ISSN 2353-8929
e-ISSN 2449-9757

Ewelina Niedzielska
Uniwersytet Łódzki, Wydział Zarządzania
e-mail: ewelina.niedzielska@unilodz.eu
ORCID: 0000-0002-2919-3200

WYKORZYSTANIE GOOGLE TRENDS DO PREDYKCJI STOPY ZWROTU INDEKSU WIG20

USING GOOGLE TRENDS TO PREDICT THE RATE OF RETURN OF WIG20 INDEX

DOI: 10.15611/e21.2018.3.06
JEL Classification: O16, G17, G41

Streszczenie: W artykule podjęto próbę wykorzystania danych, będących wynikiem aktywności użytkowników Internetu, do predykcji stopy zwrotu indeksu WIG20. Jako źródło tego typu danych przyjęta została baza Google Trends, która umożliwia przeglądanie i pobieranie zagregowanych wskaźników zapytań w wyszukiwarce Google. W pierwszej części artykułu autorka zwraca uwagę na to, że w konsekwencji rewolucji technologicznej zmianie uległ sposób pozyskiwania i kreowania informacji. Z punktu widzenia nauki o finansach jest to o tyle istotne, że informacja stanowi centralną oś hipotezy efektywności rynku. „Zalew informacyjny", którego jesteśmy uczestnikami, skłania do ponownego podjęcia refleksji nad możliwościami dyskontowania informacji przez rynek i jego graczy. W tym kontekście autorka kieruje się ku paradygmatowi finansów behawioralnych. Badacze z tego nurtu zwracają bowiem szczególną uwagę na ograniczone zdolności poznawcze człowieka. Ich zdaniem nie jest możliwe, aby inwestorzy byli w stanie zabsorbować całą dostępną na temat zdarzeń rynkowych wiedzę. Taki tok rozumowania skłania do zadania pytań o to, w jaki sposób inwestorzy dokonują selekcji informacji, co w danym momencie może znajdować się centrum ich zainteresowania oraz jakie może mieć to konsekwencje dla wyników rynkowych. W części empirycznej eksploracji podlegały dwie kwestie. Po pierwsze, sprawdzone zostało występowanie korelacji między zapytaniami użytkowników wyszukiwarki Google a zamknięciem indeksu WIG20. Spośród niemal 30 haseł, tematycznie związanych z giełdą i finansami, wyselekcjonowano siedem, które charakteryzowały się najwyższymi współczynnikami korelacji. Po drugie, zbadana została możliwość wykorzystania tego typu danych w celach predykcyjnych. W tym kontekście zastosowane zostały dwa algorytmy klasyfikacyjne uczenia maszynowego: regresja logarytmiczna oraz naiwny klasyfikator Bayesa. Badania przeprowadzano w trzech próbach. Pierwsza liczyła 113 tygodni. Jej celem było sprawdzenie zdolności predykcyjnych wskaźników wyszukiwań przy założeniu, że ich wartości przeliczane były w tym samym tygodniu co stopa zwrotu indeksu. Druga próba liczyła 112 tygodni i uwzględniała różnicę wyszukiwań w ramach tygodniowego opóźnienia. Trzecia próba składała się ze 111 tygodni i uwzględniała różnicę wyszukiwań z dwutygodniowym opóźnieniem. Szczególnie wysokimi wartościami predykcyjnymi charakteryzował się klasyfikator Bayesa w trzeciej próbie. Wyniki badań stanowić mogą przesłankę do stwierdzenia,

że dane pochodzące z Google Trends niosą ze sobą walor predykcyjny w kontekście rodzimego rynku kapitałowego. Dokonany przez autorkę przegląd literatury może wskazywać na to, że poruszana w artykule problematyka charakteryzuje się aktualnością, zarówno w rozumieniu wspomnianych zmian technologicznych, jak i światowego dorobku badawczego. Jednocześnie, zgodnie z wiedzą autorki, na gruncie polskiej nauki nie podejmowano jak dotąd prób stosowania uczenia maszynowego do predykcji finansowych przy wykorzystaniu danych pochodzących z wyszukiwarek internetowych.

Słowa kluczowe: giełda, uczenie maszynowe, Google Trends, finanse behawioralne.

Summary: The article attempts to use data resulting from Internet users' activity to predict the rate of return of the WIG20 index. As a source of this type of data, the Google Trends database has been adopted. This tool allows to view and download data which are the result of searching information on Google. In the first part of the article, the author points out that as a consequence of the technological revolution, the way of acquiring and creating information has changed. From the financial science point of view, it is important because information is the central axis of the market efficiency hypothesis. The "Information lagoon", which we are participants in, prompts us to rethink the abilities of market and its players to discount all knowledge of market events. In this context, the author turns to behavioral finance paradigm. Researchers from this paradigm pay special attention to limited human cognitive abilities. In their opinion, it is not possible for investors to be able to absorb all available knowledge about market events. Such reasoning leads to the question of how investors select information, what the center of their interest may be at a given moment and what the consequences for market results may have. Two issues were covered in the empirical part of this article. First, correlation between the occurrence of queries Google search users, and the closure of the WIG20 index has been checked. Out of almost 30 entries, thematically related to the stock exchange and finance, seven were selected, which were characterized by the highest correlation coefficients. Secondly, the possibility of using this type of data for prediction purposes was examined. In this context, two classification algorithms of machine learning were used: logarithmic regression and the naive Bayes classifier. The tests were carried out in three trials. The first consisted of 113 weeks. Its aim was to test the ability of searches indicators to predict the rate of return, assuming that their values were calculated in the same week as the rate return of index. The second attempt consisted of 112 weeks and included searches difference within a week delay. The third trial consisted of a sample of 111 weeks and included the search difference with a two-week delay. Particularly high values of the predictive characterized Bayes classifier in the third sample. The research results may constitute a premise to conclude that data from Google Trends bring with them a predictive value in the context of the domestic capital market. The literature review carried out by the author may indicate that the issue discussed in the article is characterized by timeliness, both in the understanding of the technological changes mentioned above as well as the global research achievements. At the same time, according to the author's knowledge, no attempts have been made so far on the basis of Polish science to use machine learning for financial predictions using data from search engines.

