

Akademia Górniczo-Hutnicza

im. Stanisława Staszica w Krakowie

Wydział Informatyki, Elektroniki i Telekomunikacji

Katedra Informatyki



Projekt dyplomowy

Analiza wpływu twittów na kurs akcji

Marek Grzyb

Opiekun Projektu: dr inż. Robert Marcjan

Kraków 2019

Spis treści

1. Wstęp.....	1
2. Notowania spółek.....	2
3. Twitter.....	3
3.1. Konto deweloperskie.....	3
3.2. Autoryzacja.....	4
3.3. Pobieranie twittów.....	4
3.4. Struktura wiadomości.....	5
4. Przetwarzanie języka naturalnego.....	6
5. Analiza.....	7
5.1. Studium przypadku: PKN ORLEN.....	7
5.2. Studium przypadku: CD PROJECT.....	8
5.3. Podsumowanie.....	8
6. Referencje.....	9

1. Wstęp

Celem pracy jest zbadanie jaki wpływ na decyzje inwestorów mają informacje zamieszczane w serwisach społecznościowych. W tym opracowaniu przedstawię związki między informacjami umieszczanymi na Twitterze a kursami akcji spółek notowanych na Giełdzie Papierów Wartościowych.

Podczas analizy wykorzystam informacje o kursie akcji pobrane za pośrednictwem API Bankier.pl oraz treść Twittów.

W pracy przeprowadzę analizę twittów, sprawdzę skorelowanie ich sentymentu z kursem akcji. Będę poszukiwał twittów które wyprzedzają zmianę kursu oraz osoby mające znaczny wpływ na kurs akcji.

2. Notowania spółek

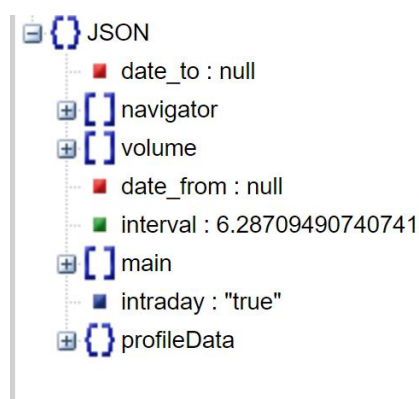
W przeprowadzonym eksperymencie wykorzystam informacje o kursach akcji spółek notowanych na Giełdzie Papierów Wartościowych uwzględnionych w indeksie WIG20.

Notowania spółek pobierane są ze strony bankier.pl. Do pobierania danych wykorzystywane jest zapytanie HTTP GET:

<https://www.bankier.pl/new-charts/get-data?symbol={symbol}&intraday=true&type=area>

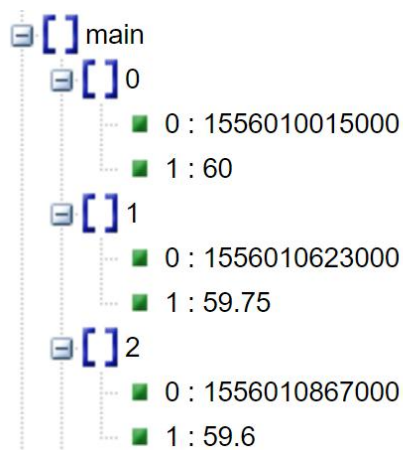
gdzie {symbol} oznacza symbol spółki giełdowej której notowania chcemy pozyskać.

Pobrane dane mają format JSON o strukturze zobrazonej na Rys. 1.



Rys. 1. Struktura wiadomości JSON zwracanej przez API bankier.pl

Atrybut "main" zawiera kolekcję kursy spółki w postaci:



Rys. 2. Struktura JSON kursów akcji zwracana przez API bankier.pl

gdzie pierwsza wartość oznacza czas wyrażony w formacie POSIX (liczba sekund od roku 1970) dodatkowo pomnożona przez 1000, drugi to cena.

Pobrane dane wyrażają kurs spółki z interwałem 1 min w dni robocze od 9:00 do 17:00 (godziny pracy GPW).

3. Twitter

Twitter udostępnia API dzięki któremu możemy w prosty sposób pobierać informacje o treściach zamieszczanych na portalu.

Wyszukiwanie dostępne jest w trzech wersjach:

- Standard

Wyszukiwanie jedynie na próbce twittów z ostatnich 7 dni.

