AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

WYDZIAŁ INFORMATYKI, ELEKTRONIKI I TELEKOMUNIKACJI

KATEDRA INFORMATYKI

PRACA DYPLOMOWA MAGISTERSKA

Automatyczna identyfikacja zachowań trollingu i spamowania na wybranych mediach społecznościowych

Automatic identification of trolling and spamming behavior on selected social media

Autor:	Łukasz Uchman
Kierunek studiów:	Informatyka
Typ studiów:	Stacjonarne
Opiekun pracy:	dr hab. inż. Jarosław Koźlak

Kraków, 2020

Oświadczenie studenta

Uprzedzony(-a) o odpowiedzialności karnej na podstawie art. 115 ust. 1 i 2 ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (t.j. Dz.U. z 2018 r. poz. 1191 z późn. zm.): „Kto przywłaszcza sobie autorstwo albo wprowadza w błąd co do autorstwa całości lub części cudzego utworu albo artystycznego wykonania, podlega grzywnie, karze ograniczenia wolności albo pozbawienia wolności do lat 3. Tej samej karze podlega, kto rozpowszechnia bez podania nazwiska lub pseudonimu twórcy cudzy utwór w wersji oryginalnej albo w postaci opracowania, artystyczne wykonanie albo publicznie zniekształca taki utwór, artystyczne wykonanie, fonogram, wideogram lub nadanie.", a także uprzedzony(-a) o odpowiedzialności dyscyplinarnej na podstawie art. 307 ust. 1 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2018 r. poz. 1668 z późn. zm.) „Student podlega odpowiedzialności dyscyplinarnej za naruszenie przepisów obowiązujących w uczelni oraz za czyn uchybiający godności studenta.", oświadczam, że niniejszą pracę dyplomową wykonałem(-am) osobiście i samodzielnie i nie korzystałem(-am) ze źródeł innych niż wymienione w pracy.

Jednocześnie Uczelnia informuje, że zgodnie z art. 15a ww. ustawy o prawie autorskim i prawach pokrewnych Uczelni przysługuje pierwszeństwo w opublikowaniu pracy dyplomowej studenta. Jeżeli Uczelnia nie opublikowała pracy dyplomowej w terminie 6 miesięcy od dnia jej obrony, autor może ją opublikować, chyba że praca jest częścią utworu zbiorowego. Ponadto Uczelnia jako podmiot, o którym mowa w art. 7 ust. 1 pkt 1 ustawy z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2018 r. poz. 1668 z późn. zm.), może korzystać bez wynagrodzenia i bez konieczności uzyskania zgody autora z utworu stworzonego przez studenta w wyniku wykonywania obowiązków związanych z odbywaniem studiów, udostępniać utwór ministrowi właściwemu do spraw szkolnictwa wyższego i nauki oraz korzystać z utworów znajdujących się w prowadzonych przez niego bazach danych, w celu sprawdzania z wykorzystaniem systemu antyplagiatowego. Minister właściwy do spraw szkolnictwa wyższego i nauki może korzystać z prac dyplomowych znajdujących się w prowadzonych przez niego bazach danych w zakresie niezbędnym do zapewnienia prawidłowego utrzymania i rozwoju tych baz oraz współpracujących z nimi systemów informatycznych.

……

(czytelny podpis studenta)

Spis treści

Spis ilustracji

Spis tabel

Spis równań

1. Wstęp

1.1. Motywacja

Współczesny rozwój technologiczny daje nam możliwości, o których w dawnych czasach nie mogliśmy nawet marzyć. W przeszłości ludzie mieli kontakt tylko z osobami ze swojego najbliższego otoczenia. Można powiedzieć, że świat dla przeciętnego człowieka był bardzo duży. Żył on w swojej lokalnej społeczności, która nie musiała liczyć wiele osób - większość jej członków mogła się wzajemnie znać. Poprzez rozwój transportu ludzie mogli migrować na stałe lub czasowo. Przełożyło się na to na zwiększenie sieci znajomości przeciętnego człowieka. Ułatwiony został kontakt między ludźmi, których dzieli znaczna odległość np. innych ras, co pozwoliło na wymianę kulturową.

Następnym krokiem rozwoju były osiągnięcia telekomunikacji pozwalające na bezpośredni kontakt ludzi bez konieczności podróżowania lub wysyłania listów. Telegram lub nawet współczesny telefon, nie zapewnia jednak możliwości prowadzenia dyskusji między grupą ludzi naraz. W większości przypadków jest to narzędzie komunikacji między dwoma osobami, które w jakimś stopniu się znają, są świadome tego kim jest ich rozmówca.

W tym miejscu możemy przejść do technologii, która ma współcześnie kluczową rolę w zwiększeniu interakcji między ludźmi - Internetu. Jego początki przypadają na koniec lat 60. XX wieku. Pierwsza strona internetowa powstała w roku 1991, w którym również miały miejsce uruchomienia pierwszych połączeń internetowych w Polsce. Jego szybki rozwój umożliwił na bezproblemową komunikację ludzi z prawie każdego miejsca na Ziemi. Z raportu Digital Report na początek 2019 roku [5] wynika, że liczba internautów w tamtym momencie wynosiła około 4.4 miliarda, co stanowi około 57% populacji - przyrost o 4% względem poprzedniego roku. Ponadto, około 45% mieszkańców naszej planety używa mediów społecznościowych.

Internet dla większości ludzi staje się niezbędny do funkcjonowania we współczesnym świecie. Interakcja międzyludzka przenosi się z codziennego życia do sieci. Informacje w Internecie mają duży wpływ na wyniki wyborów czy rynki finansowe. Powstanie wielu portali informacyjnych, for dyskusyjnych i portali społecznościowych pozwoliło na bardzo łatwą możliwość wymiany zdań na dowolny temat. Wydawałoby się, że ma to same zalety. Niestety musimy zwrócić uwagę na to, że w sieci każdy może być anonimowy. Prowadzi to do częstych "patologii". Ludzi przestają obowiązywać granice, które respektują w codziennym życiu. Znacznie obniża to poziom kultury konwersacji i prowadzi do występowania negatywnych zjawisk takich jak trolling i spam, które są przedmiotem badania tej pracy. Trolling to antyspołeczne zachowanie polegające na świadomym uprzykrzaniu życia innym dyskutującym osobom, a spamem można nazwać publikowanie dużej liczby podobnych wiadomości, które są niechciane przez odbiorców.

1.2. Pytanie badawcze

Portale społecznościowe takie jak Twitter czy fora dyskusyjne portali informacyjnych takich jak ABC News zmagają się z problemami trollingu i spamu. Niepożądane wiadomości łamiące regulaminy portali nie dość, że nie wnoszą nic do rozmowy, to w dodatku bardzo

często mają frustrujący wpływ na innych użytkowników. Manualne kasowanie nieregulaminowych postów jest niemożliwe przy takiej skali problemu. Wymagane jest opracowanie rozwiązań pozwalających na automatyzację tego procesu. Możemy postawić pytanie: **Czy możliwe jest wybranie zestawu cech pozwalających na identyfikację zachowań trollingu i spamowania w obrębie portali społecznościowych z wysoką skutecznością?** Mówiąc o wysokiej skuteczności mamy na myśli taką, aby wykryte wiadomości i użytkownicy mogli być usuwani bez udziału osób nadzorujących.

1.3. Hipoteza badawcza

Pierwszym krokiem w celu walki z niepożądanymi zachowaniami musi być analiza tych zachowań. Przyjmujemy hipotezę: **Poprzez zbadanie zachowań niepożądanych użytkowników portalu społecznościowego możliwe jest wybranie zestawu cech, które pozwolą na ich identyfikację w całej społeczności portalu**. Mamy wiele źródeł danych, z których można pozyskać cechy. Skupiając się na relacjach użytkownika z innymi, możemy przykładowo badać jego pozycję w społeczności, relacje innych na publikowane przez niego treści, sieć jego znajomości, częstość współudziału w tematach dla poszczególnych użytkowników, częstość wypowiedzi i zwyczaje obserwowanego użytkownika. Z drugiej strony cechy mogą być związane z analizą samej treści wiadomości. Można zwrócić uwagę na czytelność, wartość merytoryczną czy długość publikowanych treści. Możemy zwracać uwagę na liczbę linków w wiadomościach lub wyłapywać kluczowe słowa. Kluczowa może być analiza sentymentu wypowiedzi, która ma na celu określić charakter wpisu: negatywny, neutralny, pozytywny. Może być to uzyskane z wykorzystaniem słowników i stałych wzorców lub z użyciem metod uczenia maszynowego.

1.4. Cele pracy

Celem pracy jest **dokonanie analizy zachowań użytkowników portalu społecznościowego i na jej podstawie opracowanie oraz zbadanie cech identyfikujących niepożądanych użytkowników (trolli i spamerów), z których zostaną zaproponowane różne zestawy pozwalające na stworzenie skutecznych klasyfikatorów.** Kluczowymi elementami pracy jest wybór odpowiednich cech, ich analiza oraz wyciągnięcie poprawnych wniosków z analizy. Pozwoli to na utworzenie zestawu danych z wartościami współczynników cech, potrzebnego w uczeniu maszynowym. Bardzo ważny jest również wybór odpowiedniego źródła danych do badania. Wiele portali społecznościowych udostępnia API pozwalające na pobranie informacji o konwersacjach lub użytkownikach. Takie dane wymagają jednak czasochłonnego opracowania. Lepszym wyborem może być znalezienie już gotowego i opisanego publicznie udostępnionego zestawu danych. Rodzaj danych ma znaczący wpływ na cechy poddawane badaniu. Do klasyfikacji zostaną użyte różne algorytmy, których część będzie pokrywać się z wyróżniającymi się algorytmami użytymi w innych pracach. Zestawienie wyników klasyfikatorów stworzonych różnymi podejściami pozwoli na wybranie najbardziej przydatnych w badanym problemie oraz porównanie czy są zgodne z innymi rozwiązaniami.