Keywords: stock exchange, machine learning, Google Trends, behavioral finance.

1. Wstęp

Specyficzna dla okresu, w jakim żyjemy, relacja sprzężenia zwrotnego między człowiekiem a technologią jest efektem użytkowania globalnej sieci w taki sposób, że miliony ludzi stają się nie tylko odbiorcami informacji, ale również jej twórcami. W wyniku aktywności internetowej każdego dnia kreujemy blisko 2,5 tryliona bajtów danych [Connors i in. 2013]. Dostrzeżenie ich znaczenia, to jest wydobycie z nich wiedzy na temat stojących za nimi zachowań, zmusza badaczy do zmagania się z niespotykaną dotąd ilością heterogenicznych i częstokroć nieustrukturyzowanych danych [Blazquez, Domenech 2017, s. 1]. Rozwój technologii zaowocował jednak nie tylko powstaniem, ale również dostępem do nowego typu zasobów, ponadto z naukowego punktu widzenia stwarza rozliczne możliwości narzucania perspektyw analizowania procesów społecznych i ekonomicznych, które dotąd mogły być trudne do dostrzeżenia.

W przedstawianym artykule za źródło opisanych powyżej danych przyjęta została aktywność użytkowników wyszukiwarki Google. W związku z powyższym celem opracowania jest poddanie pod rozwagę przesłanki mówiącej o możliwości utożsamienia liczby zapytań użytkowników Internetu, związanych z rynkiem finansowym, z poziomem uwagi inwestorów oraz wskazanie na potencjał badawczy danych niefinansowych w kontekście predykcji stóp zwrotu na rynku kapitałowym. Biorąc pod uwagę rozwój nurtu finansów behawioralnych, które wskazują na konieczność włączenia do nauk ekonomicznych czynników psychologicznych, podjęcie refleksji nad tym, skąd czerpać wiedzę dotyczącą tego, co przykuwa uwagę graczy rynkowych, staje się niebywale aktualnym problemem. W dalszej części tekstu przedstawione zostały wyniki autorskich badań, których celem była odpowiedź na pytanie o to, czy tego typu dane niosą ze sobą walor predykcyjny dla rodzimego rynku kapitałowego. Do tego celu posłużyły dwa algorytmy technik uczenia maszynowego. Zgodnie z wiedzą autorki jak dotąd na polskim gruncie nie prowadzono badań w tym zakresie, co stanowi wartość dodaną przedstawianych analiz.

2. Przegląd literatury

Biorąc zatem pod uwagę, że żyjemy w „epoce zalewu informacyjnego" [Cheikh i in. 2015, s. 78], należy stwierdzić, że taki stan rzeczy motywuje do ponownego poddania namysłowi hipotezy efektywności rynku, której *spiritus movens* stanowiła właśnie informacja. Należy zaznaczyć, że obecnie wciąż trudno pokusić się o stwierdzenie darmowości przekazu i wiedzy na temat zdarzeń rynkowych. Mimo to, na tym etapie rozwoju cywilizacyjnego istnieje paląca potrzeba odpowiedzi na pytanie o to, jaki wpływ na efektywność rynków ma zwiększająca się dostępność i wszechobecność informacji.

W latach 80. badacze Grossman i Stiglitz [1980], konstatując logikę wywodu Famy, wskazywali, że większa dostępność informacji skutkować będzie głębszym

zakorzenieniem ceny instrumentów w aktualnej wiedzy na temat rynku, a tym samym doprowadzi do jego wyższej efektywności. Rysę na tym twierdzeniu kreśli jednak problem dyskontowania tychże informacji lub, ujmując tę kwestię precyzyjniej, poznawczych zdolności człowieka w ich przyswajaniu i przetwarzaniu. Kontrargumentów do propozycji Grossmana i Stiglitza, uwzględniających kognitywny aspekt gry rynkowej, doszukiwać można się w nurcie finansów behawioralnych, w ramach których zwiększająca się ilość dostępnych inwestorom zasobów wiedzy rozumiana może być jako podłoże do potencjalnego chaosu informacyjnego, który z kolei przyczynić może się do redukcji efektywności rynku [Hu 2018, s. 188]. Przegląd literatury przedmiotu wskazuje na istnienie bogatego dorobku naukowego, analizującego problem wpływu ograniczonych zdolności poznawczych człowieka na jego decyzje związane z działalnością inwestycyjną. Przyjęcie założeń tego nurtu prowokuje do postawienia kilku pytań. Po pierwsze, jeśli nie jest możliwe przetworzenie przez nasz umysł wszystkich dostępnych informacji, to na jakich zasadach dokonujemy ich kategoryzacji na istotne i nieistotne? Po drugie, w jaki sposób selekcję „szybkiego myślenia”[1] poddawać analizie? Innymi słowy, jak wskazywać i dokonywać kwantyfikacji uwagi graczy rynkowych? O ile odpowiedź na pierwsze pytanie zdaje się znajdować w obszarze zainteresowania naukowców z dziedziny psychologii, o tyle pytanie drugie powinno zajmować również badaczy z zakresu finansów. Jak wskazują Da i zespół, otrzymywana przez nas wiedza na temat zachowań inwestorów ma z reguły charakter pośredni. Częstokroć bowiem analiza dynamiki rynków kapitałowych bazuje na standardowych danych, tj. cenie otwarcia, zamknięcia czy też wolumenie obrotu. Tymczasem tego typu wartości są wtórne wobec działań, jakie podejmują gracze, które to działania są pochodną ich uwagi. W ten sposób w centrum rozważań postawić można pytanie o to, co i w jakim stopniu w określonym czasie stanowi przedmiot uwagi uczestników rynku [Da i in. 2011, s. 1497].