- Premium

Wyszukiwanie na pełnym zbiorze twittów. Darmowe konto ma jednak ograniczenia, 250 zapytań miesięcznie albo 1MB danych miesięcznie. Postaram się zmieścić w tym limicie dla celów eksperymentu.

- Enterprise

Dostępne tylko po podpisaniu stosownej umowy.

Cennik dodatkowych wyszukiwań w opcji Premium został zamieszczony na Rys 3.

	Total Requests PER MONTH ?	Month-to-month PRICE PER MONTH ?	
Paid			
	Up to 500	\$149.00	Select
	Up to 1000	\$289.00	Select
	Up to 2,500	\$699.00	Select
	Up to 5,000	\$1,299.00	Select
	Up to 10,000	\$2,499.00	Select

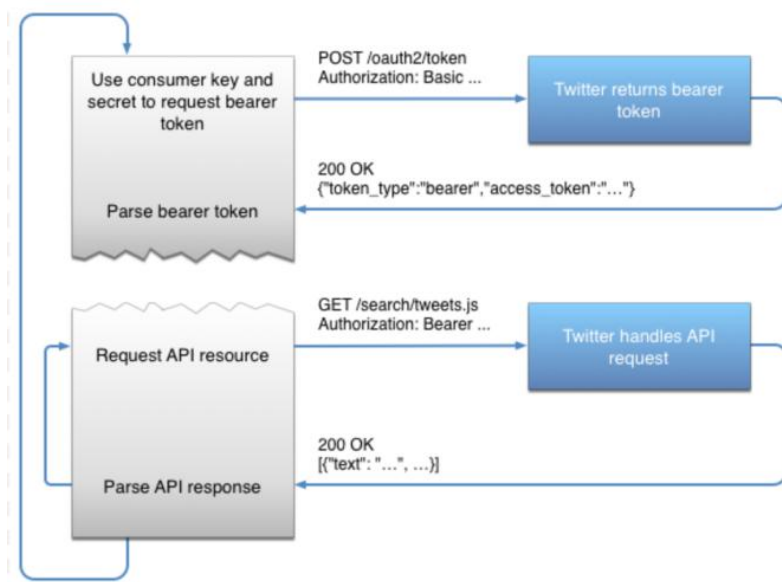
Rys. 3. Cennik usługi Premium Twittera

3.1. Konto deweloperskie

Aby móc korzystać z API twittera potrzebne jest konto deweloperskie, aby je założyć należy szczegółowo wyjaśnić w jakim celu dane pochodzące z twistera będą wykorzystane.

3.2. Autoryzacja

Dostęp do API wymaga autoryzacji protokołem OAuth 2.0, polega on na przesłaniu nazwy aplikacji oraz sekretnego klucza, w odpowiedzi otrzymujemy token który pozwala wykonywać udostępniane metody API.



Rys. 4. Diagram obrazujący sposób wymiany wiadomości podczas autoryzacji OAuth 2.0

3.3. Pobieranie twittów

Wykorzystane API dostępne jest pod adresem:

<https://api.twitter.com/1.1/tweets/search/30day/DEV.json>

Mechanizm komunikacji z tym API został zaimplementowany w wielu bibliotekach, jedna z nich to 'tweepy' (strona projektu <http://www.tweepy.org>), która zostanie wykorzystana w tym opracowaniu.

3.4. Struktura wiadomości

Pobrane dane mają format JSON. Z pośród kilkunastu dostępnych atrybutów do dalszej analizy wybierzemy tylko kilka najważniejszych:

- id - identyfikator
- created_at – data utworzenia
- full_text - pełny tekst wiadomości
- user.id, user.name - identyfikator i nazwa użytkownika
- user.followers_count, user.friends_count, user.listed_count, user.favourites_count – parametry wskazujące jak popularny i aktywny jest to użytkownik
- retweet_count, favorite_count – parametry wskazujące jak popularny jest to twitt



Rys. 5. Struktura wiadomości JSON opisująca Twitt

4. Przetwarzanie języka naturalnego

Pierwszym etapem procesowania tekstu Twittów to jego podział na wyrazy (tokenizacja). Najprostszym sposobem podziału jest rozdzielenie tekstu po spacjach, strategia ta nie jest doskonała gdyż nie uwzględnia wyrazów rozdzielonych myślnikiem np.: czarno-biały. W tym opracowaniu tokenizacja została wykonana przy użyciu biblioteki NLTK (strona projektu <http://www.nltk.org>).

