

NLP dokumentacja końcowa

Kamil Dąbrowski, Michał Kędzierski, Maksymilian Szymanowicz, Tomasz Ziemnicki

15.06.2021

Spis treści

1. Opis zadania	2
2. Celowość projektu	2
3. Synteza problemów z treści opisu zadania	2
4. Osiągnięta funkcjonalność	2
5. Sposoby rozwiązania problemów	3
6. Topologia modelu	3
7. Narzędzia	4
8. Zbiory danych	4
9. Trening i parametry sieci	5
10. Rezultaty	5
11. Aplikacja StockRadar	8

1. Opis zadania

Celem projektu było stworzenie modelu, który na podstawie aktualnych informacji o wybranej spółce dokona oceny towarzyszącego jej sentymentu (pozytywny/neutralny/negatywny). Informacja ta może zostać uwzględniona w strategii giełdowej inwestora, może być też niezwykle przydatna dla akcjonariuszy oraz kierownictwa przedsiębiorstwa tworząc predykcje na temat przyszłej wyceny akcji.

2. Celowość projektu

Motywacją do stworzenia narzędzia działającego w sposób przedstawiony w niniejszym dokumencie są dwa czynniki:

Rosnąca w ostatnim czasie łatwość nabycia/sprzedaży akcji przez osoby niezwiązane profesjonalnie z obrotem papierami wartościowymi (dzięki portalom jak eToro czy Plus500) otworzyła bardzo wąski i sprofesjonalizowany do tej pory rynek narzędzi przewidujących zachowanie spółek na giełdzie. Zapotrzebowanie na tego typu narzędzia przejawiają teraz nie tylko korporacje z wielkim budżetem, ale również indywidualni użytkownicy, często początkujący gracze na giełdzie, poszukujący prostego i taniego narzędzia które zwiększyłoby prawdopodobieństwo podejmowania przez nich trafnych decyzji.

Kolejnym powodem dla którego zdecydowaliśmy się stworzyć to narzędzie jest rosnąca liczba powszechnie dostępnych informacji posiadających wartość jeśli chodzi o podejmowanie decyzji takich jak sprzedaż/kupno akcji.

Ilość portali i newsów na nich zdecydowanie wykracza poza możliwości manualnej analizy, zwłaszcza zważywszy na dynamiczny charakter rynku papierów wartościowych, gdzie często minuty a nawet sekundy decydują o rentowności danej inwestycji.

3. Synteza problemów z treści opisu zadania

Analiza treści zadania oraz poszukiwanie sposobów na jego wykonanie pozwoliło wyznaczyć największe problemy które musiały zostać rozwiązane do stworzenia sprawnego projektu:

1. Brak zbioru uczącego składającego się z tekstów które końcowy produkt miał analizować ocenionych tylko pod względem przyszłego zachowania spółki
2. Wybór architektury sieci wykorzystanej do oceny tekstów
3. Forma aplikacji udostępnionej dla użytkownika końcowego

Kolejność problemów jest zgodna z kolejnością ich rozwiązania.

4. Osiągnięta funkcjonalność

Finalne narzędzie, które jest efektem naszego projektu, służy do analizy i oceny danego newsa dotyczącego wybranej spółki giełdowej.

Ocenie jest poddawana treść jak i wydźwięk samego artykułu. Na podstawie wczytanego fragmentu tekstu nasze narzędzie ocenia dany artykuł pod kątem prognozowanego zachowania opisywanej we fragmencie spółki. Możliwe są trzy następujące werdykty:

- Ocena pozytywna
- Ocena neutralna

- Ocena negatywna

Nasze narzędzie ma charakter doradczy, użytkownik będzie w stanie na podstawie wydanej oceny zdecydować czy kupić (jeśli perspektywa danej spółki jest pozytywna, prognozowany jest wzrost ceny akcji), zachować (jeśli perspektywa jest neutralna) lub sprzedać (jeśli perspektywa jest negatywna i prognozuje się spadek wartości akcji) akcje interesującej dla użytkownika narzędzia spółki. Ułatwi też mniej doświadczonym inwestorom utrzymanie w portfelu aktywa mającego potencjał wzrostowy po spadku cenu z powodu np. słabego raportu kwartalnego.

