**Akademia Górniczo-Hutnicza**
**im. Stanisława Staszica w Krakowie**

Wydział Informatyki, Elektroniki i Telekomunikacji

KATEDRA INFORMATYKI


PRACA MAGISTERSKA

DARIUSZ MYDLARZ

# MOŻLIWOŚCI POWIĄZANIA DANYCH GEOLOKACYJNYCH I ANALIZY SENTYMENTU W ANALIZIE ZACHOWAŃ UŻYTKOWNIKÓW W WYBRANYCH PORTALACH SPOŁECZNOŚCIOWYCH

OPIEKUN:
dr inż. Anna Zygmunt
KIERUNEK:
Informatyka

Kraków 2014

## OŚWIADCZENIE AUTORA PRACY

OŚWIADCZAM, ŚWIADOMY ODPOWIEDZIALNOŚCI KARNEJ ZA POŚWIADCZENIE NIEPRAWDY, ŻE NINIEJSZĄ PRACĘ DYPLOMOWĄ WYKONAŁEM OSOBIŚCIE I SAMODZIELNIE I ŻE NIE KORZYSTAŁEM ZE ŹRÓDEŁ INNYCH NIŻ WYMIENIONE W PRACY.

.....

PODPIS

**Streszczenie**

Tutaj będzie streszczenie pracy

# Spis treści

# Spis rysunków

# Spis tablic

# 1. Wstęp

W dzisiejszych czasach wpływ Internetu na życie codzienne jest niepodważalny. Już od kilku lat świat globalnej wioski przenika się z życiem realnym. Nikogo nie dziwią prezentowane w kanałach informacyjnych komentarze pochodzące z sieci, których autorami są zarówno osoby znane jak i zwykli internauci. Rozrost Internetu przebiega w błyskawicznym tempie, a wydarzenia na świecie komentowane są na żywo przez wielu ludzi. Aktualne trendy tworzone są na blogach, mikroblogach czy serwisach społecznościowych.

Wyzwanie wobec ogromu tych informacji podejmuje dzisiejsza informatyka. Przetwarzanie tak dużej ilości danych wymaga wielu zautomatyzowanych procesów. W dzisiejszych czasach nie wystarczy już dowiedzieć się kto z kim najczęściej się komunikuje, ale dużo bardziej interesujące jest to, o czym dany internauta pisze i w jaki sposób to czyni.

Wielkie firmy chcą wiedzieć jak odbierane są ich produkty, jakie emocje wzbudzają wśród klientów ich usługi i czy udaje im się spełniać ich oczekiwania. Analiza użytkowników serwisów społecznościowych może być także bardzo interesującym przedmiotem badań socjologów nad zmieniającym się społeczeństwem i wpływem Internetu na ten proces. Dodatkowo, analiza geolokalizacji może pozwolić marketingowcom na odkrywanie nowych rejonów świata, w których mogliby oferować swoje produkty i usługi.

Naprzeciw tym potrzebom budowane są systemy informatyczne, które potrafią takie informacje uzyskać, przetwarzać i prezentować. Przykład takiego systemu został zrealizowany w ramach tej pracy magisterskiej.

## 1.1. Cel pracy

Niniejsza praca skupia się na analizie zachowań użytkowników w wybranych portalach społecznościowych. Przedmiotem badań są użytkownicy serwisu mikroblogowego Twitter. W ramach pracy staram się odpowiedzieć na pytania:

– jak internauci korzystają z mediów społecznościowych,

– kiedy są najaktywniejsi,

– jakie wyrażają emocję,

– z jakich miejsc komentują,

– czy i w jakie grupy się łączą.

Analiza serwisów społecznościowych niesie ze sobą wiele wyzwań. Jako główne można wymienić:

– przetwarzanie języka naturalnego – wiele skrótów, wyrażeń slangowych, błędów ortograficznych czy typograficznych, sklejanie wyrazów, używanie słów zapożyczonych z obcych języków, itp.,

– ogromna ilość przetwarzanych informacji,

– duża liczba krótkich wiadomości,

– duża liczba danych zaszumionych – wpisy reklamowe (SPAM), automatycznie wklejanie linków do blogów, innych serwisów społecznościowych, itp.

W związku z powyższym zebrane dane muszą być odpowiednio przetworzone i przefiltrowane zanim zostaną przeprowadzone na nich jakiekolwiek operacje.

W ramach tej pracy pobrałem z serwisu Twitter w ciągu 3 miesięcy blisko 8 milionów wpisów (w tym także tych zawierających informacje o geolokalizacji), opracowałem metodę analizy sentymentu – czyli wydźwięku wypowiedzi (pozytywna, negatywna lub neutralna), stworzyłem narzędzie wspomagające analizę zebranych danych – wyświetlanie szerokiej gamy wykresów, prezentowanie wpisów na mapie, informowanie o sentymencie. Ponadto przeprowadziłem analizę sieci społecznej jaka wyłoniła się z zebranych danych.

## 1.2. Zawartość pracy

krótkie opisanie zawartości rozdziałów

# 2. Przegląd badań

W niniejszym rozdziale znajduje się aktualny stan badań dotyczący 4 tematów, które składają się na tę pracę. Na początku opisana jest dziedzina sieci społecznych, czym ta nauka się zajmuje, w jakich przypadkach może zostać zastosowana. Następnie omówiona zostaje analiza sentymentu wypowiedzi i przetwarzanie tekstu celem ekstrakcji jego wydźwięku. Później skupiam się nad temat związanym z geolokacją i opisem, co można dzięki niej się dowiedzieć, a rozdział kończę omówieniem serwisu społecznościowego Twitter, który został wykorzystany jako źródło danych do analizy sieci społecznych.