1.5. Streszczenie pracy

Pierwszym krokiem pracy jest analiza dotychczasowych podejść i rozwiązań badanego problemu, a zarazem głębsze zapoznanie się z ideami, celami oraz zwyczajami niepożądanych użytkowników, których staramy się identyfikować. Analizowane w dotychczasowych rozwiązaniach cechy niepożądanych użytkowników możemy podzielić na dwie grupy: cech bazowych i oryginalnych. Cechy bazowe, z reguły bardzo proste, wykorzystywane są przez wszystkie rozwiązania. Są one bazą, do której w różnych podejściach dobudowane są autorskie, oryginalne rozwiązania, które są przedstawicielami drugiej grupy. Do cech bazowych możemy zaliczyć między innymi liczbę słów w wiadomościach, liczbę linków, polubień, wiadomości, stosunek obserwujących i obserwowanych na Twitterze lub znajomych na innych portalach. Do grupy oryginalnych podejść zaliczają się grafowe analizy użytkowników odpowiadających sobie na wiadomości, analiza odpowiedzi na badaną wiadomość w celu określenia reakcji na nią innych użytkowników, analiza intencji i zamiarów użytkownika, różne analizy czasu pisania wiadomości czy analiza stron internetowych, do których prowadzą linki. W kwestii budowy klasyfikatora, najlepsze wyniki osiągane były z wykorzystaniem algorytmu lasów losowych oraz naiwnego algorytmu bayesowskiego.

Do analizy został wybrany Twitter. Wykorzystywane są dwa zbiory: niepożądanych i normalnych użytkowników. Pierwszym z nich jest zbiór IRA zawierający wiadomości rosyjskich trolli z czasu kampanii prezydenckiej w Stanach Zjednoczonych z 2016 roku. Udostępniany on jest publicznie przez Twittera i liczy prawie 2 miliony wiadomości, w tym 600 tysięcy w analizowanym języku angielskim. Drugi zbiór, został stworzony z wykorzystaniem bazy Harvard Dataverse. Dostarczyła ona pliki tekstowe zawierające ID wiadomości opublikowanych w czasie tej samej kampanii wyborczej. Ponad 11 milionów wiadomości zostało pobranych z wykorzystaniem Twitter API. Oba zbiory zostały zapisane w relacyjnej bazie danych i poddane drobnym korektom i uzupełnieniom.

Stworzone środowisko analizy i przetwarzania danych składa się z pięciu głównych elementów, do których zaliczają się: relacyjna baza danych, moduł przygotowania danych służący do ich pobierania do bazy oraz uzupełniania, modułu skupiającego i udostępniającego implementacje cech, modułu przeprowadzania analiz, z którego pochodzą wszystkie dane zwizualizowane w tabelach i na diagramach, modułu tworzenia zbiorów testowo-treningowych pozwalającego na stworzenie zbioru wiadomości, jego skalowanie z zachowaniem proporcji klastrów, filtrowanie i scalanie zbiorów oraz z ostatniego modułu klasyfikacji i badania rozwiązań, w którym badane są zbiory cech oraz ich zdolności klasyfikacyjne na wcześniej opracowanych zbiorach. Wszystkie moduły zostały stworzone w języku Python.

Opracowane cechy można podzielić na trzy kategorie ze względu na źródło ich pochodzenia: pochodzące z analizy treści pisanych wiadomości, analizy zachowań i statystyk użytkowników oraz analizy zależności między użytkownikami. Do wyróżniających się analiz należą: analiza emocji z wykorzystaniem biblioteki trenowanej na ponad miliardzie wiadomości z Twittera, analiza częstości pisania wiadomości w postaci wyszukiwania najdłuższych serii publikacji użytkownika, analiza grafów retweetujących się użytkowników pod kątem liczby obecności danego użytkownika w dowolnych maksymalnych klikach oraz sam rozmiar maksymalnej kliki dla jego samego, analizy związane z kategoryzacją źródeł

publikacji wiadomości oraz kategoryzacją linków zawartych w wiadomościach. Wszystkie cechy zostały przebadane pod kątem rozkładu ich wartości dla obu zbiorów użytkowników, co zostało przedstawione na diagramach i odpowiadającym im wykresach.

Zbiór testowo treningowy został utworzony ze wszystkich dostępnych wiadomości ze zbioru niepożądanych użytkowników. W przypadku o wiele większego zbioru potencjalnie normalnych użytkowników na początku należało odfiltrować z niego część użytkowników, którzy nie powinni się w nim znajdować poprzez wykorzystanie klastrowania metodą k-means oraz manualną analizę, która była możliwa ze względu na małą liczebność większości klastrów. Zbiór normalnych użytkowników został następnie przeskalowany, tak aby jego liczebność była taka sama jak niepożądanych użytkowników. Przed skalowaniem dokonano klastrowania metodą k-means, a samo skalowanie odbyło się w taki sposób, aby zachować proporcje klastrów w przeskalowanym zbiorze. Finalny zbiór testowo-treningowy liczy 1 191 646 wiadomości z proporcją normalnych i niepożądanych użytkowników równą 1:1.

Zostały uwzględnione dokładnie 34 cechy. 11 z nich zalicza się do analizy emocji oraz 2 do analizy sentymentu. Cechy zostały przebadane pod kątem indywidualnych zdolności klasyfikacyjnych algorytmem RandomForest z 10-krotną kros-walidacji. Najlepsze wyniki zostały osiągnięte przez cechy oparte o liczby obserwujących i obserwowanych. Dobrze wypada również analiza emocji, w której uwzględniamy wszystkie wchodzące w jej skład cechy. Wnosi ona wyraźnie więcej niż analiza sentymentu, która osiągnęła poprawne, lecz niższe wyniki. Bardzo dobre rezultaty osiągane są również przez kategoryzację źródła publikacji wiadomości oraz kategoryzację ulubionego źródła publikacji użytkownika. W identyczny sposób zostały przebadane różne grupy cech, między innymi dotyczące 3 kategorii źródeł ich pochodzenia. Najgorszej wypadły cechy związane z analizą tekstu osiągając wynik miary F1 równy 0,92, wyraźnie lepiej cechy z kategorii analizy zachowań i statystyk osiągając 0,98 F1, a najlepiej cechy związane z analizą zależności przekraczając próg 0,99. Na wynik cech dotyczących analizy zależności miała wpływ obecność w nich liczb obserwujących i obserwowanych, które jak już wcześniej wspomniano indywidualnie wypadały najlepiej.

Na podstawie analiz z wykorzystaniem korelacji Pearsona, metody Lasso, zachłannej metody eliminacji najmniej wnoszących cech w kolejnych krokach oraz analizy zgodności klasyfikacji cech opracowano finalne zbiory. Ich wyniki zostały ze sobą zestawione. Podzbiór składający się z 8 cech wybranych zachłanną metodą eliminacji osiągnął najlepszy wynik F1 wynoszący 0,9997 na zbiorze liczącym prawie 1,2 miliona wiadomości. 4 cechy wybrane metodą Lasso osiągnęły wynik 0,995. Zbiór stworzony po ponownym przeprowadzeniu metody Lasso na cechach odrzuconych za pierwszym razem, osiągnął wynik 0,986. Zbiór składający się z 12 najgorszych indywidualnie cech o F1 poniżej 0,7 osiągnął wynik 0,883. Bardzo duży wpływ na wyniki mają liczby obserwujących i obserwowanych. Gdy weźmiemy pod uwagę zbiór wszystkich cech z ich wyłączeniem, osiągany jest wynik 0,992. Osiągane wyniki klasyfikacji są bardzo wysokie, co może budzić podejrzenia co do badanego zbioru danych. Najprawdopodobniej znajduje się w nim mało trudnych do wykrycia niepożądanych użytkowników. W kwestii wyników różnych klasyfikatorów potwierdziła się widoczna w innych pracach zasadność użycia drzew decyzyjnych osiągających najlepsze wyniki. Nieznacznie gorsza okazała się klasyfikacja z wykorzystaniem

algorytmu k-najbliższych sąsiadów oraz AdaBoost. W przeciwieństwie do niektórych prac, które pokazały, że naiwny algorytm bayesowski może przynosić zbliżone wyniki do drzew decyzyjnych, a czasami nawet lepsze, nie znalazło to potwierdzenia w tym przypadku. Wyniki naiwnego algorytmu bayesowskiego były wyraźnie gorsze.

Do celów pracy należała analiza zachowań użytkowników portalu społecznościowego i na jej podstawie opracowanie zestawu cech, który pozwoli na stworzenie klasyfikatora identyfikującego niepożądanych użytkowników, czyli trolli i spamerów. Postawione cele zostały zrealizowane. Stworzone zbiory cech zawierające wiele własnych podejść, pozwoliły na osiągnięcie zadowalających wyników. Nie znaczy to jednak, że praca nie może być dalej rozwijana przez usprawnienie aktualnych cech i implementację nowych lub przystosowanie środowiska do analizy innego portalu z wykorzystaniem kompatybilnych z nim cech.