5. Sposoby rozwiązania problemów

Do oceny sentymentu towarzyszącego danej spółce, bądź aktywowi finansowemu, wykorzystany został zaczerpnięty z dziedzin przetwarzania języka naturalnego oraz uczenia maszynowego, model finBERT, będący adaptacją szerzej znanego modelu BERT (*Bidirectional Encoder Representation from Transformers*) do terminów z zakresu finansów i ekonomii. Projekt bazuje na gotowej implementacji modelu [finBERT](#), jednakże fine-tuning odbywa się z wykorzystaniem dodatkowych baz danych, przede wszystkim tych stworzonych przez Autorów. Hiperparametry sieci dostosowano do zadania. Ponadto, opracowano również prostą aplikację, wewnątrz której dokonywana jest ocena wprowadzanego tekstu przez sieć.

6. Topologia modelu

Do rozwiązania zadania została użyta architektura finBERT. Wykorzystywane są w niej sieci typu Transformer, który działa wykonując małą, stałą liczbę kroków. W każdym z nich stosowany jest mechanizm uwagi pozwalający na zrozumienie znaczenia relacji kontekstualnych pomiędzy słowami w sentencji, niezależnie od ich pozycji oraz wieloznaczności. Najważniejszymi warstwami sieci typu Transformer są enkoder, pozwalający na wczytanie zamienionych w sekwencję tokenów, danych tekstowych oraz dekodery, podający na wyjściu predykcję. Co więcej, podczas pretreningu w ramach wykorzystującego Transformers modułu Masked Language Model (MLM), maskowane są przypadkowe tokeny (reprezentujące wyrazy), co pozwala na znaczne załagodzenie więzów związanych z odczytywaniem kontekstu w jednym tylko kierunku. Moduł Next Sentence Prediction (NSP) odpowiada z kolei za zrozumienie przez finBERT relacji kontekstualnej między sąsiednimi zdaniami. Podczas fine-tuning, stosowane są warstwy dostosowane do naszego zadania, wraz z technikami zapobiegającymi katastroficznemu zapomnieniu. W porównaniu z pretreningiem, fine-tuning wymaga znacznie mniej mocy obliczeniowej.

7. Narzędzia

- Python3.7
- PyTorch
- transformers
- finBERT
- Beautiful Soup
- sELENIUM
- TRC2-financial oraz Financial PhraseBank - dane uczące dostępne w sieci

8. Zbiory danych

Na potrzeby prezentacji poprawności konceptu zostały wykorzystane dwa zbiory danych:

- [FinancialPhraseBank](#)
- Własnoręcznie zebrane artykuły ze strony internetowej

Pierwszy ze zbiorów danych został opisany i wykorzystany w artykule opisującym wyniki pracy [finBERT](#). Składa się on z par „wiadomość-ocena” tworzonych ręcznie z perspektywy ekonomicznej.

Zbiór ten nie spełniał założeń stawianych przed produktem końcowym:

- Ocena wiadomości na podstawie ich treści, nie zachowania spółki w przyszłości
- Długość jednego zdania - za mało wiadomości zawierających więcej niż jeden fakt

Nadrzędnym problemem niesformułowanym wprost dla narzędzia oraz danych było oszacowanie jakie informacje powodują wzrost wartości kapitalizacji spółek a jakie przeciwnie. Początkujący inwestor nie posiada tej intuicji, na skutek czego około 80% inwestorów traci grając na giełdzie.

Te problemy z gotowym zbiorem danych uczyniły niezbędnym utworzenie zbioru danych których powyższe problemy nie dotyczą, wykorzystując powyższy zbiór tylko jako dodatek podczas uczenia poprawiający jakość oceny trywialnych tekstów.

Została ściągnięta lista symboli wystawionych na *The New York Stock Exchange*. Zostało wytypowane źródło artykułów:

- artykuły z czystymi faktami, bez przemyśleń autora
- strona umożliwiająca trywialnymi metodami pobranie w wolnym tempie dużej ilości stron (zachowując przyzwoitość - niewielkie tempo nawiązywania połączeń)
- prosta w pracy z nią struktura pliku HTML strony w celu wyciągnięcia danych

Wybrana strona z czasem potwierdziła spełnienie tych trzech warunków. Pierwszy warunek był najważniejszy dla następnych faz realizacji projektu. Kolejne dwa warunki umożliwiły stworzenie zbioru danych jako takiego.

Z wykorzystaniem Selenium Chrome Web Driver został stworzony web-scraper archiwizujący każdy artykuł zapisując wszystkie związane z nim informacje dla każdej spółki o której były artykuły na witrynie.

Metodą oceny artykułów była zmiana ceny akcji po trzech miesiącach od daty napisania artykułu.