Podążając za sugestią wyżej wskazanych badaczy, autorka korzysta w tym aspekcie ze źródła, jakim jest Google Trends (GT), czyli narzędzia umożliwiającego obserwację i pobieranie danych na temat zapytań użytkowników Internetu w wyszukiwarce Google. Przedstawione poniżej badania mogą bowiem wskazywać na istnienie związku między aktywnością użytkowników Internetu a wynikami rynkowymi. Przesłankę do takiej diagnozy stanowić może fakt, iż wyszukanie informacji jest typem działania podejmowanego celowo. W związku z tym pokusić się można o stwierdzenie, że informacja, będąca przedmiotem takiego działania, znajduje się w polu zainteresowania osoby wyszukującej. Taki tok rozumowania motywować może zatem do podjęcia badań na temat tego, czy GT stanowi istotne źródło informacji na temat uwagi inwestorów (por. [Da i in. 2011, s. 1462]).

Dokonując przeglądu literatury na temat wykorzystania danych pochodzących z GT do działań predykcyjnych, należy zwrócić uwagę na wielość dziedzin, w ra-

[1] Autorem sformułowania „szybkie myślenie” jest D. Kahneman.

mach których badaczom udaje się znaleźć dla nich zastosowanie. Najpopularniejsze przykłady zdają się pochodzić z nauk medycznych. Prekursorskie badania Ginsberga i zespołu pozwoliły zauważyć, że śledzenie zapytań w wyszukiwarce Google daje możliwość wskazywania populacji, w której panuje grypa [Ginsberg i in. 2009]. Co więcej, wykorzystanie takich danych pozwalało na wykrywanie rozprzestrzeniającej się grypy szybciej, niż robiły to Centra Kontroli i Prewencji Chorób[2] [Jun i in. 2017, s.1]. Sami zresztą programiści Google'a udostępniali przez pewien czas narzędzia służące śledzeniu choroby, oparte na wzorcach wyszukiwań symptomów[3].

Korzyści, jakie niesie ze sobą dostęp do danych GT, dostrzeżone zostały również w biznesie i ekonomii. Badania Choi i Variana [(red.) 2009] dowodzą, że wskaźniki wyszukiwań internetowych niosą ze sobą wartość dodaną w kontekście przewidywania poziomów sprzedaży różnego typu dóbr. Przykładowo wykazali oni, że wzrost zapytań związanych z branżą motoryzacyjną o 1% wiąże się ze wzrostem detalicznej sprzedaży samochodów marki Ford w USA o 0,5%. Ci sami badacze podjęli podobne próby na przykładzie rynku nieruchomości oraz sprzedaży detalicznej.

Jednym z zagadnień związanych z wykorzystaniem GT, które w tematyce ekonomicznej wiedzie prym, jest problem bezrobocia. Próby sprawdzenia korelacji między wyszukiwaniem haseł związanych z aktywnością zawodową a tym zjawiskiem przeprowadzone zostały w USA, Francji i Włoszech. W każdym z badań wykazano predykcyjny walor tego typu danych. Fondeur i Karamé [2017] (Francja) doszli do wniosku, że włączanie do badań danych GT wzmocniło predykcję bieżącego bezrobocia populacji osób w wieku 15-24 lata w sensie zarówno poziomu prognozy, jak i jej bezbłędności. Do podobnych wniosków doszli badacze z Włoch. Wyniki ich analiz potwierdziły istnienie związku między aktywnością internetową a poziomem bezrobocia wśród osób młodych. W konkluzji stwierdzili oni, że GT może być rozumiane jako pomocnicze źródło danych, służące prognozowaniu bezrobocia w krótkim i średnim ujęciu czasowym [Naccarato i in. 2017, s. 7]. Z kolei analogiczne badania przeprowadzone na amerykańskim rynku pracy wykazały użyteczność GT w prognozach o średnim i długim horyzoncie [D'Amuri, Marcucci 2017][4].

Zainteresowanie wykorzystaniem danych z wyszukiwań internetowych zauważyć można również na przykładzie rynku surowców. W 2017 r. Yao wraz z zespołem sprawdzili wpływ uwagi inwestorów na ceny ropy naftowej. Wykorzystując GT jako kwantyfikator tejże, oszacowali wartość dodaną zagregowanych wskaźników wyszukiwań dla predykcji cen ropy naftowej na 15,2%. Wyniki ich badań wskazują na istnienie silnego negatywnego wpływu uwagi inwestorów na ceny ropy, który zauważalny jest w dwumiesięcznym odstępie i trwa około dziewięciu miesięcy. Wartością dodaną ich badań było stworzenie zagregowanego wskaźnika GSVI (*Google Search Volume Index*), na który składały się wolumeny wyszukiwań słów związa-

[2] Agencje federalne USA.

[3] Google Flu Trends (GFT).

[4] Badania prowadzone były przy zastosowaniu *Google Job Index*.

nych ze zwrotem „cena ropy naftowej" [Yao i in. 2017]. Przeprowadzone rok później przez Yu i zespół badania nad podobnym problemem potwierdziły możliwość określenia GT jako efektywnego predykatora zapotrzebowania na ten surowiec. Każdy z proponowanych przez nich modeli zwiększał swoją zdolność predykcyjną, ilekroć dodawano do niego zmienne pochodzące z wyszukiwarki internetowej [Yu i in. 2018].

