# Internetowe predykcje notowań spółek giełdowych

**Agnieszka Woch, Michał Wójcikiewicz**

W ciągu każdej sekundy przez internet transferuje się 22574 GB danych, powstaje 5700 tweetów, 55 tysięcy postów na Facebooku, a na portal YouTube dodawane są dwie godziny materiału[1]. Ten „cyfrowy wszechświat" co dwa lata podwaja swoje rozmiary. W 2020 r. liczba bitów informacji wygenerowanych przez ludzkość przekroczy liczbę gwiazd we Wszechświecie[2]. Obecnie skutecznie analizuje się tylko 0,5 proc. danych zgromadzonych w postaci cyfrowej[3].

Jednym ze sposobów badania internetowych treści jest rafinacja informacji, które pochodzą z zasobów Big Data. Według definicji McKinsey Global Institute pod pojęciem Big Data rozumiemy „duże zbiory danych, które z powodu swojej wielkości nie są możliwe do przechwycenia, przechowywania, zarządzania i analizowania przy użyciu typowych narzędzi programowania"[4]. Charakteryzuje ją kilka cech: objętość (*volume*), polegająca na „znaczącej dynamice wzrostu danych"; różnorodność (*variety*), cechująca się tym, że analizowane dane pochodzą z różnych źródeł, występują w różnej postaci; strumień (*velocity*), oznaczający fakt, że informacje napływają strumieniowo, niezwykle szybko, co wymaga dodatkowych mocy obliczeniowych. Ponadto Big Data odznacza się złożonością (*complexity*) – dane są ustrukturyzowane w różnoraki sposób lub nie są uporządkowane w ogóle, należy poznać łączące je relacje i poddać je zintegrowaniu; zmiennością (*variability*) – ich natężenie jest zmienne w czasie – i wartością (*value*), indywidualną jakością informacyjną pozwalającą na wyciąganie wniosków ze zbiorów danych[5].

Rafinacja zasobów sieciowych umożliwia pozyskiwanie informacji wtórnych z analizy zbiorów informacji pierwotnych (Big Data)[6]. Może służyć jako niezwykle przydatne narzędzie dla dziennikarstwa – dzięki niej pracownicy mediów mają dostęp do niedocenianych i niepoznanych wcześniej treści. Wykorzystuje się ją także w różnych gałęziach gospodarki,

---

[1] *The Internet in real time,* http://pennystocks.la/internet-in-real-time/ [dostęp: 25.04.15].

[2] *The digital universe of opportunities: Rich Data and the inreasing value of the Internet of things,* http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm [dostęp: 25.04.15].

[3] *Big Data, bigger digital shadows, and biggest growth in the Far East,* https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf [dostęp: 25.04.15].

[4] Y. Liu, *Big Data and predictive business analytics,* http://eds.b.ebscohost.com.atoz.han.buw.uw.edu.pl/eds/pdfviewer/pdfviewer?sid=3149252a-86a3-462d-87c0-690b06fda804%40sessionmgr110&vid=6&hid=112 [dostęp: 09.05.15].

[5] M. Tabakow, J. Korczak, B. Franczyk, *Big Data – definicje, wyzwania i technologie informatyczne*, „Informatyka Ekonomiczna" 2014, nr 1 (31).

[6] W. Gogołek, P. Kuczma, *Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych,* „Studia Medioznawcze" 2013, nr 2 (53).

począwszy od ekonomii, przez reklamę, medycynę, ubezpieczenia, aż do nauki[7]. Jako przykład jej zastosowania może posłużyć działalność firm Brand24 i SentiOne zajmujących się pozyskiwaniem informacji z zasobów Big Data. Służą one do śledzenia sentymentów wpływających na pozycję danej marki. Rafinacja informacji może znaleźć zastosowanie również w monitoringu mediów, ocenie skuteczności kampanii marketingowych oraz poznawaniu profili reklamobiorców[8].

Badania w zakresie współzależności pomiędzy zasobami Big Data a notowaniami giełdowymi dopiero zaczynają stawać się obiektem zainteresowania naukowców. Dotychczasowe analizy zależności pomiędzy opiniami użytkowników internetu a zmianami cen akcji nie przyniosły jednoznacznych wniosków w zakresie tworzenia modeli predykcyjnych z wykorzystaniem zasobów Big Data. Tym zagadnieniem zajął się w swojej pracy Vasant Dhar[9]. Po analizie porównawczej sentymentów zawartych w ponad 13 tysiącach artykułów o tematyce biznesowej publikowanych przez Reutersa i notowań spółek giełdowych z indeksu S&P500 postawił kilka tez. Po pierwsze, zmniejszenie się liczby wypowiedzi o charakterze pozytywnym (pozytywne sentymenty), równoważącym zwykle stale pojawiające się w dyskusji nastroje pesymistyczne, może być zapowiedzią zbliżających się spadków na giełdzie. Z drugiej strony niski poziom pozytywnych sentymentów przy normalnym, nieodbiegającym od normy poziomie pejoratywnych komentarzy może świadczyć o tym, że inwestorzy sprzedali już swoje udziały, może więc być to przedsmak nadchodzących wzrostów – pieniądze pochodzące z owych transakcji będą prawdopodobnie niedługo ponownie włączone do obiegu. Ta hipoteza może także zakładać, że przy wysokim poziomie sentymentów pozytywnych i małej liczbie wypowiedzi negatywnych istnieje rosnące ryzyko spadku na giełdzie. W przypadku pojawienia się niepokojącej informacji (sentymenty negatywne) o danej spółce inwestujące w nią osoby mogą zacząć zbiorowo sprzedawać jej udziały, co spowoduje gwałtowny spadek jej wartości. Analiza haseł z wyszukiwarki Google, przeprowadzona w Warwick Business School przez Chestera Curme'a, Tobiasa Preisa, Harry'ego Eugene'a Stanleya i Helen Susannah Moat, dowiodła natomiast, że kryzys giełdowy może być poprzedzony wzmożonym wyszukiwaniem informacji związanych z polityką i biznesem.