Głównymi problemami początkujących inwestorów jest niecierpliwość oraz płochliwość - kupowanie aktywów w okresie dużego wzrostu, często po ważnym dla spółki wydarzeniu oraz sprzedaż po pierwszej korekcie ceny akcji. Krótki okres „testu” wiadomości jest podyktowany takim zachowaniem - jest to okres między dwoma sprawozdaniami kwartalnymi, które mają duży wpływ na kształtowanie się ceny spółek.

9. Trening i parametry sieci

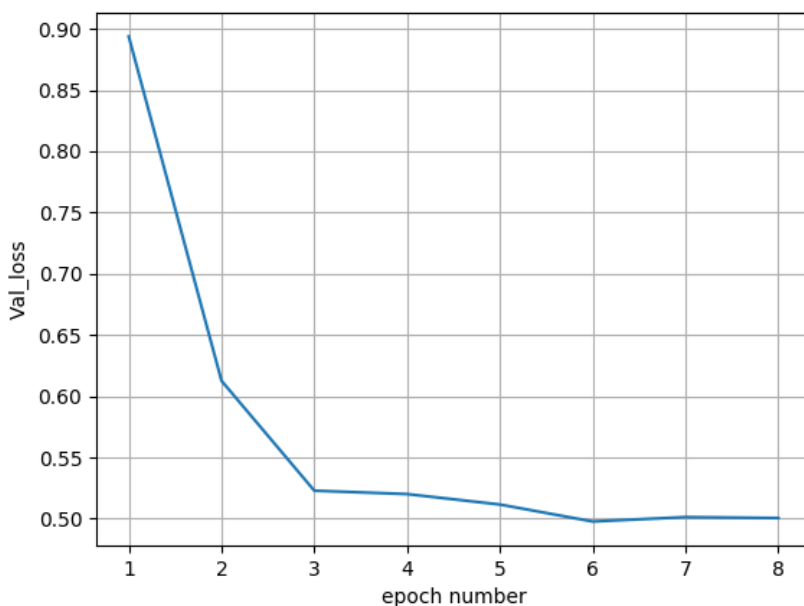
Najważniejsze parametry sieci:

- Ilość epok: 8;
- Maksymalna długość sekwencji: 48;
- *Batch_size*: 32;
- Współczynnik uczenia: 10^{-5}
- Liniowy współczynnik rozgrzewki procedury uczenia *Warmup proportion rate*: 0.2
- Współczynnik penalizujący złożoność sieci *Weight decay*: 0.01

Trening przeprowadzано na podzespołach: GeForce GTX960M, Intel Core i7-6700HQ CPU @ 2.60GHz, 16GB RAM, dysk HDD. Przy takiej konfiguracji oraz podanych wyżej parametrach sieci, czas treningu wyniósł ~35min.

10. Rezultaty

Najlepsze rezultaty osiągnięto, gdy dane uczące były pojedynczym zdaniem, a w których to danych dodatkowo usunięto kropki z przeważającej większości skrótów. Metryki są oczywiście obliczane dla danych weryfikujących. Przyporządkowano 0 dla pozytywnego sentymentu, 1 dla negatywnego sentymentu, 2 dla neutralnego sentymentu.



Rys. 1: Wartość funkcji strat w zależności od epoki

Loss:0.51
Accuracy:0.73

Classification Report:

	precision	recall	f1-score	support
0	0.77	0.57	0.65	747
1	0.58	0.77	0.66	510
2	0.87	0.92	0.89	575
accuracy			0.73	1832
macro avg	0.74	0.75	0.74	1832
weighted avg	0.75	0.73	0.73	1832

Rys. 2: Najlepszy model - rezultaty

Mając na uwadze przyjętą automatyczną ocenę sentymentu, uzależnioną wyłącznie od zmian ceny akcji po trzech miesiącach, otrzymane wyniki są bardzo dobre.

W przypadku użycia kapitału do zatrudnienia doświadczonych ekspertów np. z branży funduszy hedgingowych lub doradztwa finansowego w celu ręcznej anotacji zbioru danych, rezultaty mają ogromny potencjał wzrostowy.