Badacze zainteresowani rynkami kapitałowymi również dostrzegają potencjał, jaki mogą ze sobą nieść dane GT. Da i in. badali relację między wyszukiwaniem takich słów, jak: recesja, bankructwo, bezrobocie itp., a cenami akcji. Efektem ich prac było stworzenie indeksu FEARS[5]. Wysoki poziom tego wskaźnika skorelowany był z niższymi zwrotami w ujęciu dziennym przy wysokich zwrotach w odstępie kilku dni. Jednocześnie rosnący poziom indeksu wiązał się ze wzrostem wahań rynkowych oraz zmniejszaniem kapitału rynku akcji na rzecz rynku obligacji [Da i in. 2013]. Joseph z zespołem [2011] analizowali z kolei zdolność danych, będących wynikiem wyszukiwania tickerów spółek giełdowych (S&P 500), do przewidywania ponadprzeciętnych stóp zwrotu z akcji. Wyniki ich badania wykazały, że ponadprzeciętne zwroty oraz wolumen obrotu w ujęciu tygodniowym poprzedzane są zmianami intensywności wyszukiwań w Google. Korelacje między intensywnością wyszukiwań a wolumenem obrotu akcjami potwierdził również w 2010 r. Preis z zespołem [Bijl i in. 2016, za Preis i in. 2010], trzy lata później wykorzystując GT do zbudowania portfela inwestycyjnego opartego na wartościach wyszukiwań konkretnych słów. Tak skonstruowany koszyk instrumentów pozwolił im w ciągu 7 lat przekroczyć stopę zwrotu indeksu o ponad 300%. Badacze ci zasugerowali, że wzorce wyszukiwań potraktować można jako „narzędzia wczesnego ostrzegania" zdarzeń rynkowych [Preis i in. 2013]. Analogicznie Kristoufek [2013] wykorzystał wolumen wyszukiwań w ramach GT do stworzenia strategii dywersyfikacyjnej portfela. Wychodząc z założenia, że poziom indeksu wyszukiwań jest dodatnio skorelowany z ryzykiem inwestycyjnym, eliminował z portfela akcje o wysokim wolumenie zapytań internetowych. Stworzony w ten sposób portfel pozwolił na uzyskanie od 109% (jeśli przeważanie portfela następowało w tym samym tygodniu co obliczanie zysków) do 163% skumulowanego zwrotu (jeśli obliczanie zysków następowało z tygodniowym opóźnieniem). W 2014 r. dane pochodzące z GT wykorzystane zostały do przeprowadzenia badań na rynku japońskim [Takeda i Wakao 2014]. Częstotliwość wyszukiwań mierzona była liczbą zapytań o nazwy spółek (indeks Nikkei 225). Badacze znaleźli silną pozytywną korelację, która występowała między wartością wyszukiwań a wolumenem obrotów, oraz słabą pozytywną korelację ze zwrotami z akcji.

Należy jednak zwrócić uwagę, że początkowy optymizm w wykorzystywaniu danych GT do predykcji ustąpił miejsca bardziej krytycznemu podejściu. Nie wszystkie narzędzia i aplikacje zbudowane na kanwie tego podejścia okazały się trwałe. W roku 2009 zweryfikowano użyteczność narzędzia GFT, które nie wykryło

[5] *Financial and Economic Attitudes Revealed by Search Index.*

pandemii świńskiej grypy. Co więcej, z czasem jego wartości predykcyjne okazywały się znacznie przeszacowane wobec wskazań CDS. W efekcie projekt został zamknięty. Jednocześnie nie wszystkie badania potwierdzają predykcyjne walory GT. Przykładowo analizy Preis i zespołu z 2010 roku nie wskazały żadnych istotnych korelacji między wyszukiwaniem nazw przedsiębiorstw w wyszukiwarce Google a zwrotami na ich akcjach. Przytaczane badania z rynku japońskiego wykazały z kolei jedynie słabą korelację między intensywnością wyszukiwania a zwrotami na akcjach. Co więcej, wyniki eksperymentu przeprowadzonego w 2013 r. przez Challet i Ahmed pokazały, że efekty strategii inwestycyjnej opartej na indeksie wyszukiwań słów związanych z finansami nie przekraczają efektów strategii opartej na kompletnie przypadkowych zwrotach [Bijl i in. 2016, za Challet, Ahmed 2013]. Mimo to, zdaniem autorki, różnorodność dziedzin i wielość badań, które podjęte zostały w celu sprawdzenia przydatności danych GT do celów predykcyjnych, stanowi jedną z przesłanek do weryfikacji tej hipotezy na rodzimym rynku kapitałowym.

3. Metodyka badań

Wartościowość opisywanych danych zdeterminowana jest możliwością zdobycia na ich podstawie wiedzy o rzeczywistości. Jak zostało już wspomniane, konieczne do tego procesu jest zaangażowanie narzędzi i metod, które nie są tradycyjnym aparatem badawczym naukowców z dziedzin ekonomicznych i społecznych (por. [Blazquez, Domenech 2017, s. 13]). W poniższym badaniu wykorzystane zostało podejście zbiorczo określane mianem uczenia maszynowego (ML). W ogólnym ujęciu tego typu narzędzia wykorzystywane są do rozwiązywania problemów trojakiego rodzaju: generacji (tj. tworzenia treści), klasyfikacji oraz predykcji [McKinsey Global Institute 2016, s. 12-13]. Celem niniejszej analizy była odpowiedź na pytanie o predykcyjny walor danych GT w kontekście GPW na przykładzie indeksu WIG20. W badaniu wykorzystane zostały dwa popularne algorytmy klasyfikacyjne maszyn uczących się: regresja logistyczna (RLog) oraz naiwny klasyfikator Bayesa (NB). W przypadku obu algorytmów zastosowane zostało uczenie nadzorowane. Oznacza to, że próby podzielone zostały na część treningową, pozwalającą „nauczyć się" modelom oczekiwanych wartości wyjściowych (y) na podstawie zapodanych wartości wejściowych (X), i testową, na której sprawdzana była zdolność modelu do wskazania nieznanych wcześniej wartości y na podstawie wartości X niepochodzących ze zbioru treningowego.

Zarówno dane finansowe, jak i dane pochodzące z GT składały się ze 113 obserwacji. Dane pobierane były w okresach tygodniowych. Zakres czasowy, jak również gęstość danych podyktowane były możliwościami stwarzanymi przez bazę źródłową[6]. W celu zapewnienia porównywalności danych badaniu podlegał okres od

[6] Google Trends umożliwia pobieranie danych z okresów dłuższych niż 90 dni jedynie w okresach tygodniowych. Jednocześnie od 1 stycznia 2016 r. programiści GT zmienili system gromadzenia danych.