## 2.1. Sieci społeczne

Termin ten został użyty po raz pierwszy w 1954 roku przez Johna Arundela Barnesa [2]. Oznacza strukturę społeczną, którą tworzą jednostki (np. osoby lub organizacje) i połączenia między nimi. Analiza sieci społecznych jest badaną od wielu lat dziedziną nauki. Szybki rozwój Internetu w XXI wieku wzbogacił ją o bogate źródło danych. Główne obszary badań [1] to między innymi:

– statystyczna analiza sieci społecznych – opisuje jak wygląda typowa sieć społeczna, badane są połączenia między jednostkami, aby sprawdzić czy posiadają kilka połączeń, czy sieć zbudowana jest z *hubami*, czy może liczba połączeń rozłożona jest równomiernie,

– odkrywanie grup/społeczności – jest jednym z głównych tematów analizy sieci społecznych; szukanie grup związane jest z klasteringiem i odkrywaniem regionów sieci, które są odpowiednio gęste; problem powiązany jest z badaniem grafów, określaniem jak dzielić sieć na regiony,

– klasyfikacja wierzchołków - w niektórych sieciach część wierzchołków może być oznaczona; badania skupiają się na tym by korzystając z atrybutów danego wierzchołka móc ocenić jakie inne również mogłyby przyjąć daną etykietę,

– odnajdywanie ekspertów – sieci społeczne mogą być używane jako narzędzia w celu odkrywania ekspertów do danego zadania,

– przewidywanie przyszłych połączeń wewnątrz sieci – wiele badań skupia się na połączeniach wierzchołków celem odkrycia interesujących informacji na temat sieci społecznej; w wielu sieciach połączenia między węzłami są dynamiczne i wynikiem tych badań może być poprawna predykcja przyszłych połączeń wewnątrz sieci,

– ekstrakcja wiedzy z sieci – polega na eksploracji danych z mediów społecznościowych i eksploracji tekstu z serwisów społecznościowych; eksploracja danych dostarcza naukowcom narzędzia do analizy dużych, złożonych i często zmieniających się danych wewnątrz sieci, a eksploracja tekstu może prowadzić do odkrycia nowych połączeń między węzłami i nowych charakterystyk je łączących; jej użycie wpływawpłynąć na poprawę jakości badanej sieci.

### 2.1.1. Przykłady zastosowania sieci społecznych

Wyniki badań nad sieciami społecznymi stwarzają wiele możliwości dla różnych dziedzin życia. Ich zastosowanie może być zaaplikowane przez:

– służby porządkowe – policja może przy ich pomocy odkrywać powiązania między przestępcami i dochodzić do zależności między grupami przestępczymi, a także odkrywać, kogo dane grupy mogłyby zwerbować,

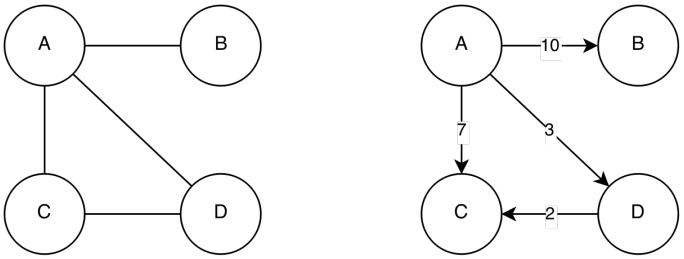
– badania naukowe – odkrywanie naukowców zajmujących się podobnymi tematami celem opracowania bardziej kompletnych wyników lub podjęcia nowego, wspólnego tematu,

– przedsiębiorstwa handlowe – odkrywanie zbliżonych typów klientów i oferwanie im produktów lub usług do nabycia przy użyciu systemów rekomendujących,

– służby zdrowotne – użycie sieci społecznych może pomóc w określaniu obszarów, w które rozprzestrzeniają się wirusy groźnych chorób, dzięki czemu możliwe może być zapobieganie ich dalszej ekspansji.

### 2.1.2. Reprezentacja sieci społecznych

Najczęściej spotykaną reprezentacją sieci społecznych jest reprezentacja grafowa. W naturalny sposób modeluje ona jednostki jako węzły grafu i relacje między nimi jako krawędzie. W zależności od rodzaju sieci graf taki może być skierowany lub nieskierowany oraz o krawędziach ważonych lub nieważonych. W przypadku krawędzi ważonych waga danego połączenia może reprezentować na przykład liczbę wiadomości wymienionych między węzłami. Kierunek krawędzi reprezentuje, w która stronę dana komunikacja przebiegała. Wygląd takich grafów przedstawia rysunek 2.1.



