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

POSSIBILITIES OF USING GEOLOCATION AND SENTIMENT IN THE ANALYSIS OF USERS' BEHAVIOUR IN SOCIAL NETWORKS

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

Składam serdecznie podziękowania
Pani dr inż. Annie Zygmunt
za pomoc i udzielone wsparcie przy realizacji niniejszej pracy.

Osobne podziękowania składam również moim rodzicom,
rodzeństwu i narzeczonej.

Streszczenie

Analiza sieci społecznych to nauka o kilkudziesięcioletniej historii. Skupia się na badaniu połączeń między użytkownikami, odkrywaniu grup między nimi, a także rodzajach relacji jakie między sobą tworzą.

W niniejszej pracy staram się pokazać w jaki sposób można połączyć ze sobą analizę sentymentu (badanie wydźwięku – określanie wypowiedzi jako pozytywne lub negatywne) i geolokalizację w celu bogatszego badania sieci społecznych. Ich zastosowanie otwiera możliwość odkrycia ciągów przyczynowo-skutkowych w procesie tworzenia się sieci społecznych.

W tym celu zebrane zostało blisko 8 milionów wpisów z serwisu Twitter, których autorami było ponad 1,5 miliona internautów. Zebrane dane były ukierunkowane na środowisko piłkarskie związane z najwyższą klasą rozgrywkową w Anglii i Walii – Barclays Premier League w sezonie 2013/2014.

Praca zawiera opis zastosowanych technik i przykłady ich wykorzystania na zebranych danych. Udowadnia, że przy ich pomocy można wysoce zautomatyzować sposób badania nastrojów społecznych z uwzględnieniem lokalizacji geograficznej czy grup, które wspólnie tworzą. Może być zastosowana jako *proof-of-concept* (z ang. potwierdzenie, dowód koncepcji) w badaniach na szerszą skalę niż tu zaprezentowana.

Spis treści

Spis rysunków

1. Wstęp

W dzisiejszych czasach wpływ Internetu na życie codzienne jest niepodważalny. Już od kilkunastu 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. Nie wystarczy już dowiedzieć się kto z kim najczęściej się komunikuje, ale 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. 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

Celem niniejszej pracy jest przedstawienie możliwości wykorzystania analizy sentymentu oraz geolokacji w analizie zachowań użytkowników sieci społecznościowych. Praca ta ma przedstawić czy możliwe jest połączenie tych trzech dziedzin ze sobą i jeśli tak, to w jaki sposób to zrobić i co dzięki temu można się dowiedzieć. Analiza sentymentu polega na badania wydźwięku wypowiedzi. Polega ona na określaniu danego tekstu jako pozytywnego, negatywnego lub neutralnego. Zastosowanie komputerowych technik do badania sentymentu pozwala na dużo szybszą ocenę pojedynczego tekstu lub grupy tekstów w porównaniu do ręcznego badania. Geolokacja to technika identyfikacji geograficznego położenia osoby lub urządzenia za pomocą cyfrowych danych przetwarzanych przy pomocy Internetu. Poprzez zastosowanie nadajników GPS w nowoczesnych urządzeniach elektronicznych (telefony, tablety) ich użytkownicy mogą udostępniać innym informacje o swojej lokalizacji. Zastosowanie geolokacji po-

zwala odkrywać charakterystyczne zachowania w różnych rejonach świata. Analiza sieci społecznych zajmuje się badaniem struktury społecznej tworzonej przez jednostki (osoby lub organizacji) i połączeniami między nimi. Spośród trzech wymienionych dziedzin jest najdłużej badaną przez naukowców. Pomaga odkrywać rodzaje połączeń między jednostkami, kierunki ewolucji danych grup oraz przewidywać zmiany w nich zachodzące.

Przedmiotem niniejszych badań są użytkownicy serwisu mikroblogowego Twitter[1]. Skupiono się na osobach komentujących mecze piłkarskie angielskiej Premier League. Praca podejmuje następujące tematy:

– jak internauci korzystają z mediów społecznościowych – jakie charakterystyczne zachowania można zauważyć,

– kiedy są najaktywniejsi i tworzą najwięcej wpisów i komentarzy,

– jakie wyrażają emocję – kiedy i dlaczego tworzą wpisy pozytywne, a kiedy negatywne i jaki ma to wpływ na innych,

– z jakich miejsc komentują i dlaczego, jak daleko znajdują się od siebie, jak to wpływa na interakcje między nimi,

– czy i jakie grupy tworzą, jak te grupy się zmieniają w czasie i dlaczego?

Do realizacji postawionych celów zbudowany został system zbierający i przetwarzający dane z serwisu Twitter. Przeprowadzonych zostało wiele eksperymentów mających odpowiedzieć na zadane pytania. Zaprezentowane zostały ich wyniki i wyciągnięte wnioski.

1.2. Struktura i zawartość pracy

Pracę można podzielić trzy części. W pierwszej zaprezentowany jest aktualny przegląd badań dotyczący poruszanych tematów, który opisany został w rozdziale 2. Są to informacje na temat sieci społecznych, analizy sentymentu, geolokacji a także dotyczące Twittera w kontekście aktualnego stanu wiedzy na ich temat i sposobu wykorzystania ich w nauce.

Drugą część stanowią rozdziały 3 oraz 4, gdzie zaprezentowano koncepcję rozwiązania oraz architekturę systemu. W pierwszym z nich przedstawiona jest koncepcja, metody i algorytmy, które zostały zastosowane w procesie przeprowadzania badań. Omówiony tam jest sposób w jaki dane były zbierane i przetwarzane. Opisany jest także algorytm badania wydźwięku wypowiedzi oraz sposoby badania sieci społecznych i geolokacji. W drugim zaś przedstawiona jest zastosowana architektura systemu i wykorzystane narzędzia.

Trzecia część pracy to rozdział 5, w którym zaprezentowane są przeprowadzone eksperymenty. Skupiają się one na pokazaniu zastosowania Twittera w analizie sentymentu, sieci społecznych i geolokalizacji. Nacisk położono na to by pokazać jak połączyć ze sobą te trzy dziedziny.

[1] www.twitter.com

2. Przegląd badań

W niniejszym rozdziale znajduje się aktualny stan badań dotyczący trzech tematów, które składają się na tę pracę. Na początku omówiona jest dziedzina sieci społecznych – czym ta nauka się zajmuje i w jakich przypadkach może zostać zastosowana. Następnie opisana jest analiza sentymentu wypowiedzi i jej zastosowanie. Po niej omówiona jest geolokacja wraz z opisem tego co można dzięki niej się dowiedzieć. Na końcu skupiono się na opisie 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 [1]. 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 [2] 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 hubów[1], czy może liczba połączeń rozłożona jest równomiernie,

– odkrywanie grup/społeczności – jest jednym z głównych tematów analizy sieci społecznych; szukanie grup związane jest z klastrowaniem i odkrywaniem obszarów sieci, które są bardziej zagęszczone (czyli takie, w których stosunek liczby krawędzi do liczby wierzchołków jest większy niż na zewnątrz); problem powiązany jest z badaniem grafów, określaniem jak dzielić sieć na regiony,

– klasyfikacja wierzchołków – polega na opracowaniu metody, dzięki której możliwe jest zaklasyfikowanie wierzchołków do wcześniej zdefiniowanych klas na podstawie podobieństwa z innymi jednostkami do tych klas już należących,

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

– predykcja przyszłych połączeń wewnątrz sieci – wiele badań skupia się na statycznych połączeniach wierzchołków; w dużej liczbie sieci połączenia między węzłami są jednak dynamiczne i badania te koncentrują się na tym by przewidzieć nowe połączenia wewnątrz takich sieci,

[1] hub – jednostka mająca dużą liczbę połączeń, łącząca ze sobą różne części sieci społecznej; wokół niej koncentrują się inne jednostki

– 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 wpływa na poprawę jakości badania danej sieci.

2.1.1. Przykłady zastosowania analizy sieci społecznych

Wyniki badań nad sieciami społecznymi stwarzają wiele możliwości dla różnych dziedzin życia. Mogą być zastosowane między innymi 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 [3],

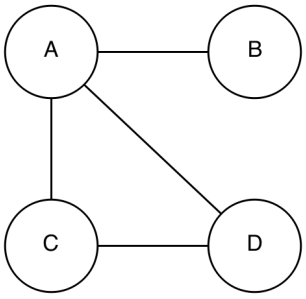
– badania naukowe – odkrywanie naukowców zajmujących się podobnymi tematami celem opracowania bardziej kompletnych wyników lub podjęcia nowego, wspólnego tematu [4],

– przedsiębiorstwa handlowe – odkrywanie zbliżonych typów klientów i oferowanie im produktów lub usług do nabycia przy użyciu systemów rekomendujących [5],

– 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 [6].

2.1.2. Reprezentacja sieci społecznych

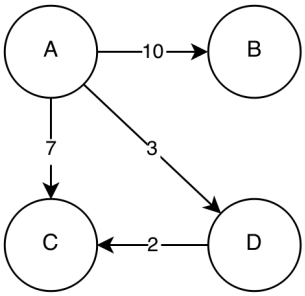
Najczęściej spotykaną reprezentacją sieci społecznych jest reprezentacja grafowa. Grafem nazywamy strukturę G = (V, E) składającą się z węzłów (wierzchołków, oznaczonych przez V) wzajemnie połączonych za pomocą krawędzi (oznaczanych przez E). Przykładowy graf został zaprezentowany na rysunku 2.1.



Rysunek 2.1: Graf o 4 wierzchołkach i 4 krawędziach

Reprezentacja grafowa w naturalny sposób modeluje jednostki jako węzły i relacje między nimi jako krawędzie. W zależności od rodzaju sieci graf taki może posiadać krawędzie skierowane lub nieskierowane oraz ważone lub nieważone. Krawędź skierowana reprezentuje kierunek, w którym przebiega

komunikacja między węzłami. Krawędź, która jest ważona zawiera w sobie dodatkową informację o jakości danego połączenia (reprezentuje na przykład liczbę wymienianych wiadmości między jednostkami czy odległość między nimi). Przykładowy graf skierowany o krawędziach ważonych przedstawia rysunek 2.2.



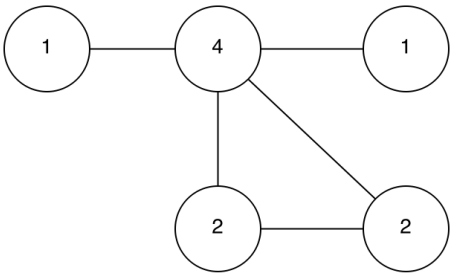
Rysunek 2.2: Graf skierowany o krawędziach ważonych

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 to [7]:

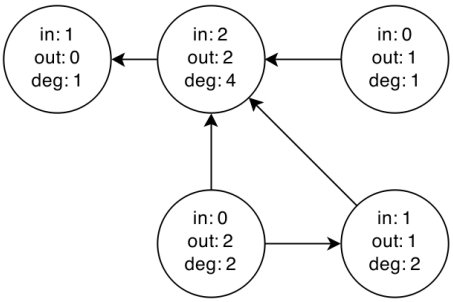
Stopień wierzchołka

Miara określająca liczbę krawędzi wchodzących i wychodzących z wierzchołka (rys. 2.3).



Rysunek 2.3: Graf z oznaczonymi stopniami wierzchołków

W przypadku grafów skierowanych możemy jeszcze mówić o stopniu wchodzącym (ang. *in degree*) oraz wychodzącym (ang. *out degree*) (rys. 2.4).



Rysunek 2.4: Graf skierowany z oznaczonymi stopniami wierzchołków

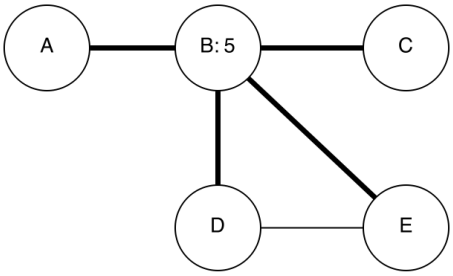
Pośrednictwo (ang. *betweenness*)

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. Wielkość tę wyraża się wzorem:

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

gdzie:
$\rho(i,k,j)$ – liczba najkrótszych ścieżek pomiędzy i oraz j przechodząca przez wierzchołek k,
$\rho(i,j)$ – liczba wszystkich najkrótszych ścieżek pomiędzy i oraz j.

Wartośc *betweenness* dla wierzchołka B (rys. 2.5) oblicza się przy pomocy najkrótszych ścieżek między węzłami innymi niż B, to jest: ABC, ABD, ABE, CBD, CBE, DE. W 5 z 6 z nich znajduje się węzeł B, stąd jego wartość *betweenness* jest równa $5/6$.



Rysunek 2.5: 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. Dlatego też mogą być skutecznym celem ataków.

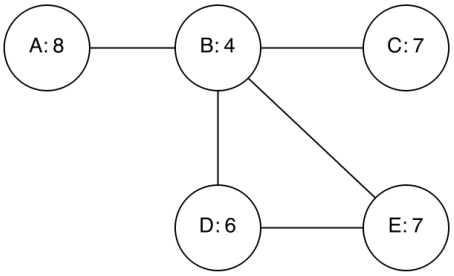
Bliskość (ang. *closeness*)

Znormalizowana odwrotność sumy odległości między węzłami w grafie.

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

gdzie:
$d(k,j)$ – odległość między wierzchołkami k oraz j,
N – liczba wszystkich wierzchołków.

Wartość *closeness* dla pojedynczego wierzchołka liczymy sumując odległości między nim a pozostałymi wierzchołkami, a następnie wartość $N-1$ dzielimy przez wyliczoną przed chwilą sumę. Przykładowy graf (rys. 2.6) i tabela (tab. 2.1) z obliczeniami tej wielkości znajdują się poniżej.



Rysunek 2.6: Suma odległości do pozostałych węzłów w grafie

	Wierzchołki					Suma odległości	Bliskość (*closeness*)
	A	B	C	D	E	$S = \sum_j d(i,j)$	$CC(i) = \frac{N-1}{S} = \frac{4}{S}$
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. Jest to więc wierzchołek najszybciej rozsyłający informacje wewnątrz sieci.

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:

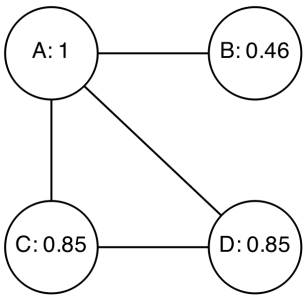
$$EV(k) = \frac{1}{\lambda} \sum_{j} A_{kj} x_j \tag{2.3}$$

gdzie:

λ – stała, równa największej wartości własnej macierzy sąsiedztwa grafu,

$$A_{kj} = \begin{cases} 1, \text{gdy wierzchołki } k \text{ oraz } j \text{ mają wspólną krawędź} \\ 0, \text{w przeciwnym wypadku} \end{cases} \tag{2.4}$$

Przykładowe wartości wielkości *eigenvector* zaprezentowano na rysunku 2.7[2].



Rysunek 2.7: Wartości wielkości *eigenvector* w przykładowym grafie

2.1.4. Wykrywanie grup w sieciach społecznych

Grupa to dwie lub więcej osób, między którymi dłużej niż przez parę chwil dochodzi do interakcji, które wzajemnie na siebie oddziałują oraz spostrzegają się w kategorii „my” [8, s. 268].

Jednym z elementów analizy sieci społecznych jest wyodrębnianie grup spośród wszystkich członków sieci. Użytkownicy w naturalny sposób tworzą między sobą takie grupy. Wykrycie ich daje szansę na rożne podejście do analizy różnych grup i ich zachowań. Ekstrakcja grup z pośród sieci społecznych może wymagać wielu długotrwałych obliczeń. Istnieją jednak prostsze metody, takie jak między innymi:

– losowe przejścia po grafie (ang. *random walks*) badając podobieństwo wierzchołków [9],

– zastosowanie 3 rodzajów metryk i balansowanie ich istotnością [10],

– szybkie wykrywanie społeczności (ang. *fast unfolding of communities in large networks*) [11].

Ostatni algorytm wykorzystano w zaprezentowanych w tej pracy badaniach. Jego implementacja znajduje się w programie Gephi. Algorytm jest heurestyczną metodą opartą na wyliczaniu modularności

[2]Obliczone przy pomocy narzędzia Gephi – https://gephi.github.io/

wierzchołków i jej optymalizacji. Opiera się na dekompozycji dużej sieci na podzbiory społeczności, w których wierzchołki są ze sobą ściśle połączone. Identyfikacja tych społeczności jest niezwykla ważna i prowadzi do odkrywania nieznanych podsieci.

Algorytm jest podzielony na dwie fazy, które powtarzane są iteracyjnie. W pierwszej fazie, każdy wierzchołek przypisany jest do innej społeczności. Następnie dla każdego węzła badani są jego sąsiedzi i oceniana jest jakość modularności poprzez usunięcie aktualnej społeczności węzła i włączenie go do społeczności sąsiada. Węzeł dołącza do tego sąsiada, u którego zysk modularności jest największy. Ten proces jest wielokrotnie powtarzany aż do momentu, w którym nie zauważa się żadnego zysku. Drugi etap polega na budowie nowej sieci, której wierzchołkami są społeczności wykryte w pierwszej fazie. Wagi krawędzi między nowymi wierzchołkami wyliczane są jako suma wag krawędzi między wierzchołkami w dwóch badanych społecznościach. Gdy drugi etap się skończy algorytm ponownie wraca do kroku pierwszego. W każdej iteracji zmniejsza się więc liczba różnych społeczności. Kroki te są powtarzane do momentu, w którym nie dochodzi już do żadnych zysków i maksymalna modularnośc zostaje osiągnięta.

2.2. Sentyment wypowiedzi

Sentyment (inaczej wydźwięk wypowiedzi) to stosunek lub postawa wobec jakiejś sytuacji, zdarzenia. 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 daje możliwość zapoznania się z opiniami innych ludzi czy ekspertów. Możliwość analizy sentymentu wypowiedzi może być bardzo pomocna. Jak wynika z badań przeprowadzonych na ponad 2000 dorosłych Amerykanów [12] 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 [13]:

– 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 tego by nie rekomendować produktów, które otrzymały 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 przez społeczeństwo 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 reakcji 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. Wiąże się bezpośrednio z przetwarzaniem języka naturalnego, przy którym spotkać można między innymi następujące trudności [14, 15]:

– 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 rozwiązań – inaczej trzeba podejść do badania sentymentu w języku polskim a inaczej w angielskim; trzeba pamiętać też, że język naturalny nie jest martwy i ciągle się rozwija,

– 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 na internetowych forach czy serwisach z opiniami również stanowią wyzwanie przy budowie narzędzia do analizy sentymentu.

2.2.1. Zastosowanie sentymentu w mediach społecznościowych

Analiza sentymentu zaczyna mieć szerokie zastosowanie w analizie mediów społecznościowych. Możliwość oceny wydźwięku wypowiedzi zamieszczanych przez internautów w sieci prowadzi do interesujących wyników i na jej podstawie wyciągać można ciekawe wnioski.

W [16] poprzez crawlowanie (z ang. przeszukiwanie stron i selekcjonowanie z nich informacji) blogów związanych z Egiptem w 2011 roku badano nastroje społeczeństwa w trakcie masowych demonstracji w tym kraju. Sentyment pokazywał, że jeszcze przed ich rozpoczęciem nastroje były bardzo negatywne.

W [17] porównany jest wpływ na giełdę artykułów w *Wall Street Jorunal* (nowojorska gazeta o tematyce gospodarczej) z wpisami na *Seeking Alpha* (popularny serwis społecznościowy skierowany do osób grających na giełdzie). Artykuł udowadnia, że wpisy w *Seeking Alpha* mają większy i bardziej trwały wpis na zwracanie akcji niż opinie wyrażane w *Wall Street Journal*.

Z kolei w [18] zostały przeprowadzone badania nad sentymentem wpisów na Twitterze w kontekście wyborów prezydenckich we Francji i Stanach Zjednoczonych. Artykuł pokazuje, że sympatia do kandydatów wyrażana w internecie była wprost propocjonalna do ich wyników sondażowych. W badaniach nad wyborami Francuskimi więcej uwagi skupiał na sobie przegrany Nicolas Sarkozy (mogłoby się wydawać więc, że powinien wygrać), natomiast wielce istotne jest to, że skierowane w jego kierunku było wiele wpisów o wydźwięku negatywnym. Jeśli chodzi o USA to przez cały czas Barack Obama wyprzedzał swojego konkurenta Mitta Romneya zarówno w sondażach jak i w opiniach w sieciach społecznościowych. To badanie również było ciekawe z tego względu, iż kandydaci korzystali w czasie kampanii z Twittera czym wzmagali liczbę rozmów na ich temat.