Wyszukiwarka Google'a, podobnie jak inne tego typu narzędzia, przechowuje informacje o zachowaniach swoich użytkowników, niektóre z nich są publicznie dostępne dzięki Google Trends, z którego korzystano podczas prezentowanych tu badań. Okazało się, że wzmożone wyszukiwanie haseł związanych z tematami politycznymi i ekonomicznymi łączy się z niepokojem zarówno indywidualnych inwestorów, jak i ogółu społeczeństwa. Ten stan rzeczy ma odzwierciedlenie w spadkach na giełdzie, które są rezultatem zwątpienia w wysoką wartość akcji i przeprowadzania transakcji na mniejsze kwoty. Ta zależność zmniejsza się w ostatnich latach, prawdopodobnie z powodu włączenia internetowych danych do zautomatyzowanych strategii inwestowania[10].

Badania Eduarda Jose Ruiza, Vagelisa Hristidisa, Carlosa Castilla, Aristidesa Gionisa

[7] K. Smith, *Big Data discoveries,* „Best's Review" 2015, nr 7.

[8] H. Chen, R. Chiang, V. Storey, *Business intelligence and analytics: From Big Data to big impact*, „MIS Quarterly" 2012, s. 1165–1188.

[9] V. Dhar, *Can Big Data machines analyze stock market sentiment?,* „Big Data" 2014, nr 4, s. 178.

[10] C. Curme i in., *Quantifying the semantics of search behavior before stock market moves*, „PNAS" 2014, nr 32.

i Alejandra Jaimesa wykazały, że nierealna jest predykcja notowań danej spółki giełdowej na podstawie wyników uzyskanych z Twittera. Analiza sentymentów jego użytkowników okazała się niewystarczająca – brak stałości w hashtagach i zbyt wiele nieużytecznych danych w wynikach, takich jak niepowiązane, niedotyczące spółki informacje, odwołujące się do wieloznaczności jej nazwy, uniemożliwiły osiągnięcie zadowalających rezultatów. Badacze zaznaczyli, że niewykluczone jest, że analiza Twittera pod kątem całego rynku giełdowego może przynieść rezultaty[11]. W innych badaniach, przeprowadzonych przez Jasminę Smailović, Mihę Grčara, Nadę Lavrač i Martina Žnidaršiča, okazało się jednak, że możliwe jest przewidywanie zmian wartości spółek giełdowych. Naukowcy zastosowali mechanizm klasyfikujący maszyny wektorów nośnych do analizy tweetów – podzielili je na trzy grupy pod względem ich zabarwienia (pozytywne, negatywne, neutralne). Predykcja notowań spółek giełdowych była możliwa przy analizie zmian liczby wypowiedzi o zabarwieniu pozytywnym – zapowiadały one analogiczne zmiany wartości akcji poszczególnych spółek[12]. Johan Bollen, Huina Mao i Xiao-Jun Zeng dowiedli natomiast, że możliwa jest predykcja notowań spółki za pomocą analizy sentymentów użytkowników Twittera nieodnoszących się bezpośrednio do rynku giełdowego. Okazało się, że dzięki tej metodzie udało się w 86,7 proc. przewidzieć dzienną wartość akcji Dow Jones Industrial Average – spółki notowanej na nowojorskiej Giełdzie Papierów Wartościowych[13].

Oczekiwanych rezultatów nie przyniosły natomiast predykcje notowań na podstawie danych zawartych na forum Yahoo, które zostały przebadane przez Michaela Rechenthina, Nicka Streeta i Padmini Srinivasan – analiza pokrywała się tylko w ok. 50 proc. z wartością akcji. Badacze sprawdzili, czy wpływ na zdolność przewidywania wahań notowań ma uczciwość treści publikowanych przez graczy giełdowych. Autorzy pracy założyli, że mogą oni osłabiać wyniki analizy sentymentów przez stawianie nieprawdziwych tez, wprowadzanie innych użytkowników w błąd przez granie w inny niż opisany przez siebie sposób – wszystko po to, by zmanipulować ich działania i spowodować, by ci grali zgodnie z ich planem. Z badań wynika jednak, że osoby próbujące wpłynąć na rynkowe działania innych graczy nie oddziałują znacząco na ruchy innych inwestorów, a więc także na rezultaty badania sentymentów[14]. Forum Yahoo! Finance and Raging Bull zostało także przebadane przez Wernera Antweilera i Murraya Franka – ich analiza dowiodła, że treści zawarte na portalu wykazują nieduże możliwości predykcji późniejszych notowań[15].