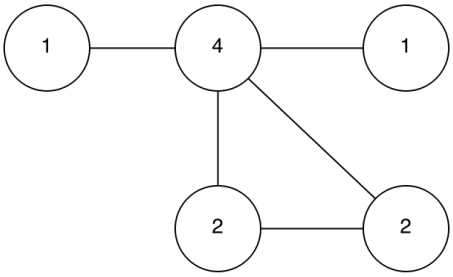

Rysunek 2.1: Przykład grafu nieskierowanego bez krawędzi ważonych (po lewej) oraz grafu skierowanego z krawędziami ważonymi (po prawej)

### 2.1.3. Miary i pojęcia grafowe

Zamodelowanie sieci społecznych w postaci grafów pozwala na skorzystanie z szeregu miar związanych z tą dziedziną wiedzy. Dzięki nim możliwe jest odnajdywanie cech charakterystycznych danej sieci. Najważniejsze miary pomagające odnaleźć najważniejsze węzły to to [4]:

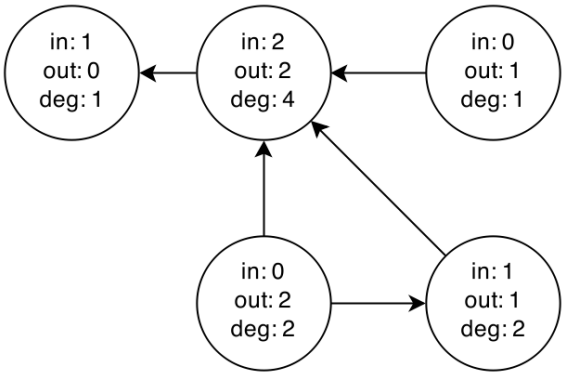
**Stopień wierzchołka**

Miara określająca liczbę krawędzi wchodzących i wychodzących z wierzchołka (patrz rys. 2.2)




Rysunek 2.2: Graf z oznaczonymi stopniami wierzchołków

W przypadku grafów skierowanych możemy jeszcze mówić o stopniu wchodzącym (ang. *in degree*) oraz wychodzącym (ang. *out degree*) (patrz rys. 2.3).



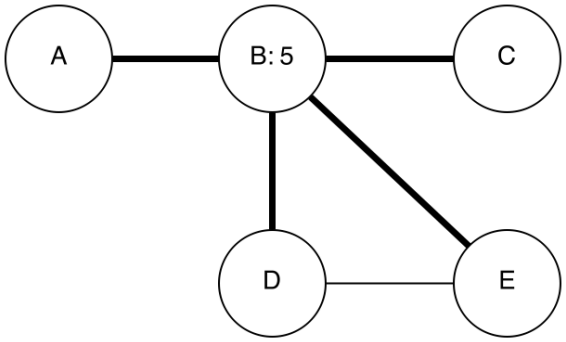

Rysunek 2.3: Graf skierowany z oznaczonymi stopniami wierzchołków

**Pośrednictwo (ang. *betweenness*)**

Liczba najkrótszych ścieżek w grafie, które przechodzą przez dany węzeł podzielona przez liczbę wszystkich najkrótszych ścieżek grafu. Przez najkrótszą ścieżkę rozumie się taką ścieżkę między dwoma węzłami grafu, dla której liczba krawędzi jest najmniejsza.

$$BC(k) = \sum_i \sum_j \frac{\rho(i,k,j)}{\rho(i,j)}, \quad i \neq j \neq k \tag{2.1}$$

Przykładowo aby obliczyć wartość tej miary dla wierzchołka $B$ posłużmy się rysunkiem 2.4. Najkrótsze ścieżki między węzłami innymi niż $B$ to: $ABC$, $ABD$, $ABE$, $CBD$, $CBE$, $DE$. W 5 z 6 z nich znajduje się węzeł $B$, stąd wynika jego wartość *betweenness* równa $5/6$.




Rysunek 2.4: Najkrótsze scieżki przechodzące przez węzeł $B$

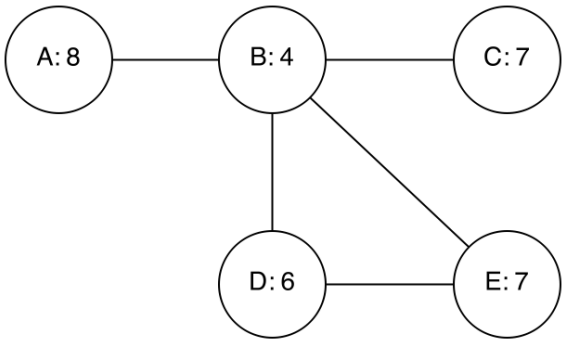
Węzły o wysokiej wartości współczynnika *betweenness* są interesujące ponieważ mogą kontrolować przepływać informacji wewnątrz sieci oraz mogą być zmuszone do przetwarzania większej ilości informacji. Z tego wynika też, że mogą być skutecznym celem ataków.

**Bliskość (ang. *closeness*)**

Znormalizowana odwrotność sumy odległości między węzłami w grafie.

$$CC(i) = \frac{N-1}{\sum_j d(i,j)} \tag{2.2}$$

Dla każdej pary węzłów liczymy odległość między nimi (liczoną jako liczbę krawędzi), a następnie dla każdego wierzchołka dzielimy tę wartość przez $N-1$, gdzie $N$ to liczba wierzchołków. Przykładowy graf 2.5 i tabela 2.1 z obliczeniami znajdują się poniżej.




