmy o kilkanaście proces rok do roku, to ta informacja jedynie pozornie może podnieść ceny akcji tej spółki. Jeśli analitycy biznesowi uznali, że spodziewany wzrost cen akcji jest rzędu kilkudziesięciu procent, to znacząca część inwestorów przeniesie swój kapitał do innych instrumentów finansowych, co spowoduje spadek cen akcji badanej firmy. Innymi słowy, wzrost zarobków danej firmy jest warunkiem koniecznym, ale niewytarczającym do wyznaczenia wydźwięku.

10.2 Analiza depesz dotyczących zmian struktury spółki

Podczas początkowych etapów realizowania projektu, natrafiliśmy na depesze giełdowe, które posiadały sentyment wskazujący na zmianę cen akcji (inny niż neutralny, zarówno pozytywny jak i negatywny). Sama treść depeszy nie zawierała informacji związanych z zarobkami firmy, ale przedstawiała informacje o zmianach w zarządzie albo restrukturyzacji spółki. Bez wiedzy wychodzącej poza raporty ESPI dotyczącej możliwych skutków takich działań, wyciągnięcie wniosku dotyczącego zmiany sentymentu na podstawie samej treści raportu było praktycznie niemożliwe do wykonania.

10.3 Pobieranie treści depesz

Podejrzewamy, że po stronie serwera, który udostępniał depesze giełdowe znajdował się system przeciwdziałający atakom typu DoS. Podczas wysyłania zapytań ze stacji roboczej po wysłaniu około 100 otrzymywaliśmy błąd serwera świadczący o blokadzie kolejnych zgłoszeń. Blokada mijała po około kilkudziesięciu minutach. Aby rozwiązać ten problem stworzyliśmy skrypt, który okresowo co kilkanaście sekund wysyłał zgłoszenia. Uruchomiony był na Google Cloud

Platform przez kilka dni, w które zebrał dane wystarczające do wytrenowania modelu.

10.4 Wyznaczanie głównie sentymentu neutralnego

Z uwagi na brak dostarczenia wiedzy dziedzinowej (opisane w punktach 10.1 oraz 10.2) model wytrenowany zarówno na zbiorach finansowych, jak i PolElmo, przyporządkowywał zdecydowanej większości danych etykietę *neutral*. Oznacza to, że treść depesz nie wskazywała jednoznacznie na silną zmianę zainteresowania daną spółką. Jest to dowód na to, że same depesze bez dodatkowych informacji dotyczących zachowania rynku, kondycji spółki, spodziewanych rezultatów finansowych czy planowanych inwestycji nie są wystarczające do przewidzenia ich wpływu na cenę akcji.

11 Narzędzia

- Język programowania: Python 3.8.
- Instalacja pakietów za pomocą narzędzia Anaconda [3].
- Biblioteki oraz inne narzędzia wykorzystane w projekcie:
  - requests [4] - w celu pozyskiwania stron html zawierających depesze giełdowe.
  - BeautifulSoup4 [5] - w celu przetwarzania stron html zawierających treść depesz giełdowych wraz z nazwą spółki oraz datą opublikowania.
  - Quandl [6] - do pobierania kursów akcji spółek.
  - tqdm [11]- do wyświetlania paska progresu.
  - PyTorch oraz transformers [7] - do stworzenia modelu opartego o BERT dla zadania analizy sentymentu.
  - Scikit-learn [8] - do operacji na zbiorach danych (podział na zbiór trenujący, walidujący i testowy itp.).
  - Google Cloud Platform [9] do pobierania danych za pomocą wielu maszyn
  - Google Colab [10] do trenowania modelu na wydajnej maszynie z GPU.

12 Literatura

Literatura

[1] https://klejbenchmark.com/leaderboard/

[2] http://infostrefa.com/infostrefa/pl/raporty/espi/firmy/

[3] https://anaconda.org

[4] https://docs.python-requests.org/en/master/

[5] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[6] https://www.quandl.com/tools/python

[7] https://pytorch.org

[8] https://matplotlib.org

[9] https://cloud.google.com

[10] https://colab.research.google.com/?utm_source=scs-index

[11] https://github.com/tqdm/tqdm