## Metodologia

Dowodząc potencjału informacyjnego dużych zasobów sieciowych, podjęto pracę badawczą mającą na celu potwierdzenie lub wykluczenie współzależności pomiędzy wydźwiękiem opinii zawartych w zasobach sieciowych a notowaniami wybranych spółek. Przyjęto hipotezę, że możliwe jest wykazanie zależności między treściami publikowanymi w internecie przez

[11] M. Rechenthin, W.N. Street, P. Srinivasan, *Stock chatter: Using stock sentiment to predict price direction*, „Algorithmic Finance" 2013, nr 3–4, s. 170.

[12] J. Smailović i in., *Predictive sentiment analysis of Tweets: A stock market application*, „Human-computer interaction and knowledge discovery in complex, unstructured, Big Data" 2013, s. 77–88.

[13] J. Bollen, H. Mao, X. Zeng, *Twitter mood predicts the stock market*, „Journal of Computational Science" 2011, Vol. 2, wyd. 1.

[14] M. Rechenthin, W.N. Street, P. Srinivasan, *Stock chatter*…, dz. cyt., s. 169–196.

[15] S. Martínez Bustos i in., *Pricing stocks with yardsticks and sentiments*, „Algorithmic Finance" 2011, nr 1.

jego użytkowników a zmianą wartości akcji poszczególnych spółek.

Założono, że treści o zabarwieniu pozytywnym, a zatem informacje o sukcesach firmy, opinie o wzrostach cen spółki i rekomendacje kupna akcji spółki będą poprzedzać trend rosnący akcji, zaś negatywne komentarze, odnoszące się do wiadomości o porażkach spółki, spadkach jej wartości lub rekomendacji sprzedaży będą przewidywać tendencję spadkową. Do udowodnienia owej tezy zostały użyte dane czterech spółek giełdowych, których notowania porównano z wynikami analizy zbiorów danych z różnych źródeł internetowych: portali społecznościowych, forów internetowych, stron i blogów. Pod uwagę wzięto spółki notowane na WIG20: KGHM, Tauron, Enea i Synthos. Zakres czasowy badania wyniósł trzy tygodnie w styczniu i marcu 2015 r. Dobór terminu analizy miał na celu sprawdzenie skuteczności badania w dwóch odmiennych okresach.

Narzędzia potrzebne do realizacji badań, w zakresie „wyłuskania" informacji z zasobów sieciowych zostały udostępnione dzięki uprzejmości firmy SentiOne, której należą się szczególne podziękowania.

Analiza wydźwięku (sentymentu) polega na identyfikacji ładunku emocjonalnego wypowiedzi i zaszeregowaniu go do jednej z trzech kategorii – wypowiedź może być pozytywna, negatywna lub neutralna. Skuteczność istniejących narzędzi analizy sentymentów nie wystarcza, ze względu na trudności płynące ze złożoności języka i kontekstu. Z tego powodu w badaniu przeprowadzono analizę wypowiedzi, która w związku z brakiem odpowiednich narzędzi badawczych (dotyczących sentymentów) została oparta na opinii ekspertów.

Pojawiające się trudności polegały na prawidłowym rozpoznaniu ironii i sarkazmu, a także na poprawnym rozszyfrowaniu różnych odmian języka, np. slangu i żargonu. Kolejną komplikacją okazała się klasyfikacja wpisów – powstało pytanie, czy każda wypowiedź na temat spółki jest znacząca dla analizy, czy też należałoby ograniczyć się do informacji dotyczących wyłącznie jej notowań. Zwrócono także uwagę na problem komentarzy wirusowych, rozumianych jako wypowiedzi powielane przez internautów.

W analizie został użyty współczynnik korelacji Pearsona określający siłę liniowej zależności między dwiema zmiennymi – liczbą negatywnych i pozytywnych wypowiedzi w internecie i notowaniami spółki. Dzięki wielkości korelacji można ocenić podobieństwo zmian między dwoma cechami[16].

## Przebieg badań

W badaniu przeanalizowano wszystkie wypowiedzi na temat uwzględnianych spółek, które zostały zebrane za pomocą narzędzia SentiOne. Program agreguje zawężoną liczbę treści pochodzących z ogólnodostępnych źródeł, dlatego liczba wzmianek wyniosła 6318. Najczęściej pojawiające się słowa kluczowe oscylowały wokół inwestycji spółek, ich pozycji na giełdzie, decyzji podejmowanych przez ich władze oraz jakości świadczonych przez nie usług.