1 stycznia 2016 r. do 25 lutego 2018 r. Statystyki opisowe dla danych finansowych zaprezentowane są w tab. 1, natomiast statystyki opisowe dla danych GT w tab. 3.

Udostępniane przez GT dane przedstawiane są jako względna popularność wyszukiwania, co oznacza, że każdy punkt na osi dzielony jest przez ogólną liczbę wyszukiwań w badanym obszarze geograficznym i zakresie czasowym. Narzędzie automatycznie dokonuje skalowania w ten sposób, że wszystkie dane przybierają wartości od 0 do 100 (przy czym wartości te mogą się powtarzać w ramach jednego zbioru danych), gdzie 0 oznacza najmniejszą, a 100 największą popularność wyszukiwań [Google Trends 2018].

Tabela 1. Statystyki opisowe danych indeksu WIG20

Wyszczególnienie	Zamknięcie	Stopa zwrotu
Próba	113	113
Średnia	2101,29	–0,0063
Odchylenie standardowe	294,40	0,0970
Min.	1705,43	–0,0721
25%	1796,75	–0,0131
50%	2076,12	0,0022
75%	2380,16	0,0163
Max.	2601,82	0,0652

Źródło: opracowanie własne.

W pierwszej fazie badania pod uwagę brane było 30 haseł tematycznie związanych z giełdą, które zdaniem autorki mogły być przedmiotem wyszukiwania osób zainteresowanych notowaniami na GPW[7]. Warto zwrócić uwagę, że przy tego typu badaniach trudno wskazać jednoznaczną metodę gromadzenia danych. Korzystanie z wyszukiwarki internetowej nie jest bowiem działaniem sformalizowanym w tym sensie, że używane w tym celu zapytania nie muszą zamykać się w zbiorze oficjalnej nomenklatury. Częstokroć wpisywanie zwrotów potocznych, skrótowych bądź symbolicznych może dać użytkownikom lepsze (tj. bardziej adekwatne do ich potrzeb informacyjnych) rezultaty. Dlatego też poszukiwanie haseł i tematów, które pozwoliłyby wskazać kierunek uwagi inwestorów, wiąże się z wieloma nieformalnymi zabiegami, takimi jak na przykład wstawianie bądź usuwanie znaków interpunkcyjnych. Przykładem tego typu kwestii jest wykorzystane w badaniu hasło „WIG40" oraz „WIG 40".

W celu zdiagnozowania zmiennych niezależnych mogących charakteryzować się walorem predykcyjnym, analizie poddana została korelacja występująca między wartościami zamknięcia indeksu WIG20 a poszczególnymi hasłami. Do tego celu

[7] Pod uwagę wzięty został jedynie obszar Polski.

wykorzystano współczynnik korelacji Pearsona. W dalszej części badania analizie podlegały jedynie te hasła, dla których współczynnik wynosił 0,3 lub więcej i był istotny statystycznie. W ten sposób liczba analizowanych wyszukiwań zmniejszyła się do siedmiu (tab. 2).

Tabela 2. Tabela korelacji

Wyszczególnienie	Współczynnik korelacji	*p value*
WIG20 (hasło)	0,55	0,0000
Notowania (hasło)	0,49	0,0000
WIG20 (indeks)	0,46	0,0000
Akcje (hasło)	0,42	0,0000
Notowania giełdowe (hasło)	0,41	0,0000
WIG40 (hasło)	0,39	0,0000
WIG 40 (hasło)	0,37	0,0000

Źródło: opracowanie własne.

Analiza korelacji wskazała, że wśród branych pod uwagę haseł najwyższym współczynnikiem charakteryzowało się hasło „WIG20", najniższym zaś „WIG 40". Zapytania dla haseł „WIG30" oraz „WIG 30" cechowały się bardzo niskim współczynnikiem korelacji z danymi zamknięcia WIG20, przy jednoczesnym niespełnieniu warunku istotności statystycznej.

Tabela 3. Statystyki opisowe danych GT

Wyszczególnienie	Średnia	std	min	25%	50%	75%	max
Notowania (hasło)	68,35	9,12	45,00	63,00	69,00	74,00	100,00
Notowania giełdowe (hasło)	60,48	15,66	22,00	49,00	62,00	72,00	100,00
Akcje (hasło)	66,34	13,65	41,00	53,00	68,00	77,00	100,00
WIG20 (indeks)	43,19	12,34	24,00	34,00	42,00	51,00	100,00
WIG20 (hasło)	44,11	13,72	21,00	34,00	40,00	52,00	100,00
WIG40 (hasło)	21,84	22,75	0,00	0,00	17,00	36,00	100,00
WIG 40 (hasło)	17,43	20,04	0,00	0,00	18,00	22,00	100,00

Źródło: opracowanie własne.

W dalszej kolejności obliczone zostały stopy zwrotu indeksu WIG20, które zakodowane zostały jako „0" (dla tygodni, w których zwrot był ujemny) oraz „1" (dla tygodni, w których zwrot był dodatni).

4. Wyniki badań własnych

W każdym z modeli dane podzielone zostały na próbę treningową i testową, w stosunku 80% do 20%. Po wprowadzeniu do modeli danych treningowych testom podlegały ich zdolności predykcyjne, polegające na przyporządkowaniu danych testowych do odpowiedniej klasy: „0" lub „1". Poszczególne zmienne wprowadzane były do modeli metodą krokową postępującą. W ten sposób zakwalifikowane zostały zmienne prezentowane w tab. 4.

Tabela 4. Zmienne dla modeli bez przesunięcia czasowego

Model	Notowano (hasło)	Notowania giełdowe (hasło)	Akcje (hasło)	WIG20 (indeks)	WIG20 (hasło)	WIG40 (hasło)	WIG 40 (hasło)
$RLog_0$		x	x	x	x	x	x
NB_0		x			x	x	x

Źródło: opracowanie własne.