W opracowaniu [19] została zbadana korelacja sentymentu wpisów na blogach z wpisami na portalach społecznościowych. Teksty zostały podzielone na dwie kategorie: publiczne (takie jak afery, newsy) oraz prywatne (wydarzenia z życia autorów wpisów). Okazało się, że sentyment wyrażany w postach publicznych miał bardzo duży wpływ na to jakie wpisy pojawiały się na Twitterze czy w komentarzach. Co ciekawe, dużo większą uwagę internautów skupiały wpisy o wydźwięku negatywnym. Tymczasem we wpisach prywatnych sentyment odpowiedzi był raczej pozytywny. Gdy wpis miał pozytywny ton, odpowiadający również wyrażali się pozytywnie wspierając wyrażoną opinię. Otrzymane wyniki związane są z tym, że wpisy publiczne gromadzą wokół siebie większą liczbę ludzi, niekoniecznie blisko związaną z danym tematem. Ludzie ci wypowiadają się częściej wtedy, gdy pisze się o jakiejś krzywdzie, która dotyka również ich. We wpisach o charakterze prywatnym komentujący bardzo często w jakiś sposób identyfikują się z autorem, dlatego też w większości przypadków wspierają go i podzielają jego opinie.

2.2.2. Reprezentacja tekstu w przetwarzaniu języka naturalnego

Zanim zacznie się badać sentyment należy w jakiś sposób reprezentować tekst w komputerze. Wymienić można kilka podejść, między innymi:

– *bag of words* [20] – najpopularniejszy sposób reprezentacji tekstu, polega na podzieleniu tekstu na termy (napisy), niezależnie od tego czy powinny występować razem czy nie; nie zachowuje żadnego połączenia między wyrazami w tekście;

  na przykład tekst „`nawet nie powiedział do widzenia`" jest reprezentowany w postaci: {„`nawet`", „`nie`", „`powiedział`", „`do`", „`widzenia`"},

– *bag of concepts* [20] – rozwinięcie idei *bag of words*; polega na reprezentacji tekstu w postaci konceptów, a nie pojedynczych słów; główna zaleta polega na tym, iz zachowuje znaczenie semantyczne i połączenia między wyrazami pojawiającymi się w dokumencie; do odkrywania konceptów konieczna jest podstawowa wiedza semantyczna dostarczona przez słowniki, listy konceptów, i tym podobne,

– reprezentacja wektorowa (ang. *vector space model*) [21] – reprezentuje dokument tekstowy w oparciu o reprezentację wektorową dokumentu; polega na tym, że dowolny dokument reprezentowany jest w postaci wektora częstości występowania słów kluczowych; słowa kluczowe tworzą bazę, a zbiór dokumentów tekstowych można przedstawić w formie macierzy informującej ile razy, który term wystąpił w badanym dokumencie; minusem tego podejścia jest fakt, że zbiór słów kluczowych może być bardzo duży; reprezentacja wektorowa znajduje swoje zastosowanie w szukanie podobieństw między tekstami, próbami kategoryzacji dużych zbiorów tekstów, itp.,

– reprezentacja grafowa [22] – reprezentacja, która rozszerza reprezentację wektorową o to, że interesuje się również kolejnością termów w tekście, zachowując bogatsze informacje na temat przetwarzanego tekstu, które w VSM są zapominane; dzięki temu zachowuje semantyczne znaczenie przechowywanych dokumentów; wyrazy są węzłami grafu, połączone krawędziami jeśli występują wspólnie w tekście; krawędzie mogą być skierowane aby pokazać kolejność słów. Reprezentacja grafowa jest użyta między innymi w mechanizmie Google PageRank [23], wspomagającym wyszukiwarkę i pozycjonowanie w niej stron internetowych.

Ze względu na prostotę i małą długość tekstów do analizy zdecydowano się na reprezentowanie i przetwarzanie tekstu w postaci *bag of words*.

2.2.3. 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).

Techniki badania sentymentu można podzielić na dwie klasy: oparte o słowniki oraz bazujące na klasyfikatorach. Te pierwsze bazują na zbudowanych wcześcniej ręcznie bądź mechanicznie listach słów z określonym sentymentem, a te drugie oceniają tekst używając technik maszynowego uczenia się.

Niektóre ze sposobów oceny sentymentu to [24]:

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:

$$X_t = \frac{P(pos|topic, t)}{P(neg|topic, t)} \tag{2.5}$$

gdzie:

$P(pos|topic, t)$ – prawdopodobieństwo zdarzenia, że słowo t w temacie $topic$ wystąpi z sentymentem pozytywnym,

$P(neg|topic, t)$ – prawdopodobieństwo zdarzenia, że słowo t w temacie $topic$ wystąpi z sentymentem negatywnym.

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 założeniu 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 uczy się klas wyrazów sukcesywnie analizując kolejne teksty [25].

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

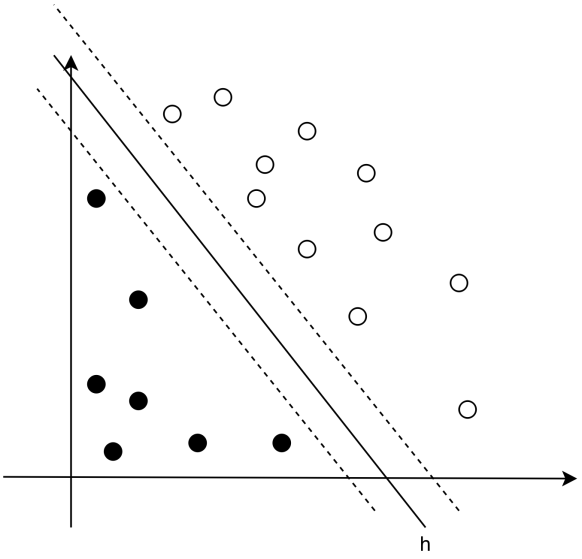
Technika estymacji rozkładu prawdopodobieństwa. Główna zasada polega na tym, że jeśli dane nie są dobrze znane, rozkład powinien być jak najbardziej jednolity, to znaczy mieć maksymalną entropię. Do tej techniki mogą dochodzić ograniczenia, które pozwalają by rozkład nie był maksymalnie jednolity. Ograniczenia takie mogą pochodzić z oznaczonych już danych treningowych 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.

Metoda wektorów nośnych (ang. *support vector machines*)

Metoda wektorów nośnych to metoda klasyfikacji, której główna idea polega na znalezieniu hiperpłaszczyzny (rys. 2.8), która podzieli teksty na pozytywne i negatywne z marginesem pomiędzy klasami tak dużym jak to tylko możliwe.



Rysunek 2.8: Hiperpłaszczyzna h dzieląca zbiór na dwie części

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. Można ją opisać wzorem:

$$\vec{h} = \sum_i \alpha_i C_i \vec{t_j}, \quad \alpha_i \geq 0 \tag{2.6}$$

gdzie:
$\vec{h}$ – szukana hiperpłaszczyzna,
$\vec{t_j}$ – badany tekst (pojedynczy wpis),
$C_i \in \{1, -1\}$ – klasy, do których może trafić wpis (pozytywna/negatywna),
α_i – wartość, która może być znaleziona przez rozwiązanie problemu podwójnej optymalizacji.

Teksty o α_i większym od zera, to te które biorą udział w szukaniu funkcji h i nazywa się je wektorami nośnymi (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, a także z dużą 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 tej metody.

Metoda Alexandra Paka i Patricka Paroubek'a [26]

Technika jest odpowiedzią na problemy związane z brakiem odpowiedniego słownika do oceny sentymentu. Została opracowana z uwzględnieniem Twittera i korzysta założeń z nim związanych. 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 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 obliczana jest jako:

$$valence(T) = \frac{\sum_{i=1}^{n} valence(w_i)}{n} \tag{2.7}$$

gdzie:
T – tekst poddany analizie,
w_i – pojedyncze słowo w tekście T,
$valence(w_i)$ – wartość $valence$ dla słowa w_i,
n – liczba słów w tekście T.

Wartość $valence(w_i)$ obliczana jest przy zastosowaniu skonstruowanego leksykonu za pomocą następującego równania:

$$valence(w_i) = log\frac{N(w_i, M^+) + 1}{N(w_i, M^-) + 1} \tag{2.8}$$

gdzie:
$N(w_i, M^+)$ – liczba wystąpień słowa w_i w zbudowanym leksykonie w kontekście pozytywnym,
$N(w_i, M^-)$ – liczba wystąpień słowa w_i w zbudowanym leksykonie w kontekście negatywnym.

Zastosowanie takiego wzoru prowadzi do tego, że niezależnie jak często dany wyraz pojawia się w zbiorze treningowym, najważniejsza jest jego polaryzacja. Gdy na przykład słowo *świetny* pojawia 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 sposób, technika identyfikacji geograficznego położenia osoby lub urządzenia za pomocą cyfrowych danych przetwarzanych przy pomocy Internetu[3].

Główne sposoby pozyskiwania takich danych to:

– korzystanie z urządzeń GPS – wbudowanych we współczesne telefony komórkowe, tablety, itp.,

[3]Oxford Dictionaries – www.oxforddictionaries.com

– pozycjonowanie względne – ustalanie pozycji na podstawie bazowych stacji telefonii komorków, ruterów WI-FI,

– użycie 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 lokalnych reklam,

– zapobieganie nadużyciom zakupowym – sprawdzenie geolokacji klienta sklepu internetowego i porównanie jej z danymi z karty kredytowej w celu ochrony osób, którym taka karta mogła zostać 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. Prowadzi to do uzupełnienia zebranych danych. Dzięki temu analiza danej społeczności może być wzbogacona. Ciekawym zastosowaniem może być na przykład zbadanie reakcji społeczeństwa w różnych regionach kraju na planowane przez rząd zmiany w prawie.

Główne zalety dzielenia się informacją o geolokacji [27] (z punktu widzenia użytkowników telefonów komórkowych i sieci społecznych) to to udostępnianie jej z ogólnie rozumianą społecznością (56%) oraz ze ze znajomymi (41%). Głównymi problemami są obawy o prywatność (33%) oraz brak korzyści z niej wynikających (26%).

2.3.1. Zastosowanie geolokacji w sieciach społecznościowych

Możliwość odkrywania pozycji osób korzystających z sieci społecznościowych stała się przyczynkiem do tworzenia serwisów społecznościowych opartych właśnie o geolokację, tak zwane *location based social networks*. Niektóre ze sposobów wykorzystania geolokacji w sieciach społecznościowych to:

– rekomendacje nowych miejsc spotkań w mieście [28],

– przewidywanie rodzaju znajomości jaką stworzy między sobą para osób (odkrywanie czy zostaną znajomymi, czy partnerami) [29],

– odkrywanie pokrywających się społeczności przy wykorzystaniu geolokacji [30],

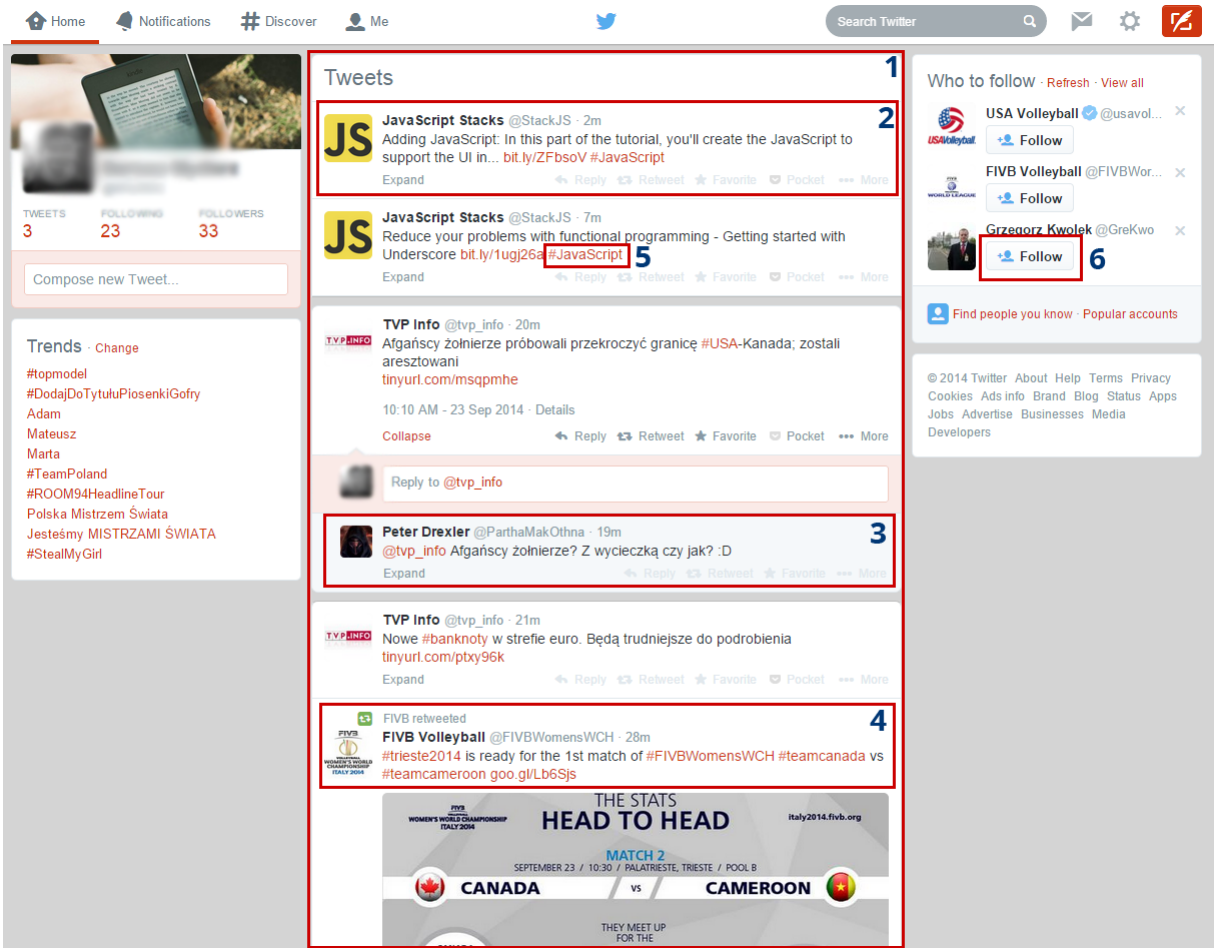
– połączenie podobieństwa fizycznych lokalizacji oraz sieci przyjaciół w celu rekomendacji nowych znajomości między osobami [31],

– odkrywanie preferencji kulinarnych wśród internautów [32].

2.4. Twitter

Twitter[4] to serwis społecznościowy o charakterystyce mikrobloga zorientowany na szybką i bezpośrednią komunikację. 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.

Podstawowe pojęcia związane z tym serwisem to (oznaczone na rysunku 2.9):



Rysunek 2.9: Budowa serwisu Twitter

1. Ściana wpisów (ang. *newsfeed*) – inaczej strona główna użytkownika, na której widzi wszystkie tweety wysłane przez osoby, które śledzi.

2. Wpis (ang. *tweet*) – pojedynczy wpis/post na Twitterze; maksymalna długość to 140 znaków; może dodatkowo zawierać zdjęcie lub informację o geolokalizacji.

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

[4] www.twitter.com

4. Podanie wpisu dalej (ang. *retweet*) – oznacza przekazanie jakiegoś wpisu dalej; jeśli użytkownik A użyje funkcji retweet dla dowolnego wpisu w serwisie, wówczas osoby śledzące użytkownika A również zobaczą ten wpis na swojej stronie głównej.

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

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

2.4.1. Twitter jako źródło danych

Twitter używany jest przez ponad 270 milionów użytkowników wysyłających ponad 500 milionów wpisów dziennie[5] i liczby te szybko rosną. Szybkość komunikacji i łatwość publikacji wpisów sprawia, że staje się medium komunikacyjnym dla wielu grup ludzi. Odgrywał ważną rolę w wydarzeniach społeczno-politycznych takich jak Arabska Wiosna w 2010 czy okupowanie Wall Street w 2012 [33]. Serwis ten jest również bardzo często wykorzystywany do komentowania wydarzeń sportowych. W trakcie mundialu w Brazylii użytkownicy wysłali 672 miliony wpisów z tagiem *#WorldCup* [34].

Popularność Twittera jako źródła informacji doprowadziła do rozwoju badań w różnych dziedzinach. Pomoc humanitarna i w klęskach żywiołowych jest jedną z takich dziedzin, w których informacje z Twittera są używane w celu zapewnienia odpowiedniej pomocy. Naukowcy wykorzystują go by przewidzieć występowanie trzęsień ziemi i śledzić odpowiednich użytkowników, dzięki którym można na bieżąco otrzymywać informacje związane z katastrofą [33].

Sposób dostępu do danych

Dostęp do danych z Twittera możemy uzyskać na dwa sposoby. Pierwszy to Streaming API[6]. Aby z niego korzystać należy napisać program, który nasłuchuje pojawiających się na żywo wpisów. Co ważne ich liczba jest ograniczona do maksymalnie 1% wszystkich wpisów na Twitterze. Streaming API pozwala na zbieranie danych przy użyciu słów kluczowych, określenia języka wpisów, lokalizacji i tym podobnych.

Drugim sposobem zbierania danych z Twittera jest REST API[7]. Służy ono do wykonywania zapytań restowych, za pomocą których możliwe jest pobranie listy wpisów danego użytkownika, listy jego znajomych, listy wpisów z zadanymi słowami kluczowymi i tym podobne. W przeciwieństwie do poprzedniej metody REST API nie nasłuchuje wpisów na żywo, a jedynie zwraca dane dostępne w momencie wysłania żądania. Zapytania do API można wykonywać w 15 minutowych oknach, w których możliwe jest wysłanie maksymalnie 180 zapytań, w wyniku których można uzyskać nie więcej niż 100 wpisów (to daje nam maksymalnie 7200 wpisów na godzinę). Co więcej wyszukiwanie wpisów przez REST API przy użyciu słów kluczowych w rezultacie zwraca jedynie wyniki z ostatnich 6-9 dni. To API bardziej

[5] www.about.twitter.com/company – na dzień 16 lipca 2014 r.
[6] www.dev.twitter.com/streaming/overview
[7] www.dev.twitter.com/rest/public

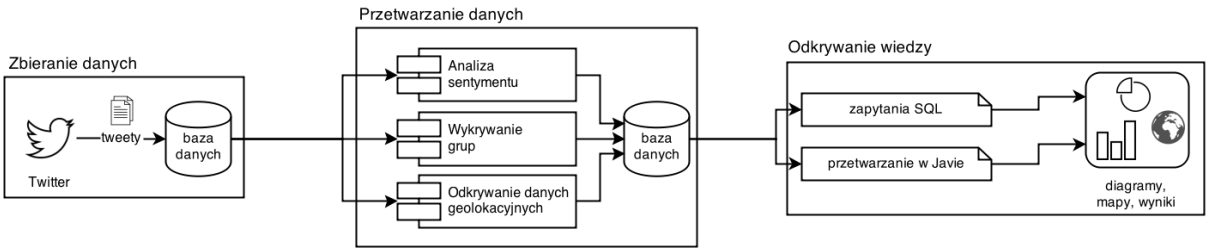
przydaje się więc do tworzenia aplikacji klienckich do Twittera niż do zbierania danych (konsumowania ich na żywo), gdzie lepszym rozwiązaniem jest Streaming API.

3. Koncepcja powiązania analizy sentymentu i geolokacji w badaniu zachowań użytkowników sieci społecznościowych.

W tym rozdziale opisany jest sposób realizacji celów mojej pracy magisterskiej – czyli sposób powiązania ze sobą analizy sentymentu i geolokacji w kontekście badania zachowań użytkowników sieci społecznościowych. W rozdziale 3.1 zaprezentowany został model systemu z uwzględnieniem przetwarzania wstępnego (r. 3.1.1), modelu analizy sentymentu (r. 3.1.2), modelu grup społecznych (r. 3.1.3) i modelu geolokacji (r. 3.1.4). Następnie przedstawione są miary i algorytmy wykorzystywane w eskperymentach (r. 3.2), a na końcu szczegółowo została opisana koncepcja algorytmu analizy sentymentu (r. 3.3).

3.1. Model systemu

Ogólny model systemu został przedstawiony na rysunku 3.1.



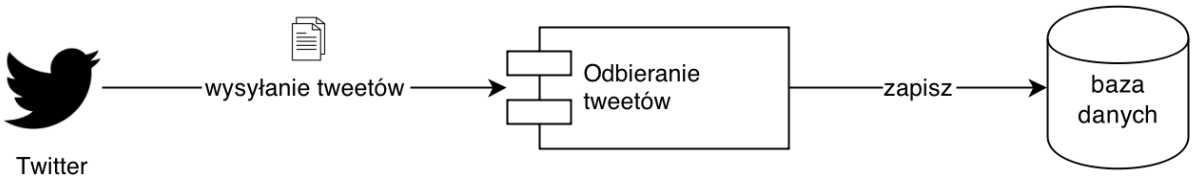
Rysunek 3.1: Ogólny model systemu

Można go podzielić na trzy następujące części:

1. Zbieranie danych.
2. Przetwarzanie i uzupełnianie danych.
3. Zdobywanie wiedzy z zebranych i przetworzonych danych.

3.1.1. Zbieranie danych

Zbieranie danych jest pierwszym elementem składowym całego systemu. Jego koncepcje przedstawia rysunek 3.2. Jest to bardzo ważny składnik systemu. Aby móc przeprowadzić jakiekolwiek analizy potrzeba najpierw zebrać dane.