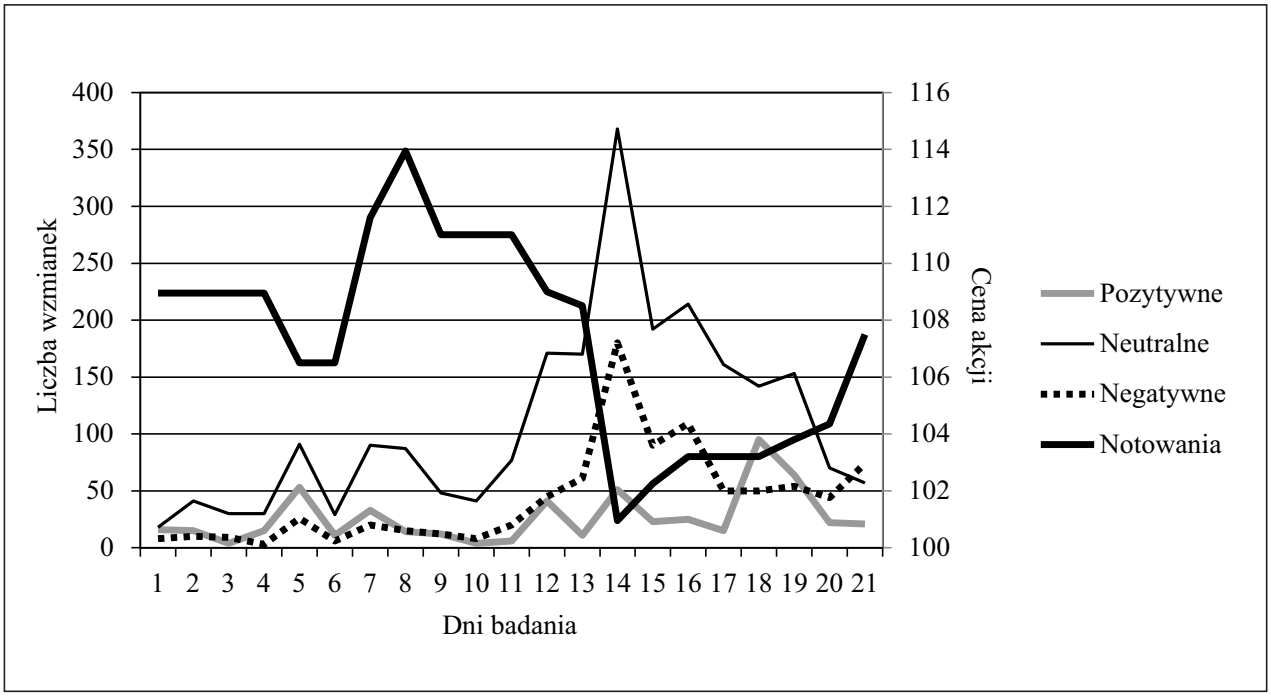
Za wyrazy o zabarwieniu pozytywnym – wyróżnione na podstawie opinii eksperckich – uznano: sukces, najlepszy interes, korzyści, dobra jakość, poprawa, odbicie, wzrost, zyski, spore inwestycje. Wpisy o charakterze negatywnym zawierały często wyrażenia: oszustwo, złodziejstwo, brak profesjonalizmu, trudna sytuacja, spadek, strata, słabe wyniki, chude

[16] Wartość $R^2$ waha się od -1 do 1. Im jest ona wyższa, tym większą analogią wykazują się oba czynniki. Ujemna wartość oznacza związek ujemny – zmienne charakteryzują się odwrotnie proporcjonalną zależnością, jedna z nich ($X$) rośnie, podczas gdy druga ($Y$) maleje.

lata, patologia, choroba, upadek, wygaszanie zakładów, podatek, nierentowność. Neutralne komentarze okazały się najtrudniejsze do usystematyzowania – wiele z nich charakteryzowało się treściami trudno klasyfikowalnymi, w dużym stopniu nieistotnymi dla analizy (co widać po wskaźnikach korelacji). Najczęściej pojawiającymi się słowami kluczowymi w tej kategorii były zwroty: zainteresowanie oraz inwestycje.

## *KGHM*

W okresie od 1 stycznia 2015 do 21 stycznia 2015 uzyskano 2377 wyników dotyczących spółki akcyjnej KGHM, z czego 19,44 proc. stanowiły wypowiedzi o zabarwieniu pozytywnym – odnotowano 462 tego typu wzmianki. Komentarze o nacechowaniu pejoratywnym, których zliczono 698, stanowiły 29,36 proc. ogółu treści. Przeprowadzona analiza wskazuje na to, że istnieje zależność między zabarwieniem wypowiedzi internautów dotyczących spółki KGHM a zmianami wartości jej akcji. W pierwszym badaniu, w czasie którego wzięto pod uwagę wpisy wirusowe, wskaźnik korelacji Pearsona dla sentymentów pozytywnych wyniósł -0,41, dla negatywnych: 0, a dla neutralnych: -0,65. W tej próbie wyeliminowano informacje na temat sponsorowania przez firmę drużyn sportowych, a także komentarze uznane za nieistotne dla analizy (wpisy wirusowe). W drugim badaniu postanowiono sprawdzić, jak wpisy wirusowe wpływają na korelacje między zmiennymi. Podjęto próbę odrzucenia trzech tego typu wypowiedzi, a także, podobnie jak wcześniej, informacji o sponsorowaniu drużyn. Zachowano jednak inne wpisy, uznawane wcześniej za nieistotne. Okazało się, że najsilniejszy związek między notowaniami KGHM-u a sentymentami użytkowników internetu wykazano w momencie eliminacji jednego komentarza wirusowego dotyczącego zarządzania polskim górnictwem. W tym przypadku osiągnięto dużo lepsze wyniki niż przy pierwszej analizie – korelacja dla sentymentów pozytywnych wyniosła -0,45, negatywnych -0,65, a neutralnych -0,78.