Proces selekcji wskazał na konieczność włączenia do modelu bazującego na regresji logistycznej większej liczby zmiennych niż do modelu bayesowskiego, przy czym pozwoliło to na uzyskanie jedynie nieznacznie wyższego wyniku średniego *f1-score* w stosunku do modelu NB_0. Zdolności predykcyjne obu modeli ulegały osłabieniu, ilekroć podejmowana była próba włączenia do nich zmiennej „notowania", która wykazywała relatywnie wysoki współczynnik korelacji z danymi zamknięcia indeksu WIG20.

Tabela 5. Wyniki analizy predykcyjnej modeli bez przesunięcia czasowego

Model	Klasa	*precision*	*recall*	*f1-score*	Próba
$RLog_0$	0	1,00	0,38	0,55	16
	1	0,41	1,00	0,58	7
	avg	0,82	0,57	0,56	23
NB_0	0	0,86	0,38	0,52	16
	1	0,38	0,86	0,52	7
	avg	0,71	0,52	0,52	23

precision – prawdziwe pozytywne / (prawdziwe pozytywne + fałszywe pozytywne); *recall* – prawdziwe pozytywne / (prawdziwe pozytywne + fałszywe negatywne); *f1-score* – 2 * (*precision* * *recall*) / (*precision* + *recall*)

Źródło: opracowanie własne.

Warto zwrócić uwagę, że zarówno model $RLog_0$, jak i NB_0 charakteryzowały się wysokimi zdolnościami do nieklasyfikowania dodatnich stóp zwrotu jako ujemne

oraz do przyporządkowywania dodatnich stóp zwrotu do odpowiedniej klasy. Mimo to oba odznaczały się wysokim poziomem błędu I rodzaju, co oznacza, że ujemne stopy zwrotu klasyfikowane były jako dodatnie. W efekcie w zarówno w pierwszym, jak i drugim przypadku średni wskaźnik *f1-score* osiągał wartości nieprzekraczające 60%.

Ze względu na niesatysfakcjonujący wynik wyżej opisanych modeli w kolejnym kroku rangi przypisane w ramach GT poszczególnym hasłom w ujęciu czasowym przekształcone zostały w taki sposób, aby uwzględniały dynamikę wyszukiwań w ujęciu jednego tygodnia:

$$\Delta x_{t_n} = x_{t_n} - x_{t_{n-1}}.$$

W efekcie takiego zabiegu próba uległa zmniejszeniu o jedną obserwację i wynosiła 112 tygodni. Dalszy ciąg badania przebiegał analogicznie do sposobu opisanego wyżej. W wyniku selekcji do obu modeli zakwalifikowane zostały zmienne prezentowane w tab. 6.

Tabela 6. Zmienne dla modeli z przesunięciem jednego tygodnia

Model	Notowano (hasło)	Notowania giełdowe (hasło)	Akcje (hasło)	WIG20 (indeks)	WIG20 (hasło)	WIG40 (hasło)	WIG 40 (hasło)
$RLog_1$	x	x					x
NB_1				x	x		

Źródło: opracowanie własne.

Oba modele wykazały wyższe zdolności predykcyjne przy uwzględnieniu mniejszej liczby zmiennych objaśniających. Jednocześnie tylko do modelu NB_1 zaklasyfikowane zostały zmienne „WIG20 (indeks)" oraz „WIG20 (hasło)", które charakteryzowały się względnie wysokimi współczynnikami korelacji z danymi zamknięcia indeksu. Do modelu $RLog_1$ dobrane zostały zmienne „notowania", „notowania giełdowe" oraz „WIG 40". W efekcie oba modele bazowały na różnych zmiennych niezależnych.

Tabela 7. Zdolność predykcyjna modeli z przesunięciem jednego tygodnia

Model	Klasa	*precision*	*recall*	*f1-score*	Próba
$RLog_1$	0	0,83	0,45	0,59	11
	1	0,65	0,92	0,76	12
	avg	0,74	0,70	0,68	23
NB_1	0	1,00	0,27	0,43	11
	1	0,60	1,00	0,75	12
	avg	0,79	0,65	0,60	23

Źródło: opracowanie własne.

Podobnie jak w poprzedniej próbie, modele poprawnie klasyfikowały dodatnie stopy zwrotu. Jednocześnie cechowały się niską czułością w kontekście ujemnych stóp zwrotu. W konsekwencji, chociaż zarówno dla NB_1, jak i $RLog_1$ średni *f1-score* osiągnął wyższy poziom niż w modelach nieuwzględniających dynamiki, to ponownie odznaczały się one błędem I rodzaju.

Przesłanki merytoryczne wynikające z przeglądu literatury skłoniły autorkę do podjęcia próby sprawdzenia zdolności predykcyjnych modeli w dłuższym horyzoncie, tj. dwóch tygodni. W tym celu dane pobrane z GT przekształcone zostały w następujący sposób:

$$\Delta x_{t_n} = x_{t_{n-1}} - x_{t_{n-2}}.$$

Próba podzielona została w analogiczny sposób, to jest 80% stanowiła próba treningowa, a 20% testowa, przy czym cała próba wynosiła 111 obserwacji. Dane do poszczególnych modeli również dobierane były metodą krokową postępującą (tab. 8).

Tabela 8. Zmienne dla modeli z przesunięciem czasowym dwóch tygodni

Model	Notowano (hasło)	Notowania giełdowe (hasło)	Akcje (hasło)	WIG20 (indeks)	WIG20 (hasło)	WIG40 (hasło)	WIG 40 (hasło)
$RLog_2$	x	x	x	x	x	x	x
NB_2		x		x		x	

Źródło: opracowanie własne.

Tak dobrane zmienne pozwoliły na osiągnięcie wyników predykcyjnych zaprezentowanych w tab. 9. Warto zwrócić uwagę, iż model $RLog_2$ osiągał najwyższe wartości predykcyjne przy wykorzystaniu wszystkich dostępnych zmiennych. Mimo to wykazywał się niższą zdolnością predykcyjną niż model NB_2.