	sentence	logit	prediction	sentiment_score
0	This week, the cryptocurrency markets displayed a more sophisticated understanding of regulatory and technology risk.	[0.8139987, 0.012689185, 0.17331214]	positive	0.801310
1	Tunisian Finance Minister says Bitcoin ownership should be decriminalized.	[0.07808677, 0.070130914, 0.8517823]	neutral	0.007956
2	The Netherlands must ban the mining, trading and holding of Bitcoin because it doesn't meet any of the three functions of money and is handy for criminals, one Dutch official argued.	[0.032660235, 0.7711261, 0.19621362]	negative	-0.738466
3	The Netherlands should regulate crypto instead of banning it, says finance minister.	[0.14047715, 0.035290185, 0.82423264]	neutral	0.105187
4	Some attendees of the Bitcoin 2021 event in Miami have tested positive for COVID-19 after returning home from the conference, leading to a wave of negative media coverage and social media speculation it could turn into a super spreader event.	[0.01676907, 0.9570232, 0.026207715]	negative	-0.940254
5	Bitcoin sell pressure may hit zero in July thanks to Grayscale's giant 16K BTC unlocking.	[0.47732934, 0.14205186, 0.3806188]	positive	0.335277
6	Death knell for Chinese crypto miners, rigs on the move after gov't crackdown.	[0.020483505, 0.9159391, 0.063577406]	negative	-0.895456
7	El Salvador's move to make bitcoin legal tender offers an opportunity to prove that cryptocurrency can power renewable energy development, says CoinDesk's chief content officer.	[0.8863148, 0.007840498, 0.105844736]	positive	0.878474
8	Without proper oversight, there could be a worsening in market transparency, the basis of legality and rational choice for (market) operators, Consob Chairman Paolo Savona said.	[0.0212239, 0.91607565, 0.06270048]	negative	-0.894852
9	Investment bank JPMorgan Chase has warned of a further bitcoin price decline, expecting an incoming bear market.	[0.014326043, 0.9671445, 0.01852952]	negative	-0.952818

Rys. 3: Ocena sentymentu dla przykładowych doniesień dot. kryptowaluty Bitcoin

Ocena przez sieć komunikatów prasowych, niemal w całości zgadza się z oceną ludzką.

	sentence	logit	prediction	sentiment_score
0	Nvidia asks Chinese regulators to approve \$40 billion Arm deal.	[0.72471833, 0.029226627, 0.24605505]	positive	0.695492
1	Nvidia Corp forecast second-quarter revenue above analysts' estimates on Wednesday, but shares fell 1% after-hours as the company could not say for certain how much of its recent revenue rise was driven by the volatile cryptocurrency-mining market.	[0.013428428, 0.97854006, 0.008031511]	negative	-0.965112
2	Nvidia Corp forecast second-quarter revenue above analysts' estimates on Wednesday, betting on strong demand for its flagship gaming and artificial intelligence chips for data centers.	[0.84503716, 0.118082665, 0.036880203]	positive	0.726955
3	Chipmaker Nvidia Corp on Friday announced a four-for-one stock split.	[0.041464094, 0.063010566, 0.89552534]	neutral	-0.021546
4	Hedge funds are cashing out of NVIDIA Corporation.	[0.04369671, 0.41451225, 0.5417911]	neutral	-0.370816

