

Uniwersytet Warszawski  
Wydział Nauk Ekonomicznych

Paweł Elert  
Nr albumu: 203027

**Analiza wpływu wiadomości prasowych na cenę  
akcji z wykorzystaniem narzędzi text-miningu**

Praca magisterska  
na kierunku: Informatyka i Ekonometria

Praca wykonana pod kierunkiem  
dr. hab. Ryszarda Kokoszczyńskiego, prof. UW  
z Katedry Statystyki i Ekonometrii  
WNE UW

Warszawa, maj 2008

*Oświadczenie kierującego pracą*

Oświadczam, że niniejsza praca została przygotowana pod moim kierunkiem i stwierdzam, że spełnia ona warunki do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data

Podpis kierującego pracą

*Oświadczenie autora pracy*

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa (magisterska) została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data

Podpis autora pracy

## **Streszczenie**

Praca podejmuje problematykę przetwarzania informacji zawartych w nagłówkach wiadomości prasowych za pomocą narzędzi text miningu. We wstępie teoretycznym przedstawiono hipotezę o efektywności rynków finansowych, stanowiącą szkielet teoretyczny tej pracy. Następnie zaprezentowano przegląd wybranych pozycji z literatury dotyczącej omawianego zagadnienia, wyczerpującą dyskusję na temat idealnych danych wejściowych oraz problem ich wyboru. W rozdziale badawczym zweryfikowano hipotezy, czy wiadomości prasowe niosą dodatkową informację, czy wiedzę tę można wydobyć za pomocą zastosowanego narzędzia – text miningu, oraz czy na podstawie tej wiedzy można osiągnąć ponadprzeciętne zyski na giełdzie.

## **Słowa kluczowe**

<ceny, akcje, tekst, text mining, wiadomości, prasa,  
giełda, prognoza, efektywność>

## **Dziedzina pracy (kody wg programu Socrates-Erasmus)**

Ekonomia (14300)

## **Klasyfikacja tematyczna**

## SPIS TREŚCI

WSTĘP.....	5
ROZDZIAŁ I. Część teoretyczna.....	9
1.1. Hipoteza o efektywności rynków finansowych (EMH).....	9
1.1.1. Istota i formy efektywności rynku kapitałowego.....	9
1.1.2. Kontrowersje dotyczące hipotezy.....	11
1.1.3. Praktyczna ocena efektywności rynków finansowych.....	12
1.1.4. Zastosowanie text miningu w kontekście EMH.....	14
1.2. Text mining.....	14
1.2.1. Analiza tekstu.....	16
1.2.2. Budowa macierzy wystąpień.....	19
ROZDZIAŁ II. Przegląd wybranych pozycji z literatury.....	22
2.1. Zespół Wütricha.....	22
2.2. Zespół Lavrenki.....	23
2.3. Peramunetilleke i Wong.....	24
2.4. Thomas.....	25
2.5. Porównanie prac.....	26
ROZDZIAŁ III. Dane wejściowe.....	28
3.1. Pożądane cechy tekstu.....	28
3.2. Źródła danych.....	31
3.3. Studium przypadku.....	32
ROZDZIAŁ IV. Praca badawcza.....	38
4.1. Plan badania.....	39
4.2. Zdefiniowanie i przygotowanie danych wejściowych.....	41
4.3. Analiza morfologiczna.....	42
4.4. Połączenie danych tekstowych z szeregami czasowymi.....	44
4.5. Istotność wiadomości.....	47

4.6. Słownik fraz .....	48
4.6.1. Algorytm korekcji danych.....	53
4.7. Budowa macierzy wystąpień.....	54
4.8. Zastosowanie uczenia maszynowego.....	55
4.9. Wyniki badania .....	58
4.9.1. Klasyczne miary data miningu .....	58
4.9.2. Walidacja krzyżowa (krosvalidacja) .....	61
4.9.3. Symulacja rynkowa .....	62
4.10. Koszty transakcyjne .....	65
PODSUMOWANIE .....	67
BIBLIOGRAFIA.....	71
ZESTAWIENIE SPISÓW.....	73

## WSTĘP

*“A Modus—what is that?”*

*“It is a legend in sporting circles, Dr. Mallory. A Modus is a gambling-system, a secret trick of mathematical Enginery, to defeat the odds-makers. Every thieving clacker wants a Modus, sir. It is their philosopher’s stone, a way to conjure gold from empty air!”*

*“Can that be done? Is such an analysis possible?”*

[...]

*Fraser offered Mallory a look of pity for such naivete. “A true Modus would destroy the institutions of the Turf! It would wreck the livelihood of all your sporting-gents . . . Ever seen a track-crowd mill-up about a welsher? That’s the sort of stir a Modus would bring.”<sup>1</sup>*

Jednym z bardziej nośnych tematów w całej makroekonomii jest próba prognozowania przyszłych cen akcji za pomocą wszelkich możliwych zestawów zmiennych. Wydaje się, że prawie każdy model rynku niemal w każdym zakątku świata został już dokładnie zweryfikowany, a mimo to cały czas pojawiają się kolejne publikacje, których autorzy usiłują wypracować jak najtrafniejszą prognozę, prognozę, która okaże się lepsza niż prosty model błędzenia przypadkowego. Prognozowanie cen akcji tak bardzo pobudza wyobraźnię naukowców głównie ze względu na jego praktyczny charakter. Trafna prognoza dałaby bowiem odpowiedź na pytanie o prawidłowość alokacji kapitału w gospodarce oraz miałaby ważne implikacje dla różnych strategii inwestycyjnych dokonywanych zarówno przez drobnych inwestorów, jak i wielkie instytucje finansowe. Ponadto, jeśli można pozwolić sobie na pewien prywatny komentarz, temat ten trafia bardzo do młodzieńczej (i nie tylko) wyobraźni. Któż z czytelników, patrząc na wykresy akcji, nie marzył o znalezieniu jakiejś zależności, dzięki której zdobyłby uznanie, a może nawet sławę, i fortunę. Jak zatem bezlitosna dla człowieka, którego mózg ma wrodzoną skłonność do wyszukiwania powtarzających się wzorców (nawet jeśli one nie istnieją), jest sztandarowa teoria ekonomiczna uznawana przez większą część środowiska akademickiego. Teoria, mówiąca, iż wszelkie wahania cen akcji są jedynie białym szumem – realizacją błędu losowego.

Metodologia prognozowania cen akcji ciągle jednak ewoluuje. Od niedawna do standardowego arsenału informacji objaśniających przyszłe ceny akcji (np. zmienne makroekonomiczne oraz przeszłe ceny) dołączyła nowa możliwość – wykorzystywanie informacji nieustrukturyzowanych, czyli analiza tekstu zapisanego w języku naturalnym. Przeważająca część odbieranych przez nas informacji przekazywana jest w języku

---

<sup>1</sup> W. Gibson, B. Sterling. *The Difference Engine*. Bantam Books 1991. Pomysł umieszczenia tego cytatu w kontekście badań poświęconych efektywności rynków finansowych pochodzi z pracy: J. Thomas: *News and Trading Rules*. Pittsburgh, 2003.

naturalnym, ale do niedawna występowała wyraźna tendencja do ignorowania tych danych na rzecz danych ilościowych. Preferencja ta jest w pewnej części uzasadniona trudnościami związanymi z wydobyciem i przetworzeniem informacji zawartych w tekście, jednak wraz z rozwojem nauki oraz zwiększaniem się mocy obliczeniowej komputerów argument ten zaczyna tracić na wartości. Ponadto zaletą danych tekstowych jest to, iż możemy poznać przyczynę wystąpienia danego zjawiska a nie tylko skutek – jak w przypadku danych ilościowych.

Jednocześnie powszechnie wiadomo, iż zmiany cen na rynku wywoływane są przez wydarzenia, do których dochodzi w rzeczywistym świecie, a komunikaty prasowe są jednym z najlepszych źródeł informacji o tych wydarzeniach. Wyzwaniem dla autora tej pracy, a właściwie jego celem, jest właśnie analiza wpływu wiadomości prasowych na ceny papierów wartościowych.

### **Przedstawienie hipotez**

Badania mające na celu określenie, jaki wpływ na ceny akcji mają informacje prasowe, powinno się rozpocząć od postawienia trzech związanych ze sobą hipotez:

- *H1*: Wiadomości prasowe mają pewną wartość informacyjną;
- *H2*: Narzędzia text miningu potrafią tę informację wydobyć;
- *H3*: Dzięki wydobyciu tej informacji można osiągnąć ponadprzeciętne zyski.

Należy podkreślić, iż hipotezy te występują w ścisłym porządku i każda następna zawiera również poprzednią, a zatem udowodnienie prawdziwości którejś z nich oznacza automatycznie udowodnienie prawdziwości hipotez znajdujących się powyżej.

Pierwsza hipoteza wyraża zdroworozsądkowe przekonanie, że wiadomości prasowe zawierają pewien zasób informacji. Hipoteza ta nie wymaga formalnych testów, można bowiem wnioskować, iż profesjonalni gracze giełdowi, spekulanci oraz drobni inwestorzy przed podjęciem decyzji o zajęciu pozycji krótkiej bądź długiej powinni śledzić uważnie ostatnie wiadomości ekonomiczne oraz finansowe, studiować raporty, analizy i komentarze publikowane w różnych źródłach. Powyższe rozumowanie wydaje się być również potwierdzone empirycznie. Liczba gazet oraz serwisów internetowych oferujących dostęp do wiadomości jest przeogromna. Ponadto istnieją wyspecjalizowane instytucje zajmujące się przetwarzaniem wiadomości tekstowych na gotowe rekomendacje. Skoro ludzie płacą za te usługi, oznacza to, iż muszą mieć one wartość dodatkową. W pracy tej jednak, niejako na marginesie, zostanie przeprowadzona weryfikacja omawianej hipotezy. W rozdziale 4.5

zostanie sprawdzone, czy sam fakt pojawienia się wiadomości wpływa na rozkład cen w najbliższej godzinie.

Druga hipoteza została sformułowana jako pytanie, czy (i na ile) text mining jako narzędzie potrafi wydobyć informację z tekstu. W największym uproszczeniu omawiane w tej pracy zastosowanie text miningu ogranicza się do automatycznej identyfikacji prostych reguł w celu odpowiedniej klasyfikacji nowych wiadomości. Reguły te składają się z występujących w wiadomościach wyrazów oraz łączników logicznych. Przykładowa reguła może brzmieć następująco: jeżeli nagłówek wiadomości zawiera słowo „wojna” oraz słowo „Irak”, wtedy indeks WIG spadnie. Hipoteza ta zakłada, iż *H1* jest spełniona – w końcu nie da się za pomocą jakiegokolwiek narzędzia wydobyć informacji, która nie istnieje. Hipoteza dotycząca skuteczności text miningu zostanie zweryfikowana za pomocą testów, pozwalających wyjaśnić, w jakim stopniu jesteśmy w stanie lepiej prognozować zmiany cen akcji, analizując wiadomości prasowe.

Trzecia hipoteza zakłada, iż spełnione są hipotezy *H1* i *H2*. Mówiąc inaczej, sprawdzana jest możliwość osiągnięcia ponadprzeciętnych zysków na giełdzie przy założeniu, iż wiadomości prasowe mają pewną wartość informacyjną, a informacja ta może zostać wydobyta przez narzędzia text miningu. Hipoteza ta zostanie zweryfikowana za pomocą symulacji rynkowej. Uwzględnione zostaną również koszty transakcyjne.

## **Streszczenie**

W rozdziale pierwszym zostały przedstawione cele pracy i hipotezy badawcze.

Rozdział drugi zawiera wstęp teoretyczny, przy czym w pierwszym podrozdziale została szczegółowo przedstawiona hipoteza o efektywności rynków finansowych, kontrowersje dotyczące tej hipotezy oraz zarys badań empirycznych. Podrozdział ten zakończony jest opisem zastosowania text miningu w kontekście wspomnianej hipotezy, czyli próbą „ulokowania” tematu tej pracy względem dotychczasowego dorobku teoretycznego.

W drugiej części rozdziału teoretycznego opisany jest proces przetwarzania tekstu naturalnego, czyli próby wydobycia z niego pewnej wiedzy za pomocą narzędzi text miningu. Zaprezentowane są tu dwa stosunkowo niezależne komponenty text miningu: analiza tekstu, czyli proces przetworzenia dokumentu tekstowego, który dla komputera jest szeregiem znaków, w listę słów, i budowa macierzy wystąpień, będącej numerycznym odpowiednikiem dokumentów tekstowych, która w swej ilościowej postaci „pasuje” do klasycznych metod data miningu.

W rozdziale trzecim zostały przedstawione cztery publikacje opisujące podobne zagadnienie, jednak dotyczyły one rynków zagranicznych. W każdym podrozdziale scharakteryzowana została konkretna praca i uwagi krytyczne do niej się odnoszące. Najważniejszym podrozdziałem wydaje się podrozdział ostatni, zawierający podsumowanie zastrzeżeń wysuwanych pod adresem tych prac. W rozdziale piątym, będącym głównym rozdziałem empirycznym tego opracowania, dołożono wszelkich starań, aby nie popełnić tych samych błędów.

Rozdział czwarty opisuje dane wejściowe. Rozpoczyna się od rozważań na temat „idealnego tekstu” do przetwarzania automatycznego. Następnie przedstawione są możliwe źródła oraz wybór najlepszego z nich. Ostatnie dwa podrozdziały zawierają uzasadnienie doboru danych źródłowych oraz sposoby przekształcania tych danych. Przytoczono również fragmenty danych źródłowych w uzasadnionych przypadkach.

Rozdział piąty jest głównym rozdziałem empirycznym tej pracy, rozdziałem, zawierającym szczegółowy opis badania. Na wstępie zostały przedstawiane cele, charakterystyka i plan badania wraz z diagramem, który przybliży istotę działania zastosowanego algorytmu. Następnie zostały zweryfikowane hipotezy, czy wiadomości prasowe niosą dodatkową informację, czy wiedzę tę można wydobyć za pomocą zastosowanego narzędzia – text miningu, oraz czy na podstawie tej wiedzy można osiągnąć ponadprzeciętne zyski na giełdzie. Ostatni podrozdział zawiera omówienie wyników uzyskanych za pomocą klasycznych miar data miningu (liczba poprawnych odpowiedzi, walidacja krzyżowa) oraz symulacji rynkowej uwzględniającej koszty transakcyjne.

Rozdział szósty stanowi podsumowanie pracy.

## **Podziękowania**

Na koniec chciałbym gorąco podziękować Panu dr. hab. Ryszardowi Kokoszcyńskiemu prof. UW, mojemu promotorowi, za cenne rady oraz wskazówki, bez których z pewnością utknąłbym w ciemnym zakamarku błędnych decyzji.

Ponadto chciałbym podziękować Panu dr. Adamowi Przepiórkowskiemu za wykłady oraz konsultacje z zakresu inżynierii lingwistycznej, Panu dr. inż. Maciejowi Piaseckiemu za specjalne, przedpremierowe udostępnienie dezambiguatora, który był niezbędnym narzędziem tej pracy, oraz administratorom serwisu Money.pl, Reuters i bossa.pl za zgodę na wykorzystanie ich danych.

# ROZDZIAŁ I

## Część teoretyczna

### 1.1. Hipoteza o efektywności rynków finansowych (EMH)

W rozdziale tym opisany zostanie szkielet teoretyczny rozważań na temat wyceny akcji na giełdzie, czyli hipoteza o efektywności rynków finansowych. Wedle tej teorii cała dostępna wiedza i oczekiwania na temat przyszłości zawarte są w aktualnych cenach akcji. Jedynie nowe nieantycypowane wiadomości powodują takie zmiany cen akcji, które możemy przewidzieć, a tym samym zmiany niemające charakteru losowego.

#### 1.1.1. Istota i formy efektywności rynku kapitałowego

Najogólniej ujmując problem, pytanie o efektywność rynków finansowych jest sformalizowanym pytaniem o to, czy rynki działają w sposób prawidłowy – czy akcje spółek notowanych na Warszawskiej Giełdzie Papierów Wartościowych odpowiadają „prawdziwej” wartości tych spółek bądź kształtują się w przybliżeniu tak jak na innych dojrzałych rynkach. Istotę oraz implikację teorii EMH przedstawia Jan Czekaj w następujący sposób: „Giełda papierów wartościowych jest instytucją gospodarki rynkowej, bez której sprawne funkcjonowanie gospodarki nie jest możliwe. W warunkach gospodarki rynkowej rynek kapitałowy, w tym giełda, jest głównym mechanizmem alokacji zasobów rzeczowych i finansowych. Istnienie i sprawne działanie instytucji rynku kapitałowego jest warunkiem efektywnej alokacji tych zasobów. Efektywność procesów alokacji zasobów rzeczowych i finansowych pomiędzy różne zastosowania jest czynnikiem determinującym stabilność długookresowego wzrostu gospodarczego”<sup>1</sup>.

Efektywność rynku można rozpatrywać w trzech różnych kategoriach: jako efektywność alokacyjną, transakcyjną oraz informacyjną. Ostatnia z nich, efektywność informacyjna, zapewnia szybki przepływ informacji pomiędzy uczestnikami giełdy, co powoduje, iż informacje te w pełni i bez zwłoki znajdują odbicie w cenach papierów wartościowych. Tak rozumiana efektywność jest głównym tematem tego rozdziału oraz całej pracy.

---

<sup>1</sup> J. Czekaj, M. Woś, J. Żarnowski: *Efektywność giełdowego rynku akcji w Polsce z perspektywy dziesięciolecia*. Warszawa 2001.

Fundamentalną pracą naukową definiującą pojęcie „efektywności informacyjnej” jest artykuł Eugene F. Fama. W tekście tym autor wyróżnia trzy rodzaje hipotez efektywności rynków finansowych<sup>1</sup>:

- **Efektywność słaba** – ceny aktywów notowanych na rynku odzwierciedlają całą dostępną informację zawartą w przeszłych cenach. Oznacza to, iż analiza techniczna w przeciwieństwie do analizy fundamentalnej nie pozwala na osiągnięcie ponadprzeciętnych zysków.
- **Efektywność półsilna** – ceny aktywów oprócz przeszłych cen odzwierciedlają dodatkowo wszelką publicznie dostępną informację. Zarówno analiza fundamentalna, jak i techniczna nie pozwala na osiągnięcie ponadprzeciętnych zysków w długim okresie. Jediną możliwością „wygrania” z rynkiem jest użycie wiadomości dostępnych prywatnie (tzn. insider trading), strategie te są jednak w przeważającej liczbie krajów nielegalne.
- **Efektywność silna** – ceny aktywów zawierają wszelkie dostępne informacje, wliczając w to informacje poufne oraz prywatne. Nie ma możliwości osiągnięcia ponadprzeciętnych zysków w długim okresie.

Według autora hipoteza ta jest zbudowana na następujących założeniach:

- Brak kosztów transakcyjnych;
- Nieograniczona dostępność informacji dla uczestników rynku kapitałowego, przy czym informacja ta winna być dostępna za darmo;
- Zgodność poglądów uczestników rynku kapitałowego co do wpływu nowych informacji na ceny.

Oczywiście współczesne rynki nie spełniają tych założeń, należy jednak pamiętać, iż są to tylko warunki wystarczające i nawet w przypadku niespełnienia części postulatów można twierdzić, że rynek jest efektywny w sensie informacyjnym bądź jest bliski tego pojęcia.

Najistotniejszy argument teoretyczny popierający te hipotezy brzmi następująco: jeśli rynek byłby nieefektywny, to powinna istnieć możliwość osiągnięcia ponadprzeciętnych zysków. Gracze giełdowi z kolei, wykorzystując ponadprzeciętne zwroty, zawieraliby odpowiednie transakcje, a tym samym wpływali na zmiany ceny akcji w kierunku wyeliminowania nieefektywności. Konkludując: „Jedynymi zmianami cen, jakie będą miały miejsce, są zmiany spowodowane nowymi informacjami. Ponieważ nie ma powodów, dla

---

<sup>1</sup> E.F. Fama: Efficient capital markets: A review of theory and empirical work. *Journal of Finance*. 1970, tom 25, s. 383-417.

których moglibyśmy oczekiwać, że informacje te nie mają losowego charakteru, zmiany cen akcji z okresu na okres winny mieć charakter losowy, winny być statystycznie niezależne”<sup>1</sup>.

### 1.1.2. Kontrowersje dotyczące hipotezy

Według Eugene F. Famy wyniki badań empirycznych wskazują na występowanie efektywności słabej i półsilnej. Natomiast brak faktów potwierdzających hipotezę EMH w wersji silnej. Według autora nieefektywność w wersji silnej może występować, gdyż istnieją Market-makers oraz insiders. Market-makers to osoby, które mają dostęp do technicznych informacji, np. limitów cen dla niezrealizowanych transakcji, a przez to mogą zawierać bardziej korzystne transakcje. Insiders to grupa ludzi dysponujących poufną informacją, najczęściej są to pracownicy zatrudnieni w tych spółkach.

Następne prace empiryczne sugerowały inne anomalie ekonomiczne, jak na przykład efekt poniedziałku, efekt stycznia, efekt kapitalizacji i wiele innych. Zasadne wydaje się pytanie, czy w obliczu takich „dowodów” nie powinniśmy odrzucić hipotezy o efektywności rynku? Kolejnym kontrargumentem dla EMH są jednostki, które swoją fortunę zgromadziły, osiągając ponadprzeciętne zwroty na giełdzie w długim okresie. Nieco ironicznie zaznaczył to Warren Buffet: „Byłbym bezdomnym z niewielkim plastikowym kubeczkim gdyby rynki finansowe były zawsze efektywne”<sup>2</sup>. Jan Czekaj ustosunkował się do tej debaty w sposób następujący: „Jakkolwiek istniejący stan wiedzy nie pozwala na zajęcie jednoznacznego stanowiska w tej kwestii, to należy stwierdzić, że doskonała efektywność rynku jest konceptem teoretycznym, wymagającym spełnienia licznych nierealistycznych założeń”<sup>3</sup>.

Co więcej dyskusja ta wydaje się akademicka, jeśli weźmiemy pod uwagę rozwiązanie zaproponowane przez Grossmana oraz Stiglitz. Twierdzą oni, że prawda leży gdzieś pośrodku – anomalie związane z ponadprzeciętnymi zwrotami muszą występować, ponieważ koszty zdobycia oraz przetworzenia informacji nie są zerowe. Istnieją koszty transakcyjne, które na ogół są pomijane w badaniach ekonomicznych. Podsumowując, wszelkie strategie budowane na dostępnych nieefektywnościach mogą być istotne statystycznie, ale nie ekonomicznie, ponieważ koszty transakcyjne oraz przetwarzania informacji mogą przewyższyć zyski osiągnięte dzięki tym strategiom. W tym momencie dochodzimy do fundamentalnego prawa ekonomii: poszukiwanie i przetwarzanie informacji, a następnie zawieranie odpowiednich transakcji jest opłacalne do momentu, kiedy krańcowy zysk jest nie

---

<sup>1</sup> J. Czekaj, M. Woś, J. Żarnowski: *op.cit.*

<sup>2</sup> Cytowane za: P. Russel, V. Torbey: *The Efficient Market Hypothesis on Trial: A Survey*. Philadelphia, 2002.