Drugim etapem jest sprowadzenie rozpoznanych wyrazów do formy podstawowej np.: ropę -> ropa, w tym celu wykorzystamy słownik fleksyjny pobrany ze strony <https://sjp.pl/sownik/odmiany/>, zawiera on 224 tysiące form bazowanych wyrazów. Dla każdej z form bazowych słownik zawiera formy odmienione np.: ropa, rop, ropach, ropami, ropą, ropę, ropie, ropo, ropom, ropy. Wycinek słownika fleksyjnego pokazuje Rys.6.

Z uwagi na fakt, że twitty mogą nie zawierać akcentów, przy dopasowaniu uwzględnimy mapowanie:

a - ą, c - ć, e - ę, l - ł, n - ń, o - ó, s - ś, z - ź, z - ż

Tak przygotowane dane zostaną wykorzystane do dalszej analizy. W kolejnych krokach sprawdzimy powiązanie twittu ze spółką oraz ich wydźwięk (sentyment).

dąsać, dąsa, dąsacie, dąsaj, dąsają, dąsając, dąsająca, dąsającą, dąsające, dąsającego, dąsającej, dąsającemu, dąsający, dąsających, dąsającym, dąsającymi, dąsajcie, dąsajcież, dąsajmy, dąsajmyż, dąsajże, dąsali, dąsaliby, dąsalibyście, dąsalibyśmy, dąsaliście, dąsaliśmy, dąsał, dąsała, dąsałaby, dąsałabym, dąsałabyś, dąsałam, dąsałaś, dąsałby, dąsałbym, dąsałbyś, dąsałem, dąsałeś, dąsało, dąsałoby, dąsały, dąsałyby, dąsałybyście, dąsałybyśmy, dąsałyście, dąsałyśmy, dąsam, dąsamy, dąsania, dąsaniach, dąsaniami, dąsanie, dąsaniem, dąsaniom, dąsaniu, dąsano, dąsań, dąsas, dąsalska, dąsalską, dąsalskich, dąsalskie, dąsalskiej, dąsalskim, dąsalskimi, dąsalski, dąsalscy, dąsalska, dąsalską, dąsalskich, dąsalskie, dąsalskiego, dąsalskiej, dąsalsk iemu, dąsalskim, dąsalskimi, dąsalsko, dąsalsku, dąsalski, dąsalscy, dąsalskich, dąsalskie, dąsalskiego, dąsalskiemu, dąsalskim, dąsalskimi, dąsalskość, dąsalskości, dąsalskościach, dąsalskościami, dąsalskością, dąsalskościom, Dąsal, Dąsala, Dąsalach, Dąsalami, Dąsale, Dąsałem, Dąsalom, Dąsalowi, Dąsalowie, Dąsalów, Dąsalu, dąsik, dąsikach, dąsikami, dąsiki, dąsikiem, dąsikom, dąsikowi, dąsików, dąsiku, dąsy, dąsach, dąsami, dąsom, dąsów, dąs, dąsach, dąsami, dąsem, dąsie, dąsom, dąsowi, dąsów, dąsu, dąsy, dążenia, dążeniach, dążeniami, dążeniom, dążień, dążenie, dążenia, dążeniach, dążeniami, dążeniem, dążeniom, dążeniu, dążień, dążność, dążności, dążnościach, dążnościami, dążnością, dążnościom, dążyć, dąż, dążą, dążąc, dążąca, dążącą, dążące, dążącego, dążącej, dążącemu, dążący, dążących, dąż

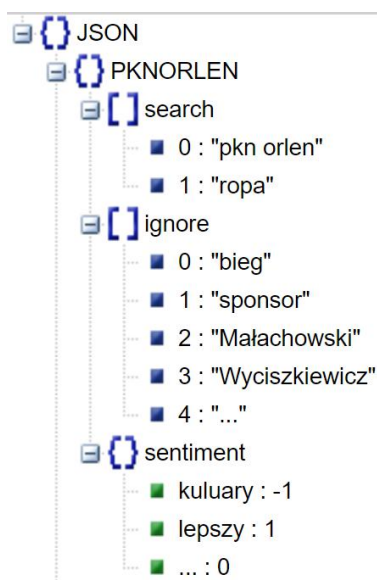
Rys. 6. Wycinek słownika fleksyjnego

5. Analiza

W tym opracowaniu przedstawie szczegółową analizę dla kilku spółek notowanych na GPW. Wybrałem spółki z index-u WIG 20, są to jedne z bardziej znanych w swojej kategorii:

- PKN ORLEN - największa polska firma petrochemiczna
- CD PROJECT - najbardziej rozpoznawalny twórca gier komputerowych w Polsce, twórca kultowej gry Wiedźmin,
- GRUPA AZOTY - największy koncern chemiczny w Polsce
- LPP - polskie przedsiębiorstwo odzieżowe zajmujące się projektowaniem, produkcją i dystrybucją odzieży, mające w swoim portfolio marki: Reserved, House, Cropp, Mohito i Sinsay
- mBank - czwarty co do wielkości bank w Polsce

Dla każdej przeanalizowanej spółki została przygotowana konfiguracja przedstawiona na Rys. 7.

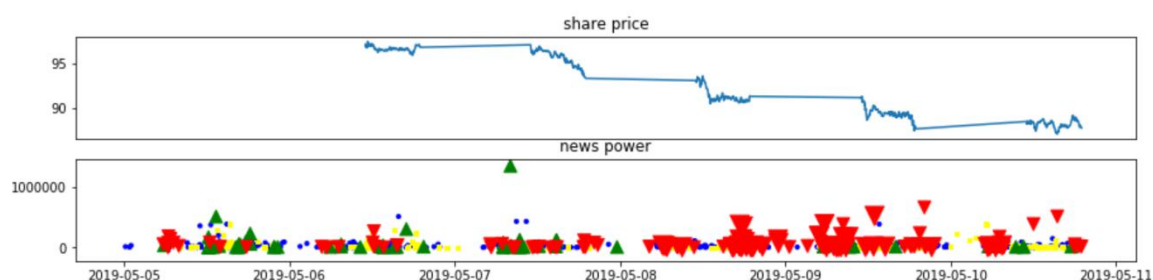


Rys. 7. Struktura konfiguracji dla jednej ze spółek

Gdzie

- search - słowa kluczowe wykorzystane przy zapytaniach o twitty
- ignore – słowa które powodują ignorowanie danego twittu, dla PKN ORLEN informacje o sportowcach sponsorowanych przez firmę
- sentiment - słowa kluczowe oraz wskaźnik sentymentu z jakim one się wiążą

Wyniki analizy przedstawie są na wykresie, przykładowy wykres na Rys. 8.



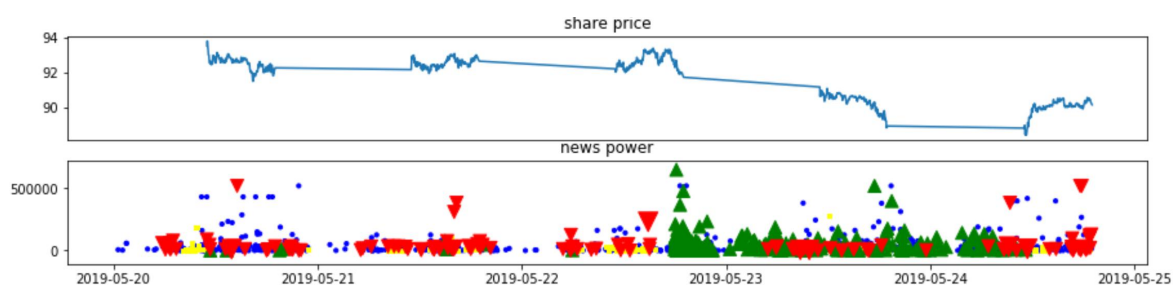
Rys. 8. Wyniki analizy

Górna część wykresu przedstawia cenę akcji w czasie dla analizowanej spółki, cena zmienia się tylko w godzinach 9-17, w dni robocze, jest to związane z godzinami pracy Giełdy Papierów Wartościowych. Wykres oparty jest o dane godzinowe.

Dolny wykres przedstawia twitty w czasie. Os X jest wspólna z wykresem kursu akcji. Os Y przedstawia moc oddziaływania twitta oraz osoby która go opublikowała. Kształty i kolory twittów oznaczają:

- niebieskie koło – twitt niesklasyfikowany
- żółty kwadrat - twitt zignorowany
- czerwony trójkąt - informacja negatywna
- zielony trójkąt - informacja pozytywna

5.1. Studium przypadku: PKN ORLEN



Rys. 9. Korelacja cen akcji i sentymentu twittów dla PKN ORLEN

Dla PKN ORLEN na twitterze znajdowało się sporo informacji o sponsorowanych sportowcach: Orlen Team, twitty te jednak nie są zbyt popularne, mieszczą się w dolnej części wykresu.