1.6. Organizacja tekstu pracy

Rozdział 2 zawiera wprowadzenie w tematykę pracy. Pozwala on na zapoznanie się z podstawowymi pojęciami takimi jak portal społecznościowy, trolling czy spam. Opisuje również kluczowe miary oceny rozwiązań oraz wykorzystywane algorytmy klasyfikacji. W rozdziale 3 zostaną opisane wyróżniające się aspekty dotychczasowych osiągnięć w badanym problemie. Zostanie on podsumowany wyciągnięciem wniosków. Rozdział 4 zawierać będzie propozycję własnego zestawu cech, a w 5 zostaną opisane badane zbiory danych oraz system, w którym będą one przetwarzane i analizowane. Sama analiza każdej z zaproponowanych cech zostanie przedstawiona w punkcie 6. Wyniki osiągnięte z wykorzystaniem różnych algorytmów oraz zestawów cech zostaną przedstawione w rozdziale 7. Ostatni rozdział 8 dokonuje podsumowania prac oraz opisuje wyciągnięte wnioski.

2. Wprowadzenie w tematykę pracy

W rozdziale zostaną zdefiniowane kluczowe pojęcia z puntu widzenia pracy takie jak portal społecznościowy, zjawisko trollingu oraz spamowania. Opisane zostaną również proste metryki, które posłużą w celu oceny rozwiązań oraz algorytmy klasyfikacji, które zostaną wykorzystane.

2.1. Media społecznościowe

Według słownika języka polskiego PWN media społecznościowe można zdefiniować jako "technologie internetowe i mobilne, umożliwiające kontakt pomiędzy użytkownikami poprzez wymianę informacji, opinii i wiedzy" [4]. Jest to bardzo ogólna definicja nie zagłębiająca nas w szczegóły pojęcia. W artykule autorstwa A.M. Kaplan'a i M. Haenlein'a [35] znajdujemy bardziej dokładną definicję mediów społecznościowych. Autorzy definiują je jako grupę aplikacji internetowych, które zbudowane są na bazie założeń ideologicznych i technologicznych koncepcji sieci Web 2.0, i przez to pozwalają na tworzenie i wymianę treści generowanej przez użytkowników. Definicja zawiera dwa pojęcia, które mogą być niezrozumiałe. Pierwsze z nich, Web 2.0 jest pojęciem określającym serwisy internetowe, w których tworzeniu dużą rolę odgrywają sami użytkownicy. Kiedyś użytkownicy nie mieli wpływu na treści prezentowane przez aplikację. Odpowiedzialni za to byli głównie właściciele serwisu. W Web 2.0 każdy użytkownik może współuczestniczyć w rozwoju treści portalu. Drugie pojęcie jest mocno związane z pierwszym. Treści generowane przez użytkowników to szerokie pojęcie zawierające w sobie wszystko to co jest tworzone i publikowane przez użytkowników w Internecie, między innymi tekst, muzykę czy obrazy.

Media społecznościowe według [35] możemy skategoryzować, ze względu na rolę aplikacji na następujące grupy:

- współtworzone projekty takie jak encyklopedie internetowe, posiadające możliwość edycji i dodawania haseł,
- blogi, które są najstarszą z kategorii,
- społeczności związane z danym typem treści. Przykładem może być portal YouTube związany z filmami lub portal SlideShare do publikowania prezentacji,
- sieci społecznościowe takie jak Facebook czy MySpace,
- wirtualne światy gier internetowych pozwalające na stworzenie własnej wirtualnej postaci i interakcję z innymi,
- wirtualne światy społecznościowe, które w przeciwieństwie do gier, starają się odtworzyć rzeczywisty świat.

2.2. Twitter jako przykład portalu społecznościowego

Twitter jest jednym z przykładów popularnej platformy społecznościowej. Został założony w 2006 roku. Z wyników finansowych Twittera na 2 kwartał 2019 roku wynika,

że portal ma około 139 milionów aktywnych użytkowników piszących dziennie w granicach 500 milionów wiadomości. Liczby te ciągle wzrastają. 21% kont zarejestrowanych na platformie pochodzi ze Stanów Zjednoczonych, które są największym źródłem dochodu firmy. W naszym kraju internauci nie są do niej aż tak przekonani i nie odnosi ona takiego sukcesu wśród zwykłych użytkowników jak chociażby Facebook. W Polsce głównie korzystają z niego osoby publiczne takie jak politycy czy publicyści. Twitter może być skategoryzowany również jako serwis informacyjny. Jest on jednak tworzony przez samych użytkowników.

Wiadomości na Twitterze nazywamy "tweetami". Tweety mogą zwierać odnośniki i obrazki. Mają ograniczoną długość - aktualnie jest to limit 280 znaków, lecz do roku 2017 wynosił on 140. W wiadomościach znajdują się również znaki specjalne "#" oraz "@". Znak kratki oznacza tag, do którego przypisywana jest wiadomość. Znak małpy służy do odpowiadania na tweety innych użytkowników: "@nazwa-użytkownika". Opublikowana wiadomość jest widoczna na profilu autora oraz jest prezentowana wszystkim użytkownikom, którzy go obserwują. Każdy tweet może zostać skomentowany, podany dalej (nazywane jest to "retweetem") lub polubiony.

Dyskusje na Twitterze mogą obrać dowolny temat lecz w większości mają charakter informacyjny lub polityczny. Portal jest popularny wśród osób publicznych takich jak politycy, publicyści czy dziennikarze. Swoje profile mają również celebryci oraz znane osobistości takie jak papież Franciszek czy były prezydent Stanów Zjednoczonych Barack Obama. Na podstawie danych z artykułu [32] 31% użytkowników to kobiety. Z Twittera najczęściej korzystają ludzie młodsi. 31% użytkowników na świecie znajduje się w grupie wiekowej 18-24, a 27,3% w grupie 25-34. 12% użytkowników wykorzystuje Twittera jako źródło codziennych informacji.

Twitter dostępny jest zarówno w wersji przeznaczonej na komputery jak i wersji mobilnej. Jak wynika ze statystyk [32] w około 80% przypadków używana jest wersja przeznaczona na telefony.

2.3. Badane zjawiska

Spam i trolling są pojęciami znanymi od dawna. Zostały zdefiniowane w wielu źródłach. Cechy obu zjawisk ciągle ewoluują i są często zależne od środowiska ich występowania.

2.3.1. Definicje badanych zjawisk

Spam

Encyklopedia PWN [4] definiuje spam jako "reklamy i niechciane wiadomości w poczcie elektronicznej lub artykuły wieloadresowe wysyłane na wiele grup dyskusyjnych".

Pojęciem spamu nazywamy niepożądane lub niepotrzebne treści elektroniczne rozsyłane do dużej ilości odbiorców bez ich zgody. Wiadomości, które są spamem są bezwartościowe dla odbiorców, w dodatku bardzo często mogą być dla nich zagrożeniem. Spam w większości przypadków ma charakter reklamowy, lecz może również służyć wyłudzeniom lub atakom hakerskim. Wiadomości mogą zachęcać do wejścia w niebezpieczne linki, lub w przypadku e-maili być próbą podszycia się i wyłudzenia informacji. Korzyść jest zawsze jednostronna - zyskuje nadawca wiadomości. Pierwszy spam odnotowano w 1978 roku.

Jego celem było zaproszenie na urodziny, a grupa, która otrzymała wiadomość liczyła około 1000 osób. W tym samym roku został rozesłany również pierwszy spam reklamowy. Aktualnie większość wiadomości będących spamem generowana jest przez boty.

Spamowanie nie jest legalne w wielu kodeksach prawnych. W Unii Europejskiej zakazane jest rozsyłanie wszelkich form spamu w postaci reklamy, lecz wiadomości o charakterze społeczno-politycznym są już legalne. W przypadku prawa polskiego zakazane są wszelkie formy spamu. Rozpowszechnianie niezamówionych informacji handlowych podlega karze grzywny i jest ścigane na wniosek poszkodowanego, czyli odbiorcy wiadomości.

Spam występuje w wielu miejscach. Z początku najczęściej spotykany był w poczcie elektronicznej, lecz wraz z rozwojem Internetu można doświadczyć go na komunikatorach internetowych, blogach, forach dyskusyjnych, portalach społecznościowych i informacyjnych z możliwością komentowania artykułów. Walka ze spamem spowodowała rozwój różnych technik stosowanych przez spamerów mających na celu zamaskowanie swojej działalności. Można to szczególnie zaobserwować na portalach społecznościowych, w tym na Twitterze.

Trolling

Zjawisko było wielokrotnie badane. Artykuł C. Hardaker [30] przedstawiający badania na temat trollingu definiuje trolla jako użytkownika, który tworzy pozorną tożsamość szczerej osoby o dobrych intencjach w celu wzięcia udziału w dyskusji i stworzenia w niej konfliktów wyłącznie w celach własnej rozrywki. Autorzy artykułu [33] opracowujący narzędzie do detekcji trollingu, definiują go jako fenomen obserwowany w relacjach pomiędzy użytkownikami Internetu, nastawiony na uzyskanie "mocnych" odpowiedzi na publikowane ofensywne i emocjonujące treści od jak największego grona osób.

Trollingiem nazywamy antyspołeczne zachowanie, które polega na świadomym uprzykrzaniu życia innym osobom biorącym udział w dyskusji. Trolling najczęściej występuje na forach dyskusyjnych, blogach, portalach społecznościowych, portalach tematycznych. Jednym z głównych celów trollingu jest prowokowanie, skłócanie i zabawa innymi użytkownikami. Jest to osiągane przez trolli poprzez publikowanie kontrowersyjnych, często obraźliwych i wyzywających wypowiedzi mających skupić na sobie uwagę. W przypadku sukcesu trolla, prowadzi to do powstania nowego, nic nie wnoszącego wątku w dyskusji, który w znaczącym stopniu może ją zdezorganizować. Im lepszy trolling tym więcej osób bierze udział w dyskusji. Pojęcie wywodzi się z wędkarstwa, w pewnym sensie troll zastawia pułapkę w postaci wiadomości, a następnie "łowi" swoje ofiary irytując je i wciągając do dyskusji. Cała idea trollingu opiera się na podtrzymywaniu złych emocji odbiorców, którzy jeszcze nie zrozumieli, że dalsza konwersacja nie ma sensu.