Rysunek 2.5: Suma odległości do pozostałych węzłów w grafie

| | Wierzchołki | | | | | Suma odległości | Bliskość (*closeness*) |
|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | $\sum_j d(i,j)$ | $CC(i)$ |
| A | 0 | 1 | 2 | 2 | 3 | 8 | 0.5 |
| B | 1 | 0 | 1 | 1 | 1 | 4 | **1.0** |
| C | 2 | 1 | 0 | 2 | 2 | 7 | 0.57 |
| D | 2 | 1 | 2 | 0 | 1 | 6 | 0.67 |
| E | 3 | 1 | 2 | 1 | 0 | 7 | 0.57 |

Tablica 2.1: Odległości między węzłami i wartości miary *closeness*

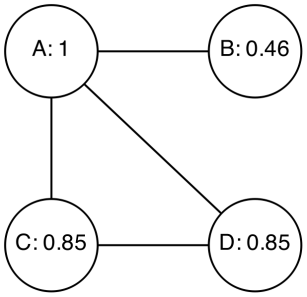
Węzłem o najmniejszej sumie odległości do innych wierzchołków – a co za tym idzie – o największej wartości bliskości jest węzeł $B$. Wynika z tego, że jest to wierzchołek najszybciej rozsyłający informacje wewnątrz sieci pomiędzy jej elementami.

**Wektor własny (ang. *eigenvector*)**

Miara centralności węzła, która oceniając dany węzeł bierze także pod uwagę wartości jego sąsiadów (bezpośrednio przyległych węzłów). Zastosowanie tej wielkości pozwala wskazać najważniejszy węzeł w sytuacji, gdy poprzednie miary zwracają równe wyniki. Wartość tej wielkości wyraża się wzorem:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j \tag{2.3}$$

gdzie $\lambda$ jest stała i równa największej wartości własnej macierzy sąsiedztwa danego grafu, a wartość $A_{ij} = 1$, gdy wierzchołki mają wspólną krawędź, w przeciwnym wypadku wynosi $0$. Przykładowe wartości wielkości *eigenvector* zaprezentowano na rysunku 2.6. Wartości te zostały obliczone przy pomocy narzędzia Gephi[1].

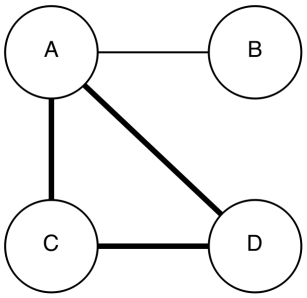



Rysunek 2.6: Wartości wielkości *eigenvector* w przykładowym grafie

Dla uzupełnienia terminologii związanej z traktowaniem sieci społecznej jako grafu chciałbym przypomnieć jeszcze dwa pojęcia:

– klika – podgraf grafu, w którym wszystkie wierzchołki połączone są krawędzią

– k-klika – klika składająca się z dokładnie $k$-wierzchołków. Na przykład k-klika o $k = 3$ to podgraf zbudowany z 3 wierzchołków, gdzie między każdym z nich znajduje się krawędź (patrz rys. 2.7).




Rysunek 2.7: Wierzchołki $ABC$ tworzą klikę o rozmiarze 3

[1] https://gephi.github.io/

## 2.2. Sentyment wypowiedzi

Recenzje, komentarze i opinie odgrywają istotną rolę w ocenie satysfakcji z produktu lub usługi czy w badaniu reakcji na wydarzenia. Dane, które zawierają takie informacje mają bardzo wysoki potencjał w odkrywaniu wiedzy. Dowiadywanie się, co myślą inni ludzie zawsze było bardzo istotne w procesie podejmowania decyzji. Internet dał nam możliwości zapoznania się z opiniami innych zwykłych ludzi, ale także pozwolił na poznanie komentarzy ekspertów w swoich dziedzinach. Badanie sentymentu – a więc wydźwięku wypowiedzi (ocena wypowiedzi jako pozytywnej, negatywnej lub neutralnej) – odgrywa bardzo istotną rolę. Jak wynika z badań przeprowadzonych na ponad 2000 dorosłych Amerykanów [6] 81% użytkowników Internetu przynajmniej raz poszukiwało w Internecie informacji o jakimś produkcie z czego od 73% do 87% osób twierdzi, że recenzje innych miały wpływ na ich wybory.

Zastosowanie analizy sentymentu jest bardzo szerokie. Niektóre z obszarów jej użycia to [7]:

– portale internetowe z opiniami – zastosowanie analizy sentymentu może być użyte do poprawy błędów popełnionych przez użytkowników (gdy opinia jest pozytywna, a użytkownik omyłkowo wybrał niską ocenę) lub gdy opinie są ewidentnie stronnicze, mogą pomóc w faktycznej ocenie danego przedmiotu czy usługi

– jako technologia wspomagająca większe systemy – analiza sentymentu może być wsparciem dla systemów rekomendacji; na przykład może służyć do nie rekomendowania produktów, które otrzymują negatywne opinie; w systemach serwujących reklamy kontekstowe, wykrycie pozytywnego sentymentu na stronie może być powodem wyświetlenia jakiejś reklamy, a wykrycie negatywnego sentymentu powodem jej ukrycia; innym zastosowaniem jest ekstrakcja informacji, która może być polepszona poprzez pomijanie zdań subiektywnych, zawierających sentyment

– biznes – poprzez dostarczenie informacji o odbiorze sprzedawanych produktów i serwowanych usług; gdy na przykład sprzedawany laptop ma negatywny odbiór stosując analizę sentymentu można to bardzo szybko wykryć i dowiedzieć się dlaczego zaistniała dana sytuacja; firma może badać swój ogólny odbiór w społeczeństwie – szybko reagować na niezadowolenie klientów, lub wprowadzać poprawki do swoich produktów; wykrywanie sentymentu może również pomóc przewidzieć wyniki sprzedaży

– polityka – użycie analizy sentymentu jest wręcz naturalne dla tego obszaru życia; partie czy politycy mogą badać odbiór społeczeństwa swoich programów i decyzji; badanie sentymentu może im na przykład wskazać w jakich miejscach, czy przy jakich postaciach się pokazać by zyskać sympatię wyborców; istotne również mogą być informacje na temat rekacji społeczeństwa na planowane przez rząd zmiany w prawie.