Tabela 9. Zdolność predykcyjna modeli z przesunięciem dwóch tygodni

Model	Klasa	*precision*	*recall*	*f1-score*	Próba
$RLog_2$	0	0,50	0,44	0,47	9
	1	0,67	0,71	0,69	14
	avg	0,60	0,61	0,60	23
NB_2	0	0,86	0,67	0,75	9
	1	0,81	0,93	0,87	14
	avg	0,83	0,83	0,82	23

Źródło: opracowanie własne.

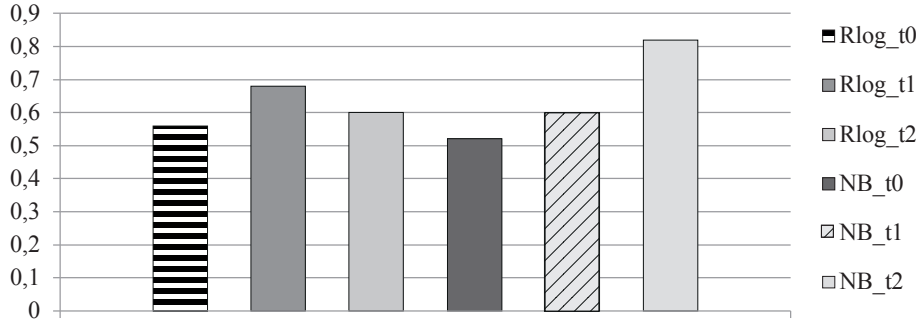
Zauważyć można, że model NB_2 prezentuje średnie ważone wskaźniki czułości i precyzji na wyższym poziomie niż model logarytmiczny, który odznaczał się wyż-

szymi błędami zarówno I, jak i II rodzaju. Taki stan rzeczy wynika z faktu, iż klasyfikator Bayesa charakteryzował się mniejszą skłonnością do kwalifikowania dodatnich stóp zwrotu jako ujemne i odwrotnie, a przez to osiągał niższy poziom błędu I rodzaju (prawidłowo zaklasyfikował 6 na 9 ujemnych stóp zwrotu). W przypadku regresji logarytmicznej wyników nie można uznać za satysfakcjonujące w tym sensie, że przesunięcie czasowe nie poprawiło jego rezultatów.

Mimo znacznie wyższego wyniku zdolności predykcyjnej klasyfikatora bayesowskiego w ostatniej próbie warto zwrócić uwagę, że regresja logarytmiczna w niektórych podejściach przynosiła podobne wyniki w kontekście średnich wskaźników. I tak średnia precyzja modelu $RLog_0$ wykazała zbliżone rezultaty do średniej precyzji modelu NB_2 (odpowiednio 0,82 i 0,83). Do podobnego wniosku można dojść, zestawiając średnią precyzję modeli z jednotygodniowym przesunięciem. Najniższą średnią precyzją charakteryzował się jednak model $RLog_2$, który z kolei w przypadku klasyfikacji bayesowskiej osiągał najlepsze rezultaty.

W przypadku obu klasyfikatorów, niezależnie od ujęcia czasowego, do zaniżania wartości przyczyniał się błąd I rodzaju. Oznacza to, że modele miały skłonność do przyporządkowywania ujemnych stóp zwrotu do klasy „1". Wprawdzie model NB_2 osiągnął w tym kontekście najwyższą wartość, ale nie przekroczyła ona 70%.

W ogólnym ujęciu przesunięcie czasowe w modelach logarytmicznych jedynie w nieznacznym stopniu poprawiało średni wskaźnik *f1-score*, który w każdej próbie wynosił blisko 60%. Odwrotna sytuacja miała miejsce w przypadku klasyfikatora bayesowskiego. Dla większości wyników przesunięcie o kolejny tydzień powodowało uzyskanie wyższych wartości każdego z analizowanych wskaźników.



Rys. 1. Wyniki średniego wskaźnika *f1-score* dla wszystkich modeli

Źródło: opracowanie własne.

Należy jednak zaznaczyć, że przedkładane badania obarczone są pewnymi ograniczeniami. Po pierwsze, w żadnej z prób liczba klas „0" oraz „1" nie była identyczna, co może rzutować na wyniki prognoz. Pod względem liczbowym klasy miały zbliżony charakter w modelach z NB_1 i $LRog_1$, jednak nie wykazały one

najwyższych wyników w kontekście całego badania. Po drugie, w badaniu przyjęte zostało tygodniowe zestawienia danych. Zdaniem autorki wartość dodaną wniosłyby również badania bazujące na danych dziennych (przy świadomości, że próba nie mogłaby być wówczas większa niż 60 obserwacji). Po trzecie, należy zwrócić również uwagę na to, że metody klasyfikacyjne, w których zastosowane zostały binarne zmienne objaśniane, nie odpowiadają na pytanie o wpływ liczby zapytań na deltę stóp zwrotu. W przyszłości warto zatem zastosować również metody bazujące na regresji.

5. Zakończenie

Trudno nie zgodzić się z tokiem rozumowania Da i zespołu [2011], zgodnie z którym informacja będąca przedmiotem wyszukiwania znajduje się w polu zainteresowania osoby wyszukującej. Dane pochodzące z wyszukiwarki internetowej mogą, zdaniem autorki, stanowić bezpośredni miernik uwagi użytkowników Internetu. Mimo to nierozstrzygnięte pozostaje, kim są owi użytkownicy. Źródło danych, jakim jest GT, nie wskazuje na to, kto w rzeczywistości jest autorem wyszukiwań. Nie istnieją również żadne szacunki, które określałyby udział wiedzy pochodzącej bezpośrednio z wyszukiwania w ogólnej liczbie dyskontowanych przez inwestorów informacji (por. [Takeda, Wakao 2014, s. 3]). Stąd nie można stwierdzić, czy zdolność modeli do klasyfikowania stóp zwrotu jest wynikiem uczenia się na zmiennych objaśniających, będących wynikiem aktywności inwestorów, czy może dowolnych użytkowników Internetu. Co za tym idzie, autorka podaje w wątpliwość stwierdzenie mówiące o tym, że wartości pochodzące z GT mogą być bezpośrednim miernikiem uwagi graczy rynkowych. Można sobie bowiem wyobrazić sytuację, w której niektóre zdarzenia rynkowe wywołują zainteresowanie wśród osób niezwiązanych na co dzień z giełdą. Za tego typu wątpliwością przemawia fundamentalna dla finansów behawioralnych teoria perspektywy, zgodnie z którą głównym motywatorem działania jest potencjał straty (zob. [Kahneman, Tversky 1989]). Zatem, gdyby w istocie za wskaźniki wyszukiwań odpowiedzialni byli jedynie inwestorzy, spodziewać można by się, że korelacja występująca między zapytaniami a wynikami indeksu przyjmie kierunek ujemny.