<sup>3</sup> J. Czekaj, M. Woś, J. Żarnowski: *op.cit.*

mniejszy niż krańcowy koszt takiej strategii. Trafnie ujął to Jan Czekaj stwierdzając: „W istocie rzeczy zatem w sensie technicznym pewien stopień nieefektywności jest możliwy, a nawet konieczny. Istotne jest natomiast pytanie, czy rynek jest efektywny w sensie ekonomicznym”.

### 1.1.3. Praktyczna ocena efektywności rynków finansowych

W miejscu tym należy zwrócić uwagę na poważny problem metodologiczny związany z empiryczną weryfikacją hipotezy EMH, nie istnieje bowiem jeden model, jedna metoda, za pomocą której dałoby się przetestować pojęcie efektywności na rynkach finansowych. Każdy test efektywności bada koniunkcję dwóch hipotez: „dobrano poprawny model” oraz „rynek jest efektywny”. A zatem wyniki sugerujące odrzucenie hipotezy mogą być skutkiem błędnego modelu, nieefektywności rynku lub obu tych przyczyn. Problem ten jest szczególnie dotkliwy w przypadku prowadzenia badań w polskich warunkach, gdzie do dyspozycji mamy dość krótkie finansowe szeregi czasowe.

Ocena słabej efektywności rynku może być dokonana w dwojaki sposób: za pomocą statystycznych właściwości rozkładów cen instrumentów bądź przy zastosowaniu metod opartych na analizie technicznej. Pierwszy sposób polega na przeprowadzeniu testów za pomocą korelacji zmiany cen. Jest to jednak nie najlepsze rozwiązanie, ponieważ zmiany cen mogą być zależne, ale nieskorelowane. Lepszą metodą jest zastosowanie modelu błędzenia przypadkowego według wzoru:

$$P_t = P_{t-1} + \varepsilon_t$$

a następnie sprawdzenie, czy składnik  $\varepsilon_t$  jest białym szumem, mającym jednakowy rozkład IID (ang. independently and identically distributed).

Metody oparte na analizie technicznej, nie czyniąc założeń o rozkładzie cen, sprowadzają hipotezę EMH w wersji słabej do pytania o istnienie strategii technicznej, która korzystając jedynie z danych transakcyjnych, np. historyczne ceny akcji czy wolumen obrotu, pozwoli osiągnąć ponadprzeciętne stopy zwrotów. Do weryfikacji tej hipotezy wykorzystano algorytmy komputerowe generujące automatyczne sygnały dotyczące kupna/sprzedaży, co umożliwiło obiektywne sprawdzenie skuteczności tychże strategii. Wyniki pracy zespołu Jana Czekaja sugerują, że zyski w wysokości 1,5%, osiągnięte za pomocą różnych strategii oraz filtrów, można wytłumaczyć już kosztami transakcyjnymi równymi 0,1%. W rzeczywistości

koszty transakcyjne są wyższe, co oznacza, iż strategie te nie prowadzą do osiągnięcia zysku, co więcej, powodują stratę.

W przypadku hipotezy o efektywności półsilnej mamy dodatkowo do dyspozycji publicznie dostępne informacje. Weryfikacja tej hipotezy polega na ogół na sprawdzeniu, czy rynek reaguje dostatecznie szybko oraz adekwatnie na pojawienie się nowych wiadomości. Hipotezę tę można badać w sposób bezpośredni poprzez analizę reakcji rynku na napływające informacje ze spółek i ich otoczenia – tak zwane event study. Inny sposób polega na symulacji zysków osiągniętych za pomocą różnych wirtualnych strategii inwestycyjnych przeprowadzanych przy użyciu komputerów. Strategie inwestycyjne muszą być jednak realne, to znaczy podejmujące decyzje w oparciu o tę wiedzę, którą dysponowałby decydent w określonym momencie czasu  $t$ . W szczególności oznacza to niewykorzystywanie wiedzy nabytej po czasie  $t$ . Zaletą symulacji jest możliwość wykrycia siły zjawiska – wartości ponadprzeciętnej stopy zwrotu, a nie tylko kierunku zjawiska.

W początkowych etapach badań stwierdzono istnienie znacznej ilości anomalii w modelu CAPM. W latach dziewięćdziesiątych Fama i French<sup>1</sup> opracowali model trójczynnikiowy, będący rozszerzeniem modelu CAPM, który na tyle dobrze objaśniał stopy zwrotu z akcji, iż uznano, że rynek jest efektywny w sensie półsilnym.

Efektywność silna nie może być badana za pomocą odpowiednich strategii inwestycyjnych z uwagi na niepubliczność wykorzystywanych informacji. Dlatego analizuje się wyniki osiągane przez poszczególne osoby, co do których istnieje podejrzenie, iż dysponują poufnymi informacjami. Ponadto porównywane są zwroty z dużych instytucji finansowych takich jak fundusze inwestycyjne. Jako benchmark najczęściej wykorzystywana jest strategia buy&hold realizowana na danym indeksie. Ważnym elementem tych badań jest wybór odpowiedniego benchmarku jako indeksu do porównania.

Rezultaty badań wskazują wprawdzie na jednostki osiągające ponadprzeciętne zyski, jednak po uśrednieniu wyniki te nie są statystycznie istotne, co implikuje, iż rynek jest w zasadzie efektywny w sensie silnym.

Dokładne testy hipotezy EMH w wersji słabej, półsilnej oraz silnej można odnaleźć w książce: „Efektywność giełdowego rynku akcji w Polsce z perspektywy dziesięciolecia”<sup>2</sup>.

---

<sup>1</sup> E. F. Fama, K.R. French: Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*. 1993, tom 33, s. 3-56.

<sup>2</sup> J. Czekał, M. Woś, J. Żarnowski: *op.cit.*

#### **1.1.4. Zastosowanie text miningu w kontekście EMH**

Po krótkim omówieniu hipotezy o efektywności rynków finansowych chciałbym zwrócić uwagę na skutki zastosowania text miningu w kontekście hipotezy EMH, czyli w pewnym sensie „ulożować” tematykę tej pracy względem tej hipotezy. Zakładając pełną i doskonałą efektywność rynku nawet w wymiarze silnym, teoria EMH mówi, iż ceny akcji zawierają wszelkie dostępne informacje a tym samym prawidłowo wyceniają wartość akcji w danym momencie  $t$ . Jednakże pojawienie się nowych informacji, nieznanych wcześniej, może zmienić położenie równowagi, a tym samym wpłynąć na cenę akcji. Na przykład ogłoszenie raportu przez spółkę o wypracowanie zysku powyżej oczekiwań bądź nieoczekiwane informacje o fuzji muszą wpłynąć na cenę akcji. Według teorii EMH zmiana ceny jest wywołana przez arbitraż, jednakże dokonanie arbitrażu musi przynieść ponadprzeciętne zyski inwestorom, mówiąc prościej, ktoś musi zarobić na zmianie. Rozmiary tak powstałej nieefektywności są zatem zależne od szybkości rozprzestrzenienia się informacji. Empirycznym dowodem na możliwości osiągnięcia ponadprzeciętnych zysków w przypadku nowej wiadomości jest fakt, iż wszystkie serwisy giełdowe udostępniają wyniki indeksów oraz rekomendacje analityków z ostatnich 15 minut odpłatnie, natomiast informacje te stają się bezpłatne dopiero po tym czasie. Jeśli nowa informacja rozprzestrzeni się w ciągu paru minut, informacja po 15 minutach może stracić swoje komercyjne znaczenie.

Zastosowanie narzędzi text miningu może skrócić czas przetworzenia informacji oraz zmniejszyć związane z nią koszty. W realiach polskiego rynku zysk z użycia narzędzi text miningu może być jeszcze bardziej widoczny, szczególnie dla mniejszych spółek notowanych na giełdach. Głównym argumentem jest założenie, iż koszt pracy analityków nie zależy od rozmiarów analizowanej spółki, natomiast możliwy do zrealizowania zysk jest proporcjonalny do wielkości spółki. Sprawia to, iż relatywny koszt analizy wiadomości dla niewielkich spółek jest wysoki.

### **1.2. Text mining**

Text mining, znany również jako text data mining, oznacza proces poszukiwania informacji oraz nietrywialnych zależności w nieustrukturyzowanych dokumentach tekstowych. W pewnym sensie text mining jest rozszerzeniem technik oraz metodologii znanych jako data mining użytych do danych nieustrukturyzowanych. Jak zaznaczono we wstępie, najbardziej naturalną formą przechowywania informacji jest tekst. Według Ah-

Hwee<sup>1</sup> Tana ponad 80% informacji przechowywanych przez przedsiębiorstwa ma formę tekstu, co powoduje, iż text mining posiada znaczny potencjał komercyjny. Proces text miningu jest jednak dużo bardziej skomplikowany niż proces data miningu, ponieważ celem jest przetworzenie danych nieustrukturyzowanych, nieprecyzyjnych oraz często wieloznacznych. Przetwarzanie tekstu jest interdyscyplinarną nauką, korzystającą z takich dziedzin jak wyszukiwanie informacji (information retrieval), analiza tekstu, klasteryzacja, kategoryzacja, wizualizacja, interakcje z bazami danych, ekstrakcja informacji, systemy uczące się (ang. machine learning) czy data mining.

Zaawansowane przetwarzanie tekstu przez maszyny, potrafiące zrozumieć język w takiej postaci, jak rozumie go człowiek, zawsze było wielkim marzeniem ludzkości. Jednym z ciekawszych i bardziej znanych projektów jest „projekt Cyc”<sup>2</sup>, prowadzony przez Douglassa Lenata. Stworzył on zaawansowany model przetwarzania języka, moduł wnioskowania, oraz co istotniejsze, wielką bazę faktów i reguł zawierającą ponad 3,2 miliona pozycji. Metody takie jak ta opracowana przez Lenata stwarzają największe nadzieje na efektywne przetwarzanie tekstu i jeśli się sprawdzą, klasyfikacja wiadomości ekonomicznych okaże się dość łatwym zadaniem. Niestety wysiłek niezbędny do stworzenia takiego systemu jest przeogromny, a „Projekt Cyc”, nad którym rozpoczęto prace w latach osiemdziesiątych ubiegłego wieku, ciągle nie przynosi oczekiwanych wyników.

Obecne podejście jest znacznie prostsze – większość działających narzędzi text miningowych nie analizuje znaczeń wyrazów oraz zdań, a jedynie próbuje znaleźć pewne reguły oraz prawidłowości związane z występowaniem określonych ciągów znaków, które dla ludzi są słowami.

Proces szukania reguł oraz prawidłowości, czyli inaczej mówiąc pewnej wiedzy, można rozbić na trzy stosunkowo niezależne części: analizę tekstu, budowę macierzy wystąpień oraz użycie klasycznych metod data miningu. Analiza tekstu to proces przetworzenia dokumentu tekstowego, który dla komputera jest szeregiem znaków, w listę słów, a następnie wyrazów, zwaną „workiem wyrazów” (bag of words). Kolejny etap to budowa macierzy wystąpień, czyli specjalnej tablicy, będącej numerycznym odpowiednikiem dokumentów tekstowych, która w swej ilościowej postaci „pasuje” do klasycznych metod data miningu.

Niezwykle ważne jest odpowiednie zrozumienie i zdefiniowanie używanych w tej pracy terminów „wyraz” oraz „słowo”. Termin „słowo” będzie używany w sposób intuicyjny,

---

<sup>1</sup> Tan A.: Text Mining: The state of the art and the challenges. *Pacific Asia Conference on Knowledge Discovery and Data Mining PAKDD*. 1999, s. 65-70.

<sup>2</sup> C. Matuszek, J. Cabral, M. Witbrock [et al.]: An Introduction to the Syntax and Content of Cyc. *Proceedings of the 2006 AAAI Spring Symposium*. Stanford 2006, s. 44 - 49.

natomiast termin „wyraz” – jako komputerowy odpowiednik „słowa”, oznaczający pewne pojęcie, najmniejszą cząstkę informacji, którą możemy przechować w naszym umyśle i która w sposób samodzielny ma określone znaczenie. Zbiór „wyrazów” jest bliski zbiorowi „słów” bez końcówek fleksyjnych, pomniejszonemu o przyimki, spójniki, zaimki, partykuły oraz wykrzykniki – a zatem o wszelkie cząstki, które samodzielnie niosą znikomą informację.

### 1.2.1. Analiza tekstu

Analiza tekstu to proces przetwarzania dokumentu tekstowego, który dla komputera jest szeregiem znaków, w listę słów, a następnie listę wyrazów, zwaną „workiem wyrazów” (bag of words). Znaczna część działań związanych z tym procesem jest ukierunkowana na utożsamianie słów o tym samym znaczeniu. Celem analizy tekstu jest przedstawienie jednego dokumentu tekstowego w postaci wielozbioru wyrazów (ang. bag of words), w którym możliwa przestrzeń potencjalnych wyrazów została ograniczona do  $n$ . Oznacza to, iż kolejność wyrazów nie ma znaczenia, a zatem przykładowa wiadomość „TP S.A. wyprzedziła Netię pod względem użytkowników” będzie równoznaczna z wiadomością „Netia wyprzedziła TP S.A. pod względem użytkowników”. Formalnie wielozbiór wyrazów możemy zapisać za pomocą wektora:

$$\vec{d}_j = (w_{1,j}, w_{2,j}, w_{3,j}, \dots, w_{n,j})$$

gdzie:

$w_{i,j}$  to waga (np. liczebność) wyrazu  $w_n$  w dokumencie  $j$ <sup>1</sup>.

Pierwszym etapem przetwarzania dokumentu jest podział strumienia znaków na słowa i najczęściej proces ten przeprowadza się, wykorzystując spację, przecinki oraz znaki specjalne (ogólnie tzw. białe znaki). Transformacja ta nie jest aż tak trywialna, jak mogłoby się wydawać. Co prawda język polski nie zawiera apostrofów, jak np. język angielski, jednak istnieją takie ciągi znaków jak c++, B-52 czy M\*A\*S\*H, które powinny być traktowane jako jedno słowo. Kolejną grupę stanowią adresy poczty elektronicznej bądź stron internetowych, które również należy traktować jako jedno słowo, mimo iż zawierają znak małpki (@) lub

---

<sup>1</sup> A. Przepiórkowski: *Slajdy z wykładu inżynieria lingwistyczna*. Warszawa 2008.

kropkę. Odrębną kategorię tworzą dywizy stosowane przy przenoszeniu słów, co praktykuje się w wiadomościach prasowych.

Kolejnym etapem jest pominięcie słów często występujących lub słów, które nie mają semantycznego znaczenia, ponieważ ich wartość informacyjna jest znikoma. Poniższa tabela zawiera spis takich słów.

Tablica 1. Słowa semantycznie puste

w	o	są	gdy
i	że	dla	jej
się	jak	tym	jako
na	co	ale	aby
z	po	jego	już
do	od	u	ze
nie	za	e	lub
jest	tak	tej	ten
to	tego	go	ma
a	przez	ich	będzie

Źródło: A. Przepiórkowski: *op.cit.*

Słowa te w języku polskim są zwykle używane jako łączniki a słowa o samodzielnym znaczeniu. Przykładowy nagłówek artykułu prasowego „Ekspertyza ma pograżyć Giertycha oraz Leppera” zostanie przekształcony do postaci: „Ekspertyza pograżyć Giertycha Leppera”. Można zauważyć, iż ilość słów zmniejszyła się o 1/3, nie powodując istotnego pogorszenia znaczenia zdania. Lista słów, które powinny zostać usunięte nazwana jest stoplistą. Może ona zostać skonstruowana ręcznie przez lingwistów, automatycznie przy użyciu komputera poprzez wyszukanie najczęściej występujących słów albo przez połączenie tych metod. Użycie stoplisty znacznie zmniejsza wymagania pamięciowe oraz zwiększa szybkość przetwarzania. Ostatnie prace z zakresu analizy tekstów sugerują jednak, by rezygnować ze stosowania stoplisty, ponieważ utrudnia ona wyszukiwanie pewnych fraz. Przykładem może być fraza „Prawo i Sprawiedliwość”, która ma inne znaczenie niż słowa „Prawo”, „Sprawiedliwość”, gdy wyrzucony zostanie spójnik „i”.

Następnym etapem jest łączenie słów, mających bardzo podobne znaczenia, np. samochód oraz automobil. Łączenie słów zmniejsza rozmiar macierzy wystąpień, poprzez redukcję liczby kolumn, a zarazem poprawia jakość danych.

Preferowane jest, aby znaczenie słowa nie zależało od wielkości liter składających się na to słowo. Istnieje jednak potrzeba wyodrębniania nazw własnych – system powinien zatem dostrzegać różnicę między określeniem „Polska Agencja Prasowa”, a trzema słowami

„polska” „agencja” oraz „prasowa”. Najprostszą heurystyką rozwiązującą ten problem jest zamiana pierwszej litery w każdym zdaniu dokumentu na małą.

Ustalanie rdzenia słowa jest bardzo ważnym elementem analizy tekstu. Z przyczyn gramatycznych te same słowa mają różne formy, co w języku polskim jest szczególnie istotne ze względu na bardzo bogatą fleksję. Słowa: „pójść”, „iść”, „poszedłem”, mają to samo znaczenie, mimo iż są inaczej zapisane. Jestem świadomy, iż część wnikliwych czytelników może zaprotestować przeciwko stwierdzeniu, iż słowa niezależnie od formy znaczą to samo. Zdania: „Stopy procentowe poszły w górę, ponieważ ...” oraz: „Stopy procentowe poszłyby w górę, gdyby nie ...” mają przecież przeciwne znaczenia. Istnieją jednak dwa argumenty przemawiające za redukcją form fleksyjnych: efektywność (traktowanie każdej formy fleksyjnej jako słowa o oddzielnym znaczeniu może wydłużyć czas obliczeń o kilka rzędów wielkości) oraz prostota. W idealnym świecie text miningu powinien istnieć mechanizm (metryka), mierzący podobieństwo jednego słowa do drugiego. Obecnie jednak podobieństwo między słowami jest mierzone zerojedyńkowo, co oznacza, iż rozsądniejsze jest przyjęcie założenia, iż słowa o tym samym rdzeniu są bardziej podobne do siebie niż różne.

Zadanie znalezienia rdzenia słowa można wykonać za pomocą dwóch typów narzędzi: lematyzacji oraz stemmingu. Lematyzacja ma tę przewagę nad stemmingiem (który jest często nazywaną lematyzacją dla ubogich), iż korzysta ze słownika i analizuje kontekst słów, zwłaszcza zaś formy słów stojących obok. Wynikiem tego działania jest znalezienie form hasłowych.. Stemming natomiast jest prostym zbiorem reguł, operujących na poziomie liter, którego wynikiem są rdzenie słów<sup>1</sup>. W języku angielskim sprawa znalezienia rdzenia jest stosunkowo prosta – wystarczą odpowiednie reguły, modyfikujące końcówkę słowa. W języku polskim, charakteryzującym się dużo bogatszą odmianą, nie ma możliwości użycia tak prostego zbioru reguł, ponieważ niektóre słowa, jak np. „iść” oraz „poszedłem”, nie zawierają nawet wspólnych liter. W pracy tej do uzyskania odpowiednich form hasłowych użyte zostały specjalne narzędzia do lematyzacji.

Kolejną kwestią jest wyszukiwanie fraz, ponieważ pewne pary słów stanowią jedną całość i nie należy ich rozbijać, np. „Lewica i Demokraci”. Innym przykładem są związki czasowników z rzeczownikami, które znaczą dużo więcej niż pojedyncze słowa składające się na tę frazę. W tym przypadku ważnym czynnikiem jest efektywność przetwarzania; łatwo można policzyć, że traktowanie łączne każdej pary słów spowoduje zwiększenie słownika możliwych znaczeń oraz czas przetwarzania o kilka rzędów. Łączenie każdych trzech

---

<sup>1</sup> A. Przepiórkowski: *op.cit.*

wyrazów we frazy może być niewykonalne dla dzisiejszych komputerów. Możliwe jest jednak emulowanie fraz złożonych z trzech wyrazów za pomocą dwóch fraz dwuwyrzowych, np. frazę „Polskie Koleje Państwowe” można zapisać jako dwie frazy: „Polskie Koleje” oraz „Koleje Państwowe”. Istnieje sporo pomysłów na optymalizację budowania fraz, kluczową ideą jest jednak identyfikacja słów, które łącząc się w pary niosą dodatkowe znaczenie. Najprostszym pomysłem jest łączenie w pary jedynie rzeczowników.

W dyskusjach nad procesem analizy tekstu często poruszany jest temat poprawiania błędów typograficznych, czyli prostych literówek. W pracy tej jednak zakładamy dobrą jakość danych wejściowych i nie będziemy wyszukiwali błędów.

Osoby szukające dodatkowych informacji na temat przetwarzania tekstu powinny zapoznać się z pracą *Introduction to Information Retrieval*<sup>1</sup>. W książce tej można znaleźć wyczerpujący opis takich zagadnień jak indeksowanie dużych zbiorów danych (niemieszczących się w pamięci komputera) czy wyszukiwanie binarne.

### 1.2.2. Budowa macierzy wystąpień

Drugim krokiem text miningu jest budowa macierzy wystąpień, czyli specjalnej tablicy będącej numerycznym odpowiednikiem dokumentów tekstowych, która w swej ilościowej postaci może być użyta jako wejście do algorytmów eksploracji danych (data mining).

Macierz wystąpień jest macierzą klasyczną. W macierzy wystąpień kolumny oznaczają wyrazy bądź zbiory wyrazów, natomiast wiersze poszczególne dokumenty tekstowe. Komórka macierzy  $A_{ij}$  mierzy liczbę wystąpień danego wyrazu w danym dokumencie tekstowym. Jeden wiersz jest ilościowym odpowiednikiem jednego dokumentu w postaci wektora wyrazów.

Istotą budowy macierzy wystąpień jest stworzenie kolumn ze wszystkich unikalnych wyrazów występujących w „worku wyrazów”. Następnie każdą komórkę macierzy  $A_{ij}$  wypełnia się za pomocą wartości funkcji  $f(x)$ , gdzie argumentem  $x$  jest liczba wystąpień wyrazu  $j$  w dokumencie  $i$ . Poniższa tabela przedstawia przegląd najpopularniejszych postaci funkcji  $f(x)$ :

---

<sup>1</sup> C. D. Manning, P. Raghavan, H. Schütze: *Introduction to Information Retrieval*. Cambridge 2008.