Krótko mówiąc największym zyskiem związanym z badaniem sentymentu jest możliwość zbadania opinii bardzo dużej liczby osób w sposób mechaniczny. Nie ma potrzeby przeprowadzania ankiet, pytania ludzi co sądzą na dany temat. Internauci samodzielnie przedstawiają swoje opinie w Internecie, a przy pomocy analizy sentymentu bardzo łatwe staje się zbadanie nastrojów.

Badanie sentymentu nie jest trywialne. Związane jest bezpośrednio z przetwarzaniem języka naturalnego, które niesie ze sobą szereg problemów:

– złożoność języka naturalnego – bardzo trudnym zadaniem jest nauczenie programu komputerowego pełnego rozumienia języka naturalnego; co więcej każdy język jest inny, więc dla każdego konieczne jest zastosowanie różnych podejść – inaczej trzeba zabrać się za badanie sentymentu w języku polskim a inaczej w angielskim; język ciągle się rozwija, nie jest martwy,

– trudność w analizie kontekstu wypowiedzi – wykrycie ironii nie jest zadaniem prostym; bardzo często wypowiedzi mogą mieć związek z jakimś pojęciem zupełnie niezrozumiałym dla programu komputerowego, a oczywistym dla człowieka (np. idiomy, odniesienia do wydarzeń na świecie)

– slang w Internecie, skrótowce, literówki – wszystkie te elementy dodatkowo utrudniają analizę sentymentu; użytkownicy Internetu nie zawsze dbają o jakość swojego języka, często stosują skróty, czy wyrażenia slangowe, które mogą być niezrozumiałe dla automatycznego analizatora sentymentu;

– SPAM, szum – wszystkie wpisy, które nie niosą ze sobą żadnej wartości a pojawiają się w internetowych forach, serwisach z opiniami również stanowią wyzwanie przy budowie narzędzia do analizy sentymentu.

### 2.2.1. Techniki badania sentymentu

Podejść do badania sentymentu jest wiele. Poniżej przedstawione są te, które najlepiej nadają się do badania sentymentu na Twitterze (w związku z tym, że to ten serwis jest źródłem danych w tej pracy), a które zostały opisane w artykule [3]. Oprócz nich przedstawię także metodę opracowaną przez Alexandra Paka i Patricka Paroubek'a [5], którą zastosowałem w swoich badanich. Techniki te to:

**Podejście oparte na słowniku (ang. *lexicon based approach*)**

Podejście polega na zastosowaniu słownika z wyrazami oznaczonymi jako pozytywne i negatywne. Klasyfikator ocenia tekst na podstawie liczby wystąpień odpowiednich słów. Niestety podejście to ma bardzo wysoki stopień błędów. Przykładowa funkcja oceniająca sentyment słowa to:

$$X_t = \frac{p(pos|topic, t)}{p(neg|topic, t)} \tag{2.4}$$

w tym przypadku wyrazy mają przypisany odpowiedni sentyment w zależności od tematu, którego dotyczą.

Największym problemem tego podejścia jest brak mechanizmu radzenia sobie z kontekstem słów.

**Naiwny klasyfikator Bayesa (ang. *naive Bayes classifier*)**

Jest to podejście probabilistyczne. W ramach tej metody zakłada się, że dana kategoria tekstów $k_1$ (np. pozytywne) charakteryzuje się określonym słownictwem, a inna $k_2$ (negatywne) innym słownictwem. Na tej podstawie określamy prawdopodobieństwo jeszcze przed przeprowadzeniem jakiejkolwiek klasyfikacji tekstu. Zakłada się także, że tekst, który posiada słownictwo z kategorii $k_1$ w większej liczbie

niż z kategorii $k_2$, powinien być zaklasyfikowany do tej pierwszej. W tym przypadku jest to określenie klasyfikacji posiadając pewną wiedzę na temat badanego tekstu.

Naiwny klasyfikator Bayesa opiera się na założenia o wzajemnej niezależności słów. Oznacza to, że wyrazy, które identyfikują określoną kategorię mogą występować niezależnie w różnych lub tym samym tekście. Taki naiwny klasyfikator może więc identyfikować i klasyfikować słowa, nie biorąc pod uwagę kontekstu w jakim one występują. Pomimo, że jest to podejście naiwne, okazuje się skuteczne ze względu na swoją prostotę. Wzór Bayesa określa bowiem prawdopodobieństwo tego, że szanse przypisania tekstu do odpowiedniej klasy zależą od tego jak często jego słowa należą do różnych klas i jak często do nich nie należą.