Artykuł [39] identyfikuje dwa typy trolli w badanych portalach informacyjnych: trolla klasycznego i hybrydowego. Troll hybrydowy różni się od zwykłego następującymi cechami: intensywnie publikuje te same wiadomości, powtarza wiadomości, które były już publikowane z innych kont lub adresów IP, tworzy wiadomości zawierające te same informacje i linki. Troll hybrydowy ma określony cel i zysk z tego co robi. Często jego działania mają charakter polityczny.

W większości przypadków troll poza własną satysfakcją nie osiąga więcej korzyści ze swoich działań. Zjawisko to jest w pełni nieszkodliwe jeśli wiadomości tego typu są

ignorowane. W przeciwieństwie do spamu, trolling nie jest zautomatyzowany, a opracowanie metod jego wykrywanie jest trudniejsze. Trolling nie jest związany jedynie z samymi wiadomościami. Osoby zajmujące się trollingiem często pracują nad wykreowaniem profilów na portalach społecznościowych, które pozwolą zwiększyć ich wiarygodność. Jeden troll może posługiwać się wieloma kontami. Klasyfikacja wiadomości jest trudniejsza niż w przypadku spamu, w którym wiadomości były bardziej jednoznaczne, ponieważ troll często ukrywa swoje rzeczywiste zamiary.

2.3.2. Cechy badanych zjawisk

Cechy zjawisk opisane są z punktu widzenia potencjalnego użytkownika portalu społecznościowego lub tematycznego. Pokazują po czym można poznać niepożądane wiadomości.

Spam

Wiadomości rozsyłane są na duża skalę. Są identyczne lub nieznacznie różniące się między sobą. W przypadku konwersacji nie są związane z jej tematem. Najczęściej są reklamą, zachęcają do wejścia w linki. Wiadomości posiadają wiele wyróżniających się cech związanych ze swoją zawartością: mogą być nieczytelne, zawierać dużą ilość dużych liter, być niestandardowo długie, zawierać charakterystyczne dla spamu słowa lub wiele linków.

Bardzo dużo cech wiadomości zawierających spam jest powiązanych z platformą, na której występują. Przykładem jest Twitter, na którym użycie dużej ilości nadmiarowych tagów lub odniesień do innych użytkowników może być kryterium klasyfikującym wiadomość.

Trolling

Wiadomości będące trollingiem są często obraźliwe i prowokujące. Wiele trolli jednak próbuje maskować swoją działalność i działać w dyplomatyczny sposób. Należy zwrócić uwagę czy wiadomość odnosi się do emocji odbiorców, jaki jest jej sentyment - czy nastawienie autora jest pozytywne czy negatywne? Gdy główny temat wiadomości jest różny od tematu dyskusji, wiadomość może być już podejrzana o trolling.

Tak samo jak w przypadku spamu, wiele cech powiązanych jest z miejscem publikacji wiadomości.

2.4. Miary ocen i metody analizy rozwiązań

Zostaną przedstawione podstawowe miary pozwalające na ocenę rozwiązań. Zostaną one wykorzystane w rozdziale poświęconym badaniu osiągniętych wyników.

2.4.1. Tablica pomyłek

Tablica pomyłek (*Confusion matrix*) [38] - jest stosowana w przypadku klasyfikacji binarnych. W postaci tablicy zestawia ona liczbę prób klasyfikacyjnych, które zakończyły się pomyślną i błędną klasyfikacją danych. W dodatku próby te są podzielone na przypadki,

w których dane były pozytywne albo negatywne. Poniżej w tabeli 1 znajduje się przykład takiej tablicy.

Tabela 1: Tablica pomyłek.

Data Class	Classified as positive	Classified as negative
positive	true positive (t_p)	false negative (f_n)
negative	false positive (f_p)	true negative (t_n)

gdzie:
t_p - liczba skutecznie rozpoznanych jako przynależące do klasy
t_n - liczba skutecznie rozpoznanych jako nieprzynależące do klasy
f_p - liczba błędnie rozpoznanych jako przynależące do klasy
f_n - liczba przynależących do klasy, których nie udało się wykryć

2.4.2. Dokładność

Dokładność (*Accuracy*) [38] - jest jedną z najprostszych i najbardziej intuicyjnych miar, wyraża się ją w procentach. Jest to stosunek liczby dobrze dokonanych klasyfikacji do liczby wszystkich prób pomnożona przez 100. Przekładając to na składowe tablicy pomyłek otrzymujemy wzór

$$Accuracy = \frac{t_p \ + \ t_n}{t_p \ + \ f_p \ + \ t_n \ + \ f_n} \ * \ 100. \tag{1}$$

Dokładność posiada jednak dużą wadę, która jest bardzo widoczna, gdy zbiór jest niezrównoważony. Możemy wtedy uzyskać dużą dokładność bez osiągnięcia wysokiej skuteczności.

2.4.3. Czułość

Czułość (*Recall*) [38] - to efektywność klasyfikatora w identyfikowaniu danych z pozytywnymi etykietami. Pokazuje ile obserwacji zostało pominiętych dla danej klasy. Czułość można zobrazować jako prawdopodobieństwo wykrycia choroby u osoby chorej:

$$Recall = \frac{t_p}{t_p \ + \ f_n}, \tag{2}$$

gdzie t_p, f_n to odpowiednio *True Positive i False Negative*. Czułość nie zwraca uwagi na liczbę *True Negative*(t_n).

2.4.4. Precyzja

Precyzja (*Precision*) [38] - metryka podobna do czułości, pokazująca ile z danych zakwalifikowanych jako pozytywne rzeczywiście nimi jest. Można ją zobrazować jako proporcja

liczby pacjentów sklasyfikowanych jako chorych do ilości naprawdę chorych. Opisywana jest następującym wzorem:

$$Precision = \frac{t_p}{t_p \ + \ f_p}, \tag{3}$$

gdzie t_p, f_p to odpowiednio *True Positive i False Positive*. Precyzja nie zwraca uwagi na liczbę *True Negative*(t_n).

2.4.5. F1

F1 [26] - jest to miara uwzględniająca równomiernie wyniki dwóch innych miar: precyzji i czułości. Wynik z zakresu od 0 do 1 obliczany jest na podstawie średniej harmonicznej:

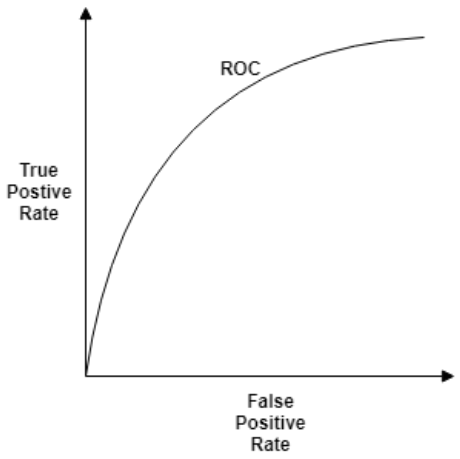
$$F1 = 2 * \frac{recall \ * \ precision}{recall \ + \ precision} = \frac{2t_p}{2t_p \ + \ f_p \ + \ f_n}, \tag{4}$$

gdzie t_p, f_p, f_n to odpowiednio *True Positive, False Positive i False Negative.*

Miara koncentruje się na klasie dodatniej. Negatywne cechy są zdewaluowane w porównaniu do pozytywnych.

2.4.6. ROC/AUC

AUC (*Area Under the Curve*) [19] [38] - określa zdolność klasyfikatora do unikania nieprawidłowych klasyfikacji. AUC to pole powierzchni pod krzywą ROC (Receiver Operating Characteristic). Krzywa ROC przedstawiona na wykresie 1 jest funkcją przedstawiającą zależność True Positive Rate od False Positive Rate.



Rysunek 1: Krzywa ROC obrazująca zależność między True Positive Rate, a False Positive Rate. AUC, czyli pole pod wykresem krzywej ROC określa zdolność klasyfikatora do unikania pomyłek. Krzywe ROC mogą służyć do porównywania klasyfikatorów. (rysunek na podstawie [19])

Im bardziej wykres jest wypukły tym klasyfikator jest lepszy. Krzywe ROC mogą służyć do porównywania różnych klasyfikatorów. AUC to pole powierzchni pod krzywą, które może zostać opisane wzorem

$$AUC = \frac{1}{2} * (\frac{t_p}{t_p \ + \ f_n} + \frac{t_n}{t_n \ + \ f_p}), \tag{5}$$

gdzie t_p, t_n, f_p, f_n to odpowiednio *True Positive, True Negative, False Positive i False Negative.*

W przeciwieństwie do precyzji i czułości AUC nie jest niezmienny przy zmianie prawidłowo sklasyfikowanych negatywnych przykładów i lepiej wykryje możliwe zmiany w klasie negatywnej niż miara F1.

2.5. Różne algorytmy klasyfikacji

Poniżej zostaną opisane algorytmy klasyfikacji, które będą wykorzystane do budowy klasyfikatorów. Dostarczone są podstawowe informacje o algorytmach tak, aby można było zrozumieć ideę ich działania. Więcej szczegółów można znaleźć w źródłach, na podstawie których bazują stworzone opisy.