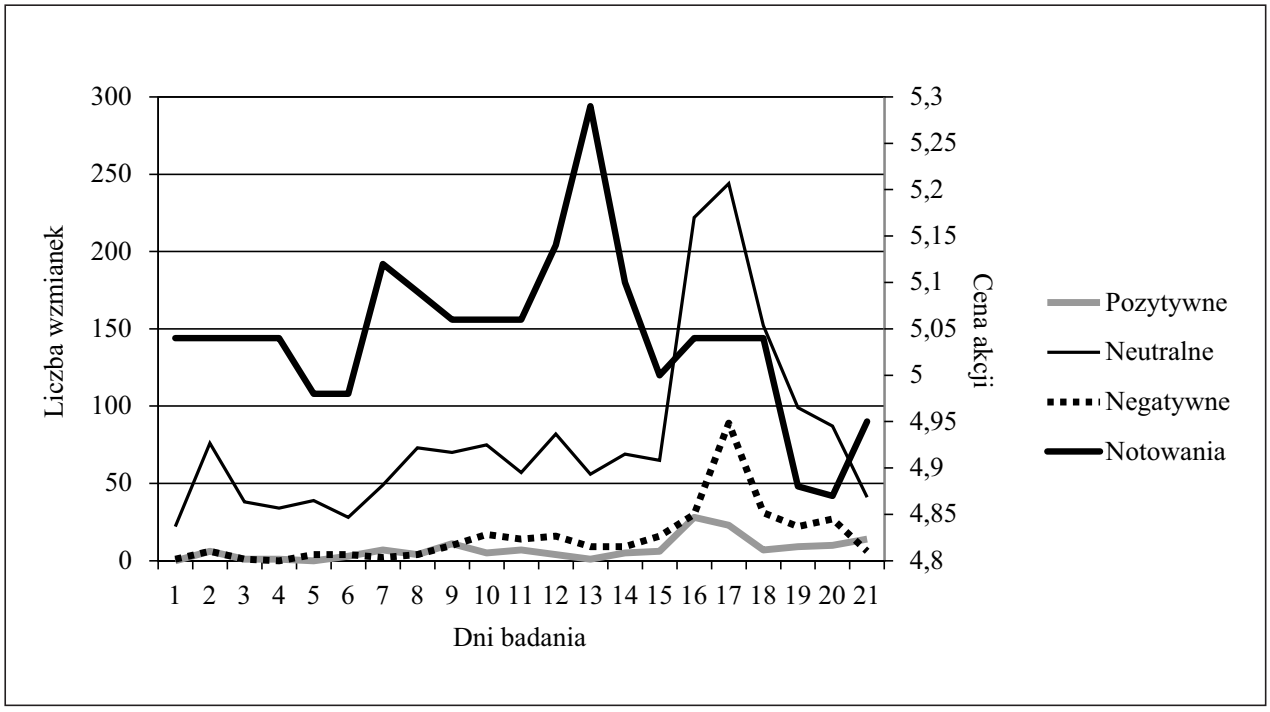

Wykres 1. Porównanie notowań spółki KGHM i dotyczących jej sentymentów z całego badanego okresu (1 stycznia–25 stycznia 2015)

Źródło: badanie własne

***Tauron***

W badanym okresie pojawiło się 2148 wypowiedzi na temat spółki Tauron. Najwięcej komentarzy miało charakter neutralny – stanowiły aż 78,12 proc. wszystkich opinii. Negatywne sentymenty wyrażało 14,80 proc. wszystkich wpisów, zaś pozytywne – 7,08 proc. Badanie wskazuje, że w przypadku spółki Tauron nie zaszła tak wysoka korelacja między jej notowaniami a wypowiedziami na jej temat, jak w przypadku KGHM. Korelacja dla sentymentów pozytywnych wyniosła -0,31, dla negatywnych -0,41, zaś dla neutralnych -0,11. Prawdopodobnie było to spowodowane niezwykle dużą liczbą danych nieodnoszących się bezpośrednio do sytuacji finansowej firmy czy jej pozycji na giełdzie. Bardzo wysoki odsetek wpisów odnosił się do kondycji firmy i jakości usług świadczonych przez przedsiębiorstwo – okazało się, że nie wywierają one zbyt dużego wpływu na wartość akcji Tauronu, zaburzają wręcz możliwość ich predykcji. Szczególnie neutralne komentarze, których odnotowano najwięcej, zawierające w większości informacje właśnie na temat poziomu satysfakcji z oferowanych usług, nie wykazują znaczącej korelacji z cenami udziałów spółki. W badanym okresie nie pojawiło się także zbyt wiele opinii inwestorów na temat przedsiębiorstwa, co również mogło znacząco wpłynąć na wynik badania.

Wykresy sentymentów negatywnych i pozytywnych pokazują, że w niemal całym badanym okresie ich liczby były na podobnym poziomie, dopiero 10 stycznia wypowiedzi o charakterze negatywnym zaczęły liczbowo przewyższać te pozytywne. Ich gwałtowny wzrost odnotowano 17 stycznia, kiedy nastąpił spadek wartości akcji Tauronu. Tendencja spadkowa zaczęła się w tym przypadku 13 stycznia, nie można więc mówić o predykcji tego trendu, chyba że uzna się za słuszną interpretację Dhara, że negatywne komentarze inwestorów są poprzedzone przez sprzedawanie przez nich akcji, przez co stają się zapowiedzią odbicia – niskie ceny akcji mobilizują inwestorów do ich kupna.