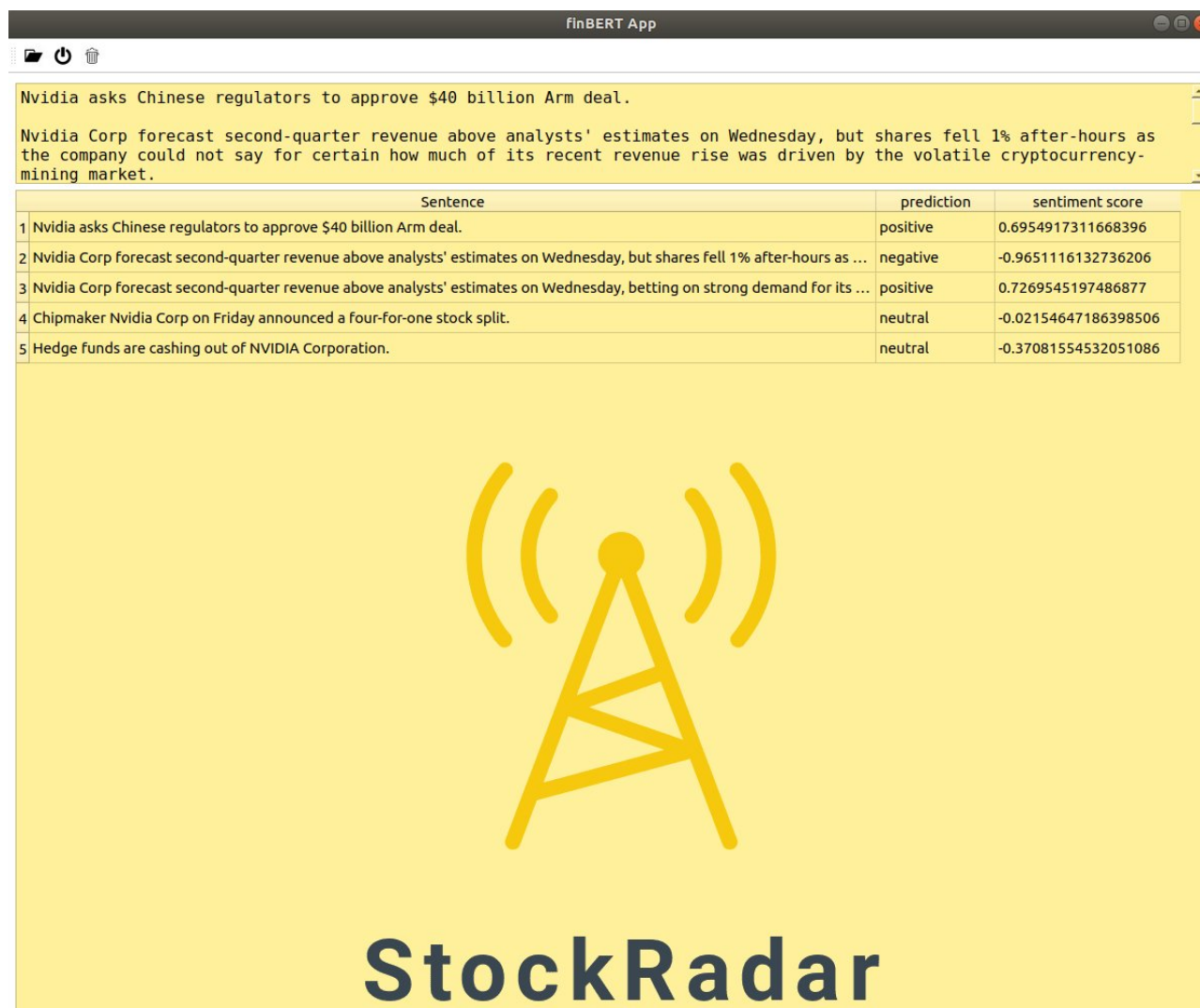
Rys. 4: Ocena sentymentu dla przykładowych doniesień dot. akcji spółki Nvidia

Ocena przez sieć komunikatów prasowych, zgadza się z oceną ludzką, być może za wyjątkiem pozycji 3.

Taka automatyzacja czytania i skupiania się na komunikatach medialnych to potencjalnie ogromne ułatwienie i oszczędność czasu dla (zwłaszcza początkującego) inwestora.

11. Aplikacja StockRadar

Wykonano również aplikację StockRadar z przejrzystym interfejsem, aby umożliwić intuicyjną i niezwykle szybką ocenę komunikatów prasowych zawartych w pliku *.txt* . Wartości sentymentu można odczytać z wyświetlanej w aplikacji tabelki:



	Sentence	prediction	sentiment score
1	Nvidia asks Chinese regulators to approve \$40 billion Arm deal.	positive	0.6954917311668396
2	Nvidia Corp forecast second-quarter revenue above analysts' estimates on Wednesday, but shares fell 1% after-hours as ...	negative	-0.9651116132736206
3	Nvidia Corp forecast second-quarter revenue above analysts' estimates on Wednesday, betting on strong demand for its ...	positive	0.7269545197486877
4	Chipmaker Nvidia Corp on Friday announced a four-for-one stock split.	neutral	-0.02154647186398506
5	Hedge funds are cashing out of NVIDIA Corporation.	neutral	-0.37081554532051086

Rys. 5: Ocena sentymentu dla przykładowych doniesień dot. akcji spółki Nvidia - GUI

Menu zawiera 3 przyciski (od lewej) : *Open file*, *Run*, *Reset*.

- Opcja *Open file* pozwala wybrać plik *.txt* z doniesieniami prasowymi, po czym automatycznie przenosi jego zawartość do pola tekstowego.
- Opcja *Run* uruchamia ocenę sentymentu dla poszczególnych sentencji przez sieć. Po zakończeniu tego procesu, wyświetla się widoczna na rysunku tabelka z rezultatami dla każdej sentencji.
- Opcja *Reset* czyści pola, przywracając aplikację do stanu początkowego.

Aplikacja jest na tyle prosta w obsłudze, że skorzystać z niej może każda osoba nietechniczna.