2.5.1. CART

CART (*Classification and Regression Trees*) [44] jest binarną rekurencyjną procedurą partycjonowania. Tworzone jest drzewo binarne aż do osiągnięcia maksymalnego rozmiaru lub założonych warunków stopu (preferowany jest brak stopu) z kryterium podziału w wersji domyślnej opartym na zanieczyszczeniu Giniego (Gini impurity), a następnie drzewo jest przycinane. Przycięciu w danym kroku podlega podział, który ma najmniejszy wpływ na ogólną wydajność drzewa na danych uczących. Algorytm CART nie generuje jedynie jednego drzewa, lecz sekwencję zagnieżdżonych przyciętych drzew, z których każde jest kandydatem na optymalne drzewo. Algorytm nie ma żadnej miary oceny rozwiązań na podstawie danych treningowych. Badana jest predykcyjna zdolność każdego drzewa z sekwencji na niezależnych danych testowych (możliwe jest wykorzystanie kros-walidacji). Wybierane jest drzewo osiągające najlepszy wynik.

2.5.2. Lasy losowe

Lasy losowe (*Random Forest*) [41] podejmują decyzje klasyfikacyjne na podstawie większości głosów zwracanych przez N stworzonych drzew decyzyjnych. Budowanie drzew można opisać w trzech krokach: krok wybrania próby bootstrap z danych treningowych, zbudowania drzewa o maksymalnej wysokości dla wybranych danych dokonując decyzji o podziale w danym wierzchołku na bazie losowo wybranego podzbioru cech obiektów, powtórzenia wcześniejszych kroków aż do chwili, gdy zbudowane zostanie N drzew. Dzięki losowaniu podzbiorów cech przy podziałach unikana jest korelacja pomiędzy drzewami. Algorytm lasów losowych, mimo że buduje N drzew decyzyjnych, nie jest słaby wydajnościowo. Jego

wydajność bardzo mocno poprawiana jest poprzez wcześniej przytoczone decydowanie o podziałach na podstawie podzbiorów cech oraz brak przycinania stworzonych drzew.

2.5.3. Extra Tree

Algorytm Extra Tree [28] podobnie jak lasy losowe podejmuje decyzje na bazie wielu drzew decyzyjnych. Sposób budowy drzew różni się jednak dwiema ważnymi decyzjami. W przeciwieństwie do lasów losowych każde drzewo jest budowane na pełnym zbiorze danych, a nie na jego próbie bootstrap. Ponadto zamiast obliczać najbardziej optymalny punkt odcięcia dla każdej rozważanej cechy, wybór dokonywany jest losowo. Wykonywany jest podział węzła, który ze wszystkich losowo wygenerowanych odnosi najlepszy wynik.

2.5.4. Naiwny algorytm bayesowski

Naiwny algorytm bayesowski [44] jest prostym algorytmem, który nie wymaga dużych nakładów obliczeniowych, co za tym idzie możemy go wykorzystać na dużym zbiorze danych. Klasyfikator opiera się o regułę Bayesa wyrażającą prawdopodobieństwo tego, że obiekt należy do określonej klasy oraz założenie, że warunkowe prawdopodobieństwa dla cech opisujących ten obiekt są wzajemnie niezależne. N-wymiarowy problem wielowymiarowy jest redukowany do N jednowymiarowych problemów estymacji. Szacowanie jednowymiarowe jest znacznie prostsze i szybsze. Algorytm jest często używany do klasyfikacji tekstów oraz filtrowania spamu.

2.5.5. AdaBoost

AdaBoost [44] jest przykładem z grupy algorytmów łączących współpracę wielu klasyfikatorów w celu osiągnięcia jak najlepszego lepszego wyniku (tak zwane ensemble methods). Algorytm rozpoczynany jest od budowy bazowego prostego modelu opartego na danych treningowych z jednakowymi wagami. Następnie model jest testowany, a wagi obiektów, które zostały źle zaklasyfikowane są zwiększane. W kolejnej iteracji budowany jest nowy klasyfikator, który dzięki zwiększonym wagom zwróci większą uwagę na poprzednio popełnione błędy. Proces wykonywany jest N razy, co prowadzi do stworzenia N klasyfikatorów z różnymi wagami przydzielanych w oparciu o ich przydatność. Wynik klasyfikacji oparty jest na większości ważonej decyzji każdego z N klasyfikatorów. Algorytm jest podatny na zaszumione dane i odstające obserwacje.

2.5.6. K-najbliższych sąsiadów

Algorytm K-najbliższych sąsiadów [44] jest jednym z prostszych algorytmów, które mogą być wykorzystane w klasyfikacji. Jego zasada działania jest bardzo prosta w zrozumieniu. Mamy dane z przypisanymi klasami, dane do sklasyfikowania oraz metrykę do obliczania odległości. Możemy to przedstawić jako n-wymiarową przestrzeń, gdzie każdy element to punkt opisywany przez n cech. Gdy chcemy dokonać klasyfikacji punktu szukamy K najbliższych punktów w jego otoczeniu. Przydzielamy mu klasę, do której należy większość punktów ze znalezionej grupy sąsiadów. Należy zwrócić uwagę na dobranie odpowiedniej wartości parametru K oraz odpowiedniej metryki. Zarówno wybranie małej, jak i dużej

wartości parametru K może mieć negatywny wpływ na wyniki. Istnieją również ulepszenia metody dotyczące uwzględniania wag najbliższych sąsiadów w postaci ich odległości przy wyborze klasy dla przydzielanego punktu.

2.5.7. MLP

MLP (*Multi-layer perceptron*) [34] jest to popularny typ sieci neuronowej. Sieć składa się z warstwy wejścia, k warstw ukrytych oraz warstwy wyjścia. Od liczby warstw ukrytych zależy stopień skomplikowania sieci. Zadaniem sieci jest aproksymacja pewnej funkcji, przykładowo funkcji klasyfikacji przekształcającej dane wejściowe na kategorię klasyfikacji. Trenowanie modelu odbywa się w trzech krokach: podaniu danych do modelu, kalkulacji błędu oraz wprowadzeniu poprawek i cofnięciu się do kroku pierwszego.

2.5.8. Regresja logistyczna

Regresja logistyczna [20] jest statystyczną metodą klasyfikacji binarnych klas. Wykorzystywana jest w niej funkcja sigmoidalna, do której jako argument podajemy wyliczoną wartość funkcji zawierającej ilorazy wartości cech z ich współczynnikami. Współczynniki muszą zostać dobrane w taki sposób, aby jak najbardziej odwzorować podział między klasami na danych uczących. Funkcja sigmoidalna zwraca prawdopodobieństwo przynależności obiektu do klasy. Przykładowo jeśli zwrócone prawdopodobieństwo jest mniejsze od 0,5 badany użytkownik jest normalny, a gdy wynosi minimum 0,5 oznacza to, że jest niepożądanym użytkownikiem.

3. Analiza dotychczasowych dokonań w identyfikacji niepożądanych użytkowników

Dokonana zostanie analiza różnych dotychczasowych badań o podobnej tematyce. Wyniki przeglądu dziedziny podzielone są na kategorie. Ostatni punkt rozdziału zawiera wnioski, które zostały wyciągnięte z analizy.

3.1. Portale społecznościowe poddawane analizie

Większość prac o podobnej tematyce zajmowało się badaniem serwisu Twitter. Powodem tego mogło być to, że portal ten udostępnia specjalne API, które pozwala na bardzo łatwe pobieranie wiadomości oraz informacji o kontach autorów. Dane są anonimizowane. Ponadto Twitter sam udostępnia niektóre zbiory danych. Przykładem może być zbiór związany z wyborami prezydenckimi w Stanach Zjednoczonych z 2016 roku, który jest badany między innymi przez B. Ghanem'a, D. Buscaldi'ego i P. Ross'a w pracy pod tytułem "Textrolls" [29]. Twórcy jednego z artykułów [40], nawiązali nawet współpracę z Twitterem. Wysłali listę wykrytych niepożądanych użytkowników do Twittera, od którego otrzymali informacje na temat blokad zgłoszonych kont.

Analizie mogą również podlegać dane udostępniane poprzez Disqus API. Disqus to platforma hostingowa czatów, używana przez rozległą liczbę portali i blogów. Wykorzystana została w pracy "Trollspot: Detectingmisbehavior in commenting platforms." [36], w której badaniu podlegały następujące serwisy: CNBC News, ABC News, Bloomberg Views oraz kanał dyskusyjny Disqus'a Breaking News. Disqus posiada system zgłaszania wiadomości, który może pomóc w tworzeniu zbioru badawczego.

Inne popularne portale społecznościowe lub dyskusyjne takie jak Facebook czy Reddit nie są tak często wybierane. Prace, które odnoszą się do lokalnych portali z językiem innym niż angielski są marginalne.

3.2. Opracowane zbiory danych

Najwięcej zbiorów danych dotyczy Twittera, wykorzystuje Twitter API do pozyskania danych i skupia się na języku angielskim. Z wyróżniających się prac w innych językach można wspomnieć o pracach [27] i [33], w których badano kolejno zbiory danych w języku hiszpańskim i fińskim.