Tabela 1. Przegląd najpopularniejszych wartościowań dla wielozbiorów wyrazów

Nazwa przekształcenia	Wzór
<i>Boolean weighting</i>	$B_i = \begin{cases} 0, & \text{gdy } TF_i = 0 \\ 1, & \text{gdy } TF_i > 0 \end{cases}$
<i>Term frequency (TF):</i>	liczba wystąpień wyrazu w całym dokumencie
<i>Normalization</i>	$NTF_i = \frac{TF_i - E(TF_i)}{Var(TF_i)}$
<i>Inverse document frequency (IDF)</i>	$IDF_i = \log \frac{N}{n_i}$ gdzie N – liczba dokumentów w zbiorze, ni liczba dokumentów zawierających wyraz i
<i>TF × IDF</i>	$TF \times IDF_i = TF_i \times IDF_i$
<i>Argumented normalized term frequency</i>	$\frac{TF_i}{2 \max(TF_i)} + 0,5$

Źródło: V. Cho, B. Wüthrich, J. Zhang: Text Processing for Classification. Journal of Computational Intelligence in Finance. 1999, tom 7, s. 6-22.

Funkcja  $f(x)$  przekształca liczbę wystąpień danego znaczenia w dokumencie na odpowiednią wagę, niestety nie istnieje jeszcze ogólnie przyjęta postać funkcji, gdyż każda z proponowanych przez specjalistów ma swoje wady oraz zalety. Najprostsza z nich to wartościowanie binarne (boolean weighting), gdzie przyporządkowujemy wartość 1, gdy wyraz występuje oraz 0 w przeciwnym wypadku. Kolejną, właściwie intuicyjną, funkcją jest funkcja liniowa (term frequency), gdzie wartością funkcji jest liczba występujących wyrazów w dokumencie. Przekształcenie to ma poważną wadę, ponieważ najczęściej występujące w języku polskim słowa: „w”, „oraz”, „i”, niosą znikomą informację. Rozwiązaniem eliminującym powyższą wadę jest funkcja inverse document frequency, w której najwyższą wartość osiągają wyrazy występujące jedynie w badanym dokumencie. Istnieje wiele innych postaci funkcji, ale wymienienie ich wszystkich przekracza zakres tej pracy.

Poniższa tabela ilustruje proces budowy macierzy wystąpień:

Tabela 2. Przykładowa macierz wystąpień

<b>Nagłówek wiadomości</b>	<b>Premier</b>	<b>Kaczyński</b>	<b>Zarządzać</b>	<b>Wybory</b>	<b>Zabójstwo</b>	<b>Irak</b>	<b>Demokracja</b>
Premier Kaczyński zarządza wybory	$f(1)$	$f(1)$	$f(1)$	$f(1)$	$f(0)$	$f(0)$	$f(0)$
Zabójstwo premiera w Iraku to nie zabójstwo	$f(1)$	$f(0)$	$f(0)$	$f(0)$	$f(2)$	$f(1)$	$f(0)$
Wybory to zabójstwo Demokracji!	$f(0)$	$f(0)$	$f(0)$	$f(1)$	$f(1)$	$f(0)$	$f(1)$

Źródło: Obliczenia własne

Jak stwierdzono w podrozdziale 1.2.1 słowa „w”, „to”, „nie” należało odrzucić jako wyrazy bez semantycznego znaczenia. Słowa „Premier” oraz „premiera” zostały uznane za ten sam wyraz mający jedno znaczenie. Liczba kolumn powyższej macierzy jest równa liczbie wyrazów występujących we wszystkich dokumentach – przeważnie jest rzędu kilkunastu tysięcy. Należy zauważyć, iż zazwyczaj każdy wektor zawiera niemal same zera i kilkanaście do kilkudziesięciu wartości dodatnich, co sprawia, iż efektywne jest użycie jakiejś formy kompresji.

Istotnym problemem w procesie budowy macierzy jest potencjalnie ogromna liczba kolumn macierzy. Rozwiązaniem tego problemu jest zignorowanie wyrazów występujących bardzo rzadko, np. poniżej dziesięciu razy, bądź użycie algorytmów redukujących liczbę wymiarów danych na przykład poprzez analizę czynnikową. Oczywiście jest to tylko zarys wiadomości z zakresu text miningu, ale mam nadzieję, iż czytelnik zrozumie ogólną ideę procesu.

## ROZDZIAŁ II

### Przegląd wybranych pozycji z literatury

Zagraniczne rynki papierów wartościowych wydają się dość dobrze zbadane, o czym świadczą prace omawiające to zagadnienie. Poniżej przedstawiono cztery najbardziej znaczące pozycje z tego obszaru.

#### 2.1. Zespół Wütricha

Pierwszy model wykorzystujący text mining został opisany w artykule „Text Processing for Classification” opracowanym przez zespół V. Cho, B. Wüthrich oraz J. Zhang<sup>1</sup>.

W pracy tej autorzy opisali przebieg analizy głównego indeksu Hang Seng Index. Źródłem danych były wiadomości publikowane w ciągu nocy na witrynach internetowych The Wall Street Journal, Financial Times i Reuters przez 179 dni. Dane te zawierały globalne, lokalne, polityczne oraz ekonomiczne wiadomości, cytaty wpływowych polityków i rekomendacje finansowych analityków. Według autorów artykułu dane te dostarczyły więcej informacji niż tradycyjne dane numeryczne, ponieważ oprócz efektów w postaci określonych zmian na rynku papierów wartościowych pozwalały poznać przyczynę tych zmian.

W celu odpowiedniej klasyfikacji wiadomości autorzy użyli zbioru reguł, stworzonego przez ekspertów, zawierającego 423 reguły w formie implikacji (np. „obligacje AND rosnąć WTEDY indeks ROŚNIE”). Listy reguł zostały dobrane tak, aby dopasować je do zbioru najczęściej występujących sytuacji na rynku. Każda fraza mogła składać się z kilku słów. Następnym krokiem było sprawdzenie częstotliwości wystąpienia każdej reguły w danym dniu. Częstotliwości te były przekształcone za pomocą funkcji ITF (inverted term frequency) a następnie normalizowane.

Prognoza odbywała się za pomocą klasyfikacji danego dnia do jednej z trzech kategorii: indeks giełdowy spadnie o ponad x%, indeks wzrośnie o ponad x% oraz wartość indeksu nie zmieni się. Wartość x została ustalona na poziomie 0,5, co pozwoliło uzyskać równomierny rozkład w tych trzech kategoriach. Spadek bądź wzrost indeksu był mierzony różnicą cen zamknięcia z dnia poprzedniego i zamknięcia dnia aktualnego. System został wytrenowany na pierwszych 100 dniach. Autorzy sprawdzili następujące rodzaje estymacji: naiwny klasyfikator Bayesa (naïve Bayes classifier), algorytm najbliższych sąsiadów (nearest neighbour classifier) oraz sieć neuronową. W fazie operacyjnej prototyp przetwarzał dane,

---

<sup>1</sup> V. Cho, B. Wüthrich, J. Zhang: *op.cit.*

które pojawiły się w ciągu nocy, i na podstawie odpowiednich rozkładów prawdopodobieństwa fraz dawał gotowe rekomendacje do zajęcia odpowiedniej pozycji rano. Wartość akcji była spieniężana pod koniec dnia. Według autorów model ten prawidłowo prognozował dane w 46,8% przypadków, dla porównania proste losowanie dałoby jedynie 33,3% skuteczności.

Ekonomiści zajmujący się tym zagadnieniem uznali, iż jednym z mankamentów tej pracy było użycie znacznej liczby wiadomości prasowych „po fakcie”. Wiadomości te opisywały starsze wydarzenia, a tym samym nie niosły za sobą żadnych nowych informacji. Omawiane badanie nie do końca można zaliczyć do kategorii text miningu, ponieważ pierwszy słownik został stworzony manualnie. Poważniejszą wadą metodologiczną modelu było założenie, iż cena zamknięcia równa jest cenie otwarcia, co powodowało, iż model sprzedawał zbyt drogo, a kupował zbyt tanio.

## 2.2. Zespół Lavrenki

Głównym celem modelu opracowanego przez zespół kierowany przez V. Lavrenkę<sup>1</sup> była prognoza cen akcji w okresie 115 dni przy użyciu bardzo krótkich okien czasowych (intraday). W czasie badania zespół automatycznie przetworzył 38 469 wiadomości dla 127 spółek giełdowych pochodzących ze strony Biz Yahoo. Znaczną różnicą w porównaniu do modelu opisanego w „Text Processing for Classification”<sup>2</sup> jest fakt, iż zbiór reguł został stworzony automatycznie, podczas gdy V. Cho, B. Wüthrich oraz J. Zhang korzystali ze słownika opracowanego przez ekspertów.

Podstawową ideą modelu zespołu Lavrenki było odpowiednie przyporządkowanie artykułów prasowych do występujących w szeregach czasowych trendów. System ten bazuje na założeniu, że szereg czasowy reprezentujący ceny akcji może być rozbity na trendy (trend jest monotoniczną funkcją 3 lub więcej okresów), a trendy można zakwalifikować do odpowiednich kategorii. Zespół wprowadził pięć możliwych kategorii: skok w górę (nachylenie 75% - 100%), niewielki wzrost (50% - 75%), bez zmian (absolutna wartość nachylenia mniejsza od 50%), niewielki spadek (50% - 75%) oraz nurkowanie (75% - 100%). Liczby w procentach oznaczają relatywne nachylenie określonego segmentu do segmentu o największym (i najmniejszym odpowiednio) nachyleniu. Autorzy zasygnalizowali problem doboru wiadomości, gdyż, jak stwierdzili, znaczna część wiadomości analizowanych w

---

<sup>1</sup> V. Lavrenko, M. Schmill, D. Lawrie [et al.]: Language Models for Financial News Recommendation. *Conference on Information and Knowledge Management*. McLean 2000, 9 konferencja, s. 389 - 396.

<sup>2</sup> V. Cho, B. Wüthrich, J. Zhang: *op.cit.*

klasycznych badaniach text miningowych ma bardzo znikomy wpływ na cenę akcji, stanowiąc biały szum, który może utrudnić badanie. Autorzy rozwiązali ten problem, używając zewnętrznego przypisania stworzonego przez Biz Yahoo, które przyporządkowuje każdej wiadomości konkretną nazwę firmy, do której ta wiadomość się odnosi. Kolejnym krokiem było połączenie artykułów prasowych z trendem czasowym. Artykuł prasowy mógł spowodować trend, jeśli ukazał się w ciągu pięciu do dziesięciu godzin przed trendem. Wartość z przedziału 5 – 10 godzin została uznana za tę, która przynosi najlepsze wyniki.

Według autorów zysk w ciągu 40 dni wyniósł przeciętnie 210%(!), największy zysk osiągnął IBM – 470%, największą stratę Disney – 530%. Autorzy określili zysk w wysokości 210% jako „bardzo skromny”. Niestety, omawiane badanie ma pewne wady. Zespół Lavrenki wybrał 127 spółek, które miały największe zyski bądź straty, a taki wybór nie powinien być dokonywany *ex ante*, gdyż prowadzi do obciążenia w kierunku spółek o wysokiej wariancji. Zastrzeżenia budzi również fakt, iż w ciągu badania spółki wygenerowały 38 000 różnych wiadomości, które stanowiły sygnał do sprzedaży bądź kupna. Liczba ta jest zbyt wysoka jak na okres 115 dni. Możliwe, że niektóre zdarzenia zostały błędnie przypisane do wszystkich badanych spółek. Autorzy założyli również, iż system może pożyczyc w nieskończoność. Z przedstawionych wyliczeń wynika, iż system pożyczał średnio 400 000 dolarów, czyli operował kwotą 40 razy większą niż posiadał, podczas gdy większość godnych zaufania instytucji finansowych nie operuje kwotą przekraczającą dziesięciokrotnie wartość posiadanego kapitału. Największą wadą modelu jest jednak pominięcie kosztów transakcyjnych, co z uwagi na olbrzymią liczbę transakcji może znacząco zmienić wyniki symulacji.

### 2.3. Peramunetilleke i Wong

W 2001 został opracowany przez D. Peramunetilleke'a i R.K. Wonga<sup>1</sup> przy współudziale specjalistów z UBS, zajmujących się handlem walutami, prototyp systemu. Jego zadaniem była prognoza kursów akcji w ciągu 60 minut, na podstawie przeszłych 120 minut. Celem prognozy była odpowiedź na pytanie, czy wejść w pozycję długą, krótką czy bez zmian na rynku walutowym USD/DEM oraz USD/JPY. Jak stwierdzili autorzy na podstawie testów empirycznych, wybór 60 oraz 120 minut dał najlepsze rezultaty. Wybór ten był sporządzony *ex ante*, co powoduje, iż wyniki symulacji mogą być obciążone.

---

<sup>1</sup> D. Peramunetilleke, R.K. Wong: Currency Exchange Rate Forecasting from News Headlines. *ACM International Conference Proceeding Series*. Melbourne 2002, 13 konferencja, s. 131 - 139.

Słownik składał się z ręcznie stworzonych 400 fraz, wyrażających od 2 do 5 słów połączonych operatorem „i”. Dane, na których system został przetestowany, pochodziły z września 1993 z Olsen & Associates in Zurich, było ich relatywnie niewiele i były trochę nieaktualne. Klasyfikacja odbywała się poprzez przypisanie jednej z trzech kategorii: dolar wzrośnie, dolar spadnie oraz dolar pozostanie bez zmian. Wartość 0,23% została wybrana jako wartość progowa, powodująca, iż wszystkie kategorie zawierają tę samą liczbę obserwacji. Według autorów system dokonał poprawnej prognozy w 50% przypadków, podczas gracz losowy miałby rację jedynie w 33,3% przypadków. Autorzy zaznaczają również, iż profesjonalni gracze giełdowi zajmujący się handlem walutami dokonują trafnej prognozy również w 50% przypadków, jednak automatyczne przetwarzanie danych przez maszynę ma przewagę w prędkości obliczeń.

#### **2.4. Thomas**

Cechą charakterystyczną prototypu stworzonego przez J. Thomasa<sup>1</sup>, prototypu przedstawionego w pracy „News and Trading Rules”, jest modelowanie zmienności cen akcji, a nie ich wartości. Wiadomości są dzielone na te, które po ogłoszeniu wywołują duże wahania cen, oraz na pozostałe, które nie powodują tego zjawiska. Badanie to jest bardzo interesujące z punktu widzenia ekonomisty, ponieważ często informacja o przyszłych wahaniami cen jest tak samo ważna jak sam poziom cen. Publikacja Thomasa jest pracą zawierającą wiele wątków dotyczących automatycznych strategii inwestycyjnych, takich jak analiza techniczna oraz algorytmy uczenia maszynowego (np. algorytmy genetyczne). Z mojego punktu widzenia najbardziej interesujące są rozdziały dotyczące prognozy zmienności cen akcji za pomocą liczby wątków na forach dyskusyjnych oraz wiadomości prasowych. W swojej publikacji Thomas często podkreśla, iż w ekonomii najprostsze rozwiązania są często najlepsze z powodu patologicznej ilości szumów w danych finansowych – według hipotezy o efektywności rynków finansowych niektóre finansowe szeregi czasowe to jedynie biały szum. Wszystko to powoduje niebezpieczeństwo przeuczenia algorytmów data miningu. Warto zaznaczyć również, iż Thomas w przeciwieństwie do wcześniej wymienionych autorów uwzględnia różnicę w cenie otwarcia akcji oraz cenie zamknięcia akcji.

Prognoza zmienności cen akcji za pomocą liczby wątków dotyczących danych spółek jest bardzo ciekawym pomysłem, a zarazem prostszym w realizacji niż klasyczny text mining. Dane dotyczące tego badania pochodzą z grup dyskusyjnych Yahoo zbieranych w latach 1998

---

<sup>1</sup> J. Thomas: *News and Trading Rules*. Pittsburgh, 2003.

– 2001. Autor rozważał odchylenia od średniego poziomu dla danego dnia, aby uwzględnić fakt, iż w weekendy jest zazwyczaj więcej wiadomości. Według Thomasa jakość wątków w grupach dyskusyjnych jest bardzo wątpliwa – tematy rozmowy często odbiegają od merytorycznych zagadnień, sprowadzają się do osobistych ataków, często informacje są sprzeczne, nierzetelne, a czasem specjalnie błędne. Wszystko to powoduje, że analiza liczby wiadomości jest znacznie bardziej sensowna od analizy szczegółów wiadomości.

Głównym wnioskiem autora jest występowanie korelacji pomiędzy liczbą postów pojawiających się w ciągu dnia dotyczących danego papieru wartościowego a wolumenem obrotu tegoż papieru wartościowego. Ponadto liczba postów może tłumaczyć tę część cen akcji w analizie technicznej, która nie jest tłumaczona bezpośrednio przez wolumen transakcji.

Poważnym wkładem Thomasa w rozwój text miningu jest analiza nagłówków wiadomości z serwisu Yahoo dokonana pomiędzy majem 2001 a kwietniem 2002 roku. Autor stworzył strategię, według której w momencie publikacji wiadomości, mogących zwiększyć zmienność akcji, należy czasowo opuścić rynek akcji. Decyzja o powrocie na rynek jest uzależniona od analizy technicznej. Rezultatem tej strategii jest lepszy wskaźnik zysk/ryzyko, ponieważ długookresowy zysk z papierów wartościowych jest osiągnięty przy mniejszej zmienności. Obecna wersja prototypu zakłada klasyfikację wiadomości do jednej z 39 predefiniowanych kategorii, reprezentujących specyficzny typ wiadomości, np. przejęcie, sprawa sądowa. Zarówno podział na kategorie jak i zbiory wyrazów działające jako reguły klasyfikacyjne zostały stworzone ręcznie przez ekspertów. W technicznym sensie ma to niewiele wspólnego z text miningiem. Jeżeli przynajmniej jedna z reguł „zadziała”, wiadomość zostaje przyporządkowana do odpowiedniej kategorii.

Niestety w pracy tej brakuje pewnych ważnych szczegółów dotyczących np. zachowania się systemu w przypadku, kiedy wiadomość będzie pasowała do kilku kategorii jednocześnie. Poważniejszym zastrzeżeniem merytorycznym jest wybór dziennej częstotliwości. Największa zmienność papierów wartościowych może wystąpić w krótkim czasie po publikacji wiadomości.

## **2.5. Porównanie prac**

Podsumowując, w rozdziale tym zostały omówione cztery wybrane prace, prezentujące praktyczne zastosowanie narzędzi text miningu w celu predykcji finansowych szeregów czasowych. W trzech z nich, Wütricha, Lavrenki, Peramunetilleke’a, prognozowany jest

kierunek zmian cen akcji, odbywający się poprzez klasyfikacje nowych obserwacji do jednej z trzech (bądź pięciu) predefiniowanych kategorii: wartość aktywa rośnie, spada bądź pozostaje bez zmian. W przypadku pięciu kategorii spadek oraz wzrost jest dzielony na dalsze dwie kategorie: silna zmiana oraz niewielka zmiana. Czwarty prac autorstwa Thomasa prognozuje nie kierunek a zmienność cen akcji, klasyfikując nowe obserwacje do jednej z dwóch kategorii: duża zmienność oraz niewielka zmienność. W znacznej części prac ich autorzy używają zbioru reguł stworzonych ręcznie, przy pomocy ekspertów z danej dziedziny. Niestety zbiory tych reguł, pomijając kilka prostych przykładów z każdego zbioru, nie zostały opublikowane publicznie. Reguły zazwyczaj były zbiorem słów połączonych operatorami logicznymi: „i”, „oraz”, „nie”.

Krytycy uznali, że w przedstawionych publikacjach autorzy popełnili błąd decydując się na wykorzystanie danych o częstotliwości dziennej. W części teoretycznej niniejszej pracy przedstawiony został pogląd, według którego nowe informacje wpływają na rynek w czasie krótszym niż 15 minut. A zatem w przypadku operowania na częstotliwościach dziennych dane z poprzedniego dnia mogą być nieaktualne, więc nie powinny być użyte jako zmienne decyzyjne. Kolejnym zastrzeżeniem jest użycie zbyt krótkich szeregów czasowych, wynoszących czasem kilka miesięcy. Rozbicie zbioru na część treningową oraz walidacyjną zmniejsza jeszcze bardziej liczbę informacji dostępnej dla algorytmów uczenia maszynowego. Należy pamiętać, iż text mining nie analizuje znaczenia tekstu ani nie zawiera żadnych modułów wnioskowania logicznego, lecz próbuje w sposób automatyczny prognozować przyszłą sytuację, korzystając z wydarzeń historycznych. Oznacza to, iż długość dostępnych danych historycznych ma kolosalne znaczenie. Dla przykładu, aby poprawnie prognozować wpływ wyrazów „wojna” bądź „wybory” muszą one wystąpić w danych treningowych. Wątpliwe jest, aby w tak krótkim oknie czasowym jak pół roku zdarzyły się dwie wojny bądź dwukrotne wybory parlamentarne.

Według autorów przedstawionych prac zastosowane modele osiągają bardzo obiecujące wyniki finansowe. Jednakże modele te w przeważającej części nie uwzględniają kosztów transakcyjnych, spreadu pomiędzy ceną kupna a ceną sprzedaży oraz różnicy pomiędzy ceną zamknięcia pod koniec dnia a ceną otwarcia w dniu następnym. Uwzględnienie tych różnic najprawdopodobniej spowodowałoby, iż zyski osiągnięte przez te modele zbiegłyby do zera. Ostatnim zastrzeżeniem a zarazem wskazówką dla przyszłych badań powinien być odpowiedni dobór obserwacji. W wielu pracach wybór papierów wartościowych był dokonywany *ex ante* lub w sposób subiektywny (bez jasnego kryterium), co mogło obciążać wyniki badania, doprowadzając do przeszacowania możliwych zysków.

## ROZDZIAŁ III

### Dane wejściowe

W badaniach poświęconych dziedzinie text miningu autorzy często koncentrują swoje wysiłki na metodach estymacji, traktując dane wejściowe jako obywateli drugiej kategorii, jako rzecz egzogeniczną, często ograniczając się jedynie do podania źródła. Według mnie takie postępowanie jest błędne, ponieważ w tego typu badaniach wybór odpowiednich dokumentów stanowiących dane źródłowe może przesądzić o powodzeniu lub porażce pracy. Celem tego rozdziału jest zatem zwrócenie uwagi na problem wyboru danych źródłowych oraz przegląd potencjalnych możliwości przekształceń tych danych.