Wykres 2. Porównanie notowań spółki Tauron i dotyczących jej sentymentów z całego badanego okresu (1 stycznia–25 stycznia 2015)

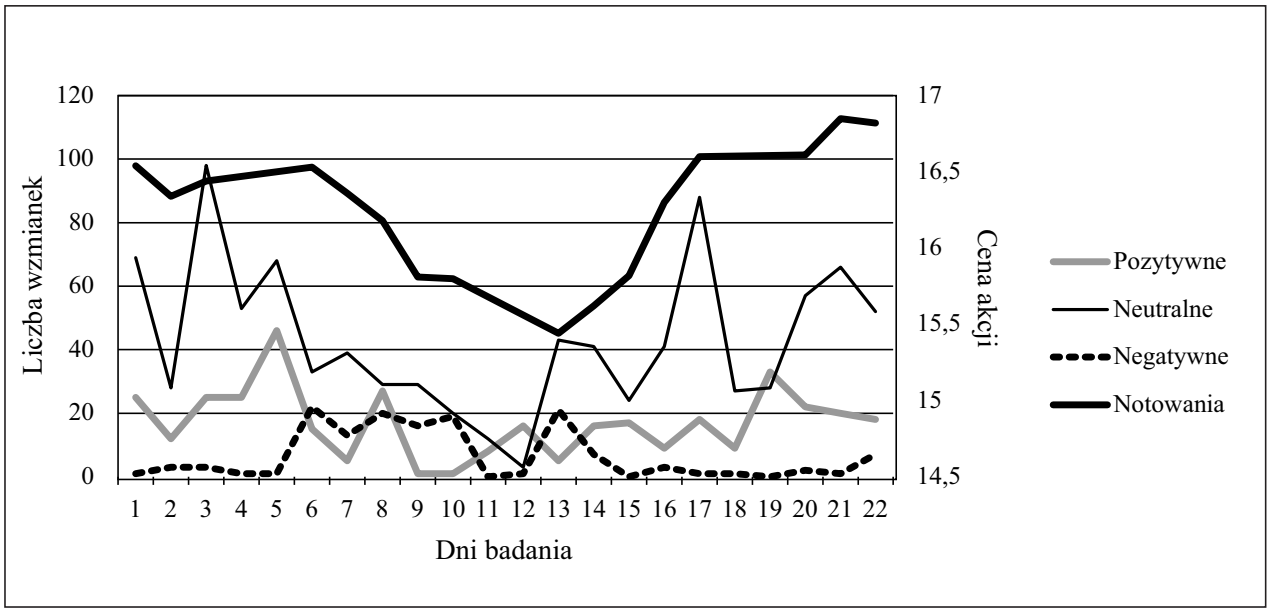
Źródło: badanie własne

***Enea***

W celu ustalenia sentymentu wypowiedzi dotyczących spółki Enea poddano analizie 1483 wypowiedzi. Badanie przeprowadzono w dniach 4 marca–25 marca 2015. Podsumowując, badanie porównawcze analizy sentymentu oraz notowań spółki Enea, pozwoliło zaobserwować pewne przesłanki, które mogą świadczyć o predykcyjnym charakterze komentarzy inwestorów. Wzrost liczby pozytywnych komentarzy w dniach 7–8 marca poprzedził lekki wzrost akcji 9 marca. Gwałtowny spadek pozytywnych wypowiedzi (z 46 do 15) i wzrost negatywnych (z 1 do 22) w dniu 9 marca, kiedy akcje osiągnęły maksymalną wycenę, mógł sugerować zmianę nastawienia inwestorów spodziewających się spadków. Rzeczywiście, 10 marca nastąpiło załamanie cen akcji, które trwało do 16 marca, kiedy akcje osiągnęły najniższy objęty badaniem poziom 15,44 zł. Jednak w niedzielę 15 marca w internecie znalazło się 16 pozytywnych komentarzy, co mogło sugerować zmianę trendu. Od 17 marca kurs spółki zaczął nieprzerwanie rosnąć, aż do końca badania, osiągając 25 marca poziom 16,82 zł. Temu wzrostowi towarzyszył stale pozytywny wydźwięk komentarzy inwestorów.

Współczynnik korelacji Pearsona dla całego badanego okresu wyniósł odpowiednio: 0,54 dla komentarzy pozytywnych, 0,54 dla komentarzy neutralnych oraz -0,45 dla komentarzy negatywnych.

Wartość współczynnika Pearsona okazała się znacznie wyższa dla całego badanego okresu niż dla poszczególnych jego części. Może to świadczyć o większej skuteczności analizy sentymentu dla dłuższego okresu. Inną możliwą przyczyną była większa liczba komentarzy, co pozwoliło na precyzyjniejszą predykcję notowań spółki.