Najbardziej znaczące twitty zostały opublikowane przez:

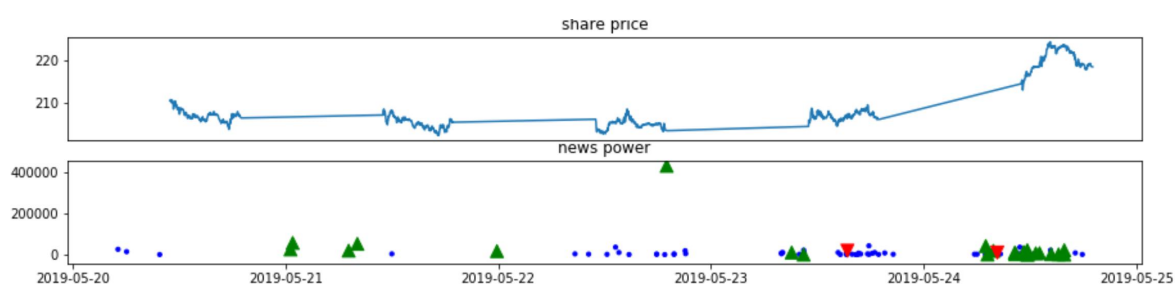
- Bartek Piekarski
- Wojciech Jakóbiak
- Rzeczpospolita Ekonomia
- Jarosław Szmyt



Rys. 10. Twitty o największej mocy

Jeden z twittów (oznaczony zieloną ramką na Rys. 10.) wydaje się być wyprzedzający zmianę kursu akcji. Jest to twitt Jarosław Szmyt który odnosi się do artykułu zamieszczonego na stronie FMF FM pod tytułem: „Zanieczyszczona ropa z Rosji: Niemcy chcą przerzucić problem na Polskę?”. Jest to informacja która mogła mieć wpływ na notowania PNK ORLEN, aczkolwiek akcjonariusze mogli pozyskać tę informację z różnych kanałów.

5.2. Studium przypadku: CD PROJECT



Rys. 11. Korelacja kursu akcji CD PROJEKT z sentymentami twittów

Dla CD PROJEKT najbardziej znaczącą informacją był twitt WPROST o tym oczekiwaniach w związku z grą „Cyberpunk 2077”, nie miał on jednak wpływu na cenę akcji.

Kurs firmy znacząco wzrósł między czwartkową a piątkową sesją. Analiza twittów z tego okresu odpowiada na pytanie dlaczego to się stało. O 2019-05-24 07:00:33 na twittterze pojawiła się informacja o nadspodziewanie dobrych wynikach finansowych firmy, opublikowana przez Wirtualnedia.pl, ta informacja pochodziła z oficjalnego raportu firmy opublikowanego 23-05 o 18:00 (już po zamknięciu rynku).

5.3. Studium przypadku: GRUPA AZOTY

5.4. Studium przypadku: LPP

5.5. Studium przypadku: mBank

5.6. Podsumowanie

Na podstawie przeanalizowanych przykładów nie możemy stwierdzić że twitter odgrywa kluczową rolę dla kursu cen akcji. Dla żadnej z przeanalizowanej spółki nie odnalazłem kluczowej osoby, informacji która była by 'breaking news' opublikowana tylko na twitterze i powodująca gwałtowną reakcję giełdy.

6. Referencje

1. <https://www.bankier.pl> - strona z informacjami o kursach akcji i nie tylko
2. <http://www.tweepy.org> - strona biblioteki implementującej komunikację z twitterem
3. https://github.com/PyConPL/Book/tree/master/2013/przetwarzanie_jezyka_naturalnego_w_praktyce - praktyczne informacje o analizie tekstu w języku polskim
4. <http://www.nltk.org> - biblioteka do procesowania języka naturalnego
5. <https://github.com/agh-glk/pydic> - narzędzie do sprowadzania wyrazów do formy podstawowej
6. <http://jsonviewer.stack.hu> - narzędzie do wizualizacji json
7. <https://sjp.pl/slownik/odmiany> - słownik fonetyczny dla języka polskiego