#### 3.1. Pożądane cechy tekstu

Dane nieustrukturyzowane, czyli tekst, występują w wielu formach, mogą mieć różny stopień niejednoznaczności, a wybór adekwatnego źródła jest dużo bardziej złożonym zagadnieniem niż wybór odpowiedniego szeregu czasowego. To właśnie w badaniach text mingowych sprawa wyboru źródła ma fundamentalne znaczenie i może zaważyć na powodzeniu całej pracy. Wybór nieodpowiednich danych wejściowych może być wąskim gardłem i spowodować, iż nawet najbardziej wyszukane metody estymacji zawiodą. Zatem konieczne wydaje się zastanowienie, jakie cechy powinny spełniać źródłowe dokumenty tekstowe, aby można było uznać je za poprawne. Wynikiem moich przemyśleń na temat idealnego tekstu jest poniższy zbiór reguł. Dla lepszego zrozumienia reguły te podzieliłem na trzy kategorie.

Pierwszy zbiór reguł przedstawia wiadomości, które wpływają na sytuację na giełdzie:

- Dokumenty tekstowe powinny być dobrane tematycznie. Najbardziej odpowiednie byłyby dane „biznesowe” zawierające wiadomości o treściach politycznych i ekonomicznych oraz dotyczące wydarzeń giełdowych.
- Powinny być one wysokiej jakości, obiektywne i rzetelne. Najlepiej, aby były publikowane przez znane i godne zaufania instytucje publiczne. Wymóg ten oznacza, iż nie powinno się korzystać z niektórych danych umieszczanych przez użytkowników i nie poddanych żadnej weryfikacji, zwłaszcza zaś ze wszystkich for internetowych oraz komentarzy.
- Każda wiadomość powinna mieć przyporządkowany dokładny punkt czasowy, w którym się pojawiła, najlepiej z dokładnością co do minuty. Oznacza to, iż nie można

wykorzystywać gazet, gdyż każdej wiadomości przez nie podanej możemy przyporządkować jedynie dzień publikacji.

Następny zbiór reguł dotyczy warunków specyficznych dla narzędzia, jakim jest text mining:

- Artykuły nie powinny opisywać przeszłości, zawierać analiz minionych wydarzeń ani sugestii, jak nie należało wówczas postępować. Powinny to być jedynie suche fakty, aktualne wydarzenia. Text mining jest stosunkowo prostym narzędziem, które nie analizuje kontekstu ani znaczenia wypowiedzi. Ponadto w rozdziale teoretycznym, dotyczącym hipotezy o efektywności rynków finansowych przedstawiono i uzasadniono pogląd, iż informacja bardzo szybko się starzeje i traci swoje komercyjne znaczenie. Dla przypomnienia rozważmy przykład dwóch artykułów o nagłówkach: „Nadzwyczajne spadki na azjatyckich giełdach” oraz „Skutki nadzwyczajnych spadków odnotowanych w poprzednim roku na azjatyckich giełdach.”. Widać, iż pierwsza informacja ma duże znaczenie i może być przyczyną zmiany cen akcji, natomiast druga nie. Istnieje duże niebezpieczeństwo, iż text mining nie rozróżni tych wiadomości.
- Artykuły powinny być proste w przekazie, mieć prostą składnię oraz powinny być napisane prostym językiem. Użycie przenośni, przykuwających uwagę dwuznaczności, ironizowanie sprawia, iż artykuł czyta się przyjemniej, jednak zmniejsza to szanse poprawnego „zrozumienia” przez maszynę.
- Artykuły nie powinny zawierać dziennikarskich komentarzy ani żadnych danych, które są wynikiem przemyśleń a nie opisem nowych wydarzeń. Dane takie wprowadzają jedynie niepotrzebny szum do badania. Wyjątkiem mogą być pewne komentarze bardzo wpływowych osób, np. przewodniczącego banku centralnego USA – System Rezerwy Federalnej. Jednakże łatwo się zgodzić, iż takie wypowiedzi są same w sobie wydarzeniem a nie komentarzem.

Kolejna grupa reguł dotyczy ograniczeń natury technicznej:

- Dane muszą być dostępne w formie elektronicznej. Liczba danych niezbędnych do wytrenowania uczenia maszynowego jest bardzo duża, co powoduje, iż ręczne ich wprowadzanie na potrzeby tej pracy jest praktycznie niewykonalne. Warunek ten w pierwszej chwili może wydawać się zbędny, istnieją jednak agencje takie jak PAP, w których większość zasobów archiwalnych utrzymywana jest w wersji papierowej.
- Koszt uzyskania danych jest również poważnym problemem. Znaczna część komercyjnych baz danych, zawierających dane bardzo przyzwoitej jakości, może okazać się zbyt droga na akademickie możliwości.

- Dane muszą pokrywać odpowiednio długi okres czasowy, przy czym odpowiednio długi okres nie powinien wyrażać się w miesiącach czy w latach. Stwierdzenie, iż trzy miesiące są niewystarczające, natomiast sześć już wystarczy, może nie być poprawne metodologicznie. Odpowiednio długi okres to taki, w którym ceny akcji na giełdzie zachowują się w sposób niejednorodny oraz występują znacznie różniące się od siebie okresy. Istnieje poważne niebezpieczeństwo, iż algorytm uczenia maszynowego wytrenowany jedynie w czasie hossy nie będzie dawał dobrych rezultatów podczas bessy. W podrozdziale 2.5, w którym dokonano porównania dotychczasowych publikacji, opisana została dokładniej krytyka zbyt krótkich szeregów czasowych. W tym miejscu należy jedynie przypomnieć, iż text mining potrafi prognozować nowe wydarzenia tylko wtedy, jeśli zaistniały podobne w przeszłości.

Oczywiście powyżej przedstawiono zbiór cech idealnych tekstów źródłowych, a w rzeczywistości należy szukać tekstów, które spełniają jak największą liczbę podanych wymienionych kryteriów. Najprostszym pomysłem jest użycie elektronicznych wersji gazet o tematyce finansowej, takich jak „Wall Street Journal”, „Parkiet” czy „Puls biznesu”. Biblioteka Uniwersytetu Warszawskiego udostępnia studentom pokaźny zbiór prasy w wersji elektronicznej nieodpłatnie. Jednakże prasa ta zawiera zbyt dużą liczbę komentarzy oraz artykułów nie będących wiadomościami, ponadto wiadomościom można co najwyżej przyporządkować dzień publikacji, brakuje natomiast dokładnego czasu pojawienia się wiadomości. Pewnym rozwiązaniem byłoby użycie serwisów takich jak wikinews<sup>1</sup> w wersji polskiej, ze względu na dostępność informacji, jednak jakość publikowanych tam danych, szczególnie pochodzących sprzed 2007 roku, jest niewystarczająca. Ponadto serwis ten zawiera zbyt dużą liczbę informacji „niebiznesowych”.

Wyżej wymienione kryteria spełniają w największym stopniu internetowe serwisy informacyjne znanych agencji informacyjnych. Niestety większość tych serwisów jest płatna w wysokości przekraczającej budżet tej pracy. Chlubne wyjątki to serwis Reuters oraz Money.pl, które zostaną opisane w następnym podrozdziale

---

<sup>1</sup> Wikinews. Wolne źródło informacji. On line. Dostęp listopad 2007. <http://pl.wikinews.org/>

### 3.2. Źródła danych

Dane tekstowe stanowiące podstawę tej pracy pochodzą z serwisów Reuters<sup>1</sup> oraz Money.pl<sup>2</sup>. Serwis Reutersa publikuje najważniejsze biznesowe informacje z Polski oraz ze świata. Serwis ten opisuje siebie w następujący sposób: „Reuters jest globalną firmą dostarczającą wiadomości i dane niezbędne do funkcjonowania rynków finansowych, mediów oraz przedsiębiorstw. Nasze informacje cechuje wiarygodność, szybkość, dokładność i bezstronność. Są one bardzo często podstawą podejmowania decyzji na całym świecie”<sup>3</sup>. Z opisu tego możemy wnioskować, iż jest to poszukiwany przez nas typ wiadomości – wiadomości, które są podstawą podejmowania decyzji przez inwestorów giełdowych. Dane zawierają 9363 wiadomości publikowane od początku stycznia 2006 do końca listopada 2007 roku. Czas publikacji jest określony co do minuty.

Drugim źródłem danych tekstowych jest portal Money.pl, zawierający wiadomości powiązane z Warszawską Giełdą Papierów Wartościowych. Na stronach „o nas” można przeczytać: „Money.pl to najpopularniejszy portal finansowy i biznesowy w polskim internecie. Korzysta z niego 1,98 mln użytkowników miesięcznie (Megapanel Gemius/PBI, próba: lipiec 2007 r.). Wielokrotnie nagradzany przez internautów oraz specjalistów (Webstarfestival, Teraz Internet) [...] Serwis finansowy www.money.pl to serwis, który dostarcza bieżących informacji ze świata finansów oraz z pogranicza polityki i biznesu. Czytelnik znajdzie tu codziennie najświeższe wiadomości, kursy i notowania, komentarze z rynku”<sup>4</sup>. Dane zawierają 12 906 wiadomości publikowanych od początku stycznia 2001 do końca listopada 2007. Czas publikacji jest określony co do sekundy. Zaletą serwisu Money.pl jest możliwość selekcji artykułów ze względu na rodzaj wiadomości, a zatem można odfiltrować komentarze oraz inne wydarzenia nie dotyczące giełdy.

Niewątpliwie jest rzeczą bardzo korzystną użycie dwóch źródeł informacji, ponieważ wzajemnie się dopełniają. Reuters podaje wiadomości makroekonomiczne z Polski i ze świata, informacje polityczne, informuje o szokach na innych giełdach, słowem przedstawia szeroki obraz gospodarki, natomiast Money.pl skupia się na wiadomościach z GPW dotyczących poszczególnych spółek. Wiadomości wykorzystane w tej pracy zostały pobrane za pomocą dwóch programów, które, symulując przeglądarkę, wchodziły na strony internetowe i zapisywały znalezione tam informacje.

---

<sup>1</sup> Reuters. The business of information. On line. Dostęp listopad 2007. <http://www.reuters.pl/news/>

<sup>2</sup> Serwis Money.pl. Wiadomości -> Giełda. On line. Dostęp listopad 2007. <http://www.money.pl/gielda/wiadomosci/>

<sup>3</sup> Reuters. O firmie. On line. Dostęp listopad 2007. <http://about.reuters.pl/firma/index.html>

<sup>4</sup> Serwis Money.pl. O portalu. On line. Dostęp listopad 2007. [http://firma.money.pl/o\\_money/o\\_portalu/](http://firma.money.pl/o_money/o_portalu/)

Warto również wspomnieć o bardzo ważnej decyzji, jaką podjąłem, to znaczy o użyciu jedynie nagłówków artykułów prasowych, a nie treści całych artykułów. Decyzja ta była w znacznej mierze zainspirowana pracą Thomasa<sup>1</sup>, który uważa, że nagłówki są wprost idealne do automatycznego przetwarzania. Treść nagłówków jest w wysokim stopniu standaryzowana przez wewnętrzne procedury wydawnictw. Dziennikarze są uczeni pisać w stylu „odwróconej piramidy”, która nakazuje między innymi podawanie informacji streszczających artykuł w nagłówku. Oczywiście cały artykuł zawiera więcej informacji niż sam nagłówek, ale nie jest już rzeczą tak oczywistą, czy zawiera więcej kluczowych informacji. Głównym argumentem przemawiającym za użyciem jedynie nagłówków są względy techniczne. Nagłówki przez swoją prostotę oraz wyłuskanie najważniejszych informacji spełniają kryteria danych odpowiednich dla celów text miningu wymienione w powyższym rozdziale. Zawierają one znacznie mniej odwołań do wydarzeń przeszłych, analiz, komentarzy i porównań niż pełne wiadomości. Pisane są również prostszym językiem, bez dwuznaczności oraz ironizowania.

Źródłem finansowych szeregów czasowych jest serwis bossa.pl<sup>2</sup>, zawierający bogatą bazę danych dotyczących przeszłych cen akcji. Wykorzystane w tej pracy dane obejmują kwotowania indeksu WIG20 oraz wszystkich dwudziestu spółek wchodzących w jego skład w okresie od 11 listopada 2000 do listopada 2007 roku. Choć kwotowania są dostępne od 2000 roku, należy pamiętać, iż ponad połowa spółek wchodząca w skład indeksu WIG20 pojawiła się na giełdzie po 2000 roku, czyli jej dane pochodzą z późniejszego okresu. Przykładem jest Bioton, który zadebiutował 31 marca 2005 roku oraz czeski koncern energetyczny CEZ, który zadebiutował 25 października 2006 roku.

### 3.3. Studium przypadku

Po przedstawieniu źródeł danych konieczne jest, jak zostało to podkreślone we wstępie, sprawdzenie adekwatności danych wejściowych. W rozdziale tym należy przedstawić fragmenty źródeł oraz zweryfikować, czy jest to ten rodzaj informacji, który może „ruszyć” giełdę, oraz czy wiadomości te są podane w wystarczająco przystępny sposób dla algorytmów przetwarzania tekstów. Poniższy wyciąg przedstawia wiadomości z dnia 27 lutego 2007 roku. Tego dnia nastąpił jeden z większych szoków na giełdzie od czasu 11 września – korekta chińska.

---

<sup>1</sup> J. Thomas: op.cit.

<sup>2</sup> Serwis bossa.pl. Katalog z danymi. On line. Dostęp listopad 2007. <http://bossa.pl/pub/intraday/omega/cgl/?C=N;O=A>

Tablica 2. Wyciąg z danych źródłowych, nagłówki wiadomości prasowych z serwisu Reuters z dnia 27 lutego 2007 do godziny 12:00

11:21	GPW - PIERWSZY FIXING: W notowaniach jednolitych WIRR spadł o 3,82 proc.
11:08	CEZ chce wydać 928 mln USD na elektrownie w Czechach
11:06	GPW - PIERWSZY FIXING: W notowaniach jednolitych 2 spółki w dół
10:59	Fixing NBP: dolar 2,9620 zł, euro 3,9139 zł
10:52	Iran i spadające zapasy w USA umacniają ceny ropy
10:00	Złoty i nastroje na świecie osłabiły dług
09:35	Koncerny wydobywcze ciągną w dół giełdy Europy
09:23	GPW - NOTOWANIA CIĄGŁE: Na otwarciu WIG20 spadł o 1,50 proc.
09:09	Nastroje na świecie pociągną GPW w dół-maklerzy
09:05	Giełda w Chinach spadła najbardziej od 10 lat
08:52	Złoty traci w oczekiwaniu na doniesienia z RPP
08:19	ATM zrezygnował z przejęcia krajowego operatora
08:04	PRZEGLĄD PRASY -- 27 lutego
07:24	NAJWAŻNIEJSZE WYDARZENIA EKONOMICZNE NA ŚWIECIE-27 II
07:24	Japońscy eksporterzy osłabili giełdę w Tokio
07:00	Wall Street spadło pod ciężarem drogiej ropy
06:30	KALENDARIUM RYNKU KAPITAŁOWEGO
06:30	*** NAJWAŻNIEJSZE WIADOMOŚCI - PONIEDZIAŁEK ***

Źródło: Reuters. The business of information. On line. Dostęp listopad 2007. <http://www.reuters.pl/news/>

Analizując powyższy wyciąg, można uznać, iż większość przedstawionych w nim wiadomości dotyczy wydarzeń makroekonomicznych i może mieć wpływ na cenę akcji. Istnieje jednak kilka wyjątków, np. nagłówki: „Przeгляд prasy” oraz „Kalendarium rynku kapitałowego”, które wydają się nie mieć wpływu na zachowanie uczestników giełdy. Przykłady te zostaną opisane w następnych częściach tego rozdziału. Najważniejsza wiadomość, dotycząca korekty chińskiej, pojawiła się o godzinie 9:05. Warto podkreślić, że nagłówek tej wiadomości jest bardzo zwięźle skonstruowany. Każde słowo ma ważne znaczenie i oddaje sedno wydarzenia. Na podstawie tych słów można by utworzyć bardzo prostą, a zarazem skuteczną regułę: „Indeks spada, gdy występują słowa giełda, Chiny, spadać”.

Dla porównania przytoczono podobny wyciąg z serwisu Money.pl z tego samego dnia.

Tablica 3. Wyciąg z danych źródłowych, nagłówki wiadomości prasowych z serwisu Money.pl z dnia 27 lutego 2007

18:20:43	Bioton wchodzi w produkty krwiopochodne
17:15:30	KNF zatwierdziła prospekt emisyjny TelForceOne
15:28:03	Czarny wtorek na GPW
14:17:07	Bioton chce mieć 400 mln zł w 2007 r.
12:13:28	Polscy akcjonariusze chcą odkupić akcje Polkomtela od Vodafone
11:42:12	Fota nie wykonała prognozy zysku netto na 2006 rok
11:05:25	Zysk netto Polimeksu niższy od oczekiwań
10:52:07	Wyniki Biotonu niższe niż się spodziewano
10:49:03	ATM zrezygnował z przejęcia jednego z krajowych operatorów telekomunikacyjnych
09:20:03	Spółka zależna Mostostalu Warszawa ma umowy na 110,9 mln zł

Źródło: Serwis Money.pl. Wiadomości -> Giełda. On line. Dostęp listopad 2007.  
<http://www.money.pl/gielda/wiadomosci/>

Z powyższego fragmentu wynika, że wiadomości te zawierają więcej informacji dotyczących polskiej giełdy i spółek wchodzących w skład GPW. Co prawda nie ma wśród nich bezpośredniej informacji dotyczącej chińskiej korekty, jednak można wnioskować, że wiadomość z godziny 15:28:03, „Czarny wtorek na GPW”, jest skutkiem tej korekty.

Ważnym zagadnieniem, pojawiającym się w większości prac naukowych, jest pytanie o stopień ingerencji w postać danych źródłowych. Dane źródłowe użyte w tej pracy wydają się dobrej jakości, co eliminuje konieczność ich przekształcania. Wszelkie modyfikacje byłyby więc po części wynikiem subiektywnych decyzji, co mogłyby mieć negatywny wpływ na wynik badania. Ponieważ praca ta ma pionierski charakter, nie chciałem ingerować w postać danych źródłowych<sup>1</sup>, stwarzając tym samym obiektywny punkt wyjścia do dalszych badań.

Prawdopodobnie pierwszą nasuwającą się na myśl modyfikacją jest usunięcie takich artykułów jak „Kalendarium rynku kapitałowego” bądź „Przegląd prasy”. Ponieważ analizowane są jedynie nagłówki wiadomości, każdego dnia mamy ten sam nagłówek, co jest zbędne. Kontrargumentem może być stwierdzenie, że pewna liczba niepotrzebnych informacji, stanowiąca biały szum, jest do zaakceptowania w badaniu, natomiast ich usunięcie mogłoby zniekształcić wynik. Ponadto po przeprowadzeniu badania mamy gotową frazę, dzięki której możemy łatwo sprawdzić, czy nie powstała jakaś błędna reguła, która prognozuje w oparciu o nic nie znaczące nagłówki.

Druga modyfikacja mogłaby wiązać się z faktem, iż pewne artykuły w serwisie Reuters występują parokrotnie, za każdym razem wzbogacone o dodatkowy opis, co zostało pokazane poniżej:

<sup>1</sup> Pomijając oczywiście takie kwestie techniczne jak uzgodnienie wspólnego standardu kodowania znaków czy odpowiednia konwersja dat do spójnej postaci

Tablica 4. Przykład zduplikowanej wiadomości

2007-11-07 07:49 Zysk netto BZ WBK w III kw. Wyniósł 226 mln zł
2007-11-07 11:19 OPIS1-Zysk netto BZ WBK w III kw. Wyniósł 226 mln zł.

Źródło: Opracowanie własne na podstawie danych z serwisu Reuters

Artykuły odwołujące się do wydarzeń przeszłych nie spełniają kryteriów podanych w rozdziale 3.1. Istnieje też duże prawdopodobieństwo, że narzędzia text miningu potraktują dodatkowy opis jako drugie niezależne zdarzenie. Jednakże, podobnie jak w poprzednim przypadku, można argumentować, iż częstotliwość pojawiania się publikacji o tym samym wydarzeniu ma istotne znaczenie, ponieważ świadczy o zainteresowaniu tym zjawiskiem.

Jedynie modyfikacje źródła, jakich dokonano w tej pracy, to wprowadzenie wspólnego kodowania znaków, zamiana myślników na spacje oraz ręczne rozwinięcie skrótów. Myślniki zostały zastąpione przez spacje, ponieważ narzędzia przetwarzające tekst mylnie interpretowały wyrazy połączone myślnikiem jako jeden wyraz. Np. w poniższym zdaniu: „Rosyjskie dostawy gazu muszą być stabilne–Niemcy”, narzędzia identyfikowały ostatnie wyrazy jako jeden wyraz „stabilne–niemcy”.

Niestety aktualna wersja narzędzi text miningu traktowała każdą kropkę, nawet kropkę przy skrócie, jako znak kończący zdanie. Mogło to powodować błędną dezambiguację części zdania oraz błędne traktowanie wyrazów skróconych. Na przykład skrót „proc.” został zinterpretowany jako dopełniacz liczby mnogiej rzeczownika proca, a zatem dla maszyny był zupełnie innym „bytem” niż słowo procent.

Ostatnią dokonaną weryfikacją było sprawdzenie, jak często wiadomości, od których oczekujemy, iż mogłyby wpływać na cenę akcji, są zawarte w danych wejściowych. Prosty przykład to analiza wiadomości dotyczących produktu krajowego brutto i formy, w jakiej informacja ta została zapisana. Słowo PKB wystąpiło w serwisie Reuters 132 razy w ciągu 72 dni. Ilustruje to poniższy wyciąg:

Tablica 5. Zbiór wiadomości zawierających słowo PKB

20070830, 14:50 OPIS1-Tempo wzrostu PKB USA w II kw. wyniosło 4 proc.
20070830, 11:56 OPIS2-Wzrost PKB wciąż wysoki, pomogły zapasy firm
20070917, 13:08 Wzrost PKB w III kw. przekroczy 6 proc.-Woźniak
20071011, 11:08 Wzrost PKB w E13 wyniósł w II kw. 0,3 proc. kw/kw
20071016, 08:04 OPIS1-Deficyt w 2007 i 2008 spadnie do 3 proc. PKB-MF
20071108, 12:03 W 2007 PKB Chin wzrośnie o ponad 11 proc.-Bank Chin

Źródło: Opracowanie własne na podstawie danych z serwisu Reuters

Widać zatem, iż serwis ten zawiera znaczną liczbę wydarzeń ekonomicznych.