Rysunek 3.2: Zbieranie danych

Zbieranie danych z serwisu Twitter odbyło się przy pomocy wspomnianego w rozdziale 2.4.1 Twitter Streaming API. Napisany został program w Javie, który nasłuchiwał w trakcie każdego śledzonego meczu wpisów w języku angielskim, które zawierały odpowiednie słowa kluczowe. Były nimi przygotowane wcześniej słowa związane z danym spotkaniem: nazwiska i przydomki piłkarzy, menadżerów, nazwy i przydomki klubów, nazwa stadionu czy nazwisko sędziego. W taki sposób uzyskiwano tweety, które można było powiązać z konkretnym meczem.

Program do nasłuchiwania uruchamiany był na czas trwania meczu z pewnym marginesem przed i po spotkaniu. Jest to o tyle ważne, że udostępnione API służy tylko do konsumowania wpisów na żywo. Pobranie tych samych wpisów już po danym wydarzeniu byłoby dużo bardziej skomplikowane. Program otrzymuje nie więcej niż 1% wszystkich wpisów w danym momencie na Twitterze [1]. Każdemu wpisowi nadawany jest identyfikator meczu (wprowadzonego wcześniej do bazy danych) i w takiej postaci jest on zapisywany do bazy.

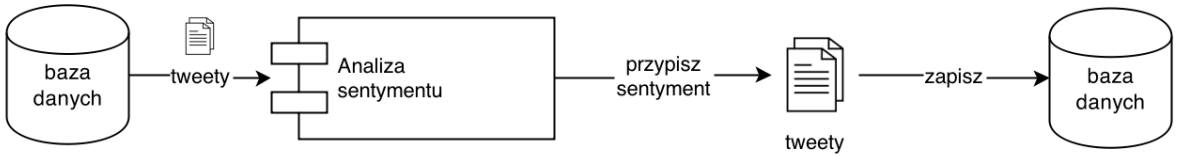
Takie podejście pozwala na elastyczną wymianę słów kluczowych do nasłuchiwania, gdyż program na wejściu potrzebuje jedynie identyfikator meczu, do którego już są przypisane słowa kluczowe.

3.1.2. Model analizy sentymentu

Analiza sentymentu należy do jednego z elementów drugiej części systemu. W tej części zebrane dane były przetwarzane pod kątem oceny ich sentymentu. Proces ten przedstawiony jest na schemacie (rys. 3.3).

Dla każego tweeta w bazie obliczana była jego wartość *valence* (r. 2.2.3). Aby do tego doszło tweet był wcześniej oczyszczany z elementów, które mogły zakłócić wyliczenie tej wartości. Proces przebiegał zgodnie z algorytmem Paka i Paroubek'a (r. 2.2.3), a szczegóły jego działania wraz z zastosowaną modyfikacją (polegającą na dodaniu wykrywania negacji) zostały opisane w rozdziale 3.3.

[1] Jeśli w danym momencie na Twitterze jest 1000 wpisów, z czego 10 z podanymi słowami kluczowymi, wówczas program otrzyma wszystkie 10 wpisów. Jeśli z określonymi słowami kluczowymi byłoby np. 100 z 1000 wpisów, wówczas program nadal otrzymałby tylko 10 z nich.



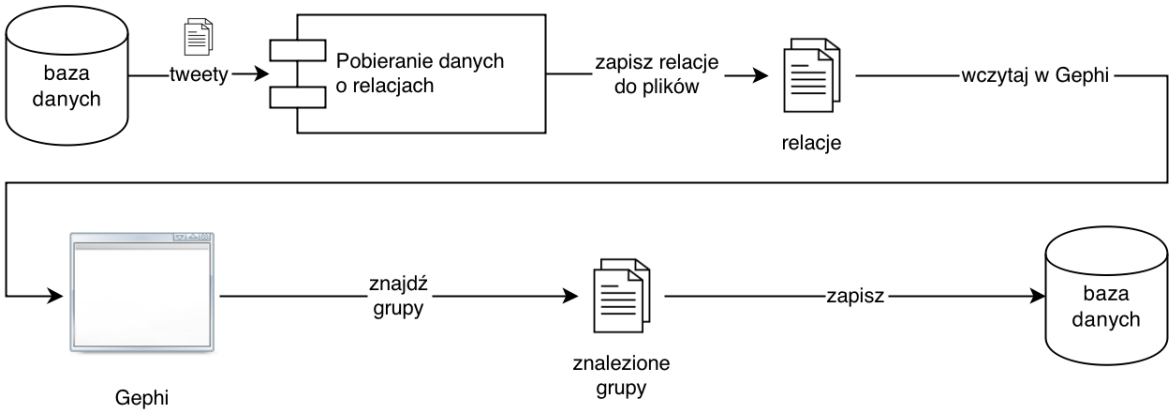
Rysunek 3.3: Model analizy sentymentu

Po wyliczeniu wartości *valence* dla zebranych tweetów obliczona została jego średnia wartość i wpisy o wartości wyższej od średniej zostały oznaczone jako pozytywne, zaś o wartości niższej jako negatywne.

Oznaczenie wartości sentymentu zebranych wpisów pozwoliło na przeprowadzenie ciekawych eksperymentów opisanych w rozdziale 5.

3.1.3. Model grup i relacji społecznych

Drugim ważnym elementem przetwarzania zebranych danych było skupienie się na odkrywaniu relacji między użytkownikami Twittera i odnajdywaniu grup, które wspólnie tworzą. Pomaga to w lepszym zrozumieniu zachowań użytkowników i możliwości osobnej analizy różnych grup. Schemat odkrywania zachowań społecznych został przedstawiony na rysunku 3.4.



Rysunek 3.4: Model grup i relacji społecznych

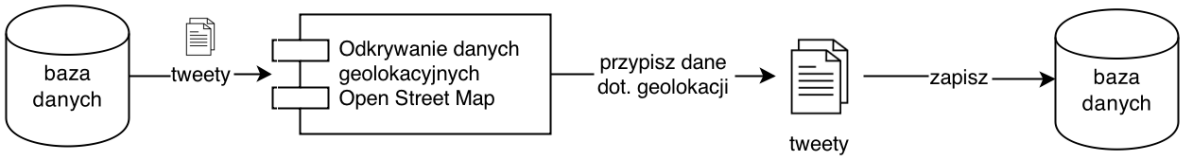
Sieć użytkowników zamodelowana jest w postaci grafu, gdzie użytkownicy to wierzchołki, a wpisy będące retweetami lub odpowiedziami są krawędziami (skierowanymi) tego grafu. Relacje te są przechowywane bezpośrednio w każdym z tweetów w polach `in_reply_to_user_id` oraz `retweeted_user_id`. Do konkretnych eksperymentów dane te były odpowiednio agregowane (np. zliczano liczbę powiązań między wierzchołkami i traktowano je jako pojedynczą relację w jakiejś jednostce czasu – na przykład w pojedynczym meczu).

Odkrywanie grup przeprowadzono przy pomocy algorytmu szybkiego wykrywania grup w dużych

sieciach (ang. *fast unfolding of communities in large networks*) (r. 2.1.4). Wyniki klasyfikacji wierzchołków do odpowiednich grup zostały następnie zapisane w bazie danych.

3.1.4. Model geolokacji

Ostatnim elementem przetwarzania zebranych danych było przetwarzanie związane z danymi geolokacyjnymi. Przechowywane w bazie tweety posiadają jedynie współrzędne określające miejsce, z którego zostały wysłane (oczywiście niewielki odsetek tweetów z całego zbioru). Schemat przetwarzania tych danych prezentuje rysunek 3.5.



Rysunek 3.5: Model geolokacji

Przy pomocy serwisu *Open Street Map*[2] i skorzystaniu z API służącego do georeversingu[3] dla każdego tweeta posiadającego współrzędne pobrano informacje szczegółowe na temat miejsca, na które te współrzędne wskazują. W ten sposób uzyskano takie dane jak: kraj, stan/województwo (ang. *state*), powiat/hrabstwo (ang. *county*) oraz miasto (ang. *city*).

Dodatkowo zaprezentowano w pracy mapy cieplne uwidaczniające częstość wpisów z różnych lokalizacji na całym świecie. Do tego zadania skorzystano z usługi *CartoDB*[4], za pomocą której dostarczając dane z geolokacją można otrzymać takiego rodzaju mapy.

3.2. Zastosowane algorytmy i miary

W tym rozdziale zaprezentowane są algorytmy i miary, które zostały zastosowane w zaprezentowanych w rozdziale 5 eksperymentach. Przedstawione są one z podziałem na moduł, którego dotyczą.

3.2.1. Analiza sentymentu

Użyte w eksperymentach użyte miary i techniki związane z analizą sentymentu.

Ocena pozytywności wpisów

Miara pozytywności wpisów określa ich sumaryczny wydźwięk. Jest to stosunek liczby wpisów pozytywnych do sumy liczby wpisów pozytywnych i negatywnych. Gdy wartość ta jest większa niż

[2] www.openstreetmap.com

[3] georeversing – odkrywanie informacji o miejscu na Ziemi (kontynent, kraj, miasto, itp.) przy pomocy jedynie współrzędnych geograficznych

[4] www.cartodb.com

50% można mówić o wydźwięku pozytywnym, a gdy mniejsza – o negatywnym. Wartość tę określa się poniższym wzorem:

$$P = \frac{|pos|}{|pos| + |neg|} \tag{3.1}$$

gdzie:

$|pos|$ – liczba wpisów oznaczonych jako pozytywne,

$|neg|$ – liczba wpisów oznaczonych jako negatywne.

Wykrywanie zwolenników i przeciwników klubu

Dzięki oznaczeniu wszystkich zebranych wpisów odpowiednim sentymentem (pozytywnym lub negatywnym) możliwe były wykorzystanie tych informacji do odkrywania nowej wiedzy. Sentyment umożliwił odkrywanie wśród wszystkich autorów wpisów ich preferencji klubowych – poprzez zliczanie liczby wpisów pozytywnych i negatywnych w meczach danego zespołu.

Proces oznaczania użytkowników jako sympatyków i przeciwników drużyny przebiegał w następujący sposób:

– dla każdego użytkownika zliczono jego wpisy w meczach każdej z badanych drużyn,

– jeśli w meczach drużyny A dla danego użytkownika przeważała liczba wpisów pozytywnych, wówczas oznaczano takiego kibica jako zwolennika drużyny A,

– jeśli natomiast przeważała liczba wpisów negatywnych, wówczas taki kibic traktowany był jako przeciwnik danej drużyny.

Zastosowanie tej techniki pozwala na przeprowadzenie eksperymentów z podziałem na zwolenników i przeciwników danego klubu i zbadanie ich zachowań w zależności od sympatii.

3.2.2. Relacje między użytkownikami i grupy

W tym rozdziale przedstawione zostały miary i techniki użyte podczas eksperymentów związanych z odkrywaniem relacji i grup między użytkownikami.

Wykrywanie grup

Do wykrywania grup wśród sieci społecznych zastosowano model opisany w rozdziale 3.1.3. Znalezione w ten sposób grupy użytkowników poddane zostały analizie – z podziałem na kolejne mecze. Grupy te zostały zagregowane w trzy przedziały: od 3 do 4 osób, od 5 do 9 osób oraz więcej niż 9 osób. Na takich zagregowanych grupach został przeprowadzony eksperyment ich podobieństwa między kolejnymi meczami.

Badanie podobieństwa grup

Tak jak wspomniano wyżej badanie podobieństwa oparte było o analizę kolejnych wydarzeń (meczów). Polegało na zliczeniu ile wierzchołków powtarza się w kolejnych spotkaniach. Oparte zostało o

poniższy wzór:

$$S = \frac{|V_1 \cap V_2|}{|V_1|} \tag{3.2}$$

gdzie:

V_1 – zbiór wierzchołków w pierwszym wydarzeniu,

V_2 – zbiór wierzchołków w następnym wydarzeniu.

Krótko mówiąc podobieństwo to iloraz liczby wspólnych wierzchołków między wydarzeniami a liczby wierzchołków w pierwszym z nich.

3.2.3. Geolokacja

W niniejszym rozdziale opisuję wielkości, które zostały zastosowane podczas przeprowadzania eksperymentów, w których najważniejszym elementem było położenie fizyczne, czyli współrzędne przypisane do wpisu.

Zmierzenie odległości między użytkownikami

Badanie odległości między użytkownikami odbyło się na podstawie odpowiedzi z geolokacją – czyli tweetów zawierających dane o współrzędnych miejsca ich wysłania oraz będących odpowiedziami (mającymi niepuste pole `in_reply_to_user_id`). Wpisy typu retweet nie zawierają informacji o położeniu użytkownika (taka informacja nie jest rejestrowana przez Twitter).

Zmierzenie odległości między użytkownikami odbyło się przy pomocy rozszerzenia PostGIS[5] do bazy PostgreSQL[6], z pomocą którego można obliczyć odległość w metrach między dwoma punktami geograficznymi.

Cały proces przebiegał według następującego schematu:

– pobierz wszystkie wpisy będące odpowiedziami (*reply*) i zawierające współrzędne geograficzne,

– pogrupuj użytkowników w pary według tego, kto z kim się komunikował,

– wylicz średnie położenie każdego z użytkowników dla każdej konwersacji (średnie położenie użytkownika zostało wyliczone jako średnia arytmetyczna długości i szerokości geograficznych jego wpisów w danej konwersacji),

– wylicz przy pomocy PostGIS-a odległość między nimi.

Zmierzenie odległości między użytkownikami zostało użyte w eksperymencie opisującym zależność odległości od częstości komunikacji (r. 5.5.1).

Badanie odległości kibiców od stadionu

Do każdego nasłuchiwanego meczu zostały przypisane współrzędne miejsca, w którym się odbywał – czyli współrzędne stadionu gospodarza danego spotkania. Badanie odległości kibiców od stadionu zostało przeprowadzone oddzielnie dla każdego meczu. Aby je przeprowadzić zastosowano poniższy schemat działania:

[5] www.postgis.net

[6] www.postgresql.org

– pobierz wszystkie wpisy posiadające współrzędne geograficzne (a więc nie tylko te będące odpowiedziami, jak w powyższym badaniu),

– wylicz średnie położenie każdego użytkownika z podziałem na spotkania,

– wylicz przy pomocy PostGIS-a odległość użytkownika od stadionu.

Badanie zostało przeprowadzone z podziałem na zwolenników i przeciwników i pokazuje zależność odległości kibiców od miejsca rozegrania meczu w zależności czy ich ulubiona drużyna gra mecz u siebie czy na wyjeździe.

3.3. Koncepcja algorytmu analizy sentymentu

W tym miejscu opisano dokładnie dokładnie koncepcję wyznaczania wydźwięku wypowiedzi i zastosowany algorytm do analizy sentymentu. Przedstawione są czynności wstępne, wybrany algorytm i zastosowane w nim modyfikacje oraz sposób aplikacji algorytmu na przykładowych tweetach.

3.3.1. Przygotowanie tekstu

W związku z tym, że wpisy są tworzone przez zwykłych użytkowników posiadają one wiele znaków i elementów, które z punktu widzenia analizy sentymentu są zbędne, a czasami prowadzące do błędów. Dlatego też tekst należy poddać oczyszczeniu, usunięciu zbędnych elementów, szumów i spamu. Przykładowe wpisy przed i po normalizacji przedstawia tabela 3.1.

Kolejne kroki, które przekształciły tweety do takiej postaci to:

1. Usunięcie skomentowanych retweetów.
   `Stay woke brah! The Arsenal is about to make everything alright soon :) ~~RT @JCphoenixx: So damn tired, So not sleepy.~~`

2. Usunięcie skomentowanych cytowań.
   `At all...~~"@dotun_somoye: Even city's first goal negredo was offside....:( the refs not helping at all"~~`

3. Usunięcie hiperlinków.
   `You up for Arsenal's match later on? - what time? maybe if i'm not busy baby sitting :) ~~http://t.co/aC5Ec8ipy1~~`

4. Usunięcie nazw użytkowników.
   `~~@abdul1haseeb~~ My arsenal is not disappointing too :P`

5. Usunięcie hashtagów.
   `Haha, you gotta agree, no one gets booed like Manchester United :D ~~#ZeDevilza~~`

	Przed normalizacją	Po normalizacji
1	RT @J_SPEKZ: Haha quality! #Fellaini #United #Moyes http://t.co/rJB4K1fvZy	–
2	Stay woke brah! The Arsenal is about to make everything alright soon :) RT @JCphoenixx: So damn tired, So not sleepy.	stay woke brah make alright
3	@abdul1haseeb My arsenal is not disappointing too :P	not_disappointing
4	@Arsenal didn't think i could respect @aaronramsey any more than i already did, bute what a gentleman he is for not to celebrate that goal:)	didnt not_respect bute gentleman not_celebrate not_goal
5	OHHHHHH!!!!! SO CLOSE!!! Wilshere!!! Good Job Ramsey keeping that move alive	ohhhhhh close good job keeping alive
6	Haha, you gotta agree, no one gets booed like Manchester United :D #ZeDevilza	haha gotta agree not_booed

Tablica 3.1: Wpisy przed i po normalizacji

6. Oznaczenie wyrazów zaprzeczonych przedrostkiem `NOT_` (opisuję dokładniej w rozdziale 3.3.2). `didn't` **`NOT_think NOT_i NOT_could NOT_respect NOT_any NOT_more NOT_than NOT_i NOT_already NOT_did,`** `bute what a gentleman he is for not` **`NOT_to NOT_celebrate NOT_that NOT_goal:)`**

7. Zachowanie tylko znaków alfabetu:

   – usunięcie zaimków dzierżawczych (`Helen's` → `Helen`),

   – usunięcie apostrofu ze skróconych zaprzeczeń (`don't` → `dont`),

   – normalizacja liter diakrytyzowanych (`José Mourinho` → `Jose Mourinho`)

   – usunięcie liczb i wszelkich znaków niealfabetycznych.

   `You up for Arsenal~~'s~~ match later on~~?~~ – what time~~?~~ maybe if i~~'~~m not busy baby sitting ~~:)~~`

8. Usunięcie wyrazów zdefiniowanych w stop liście (powszechne wyrazy danego języka, które mogą być pominięte nie tracąc jednocześnie żadnej informacji). Zastosowano stop listę z serwisu WebPageAnalyse.com [35] zawierającą 528 słów.

   ~~`You up for`~~ `Arsenal match` ~~`later on what`~~ `time` ~~`maybe if`~~ `im` ~~`not`~~ `busy baby sitting`

9. Usunięcie słów kluczowych, które użyte były do gromadzenia wpisów z Twittera – czyli nazwisk piłkarzy, menadżerów, nazw klubów, itd.

```
OHHHHHH CLOSE ~~Wilshere~~ Good Job ~~Ramsey~~ keeping alive
```

3.3.2. Zastosowanie algorytmu

W celu przeprowadzenia analizy sentymentu wykorzystano algorytm Paka i Paroubek'a 2.2.3. Pierwszym krokiem było zbudowanie słownika sentymentu ze zgromadzonego zbioru danych. Aby to osiągnąć należy skorzystać z wpisów z emotikonami wskazującymi wydźwięk.

Według artykułu [36] 20 różnych emotikon używane jest w 90% wpisów z jakimikolwiek emotikonami. Dlatego też użyto tego zbioru do badań dzieląc emotikony na wyrażające wydźwięk pozytywny i negatywny w sposób[7] przedstawiony w tabeli 3.2.

Wydźwięk	Najpopularniejsze emotikony
Pozytywny	:) :D ;) :-) :P =) (: ;-) XD =D :O =] ;D :]
Negatywny	:(:/ :-(=/ =(

Tablica 3.2: Wydźwięk emotikon

Następnie przeglądnięto wszystkie tweety z emotikonami zliczając liczbę występowania wyrazów w kontekście pozytywnym i negatywnym. Najpierw badano sentyment całego wpisu (na podstawie emotikony – gdy była ich większa ilość wybierano ten sentyment, który przeważał) a następnie dla każdego wyrazu z tego wpisu zwiększano licznik odpowiednio wystąpień pozytywnych lub negatywnych. W ten sposób uzyskano słownik sentymentu zbudowany z zebranych danych, który zawierał 34183 słowa, a najpopularniejsze z nich zaprezentowane są w tabeli 3.3, gdzie:

wartość *valence* wyliczna jest według wzoru 2.7 i używana jest w dalszych obliczeniach,

wartość pozytywności wyliczona jest według wzoru 3.1.