Krótko mówiąc, jeśli naiwny klasyfikator Bayesa w wybranym tekście znajdzie więcej słów należących do klasy pozytywnej i jednocześnie mniej należących do negatywnej, wówczas większe będzie prawdopodobieństwo zaklasyfikowania tekstu do pierwszej kategorii. Klasyfikator ten zbiera uczy się klas wyrazów sukcesywnie analizując kolejne teksty [9].

#### Technika maksymalnej entropii (ang. *maximum entropy technique*)

Technika estymacji rozkładu prawdopodobieństwa. Główna zasada polega na tym, że jeśli dane nie są dobrze znane, rozkład powinien być jak najbardziej jednolity, to znaczy mieć maksymalną entropię. Do tej techniki mogą dochodzić ograniczenia, które pozwalają by rozkład nie był maksymalnie jednolity. Ograniczenia takie mogą pochodzić z oznaczonych już danych treningów i reprezentowane jako oczekiwane wartości wybranych cech (wyrazów).

Na przykład w jakimś przypadku możemy założyć, że 50% wpisów jest pozytywnych, wówczas pozostałe klasy powinny posiadać po 25% prawdopodobieństwa (negatywne, neutralne). Taki model jest łatwy do zbudowania, ale staje się on bardziej skomplikowany wraz z rosnącą liczbą ograniczeń. Jako cechy dodawane mogą być również składniki wielowyrazowe zwiększające skuteczność tej techniki. Dlatego też podejście to nie cierpi z powodu założenia o niezależności wyrazów. Przykładowo wyrażenie „do widzenia" może być traktowane jako całościowy term, a nie jako każdy wyraz z osobna.

Niestety w związku z tym, że ograniczenia pochodzą z danych treningowych, jest duża szansa, że dane te będą relatywnie rzadkie i metoda ta może prowadzić do przeuczenia.

#### Maszyny wektorów nośnych (ang. *support vector machines*)

Support vector machines to podejście stosujące duży margines między klasami. Główna idea polega na znalezieniu hiperpłaszczyzny, która podzieli teksty na pozytywne i negatywne z marginesem pomiędzy klasami tak dużym jak to tylko możliwe. Technika ta zbudowana jest na zasadzie strukturalnej minimalizacji ryzyka (ang. *structural risk minimization principle*). Celem jest znalezienie funkcji $h$, dla której błąd klasyfikacji losowego tekstu będzie jak najmniejszy. Oznaczając hiperpłaszczyznę przez $\vec{h}$, a tekst przez $\vec{t}$ oraz klasy, do których może trafić jako $C_j \in \{1, -1\}$ wówczas możemy zapisać to postaci:

$$\vec{h} = \sum_i \alpha_i C_i \vec{t_j}, \quad \alpha_i \geq 0 \tag{2.5}$$

Wartość $\alpha_i$ może być znaleziona przez rozwiązanie problemu podwójnej optymalizacji. Teksty o $\alpha_i$ większym od zera, to te które biorą udział w szukaniu funkcji $h$ nazywa się je wektorami wspierającymi (ang. *support vectors*).

Wybór cech (wyrazów) jest bardzo ważnym zadaniem w technikach uczenia maszynowego. Musi to zostać tak wykonane by uniknąć przeuczenia i jednocześnie zwiększyć ogólną dokładność. Maszyny wektorów nośnych mają wysoki potencjał radzenia sobie z dużą liczbą wymiarów. Mierzą złożoność hipotezy którą dzielą dokumenty, a nie liczbę cech. W związku z tym liczba cech nie jest problemem. Technika ta radzi sobie z duża liczbą słów poprzez oznaczanie części z nich jako nieistotne (tych najrzadziej pojawiających się). Niestety czasami prowadzi to do utraty informacji.

Chociaż SVM przewyższa wszystkie tradycyjne metody klasyfikacji sentymentu, to niestety jest czarną skrzynką. Trudne jest zbadanie natury klasyfikacji i zidentyfikowanie, które słowa są dla niej istotne. Jest to jedna z głównych wad korzystania z tej techniki do klasyfikacji tekstów.

#### Metoda Alexandra Paka i Patricka Paroubek'a

Technika jest odpowiedzią na problemy związane z brakiem odpowiedniego słownika do oceny sentymentu. Została opracowana z uwzględnieniem Twittera i korzysta w związku z tym z pewnych założeń. Skoro nie ma żadnego idealnego słownika ze słowami oznaczonymi jako pozytywne lub negatywne, to trzeba go mechanicznie zbudować. Do budowy takiego leksykonu zostały wykorzystane wpisy na Twitterze, które zawierają emotikony podzielone na pozytywne (np. `:)`) i negatywne (np. `;(`).

Następnie spośród ściągniętych wpisów z Twittera analizowane są te, które zawierają odpowiednie emotikony i zliczana jest liczba wystąpień każdego wyrazu w każdym ze zbiorów (pozytywnym i negatywnym). W wyniku tego budowany jest leksykon zawierający wyrazy wraz z liczbą ich wystąpień w każdej z klas. W związku z tym, że wpisy na Twitterze ograniczone są do 140 znaków, autorzy przyjęli założenie, że emotikona dotyczy całego wpisu. Ocena tekstu $T$ składającego się z wyrazów $w_1, w_2, \ldots w_n$ obliczana jest jako:

$$valence(T) = \frac{\sum_{i=1}^{n} valence(w_i)}{n} \tag{2.6}$$

gdzie wartość $valence(w_i)$ obliczana jest przy zastosowaniu skonstruowanego leksykonu i równa delta IDF (ang. *inverse document frequency* – powszechnie stosowana miara ważności słowa w oparciu o liczbę wystąpień):

$$valence(w_i) = log\frac{N(w_i, M^+) + 1}{N(w_i, M^-) + 1} \tag{2.7}$$

Zastosowanie takiego wzoru prowadzi do tego, że niezależnie jak często dany wyraz pojawił się w zbiorze treningowym, najważniejsza jest jego polaryzacja. Gdy na przykład słowo *świetny* pojawiło się w zbiorach pozytywnym i negatywnym odpowiednio 1000 i 20 razy, a słowo *przezacny* odpowiednio 50 i 1 raz to ich wpływ na ocenę tekstu będzie identyczny.