W kwestii rozmiarów etykietowanych zbiorów danych można zaobserwować dwa podejścia zależne od przeznaczenia zbioru oraz typu portalu. Możemy etykietować użytkowników lub wiadomości. Pierwsze podejście jest częściej wykorzystywane. Przeciętnie liczba otagowanych użytkowników waha się pomiędzy 400 i 1000. We wcześniej wspomnianej pracy wykorzystującej zbiór danych o wyborach prezydenckich w USA [29] liczba otagowanych kont samych niepożądanych użytkowników wyniosła 2023. Sam zbiór został wcześniej przefiltrowany pod kątem użytego języka - ograniczony tylko do angielskiego. Użycie zbiorów opublikowanych w internecie pozwala na znaczne oszczędzenie czasu co widzimy w tym przypadku. Aby stworzyć rozległy zbiór potrzebne jest zaangażowanie dużej grupy ludzi. Przykładem mogą być działania w pracy "Detecting spammers on Twitter"

[18], w której stworzony zbiór danych zawierający aż 8207 etykietowanych użytkowników, powstał z pomocą wielu wolontariuszy. W przypadku drugiego podejścia spotykane jest ono zwłaszcza w kontekście analizy portali informacyjnych lub tematycznych gdzie komentarze nie są ściśle powiązane z kontami. Jest ono lepsze jeśli skupiamy się na zawartości wiadomości, a nie na ogólnej działalności podejrzanego użytkownika. Podejście to można zaobserwować w przypadku pracy zajmującej się hiszpańskim portalem z wiadomościami społecznymi o nazwie Men'eame [22] gdzie oznaczono zbiór składający się z 9044 wiadomości.

Do wyróżniających się zbiorów możemy zaliczyć zbiór opracowany przez autora pracy "Determining Trolling in Textual Comments" [23]. Wiadomości i odpowiedzi na nie były kategoryzowane oraz odpowiednio etykietowane. Autorzy opracowali więc zbiór danych pozwalający na dodanie nowych kryteriów oceny badanej wiadomości, dotyczącej odpowiedzi na nią.

Oryginalny sposób pozyskiwania danych zaprezentowali autorzy pracy opisanej w artykule "Detecting spammers on social networks" [40]. Stworzyli trzy różne zbiory badawcze dotyczące Facebooka, Twittera oraz MySpace. Na każdym z portali założyli określoną liczbę kont i używali je w sposób "bierny". Sami nie wysyłali żadnych zaproszeń czy wiadomości, jedynie odczytywali przychodzące i akceptowali otrzymane zaproszenia. Proces zbierania wiadomości i kont użytkowników z wymienionych portali trwał około rok.

3.3. Cechy zachowań użytkowników używane do ich klasyfikacji

Cechy używane w klasyfikacji trolli i spamerów możemy podzielić na dwie kategorie. Jedna związana jest z analizą tekstu, a druga z badaniem zależności między użytkownikami oraz badaniem ich zwyczajów.

Do wykrywania spamerów najczęściej spotykane są cechy związane z ilością URL w wiadomościach. Są one interpretowane na różny sposób: liczba linków w stosunku do liczby słów we wszystkich wiadomościach użytkownika, liczba odnośników w stosunku do liczby wiadomości, średnia liczba odnośników na wiadomość. W artykule [42] opisano system, który sprawdza URLe pod kątem zawartości, do której prowadzą. Rozwiązanie przeznaczone jest do wykrywania spamu i niepożądanych linków między innymi na portalach społecznościowych. Klasyfikują źródła linków na podstawie danych uzyskiwanych z przeglądarki, analizy DNS oraz adresów IP.

Wiele prac w analizie tekstu stosuje słowniki i leksykony. Przykładem jest praca TexTrolls [29], w której wykorzystywany jest leksykon NRC do wykrywania emocji oraz sentymentu wypowiedzi. Uwzględniana jest również moralność wypowiedzi ("Morality foundation theory"), stronniczość oraz kultura. Osiągnięto to z wykorzystaniem słowników kategoryzujących słowa. Używano między innymi słownik lingwistyczny LIWC. W pracy Textrolls najwyższą skuteczność miała moralność i cecha określająca postawę autora analizowanej wiadomości - to czy jest on za, czy przeciw tematowi dyskusji.

Gdy analizie poddajemy konta użytkownika wraz z wiadomościami, badane jest podobieństwo wiadomości użytkownika. Niekoniecznie wszystkich, może być to przykładowo ostatnie 20 wiadomości. A.H. Wang w swojej pracy [43] do badania duplikacji tweetów wykorzystał odległość Levensteina. Jest to miara odmienności skończonych ciągów zna-

ków stosowana między innymi w rozpoznawaniu plagiatów. W tej samej pracy widzimy wykorzystanie liczby twitterowych odniesień do tematów i innych użytkowników - zliczane są specjalne znaki "#" i "@" w wiadomościach. Jest to również bardzo częste w innych pracach dotyczących Twittera.

We wspominanej wcześniej pracy "Determining Trolling in Textual Comments" [23] źródłem cech są również odpowiedzi na badane wiadomości. Z każdej odpowiedzi starano się wyłonić parametr określający jak odpowiadająca osoba zinterpretowała zamiary potencjalnego trolla oraz jej reakcję na tą wiadomość. Ponadto z badanej wiadomości starano się określić intencje oraz to, czy autor próbuje ukryć swoje realne zamiary. Wymienione cechy nie przyniosły jednak zbyt dobrych rezultatów, co zostanie opisane w kolejnym podpunkcie pracy.

Artykuł "TrollSpot: Detecting misbehavior in commenting platforms" [36] prezentuje jedno z możliwych podejść do analizy zachowań użytkowników portali społecznościowych. Użytkownicy, którzy zbyt często współwystępują w tych samych dyskusjach są podejrzani. Budowany jest graf, w którym węzłami są użytkownicy, a łączące ich krawędzie oznaczają współuczestnictwo w co najmniej N tematach. Następnie graf jest analizowany, m.in. poszukiwane są kliki. W tej samej pracy wprowadzono cechę nazwaną "bardzo aktywne godziny" (*"Highly Active Hours"*). Jest to minimalna liczba godzin, w których użytkownik pisze połowę swoich dziennych postów. Dla 96,7% użytkowników badanego portalu wartość cechy była mniejsza niż 4h. W pracy TrollSpot podjęto się również próby klasyfikacji niepożądanych użytkowników w oparciu o stworzone grupy użytkowników, składające się z osób o podobnych zachowaniach. Przykładowymi cechami decydującymi o przydziale do grupy były m.in. godziny aktywnego korzystania z portali, stopień współpracy z innymi uczestnikami konwersacji, czy czytelność komentarzy. Najpierw użytkownik był kategoryzowany do odpowiedniej mu grupy, a następnie grupa wskazywała na prawdopodobny typ klasyfikowanego użytkownika. Dwustopniowa kategoryzacja przyniosła nieznacznie gorsze rezultaty. Z cech dotyczących tekstu oprócz czytelności i ilości odnośników autorzy analizowali długość wiadomości otrzymując wynik wskazujący, że tylko 9% przekraczało 100 wyrazów. Jest to mocna sugestia tego, że był to potencjalny spam.

Dla Twittera powszechnie używaną metryką jest liczba osób obserwujących konto podejrzanego oraz liczba kont, które on sam obserwuje. W przypadku niepożądanych użytkowników stosunek tych liczb jest bardzo nieproporcjonalny. Prace, które zajmują się portalami informacyjnymi takimi jak Men'eame [22] dodatkowo mogą badać, czy komentarze pod artykułem są powiązane z jego tematem. Jeśli są różne to wysoce prawdopodobne jest wystąpienie trollingu lub spamu. W przypadku Facebooka, można określić wiarygodność profilu poprzez liczbę jego znajomych, liczbę zdjęć i polubień pod nimi.

3.4. Wykorzystane algorytmy

W wielu pracach badana była większa liczba klasyfikatorów, a następnie efekty były zestawiane w celu wybrania najlepszego. Biblioteką do uczenia maszynowego, z której zastosowaniem można się najczęściej spotkać jest biblioteka WEKA stworzona na Uniwersytecie Waikato w Nowej Zelandii.

Studiując prace z podobnej dziedziny możemy wyróżnić dwie najczęściej używane grupy metod do klasyfikacji: drzewa decyzyjne oraz klasyfikatory bayesowskie. W przypadku drzew decyzyjnych najczęściej stosowane są lasy losowe i algorytm C4.5. W drugiej grupie często wykorzystywany jest naiwny klasyfikator bayesowski. Szczególną uwagę w kwestii zestawiania metod należy zwrócić na artykuły "Filtering Trolling Comments through Collective Classification" [22] oraz "Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying" [27]. Ich autorzy dokonują szczegółowych porównań wyników klasyfikacji wielu algorytmów. Wyniki z obu źródeł potwierdzają, że stosowanie drzew decyzyjnych i klasyfikatorów bayesowskich w rozpoznawaniu spamerów i trolli owocuje lepszymi rezultatami. Pierwszy [22] z wymienionych artykułów porównuje trzy różne podejścia do klasyfikacji bayesowskiej: naive, tree augmented naive i K2. Najczęściej stosowany algorytm naiwny wypada najgorzej z podanej trójki.

A.H. Wang w artykule "Don't follow me - spam detection in Twitter." [43] wybrał naiwny klasyfikator bayesowski i uzasadnia swój wybór. Klasyfikator bayesowski jest odporny na szumy w danych, relacja między zbiorem cech użytych w uczeniu oraz spamem jest niedeterministyczna. Nawet w przypadku bliskiego podobieństwa niektórych cech z przykładami szkoleniowymi nie można dokonać jednoznacznej klasyfikacji. Stosując klasyfikator bayesowski wyrażamy przynależności w formie prawdopodobieństw. Drugim argumentem autora jest to, że nie możemy klasyfikować każdego użytkownika w oparciu o te same reguły. Możemy dla każdego obliczać prawdopodobieństwo spamu na podstawie indywidualnych zachowań.