Wykres 3. Współzależność pomiędzy sentymentem wypowiedzi a notowaniami spółki Enea SA w okresie 4 marca– 25 marca 2015

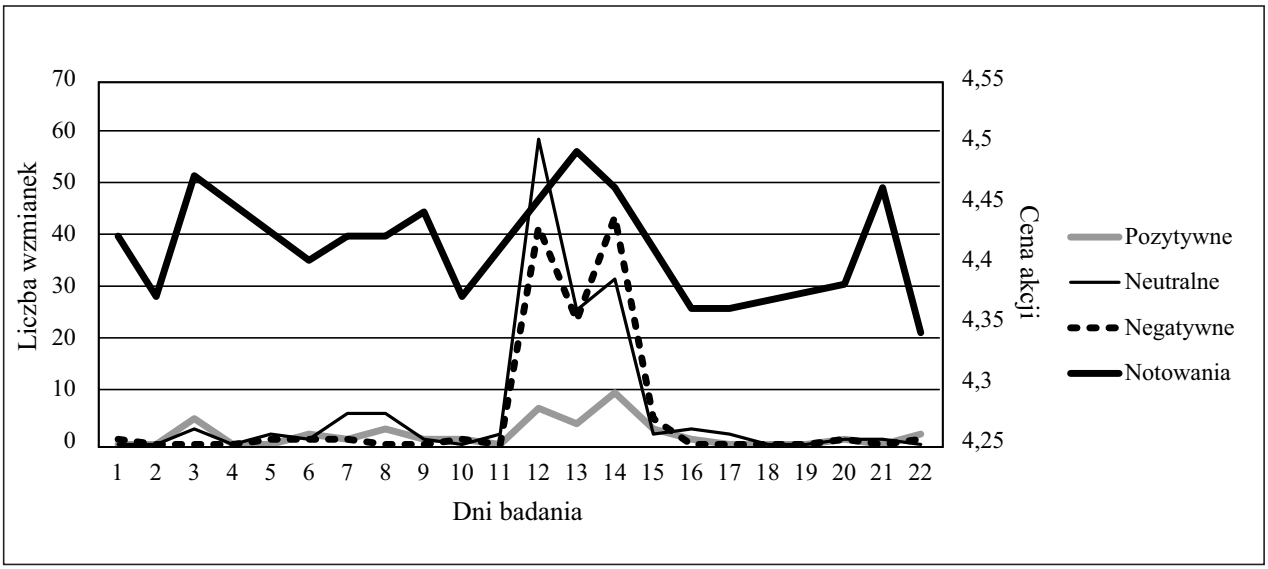
Źródło: badanie własne

***Synthos***

Badanie przeprowadzono w dniach 4 marca–25 marca 2015. W celu ustalenia sentymentu wypowiedzi dotyczących spółki Synthos analizie poddano 310 wypowiedzi. Odnotowano 41 komentarzy o wydźwięku pozytywnym, 122 – o wydźwięku negatywnym oraz 147 neutralnych.

Dynamikę wydźwięku sentymentu analizowanych wypowiedzi przedstawia wykres 4. Widoczny był niewielki wzrost liczby pozytywnych komentarzy w dniach 6, 15 i 17 marca. W ślad za nim podążał wzrost cen akcji spółki w dniach 6 i 16 marca. Jednak bardzo wyraźne zwiększenie liczby komentarzy neutralnych oraz negatywnych zaobserwowano w dniach 15–17 marca. Wiązało się to z faktem ujawnienia nagrań z tzw. „afery podsłuchowej". Informacje zawarte w tych materiałach zawierały nazwisko jednego z członków zarządu spółki. Giełda zareagowała na te informacje gwałtownym spadkiem wyceny akcji Synthos z poziomu 4,49 zł w dniu 16 marca do poziomu 4,36 zł w dniu 19 marca.

Współczynnik korelacji Pearsona wyniósł odpowiednio: 0,51 dla wypowiedzi pozytywnych, 0,58 dla wypowiedzi neutralnych oraz 0,49 dla wypowiedzi negatywnych. Podsumowując przeprowadzone badanie, na podstawie wartości współczynnika liczby Pearsona można uznać, że zaszła silna korelacja pomiędzy liczbą komentarzy a notowaniami spółki. Najsilniejsza zależność wystąpiła pomiędzy liczbą komentarzy neutralnych a notowaniami – współczynnik Pearsona wyniósł 0,58. Dla wypowiedzi pozytywnych było to 0,51, dla wypowiedzi negatywnych – 0,49. Uwzględniając niewielką liczbę materiałów wykorzystanych do analizy, jej krótki przebieg oraz eksperymentalną wersję narzędzi badawczych, można uznać, że korelacja na tym poziomie sugeruje istnienie relacji pomiędzy charakterem sentymentów a notowaniami giełdowymi spółki.




Wykres 4. Współzależność pomiędzy sentymentem wypowiedzi a notowaniami spółki Synthos SA w okresie 4 marca– 25 marca 2015

Źródło: badanie własne

## Podsumowanie

Przeprowadzone badanie wykazało, że zauważalna jest współzależność pomiędzy liczbą komentarzy w sieci o nacechowaniu emocjonalnym a notowaniami spółek. Dowodem na występowanie takiej zależności jest względnie wysoka wartość współczynnika Pearsona. Jego średnia wartość wyniosła 0,49. Należy uwzględnić czynniki wpływające negatywnie na wiarygodność badania: krótki czas analizy, brak dostępu do wypowiedzi umieszczonych na zamkniętych forach dla inwestorów i wynikają-

ca z tego stosunkowo niewielka liczba komentarzy.