## 2.3. Geolokacja

Geolokacja to identyfikacja położenia geograficznego jakiegoś obiektu. Może odnosić się do procesu zdobywania takiej informacji lub do już zdobytej wiedzy na ten temat. Główne sposoby pozyskiwania takich danych to:

- korzystanie z urządzeń GPS – wbudowanych we współczesne telefony komórkowe, tablety, itp.,
- pozycjonowanie względne – ustalanie pozycji na podstawie bazowych stacji telefonii komorków, ruterów WI-FI,
- użcyie bazy adresów przypisanych do IP.

Zastosowanie geolokacji może być bardzo szerokie, między innymi:

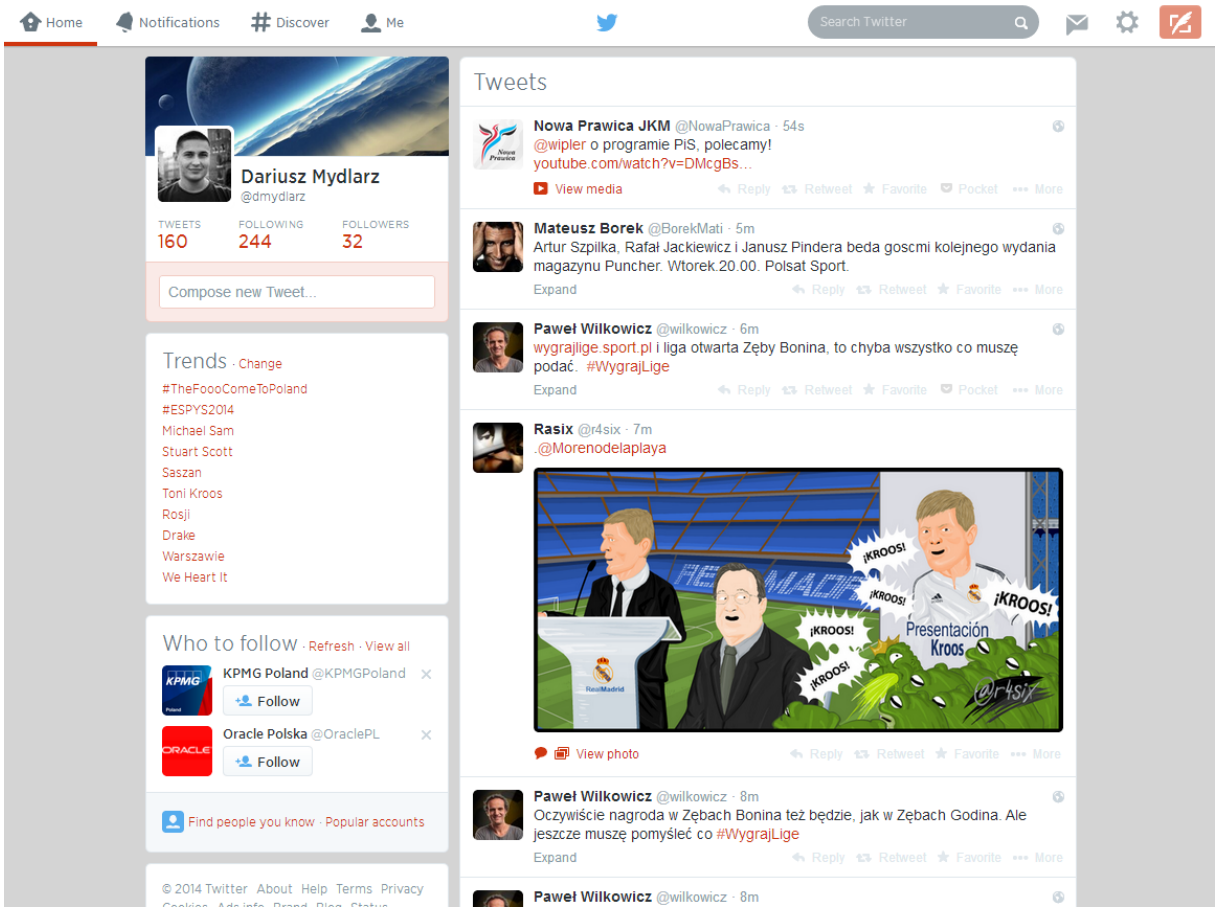
- dostarczanie lokalnych wiadomości
- dystrybucja treści cyfrowych – może być np. blokowana możliwość kupna, dla niektórych lokalizacji
- wyszukiwanie lokalnych usług, przedsiębiorstw
- wyświetlanie zlokalizowanych reklam
- zapobieganie nadużyciom zakupowym – sprawdzenie geolokacji klienta sklepu internetowego i porównanie jej z danymi z karty kredytowej, w celu ochrony osób, którym na przykład taka karta została skradziona
- prezentowanie różnych treści na stronach w zależności od lokalnego prawego (np. ukrywanie treści zabronionych w danym miejscu).