Celem niniejszego badania była weryfikacja hipotezy mówiącej o tym, że dane pochodzące z aktywności użytkowników Internetu niosą ze sobą walor predykcyjny w kontekście lokalnego rynku kapitałowego. Wyniki stanowią przesłankę do stwierdzenia, że istnieje związek między aktywnością użytkowników Internetu a wynikami indeksu WIG20 oraz że dane pochodzące z wyszukiwarki Google mogą nieść ze sobą walor predykcyjny dla GPW przy uwzględnieniu dwutygodniowej dynamiki. Dodatkowo na podstawie wyników badania pokusić się można o stwierdzenie, że podejmowanie prób analizy dynamiki rynku przez pryzmat nowego typu danych może z punktu widzenia nauki wnosić wartość dodaną.

Literatura

Bijl L., Kringhaug G., Molnár P., Sandvik E., 2016, *Google searches and stock returns*, International Review of Financial Analysis, vol. 45, s. 150-156.

Blazquez D., Domenech J., 2017, *Big Data sources and methods for social and economic analyses*, Technological Forecasting & Social Change, vol. 130, s. 99-113.

Chan E.H., Sahai V., Conrad C., Brownstein J.S., 2011, *Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance*, PLoS Neglected Tropical Diseases, vol. 5, s. 1-6.

Cheikh E, Nadine C., Nicolle C., 2015, *Understandable Big Data: A survey*, Computer Science Review, vol. 17, s. 70-81.

Choi H., Varian H. (red.), 2009, *Predicting the Present with Google Trends*, Google Inc, Mountain View.

Connors S., Courbe J., Waishampayan V., 2013, *Where have you been all my life? How the financial services industry can unlock the value in Big Data*, https://www.pwc.com/us/en/financial-services/publications/viewpoints/assets/pwc-unlocking-big-data-value.pdf (20.08.2017).

D'Amuri F., Marcucci J., 2017, *The predictive power of Google searches in forecasting US unemployment*, International Journal of Forecasting, vol. 33, s. 801-816.

Da Z., Engelberg J., Gao P., 2013, *The sum of all FEARS: investor sentiment and asset prices. Review of financial studies*, The Review of Financial Studies, vol. 28, no. 1, s. 1-32.

Da Z., Engelberg J., Gao P., 2011, *In search of attention*, The Journal of Finance, vol. LXVI, no. 5, s. 1461-1499.

Fondeur Y., Karamé F., 2013, *Can Google data help predict French youth unemployment*, Economic Modelling, vol. 30, s. 117-125.

Ginsberg J. i in., 2009, *Detecting influenza epidemics using search engine query data*, Nature, vol. 457, no. 7232, s. 1012-1014.

Google Trends, Sposób dostosowywania danych w trendach, https://support.google.com/trends/answer/4365533?hl=pl&ref_topic=6248052 (18.04.2018).

Grossman S.J., Stiglitz J.E., 1980, *On the impossibility of informationally efficient markets*, Amer. Econ. Rev. 70 (1980), s. 393-408.

Hu H., Tang L., Zhang S., Wang H., 2018, *Predicting the direction of stock markets using optimized neural networks with Google Trends*, Neurocomputing, vol. 285, s. 188-195.

Joseph K., Wintoki M.B., Zhang Z., 2011, *Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search*, International Journal of Forecasting, vol. 27, s. 1116-1127.

Jun S., Yoo H., Choi S., 2017, *Ten years of research change using Google Trends: From the perspective of big data utilizations and applications*, Technological Forecasting & Social Change, vol. 130, s. 69-87.

Kahneman D., Tversky A., 1979, *Prospect theory: an analysis of decision under risk*, Econometrica, vol. 47, no. 2, s. 263-292.

Kristoufek L., 2013, *Can Google Trends search queries contribute to risk diversification*?, Scientific Reports, vol. 3, no. 2713, s. 1-6.

McKinsey Global Institute, 2016, *The age of analytics: competing in a data-driven world*, https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20analytics/our%20insights/the%20age%20of%20analytics%20competing%20in%20a%20data%20driven%20world/mgi-the-age-of-analytics-executive-summary.ashx (18.04.2018).

Naccarato A., Falorsi S., Loriga S., Pierini A., 2017, *Combining official and Google Trends data to forecast the Italian youth unemployment rate,* Technological Forecasting & Social Change, vol. 130, s. 114-122.

Preis T., Moat H.S., Stanley E.H., 2013, *Quantifying trading behavior in financial markets using Google trends*, Scientific Reports, vol. 3, no. 1684, s. 1-6.

Takeda F., Wakao T., 2014, *Google search intensity and its relationship with returns and trading volume of Japanese stocks*, Pacific-Basin Finance Journal, vol. 27, s. 1-18.

Yao T., Zhang Y-J., Ma C-Q., 2017, *How does investor attention affect international crude oil prices*, Applied Energy Journal, vol. 205, s. 336-344.

Yu L., Zhaoa Y., Tangc L., Yangd Z., 2018, *Online big data-driven oil consumption forecasting with Google trends*, International Journal of Forecasting, w druku.