Słowo	Wyst. pozytywne	Wyst. negatywne	Pozytywność	Valence
win	4206	598	87.6 %	0.845
good	3916	435	90.0 %	0.903
game	3016	1012	74.9 %	0.301
goal	2305	584	79.8 %	0.477
today	2019	526	79.3 %	0.477
time	1844	404	82.0 %	0.602
dont	1461	775	65.3 %	0.000
match	1669	449	78.8 %	0.477
great	1885	160	92.2 %	1.041
love	1837	202	90.1 %	0.954

Tablica 3.3: Liczba występowania najpopularniejszych słów w zbudowanych słowniku sentymentu

[7] emotikona `D:` została pominięta, gdyż pokrywała więcej przypadków niż tylko użycie emotikony (np.: `Accepte`**`d:`** `Mary, John, Jane`)

Dla każdego wpisu wyliczana jest średnia arytmetyczna wartości *valence* wszystkich słów. W ten sposób wylicza sie wartość *valence* danego wpisu. W tabeli 3.4 prezentuję przykładowe wyniki wyliczania tej wielkości.

Valence	Wpis	Składowe
0.6261 POS	Just seen you in the crowd at Chelsea game! @domashman http://t.co/NjovoZdgZ3	game=0.4740, crowd=0.7782
0.5909 POS	RT @BTSP: #PRIZEDRAW If Arsenal win tonight one lucky person will win a personalised BTSP mug! Simply RT & follow to enter! http://t.co/9m...	lucky=0.4613, tonight=0.6873, btsp=0.3010, person=0.4232, personalised=0.3010, enter=0.5351, follow=1.2229, win=0.8465, mug=0.4771, simply=0.3979
0.6199 POS	Liking the brightness of this Napoli kit #tempted	liking=0.8062, kit=0.4337
0.0305 NEG	@ricktaylor1987 You're not watching Arsenal?	not_watching=0.0305
0.1234 NEG	Poor start#AFC	poor=-0.2848, start=0.5317
0.1807 NEG	Do Arsenal have any players who don't fall down with ease? #SwimmingTeam	players=0.3684, not_fall=-0.3979, not_ease=0.4771, dont=0.2751

Tablica 3.4: Wynik działania algorytmu analizy sentymentu na przykładowych wpisach

Określenie sentymentu wpisu odbywa się zgodnie z równaniem:

$$S(t) = \begin{cases} POS & valence(t) > AVG\_VALENCE \\ NEG & valence(t) <= AVG\_VALENCE \end{cases} \tag{3.3}$$

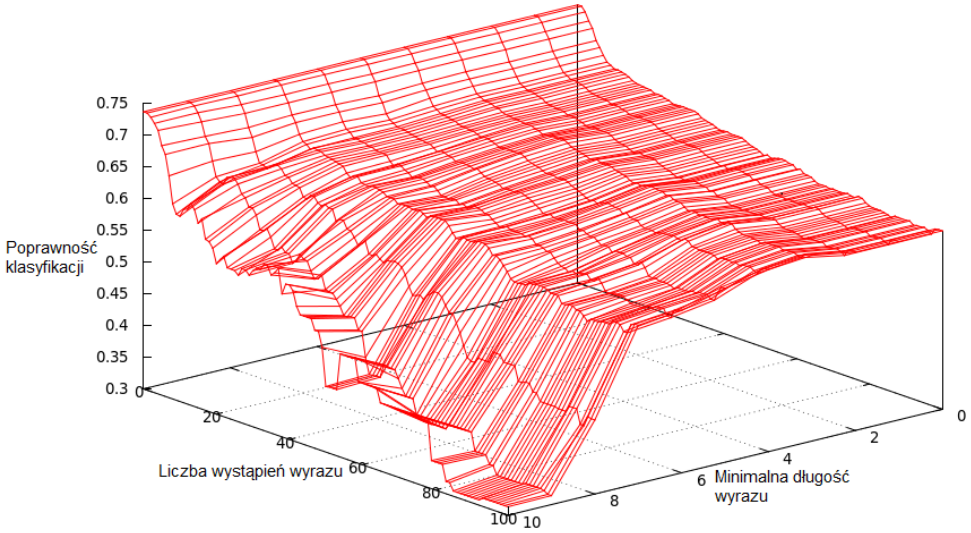
gdzie:
$AVG\_VALENCE$ jest średnią arytmetyczną wartości $valence$ wszystkich wpisów.

3.3.3. Wykrywanie i obsługa negacji

W oryginalnym algorytmie (r. 2.2.3) nie ma żadnego sposobu na wykrywanie i obsługę negacji. Oczywistym jednak jest, iż zaprzeczenia zmieniają znaczenie dalszej części tekstu i muszą być w jakiś sposób obsłużone. Wpisy na Twitterze są krótkie, więc zastosowano podejście zaprezentowane w artykule [37], które polega na dodaniu przedrostka `NOT_` do wszystkich słów pomiędzy wyrazem negującym a najbliższym znakiem przestankowym. Lista wyrazów negujących została zaczerpnięta z [38]. Słowa zanegowane miały osobno liczone liczby wystąpień w kontekstach pozytwnych i negatywnych. Dzięki temu uzyskane wyniki są bardziej dokładne, gdyż wyrazy zaprzeczeniami i bez są traktowane osobno, mając w ten sposób różny wpływ na ogólny wydźwięk wypowiedzi.

3.3.4. Dobór parametrów algorytmu analizy sentymentu

Użyty algorytm (r. 2.2.3) zakłada dobór parametrów przed przeprowadzeniem analizy sentymentu. Parametry te dotyczą słów ze zbudowanego słownika – decydując, które z nich wezmą udział w procesie analizy. Są to: minimalna długość słowa, minimalna liczba występowania słowa. Wykorzystano listę wpisów z emotikonami i podzielono je na zbiór uczący oraz testowy w stosunku 20% do 80%. Wygenerowano słownik ze zbioru uczącego. Wpisy ze zbioru testowego oznaczono spodziewanym sentymentem – był to sentyment emotikony jaką zawierały. Następnie przeprowadzono testy poprawności klasyifkacji zbioru testowego słowami ze zbioru uczącego dla wyrazów o minimalnej długości od 1 do 10 i minimalnej liczbie wystąpień od 1 do 100 – daje to w sumie 1000 testów.



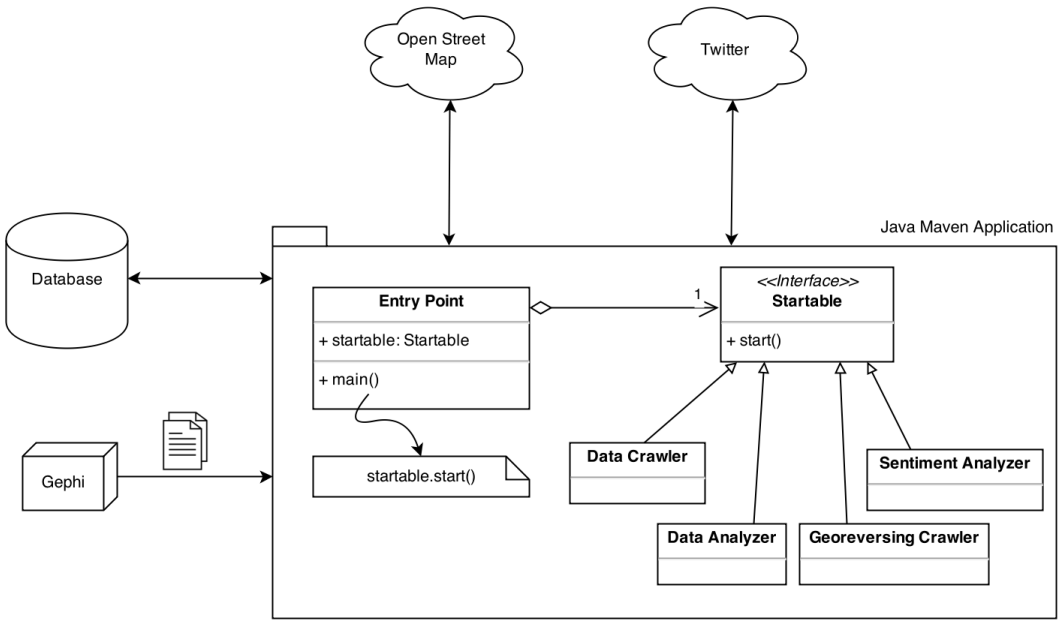
Rysunek 3.6: Dobór parametrów algorytmu analizy sentymentu

Jak widać na wykresie (rys. 3.6) najwyższy stopień poprawności klasyfikacji (oś pionowa) uzyskano dla parametrów równych: minimalna długość wyrazu – 3, minimalna częstotliwość wystąpień – 1. Wyniki poprawności klasyfikacji wyraźnie spadają, gdy minimalna długość wyrazu przekracza 6 znaków.

4. Stworzona architektura i zastosowane technologie

W tym rozdziale opisana jest architektura zbudowanego systemu, wykorzystane technologie i użyte narzędzia. W rozdziale 4.1 przedstawiono architekturę całego systemu, w 4.2 omówione są technologie, a w ostatniej części 4.3 wymienione są użyte biblioteki i narzędzia.

4.1. Architektura systemu



Rysunek 4.1: Architektura systemu

System została zbudowany jako pojedyncza aplikacji Javowa, w której zostały zaimplementowane wszystkie potrzebne mechanizmy. Aplikacja ta komunikuje się z pojedynczą bazą danych, w której zawarte są wszystkie rekordy potrzebne zarówno do zbierania danych jak i te, które są wynikiem analiz. Dodatkowo odpowiada ona za komunikację z usługami w chmurze – czyli zbieraniem danych z Twittera i *georeversingiem* lokalizacji (r. 3.1.4) z Open Street Map. Wszystkie analizy zebranych danych również zostały przeprowadzone przy użyciu aplikacji Javowej.

Oprócz niej wykorzystano także program Gephi, którego wyniki działania zapisano do plików, a następnie przy użyciu wyżej wymienionej aplikacji przeparsowano i umieszczono w bazie danych.

Wewnątrz aplikacji zaimplementowano między innymi moduł do zbierania danych z Twittera (na podstawie podanych słów kluczowych), moduł związany z przeprowadzeniem całego procesu analizy sentymentu – od budowy słownika do zanalizowania pojedynczych wpisów, moduł związany z analizą zebranych danych, a także kod odpowiedzialny za operacje związane z geolokacją.

4.2. Zastosowane technologie

Główną technologią użytą podczas prac był język Java. Oprócz niego wymienić należy także:

– baza danych – PostgreSQL[1],

– budowanie aplikacji – Apache Maven[2],

– system kontroli wersji – Git[3] z repozytorium na GitHub[4],

– wizualizacja wpisów na mapie – CartoDB[5], narzędzie webowe pozwalające na wyświetlanie danych posiadających współrzędne geograficzne na mapie świata.

4.3. Wykorzystane biblioteki i narzędzia

Najważniejsze wykorzystane biblioteki:

– Twitter4J[6] – biblioteka Javowa ułatwiająca korzystanie z Twitter API, użyta do zbierania danych z Twittera,

– Hibernate[7] – framework ORM (Object Relational Mapping – mapowanie obiektowo-relacyjne) do komunikacji z bazą danych,

– PostGIS[8] – rozszerzenie do bazy PostgreSQL dodające funkcje geograficzne, przy pomocy którego możliwe jest między innymi określenie odległości między dwoma punktami (znając ich współrzędne)

– Google Guice[9], JBoss Weld[10] – biblioteki pozwalające zastosować wstrzykiwanie zależności w desktopowej aplikacji Javowej,

[1] www.postgresql.org
[2] www.maven.apache.org
[3] www.git-scm.com
[4] www.github.com
[5] www.cartodb
[6] www.twitter4j.org
[7] www.hibernate.org
[8] www.postgis.net
[9] www.github.com/google/guice
[10] www.weld.cdi-spec.org

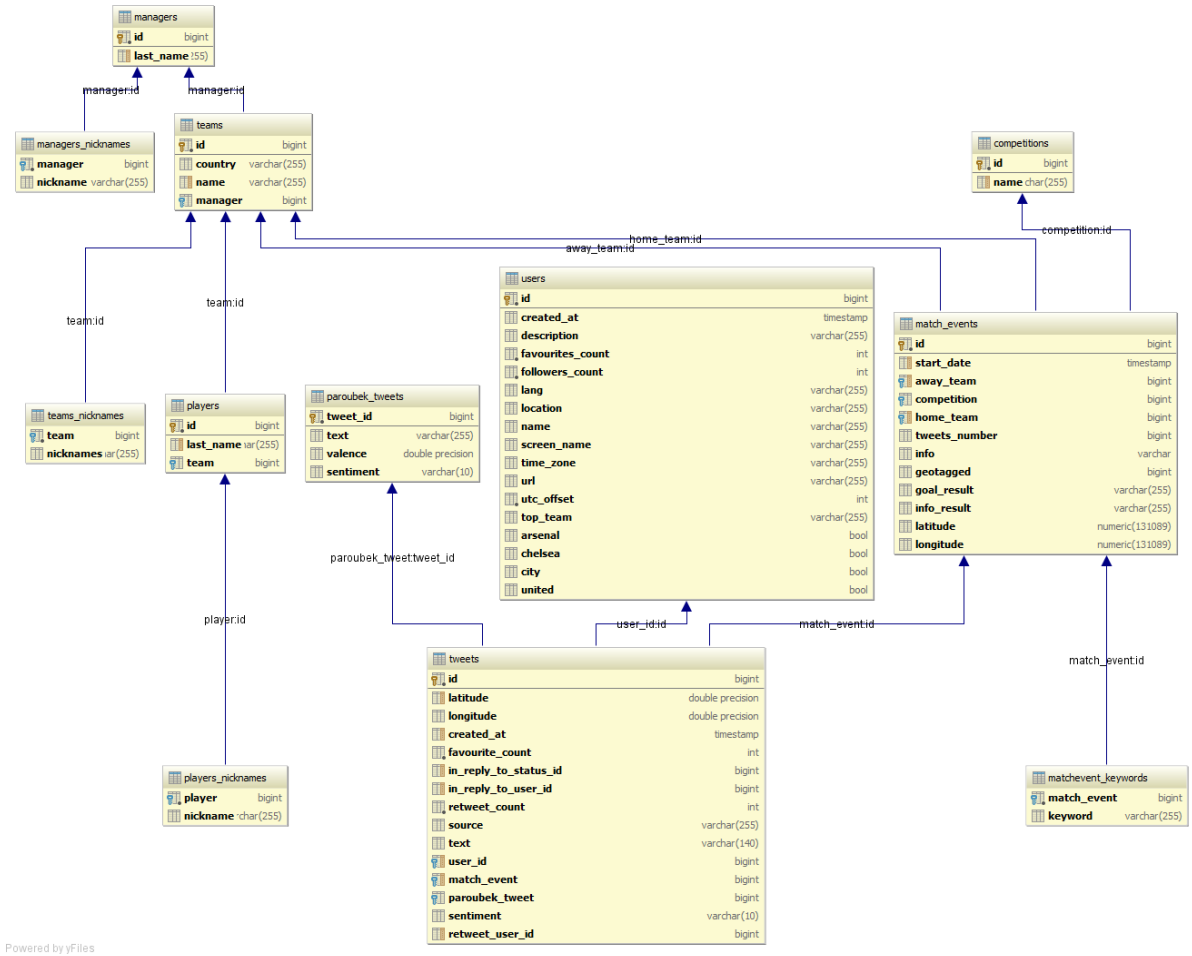
– Google Guava[11], Apache Commons[12], Apache Log4J[13], Joda Time[14] – biblioteki usprawniające programowanie w Javie,

– JUnit[15] – framework do pisania i uruchamiana testów automatycznych.

4.4. Baza danych

Jak zostało już wspomniane użyty został system zarządzania bazą danych PostgreSQL. Na rysunku 4.2 zaprezentowany jest schemat bazy danych. W przedstawionych tabelach zostały zapisane wszystkie informacje potrzebne do pobierania danych z nasłuchiwanych meczów.

Opis najważniejszych tabel został załączony na końcu pracy.



Rysunek 4.2: Schemat bazy danych

[11] www.code.google.com/p/guava-libraries
[12] www.commons.apache.org
[13] www.logging.apache.org/log4j
[14] www.joda.org/joda-time
[15] www.junit.org

5. Opis przeprowadzonych eksperymentów

W niniejszym rozdziale przedstawione są eksperymenty, które zostały przeprowadzone na zgromadzonych danych. Opisany jest sposób ich wykonania, wyniki i wnioski. Na początku w rozdziale 5.1 opisana jest charakterystyka zebranych danych, a następnie opisane są eksperymenty związane z analizą sentymentu (r. 5.3), analizą społeczną (r. 5.4) i analizą geolokacji (r. 5.5).

5.1. Opis zebranych danych

Pomiędzy październikiem a grudniem 2013 roku zebrano 7 263 523 tweety związane z piłką nożną. Pierwszy z nich ma datę 23 października 15:35:24 a ostatni 29 grudnia 19:27:27. Wszystkie wpisy są powiązane z rozegranymi w tym czasie 35 spotkaniami klubów Arsenal F.C. , Chelsea F.C., Manchester United F.C. i Manchester City F.C. Daje to średnio 207 529 tweetów na mecz i niecałe 1 815 880 tweetów na drużynę.

Do zbierania tweetów użyte zostały dane 30 drużyn z 538 piłkarzami. Dodając do tego popularne określenia menadżerów, piłkarzy czy klubów sumarycznie zebrano 777 słów kluczowych, co daje średnio 22 słowa kluczowe na mecz.

Wpisy zostały stworzone przez 1 567 435 użytkowników, co daje 4.6 wpisu na użytkownika. 222 545 wpisów zawiera informacje o geolokacji, co stanowi zaledwie 3.06% liczby wszystkich wpisów. 666 199 wpisów to odpowiedzi (ang. *replies*), to jest 9.17%, natomiast aż 3 143 060 tweetów jest retweetami pokrywając 43.27% danych.

W tabeli 5.1 zaprezentowano listę wszystkich meczów, które były nasłuchiwane wraz z podstawowymi informacjami na ich temat.

Lp.	Data	Gospodarz	Gość	Tweetów	Z geolok.
1	2013-11-23 16:00	Arsenal Londyn	Southampton FC	190028	5231
2	2013-11-26 20:45	FC Basel	Chelsea Londyn	121209	3339
3	2013-11-26 20:45	Arsenal Londyn	Olympique Marseille	185252	6255
4	2013-11-27 20:45	Manchester City	Viktoria Plzen	24990	792
5	2013-11-27 20:45	Bayer Leverkusen	Manchester United	199232	6242
6	2013-11-30 16:00	Cardiff City FC	Arsenal Londyn	233151	6316
7	2013-12-01 13:00	Tottenham Hotspur	Manchester United	166394	4628
8	2013-12-01 17:10	Chelsea Londyn	Southampton FC	241768	7536
9	2013-12-01 17:10	Manchester City	Swansea City	29977	785
10	2013-12-04 20:45	Sunderland AFC	Chelsea Londyn	60047	1997
11	2013-12-04 20:45	Manchester United	Everton FC	182406	6226
12	2013-12-04 20:45	Arsenal Londyn	Hull City	200456	5076
13	2013-12-04 21:00	West Bromwich Albion	Manchester City	17783	608
14	2013-12-07 13:45	Manchester United	Newcastle United	416647	12613
15	2013-12-07 16:00	Stoke City	Chelsea Londyn	148780	4337
16	2013-12-07 16:00	Southampton FC	Manchester City	48101	1481
17	2013-12-08 17:00	Arsenal Londyn	Everton FC	381568	13057
18	2013-12-10 20:45	Manchester United	Shakhtar Donetsk	180301	6264
19	2013-12-10 20:45	Bayern Monachium	Manchester City	145381	4957
20	2013-12-11 20:45	Chelsea Londyn	Steaua Bucuresti	66125	1767
21	2013-12-11 20:45	Napoli	Arsenal Londyn	225461	7359
22	2013-12-14 13:45	Manchester City	Arsenal Londyn	525799	15561
23	2013-12-14 16:00	Chelsea Londyn	Crystal Palace	90541	2889
24	2013-12-15 14:30	Aston Villa	Manchester United	217221	6486
25	2013-12-21 16:00	Manchester United	West Ham United	171947	3975
26	2013-12-21 16:00	Fulham FC	Manchester City	63624	1624
27	2013-12-23 21:00	Arsenal Londyn	Chelsea Londyn	622011	19033
28	2013-12-26 13:45	Hull City	Manchester United	307313	8056
29	2013-12-26 16:00	Chelsea Londyn	Swansea City	109200	3264
30	2013-12-26 16:00	West Ham United	Arsenal Londyn	275811	8276
31	2013-12-26 18:30	Manchester City	Liverpool FC	331574	12009
32	2013-12-28 16:00	Manchester City	Crystal Palace	70251	2057
33	2013-12-28 16:00	Norwich City	Manchester United	204775	5348
34	2013-12-29 14:30	Newcastle United	Arsenal Londyn	339908	10624
35	2013-12-29 17:00	Chelsea Londyn	Liverpool FC	468494	16477

Tablica 5.1: Lista meczów, które były nasłuchiwane

5.2. Plan eksperymentów

Celem eksperymentów było udowodnienie sensowności zastosowania zarówno analizy sentymentu jak i geolokacji w analizie użytkowników sieci społecznościowych. Zaprezentowana jest seria badań pokazujących w jaki sposób można połączyć te trzy dziedziny, by odkrywać wiedzę dotyczącą sieci społecznych. Zaplanowane eksperymenty obejmują następujące zagadnienia:

- wartość sentymentu w kolejnych meczach (r. 5.3.1),
- aktywność użytkowników i sentyment w ciągu meczu (r. 5.3.2),
- sposoby komunikacji między zwolennikami i przeciwnikami klubu (r. 5.4.1),
- sentyment wypowiedzi w relacjach między użytkownikami (r. 5.4.2),
- struktura grup użytkowników w kolejnych meczach (r. 5.4.3),
- odległość między użytkownikami a częstość kontaktów (r. 5.5.1),
- rozkład wpisów na mapie świata (r. 5.5.2),
- odległość kibiców od miejsca rozgrywania meczu (r. 5.5.3),
- rozkład wpisów z geolokacją (r. 5.5.4).