Podobny wniosek wynika z analizy takich wyrażen jak „stopy procentowe”, „bezrobocie” i inne, co widać wyraźnie z przedstawionej poniżej tabeli:

Tabela 3. Częstotliwość występowania kluczowych słów w danych źródłowych

<b>Kluczowe słowo</b>	<b>Reuters</b>	<b>Money.pl</b>
stopy procentowe	325	13
bezrobocie	58	0
PKO	60	149
Bioton	46	66
Kaczyński	11	2
Bush	11	0
- i Bernanke	5 (4 jednego dnia)	0
PiS	90	4
- i sondaż	40	
Irak	3	1
Iran	44	0
- i ropa	24	0

Źródło: Opracowanie własne na podstawie danych z serwisów Reuters oraz Money.pl

Podsumowując, można stwierdzić, że użycie dwóch źródeł wiadomości jest bardzo korzystne, ponieważ źródła te wzajemnie się dopełniają. Reuters zawiera więcej wiadomości (w tym dotyczących wydarzeń makroekonomicznych) z polski i ze świata, słowem przekazuje szeroki obraz gospodarki, natomiast Money.pl preferuje wiadomości z GPW dotyczące poszczególnych spółek. Wszystkie one dostarczają niezbędnych do prognozowania cen akcji informacji. Ponadto wiadomości te podane są w zwięzły i prosty sposób, co czyni je odpowiednimi do automatycznego przetwarzania tekstu. Analizując powyższe fragmenty serwisów prasowych, można również wyciągnąć wniosek, iż dane te zawierają stosunkowo niewielką liczbę informacji nieistotnych.

Ostatnią rzeczą, na którą trzeba zwrócić uwagę w tym rozdziale, są fixingi oraz okresowe informacje zawierające streszczenia najważniejszych wydarzeń z rynków światowych. Dobrym przykładem może być wiadomość: „Wall Street spadło pod ciężarem drogiej ropy”. Informację o spadku indeksu NYSE można poznać, korzystając z tradycyjnych metod analizy technicznej, jednak o przewadze danych tekstowych nad tradycyjną analizą techniczną świadczy fakt, iż możemy poznać przyczynę zjawiska, a nie tylko samo zjawisko. Znając przyczyny zjawiska, można trafniej prognozować ceny akcji w przyszłości, jeśli wystąpi ta sama przyczyna. Ponadto wiadomości tekstowe częściej pojawiają się w

nieoczekiwanych sytuacjach, a tym samym pozwalają w pewnym sensie mierzyć „nieoczekiwaność” wydarzenia.

## ROZDZIAŁ IV

### Praca badawcza

*"Everything should be made as simple as possible, but not any simpler."  
-- Albert Einstein*

Byłoby pożądanym, aby czytelnik analizujący ten rozdział wziął pod uwagę pewną specyfikę działu nauki zajmującego się automatycznym przetwarzaniem tekstu. Jest to bardzo młoda, dynamicznie rozwijająca się dziedzina, co powoduje, iż stanowi wdzięczny temat i duże wyzwanie dla badacza, z drugiej strony jednak jej „młodość” sprawia, że nie wypracowano jeszcze powszechnie uznanych narzędzi oraz ustalonych metodologii. Cały czas ukazują się nowe prace mające udowodnić, iż metoda XYZ jest lepsza od metody ABC stosowanej w poprzednim badaniu. Problem polega na porównaniu tych prac oraz wybraniu metody najlepszej.

W moim przypadku trudność dokonania wyboru odpowiedniej metodologii była tym wyraźniejsza, że moja praca jest jedną z pierwszych prac badających ceny aktywów na Warszawskiej Giełdzie Papierów Wartościowych za pomocą wiadomości tekstowych. Poza tym jest jedną z pierwszych, w której wykorzystano najnowsze oprogramowanie z zakresu inżynierii lingwistycznej języka polskiego do zastosowań nielingwistycznych. Z tych to powodów wybrałem rozwiązania prostsze, lecz cieszące się większym uznaniem, a nie metody bardziej zaawansowane, lecz niedostatecznie zweryfikowane. Celem moim jest stworzenie modelu wystarczająco prostego, aby mógł służyć jako punkt wyjścia do dalszych badań. Jednocześnie, pamiętając o słowach Alberta Einsteina, chciałbym stworzyć model wystarczająco zaawansowany, aby miał on możliwość dokonywania trafnych predykcji.

Szukając najodpowiedniejszej pracy wzorcowej, zdecydowałem się na wybór publikacji zespołu kierowanego przez Lavrenkę<sup>1</sup> oraz publikacji G. Gidófalviego i C. Elkana<sup>2</sup>. Ta ostatnia ukazała się najpóźniej, uwzględnia więc krytykę dotyczącą prac poprzednich. W wymienionych powyżej publikacjach przedstawiony został sposób automatycznej konstrukcji słownika fraz, podczas gdy w innych pracach słowniki były konstruowane przy pomocy ekspertów. Zaletą zespołu Lavrenki jest fakt, iż w swojej pracy umieścili relatywnie dużo szczegółów.

Moim dążeniem jest uniknięcie błędów wytkniętych u poprzedników, a przedstawionych w rozdziale 2.5 (między innymi pominięcie kosztów transakcyjnych). Nie

---

<sup>1</sup> V. Lavrenko, M. Schmill, D. Lawrie [et al.]: *op.cit.*

<sup>2</sup> G. Gidófalvi, C. Elkan: *Using News Articles to Predict Stock Price Movements*. California, 2001.

chciałbym też uprawiać przerzucania danych (data mining), czyli szacowania parametrów modelu na tej samej grupie danych, na której zostanie dokonana ewaluacja wyników. Aby rozwiązać ten problem, już na początku badania cały zbiór danych podzieliłem na dwie równe części – treningową oraz testową. Wszelkie decyzje dotyczące wyboru najlepszego algorytmu oraz parametrów modelu podejmowałem, wykorzystując jedynie zbiór treningowy.

#### **4.1. Plan badania**

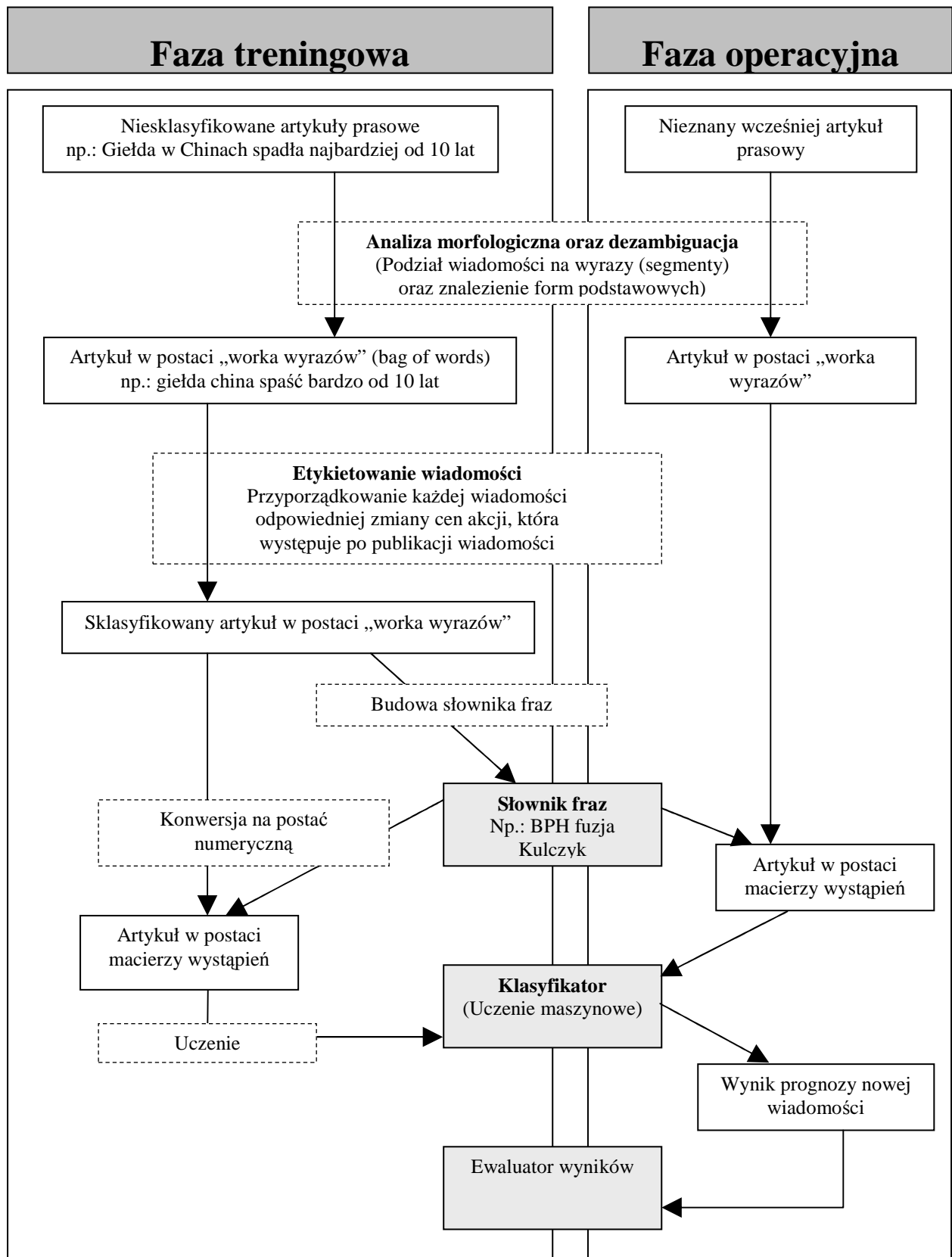
Uważna lektura opisywanych w rozdziale 2. publikacji pozwoliła na wyodrębnienie z tych prac ramowego planu czynności, stanowiącego zarazem sformalizowany opis działań, które pokrótce zostały przedstawione w dziale teoretycznym. Plan ten składa się zatem z następujących etapów:

- 1) Zdefiniowanie i przygotowanie danych wejściowych;
- 2) Dokonanie analizy morfologicznej w celu przekształcenia dokumentów na „worek wyrazów”;
- 3) Przyporządkowanie każdemu artykułowi odpowiedniego wyniku finansowego;
- 4) Stworzenie słownika fraz;
- 5) Budowa macierzy wystąpień, czyli przekształcenie artykułów na postać ilościową;
- 6) Zastosowanie uczenia maszynowego;
- 7) Przedstawienie wyników badania.

W następnych podrozdziałach zostanie przedstawiony zwięzły opis dotychczasowych rozwiązań, prezentowanych w omawianych wcześniej publikacjach, szczegóły dotyczące zastosowanych przekształceń oraz ewentualnie prosty przykład ilustrujący cel danego etapu badań. W miarę możliwości będą podane alternatywne, często bardziej zaawansowane, sposoby rozwiązania danego problemu, które mogą służyć jako punkt wyjścia do następnych prac badawczych.

Na poniższym schemacie został przedstawiony diagram sekwencji dla wiadomości z grupy treningowej (faza treningowa) oraz grupy testowej (faza operacyjna), który może służyć jako wizualizacja powyższego planu. Prostokąty o ciągłych liniach oznaczają stany, natomiast o liniach przerywanych – czynności.

Rys. 1. Przegląd działania algorytmu



Źródło: Opracowanie własne na podstawie pracy: M.A. Mittermayer, G.F. Knolmayer: *Text Mining Systems for Market Response to News: A Survey*. Bern 2006.

## 4.2. Zdefiniowanie i przygotowanie danych wejściowych

Wybranie odpowiednich danych wejściowych jest nietrywialnym zadaniem, trzeba bowiem odpowiedzieć na ważne pytanie – które spółki bądź indeksy mają być zmiennymi objaśnianymi, a które wiadomości zmiennymi objaśniającymi. Często to pytanie sprowadza się do wyboru, czy za pomocą wiadomości prasowych modelować jeden indeks czy poszczególne spółki notowane na giełdzie. W przypadku modelowania spółek giełdowych konieczna jest kolejna decyzja, czy dla każdej spółki wyznaczyć osobny rozkład słów oraz związanych z nią reguł klasyfikacji, czy tworzyć jeden zestaw reguł klasyfikacyjnych dla wszystkich spółek. Pierwsze prace analizujące wpływ wiadomości na ceny akcji modelowały indeksy, jak np. Hang Seng Index (HSI)<sup>1</sup>, natomiast w kolejnych publikacjach autorzy podejmowali się prognozy dla każdej ze spółek oddzielnie. W pracy tej chciałem dokonać weryfikacji tych metodologii, dzieląc wiadomości na ekonomiczne – mające największy wpływ na indeks WIG20, oraz dotyczące spółek – mające wpływ na notowania konkretnych spółek. Podział ten pozwoli znaleźć odpowiedź na pytanie, czy lepiej jest badać komunikaty dotyczące konkretnych spółek czy sytuacji ekonomicznej na podstawie wiadomości prasowych.

Założenie, iż wiadomość dotyczy danej spółki giełdowej, jeśli jej nazwa występuje w nagłówku artykułu, uwzględniając potencjalną odmianę tej nazwy, jest dość rozsądne. Należy przy tym pamiętać, iż nagłówki pisane są w taki sposób, aby w maksymalnym stopniu reprezentowały treść wiadomości. Algorytm zakwalifikowaniu wiadomości do odpowiednich grup składa się z następujących kroków:

- Dla każdej wiadomości należy znaleźć listę spółek, których nazwa znajduje się w nagłówku wiadomości. Trzeba również wziąć pod uwagę odmianę nazw spółek, np. dla spółki Żywiec, uwzględnić wiadomości zawierające słowa „Żywca”, „Żywcem”, „Żywcowi”;
- Jeżeli nagłówek wiadomości nie zawiera nazwy żadnej spółki, trzeba uznać, że jest to wiadomość ekonomiczna;
- Jeżeli w nagłówku wiadomości wymieniono spółkę należącą do indeksu WIG20, należy zakwalifikować ją jako wiadomość WIG20;
- W innych wypadkach ignoruj wiadomość (warunek ten polega na ignorowaniu wiadomości dotyczących spółek, które nie należą do WIG20).

---

<sup>1</sup> B. Wüthrich, S. Leung, D. Peramunetilleke [et al.]: *op.cit.*

Jestem świadomy, iż zaproponowany przeze mnie podział jest daleko idącym uproszczeniem, jednakże uważam, iż korzyści wynikające z tego uproszczenia są znaczne. Zanim przejdę do omówienia korzyści wynikających z zastosowania tego podziału, chciałbym wyjaśnić jego zasadność. Dane finansowe są często opisywane jako zawierające „patologiczną liczbę niepotrzebnych szumów” lub, jak niektórzy twierdzą, tylko biały szum. We wstępie do niniejszej pracy, wspomniano o wrodzonej tendencji ludzi do wyszukiwania zależności i wzorców, nawet gdy one obiektywnie nie istnieją. W każdym nagłówku prasowym można się doszukiwać wpływu na aktualną sytuację na giełdzie, ale trzeba wziąć pod uwagę fakt, że istnieje prawdopodobieństwo, iż większość informacji, np. wypowiedzi Busha bądź Bernanke, nie ma rzeczywistego wpływu na ceny akcji w Polsce. Oddzielenie wiadomości dotyczących konkretnych spółek od wiadomości ekonomicznych pozwoli stworzyć grupę wiadomości, które mają potencjalnie większy wpływ na ceny tychże akcji. Umożliwi również weryfikację pewnych hipotez oraz wyciągnięcie wniosków w przypadku, gdy okaże się, iż wiadomości ekonomiczne zawierają zbyt dużą liczbę szumów lub są nieistotne.

W pracy tej wszystkie spółki z grupy WIG20 zostały przetworzone przy użyciu jednego modelu. Oczywiście lepiej byłoby modelować każdą ze spółek oddzielnie, gdyż wiadomości dotyczące każdej z nich zawierają inne słowa kluczowe, i stosować dla każdej z nich odrębne reguły klasyfikacji. Głównym powodem budowy tylko jednego modelu, a w konsekwencji nierozróżnianie spółek, jest zbyt mała liczba wiadomości, które ich dotyczą. Zbudowanie rozsądnego modelu przetwarzającego tekst naturalny wymaga znacznej liczby fraz, z których każda występuje przynajmniej dwucyfrową liczbę razy.

Podsumowując, w podrozdziale tym został przedstawiony podział danych wejściowe na dwie niezależne grupy. W każdym z następnych etapów badań, budowa odpowiednich modeli oraz częstotliwości występowania wyrazów będzie dokonywana w sposób niezależny dla obydwu grup.

### **4.3. Analiza morfologiczna**

W dotychczasowej literaturze niewiele miejsca poświęcano zagadnieniu analizy morfologicznej. Najczęściej temat ten sprowadzał się do wzmianki, że dokumenty były dzielone na słowa za pomocą znaków niealfanumerycznych (wszelkie znaki przestankowe, spacje, liczby), że przeprowadzany był stemming oraz zamiana wszystkich liter na małe.

Kolejności słów nie była brana pod uwagę – dokumenty traktowano jako nieuporządkowane zbiory słów.

Takie podejście nie pasuje jednak do naszego języka, gdyż język polski powszechnie jest uważany za wybitnie trudny, i to nie z powodu fonetyki (istnieje wiele języków o znacznie bardziej skomplikowanej wymowie), ale z uwagi na skomplikowaną morfologię. W rozdziale drugim przedstawione zostały rozważania dotyczące złożoności tak prostej procesu jak podział zdania na słowa oraz porównanie lematyzacji ze stemmingiem. Ze względu na ową złożoność języka postanowiłem w pracy tej użyć narzędzi stworzonych przez lingwistów, w przeciwieństwie do autorów poprzednich prac, którzy używali prostych programów (stworzonych przez informatyków) działających na poziomie liter.

Do przeprowadzenia analizy, czyli odpowiedniej konwersji strumienia znaków na segmenty (odpowiedniki słów), ustalenia formy wyrazu, charakterystyki gramatycznej każdej formy wyrazowej oraz postaci wykładnika formy podstawowej leksemu, użyłem programu Morfeusz. Do dezambiguacji, której celem jest wybór najtrafniejszej formy wyrazowej z listy form przedstawionej przez Morfeusza na podstawie kontekstu zdania oraz technik statystycznych, użyłem dezambiguatora TaKiPi<sup>1</sup>.

Dla zastosowań przedstawionych w niniejszej pracy najbardziej użyteczna okazała jest segmentacja oraz uzyskanie formy podstawowej leksemu, co pozwala na „skojarzenie” tych samych wyrazów różniących się jedynie formą gramatyczną, np: „giełda”, „giełdzie”, „giełd”.

Poniżej przedstawiłem rezultat przetworzenia przytaczanych wcześniej artykułów:

Tablica 6. Wynik analizy morfologicznej wybranych artykułów

09:09 Nastroje na świecie pociągną GPW w dół-maklerzy → nastrój (na) świat pociągnąć GPW (w) dół makler (.)
09:05 Giełda w Chinach spadła najbardziej od 10 lat → giełda (w) china spaść bardzo (od) 10 rok (.)

Źródło: Opracowanie własne

Wyrazy w nawiasach zostały usunięte, za pomocą stoplisty – listy wyrazów semantycznie pustych.

<sup>1</sup> Piasecki M.: Polish Tagger TaKIPI: Rule Based Construction and Optimisation. *Task Quarterly*. 2007, tom 11, s. 151-167.

#### 4.4. Połączenie danych tekstowych z szeregami czasowymi

Połączenie danych tekstowych z szeregami czasowymi oznacza przyporządkowanie każdej wiadomości odpowiedniej zmiany ceny aktywa, występującej po publikacji wiadomości. W literaturze opisywano wiele sposobów realizowania tej czynności – właściwie w każdej pracy autor przedstawiał swój własny pomysł. Przed wyborem konkretnego rozwiązania należy zastanowić się nad odpowiedzią na dwa pytania. Pierwsze z nich to w jakim czasie dana wiadomość zostaje „wchłonięta” przez rynek, czyli po jakim czasie ceny akcji odzwierciedlają poprzedni stan wiedzy powiększony o nową wiadomość? Znając długość czasu rozchodzenia się wiadomości, należy zastanowić się nad postacią funkcji, która przekształca zbiór zmian cen akcji po publikacji wiadomości na jedną liczbę rzeczywistą.

W pracy „Response of Hourly Stock Prices”<sup>1</sup> P. C. Jain za pomocą szeregu regresji testuje wpływ publikacji wiadomości na ceny akcji. Według niego efekty publikacji nowych wiadomości rozchodzą się na rynku w czasie jednej godziny. Autor sprawdza swoje wyniki za pomocą regresji, w których każdą godzinę reprezentuje jeden współczynnik  $\beta_i$ , co powoduje, iż dokładność tego badania jest ograniczona do godzin. Dlatego też wyniki te należy interpretować jako górne ograniczenie prędkości rozchodzenia się informacji. Ponadto praca ta pochodzi z roku 1988, a od tego czasu nastąpił znaczny rozwój technik komunikacji i należy przypuszczać, iż czas ten uległ skróceniu.

W pracy „Intraday Market Dynamics Around Public Information Arrivals”<sup>2</sup> został przedstawiony pogląd, iż „wchłonięcie” przez rynek nowych informacji trwa około 30 – 40 minut i że największe znaczenie mają informacje o zyskach. Co więcej, zmiany cen rozpoczynają się już kilkanaście minut przed publikacją wiadomości. Autor tłumaczy to faktem, że pewne osoby mogły znać informację wcześniej lub że serwis Reuters, który też był użyty jako źródło wiadomości, mógł nie zawsze być pierwszym źródłem informacji.

Drugie wspomniane wcześniej pytanie to pytanie o postać funkcji przekształcającej zmiany w danym oknie czasowym o ustalonej długości w jedną wartość, będącą ilościowym efektem wpływu wiadomości na cenę akcji. Najprostszym rozwiązaniem jest bez wątpienia funkcja liniowa, w której finansowym efektem publikacji wiadomości jest procentowa zmiana cen akcji zachodząca od momentu publikacji wiadomości do końca okna czasowego. Jednakże można wyobrazić sobie inne nieliniowe funkcje lub proste przekształcenia ważone za pomocą czasu, który upłynął od chwili publikacji wiadomości. Przykładem może być

---

<sup>1</sup> P. C. Jain: Response of Hourly Stock Prices and Trading Volume to Economic News. *Journal of Business*. 1988, tom 61, s. 219-31.

<sup>2</sup> A. Rinaldo: *Intraday Market Dynamics Around Public Information Arrivals*. Fribourg, 2003.

funkcja przyporządkowująca relatywnie większą wagę do zmian zaobserwowanych w pierwszych minutach niż w ostatnich.