Przeprowadzone badania dowodzą, że błąd predykcji zmian wartości akcji jest możliwy nawet w przypadku odpowiedniej interpretacji zbiorów Big Data. Niekiedy może ona stwarzać duże problemy przy analizie, szczególnie w przypadku prób automatycznego badania emocjonalnego zabarwienia wpisów. Przed tym dylematem badacz staje przy klasyfikacji sentymentów do poszczególnych grup (negatywne, pozytywne, neutralne). Okazuje się, że tylko niewielka część sentymentów dotyczących danego przedsiębiorstwa ma związek z wahaniami cen jego udziałów. Podczas identyfikacji sentymentów noszących znamiona wirusowej promocji należy odrzucić wątki na temat sponsorowania przez przedsiębiorstwo drużyn sportowych lub festiwali. Istotne wydają się być dane informujące o kondycji spółek, ich sukcesach i porażkach lub wspominające o związanych z nimi wydarzeniach. Pod uwagę należy brać także sentymenty inwestorów giełdowych – ich rekomendacje czy komentarze na temat bieżących notowań i stanu konsorcjum.

Oddzielnym problemem okazały się wpisy wirusowe, które odnotowano w analizie sentymentów dotyczących spółki KGHM. Zbadano, że mają one znaczący wpływ na zależność między notowaniami spółki i sentymentami. Niektóre wpisy, źródła zliczanych sentymentów stanowiących podstawę predykcji, można uznać za spam – po ich wyeliminowaniu odnotowano o wiele wyższą korelację niż wcześniej. Stanowi to o problemie doboru sentymentów związanych z notowaniami spółki.

Na podstawie przeprowadzonego badania oraz aktualnej wiedzy teoretycznej można przypuszczać, że w przyszłości możliwości predykcji notowań giełdowych na podstawie analizy sentymentu będą coraz większe. Wzrost ogólnej liczby biznesowych komentarzy w sieci, wynikający z przenoszenia coraz większej części aktywności inwestorów do internetu, oraz doskonalsze algorytmy służące ocenie i klasyfikacji wypowiedzi umożliwią tworzenie precyzyjniejszych modeli notowań giełdowych.

## Bibliografia

*Big Data, bigger digital shadows, and biggest growth in the Far East*, https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf [dostęp: 25.04.15].

Bollen J., Mao H., Zeng X., *Twitter mood predicts the stock market*, „Journal of Computational Science" 2011, Vol. 2, wyd. 1.

Carr N., *Płytki umysł. Jak Internet wpływa na nasz mózg*, Gliwice 2013.

Chen H., Chiang R., Storey V., *Business intelligence and analytics: From Big Data to big impact*, „MIS Quarterly" 2012, s. 1165–1188.

Cukier K., Mayer-Schonberger V., *Big Data. Rewolucja, która zmieni nasze myślenie*, Warszawa 2014.

Curme C. i in., *Quantifying the semantics of search behavior before stock market moves*, „PNAS" 2014, nr 32.

Dhar V., *Can Big Data machines analyze stock market sentiment?*, „Big Data" 2014, nr 4, s. 178.

Drzewiecki R., *System wyceny człowieka. Oto jak Big Data rządzi światem*, „Forsal.pl" 2014, http://forsal.pl/artykuly/785494,system-wyceny-czlowieka-oto-jak-big-data-rzadzi-swiatem.html [dostęp: 3.05.2015].

Ferguson G.A., Takane Y., *Analiza statystyczna w psychologii i pedagogice*, Warszawa 2009.

Gogołek W., Kuczma P., *Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych*, „Studia Medioznawcze" 2013, nr 2 (53).

Hall R., Taylor J., *Makroekonomia*, Warszawa 2010.

Liu Y., *Big Data and predictive business analytics*, http://eds.b.ebscohost.com.atoz.han.buw.uw.edu.pl/eds/pdfviewer/pdfviewer?sid=3149252a-86a3-462d-87c0-690b06fda804%40sessionmgr110&vid=6&hid=112 [dostęp: 09.05.15].

Martin J., *The „Big Data" solution for Wall Street*, http://iknowfirst.com/the-big-data-solution-for-wall-street [dostęp: 24.04.2015].

Martínez Bustos S. i in., *Pricing stocks with yardsticks and sentiments*, „Algorithmic Finance" 2011, nr 1.

Pisarek W., *Polskie słowa sztandarowe i ich publiczność*, Kraków 2002.

Rechenthin M., Street W.N., Srinivasan P., *Stock chatter: Using stock sentiment to predict price direction*, „Algorithmic Finance" 2013, nr 3-4s. 169-196.

Smailović J. i in., *Predictive sentiment analysis of Tweets: A stock market application*, „Human-computer interaction and knowledge discovery in complex, unstructured, Big Data" 2013, s. 77–88.

Smith K., *Big Data discoveries*, „Best's Review" 2015, nr 7.

Tabakow M., Korczak J., Franczyk B., *Big Data – definicje, wyzwania i technologie informatyczne*, „Informatyka Ekonomiczna" 2014, nr 1 (31).

*The digital universe of opportunities: Rich Data and the inreasing value of the Internet of things*, http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm [dostęp: 25.04.15].

*The Internet in real time*, http://pennystocks.la/internet-in-real-time/ [dostęp: 25.04.15].