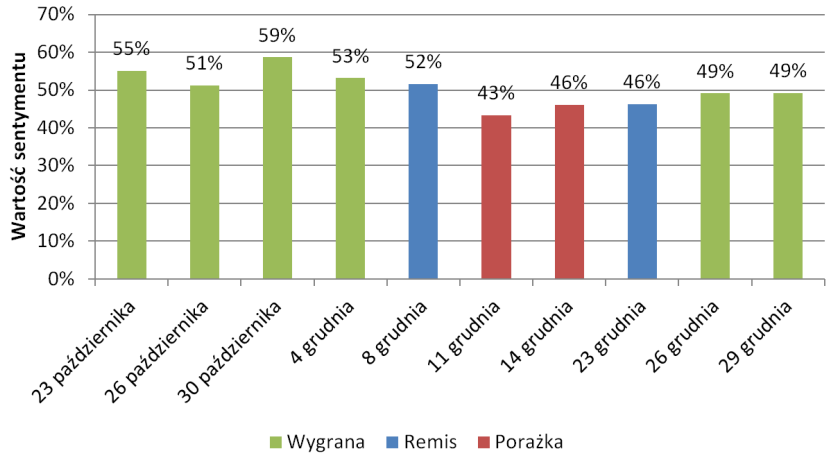
Eksperymenty zostały przeprowadzone dla wszystkich badanych drużyn. Ich wyniki były do siebie zbliżone, dlatego przedstawiam je tylko dla części z nich.

5.3. Analiza sentymentu

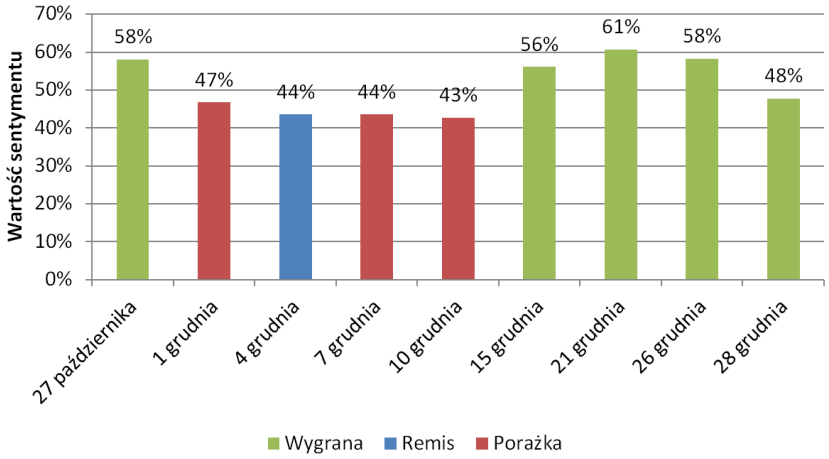
Analiza sentymentu została przeprowadzona zgodnie z algorytmem przedstawionym w rozdziale 2.2.3. W związku z tym, że wpisy typu retweet nie mogą posiadać sentymentu wszystkie zaprezentowane poniżej analizy odnoszą się do grupy wpisów nie będących retweetami. To daje nam 4 120 463 tweety, nad którymi były prowadzone badania. W tej grupie 2 005 934 wpisy zostały oznaczone jako pozytywne – 48.68%, a 1 944 448 jako negatywne – 47,19%.

5.3.1. Sentyment w meczach

Pierwszym eksperymentem jaki został przeprowadzony było zbadanie tego w jaki sposób zmienia się wydźwięk wypowiedzi pomiędzy kolejnymi meczami danej drużyny. W tym celu zbadany został ogólny sentyment podczas danego wydarzenia sportowego. Na wykresach 5.1 oraz 5.2 przedstawione są wyniki tych badań. Wartość pozytywności została wyliczone zgodnie z opisem w rozdziale 3.2.1.



Rysunek 5.1: Wyniki spotkań Arsenalu a sentyment wpisów



Rysunek 5.2: Wyniki spotkań Manchesteru United a sentyment wpisów

Posiadając informację na temat końcowego wyniku danego meczu łatwo można zauważyć, iż wartość sentymentu jest adekwatna do uzyskanego rezultatu danej drużyny. Gdy Arsenal odnosi zwycięstwo wówczas wpisy mają częściej wydźwięk pozytywny przekraczając w większości przypadków wartość 50%. Gdy jednak drużyna przegrywa nacechowanie emocjonalne wpisów wyraźnie spada poniżej 46%. Również mecz zakończony remisem powoduje raczej wpisy niezadowolenia.

Tę samą prawidłowość można zauważyć analizując wyniki badań dla meczów Manchesteru United. Gdy drużyna wygrywa, wówczas zadowolenia we wpisach sięga nawet 61%, a gdy ponosi porażkę – wynik sentymentu spada do czterdziestu kilku procent.

Po analizie tych badań nasuwają się oczywiste wnioski. Sposób reagowania internautów na wyniki ich drużyn jest dokładnie taki sam jak rezultat przez nie osiągany. Gdy drużyna wygrywa, jej kibice wykazują radość, szczęście, zadowolenie i inne pozytywne emocje. Natomiast gdy klub przegrywa mecz, wówczas wśród tweetów dużo łatwiej o wpisy o nacechowaniu negatywnym. Widać więc, że reakcje użytkowników Twittera są dokładnie takie same jak zwykłych kibiców – odzwierciedlają ich aktualny

stan ducha po meczu ulubionej drużyny. Nie ma tutaj żadnej różnicy między światem realnym a wirtualnym.

5.3.2. Liczba tweetów i rozkład sentymentu w ciągu meczu

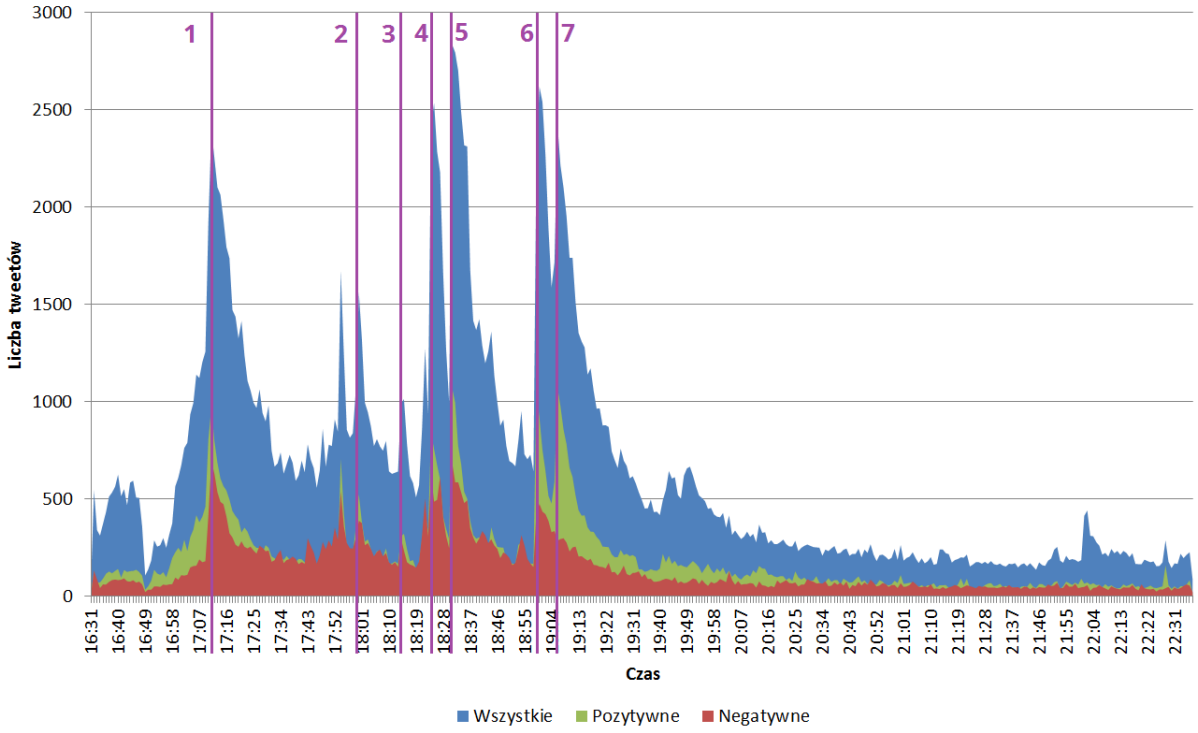
Kolejnym eksperymentem było zbadanie aktywności użytkowników Twittera w związku z wydarzeniami na boisku. Badanie takie można przeprowadzić dla każdego meczu, który znalazł się pośrod tych, które zostały pobrane. Poniżej przedstawione są rezultaty dla spotkania pomiędzy drużynami Chelsea F.C. a Southampton F.C. (01.12.2013 r.), które zakończyło się wynikiem 3-1. Są to dwa wykresy: pierwszy z liczbą tweetów na minutę (z uwzględnieniem wyrażanego przez nie sentymentu) (rys. 5.3) oraz drugi pokazujący zmianę sentymentu w trakcie spotkania (rys. 5.4).

Na obu wykresach zaznaczone są kluczowe wydarzenia:

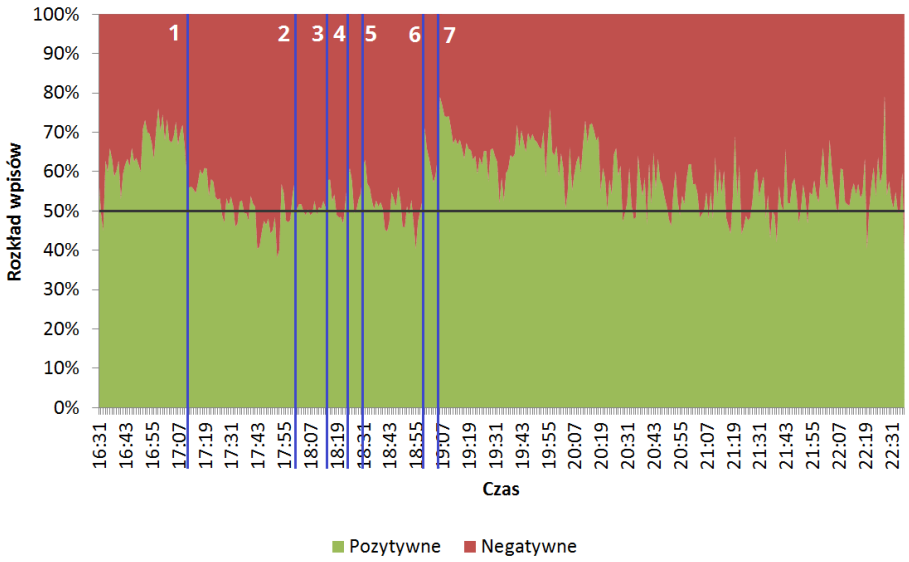
1. **17:10** – 1 min., początek meczu i gol J. Rodriguez (Southampton FC) 0-1.
2. **17:59** – 45 + 4 min., koniec pierwszej połowy.
3. **18:14** – 45 min., początek drugiej połowy.
4. **18:24** – 55 min., gol G. Cahill (Chelsea FC) 1-1.
5. **18:31** – 62 min., gol J. Terry (Chelsea FC) 2-1.
6. **18:59** – 90 min., gol D. Ba (Chelsea FC) 3-1.
7. **19:05** – 90 + 6 min., koniec spotkania.

Na pierwszym załączonym wykresie (rys. 5.3) wyraźnie widać silną korelację istotnych wydarzeń boiskowych z wysokim wzrostem liczby wysyłanych tweetów. Liczba ta rośnie nawet trzykrotnie w przypadku strzelonej bramki przez jedną z drużyn. Użytkownicy Twittera dużo mocniej angażują się w komentowanie i komunikację poprzez to medium w najważniejszych momentach danego spotkania. Można zauważyć, że początek i koniec spotkania są momentami, w których kibice generują stosunkowo dużą liczbę wpisów. W trakcie spotkania ich aktywność wzrasta, gdy na boisku dzieje się coś ciekawego. Po zakończonym meczu liczba wpisów stopniowo maleje, a dane spotkanie nie cieszy się już zainteresowaniem takiego szerokiego grona odbiorców jak wcześniej. Są to już zapewne raczej wiadomości wysyłane przez bardziej zagorzałych fanów, analizujących dane spotkanie dłużej, wyciągających z niego wnioski a nie tych osób, które były zainteresowane meczem tylko wtedy gdy ten się jeszcze odbywał.

Drugi wykres (rys. 5.4) przedstawiający zmiany sentymentu wyraźnie pokazuje, że wpisy były wyrównane w trakcie meczu (jeśli chodzi o proporcje sentymentu) aż do momentu, w którym Chelsea FC zdobyła bramkę na 3-1, ustalając w ten sposób de facto końcowy wynik spotkania. Wówczas wyraźnie przeważały wpisy o wydźwięku pozytywnym. Sam fakt tego, że to właśnie te tweety były liczniejsze może wynikać między innymi z tego, że Chelsea to klub mający więcej fanów na całym świecie niż Southampton, grający w ostatnich latach regularnie w Lidze Mistrzów, bijący się o zwycięsto w Premier League, czy jeżdżący w trakcie przygotowań do sezonu na tournee do Stanów Zjednoczonych i



Rysunek 5.3: Zmiana liczby tweetów w trakcie meczu Chelsea – Southampton



Rysunek 5.4: Rozkład sentymentu tweetów w trakcie meczu Chelsea – Southampton

Azji. Southampton natomiast to klub z dolnej części tabeli, mający zupełnie inne cele w trakcie sezonu, nieposiadający tylu gwiazd, w związku z czym nie skupiający wokół siebie takiego zainteresowania.

Bardzo podobne wykresy można uzyskać analizując także inne spotkania. Tam również internauci mocniej angażują się, gdy dzieje się coś ciekawego, a sentyment jest zgodny z wynikiem i liczbą przeważających kibiców danego klubu. Widać więc, że Twitter jest miejscem, w którym jego użytkownicy uzewnętrzniają swoje emocje błyskawicznie, gdy dzieje się coś co ich porusza. Są to takie same naturalne

emocje, których doświadczają oni na co dzień, nie są w żaden sposób wyimaginowane, przemyślane, czy sterowane, ale pokazują aktualny stan ducha danej społeczności. Można więc dojść do wniosku, że badanie Twittera może przynieść nam wyniki, do których powinniśmy podejść poważnie i nie bagatelizować ich twierdząc, że być może w internecie osoby zachowują się inaczej niż w codziennym życiu.

5.4. Analiza sieci społecznych

Do analizy sieci społecznych zostały użyte dane użytkowników pobrane równocześnie ze ściąganiem wpisów. Relacje między użytkownikami zostały zbudowane na podstawie informacji zawartych w tweetach. Są to więc dwa rodzaje relacji – odpowiedzi i retweety. Z ich pomocą przeprowadzone zostały badania nad siecią społeczną, którą tworzą użytkownicy Twittera.

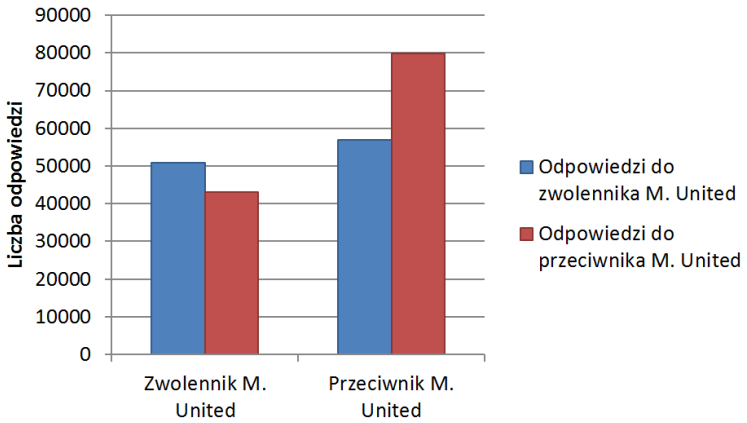
5.4.1. Liczba i rodzaje komunikacji między zwolennikami i przeciwnikami klubów

Jednym z eksperymentów przeprowadzonych w analizie sieci społecznych było sprawdzenie rodzaju komunikacji pomiędzy użytkownikami. Dodatkowo użytkownicy zostali podzieleni na zwolenników i przeciwników danego klubu zgodnie z opisem w rozdziale 3.2.1.

Charakterystyka komunikacji poprzez odpowiedzi (*replies*)

Zbadana został sposób w jaki komunikują się użytkownicy Twittera korzystając z funkcji *odpowiedz*, polegającej na możliwości komentowania wpisów innych użytkowników. Tak jak było to wcześniej zaznaczone w tym serwisie społecznościowym użytkownicy do woli mogą komentować wpisy osób, których nie mają na swoich listach znajomych.

Poprzez podzielenie użytkowników na grupy zwolenników i przeciwników – na przykładzie wpisów dotyczących Manchesteru United – można zauważyć ciekawe obserwacje, co można zaobserwować na wykresie (rys. 5.5).



Rysunek 5.5: Charakterystyka odpowiedzi wśród wpisów dotyczących Manchesteru United

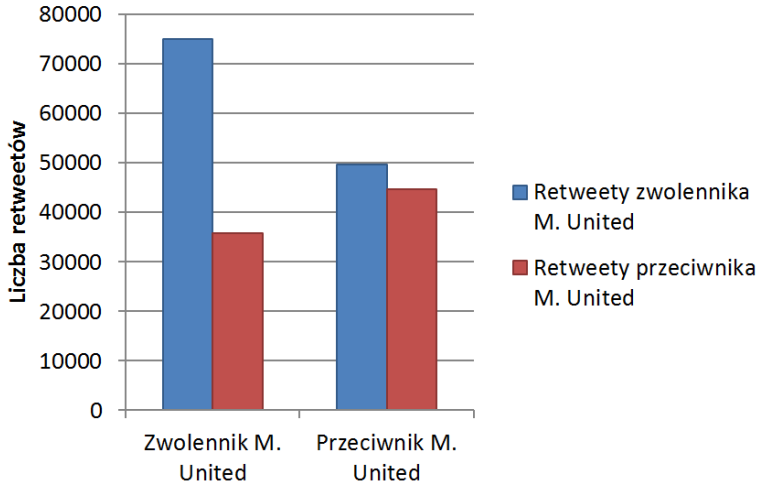
Wykres przedstawia liczbę odpowiedzi zwolenników i przeciwników Manchesteru United na wpisy zwolenników i przeciwników tego klubu. Na jego podstawie można zauważyć, że na wpisy zwolenników

częściej odpowiadają zwolennicy. Oznacza to, że ta grupa użytkowników trzyma się blisko siebie i osoby, które są sympatykami Manchesteru również obserwują i komunikują się z innymi sympatykami tego klubu produkując większą liczbę wpisów. Podobna zależność ma miejsce wśród przeciwników Manchesteru United. Wpis przeciwnika tego klubu bardziej angażuje do dyskusji również innych przeciwników. Prawdopodobnie osoby biorące udział w tej dyskusji wspólnie narzekają na grę tego zespołu, wzajemnie nakręcając się do ożywionych dyskusji.

Na podstawie powyższego wykresu widać więc, że użytkownicy Twittera lubią tworzyć grupy o podobnych zainteresowaniach czy sympatiach. Dany użytkownik częściej będzie komunikował się z osobami, które myślą podobnie jak on, umacniając w ten sposób swoje przekonanie o własnych przemyśleniach na dany temat. Gdy ktoś jest zwolennikiem danego klubu częściej dyskutuje z podobnymi sobie internautami. Tak samo przeciwnicy łatwiej znajdują nić porozumienia między sobą mając takie samo zdanie dotyczące określonego wydarzenia. Internauci lepiej odnajdują się wśród osób podzielających ich opinie.