W swojej pracy wybrałem metodę przedstawioną w rozdziale 2.3 w publikacji „Currency Exchange Rate Forecasting from News Headlines”<sup>1</sup>. Oznacza to, iż przyjąłem, że efektem finansowym danej wiadomości jest procentowa zmiana wartości akcji, która nastąpiła w ciągu godziny od momentu publikacji wiadomości. Godzina wydaje się okresem zbyt długim, prawdopodobnie pół godziny okazałoby się wystarczające, jednak według mnie korzyści wynikające z tego wyboru są wyższe niż koszty. Korzyścią jest „bezpieczeństwo” wyboru – dłuższy okres czasu pozwala precyzyjniej uchwycić efekt pojawienia się wiadomości, koszt to niebezpieczeństwo wprowadzania pewnej ilości informacji nieistotnej, czyli szumu.

Efekt oddziaływania wiadomości ekonomicznych jest mierzony za pomocą indeksu WIG20, natomiast efekt wiadomości dla każdej spółki z grupy WIG20 za pomocą odpowiedniego szeregu czasowego cen akcji tej spółki. Zarówno wiadomości tekstowe, jak i szeregi czasowe cen spółek mają rozdzielczość jednej sekundy, więc możliwe jest precyzyjne określenie wyniku finansowego. Źródłowe szeregi czasowe, niosące informację o cenach akcji, zawierają jedynie listę wszystkich transakcji w danych punktach czasu. Wartość aktywa  $i$  w danej chwili  $t$  została wyliczona jako średnia z ceny zamknięcia ostatniej transakcji oraz ceny otwarcia następującej transakcji, co można zapisać następująco:

$$cenaAktywa_{i,t} = \frac{close_{i,t} + open_{i,t}}{2}$$

Ilościowy efekt wpływu wiadomości na cenę akcji został zdefiniowany za pomocą poniższego wzoru:

$$zmianaAktywa_{i,t} = \frac{cenaAktywa_{i,T} - cenaAktywa_{i,t}}{cenaAktywa_{i,t}}$$

gdzie:  $i$  – oznacza  $i$ -te aktywo,

$t$  – czas publikacji wiadomości,

$T$  – koniec oddziaływania wiadomości równy  $t + 1$  godzina,

---

<sup>1</sup> D. Peramunetilleke, R.K. Wong: *op.cit.*

$close_{i,t}$  – ostatnia cena zamknięcia dla aktywa  $i$  przed/lub w czasie  $t$ ,

$open_{i,t}$  – pierwsza cena otwarcia dla aktywa  $i$  po/lub w czasie  $t$ .

Jeżeli wiadomość była opublikowana po zakończeniu sesji bądź w dniu wolnym (uwzględnione zostały święta narodowe oraz fakt, iż godziny sesji zostały wydłużone o godzinę w październiku 2005 r.) czas publikacji wiadomości  $t$  był przesuwany na początek otwarcia sesji w następnym dniu roboczym, a koniec wpływu wiadomości  $T$  ulegał przesunięciu do  $t + 1$  godzina. Jeżeli wiadomość była opublikowana w trakcie sesji, później jednak niż po godzinie przed jej zamknięciem, czas oddziaływania  $T$  przenosił się na następny dzień, w którym odbywała się sesja, o tyle minut, ile zabrakło przed zamknięciem sesji w poprzednim dniu roboczym.

Poniższa tabela przedstawia liczbę artykułów w każdej grupie oraz cechy grupy dla zbioru treningowego:

Tabela 4. Zmiany cen akcji po publikacjach wiadomości – miary statystyczne

Cecha	Spółki WIG20	Wiadomości ekonomiczne
Liczba wiadomości	1943	10 283
Wartość oczekiwana <sup>1</sup>	-0,0721%	-0,0021%
Standardowe odchylenie <sup>1</sup>	1,0977%	0,5925%
Nieistotne <sup>2</sup>	142 (7,37%)	270 (2,57%)

Źródło: Opracowanie własne

Z tabeli tej wynika, iż wiadomości dotyczące spółek WIG20 mają większy wpływ, mierzony za pomocą odchylenia standardowego, na cenę akcji niż wiadomości ekonomiczne. Wniosek ten jest zgodny z intuicją i rozważaniami zawartymi w tej pracy. Należy również zauważyć, iż wiadomości mają niewielki ujemny wpływ na ceny akcji. Można to tłumaczyć tym, iż nowe wiadomości zwiększają niepewność, czyli ryzyko, a tym samym mogą mieć niewielki negatywny wpływ na ceny akcji. Inne wytłumaczenie ujemnej wartości oczekiwanej to fakt, iż pewna część wiadomości jest publikowana poza godzinami pracy giełdy, a w konsekwencji występuje efekt pierwszej rannej godziny, w której ceny akcji potrafią nieznacznie spaść.

<sup>1</sup> Wartość oczekiwana oraz wariancja zmiany ceny akcji w ciągu 60 minut od publikacji wiadomości.

<sup>2</sup> Wiadomość była nieistotna, jeśli cena akcji po 60 minutach od publikacji nie zmieniła się o więcej niż 0,01%

#### 4.5. Istotność wiadomości

Mając powyższe dane, można dokonać weryfikacji postawionej we wstępie tezy, mówiącej, iż wiadomości prasowe zawierają dodatkową informację, której nie można uzyskać analizując jedynie szeregi czasowe. W tym celu sporządziłem prosty test sprawdzający, czy sam fakt pojawienia się wiadomości (bez jej analizy) jest istotny.

Termin zmiana ceny akcji można zdefiniować jak w poprzednim rozdziale, czyli jako relatywną zmianę cen akcji w ciągu godziny. Formalizując problem, załóżmy, że zmiany cen akcji mają rozkład  $F: \Delta x_1, \Delta x_2, \dots, \Delta x_n \sim F$ , natomiast zmiany cen akcji w momentach  $t$ , dla których opublikowano wiadomość, mają rozkład  $G: \Delta y_1, \Delta y_2, \dots, \Delta y_n \sim G$ . Zakładając, iż  $F$  i  $G$  jest rozkładem normalnym, problem sprowadza się do znanego problemu w statystyce – testu istotności dla dwóch średnich prób bez założenia równości wariancji w tychże próbach.

Hipoteza  $H_0$ : Obie populacje pochodzą z tego samego rozkładu (czyli sam fakt pojawienia się wiadomości nie wpływa na zmiany cen).

Hipoteza alternatywna  $H_1$ : Obie populacje pochodzą z różnych rozkładów.

Test istotności jest zdefiniowany następująco:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Wyniki tego testu wyznaczone za pomocą procedury Welcha-Satterthwaite'a znajdują się w poniższej tabeli:

Tabela 5. Zmiany cen akcji – miary statystyczne

Cecha	Spółki WIG20	Wiadomości ekonomiczne
Wartość oczekiwana <sup>1</sup>	0,0144%	0,0082%
Standardowe odchylenie <sup>1</sup>	0,974%	0,567%
Statystyka t	-3,218	-1,7425
Poziom istotności	0,13%	8,63%

Źródło: Opracowanie własne

<sup>1</sup> Wartość oczekiwana oraz odchylenie standardowe jest wyznaczone dla rozkładu  $F$ , czyli rozkładu wszystkich zmian cen (nie tylko tych, które są wynikiem publikacji wiadomości)

Tabela ta przedstawia ponadto miary statystyczne rozkładu zmian cen akcji w całym analizowanym w tej pracy przedziale czasowym, niezależnie od faktu pojawienia się wiadomości. Porównując powyższe zmiany cen akcji ze zmianami wywołanymi przez wiadomości, można wnioskować, iż pojawienie się wiadomości zmniejsza wartość oczekiwaną oraz nieznacznie zwiększa standardowe odchylenie, co wydaje się być zgodne z intuicją. Statystycznie rzecz ujmując, pojawienie się wiadomości dotyczącej spółek notowanych w WIG20 powoduje istotną zmianę rozkładu cen akcji (odp. statystyka p-value = 0,13%), natomiast pojawienie się wiadomości ekonomicznej wywołuje niewielką zmianę rozkładu cen (odp. statystyka p-value = 8,63%).

#### 4.6. Słownik fraz

Słownik fraz to zbiór krotek wyrazów, dzięki którym możemy identyfikować zdarzenia występujące w wiadomościach. Przykładowa krotka wyrazów, będąca frazą, może wyglądać następująco: „Chiny”, „spadek”, „giełda”. Opisywany model rozpoznaje tylko te wydarzenia, które są zapisane za pomocą fraz, mówiąc inaczej, słownik fraz zawiera skończoną liczbę wszelkich możliwych wydarzeń, które zbudowany w tej pracy model „rozumie”. Występowanie frazy w artykule można zdefiniować za pomocą następującego predykatu:

$$\text{występuje}(faza, \text{artykuł}) \Leftrightarrow \forall_{\text{wyraz} \in faza} \exists \text{słowo} \text{ będące odmianą } fazy \wedge \text{słowo} \in \text{artykuł}$$

Z powyższej definicji można odczytać, iż nie jest ważna kolejność słów występujących w artykule prasowym ani ich forma gramatyczna. Należy również wziąć pod uwagę, iż w jednym artykule może wystąpić kilka fraz.

W pracach zespołów Wütricha oraz Peramunetilleke’a słowniki fraz zostały stworzone ręcznie przez ekspertów z danej dziedziny i zawierały odpowiednio 423 oraz 400 fraz. Niestety nie zostały one podane do publicznej wiadomości poza kilkoma przykładami umieszczonymi w pracy zespołu Wütricha, np. „bond jump” oraz „dollar weak against mark”<sup>1</sup>. Zadanie ręcznego skonstruowania słownika fraz dla wiadomości giełdowych podawanych w języku polskim przekracza możliwości autora tej pracy, a ponadto obecnie istnieje trend do używania automatycznych słowników, które mogą wydobyć więcej informacji w sposób bardziej obiektywny. W pracy autorstwa G. Gidófalviego i C. Elkana<sup>2</sup>

<sup>1</sup> B. Wüthrich, S. Leung, D. Peramunetilleke [et al.]: *op.cit.* s. 2

<sup>2</sup> G. Gidófalvi, C. Elkan: *op.cit.*

zaproponowano metodę automatycznego konstruowania słownika fraz za pomocą miary informacji wzajemnej (mutual information). Informacja wzajemna formalnie jest definiowana jako<sup>1</sup>:

$$I_{fraza}(X; Y_{fraza}) = \sum_{y \in Y_{fraza}} \sum_{x \in X} p_{fraza}(x, y) \log \frac{p_{fraza}(x, y)}{p(x) p_{fraza}(y)}$$

gdzie zmienna  $X$  reprezentuje rozkład zmian cen akcji po przyporządkowaniu ich do jednej z trzech kategorii. Punkty graniczne tych kategorii zostały tak dobrane, aby osiągnąć rozkład równomierny, co można matematycznie zapisać jako:

$$P(X = x) = 1/3, \text{ gdzie } X = \{1, 2, 3\}$$

Kategorie te posiadają odpowiednio trzy etykiety: „cena aktywa wzrośnie”, „cena aktywa nie zmieni się”, „cena aktywa spadnie”.

Zmienna  $Y_{fraza}$  zawiera rozkład charakterystyczny dla każdej frazy,  $Y_{fraza} = 1$ , gdy fraza wystąpiła w danej wiadomości, oraz  $Y_{fraza} = 0$  w przeciwnym przypadku.

$p_{fraza}(x, y = 1)$ , gdzie  $y \in Y_{fraza}, x \in X$  oznacza prawdopodobieństwo wystąpienia artykułu zawierającego frazę *fraza*, którego publikacja powoduje zmianę ceny należącą do kategorii  $x$ . Analogicznie  $p_{fraza}(x, y = 0)$  przedstawia prawdopodobieństwo wystąpienia artykułu niezawierającego frazy *fraza*, którego publikacja powoduje zmianę ceny należącą do kategorii  $x$ .

Intuicja podpowiada, że informacja wzajemna mierzy, ile informacji o rozkładzie cen  $x$  można poznać, mając informacje o wystąpieniu frazy  $Y_{fraza}$ , czyli o ile poznanie jednej z tych zmiennych zmniejsza niepewność drugiej. Automatyczna budowa słownika fraz polega na znalezieniu takich fraz, które maksymalizują wartość miary informacji wzajemnej. Przez frazę rozumie się 1 do 3 wyrazów. Poniższa tabela przedstawia tak utworzoną listę fraz dla grupy WIG20:

---

<sup>1</sup> A. Przepiórkowski: *op.cit.*

Tabela 6. Lista fraz o największej wartości miary informacji wzajemnej – grupa WIG20

<b>Frazy</b>	<b>Spadek</b>	<b>Bez zmian</b>	<b>Wzrost</b>
, . kwartał	18 (81,8%)	0	4 (18,2%)
. i rok	12 (100%)	0	0
, z złoty	0	14 (87,5%)	2 (12,5%)
. pomóc w	0	10 (100%)	0
. lotos z	0	7 (31,8%)	15 (68,2%)

*Źródło:* Opracowanie własne

Nietrudno zauważyć, iż frazy z tej tabeli nie identyfikują żadnych istotnych wydarzeń. Najprawdopodobniej wyniki te są efektem losowych złożeń znaków przestankowych oraz słów semantycznie pustych (np. zaimki), które miały nierównomierny rozkład.

W literaturze problem ten jest dość często opisany i zwykle sugerowane jest użycie stoplisty, czyli dokonanie eliminacji słów semantycznie pustych. W pracy tej powyższe rozwiązanie uzupełnione jest o wymóg, iż dana fraza musi wystąpić przynajmniej dziesięć razy, aby została uwzględniona jako potencjalna fraza słownikowa. Zabieg ten umożliwia odrzucenie fraz, które wystąpiły zbyt rzadko, aby można było uznać, iż mają jakąkolwiek moc predykcyjną. Ogólnie preferowane jest użycie fraz występujących stosunkowo często, ponieważ można mieć nadzieję że w większej próbie wyniki będą względnie stabilnie. Walidację wyników będą przeprowadzał na zbiorze testowym, dlatego mogę dokonywać modyfikacji parametrów badania w oparciu o zbiór treningowy. Poniższa tabela zawiera listę fraz o największej wartości informacyjnej po użyciu stoplisty oraz usunięciu fraz, które wystąpiły zbyt rzadko.

Tabela 7. Lista fraz o największej wartości miary informacji wzajemnej po odfiltrowaniu informacji nieistotnych – grupa WIG20

<b>Frazy</b>	<b>Spadek</b>	<b>Bez zmian</b>	<b>Wzrost</b>
BPH fuzja	8 (33,3%)	15 (62,5%)	1 (4,2%)
BPH	46 (31,3%)	69 (46,9%)	32 (21,8%)
BPH akcja	3 (13,6%)	16 (72,7%)	3 (13,6%)
BPH Pekao fuzja	7 (33,3%)	13 (61,9%)	1 (4,8%)
fuzja Lotos	0	4 (33,3%)	8 (66,7%)
DM	9 (23,1%)	23 (59,0%)	7 (17,9%)
obligacja zł	7 (53,8%)	6 (46,2%)	0
PKN prezes	1 (6,3%)	4 (25,0%)	11 (68,8%)
Kulczyk	5 (45,5%)	0	6 (54,5%)
obligacja wartość	5 (45,5%)	6 (54,5%)	

Źródło: Opracowanie własne

Wyniki z powyższej tabeli wskazują, iż odfiltrowanie danych nieistotnych znacznie poprawia jakość słownika fraz, uzyskane dane mają już sens i ekonomiczną interpretację, która dodatkowo jest zgodna z intuicją. Pierwszy przykład, fuzja BPH, jest odczytywana jako neutralna, nie wzbudzająca emocji i znacznych ruchów akcji, w przeciwieństwie do nazwiska Kulczyk, na które reakcje rynku są bardzo nerwowe. Informacje o fuzji lotosu są ogólnie pozytywnie odbierane przez rynek. Uważne przeanalizowanie wyników z tabeli sugeruje jednak, iż frazy te opisują wydarzenia bardzo specyficzne, możliwe, iż wystąpiło przetrenowanie danych. Liczba wygenerowanych fraz po usunięciu informacji nieistotnych wyniosła jedynie 563.

Poniższa tabela przedstawia wyniki podobnej analizy listy fraz dla danych ekonomicznych:

Tabela 8. Lista fraz o największej wartości miary informacji wzajemnej – grupa ekonomiczna

<b>Frazy</b>	<b>Pozycja</b>	<b>Spadek</b>	<b>Bez zmian</b>	<b>Wzrost</b>
dolar zł	1	110 (24,9%)	224 (50,7%)	108 (24,4%)
dolar euro zł	2	110 (24,9%)	224 (50,7%)	108 (24,4%)
fixing zł	3	110 (24,9%)	224 (50,7%)	108 (24,4%)
... 7 podobnych słów	4 – 10	-,-	-,-	-,-
procent zamknięcie	11	128 (36,6%)	58 (16,6%)	164 (46,9%)
notowanie procent zamknięcie	12	128 (36,7%)	58 (16,6%)	163 (46,7%)
... 6 podobnych słów	13 – 18	-,-	-,-	-,-
WIG20 spaść	19	1 (6,3%)	4 (25,0%)	11 (68,8%)
WIG20 ciągły zamknięcie	20	5 (45,5%)	0	6 (54,5%)
WIG20 zamknięcie	21	5 (45,5%)	6 (54,5%)	0

Źródło: Opracowanie własne

Powyższa lista zawiera jedynie najistotniejsze frazy, podczas gdy cały słownik fraz zawiera ich 1000 (najmniejsza wartość miary wzajemnej informacji wyniosła 0,000004).

Pierwsze 10 fraz z analizowanej listy (w powyższej tabeli zostały przedstawione jedynie pierwsze trzy) zawiera kombinację następujących wyrazów: „dolar”, „zł”, „euro”, „fixing” oraz „NBP”. Najprawdopodobniej są to komunikaty o fixingach NBP publikowane przez serwis Reuters. Oznacza to, iż nasze założenie o nieingerowaniu w źródło (poczynione w rozdziale 3.2) może nie być słuszne. Frazy te są nieistotne z punktu widzenia tego badania – nie niosą żadnej nowej informacji, stwarzają natomiast złudny efekt dla algorytmu klasyfikacyjnego, iż działają stabilizująco na rynek, co nie jest prawdą. Kolejna grupa 8 fraz (pozycje 11 - 18) zawiera wyrazy: „notowanie” „zamknięcie” oraz „procent” wywołujące według powyższej tabeli znaczną zmienność cen akcji. Są to jednak komunikaty publikowane na zakończenie dnia, dla których wynik finansowy jest liczony jako różnica cen pomiędzy sesjami odbywającymi się w różnych dniach<sup>1</sup>. Duża zmienność cen wynika zatem nie z informacji zawartych w wiadomościach. Przykład ten pokazuje, iż poczynione w tej pracy założenie budowy jednego modelu i użycie jednego klasyfikatora dla wiadomości publikowanych w czasie sesji i poza nią może być błędne ze względu na różną wariację w obu grupach. Poczynione w tym podrozdziale uwagi są cennym punktem wyjścia do dalszych badań.

<sup>1</sup> Wynika to ze sposobu obliczenia efektów finansowych dla wiadomości, które zostały opublikowane później niż po godzinie przed zamknięciem sesji. Szczegóły opisano w rozdziale 4.4

#### 4.6.1. Algorytm korekcji danych

W podrozdziale tym proponuje się sposób automatycznego usuwania wiadomości nieniosących nowej informacji. Jest to próba oczyszczenia danych, będąca rezultatem wniosków oraz rozważań z poprzedniego podrozdziału. Jest wielce prawdopodobne, iż komunikaty o fixingu, artykuły o tytule przegląd prasy, kalendarium rynku kapitałowego czy najważniejsze wydarzenia na świecie nie zawierają nowej (a często żadnej) informacji. Należy pamiętać, iż w tym badaniu zostały przetworzone jedynie nagłówki wiadomości, tak więc nie jest to informacja zawierająca przegląd prasy, a jedynie nagłówek informujący, iż o tej godzinie został opublikowany przegląd prasy. Jak podkreśliłem już wcześniej, chciałbym ograniczyć liczbę modyfikacji źródła danych do minimum i zrobić to w sposób naukowy, a nie ograniczać się do przeglądania danych i zastanawiać nad każdym artykułem z osobna, czy go wyrzucić.

Rozwiązaniem jest identyfikacja nagłówków wiadomości, które pojawiają się dokładnie o tej samej godzinie każdego dnia. Można je znaleźć poprzez wyszukiwanie fraz o minimalnej wariancji godziny publikacji. Ze względów technicznych należy ograniczyć wyszukiwanie do fraz, które wystąpiły wielokrotnie (np. 10 razy lub więcej). Wynikiem działania algorytmu jest następująca lista:

Tablica 7. Lista nagłówków nieniosących nowych informacji

```
Fixing NBP :
(GPW) pierwszy FIXING :
(GPW) drugi FIXING :
(GPW) oficjalne wyniki sesji (GPW) : WIG20
Przegląd prasy
Kalendarium rynku kapitałowego
Kalendarium najważniejszych wydarzeń
(GPW) notowania ciągłe : na
* * * najważniejsze wiadomości
Najważniejsze wydarzenia ekonomiczne na świecie
```

Źródło: Opracowanie własne

W wyniku usunięcia powyższych nagłówków z wiadomości ekonomicznych 2739, czyli 26,078%, fraz zostało odrzuconych, co według mnie znacznie podniosło jakość danych źródłowych. Wariancja godziny publikacji artykułów z powyższego schematu była niższa od 6 sekund. Dla wiadomości z grupy WIG20 algorytm ten nie znalazł żadnej frazy do usunięcia. Frazą o najniższej wariancji czasu publikacji, równej 152,65 sekund, była fraza „PKN”, co nie wydaje się być nieistotne.

Po zastosowaniu zaproponowanej w tym podrozdziale korekcji danych w poniższej tabeli zostały zawarte frazy o największej wartości informacyjnej:

Tabela 9. Lista fraz o największej wartości miary informacji wzajemnej po usunięciu wiadomości nieniosących nowych informacji – grupa ekonomiczna

<b>Frazy</b>	<b>Spadek</b>	<b>Bez zmian</b>	<b>Wzrost</b>
miedź	60 (24%)	120 (48%)	70 (28%)
tanieć	10 (16,7%)	38 (63,3%)	12 (20,0%)
ropa	88 (26,0%)	154 (45,4%)	97 (28,6%)
miedź zapas	0	16 (69,6%)	7 (30,4%)
E12 m produkcja	0	0	10 (100%)
CeTO	53 (36,7%)	27 (17,9%)	71 (47,0%)
notowanie procent	51 (35,7%)	25 (17,5%)	67 (46,9%)
procent zamknięcie	51 (35,9%)	25 (17,6%)	66 (46,5%)

Źródło: Opracowanie własne

Powyższe wyniki są optymistyczne. Prawdą jest, że informacje o miedzi, ropie czy procesie „tanienia” mają znaczenie dla gospodarki.