3.5. Jakość wypracowanych rozwiązań

Wyniki osiągnięte przez różne rozwiązania nie mogą być ze sobą mocno porównywane. Wynika to z różnych zbiorów danych, z wykorzystaniem których były opracowywane oraz badane. Tematy badań prac również się od siebie różniły. Niektóre skupiały się wyłącznie na spamerach, inne na trollach lub obu grupach razem.

Nowatorskie podejście uwzględniające interpretacje i reakcje odpowiadających na badaną wiadomość zaproponowane przez Luisa Mojica de La Vegę [23], nie osiągnęło zbyt dobrych rezultatów. Wyniki są nieznacznie wyższe od rozwiązań trywialnych. Nie jest to jeszcze jednak finalna wersja narzędzia, dalsze prace wciąż trwają. Sama metoda wydaje się zbyt skomplikowana i wymaga dobrej kategoryzacji sporej liczby czynników.

W wielu badaniach testowano również skuteczność użycia poszczególnych cech. W przypadku badań dotyczących portalu Mene'ame [22], obliczano współczynnik wzmocnienia informacji (*information gain*) związany z każdą cechą - gdy był on bliski zeru, cecha nie była dalej używana. Z kolei artykuł "Textrolls" [29] prezentuje informacje o wynikach klasyfikacji każdej cechy z osobna, co pozwala na porównanie ich skuteczności.

Wcześniej wspominany projekt, którego autorzy współpracowali z Twitterem zaowocował stworzeniem rozwiązania, które z grupy 135 834 profili twitterowych wykryło 15 932 spamerów. Wszyscy zostali zgłoszeni. Okazało się, że 75 kont zostało zakwalifikowanych niepoprawnie. Pozostałe profile zostały zbanowane. Precyzja wyniosła zatem około 0,995.

Na 100 losowo wybranych użytkowników nie będących spamerami, sześciu z nich zostało zakwalifikowanych niepoprawnie.

Rozwiązanie zaprezentowane w artykule “Filtering Trolling Comments through Collective Classification” [22] z wykorzystaniem tradycyjnych podejść do uczenia osiągnęło z wykorzystaniem najlepszego algorytmu precyzję około 76.88%, AUC 0.68. Podejście z wykorzystaniem klasyfikacji zbiorowej (*collective classification*) dla zbioru danych otagowanego tylko w 75% osiągnęło precyzję 76.94% oraz AUC równe 0.67.

Artykuł rozwiązujący problem hiszpańskojęzycznych trolli na Twitterze [27] prezentuje bardzo dobre wyniki. W zależności od metody, AUC zawiera się w przedziale od 0.89 dla klasyfikacji z użyciem algorytmu K-najbliższych sąsiadów z 2-krotną walidacją krzyżową do 0.96 dla metody SMO-PolyKernel. Tak dobre wyniki mogą być rezultatem bardzo małego zbioru danych, składającego się z zaledwie 19 kont i 1900 tweetów.

Dobre wyniki osiągnięto w pracy “Trollspot: Detectingmisbehavior in commenting platforms.” [36]. Z wykorzystaniem lasów losowych i 10-krotnej walidacji krzyżowej udało osiągnąć się AUC równe 0.77, precyzję 86.3% oraz czułość 65.5%.

3.6. Wnioski wyciągnięte z dokonanego przeglądu prac

Większość przeanalizowanych prac w dziedzinie wykorzystywała dużą grupę takich samych cech do klasyfikacji niepożądanych użytkowników. Do tej grupy należały między innymi: liczby URLi w wiadomościach, długość wiadomości, liczba wiadomości i podobieństwo wiadomości, w przypadku Twittera liczba obserwujących i obserwowanych oraz odniesień do innych użytkowników w wiadomościach. Można powiedzieć, że powtarzające się cechy stanowią bazę, od której można zacząć. Pozwalają one wychwycić anomalie widoczne u niepożądanych użytkowników, którzy nie próbują się zbytnio kryć ze swoimi zamiarami. Następnie do bazowego zestawu cech twórcy poszczególnych rozwiązań dokładają swoje autorskie pomysły.

Jednym z wyróżniających się podejść jest to zaproponowane w artykule [23]. Twórcy pracy jako jedyni przywiązują dużą wagę do analizy odpowiedzi na podejrzane wiadomości. Wprowadzają jednak przez to dużą komplikację problemu oraz uzależniają się od innych trudnych klasyfikacji. Projekt jest nadal kontynuowany i jest to bardzo możliwe, że w przyszłości uzyska o wiele lepsze rezultaty. Na aktualny moment przedstawione wyniki nie mogą być pozytywnie oceniane. Można jednak z poczynionych prac wyciągnąć wniosek, że odpowiedzi i reakcje innych na wiadomości potencjalnego spamera lub trolla mogą być cennym źródłem informacji. Mniej skomplikowane cechy z tej kategorii mogą przynieść dobre wyniki.

Wiele prac do analizy aktywności użytkownika wylicza liczbę jego dziennych postów lub liczbę postów w danym temacie. Rozwiązanie [36] przedstawia o wiele lepsze podejście nazwane “bardzo aktywnymi godzinami” - czas, w którym użytkownik pisze 50% swoich dziennych postów. W żadnym podejściu nie jest jednak uwzględniana długość wiadomości. Wszystkie cechy mają na celu wychwycić użytkowników, którzy spędzają nienaturalnie dużo czasu na portalu. Napisanie 100 krótkich komentarzy może zająć tyle samo czasu co 30 bardziej rozległych wiadomości. Można by było oceniać dzienną aktywność po skali wytworzonej przez użytkownika treści wiadomości.

Tylko jedno z rozwiązań [36] badało zależności między użytkownikami z wykorzystaniem grafów. Zajmowało się ono analizą for portali informacyjnych. W przypadku prac badających Twittera jedyną formą badania zależności między użytkownikami była analiza ilości obserwowanych i obserwujących lub zliczanie ilości odniesień do innych użytkowników w wiadomościach. Nikt nie próbował szukać użytkowników, którzy podejrzanie często biorą udział w tych samych dyskusjach na Twitterze.

W kwestii długości wiadomości wszystkie prace nie uwzględniają tego, że pojęcie długości wiadomości może się różnić w kontekście tematu dyskusji. Jest to zrozumiałe w kontekście Twittera, gdzie długość tweeta jest ograniczona, lecz w przypadku portali bez tego limitu można by zastosować inne podejście. W przypadku wystąpienia merytorycznej i długiej odpowiedzi pozostałe odpowiedzi na nią również mogą być długie. Nie powinniśmy oceniać rozmiaru wiadomości na podstawie średniej wszystkich wiadomości w danym portalu, lecz na podstawie innych wiadomości z tego samego tematu.

W artykule [29] wykorzystano zbiór danych IRA, w którym wiele uczestników rozmów było nieanglojęzycznymi trollami. Cecha związana z NLI (*Native language identification*) dawała bardzo dobre rezultaty. W przypadku rozwiązania uniwersalnego, nie przyniosłaby ona jednak korzyści, a ogólny wynik klasyfikacji byłby gorszy. Nie zmienia to jednak faktu, że przedstawione narzędzie było bardzo dobre. W przypadku innego artykułu [43] w przypadku wielu cech analizowane jest tylko 20 ostatnich postów użytkownika. Pozwala to na skuteczne ukrycie się niepożądanych trolli i spamerów, którzy nie są bardzo radykalni w swoich działaniach, a ich zamiary nie są w pełni oczywiste. Niepożądany użytkownik może przyjąć technikę polegającą na pisaniu postów zgodnych i niezgodnych z regulaminem w odpowiednich proporcjach, aby nie wzbudzać podejrzeń.

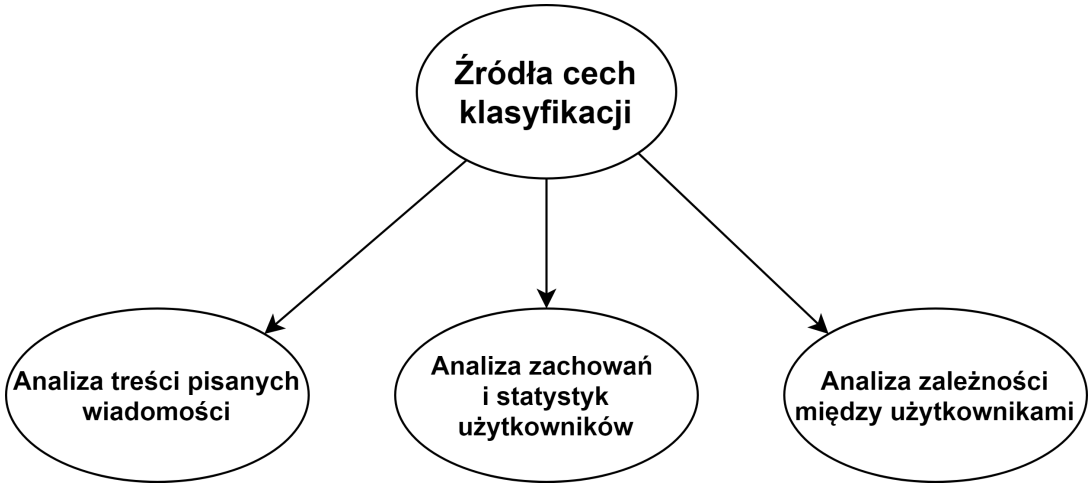
Najczęściej stosowane algorytmy uczenia maszynowego to lasy losowe oraz naiwny algorytm bayesowski. Jest to potwierdzone osiąganymi przez nie wynikami zamieszczonymi w wielu pracach z tej dziedziny. Autorzy często bezpośrednio zestawiali ze sobą różne podejścia i na tej podstawie wyłaniali najlepsze rozwiązania.