Charakterystyka komunikacji poprzez retweety

Podobnie jak powyższe badanie został przeprowadzony eksperyment na temat charakterystyki komunikacji między użytkownikami korzystający z opcji *retweet*. Polega ona na podaniu dalej wpisu, który uważamy za ciekawy. Wyniki, które w ten sposób uzyskano różnią się od poprzednich (rys. 5.6).



Rysunek 5.6: Charakterystyka retweetów wśród wpisów dotyczących Manchesteru United

Tym razem widać zupełnie inne ułożenie słupków na wykresie (rys. 5.5 i 5.6). Na pierwszy plan wyraźnie wysuwa się słupek pierwszy z lewej, który oznacza liczbę wpisów zwolenników podanych dalej przez innych zwolenników. Widać więc, że sympatycy Manchesteru United bardzo chętnie przekazują dalej wpisy pozostałych sympatyków. Mogą to być na przykład wpisy o strzelonym golu, czy jakieś pozytywne opinie na temat danej drużyny. Najrzadziej z całej czwórki zaprezentowanych relacji dochodzi do sytuacji, gdy przeciwnik Manchseteru podaje dalej wpis zwolennika. Jeśli chodzi o retweetowanie tweetów przeciwników Manchesteru United, to dochodzi do tego mniej więcej po równo między przeciwnikami i zwolennikami.

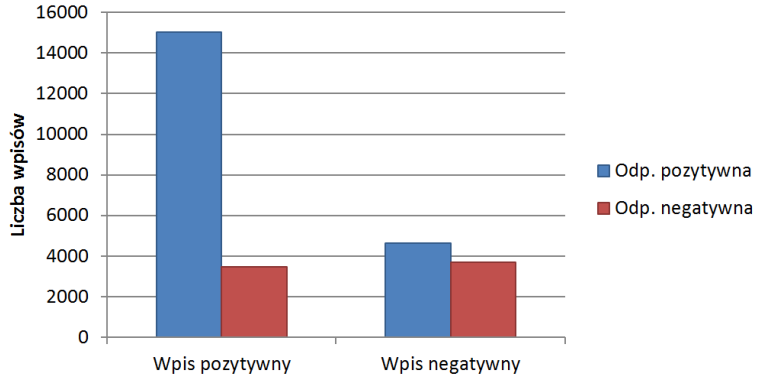
Po analizie dwóch powyższych eksperymentów nasuwają się następujące wnioski. Użytkownicy lubią gromadzić się w grupy o podobnych zainteresowaniach, wspólnie komentując wydarzenia w podobny sposób. Zwolennicy i przeciwnicy danego klubu zachowują się w charakterystyczny sposób. Ci pierwsi bardzo często retweetują wpisy innych zwolenników, a ci drudzy częściej dyskutują ze sobą.

5.4.2. Sentyment odpowiedzi między zwolennikami i przeciwnikami drużyny

Oprócz zbadania rodzaju i charakterystyki komunikacji pomiędzy użytkownikami Twittera przyjrzano się także bliżej komunikacji związanej z odpowiedziami (*replies*). Przeprowadzono eksperyment, w którym zwrócono uwagę na wydźwięk wpisów będących odpowiedziami na inne wpisy – ponownie z podziałem na zwolenników i przeciwników danej drużyny. W tym rozdziale skupiono się tylko na odpowiedziach, gdyż niemożliwe jest badanie sentymentu retweetów, które są tylko podaniem dalej innych wpisów. Poniżej prezentuję jak rozkładał się sentyment wpisów dla wszystkich kombinacji zwolenników i przeciwników Arsenalu jako autorów i odpowiadających.

Gdy odpowiada zwolennik Arsenalu

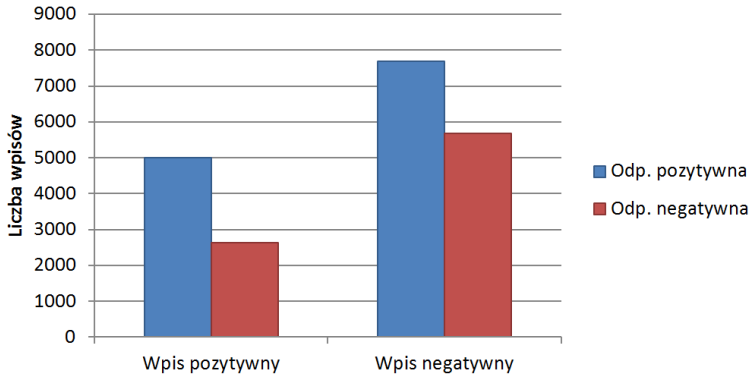
Poniżej prezentuję dwie sytuacje, w których odpowiadającym na wpis jest użytkownik będący zwolennikiem drużyny Arsenalu Londyn. W pierwszym przypadku (rys. 5.7) pokazana jest struktura odpowiedzi zwolennika na wpisy innego zwolennika, zaś w drugim (rys. 5.8) odpowiedzi odnoszą się do wpisów przeciwnika klubu.



Rysunek 5.7: Sentyment odpowiedzi. Odpowiada zwolennik Arsenalu na wpis zwolennika

To co rzuca się na pierwszy rzut oka to fakt, że zwolennicy danego klubu najczęściej tworzą wpisy pozytywne. Zauważalne jest to, że wśród zwolenników dominującym modelem komunikacji jest wymiana wiadomości o nacechowaniu pozytywnym – na wpis pozytywny odpowiedzą jest również wpis o takim samym sentymencie – wyraźnie wyróżniająca się spośród pozostałych wariantów. Zauważyć można również to, że przeciwnicy Arsenalu generują więcej wpisów negatywnych a mimo to sympatycy klubu z Londynu starają się im odpowiadać tweetami o wydźwięku pozytywnym.

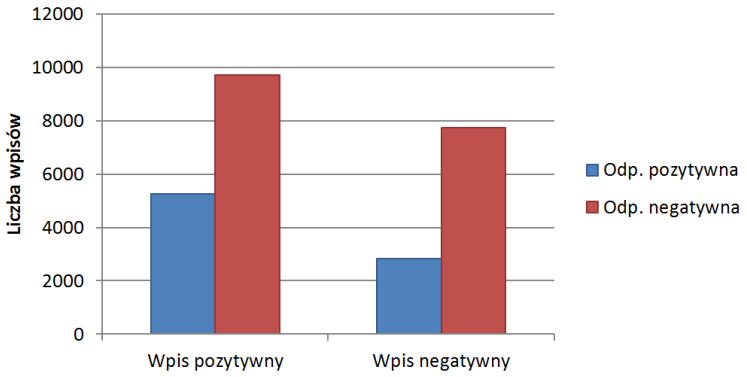
Widać więc, że osoby tworzące pozytywne wpisy na dany temat pobudzają się wzajemnie do ożywionej dyskusji a także starają się dbać o dobre imię i dobry odbiór tematu, który jest im bliski. Tworzą więc grupę, która dobrze czuje się w swoim towarzystwie a także stara się kreować pozytywny odbiór swojego ulubionego klubu na zewnątrz, wśród innych osób.



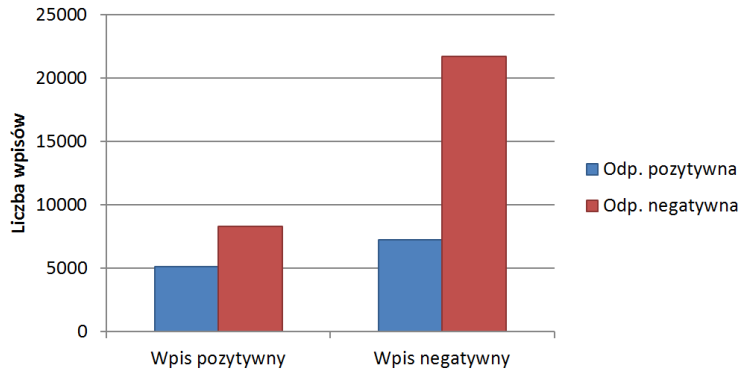
Rysunek 5.8: Sentyment odpowiedzi. Odpowiada zwolennik Arsenalu na wpis przeciwnika

Gdy odpowiada przeciwnik Arsenalu

Analogicznie do poprzednich eksperymentów w dwóch wykresach poniżej (na rysunkach 5.9 i 5.10) zaprezentowane są wyniki badań nad odpowiedziami przeciwnika Arsenalu na wpisy zwolenników i przeciwników tego klubu.



Rysunek 5.9: Sentyment odpowiedzi. Odpowiada przeciwnik Arsenalu na wpis zwolennika



Rysunek 5.10: Sentyment odpowiedzi. Odpowiada przeciwnik Arsenalu na wpis przeciwnika

Z powyższych wykresów można wywnioskować, że wpisy przeciwników mają najczęściej wydźwięk negatywny. Gdy przeciwnik odpowiada na pozytywny wpis zwolennika, to tylko 1 na 3 wpisy są również pozytywne. Zdecydowanie częściej przeciwnik odpisuje zwolennikowi wpisem o sentymencie ne-

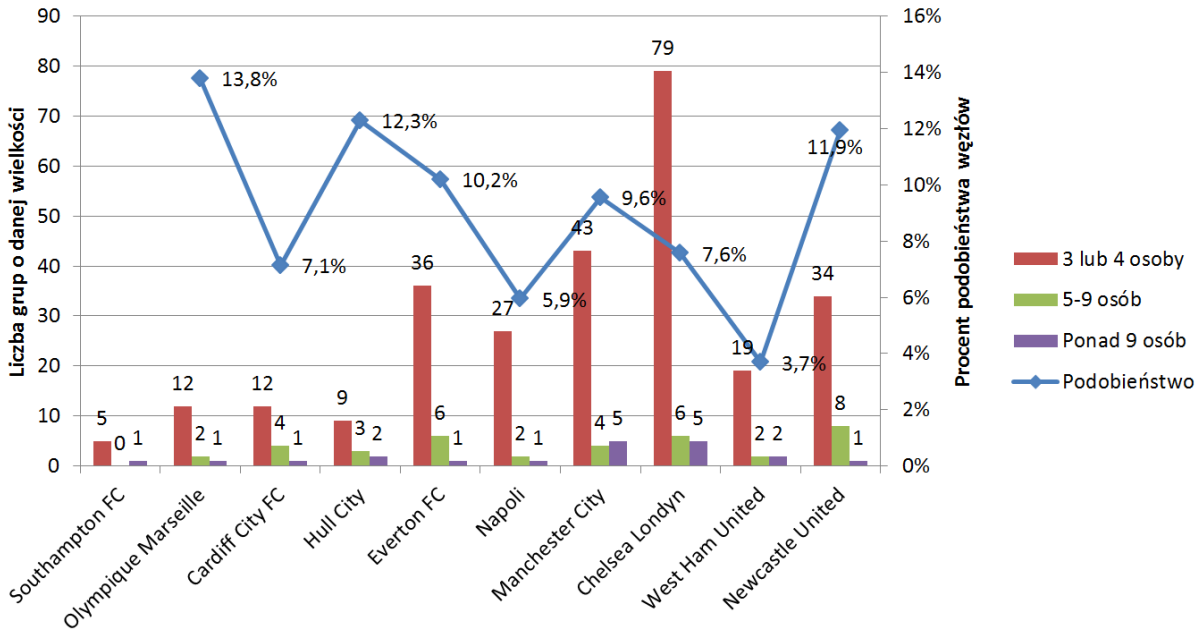
gatywnym. Co warte zauważenia najwięcej odpowiedzi przeciwnicy Arsenalu wysyłają pod wpisami negatywnymi innych przeciwników, również nacechowując swoje opinie negatywnie.

Analogicznie więc jak w poprzednim eksperymencie widać, że przeciwnicy Arsenalu tworzą wspólną grupę, w której wymieniają się swoimi krytycznymi opiniami na temat tej drużyny. I tak samo jak poprzednio wychodzą również z tymi opiniami do innych internautów starając się przekonać ich do swojego punktu widzenia.

5.4.3. Analiza grup w sieciach społecznych

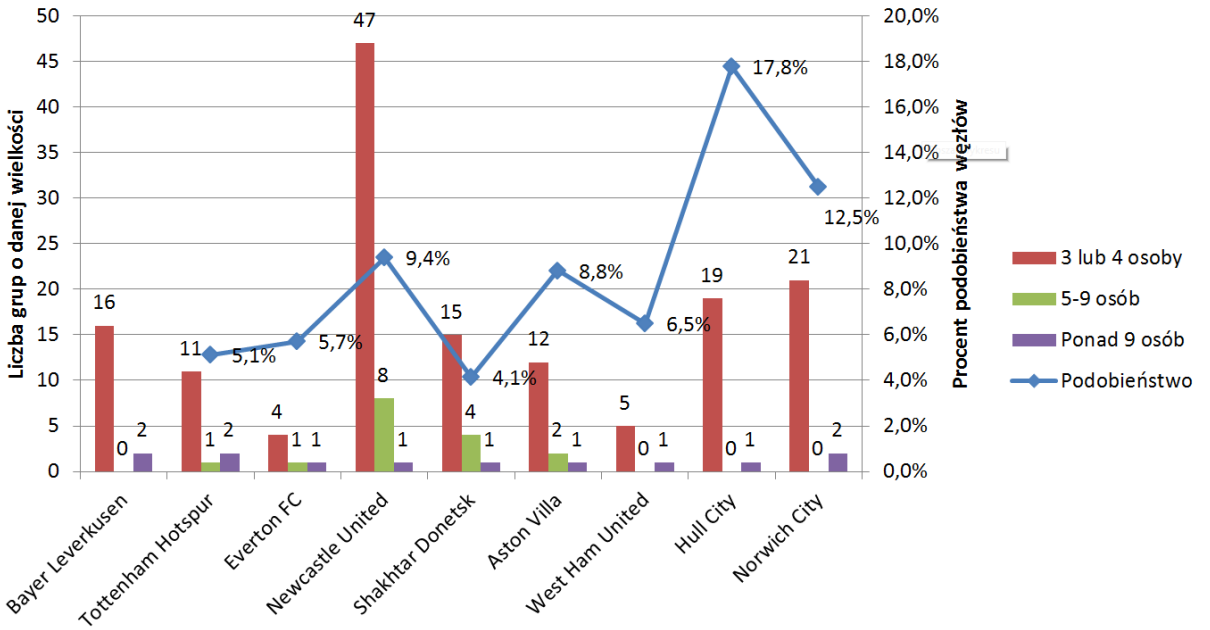
Oprócz powyższych badań nad sposobami komunikacji i relacji budowanymi między użytkownikami przeprowadzona została także analiza grup użytkowników pomiędzy meczami. Grupy te były budowane zgodnie z opisem w rozdziale 3.2.2.

Liczności tych grup zostało zaprezentowane na poniższych wykresach pokazując jak zmieniały się z meczu na mecz. Na pierwszym (rys. 5.11) przedstawione są dane dla Arsenalu, a na drugim (rys. 5.12) dla Manchesteru United. Oprócz wielkości poszczególnych grup wykresy prezentują także podobieństwo zbioru użytkowników komentującego następujące po sobie wydarzenia. Sposoby wykrywania grup i badania podobieństwa zostały opisane w rozdziale 3.2.2.



Rysunek 5.11: Struktura grup użytkowników w meczach Arsenalu

Z analizy wykresów można zaobserwować, że w każdym spotkaniu najliczniejsze są małe grupy – 3 lub 4 osobowe. Liczba tych grup dodatkowo rośnie gdy mecz odbywa się z ciekawym przeciwnikiem. Na przykład zauważyć można, że w meczach Arsenalu (rys. 5.11) bardzo dużą liczbę grup wygenerowały spotkania z Manchesterm Citym i Chelsea Londyn, a w meczach Manchesteru United skok liczby grup miał miejsce w meczu z Newcastle United. Grupy liczniejsze jeśli chodzi o wielkość tworzyły się już zdecydowanie rzadziej. I podobnie jak poprzednio ich większa liczbę również można zaobserwować w ciekawszych spotkaniach.



Rysunek 5.12: Struktura grup użytkowników w meczach Manchesteru United

Jeśli zaś chodzi o podobieństwo w kolejnych meczach to mamy tutaj do czynienia z ciekawą, aczkolwiek zrozumiała sytuacją. Otóż im ciekawszy przeciwnik, tym podobieństwo zbioru użytkowników w kolejnym meczu mniejsze. Gdy drużyna – w przypadku Arsenalu – gra najpierw z Chelsea a później z West Ham United, to podobieństwo wynosi 3.7%, a gdy najpierw gra z West Ham United – w przypadku Manchesteru Untied – a potem z Hull City to podobieństwo między tymi meczami sięga 17.8%.

Widać więc pewną charakterystyczną zależność. Gdy drużyna gra mecze z popularnymi drużynami, wówczas liczba osób biorących na Twitterze udział w danym wydarzeniu piłkarskim jest duża. Spotkanie takie absorbuje większą publikę. Stąd większe słupki grup w popularnych meczach. Gdy jednak przeciwnik jest już nieco mniej ciekawy, wówczas również zainteresowanie na Twitterze takim meczem spada.

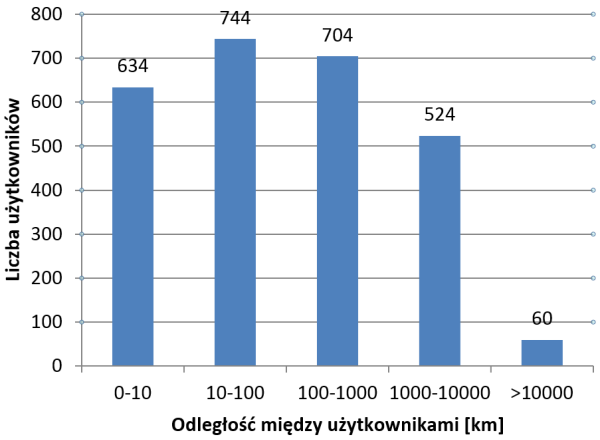
Z tych prawidłowości wynika również dlaczego osiągany był taki a nie inny wynik podobieństwa. Otóż gdy zespół gra mecz. to zawsze wśród tweetujących na jego temat jest stała grupa fanów, która w spotkaniach ze słabymi rywalami jest łatwiej dostrzegalna – stanowi większy odsetek wszystkich użytkowników Twittera. Gdy natomiast mecz jest interesujący dla większej liczby odbiorców – wówczas ta stała grupa fanów jest trudniej dostrzegalna i odsetek podobieństwa drastycznie spada. Mecze z ciekawymi rywalami przyciągają więc do siebie osoby okazjonalnie zainteresowane danym temat – te osoby następnym razem nie będą komentować meczu tej drużyny, gdy ta będzie rozgrywać spotkanie z mało ciekawym rywalem. Wówczas jednak łatwiej będzie dostrzec tę grupę, która stanowi trzon zainteresowanych daną drużyną i która jest jej stałym fanem.

5.5. Analiza geolokacji

W badaniach skupiono się także na zachowaniu użytkowników z wykorzystaniem geolokalizacji. Tylko niewielka część wpisów zawierała informacje o tym, z którego miejsca została wysłana. Mimo to przy ich pomocy możliwe było dokonanie analizy danych związanych z położeniem geograficznym.

5.5.1. Odległość między użytkownikami a częstość kontaktów

Pierwsze badanie polegało na zmierzeniu odległości fizycznej pomiędzy użytkownikami, którzy się ze sobą kontaktują i zbadaniu korelacji tej odległości do liczby wiadomości jakie między sobą wymieniają. Sposób przeprowadzenia tego badania został opisany w rozdziale 3.2.3.