Podsumowując podrozdział, chciałbym zaznaczyć, iż są to jedynie frazy, dzięki którym można identyfikować zdarzenia występujące w wiadomościach. Problem przetrenowania danych bądź nieintuicyjne wartości rozkładu słów na tym etapie nie są istotne. Ważna natomiast jest odpowiedź na pytanie, czy jest to ten typ fraz, dzięki którym, znając liczbę ich wystąpień w dokumentach, możemy skuteczniej prognozować ceny akcji. Moim zdaniem odpowiedź jest pozytywna zarówno dla spółek WIG20, jak i wiadomości ekonomicznych.

#### **4.7. Budowa macierzy wystąpień**

Mając wyniki obliczeń przeprowadzonych w poprzednich podrozdziałach, budowa macierzy wystąpień, czyli przekształcenie artykułów na postać ilościową, jest rzeczą stosunkowo prostą. Co więcej, we wszystkich omówionych przeze mnie pracach badawczych krok ten jest wykonywany podobnie. Zbiór artykułów wraz z przyporządkowanymi wynikami finansowymi jest przekształcany na macierz wystąpień. Używając terminologii data miningu, wiersz jest pojedynczą obserwacją, odpowiadającą jednemu dokumentowi tekstowemu – nagłówkowi artykułu prasowego. Kolumny reprezentują odpowiednie frazy (ze słownika fraz), identyfikujące zdarzenia, które system potrafi rozpoznać. Komórkę macierzy wystąpień można zdefiniować w sposób następujący:

$$\text{MacierzWystapien}_{i,j} = f(\text{liczba wystapien frazy } j \text{ w dokumencie } i)$$

Ostatnią kolumną macierzy wystąpień jest kolumna wyników, przedstawiająca zmiany cen po publikacji wiadomości (sposób wyliczenia tej wartości został opisany w podrozdziale 4.4). Jediną decyzją, którą należy podjąć na tym etapie badań, jest wybór odpowiedniej funkcji przekształcającej  $f$  (możliwe postaci funkcji  $f$  zostały opisane w podrozdziale 1.2.2). W moim badaniu zdecydowałem się użyć wartościowania binarnego, przypisującego 1, gdy fraza wystąpi przynajmniej raz, oraz zero, kiedy nie wystąpi w ogóle. W większości przypadków fraza będzie występowała w dokumencie tylko raz, więc użycie wartościowania binarnego jest zupełnie wystarczające.

Po przekształceniu wiadomości tekstowe są już „zwykłą” macierzą, którą można potraktować jako wejście do większości algorytmów uczenia maszynowego. Poniższa tabela przedstawia część macierzy wystąpień dla grupy WIG20

Tabela 10. Przykład konwersji wiadomości prasowych na postać ilościową – WIG20

Nagłówek wiadomości	zdecydować	NWZA	Pekao akcja	akcja zł	dywidenda	dywidenda wypłata	BZ WBK
Żywiec, BZ WBK i Pekao zdecydowały o wypłacie dywidendy	$f(1)$	$f(0)$	$f(0)$	$f(0)$	$f(1)$	$f(1)$	$f(1)$
NWZA TP SA zdecydowało o niewypłaceniu dywidendy	$f(1)$	$f(1)$	$f(0)$	$f(0)$	$f(1)$	$f(1)$	$f(0)$
Pekao zdecydowało o wypłacie 9 zł dywidendy na akcję	$f(1)$	$f(0)$	$f(1)$	$f(1)$	$f(1)$	$f(1)$	$f(0)$

Źródło: Opracowanie własne

#### 4.8. Zastosowanie uczenia maszynowego

Ostatnim elementem badania jest wybór odpowiedniego algorytmu klasyfikującego. W omawianych w rozdziale 2. publikacjach najczęściej używanymi metodami estymacji były naiwny estymator Bayesa oraz algorytmy tworzące reguły. W mojej pracy wybrałem naiwny estymator Bayesa, ponieważ jest on często wykorzystywany jako metoda bazowa przy porównaniach bardziej zaawansowanych estymatorów.

W pracy zespołu kierowanego przez V. Lavrenkę<sup>1</sup> zaproponowany został formalny model estymacji modelu  $M_t$  za pomocą dokumentów  $D_1, \dots, D_m$ . Każdy dokument  $D_i$  jest reprezentowany za pomocą zbioru fraz  $S_{i,1}, \dots, S_{i,n}$ . Celem użycia tego modelu jest wybór właściwego  $t$ , gdzie  $t$  odpowiada zachodzącym trendom (wzrost ceny, cena bez zmian, spadek ceny), wiedząc, że obserwowany jest dokument  $D_i$ . Formalnie wybór trendu można zapisać jako:

$$M_{best} = \arg \max_{t \in trends} P(M_t | \{S_1 \dots S_m\}) = \\ \arg \max_{t \in trends} \frac{P(\{S_1 \dots S_m\} | M_t) P(M_t)}{P(\{S_1 \dots S_m\})}$$

Ponieważ granice trendów zostały tak dobrane, aby uzyskać równomierny rozkład we wszystkich trzech grupach,  $P(M_t) = 1/3$  dla każdego  $t \in trends$ , oraz założono, iż rozkład fraz  $S_{i,1}, \dots, S_{i,n}$  jest niezależny od siebie, można przekształcić powyższy wzór w następujący sposób:

$$M_{best} = \arg \max_{t \in trends} \prod_{i=1}^m \frac{P(S_i | M_t) P(M_t)}{P(S_i)} \\ = \arg \max_{t \in trends} \prod_{i=1}^m \frac{P(S_i | M_t)}{P(S_i)}$$

Prawdopodobieństwa rozkładów fraz oraz prawdopodobieństwa warunkowe zostały oszacowane za pomocą ilorazu największej wiarygodności, zastępując prawdopodobieństwa liczbą wystąpień. Oczywiście założenie, iż poszczególne wyrazy w artykule są niezależne od siebie, jest fałszywe, zespół Lavrenki sugeruje jednak, iż jest to standardowa technika stosowana w podobnych pracach, ułatwiająca obliczenia, a uwzględnienie zależności między wyrazami zazwyczaj nie poprawia wyników.

Obserwując wyniki klasyfikacji modelu przedstawionego powyżej, postanowiłem wprowadzić przedstawione poniżej usprawnienia.

---

<sup>1</sup> V. Lavrenko, M. Schmill, D. Lawrie [et al.]: *op.cit.*

### Metoda elementu najbardziej dyskryminującego

Pomysł polega na znalezieniu frazy, która ma największą wartość MI (szczegóły dotyczące miary wzajemnej informacji zawarte są w rozdziale 4.6), i prognozowaniu tylko za pomocą tej wielkości.

Metodę tę możemy zapisać jako:

$$S_{best} = \arg \max_{S \in \text{frazy w } D} MI(S)$$
$$M_{best} = \arg \max_{t \in \text{trends}} P(S_{best} | M_t)$$

### Metoda średnich MI

Koncepcja ta polega na połączeniu przedstawionej powyżej metody z metodą stosowaną przez zespół Lavrenki. Prawdopodobieństwo wystąpienia trendu oblicza się jako iloraz prawdopodobieństw poszczególnych fraz ważony miarą wzajemnej informacji. Poniższy wzór przedstawia tę ideę:

$$M_{best} = \arg \max_{t \in \text{trends}} \prod_{i=1}^m \frac{P(S_i | M_t)}{P(S_i)} * \frac{MI(S_i)}{\sum_{i=1}^m MI(S_i)}$$

### Metoda poziomu ufności

Metoda ta wykorzystuje dodatkową informację zawartą w modelu  $M_t$ . Nietrudno zauważyć, iż model  $M_t$  w pewnym sensie produkuje wektor trzech prawdopodobieństw, które po znormalizowaniu opisują prawdopodobieństwo, iż wiadomość zostanie sklasyfikowana odpowiednio do kategorii: „spadnie”, „nie zmieni się”, „wzrośnie”. A zatem nic nie stoi na przeszkodzie, aby wykorzystać tę dodatkową wiedzę i obliczyć  $M_{best}$  identycznie jak poprzednio, a dodatkowo policzyć poziom ufności tego wyniku, czyli ustalić, jak bardzo klasyfikator jest pewny, że ma rację. Poziom ufności jest liczony za pomocą entropii rozkładu prawdopodobieństwa z wektora  $M$ , co formalnie możemy zapisać w następujący sposób:

$$M_t = \prod_{i=1}^m \frac{P(S_i | M_t)}{P(S_i)}, \text{ dla } t = \{\text{Spadnie, Bez zmian, Wzrośnie}\}$$

$$M'_t = a * M_t, \text{ gdzie } a > 0 \text{ oraz } \sum_{t=1}^3 M'_t = 1 \text{ (Normalizacja)}$$

$$poziomUfności = 1 - \frac{Entropy(M'_t)}{\max_h (Entropy(M_h))} = 1 - \frac{Entropy(M'_t)}{\log_2 3}$$

Miara *poziomUfności* osiąga wartość zero dla rozkładu jednostajnego  $M$ , w którym każde  $P(M_t) = 1/3$ , oraz wartość jeden w przypadku rozkładu  $M$ , dla którego  $\exists t : P(M_t) = 1$ . Miara ta jest zgodna z intuicyjnym pojęciem „pewności”.

## 4.9. Wyniki badania

Wspomniany wcześniej problem braku ustalonych metodologii przenosi się na płaszczyznę ewaluacji porównywanych modeli – brak jasno ustalonych oraz akceptowanych metryk. W rozdziale tym przedstawię dwa typy analizy wyników: klasyczne miary data miningu (liczba poprawnych odpowiedzi) oraz symulację rynkową. Obie miary pochodzą z pracy „Language Models for Financial News Recommendation”<sup>1</sup>.

### 4.9.1. Klasyczne miary data miningu

Klasyczny test zadania klasyfikacji z nadzorem rozpoczyna się od podziału zbioru na część treningową oraz walidacyjną. Następnie model jest trenowany oraz optymalizowany na części treningowej, a później sprawdzany na części walidacyjnej. Metodę tę zastosowałem w niniejszej pracy, dokonując wszelkich wyborów formy oraz parametrów modelu na podstawie danych treningowych. Problem „przerzucania danych” był często podnoszony przez krytyków prac, które przedstawiłem w rozdziale 3.5. Ponadto w pracy tej zaprezentowałem wszystkie uzyskane rezultaty, a nie tylko te najlepsze.

Poniższe tabele przedstawiają liczbę poprawnych odpowiedzi dla danych treningowych. Model został więc wytrenowany i sprawdzony na tych samych danych, co oznacza, iż analizując wyniki można poznać jedynie górne ograniczenie możliwości predykcji tego modelu. Pozwala to również sprawdzić, czy model został poprawnie skonstruowany, oraz wybrać najlepszą formę modelu bez użycia danych testowych. Wyniki poniższych tabel, dla czytelności, zostały zapisane jako różnica między wynikami algorytmu a gracza losowego, który wybiera każdą możliwość z równym prawdopodobieństwem 1/3.

---

<sup>1</sup> V. Lavrenko, M. Schmill, D. Lawrie [et al.]: *op.cit.*

Tablica 8. Liczba poprawnych odpowiedzi dla grupy WIG20 (dane treningowe)

	<b>Spadek</b>	<b>Bez zmian</b>	<b>Wzrost</b>
Spadek	<b>6,09%</b>	-1,81%	-4,37%
Bez zmian	-3,27%	<b>8,23%</b>	-4,84%
Wzrost	3,01%	-2,12	<b>5,09%</b>
Zysk z użycia modelu: 19,41%			

Źródło: Opracowanie własne

Tablica 9. Liczba poprawnych odpowiedzi dla grupy ekonomicznej (dane treningowe)

	<b>Spadek</b>	<b>Bez zmian</b>	<b>Wzrost</b>
Spadek	<b>6,91%</b>	-2,56%	-4,27%
Bez zmian	-1,18%	<b>5,96%</b>	-4,93%
Wzrost	-0,85%	-2,79	<b>3,81%</b>
Zysk z użycia modelu: 16,68%			

Źródło: Opracowanie własne

Kolumny w tabelach oznaczają wyniki prognozy modelu, natomiast wiersze aktualne wartości. Pogrubione wartości na diagonalnej macierzy wskazują liczbę poprawnych odpowiedzi, w których model okazał się lepszy od gracza losowego. Całkowitą liczbę poprawnych odpowiedzi można otrzymać, dodając do zysku z użycia modelu 33,3% (tylko poprawnych odpowiedzi uzyska przeciętnie gracz losowy). Dla wiadomości WIG20 liczba ta wynosi 52,71% (19,41% + 33,3%), natomiast dla wiadomości ekonomicznych 49,98% (16,68% + 33,3%). Wartość z tabeli WIG20 w lewym dolnym rogu równa 3,01% oznacza, iż model relatywnie zbyt często prognozuje wzrost, podczas gdy powinien prognozować spadek.

Podobną analizę można by przeprowadzić dla wszystkich możliwych 16 kombinacji: wiadomości ekonomicznych/WIG20, danych treningowych/testowych oraz czterech zaproponowanych metod estymacji. Przedstawienie wszystkich tych kombinacji w postaci osobnych tabel zajęłoby jednak zbyt dużo miejsca i nie wniosło istotnej wartości, dlatego zdecydowałem się przedstawić skuteczność każdej kombinacji (mierzonej za pomocą dodatkowej liczby poprawnych odpowiedzi w porównaniu do gracza losowego) w jednej zbiorczej tabeli:

Tabela 11. Porównanie skuteczności predykcji modeli w porównaniu do gracza losowego

	<b>Metoda bazowa</b>	<b>Metoda elementu najbardziej dyskryminującego</b>	<b>Metoda średnich MI</b>	<b>Metoda poziomu ufności</b>
WIG20 treningowe	19,41%	10,68%	16,58%	28,38%
ekonomiczne treningowe	16,68%	12,09%	16,23%	24,91%
WIG20 testowe	2,71%	2,29%	2,13%	4,23%
ekonomiczne testowe	3,22%	2,99%	3,21%	5,26%

Źródło: Opracowanie własne

Powyższe wyniki są według mnie zadowalające. Wszystkie uzyskane wyniki są nie gorsze od tych, które osiągnięto by stosując metodę naiwną – gracza losowego. Ponadto wprowadzone przeze mnie ulepszenia pozwalają wyciągnąć spójne i logiczne wnioski: metoda elementu najbardziej dyskryminującego oraz metoda średnich MI dają gorsze rezultaty niż metoda bazowa, natomiast metoda poziomu ufności lepsze. Dowodzi to, że metoda bazowa, przedstawiona w innych publikacjach, jest metodą najlepszą, a modyfikacja jej działania poprzez odpowiednie ważenie fraz jedynie pogarsza rezultaty.

Metoda poziomu ufności nie jest w gruncie rzeczy modyfikacją, a jedynie rozszerzeniem metody bazowej. Klasyfikator metody poziomu ufności produkuje dla każdej wiadomości parę liczb: prognozę i pewność, wyrażoną w skali od zera do jedynki, z jaką prognoza jest prawdziwa,. Liczba poprawnych odpowiedzi jest zatem sumą odpowiedzi z metody bazowej ważoną przez pewność tych odpowiedzi. Za pomocą tego relatywnie prostego rozwiązania podwojona została skuteczność predykcji. Zabieg ten wykracza w pewnym sensie poza dziedzinę klasycznego data miningu, ponieważ ostatecznym celem prognozy jest identyfikacja tych wiadomości, które wpływają na zmianę cen akcji, a nie próba prognozy każdej wiadomości, która się pojawi

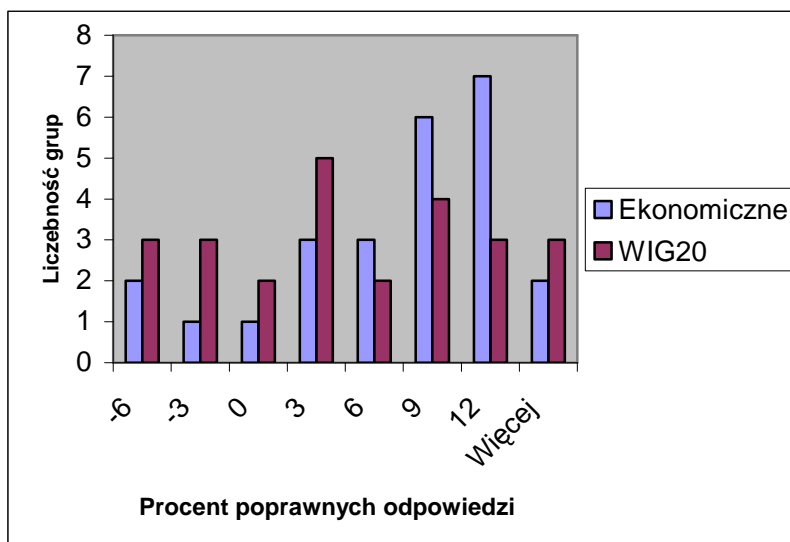
Wyniki powyższej tabeli sugerują, iż algorytm jest trochę bardziej przetrenowany dla danych WIG20 – trening oraz walidacja na tym samym zbiorze danych daje lepsze rezultaty niż dla wiadomości ekonomicznych, natomiast walidacja na zbiorze testowym gorsze. Możliwym wytłumaczeniem tego zjawiska jest fakt, iż dane dla WIG20 zawierają zbyt małą liczbę obserwacji, zwłaszcza dotyczących poszczególnych spółek, i odpowiednie frazy, reprezentujące konkretne wydarzenia, występują zbyt rzadko.

Podczas badań ujawniła się również bardzo duża zmienność wyników w zależności od podziału zbioru danych na część treningową oraz testową. Oznacza to, iż powyższe wyniki mogą nie być miarodajne i należy użyć walidacji krzyżowej.

#### 4.9.2. Walidacja krzyżowa (kroswalidacja)

W podrozdziale tym sprawdzono wyniki działania systemu poprzez walidację krzyżową. Po części jest to efektem wniosków z poprzedniego podrozdziału, w którym stwierdzono duże wahania wyników w zależności od podziału zbioru, chociaż podział ten był wykonywany zawsze w sposób losowy. Procedura kroswalidacji rozpoczyna się od podzielenia oryginalnej próby na 25 podzbiorów. Następnie kolejno każdy z nich jest traktowany jako zbiór testowy, a pozostałe jako zbiór uczący, i poddany analizie. Analiza jest więc wykonywana 25 razy. Jako metoda estymacji została wybrana metoda poziomu ufności, ponieważ daje najlepsze rezultaty. Wyniki analizy zostały przedstawione na poniższym schemacie oraz w tabeli 15.:

Rys. 2. Wyniki walidacji krzyżowej (metoda poziomu ufności)



Źródło: Opracowanie własne

Tabela 12. Wyniki walidacji krzyżowej (metoda poziomu ufności) – miary statystyczne

	<b>WIG20</b>	<b>Ekonomiczne</b>
minimum	-12,4335	-6,37869
1 kwantyl	-1,24531	2,797327
średnia	3,706506	5,806373
3 kwantyl	8,474029	9,385199
maksimum	23,65887	14,75861
wariancja	64,68502	34,58201

Źródło: Opracowanie własne

Jeden z pierwszych wniosków, wynikających zarówno z analizy schematu, jak i z tabeli to znaczna niestabilność wyników. Oznacza to, iż algorytm jest bardzo wrażliwy na wybór odpowiednich danych. Z tabeli można odczytać, iż wariancja wyniku wyniosła 64,68 dla grupy WIG20 a 34,58 dla grupy wiadomości ekonomicznych. Ponadto wiadomości ekonomiczne osiągają lepsze rezultaty – w niektórych grupach procent poprawnych odpowiedzi jest o 9 – 14% wyższy od rezultatów losowego gracza.

#### 4.9.3. Symulacja rynkowa

W rozdziale zostanie przeprowadzona klasyczna symulacja rynkowa, w której algorytm będzie generował wirtualne sygnały kupna oraz sprzedaży. W trakcie tej symulacji sprawdzona zostanie strategia, w której to typowy gracz giełdowy będzie inwestował zgodnie z wirtualnymi sygnałami dawanymi przez system. W momencie nadejścia nowej wiadomości system zaklasyfikuje ją do jednej z trzech kategorii. Jeśli wiadomość otrzyma etykietę wzrost, system kupi wskazywaną przez wiadomość akcję i sprzedaje ją po godzinie. Jeśli wiadomość otrzyma etykietę spadek, system krótko sprzeda daną akcję. Transakcja krótkiej sprzedaży to odwrócenie typowej kolejności transakcji giełdowej, operacja ta oznacza, iż inwestor sprzedaje na rynku pożyczony papier, a po godzinie go odkupuje. Zyskiem (bądź ewentualną stratą) będzie dla inwestora różnica pomiędzy ceną sprzedaży a ceną kupna.

Strategia ta jest bardzo prosta i prawdziwy inwestor może obserwować znaczną liczbę innych wskaźników, dokonywać analizy technicznej oraz fundamentalnej bądź śledzić nastroje na rynkach. Celem użycia tak prostej strategii jest próba wyeliminowania innych wskaźników w celu dokonania obiektywnej oceny zaproponowanego w tej pracy prototypu.

W wielu pracach padały propozycje, aby w przypadku gdy spadek/wzrost będzie większy niż 1%, zrealizować transakcję natychmiast, w przeciwnym razie dopiero pod koniec godziny.

Powyższa strategia przeprowadzona na wiadomościach ekonomicznych daje następujące wyniki:

Tabela 13. Wyniki symulacji rynkowej dla wiadomości ekonomicznych

Strategia	Poprawne odpowiedzi	Zysk	Średni poziom ufności	Liczba decyzji
metoda poziomu ufności	5,8%	440,5%	25,6%	7349
metoda bazowa	3,4%	178%	100%	7349
kupuj <sup>1</sup>	0,2%	4,1%	100%	11462
sprzedawaj <sup>1</sup>	-0,3%	-4,1%	100%	11462
gracz losowy (max)	1,2%	175,7%	49,6%	7786
gracz losowy (mediana)	-0,2%	0,5%	50,1%	7572
gracz losowy (min)	-0,8%	-195,8%	50,5%	7653

Źródło: Opracowanie własne

Strategia opisana jako gracz losowy polega na stworzeniu grupy tysiąca graczy, którzy podejmują losowe decyzje  $x$  z losowym prawdopodobieństwem pewności siebie  $p$ , co można zapisać:

dla  $x \in \{\text{spadek, bez zmian, wzrost}\}$ ,  $P(x) = 1/3$ ,

oraz  $p \in$  rozkład jednostajny na przedziale  $\langle 0,1 \rangle$ .