W szczególności w przypadku sieci społecznych geolokacja może być pomocna do ustalenia miejsca przebywania danych grup i może prowadzić do uzupełnienia zebranych danych o kolejne, wzbogacające analizę danej społeczności, pozwalające na wyciągnięcie bogatszych wniosków. Pomocnym może być na przykład zbadanie reakcji społeczeństwa w różnych regionach kraju na planowane zmiany w prawie przez rząd – i może to prowadzić albo do ich wprowadzenia albo wycofania.

Jako główne zalety stosowania geolokacji z punktu widzenia użytkowników telefonów komórkowych, sieci społecznych ([8] to dzielenie się ze społecznością (56%) oraz dzielenie się z osobami, które znają lub mogą spotkać (41%). Głównymi problemami, przed dzieleniem się geolokacją są obawy o prywatność (33%) oraz brak korzyści, zainteresowania (26%).

## 2.4. Twitter

Twitter to serwis społecznościowy o charakterystyce mikrobloga. Pozwala on na umieszczanie wpisów nie dłuższych niż 140 znaków. Domyślnie wszystkie wpisy są publiczne, a użytkownicy mają możliwość publicznej wymiany zdań z innymi. Każdy użytkownik ma możliwość wyboru użytkowników, których wpisy chce widzieć na swojej stronie głównej (patrz rys. 2.8).




Rysunek 2.8: Strona główna serwisu Twitter

Podstawowe pojęcia związane z tym serwisem to:

– śledzenie (ang. *follow*) – osób, organizacji; śledzenie jakiegoś użytkownika oznacza wyświetlanie wszystkich jego wpisów na swojej stronie głównej

– tweet – pojedynczy wpis/post na Twitterze; maksymalnie długość to 140 znaków; może dodatkowo zawierać zdjęcie lub informację o geolokalizacji,

– retweet – oznacza przekazanie jakiegoś wpisu dalej; jeśli użytkownik A śledzi użytkownika B i użyje funkcji retweet dla jednego z jego wpisów, wówczas osoby śledzące użytkownika A, również zobaczą ten wpis na swojej stronie głównej,

– odpowiedź (ang. *reply*) – odpisanie na jakąś wiadomość w serwisie Twitter, skomentowanie jej; serwis łączy takie wpisy w jedną grupę, wyświetlając je jeden obok drugiego,

– newsfeed – inaczej strona główna użytkownika, na której widzi wszystkie tweety wysłane przez osoby, które śledzi,

– hashtag – użycie symbolu # wraz z jakimś słowem, ułatwia rozmowy na wspólne tematy, wśród większych grup użytkowników (np. #worldcupfinal dla osób komentujących finał mistrzostw świata).

### 2.4.1. Twitter jako źródło danych

# 3. Koncepcja rozwiązania

- jak i dlaczego łącze te 3 elementy
- graf gruboziarnisty
- jak sentyment, geolokacja i sieci społeczne się łączą albo nie

# 4. Architektura systemu

- moduły
- szczegóły
- baza danych

# 5. Eksperymenty

- plan eksperymentów, opis danych
- wyniki, wnioski poszczególnych eksperymentów

# 6. Zakończenie i wnioski

## 6.1. Podsumowanie

- czy założone cele zostały osiągnięte - dyskusja
- w jakich punktach cele nie zostały osiągnięte
- dyskusja ograniczeń systemu, w jaki sposób należałoby podejść do ich rozwiązania

## 6.2. Możliwe kierunki rozwoju

- jak wyglądałby plan dalszego rozwoju systemu
- jaki jest wpływ (impact) pracy, gdzie wyniki zostały (mogą) być użyteczne, jakie może być ich zastosowanie teraz i w przyszłości

# Bibliografia


[1] C. Aggarwal. An Introduction to Social Network Data Analytics. In C. Aggarwal, editor, *Social Network Data Analytics*, pages 8–13. Springer, 2011.

[2] J. A. Barnes. Class and Committees in a Norwegian Island Parish. *Human Relations*, 7:39–58, 1954.

[3] S. Bhuta, U. Doshi, A. Doshi, and M. Narvekar. A review of techniques for sentiment analysis of twitter data. Technical report, International Conference on Issues and Challenges in Intelligent Computing Techniques, 2014.

[4] Ernesto Estrada. Introduction to network theory: Modern concepts, algorithms and application. Technical report, Scottish Insight Institute, University of Strathclyde, 2008.

[5] A. Pak and P. Paroubek. Twitter for sentiment analysis: When language resources are not available. Technical report, 22nd International Workshop on Database and Expert Systems Applications, 2011.

[6] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, pages 1–2, 2008.

[7] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, pages 11–15, 2008.

[8] W. Reese and J. Beckland. Lost in geolocation. Technical report, Spring, 2011.

[9] K. Tomanek. ąnaliza sentymentu-- metoda analizy danych jakościowych. przykład zastosowania oraz ewaluacja słownika rid i metody klasyfikacji bayesa w analizie danych jakościowych. *Przegląd Socjologii Jakościowej*, 10(2):118–136, 2014.