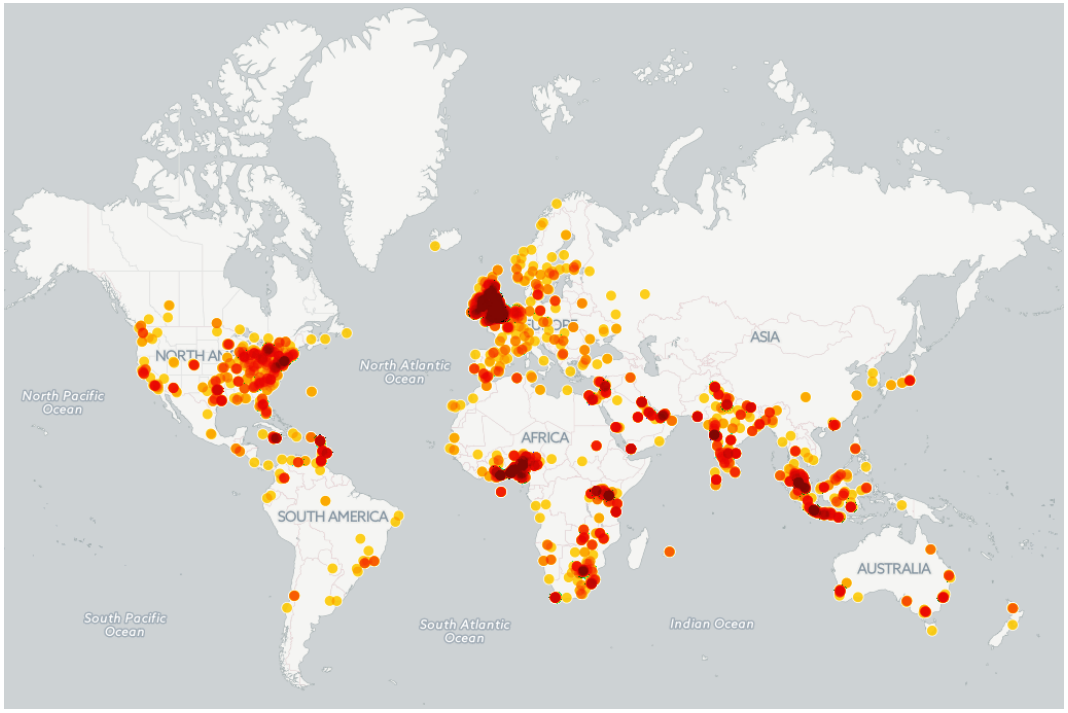
Rysunek 5.13: Odległość między użytkownikami a częstość kontaktów

Każda kolejna kolumna na powyższym wykresie pokrywa obszar o dziesięciokrotnie większym promieniu. Oznacza to, że powierzchnia, której dotyczą rośnie eksponencjalnie. Pomimo tego, liczba użytkowników którzy wchodzili ze sobą w interakcje jest za każdym razem tego samego rzędu wielkości (co widać na wykresie 5.13). Oznacza to więc, że im bliżej siebie byli użytkownicy fizycznie tym częściej odpowiadali na swoje wpisy. Gdy znajdowali się w odległości maksymalnie 10 kilometrów to do wymiany wpisów dochodziło tak samo często gdy byli od siebie oddaleni od 10 do 100 km, czy nawet od 100 do 1000 km.

Wniosek jest więc taki, iż pomimo tego, że internet nazywany jest globalną wioską łączącą ludzi z całego świata to użytkownicy komunikują się i tak z najbliższymi sobie. Może to wynikać z kilku powodów. Na przykład osoby z Manchesteru prawdopodobnie będą kibicami drużyny United lub City i większe jest prawdopodobieństwo, że będą komunikać się między sobą niż między osobami z innego miasta czy kraju. Innym powodem takiego stanu rzeczy może być również bariera językowa. Im dalej od danego kraju, tym mniej osób będzie w stanie posługiwać się językiem tam panującym. Dodatkowym ograniczeniem może być czas lokalny, czyli strefy czasowe. Gdy w jednym punkcie jest dzień, w drugim może być środek nocy jednoznacznie utrudniający komunikację między osobami z większych odległości.

5.5.2. Rozkład wpisów na mapie

Zebrane Tweety posiadające geolokalizację można przedstawić na mapie[1]. Poniżej na trzech kolejnych rysunkach zaprezentowany został rozkład wpisów, które wysłane zostały podczas meczu Chelsea F.C. z Southampton F.C. (01.12.2013 r.). Najpierw zaprezentowane są wpisy w skali całej Ziemii (rys. 5.14), następnie w skali Wielkiej Brytanii (rys. 5.15), a na końcu na mapie zawierającej obszar między Londynem a Southampton (rys. 5.16), czyli miastami obu klubów.

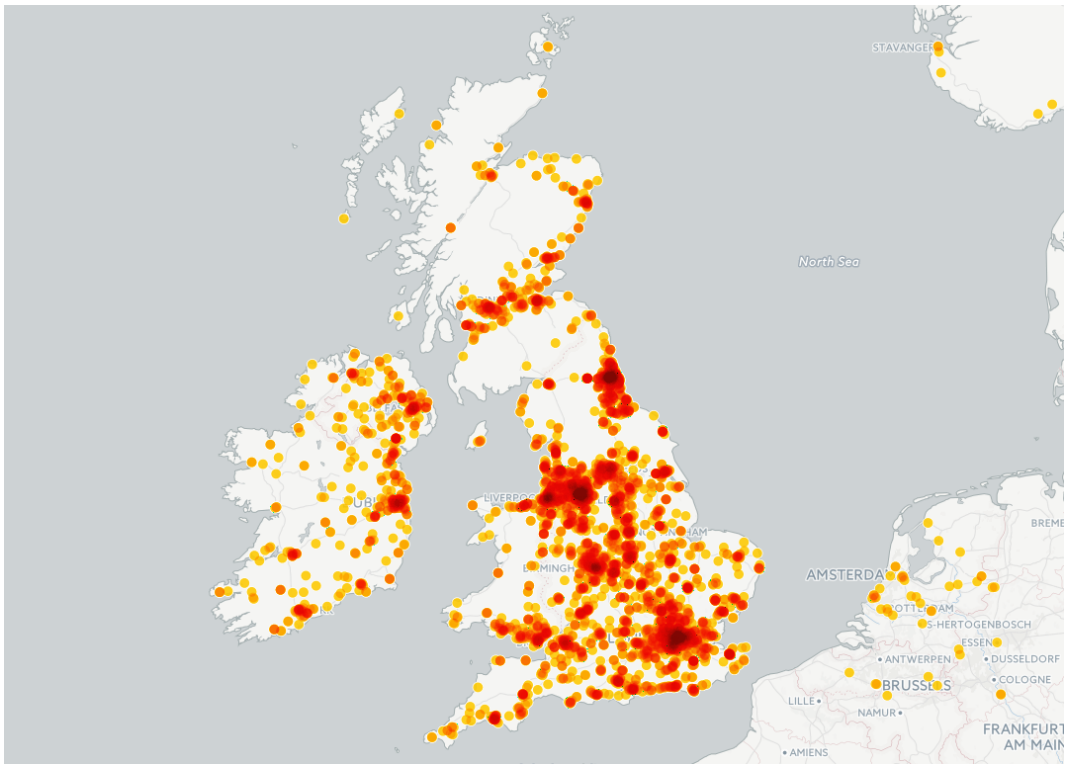


Rysunek 5.14: Rozkład wpisów na mapie świata

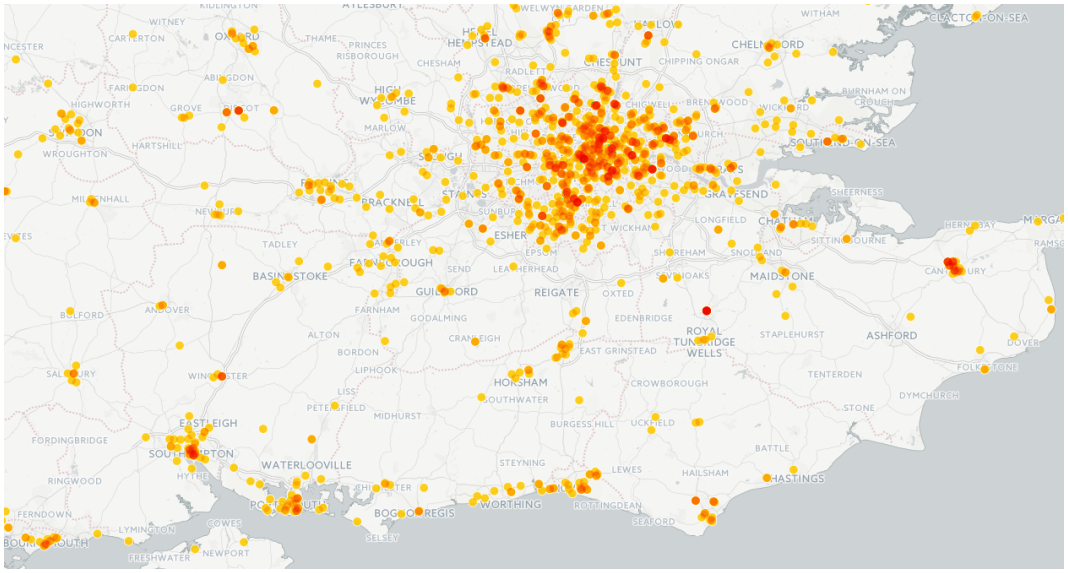
Na mapie świata (rys. 5.14) wyróżnia się kilka krajów. Przede wszystkim na pierwszym plan wyłania się Wielka Brytania – z oczywistych względów jako miejsce, w którym odbywa się mecz oraz całe rozgrywki. Oprócz niej dużą liczbę tweetów wysyła się z takich krajów jak między innymi: Stany Zjednoczone (3 kraj pod względem liczby ludności na świecie, dominujący język angielski), Indie (2 kraj pod względem ludności, język angielski), Indonezja (4 kraj pod względem ludności) a także państwa afrykańskie: Kenia i Uganda na wschodze oraz Nigeria i Ghana na zachodzie. Należą one do najbardziej zaludnionych krajów Czarnego Lądu i w każdym z nich język angielskim jest językiem urzędowym. Na mapie nie ma wpisów z Chin (najludniejszego kraju świata gdzie Twitter został zablokowany w 2009 roku[2]) oraz Rosji, w której popularniejsze są rodzime serwisy społecznościowe.

W meczu wzięli udział między innymi zawodnicy z Nigerii (John Obi Mickel – Chelsea FC) oraz Ghany (Michael Essien – Chelsea FC, Victor Wanyama – Southampton FC), co zapewne wzmogło aktywność internautów z tych regionów.

[1] Załączone wizualizacje zostały stworzone przy pomocy serwisu CartoDB (www.cartodb.com)
[2] http://www.theguardian.com/technology/2009/jun/02/twitter-china



Rysunek 5.15: Rozkład wpisów w Wielkiej Brytani



Rysunek 5.16: Rozkład wpisów wokół Londynu i Southampton

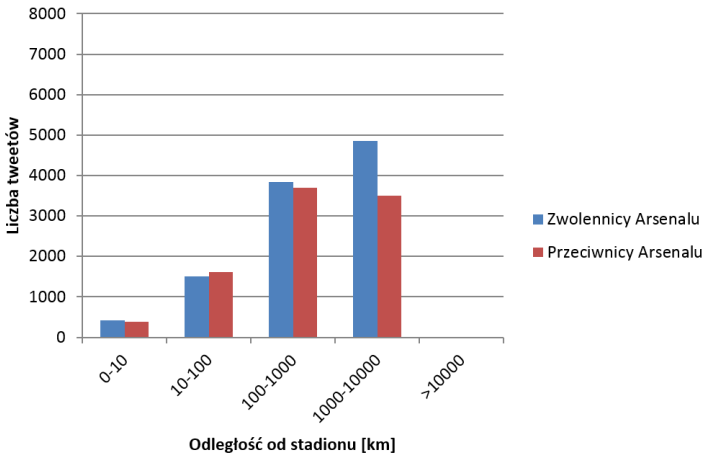
Oceniając rozkład wpisów w Wielkiej Brytanii (rys. 5.15) zauważymy dużą ich koncetrację wokół Londynu – największego miasta Zjednoczonego Królestwa (algomeracja Londynu to ponad 13 milionów ludzi[3]). Gdy zbliżymy mapę do obszaru Southampton i Londynu (rys. 5.16) to za pomocą analizy rozkładu wpisów na mapie bez problemu zlokalizujemy położenie tych miast. Właśnie w nich znajduje się najwięcej ich kibiców generujących największą liczbę wpisów. Widać więc, że dane wydarzenie an-

[3] 2012 rok. Źródło: Eurostat (http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=met_pjanaggr3&lang=en)

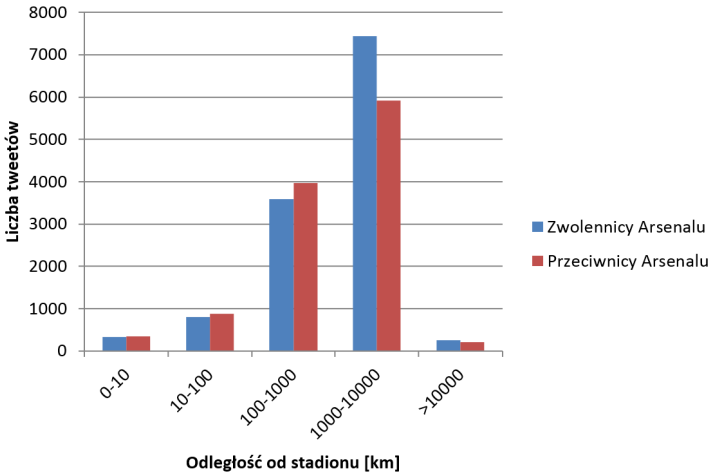
gażuje przede wszystkim osoby, którym jest ono bliskie. Zastosowanie geolokalizacji pozwala nam te miejsca odkryć. I tak jak w przypadku meczu piłarskiego można takie miejsca określić de facto jeszcze przed analizą, tak w innego rodzaju badaniach przy użyciu tej techniki możliwe może być odkrycie miejsc, obszarów geograficznych, o których nie myślałoby się w kontekście badanego tematu. Powyższe wizualizacje potwierdzają stosowność zastosowania geolokacji i jej skuteczności w różnego rodzaju badaniach społecznych.

5.5.3. Odległość od stadionu

Dane zawierające geolokalizację pozwalają wykonać szereg ciekawych eksperymentów. Kolejnym z nich było zbadanie odległości kibiców od stadionu w zależności od tego czy drużyna grała mecz u siebie czy na wyjeździe. Dodatkowo kibiców tych (dzięki analizie sentymentu (r. 3.2.1)) można było podzielić na zwolenników i przeciwników. Takie badanie zostało przeprowadzone dla wszystkich meczów Arsenalu Londyn a wyniki zaprezentowane są na rysunkach 5.17 i 5.18. Eksperyment ten został przeprowadzony zgodnie z opisem w rozdziale 3.2.3.



Rysunek 5.17: Odległość od stadionu w meczach Arsenalu u siebie

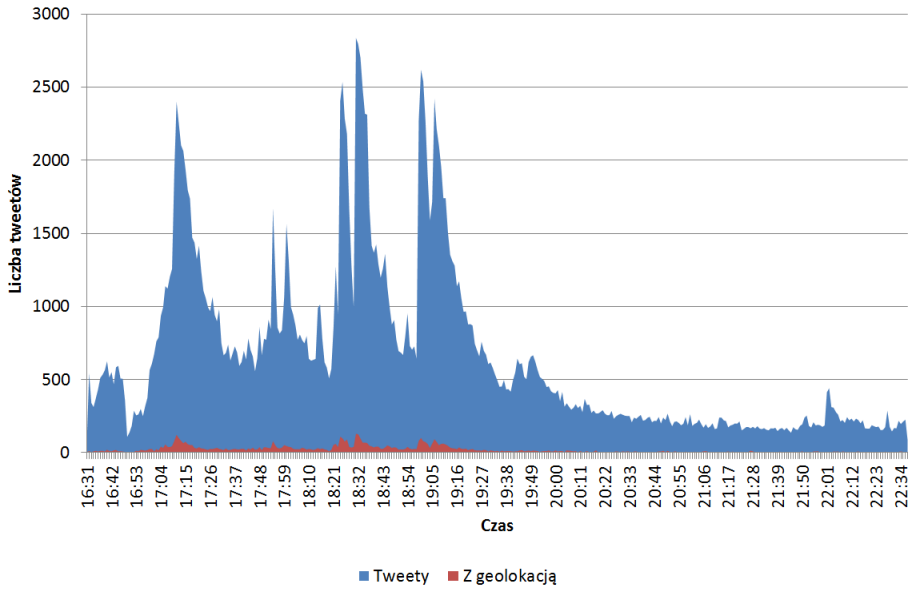


Rysunek 5.18: Odległość od stadionu w meczach Arsenalu na wyjeździe

To co możemy na nich zaobserwować to fakt, że gdy Arsenal rozgrywa spotkanie na własnym stadionie (rys. 5.17), wówczas liczba wpisów zwolenników przeważa wpisy przeciwników. Zwłaszcza w odległości 0-10 i 10-100 kilometrów od stadionu. Gdy natomiast mecz odbywa się na wyjeździe (rys. 5.18), wówczas lekko przeważające są wpisy, których autorami są przeciwnicy Arsenalu. W przypadku kiedy spojrzymy na kolumnę z wpisami wysyłanymi z odległości 1000-10000 tysięcy kilometrów od stadionu wtedy w obu przypadkach więcej wpisów wysyłanych jest przez zwolenników londyńskiego klubu. Widać więc, że spora część kibiców Arsenalu zapewne komentuje z Londynu i nie przemieszcza się za swoją drużyną na każdy mecz. Wówczas, gdy klub rozgrywa spotkanie na wyjeździe bliżej stadionu są kibice drużyny przeciwnej, którzy tworzą wpisy odznaczające się raczej niechęcią do Arsenalu. Jeśli natomiast chodzi o kibiców z dalszych zakątków świata to, można stwierdzić, że jeśli interesują się oni tą drużyną, to raczej nastawieni są do niej pozytywnie i łatwiej jest o sympatyków niż antyfanów.

5.5.4. Rozkład wpisów z geolokacją w czasie meczu

Pośród wszystkich zebranych tweetów tylko 3.06% z nich zawierało informacje o geolokacji. Poniżej na podstawie meczu Chelsea F.C. z Southampton F.C. (01.12.2013 r.) zaprezentowano porównanie (rys. 5.19) liczby wpisów z geolokacją do wszystkich wpisów.



Rysunek 5.19: Rozkład wpisów z geolokacją w trakcie meczu

Wykres jednoznacznie obrazuje o jakiej różnicy jest mowa. Nieco ponad 3% wpisów z geolokacją to bardzo niewiele. Widać, że liczba takich tweetów rośnie w tych samych momentach, gdy rośnie ogólna liczba wpisów. Niestety (z punktu widzenia badań) geotagowanie wpisów nie jest jeszcze czymś powszechnym i zapewne musi minąć trochę czasu, by użytkownicy serwisów społecznościowych chętniej dzielili się miejscem, w którym się znajdują.

5.6. Podsumowanie eksperymentów

Zaprezentowane powyżej eksperymenty potwierdziły, że zastosowane podejście do analizy dużej sieci społecznej z wykorzystaniem analizy sentymentu i geolokacji powiodło się. Przedstawiony w rozdziale 3.3 algorytm okazał się skutecznym narzędziem pokazującym zmieniające się nastroje wśród kibiców, a także pomógł zbadać sposoby interakcji między nimi z podziałem na zwolenników i przeciwników. Geolokacja, chociaż jeszcze niezbyt popularna, dostarcza interesujące wyniki.

Najwięcej wpisów pochodzi z miejsc, w których faktycznie jest zainteresowanie danym tematem. Okazało się, że kibice najchętniej komentują mecze z największymi rywalami, a podobieństwo zbioru użytkowników między kolejnymi spotkaniami jest tym większe im mniej popularny jest mecz – potwierdzając tym samym, że to najzagorzalsi fani są ze swoim klubem niezależnie od sytuacji.

Przeprowadzone eksperymenty udowodniły, że można wykorzystać nowoczesne techniki komputerowe do badania dużych sieci społecznych wzbogacając je o analizę sentymentu i geolokację, które służą do lepszego zrozumienia badanych społeczności.

6. Zakończenie i wnioski

6.1. Podsumowanie

W niniejszej pracy próbowano powiązać ze sobą analizę sentymentu i dane geolokacyjne w celu wzbogacenia analizy użytkowników sieci społecznościowych. Cel można uznać za osiągnięty.

Analiza sieci społecznych pozwala pokazać w jaki sposób osoby badane łączą się ze sobą i jaka jest charakterystyka tych połączeń. Pokazuje wielkość utworzonych grup i ich trwałość.

Zastosowanie analizy sentymentu dostarcza informacje, dzięki którym możemy odkrywać przyczyny tworzenia się takich a nie innych grup a także pomaga dowiedzieć się, dlaczego między różnymi grupami występują konkretne rodzaje relacji. Korzystając z tej dziedziny nauki możemy także skorelować nastroje społeczne z wydarzeniami na świecie. W niektórych przypadkach zauważony sentyment może być w miarę przewidywalny – tak jak w badaniu kibiców piłkarskich, gdzie związany jest z wynikiem meczu – ale w innych może pozwalać odkrywać niezauważane do tej pory ciągi przyczynowo-skutkowe.

Analiza sieci społecznych może również prowadzić do bogatszych wniosków poprzez zastosowanie danych związanych z geolokalizacją. Powiązanie informacji o fizycznym położeniu użytkowników, a także o sentymencie jaki generują daje dodatkowy obraz pozwalający zrozumieć połączenia między nimi.