W powyższej tabeli wiersze o nazwie gracz losowy pokazują najlepszy wynik, medianę oraz najgorszy wynik spośród tysiąca losowych graczy. Znając te wyniki, można wnioskować o istotności metody poziomu ufności. Metoda ta ponownie okazała się najlepszą, a ponadto jest ona istotna statystycznie na poziomie  $p\text{-value} = 0,1\%$ , ponieważ zysk z niej jest większy niż zysk każdego z tysiąca losowych graczy.

W miejscu tym uważam, że wydaje się zasadne przedstawić w szczegółach sposób wyliczania zysku. Sposób ten uwzględnia poziomy ufności generowane przez algorytm a także średni poziom ufności wszystkich odpowiedzi wygenerowany przez algorytm, co przedstawia poniższy wzór:

<sup>1</sup> Według strategii tej algorytm zawsze dokonuje zakupu (strategia kupuj) lub krótkiej sprzedaży (strategia sprzedawaj) w momencie publikacji wiadomości

$$zysk = \bar{p} * \sum_{w \in \text{wiadomości}} z_w * p_w * \Delta \text{ceny}_w$$

gdzie:  $\bar{p}$  – oznacza średni poziom ufności wszystkich odpowiedzi,

$z_w, p_w$  – odpowiednio odpowiedź oraz poziom pewności zwrócony przez zastosowaną strategię dla wiadomości  $w$ . Odpowiedź  $z = 1$  gdy strategia generuje sygnał kupna,  $z = -1$ , gdy strategia generuje sygnał sprzedaży oraz  $z = 0$  w przeciwnym przypadku.

$\Delta \text{ceny}_w$  – godzinna zmiana ceny akcji po publikacji wiadomości.

Wzór ten uwzględnia średni poziom ufności, ponieważ, gdyby tego nie uwzględniał, *ceteris paribus* algorytmy dające odpowiedzi z większym średnim poziomem ufności uzyskiwałyby automatycznie wyższe zyski.

Poniższa tabela przedstawia analogiczne wyniki symulacji rynkowej dla wiadomości WIG20:

Tabela 14. Wyniki symulacji rynkowej dla wiadomości WIG20

Strategia	Poprawnych odpowiedzi	Zysk	Średni poziom pewności	Liczba decyzji
metoda poziomu ufności	3,7%	139,6%	32,9%	2378
metoda bazowa	1,6%	38,7%	100%	2377
kupuj <sup>1</sup>	0%	-144,1%	100%	3737
sprzedawaj <sup>1</sup>	0%	144,1%	100%	3737
gracz losowy (max)	-1,1%	221,7%	50,3%	2496
gracz losowy (21.)	1,2%	138,1%	49,1%	2501
gracz losowy (mediana)	-0,1%	-2,9%	50,1%	2536
gracz losowy (min)	-0,8%	-253,8%	49,3%	2478

Źródło: Opracowanie własne

Wyniki symulacji dla wiadomości WIG20 są podobne do analogicznej symulacji dla danych ekonomicznych. Zysk jest co prawda mniejszy, jednak wykonano mniej transakcji. Wiersz gracz losowy 21. oznacza gracza, który jest na 21. pozycji listy graczy losowych uporządkowanej według kryterium zysku. Gracz ten został wybrany, ponieważ jego wynik jest porównywalny do wyniku osiągniętego przy zastosowaniu metody poziomu ufności, a zatem jego pozycję na liście po podzieleniu przez 10 możemy interpretować jako percentyl i

<sup>1</sup> Według strategii tej algorytm zawsze dokonuje zakupu (strategia kupuj) lub krótkiej sprzedaży (strategia sprzedawaj) w momencie publikacji wiadomości

użyć do oszacowanie poziomu istotności. Oznacza to, iż metoda poziomu ufności jest istotna na poziomie 2,1%.

Należy pamiętać, iż alternatywą dla zaprezentowanego w tej pracy modelu prognozy cen akcji jest prosta strategia „kup i trzymaj” (buy&hold). Strategia ta, jak wynika z powyższej tabeli, przynosi 144,1% straty w okresach, kiedy jest publikowana wiadomość. Oznacza to, iż metoda poziomu ufności zastosowana w tej pracy osiąga zysk o 283,7% wyższy niż strata generowana przez strategię „kup i trzymaj”.

#### 4.10. Koszty transakcyjne

Bardzo istotną kwestią w tego typu badaniach jest kwestia kosztów transakcyjnych. W niektórych były one po prostu pomijane, co zresztą stało się przedmiotem krytyki<sup>1</sup>. W pracy „Language Models for Financial News Recommendation”<sup>2</sup> znajduje się informacja, iż relatywne koszty transakcyjne dążą do zera, gdy wartość transakcji jest wystarczająco duża, dlatego zespół Lavrenki w swoich badaniach kosztów tych nie uwzględnił.

Postanowiłem zweryfikować tę tezę osobiście i odwiedziłem strony internetowe znanych biur maklerskich. W większości z nich stosowana była zniżka w przypadku daytradingu. Daytrading, czasem określany jako szybka sprzedaż, polega na dokonywaniu w ramach jednej sesji transakcji kupna i sprzedaży tego samego papieru wartościowego. Zniżka ta ma istotne znaczenie w przypadku tej pracy, ponieważ sednem przedstawionego modelu jest dokonywanie dwóch odwrotnych transakcji w odstępie godzinnym. Tabela 15. przedstawia wyniki moich dociekań dotyczących biur maklerskich:

Tabela 15. Porównanie kosztów transakcyjnych z osiąganymi zyskami

	<b>Koszt 1 transakcji</b>	<b>Zysk z 1 transakcji</b>
BPH	0,2%	
bossa.pl	0,2%	
MBank	0,25%	
Wiadomości ekonomiczne		0,06%
Wiadomości WIG20		0,12%

Źródło: Opracowanie własne

<sup>1</sup> M.A. Mittermayer, G.F. Knolmayer: *op.cit.*

<sup>2</sup> V. Lavrenko, M. Schmill, D. Lawrie [et al.]: *op.cit.*

Powyższe wyniki pokazują wyraźnie, iż koszty transakcyjne przy zastosowaniu przedstawionych strategii są wyższe niż osiągnięty zysk. Po osobistym sprawdzeniu opłat pobieranych przez biura maklerskie nie mogę zgodzić się z tezą zespołu Lavrenki, iż relatywne koszty transakcyjne dążą do zera, gdy wartość transakcji jest wystarczająco duża. Według mnie, wydaje się bardziej prawdopodobne, iż koszty są proporcjonalne do wartości transakcji, i choć prawdą jest, że duże instytucje mogą wynegocjować pewne zmniejszenie kosztów, to wątpię, czy mogą je zmniejszyć aż o połowę.

Prace zajmujące się prognozą cen akcji za pomocą wiadomości prasowych były krytykowane przez M.A. Mittermayera i G.F. Knolmayera za operowanie na rynku zbyt dużą kwotą w porównaniu do własnych zasobów kapitałowych (nawet czterdziestokrotnie wyższą), podczas gdy „zaufane” instytucje nie pozwalają sobie na przekroczenie jej dziesięciokrotnie. W moim badaniu maksymalna wartość transakcji jest nie większa niż odwrócony średni poziom ufności, czyli odpowiednio czterokrotnie dla danych ekonomicznych oraz trzykrotnie dla danych WIG20.

## PODSUMOWANIE

Zanim przedstawię wyniki moich badań, chciałbym uzasadnić konstrukcję tej pracy. Główny punkt ciężkości położyłem na wykonanie pracy poprawnej metodologicznie, stanowiącej fundament do rozwoju dalszych badań na temat zagadnienia prognozy cen akcji za pomocą wiadomości prasowych. Przesunięcie punktu ciężkości w stronę pracy badawczej musi się odbyć kosztem pewnej bezpośredniej korzyści praktycznej. Oznacza to, że celem tej pracy nie jest osiągnięcie w sposób bezpośredni jak najlepszych wyników mierzonych zyskiem z transakcji.

Wyniki, które osiągnąłem, uważam za zadowalające, są bowiem spójne, a ponadto zgodne z ekonomiczną intuicją. W większości badań wiadomości dotyczące konkretnych spółek giełdowych dają lepsze rezultaty od ogólnych wiadomości ekonomicznych. Zysk z zastosowanych strategii okazał się istotny statystycznie i w porównaniu ze strategią „kup i trzymaj” (buy&hold), stosowaną w ciągu 7 lat, wynosi dla wiadomości WIG20 283%, a dla wiadomości ekonomicznych 436%. Wyniki te, choć są istotne statystycznie, nie są istotne ekonomicznie, ponieważ zysk z jednej transakcji dla danych WIG20 wynosi 0,12%, a dla danych ekonomicznych 0,06%. Przy obecnych kosztach transakcyjnych w wysokości 0,2%, oznacza to, iż koszt z przeprowadzonej transakcji przewyższy zysk.

## Weryfikacja hipotez

We wstępie przedstawiłem zasadność rozbicia głównego tematu pracy na trzy powiązane ze sobą hipotezy:

- *H1*: Wiadomości prasowe mają pewną wartość informacyjną;
- *H2*: Narzędzia text miningu potrafią tę informację wydobyć;
- *H3*: Dzięki wydobyciu tej informacji można osiągnąć ponadprzeciętne zyski.

Wyjaśniłem też, że hipoteza pierwsza wydaje się bardzo intuicyjna. Można argumentować, iż skoro istnieje ogromna liczba gazet, serwisów czy wyspecjalizowanych instytucji, za których usługi ludzie słono płacą, wiadomości prasowe muszą zawierać dodatkową wartość informacyjną. Ponadto w rozdziale 4.5, niejako na marginesie, zweryfikowałem hipotezę, czy sam fakt pojawienia się wiadomości znacząco wpływa na rozkład cen w najbliższej godzinie. Odpowiedź okazała się twierdząca dla danych WIG20 (odp. statystyka p-value = 0,13%) oraz bliska twierdzącej dla danych ekonomicznych (odp. statystyka p-value = 8,63%). Poza tym należy wziąć pod uwagę, iż prosta strategia: „zawsze

dokonuj krótkiej sprzedaży akcji spółki, gdy pojawi się wiadomość dotycząca tej spółki”, przynosi zysk 144,1% w analizowanym przez mnie okresie.

Druga hipoteza, stanowiąca istotę tej pracy, została również zweryfikowana twierdząco. Przy użyciu walidacji krzyżowej (kroswalidacji) pokazano, iż korzystając z wiadomości prasowych można dokonać trafniejszej predykcji w około 5% przypadków w porównaniu z graczem losowym. Zysk z zastosowanych strategii wynosi dla wiadomości WIG20 283% oraz 436% dla wiadomości ekonomicznych ponadto, co można osiągnąć przez zastosowanie strategii „kup i trzymaj” (buy&hold) w ciągu 7 lat. Wyniki te są istotne, ponieważ odp. statystyka p-value wynosi 0,1% dla wiadomości ekonomicznych oraz 2,1% dla wiadomości WIG20.

W pracy tej nie udało się udowodnić trzeciej hipotezy. Rozważania na ten temat przedstawiono w podrozdziale 4.10 i wynika z nich, że zysk z zastosowanego algorytmu wynosi odpowiednio 0,12% dla jednej transakcji przy wykorzystaniu wiadomości WIG20 oraz 0,06% dla jednej transakcji przy wykorzystaniu wiadomości ekonomicznych. Koszty transakcyjne sięgają jednak aż 0,2%, co znacznie przewyższa zyski. Wyniki te nie oznaczają jednak, iż hipotezę trzecią należy odrzucić. Jak wielokrotnie wspominałem, zastosowany w tej pracy prototyp został skonstruowany do celów badawczych, inwestor zatem może go stosować łącznie z innymi narzędziami, np. analizą fundamentalną lub techniczną. Ponadto duże instytucje finansowe mają możliwość obniżenia kosztów transakcyjnych poprzez negocjacje.

### **Mocne strony pracy**

Moim zdaniem, główna wartość tej pracy polega na tym, iż daje ona solidne podstawy do dalszych badań w kontekście Warszawskiej Giełdy Papierów Wartościowych. Z tego co mi wiadomo, jest to pierwsza praca z tego zakresu dla języka polskiego. Ponadto dołożyłem wszelkich starań, aby nie popełnić błędów wytykanych autorom poprzednich prac (podrozdział 3.5): między innymi uwzględniłem koszty transakcyjne, nie dokonywałem przerzucania danych, wziąłem pod uwagę różnicę w cenach otwarcia i zamknięcia.

Uważam, że w pracy badawczej rzeczą niezwykle istotną, mogącą zadecydować o sukcesie lub porażce, jest dobór odpowiednich danych wejściowych. W rozdziale 3. dokonałem więc analizy danych wejściowych, co często bywa pomijane w podobnych pracach. Ponadto wykorzystałem najnowocześniejsze narzędzia lingwistyczne stworzone specjalnie dla języka polskiego, to znaczy TaKiPi oraz Morfeusz. Za mój wkład w dalsze

badania dotyczące zagadnienia przedstawionego w tytule tej pracy, wkład, który może być wykorzystany również na innych rynkach, uważam:

- Porównanie prognozy dotyczącej indeksu oraz poszczególnych spółek wchodzących w skład tego indeksu.
- Opracowanie metody automatycznej identyfikacji wiadomości niewnoszących znacznej informacji (podrozdział 4.6.1).
- Zastosowanie metody poziomu ufności znacząco poprawiającej wyniki predykcji poprzez uwzględnienie, oprócz samej odpowiedzi generowanej przez naiwny klasyfikator Bayesa, również poziomu „pewności” tej odpowiedzi.

### **Punkt wyjścia do dalszych badań**

Dopiero dokładniejsze zapoznanie się z zagadnieniem omawianym w tej pracy pozwala zrozumieć, jak jest ono złożone. Praca to od początku była pomyślana jako punkt wyjścia do dalszych badań, toteż wybierane były na ogół rozwiązania prostsze, cieszące się większym uznaniem, a nie metody bardziej zaawansowane, lecz niedostatecznie zweryfikowane. Podrozdział ten zawiera listę wniosków, które mogą stanowić podstawę do dalszych badań. Dwa najważniejsze z nich to:

- Wprowadzenie selekcji wiadomości. W przypadku giełdowych inwestycji krótkookresowych (intraday) duże znaczenie mają koszty transakcyjne. Biorąc ten fakt pod uwagę, warto sprawdzić, czy nie okaże się, że znacznie korzystniejsza jest próba identyfikacji wybranych wiadomości, które z dużą pewnością spowodują znaczną zmianę ceny, niż próba stwierdzenia zmiany ceny dla każdej z wiadomości. Dowodem na to jest fakt, iż przedstawiona w podrozdziale 4.8 metoda poziomu ufności, uwzględniająca informację o pewności zwróconego przez klasyfikator wyniku, osiąga ponad dwukrotnie lepsze rezultaty od metody bazowej.
- Wyraźne rozróżnienie wiadomości publikowanych w czasie sesji giełdowej i poza nią. Przedstawiona w tej pracy metoda, w której efekt wiadomości opublikowanej później niż godzinę przed końcem sesji bądź po zakończeniu sesji jest „przerzucany” na następny dzień pracy giełdy, okazała się błędna. Godzinna zmiana ceny, która dokonuje się w ramach jednej sesji, ma inną wariancję niż zmiana ceny, która „przechodzi” na następny dzień. Nierozróżnienie tych grup skutkuje między innymi błędną identyfikacją fraz występujących pod koniec dnia jako fraz zwiększających niepewność na rynku. Proste

podejście, odrzucające wiadomości publikowane pod koniec dnia, może już poprawić wyniki .

Pozostałe wnioski:

- Próba połączenia analizy wiadomości prasowych z innymi dostępnymi dla inwestorów narzędziami, np. analizą techniczną lub fundamentalną.
- Sprawdzenie alternatywnych klasyfikatorów, np. metody najbliższych sąsiadów, której istotą jest próba znalezienia podobnych wiadomości (metryką może być liczba podobnych wyrazów).
- Prognozowanie kierunku zmian cen oraz wariancji<sup>1</sup>. Celem badania może być optymalizacja wskaźnika zysk/ryzyko.
- Sprawdzenie alternatywnych postaci funkcji przekształcającej zmianę ceny (podrozdział 4.4), szczególnie zaś krótszych okien czasowych.

---

<sup>1</sup> Pomysł pochodzi z pracy: J. Thomas: *op.cit.*

## BIBLIOGRAFIA

- Cho V., Wüthrich B., Zhang J.: Text Processing for Classification. *Journal of Computational Intelligence in Finance*. 1999, tom 7, s. 6-22.
- Czekaj J., Woś M., Żarnowski J.: *Efektywność giełdowego rynku akcji w Polsce z perspektywy dziesięciolecia*. Warszawa 2001.
- Fama E. F., French K.R.: Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*. 1993, tom 33, s. 3-56.
- Fama E. F.: Efficient capital markets: II. *Journal of Finance*. 1991, tom 46, s. 1575-1617.
- Fama E.F.: Efficient capital markets: A review of theory and empirical work. *Journal of Finance*. 1970, tom 25, s. 383-417.
- Fung G.P.C., Yu J.X., Lu H.: The predicting power of textual information on financial markets. *IEEE Intelligent Informatics Bulletin*. 2005, s. 1-10.
- Gidófalvi G., Elkan C.: *Using News Articles to Predict Stock Price Movements*. California, 2001.
- Grossman S., Stiglitz J.: On the Impossibility of Informationally Efficient Markets. *The American Economic Review*. 1980, tom 70, s. 393-408.
- Jain P. C.: Response of Hourly Stock Prices and Trading Volume to Economic News. *Journal of Business*. 1988, tom 61, s. 219-231.
- Kobos M.: *Przewidywanie cen akcji z wykorzystaniem artykułów prasowych*. Warszawa 2007.
- Lavrenko V., Schmill M., Lawrie D. [et al.]: Language Models for Financial News Recommendation. *Conference on Information and Knowledge Management*. McLean 2000, 9 konferencja, s. 389-396.
- Manning C. D., Raghavan P., Schütze H.: *Introduction to Information Retrieval*. Cambridge 2008.
- Matuszek C., Cabral J., Witbrock M. [et al.]: An Introduction to the Syntax and Content of Cyc. *Proceedings of the 2006 AAAI Spring Symposium*. Stanford 2006, s. 44-49.
- Mittermayer M.A., Knolmayer G.F.: *Text Mining Systems for Market Response to News: A Survey*. Bern 2006.
- Niemiro W.: *Rachunek Prawdopodobieństwa i Statystyka Matematyczna*. Warszawa 1999.
- Peramunetilleke D., Wong R.K.: Currency Exchange Rate Forecasting from News Headlines. *ACM International Conference Proceeding Series*. Melbourne 2002, 13 konferencja, s. 131-139.

- Piasecki M.: Polish Tagger TaKIPI: Rule Based Construction and Optimisation. *Task Quarterly*. 2007, tom 11, s. 151-167.
- Przepiórkowski A.: *Slajdy z wykładu inżynieria lingwistyczna*. Warszawa 2008.
- Rinaldo A.: *Intraday Market Dynamics Around Public Information Arrivals*. Fribourg 2003.
- Russel P., Torbey V.: *The Efficient Market Hypothesis on Trial: A Survey*. Philadelphia 2002.
- Strychowski J.: *Zastosowanie metod uczenia maszynowego w analizie morfologicznej języka polskiego*. Wydawnictwo Uniwersytetu Śląskiego, 2004.
- Tan A.: Text Mining: The state of the art and the challenges. *Pacific Asia Conference on Knowledge Discovery and Data Mining PAKDD*. 1999, s. 65-70.
- Thomas J.: *News and Trading Rules*. Pittsburgh 2003.
- Woliński M.: System znaczników morfosyntaktycznych w korpusie IPI PAN. *Polonica*. 2003, tom XII, s. 39-55.
- Wüthrich B., Leung S., Peramunetilleke D. [et al.]: Daily Prediction of Stock Market Indices from Textual WWW Data. *Conference on Knowledge Discovery and Data Mining*. New York 1998.

## ZESTAWIENIE SPISÓW

### Spis tabel

Tabela 1. Przegląd najpopularniejszych wartościowań dla wielozbiorów wyrazów .....	20
Tabela 2. Przykładowa macierz wystąpień .....	21
Tabela 3. Częstotliwość występowania kluczowych słów w danych źródłowych .....	36
Tabela 4. Zmiany cen akcji po publikacjach wiadomości – miary statystyczne.....	46
Tabela 5. Zmiany cen akcji – miary statystyczne .....	47
Tabela 6. Lista fraz o największej wartości miary informacji wzajemnej – grupa WIG20.....	50
Tabela 7. Lista fraz o największej wartości miary informacji wzajemnej po odfiltrowaniu informacji nieistotnych – grupa WIG20.....	51
Tabela 8. Lista fraz o największej wartości miary informacji wzajemnej – grupa ekonomiczna .....	52
Tabela 9. Lista fraz o największej wartości miary informacji wzajemnej po usunięciu wiadomości nieniosących nowych informacji – grupa ekonomiczna .....	54
Tabela 10. Przykład konwersji wiadomości prasowych na postać ilościową – WIG20 .....	55
Tabela 11. Porównanie skuteczności predykcji modeli w porównaniu do gracza losowego ..	60
Tabela 12. Wyniki walidacji krzyżowej (metoda poziomu ufności) – miary statystyczne.....	62
Tabela 13. Wyniki symulacji rynkowej dla wiadomości ekonomicznych.....	63
Tabela 14. Wyniki symulacji rynkowej dla wiadomości WIG20 .....	64
Tabela 15. Porównanie kosztów transakcyjnych z osiąganymi zyskami.....	65

### Spis rysunków

Rys. 1. Przegląd działania algorytmu .....	40
Rys. 2. Wyniki walidacji krzyżowej (metoda poziomu ufności) .....	61

## Spis tablic

Tablica 1. Słowa semantycznie puste.....	17
Tablica 2. Wyciąg z danych źródłowych, nagłówki wiadomości prasowych z serwisu Reuters z dnia 27 lutego 2007 do godziny 12:00 .....	33
Tablica 3. Wyciąg z danych źródłowych, nagłówki wiadomości prasowych z serwisu Money.pl z dnia 27 lutego 2007 .....	34
Tablica 4. Przykład zduplikowanej wiadomości .....	35
Tablica 5. Zbiór wiadomości zawierających słowo PKB .....	35
Tablica 6. Wynik analizy morfologicznej wybranych artykułów .....	43
Tablica 7. Lista nagłówków nieniosących nowych informacji .....	53
Tablica 8. Liczba poprawnych odpowiedzi dla grupy WIG20 (dane treningowe) .....	59
Tablica 9. Liczba poprawnych odpowiedzi dla grupy ekonomicznej (dane treningowe).....	59