Dodatkowo w tej pracy udało się wykorzystać dane pochodzące z serwisu społecznościowego – którym był Twitter. Jest to bardzo ciekawe medium, które może być szeroko używane do automatycznego badania nastrojów i sieci społecznych. Pokazano w jaki sposób takie dane uzyskać, przetworzyć i wykorzystać do podobnych badań. Zaprezentowane zostało podejście, dzięki któremu badanie dużych grup ludzi można przeprowadzić bez ich wiedzy i w sposób automatyczny. Dzięki temu uzyskuje się bardziej wiarygodne wyniki. Użytkownicy nie wiedzą, że są obserwowani i zachowują się w sposób naturalny.

6.2. Wpływ pracy na otaczający świat

Praca ta prezentuje w jaki sposób można połączyć ze sobą analizę sieci społecznych, analizę sentymentu i analizę geolokalizacji. Udowadnia, że badanie nastrojów w społeczeństwie można zautomatyzować aplikując komputerowe techniki przetwarzania dużych zbiorów danych. Pokazuje, że zastosowanie analizy sentymentu i geolokacji może istotnie wzbogacić analizę dużych grup ludzi.

Zastosowanie automatycznych technik badania dużych sieci społecznych może uprościć sposoby komunikacji z takimi grupami i odpowiadania na ich potrzeby. Możemy wyobrazić sobie sytuację, w

której rządy, organizacje czy firmy badając sieci społeczne wraz z analizą i geolokacją błyskawicznie reagują na aktualne wydarzenia. Przykładem zastosowania takich badań może być prezentowanie spersonalizowanych reklam. Gdy na przykład dana drużyna przegrywa, jej kibicom można by wyświetlać po zakończonym meczu inny zestaw reklam niż kibicom drużyny przeciwnej. Firmy oferujące swoje usługi czy produkty na całym świecie mogą szybko reagować na opinie czy błędy zgłaszane przez sfrustrowanych internautów w serwisach społecznościowych poprawiając swój wizerunek i pokazując dbałość o klienta. Rządy czy partie polityczne mogą wykorzystać sieci społeczne, sentyment i geolokacje do odpowiednich zmian, ustaw dotyczących konkretnych grup społecznych, aby polepszyć wśród nich swoje notowania.

Praca pokazuje, że wykorzystanie mediów internetowych, czy serwisów społecznościowych może dać wymierne korzyści. Potwierdza, że świat wirtualny i realny przenikają się. Tak samo reagujemy na różne wydarzenia niezależnie od tego, czy dzielimy się swoimi opiniami z bliskimi będącymi obok nas czy z całym światem korzystając z serwisów społecznościowych. Jeśli więc zarządy firm lub organizacji nie wierzą lub wahają się nad sensem przeprowadzenia takich badań, to niniejsza praca może być dla nich dowodem, że warto bliżej przyjrzeć się zaprezentowanym tutaj aspektom. Może być więc punktem wyjścia do prowadzenia własnych badań na interesujące dany podmiot tematy.

6.3. Możliwe kierunki rozwoju

Zaprezentowana praca jest ukierunkowana na wąską dziedzinę – bada zachowania kibiców piłkarskich. W związku z tym, aby rozszerzyć jej zakres konieczne jest przeprowadzenie kilku działań celem jej rozwoju. Kilka możliwych kierunków to:

Rozszerzenie o inne języki naturalne

Zastosowany mechanizm analizy sentymentu jest skupiony jedynie na języku angielskim. Aby móc badać większe grupy ludzi koniecznym jest opracowanie techniki badającej również inne języki. Oczywiście w pewnym stopniu można wykorzystać podejście zastosowane w tej pracy, należy jednak mieć na uwadze różnice między budową używanych języków. Zupełnie inne konstrukcje językowe są w języku angielskim a zupełnie inne na przykład w języku polskim. Oczywiste jest więc, że nie można w taki sam sposób podejść do badania wpisów w różnych językach. Rozbudowanie mechanizmu o kolejne języki pozwoliłoby uzyskać bogatsze wyniki.

Odkrywanie tematów rozmów

Ciekawym rozszerzeniem badania sieci społecznych w kontekście przetwarzania języka naturalnego byłoby opracowanie i zastosowanie metody pozwalającej odkrywać tematy rozmów użytkowników. W zaprezentowanym podejściu badany jest jedynie sentyment wpisów, nie ma natomiast informacji na temat tego o czym dokładnie dyskutują użytkownicy Twittera. Użycie mechanizmu ekstrakcji tematów rozmów z wpisów z pewnością wzbogaciłoby analizę sieci społecznych.

Wzbogacenie techniki badania sentymentu

Badanie wydźwięku wypowiedzi zostało zautomatyzowane. Wykorzystuje do tego celu wpisy tworzone przez badaną grupę. Interesującym rozszerzeniem techniki badania sentymentu mogłoby być zastosowanie ręcznie stworzonego słownika. Mógłby on nieco skorygować niektóre wyniki analizy wydźwięku wypowiedzi i polepszyć ich jakość. Dodatkowo pozwoliłoby to na zastosowanie metody analizy sentymentu w szerszej dziedzinie badań. Aktualna metoda jest w pewien sposób zorientowana na środowisko piłkarskie i niekoniecznie dawałaby dobre rezultaty w innym zbiorze danych – na przykład wpisach politycznych. Skorzystanie ze słownika i zwiększenie zakresu badanych wpisów pozwoliłoby szerzej zastosować opracowaną metodę.

Bliższe przyjrzenie się grupom

Dodatkowym kierunkiem rozwoju mogło by być bliższe przyjrzenie się grupom, które się utworzyły. Można by na przykład skierować swoje badania na szukanie liderów tych grup, przyczyn ich pojawiania się i sposobów w jaki oddziałują na resztę grupy w zależności od sentymentu jaki sami generują, jaki generują dane grupy czy w jakiej lokalizacji geograficznej się znajdują.

Bibliografia

[1] J. A. Barnes, "Class and Committees in a Norwegian Island Parish," *Human Relations*, vol. 7, pp. 39–58, 1954.

[2] C. Aggarwal, "An Introduction to Social Network Data Analytics," in *Social Network Data Analytics* (C. Aggarwal, ed.), pp. 8–13, Springer, 2011.

[3] X. Shang and Y. Yuan, "Social network analysis in multiple social networks data for criminal group discovery," in *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, (Sanya), pp. 27–30, 2012 International Conference, 2012.

[4] D. Li, J. Li, Y. Tang, J. Zheng, and J. Chen, "The structure analysis of the cscwd conference's collaboration network," in *Computer Supported Cooperative Work in Design (CSCWD)*, (Wuhan), pp. 713–718, 2012 IEEE 16th International Conference, 2012.

[5] Y. Gao, C. Zhang, Y. Wang, and L. Sun, "A directed recommendation algorithm for user requests based on social networks," in *Embedded and Ubiquitous Computing (EUC)*, (Melbourne, VIC), pp. 457–462, 2011 IFIP 9th International Conference, 2011.

[6] Z. Zhang, K. Lee, H. Wang, D. Xuan, and H. Fang, "Epidemic control based on fused body sensed and social network information," in *Distributed Computing Systems Workshops (ICDCSW)*, (Macau), pp. 285–290, 2012 32nd International Conference, 2012.

[7] E. Estrada, "Introduction to network theory: Modern concepts, algorithms and application," tech. rep., Scottish Insight Institute, University of Strathclyde, 2008.

[8] D. Myers, *Social Psychology*. McGraw-Hill, 2009.

[9] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *Journal of Graph Algorithms and Applications*, vol. 10, no. 2, pp. 191–218, 2006.

[10] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, p. 066111, 2004.

[11] V. Blondel, J. Guillaume, R. Lambiotte, and E. Mech, "Fast unfolding of communities in large networks," *J. Stat. Mech*, p. P10008, 2008.

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

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

[14] R. Cole, L. Hirschman, L. Atlas, M. Beckman, *et al.*, "The challenge of spoken language systems: research directions for the nineties," *Speech and Audio Processing*, pp. 1–21, 2002.

[15] C. Surabhi, "Natural language processing future," in *Optical Imaging Sensor and Security (ICOSS)*, (Coimbatore), pp. 1–3, 2013 International Conference, 2013.

[16] R. Colbaugh and K. Glass, "Agile sentiment analysis of social media content for security informatics applications," in *Intelligence and Security Informatics Conference (EISIC)*, (Athens), pp. 327–331, 2011 European Intelligence and Security Informatics Conference, 2011.

[17] H. Chen, P. De, Y. Hu, and B. Hwang, "Sentiment revealed in social media and its effect on the stock market," in *Statistical Signal Processing Workshop (SSP)*, (Nice), pp. 25–28, 2011 IEEE, 2011.

[18] F. Nooralahzadeh, V. Arunachalam, and C. Chiru, "2012 presidential elections on twitter – an analysis of how the us and french election were reflected in tweets," in *Control Systems and Computer Science (CSCS)*, (Bucharest), pp. 240–246, 2013 19th International Conference, 2013.

[19] B. Sun and V. Ng, "Analyzing sentimental influence of posts on social networks," in *Computer Supported Cooperative Work in Design (CSCWD)*, (Hsinchu), pp. 546–551, 2014 IEEE 18th International Conference, 2014.

[20] A. Alahmadi, A. Joorabchi, and A. Mahdi, "A new text representation scheme combining bag-of-words and bag-of-concepts approaches for automatic text classification," in *GCC Conference and Exhibition (GCC)*, (Doha), pp. 108–113, 2013 7th IEEE, 2013.

[21] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[22] F. Zhou, F. Zhang, and G. Yang, "Graph-based text representation model and its realization," in *Natural Language Processing and Knowledge Engineering (NLP-KE)*, (Beijing), pp. 1–8, 2010 International Conference, 2010.

[23] A. Alahmadi, A. Joorabchi, and A. Mahdi, "A text network representation model," in *Fuzzy Systems and Knowledge Discovery*, (Jinan Shandong), pp. 150–154, FSKD '08. Fifth International Conference, 2008.

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

[25] K. Tomanek, "Analiza sentymentu – metoda analizy danych jakościowych. przykład zastosowania oraz ewaluacja słownika rid i metody klasyfikacji bayesa w analizie danych jakościowych," *Przegląd Socjologii Jakościowej*, vol. 10, no. 2, pp. 118–136, 2014.

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

[27] W. Reese and J. Beckland, "Lost in geolocation," tech. rep., Spring, 2011.

[28] A. Noulas, S. Scelatto, N. Lathia, and C. Mascolo, "A random walk around the city: New venue recommendation in location-based social networks," in *Privacy, Security, Risk and Trust (PASSAT)*, (Amsterdam), pp. 144–153, 2012 ASE/IEEE International Conference on Social Computing, 2012.

[29] M. Steurer and C. Trattner, "Acquaintance or partner? predicting partnership in online and location-based social networks," in *Advances in Social Networks Analysis and Mining (ASONAM)*, (Niagara Falls), pp. 372–379, 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013.

[30] Z. Wang, D. Zhang, X. Zhou, *et al.*, "Discovering and profiling overlapping communities in location-based social networks," *Systems, Man, and Cybernetics: Systems, IEEE Transactions*, vol. 44, no. 4, pp. 499–509, 2014.

[31] X. Qiao, J. Su, J. Zhang, and ohers, "Recommending friends instantly in location-based mobile social networks," *Communications, China*, vol. 11, no. 2, pp. 109–127, 2014.

[32] J. She, A. Vassilovski, and A. Hon, "What cuisine do you like? - improving dining preference prediction through physical social locations," in *Green Computing and Communications (GreenCom)*, (Besancon), pp. 454–457, 2012 IEEE International Conference on Green Computing and Communications, 2013.

[33] S. Kumar, F. Morstatter, and H. Liu, *Twitter Data Analytics*. New York, NY, USA: Springer, 2013.

[34] S. Rogers, "Insights into the #worldcup converation on twitter." `https://blog.twitter.com/2014/insights-into-the-worldcup-conversation-on-twitter`, July 2014. [Online: dostęp 17 lipca 2014].

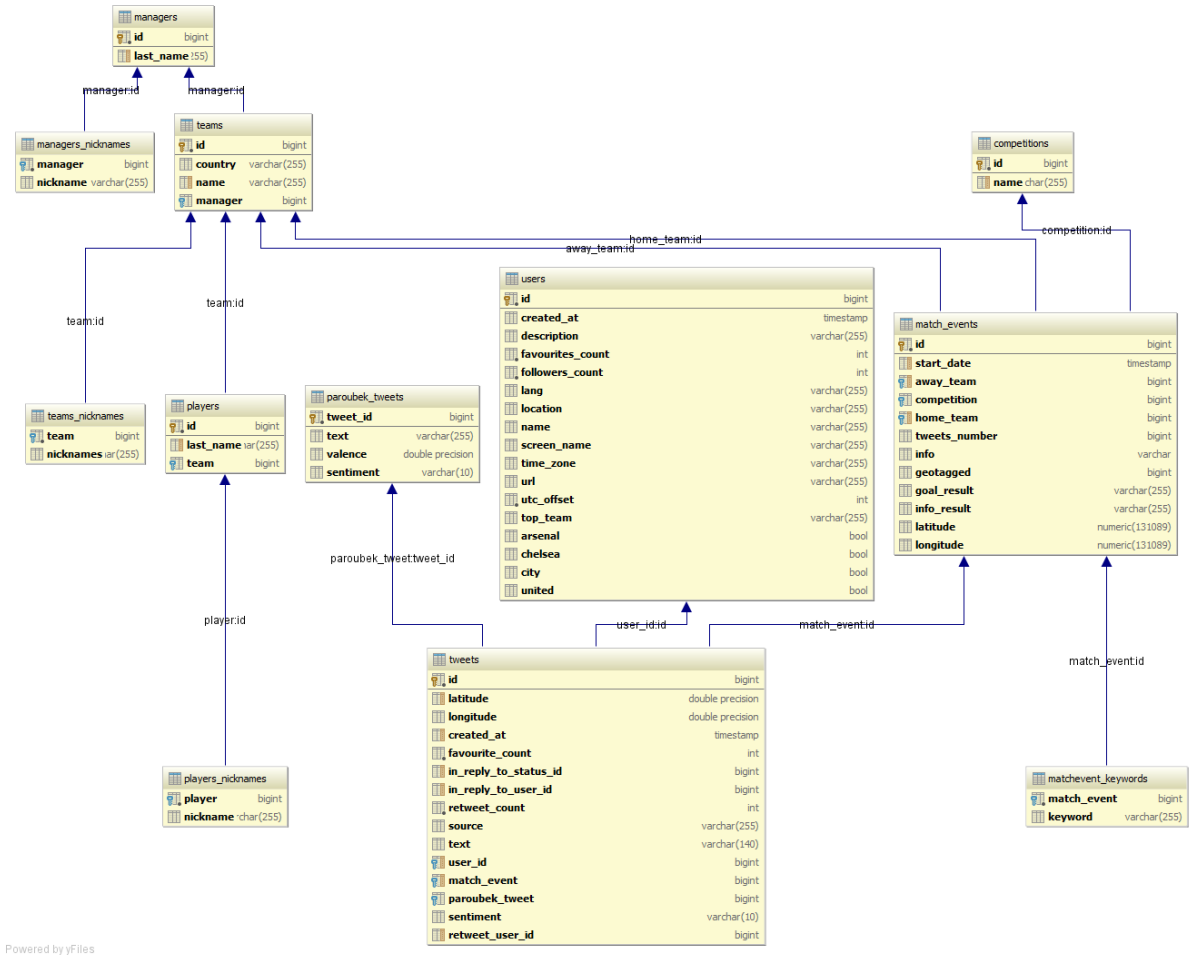
[35] "English stopwords." `http://www.webpageanalyse.com/dev/stopwords/en`, February 2014. [Online: dostęp 11 lutego 2014].

[36] D. Genetics, "Emoticon analysis in twitter." `http://datagenetics.com/blog/october52012/index.html`, 2012. [Online: dostęp 17 grudnia 2013].

[37] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, (Philadelphia, PA, USA), pp. 79–86, Association for Computational Linguistics, July 2002.

[38] A. Zwicky and G. Pullum, "Cliticization vs. inflection: English n't," *Language*, vol. 59, no. 3, 1983.

A. Opis tabel



Rysunek A.1: Schemat bazy danych

competitions rodzaje rozgrywek, w których brały udział drużyny

id – identyfikator wiersza

name – nazwa rozgrywek (np. Champions League, Liga Angielska)

managers nazwiska menadżerów (trenerów) drużyn

id – identyfikator wiersza

last_name – nazwisko menadżera

managers_nicknames popularne określenia, przezwiska menadżerów drużyn

manager – klucz obcy wskazujący rekord w tabeli *managers*

nickname – dodatkowe określenie menadżera

match_events nasłuchiwane spotkania między drużynami

id – identyfikator wiersza

start_date – data rozpoczęcia spotkania

away_team – identyfikator drużyny gości; klucz obcy do tabeli *teams*

competition – identyfikator rozgrywek; klucz obcy do tabeli *competitions*

home_team – identyfikator drużyny gospodarzy; klucz obcy do tabeli *teams*

tweets_number – liczba tweetów zebranych dla spotkania

info – krótki opis spotkania wraz z wynikiem

geotagged – liczba tweetów z geolokalizacją dla spotkania

goal_result – wynik meczu

info_result – krótki opis spotkania wraz z wynikiem

latitude – szerokość geograficzna miejsca meczu (stadionu)

longitude – długość geograficzna miejsca meczu (stadionu)

matchevent_keywords dodatkowe słowa kluczowe dla meczów (np. nazwa stadionu, nazwisko sędziego, etc.)

matchevent – klucz obcy wskazujący rekord w tabeli *match_events*

keyword – słowo kluczowe

paroubek_tweets wyniki wyliczania sentymentu tweetów

tweet_id – identyfikator tweeta; klucz obcy do tabeli *tweets*

text – oczyszczony tekst tweeta

valence – wartość *valence* (r. 2.2.3) algorytmu określania sentymentu

players zawodnicy drużyn

id – identyfikator wiersza

last_name – nazwisko zawodnika

team – identyfikator drużyny zawodnika; klucz obcy do tabeli *teams*

players_nicknames popularne przezwiska, określenia piłkarzy

player – identyfikator piłkarza; klucz obcy do tabeli *players*

nickname – określenie piłkarza

teams nasłuchiwane drużyny

id – identyfikator wiersza

country – kraj drużyny

name – nazwa klubu

manager – identyfikator menadżera; klucz obcy do tabeli *managers*

teams_nicknames popularne przezwiska, określenia drużyn

team – identyfikator drużyny; klucz obcy do tabeli *teams*

nicknames – określenie drużyny

tweets zebrane tweety

id – identyfikator wiersza

latitude – szerokość geograficzna wysłania tweeta

longitude – długość geograficzna wysłania tweeta

created_at – data utworzenia tweeta

favourite_count – liczba osób, które oznaczyły tweet jako ulubiony

in_reply_to_status_id – identyfikator tweeta, na który ten wpis jest odpowiedzią

in_reply_to_user_id – identyfikator użytkownika, na którego tweeta ten wpis jest odpowiedzią

retweet_count – liczba osób, które przekazały ten tweet dalej (ang. *retweet*)

source – źrodło wysłania tweeta (nazwa aplikacji)

text – tekst tweeta

user_id – identyfikator autora wpisu; klucz obcy do tabeli *users*

match_event – identyfikator spotkania, którego wpis dotyczy; klucz obcy do tabeli *match_events*

paroubek_tweet – identyfikator do rekordu z obliczoną wartością sentymentu (pokrywa się z kolumną *id* tej tabeli); klucz obcy do tabeli *paroubek_tweet*

sentiment – tekstowe określenie sentymentu tweeta, np. *POS* (pozytywny), *NEG* (negatywny), *NEU* (neturalny)

retweet_user_id – identyfikator użytkownika, którego tweet został podany dalej jako ten wpis

users użytkownicy Twittera będący autorami zebranych tweetów

id – identyfikator wiersza

created_at – data rejestracji użytkownika

description – opis użytkownika w serwisie Twitter

favourites_count – liczba osób, którego oznaczyły danego użytkownika jako ulubionego

followers_count – liczba osób śledzących

lang – główny język użytkownika

location – miejsce przebywania użytkownika (wpisywane ręcznie przez użytkownika)

name – nazwa użytkownika

screen_name – wyświetlana nazwa użytkownika

time_zone – strefa czasowa użytkownika

url – adres strony profilowej użytkownika w serwisie Twitter

utc_offset – różnica czasu między strefą czasową użytkownika a czasem UTC (ang. *Coordinated Universal Time* – uniwersalny czas koordynowany)
top_team – drużyna, której mecze użytkownik komentował najczęściej; klucz obcy do tabeli *teams*
arsenal – oznaczenie czy użytkownik jest zwolennikiem, czy przeciwnikiem Arsenalu (r. 3.2.1)
chelsea – oznaczenie czy użytkownik jest zwolennikiem, czy przeciwnikiem Chelsea
city – oznaczenie czy użytkownik jest zwolennikiem, czy przeciwnikiem Manchesteru City
united – oznaczenie czy użytkownik jest zwolennikiem, czy przeciwnikiem Manchesteru United