4. Koncepcja opracowania zestawu cech pozwalającego na klasyfikację użytkowników portali społecznościowych

Rozdział zawiera opis cech wybranych do identyfikacji niepożądanych użytkowników: trolli oraz spamerów. Cechy podzielone są na trzy kategorie ze względu na źródło pozyskania. Z każdą cechą powiązany jest opis jej interpretacji.

4.1. Podział źródeł cech klasyfikacji użytkowników

Aby pogrupować cechy i zarazem podejść do problemu z wielu stron, został stworzony podział zobrazowany na rysunku 2. Będzie on wykorzystywany w dalszych punktach pracy.



Rysunek 2: Rysunek przedstawia źródła cech klasyfikacji użytkowników portalu społecznościowego. Wyróżnia trzy źródła, z których możliwe jest pozyskanie danych. Cechy zawarte w każdej kategorii mogą samodzielnie być skuteczne przy identyfikacji trolli i spamerów. W przypadku połączenia cech z różnych kategorii jesteśmy w stanie otrzymać najbardziej optymalne rozwiązanie badanego problemu, pokrywające swą analizą jak najbardziej rozległą grupę niepożądanych użytkowników. Rysunek został wykonany z użyciem narzędzia draw.io [2].

Pochodzenie cech, które można wykorzystać do klasyfikacji użytkowników na normalnych i niepożądanych można podzielić na trzy grupy: pochodzące z analizy tekstu, analizy zachowań i statystyk użytkowników oraz analizy zależności między użytkownikami. Podział ten jest uniwersalny dla większości portali społecznościowych. Niektóre portale pozwalają na lepszą eksplorację cech z drugiej i trzeciej kategorii. Przykładem może być Twitter, którego mechanizm retweetów pozwala na inne zobrazowanie zależności między użytkownikami niż odpowiedzi na wiadomości innych użytkowników, które obecne są na wszystkich portalach społecznościowych. W kwestii analizy tekstu praktycznie wszystkie portale społecznościowe będą mieć wspólne cechy klasyfikujące. Różnica może wystąpić jedynie w szukaniu charakterystycznych znaków specjalnych w tekście takich jak tagi lub referencje do innych użytkowników.

Zbiór cech z każdego źródła z osobna przy wystarczająco dużym nakładzie pracy powinien poradzić sobie dobrze w wykrywaniu niepożądanych użytkowników. Z pewnością, wyniki osiągane przez różne grupy źródeł cech różniłyby się w zależności od portalu. Ukierunkowywanie się jednak na analizę pojedynczej kategorii jest rozwiązaniem, które wymaga zbyt dużo pracy, aby osiągnąć wysokie wyniki. Najlepszym podejściem jest stworzenie rozwiązania hybrydowego, składającego się z cech pochodzących ze wszystkich kategorii. Możemy takim sposobem szybko osiągnąć dobre rezultaty, a następnie zająć się doprecyzowywaniem poszczególnych kategorii cech.

4.2. Analiza treści pisanych wiadomości

Tekst wiadomości niesie bardzo wiele informacji. Można wyróżnić proste cechy dotyczące zliczania różnych elementów tekstu, jak i bardziej skomplikowane badające sentyment, emocje czy podobieństwo. Jest to kategoria wykorzystująca najbardziej zaawansowane mechanizmy, które ze względu na ich trudność niekoniecznie muszą przynieść najlepsze wyniki.

4.2.1. Linki w wiadomościach (URL)

Można się spodziewać, że spamerzy będą o wiele częściej używać w swoich wiadomościach odnośników. Mogą być to wiadomości reklamowe, zawierające odnośniki do wydarzeń społecznościowych czy mało popularnych portali. Z drugiej strony, zwykli użytkownicy mogą publikować odnośniki do artykułów powiązanych z ich wypowiedzią. Liczba linków w przypadku niepożądanych użytkowników powinna być jednak większa. Bardzo podejrzane powinny być wiadomości zawierające, więcej niż dwa odnośniki. Są one zwykle bardzo rzadko spotykane.

4.2.2. Liczba hasztagów i odniesień do użytkowników

Znak # na analizowanym portalu Twitter oznacza tag wiadomości, a @ to odniesienie do innego użytkownika portalu. Wystarczy zliczyć wystąpienia znaków specjalnych w tweecie. Dla potencjalnego spamera jest ważne, aby wiadomość miała jak największy zasięg. Osiąga to przez dodawanie wielu tagów lub referencji do jak największej ilości użytkowników. Wiadomość zawierająca dużo znaków specjalnych może więc wskazywać na potencjalnego niepożądanego użytkownika.

4.2.3. Długość wiadomości

Liczymy znaki i słowa w wcześniej przetworzonej wiadomości. Usuwane są z niej odnośniki. Retweetowane wiadomości również podlegają analizie. W ich przypadku usuwamy charakterystyczny początek w postaci “RT @username:”.

Interpretacja długości w postaci zarówno liczby znaków i słów powinna przynieść podobne wyniki. Znaki mogą być jednak lepszą miarą wkładu pracy włożonej w napisanie wiadomości. Wypowiedź może składać się z długich lub bardzo krótkich wyrazów, co przy takiej samej liczbie słów może dać nawet dwa razy więcej znaków.

4.2.4. Sentyment badanej wiadomości

Parametr określający nastawienie autora wiadomości. Poprzez analizę sentymentu odpowiadamy na pytanie, czy wiadomość ma wydźwięk pozytywny, czy negatywny. Można spodziewać się większej liczby negatywnych wypowiedzi po stronie niepożądanych użytkowników, zwłaszcza trolli. W przypadku spamerów sentyment nie powinien odegrać dużej roli.

Analiza sentymentu może być jednak mało wystarczająca w przypadku mniej oczywistych przypadków. Wypowiedzi mogą zawierać ironię, która zostanie odczytana jako wypowiedź pozytywna. W dodatku analiza sentymentu jest procesem skomplikowanym i nie zawsze przynoszącym poprawne wyniki.

Wiadomości poddane analizie muszą zostać przygotowane. Usuwamy linki i znaki specjalne (# i @). Z retweetów usuwamy wskazujący na nie prefiks "RT @username:" i analizujemy jak zwykłą wiadomość. Dodatkowo poprawiane są drobne literówki w każdym napisanym słowie.

4.2.5. Subiektywność badanej wiadomości

Określamy czy badana wiadomość jest subiektywna, czy obiektywna. W przypadku zwykłych użytkowników spodziewana jest większa liczba bardziej obiektywnych wiadomości. Określenie obiektywności jest jednak, tak samo jak w przypadku sentymentu skomplikowane, a otrzymane wyniki mogą odbiegać od rzeczywistości.

Wiadomości przed analizą poddawane są temu samemu zabiegowi jak w przypadku analizy sentymentu. Usuwamy linki i znaki specjalne (# i @), z retweetów usuwamy wskazujący na nie prefiks "RT @username:" i analizujemy jak zwykłą wiadomość, a drobne literówki w każdym wyrazie są poprawiane.

4.2.6. Emocje badanej wiadomości

Staramy się określić najbardziej wyróżniające się emocje w wiadomości. Jest to jeszcze trudniejsze zadanie niż w przypadku sentymentu czy obiektywności. O emocjach może decydować wiele niuansów językowych. Cecha samodzielnie może nie mieć dobrych właściwości rozróżniających użytkowników, lecz w połączeniu z analizą sentymentu może się to zmienić. Przykładowo wiadomości o negatywnym sentymencie, zawierające w sobie sporo skrajnie negatywnych emocji, mogą być próbą ataków na inną osobę. Do każdej wiadomości przyporządkowywana będzie jedna ze zbioru kategorii emocji. Zbiór powinien wyróżniać minimum 8 emocji z zachowaniem równowagi pomiędzy liczbą negatywnych jak i pozytywnych przypadków obecnych w zbiorze. Nie ukierunkowujemy się w szukaniu jedynie negatywnych emocji. Pozytywne również mogą być przydatne w klasyfikacji.

Wiadomości przygotowywane są do analizy w ten sam sposób jak w przypadku analizy sentymentu i obiektywności poprzez: usunięcie linków, znaków specjalnych, prefiksów retweetów oraz poprawę literówek w wyrazach.

4.2.7. Dwuwymiarowa analiza sentymentu i obiektywności

Na podstawie analiz sentymentu i obiektywności możliwe jest przeprowadzenie dwuwymiarowej analizy opartej o wyniki obliczonych wcześniej cech. Dzieląc wyliczone zakresy wartości współczynników dla sentymentu i obiektywności na przedziały możemy skonstruować tablicę, której pola będą odpowiadać konkretnym wartościom przedziałów dla obu cech. W przypadku takiej interpretacji, każda komórka tablicy to jedna z kategorii, które mogą zostać przypisane do wiadomości.

4.2.8. Średnie podobieństwo semantyczne wiadomości do innych wiadomości użytkownika

Wejściem do analizy są wszystkie wiadomości użytkownika. Dla każdej wiadomości porównujemy jej podobieństwo z każdą inną. Następnie dla wszystkich obliczonych podobieństw dla pojedynczej wiadomości wyciągana jest średnia arytmetyczna. Wynikiem jest zatem średnie podobieństwo badanej wiadomości do pozostałych wiadomości tego samego autora. Analiza ograniczone jest wyłącznie do pojedynczego użytkownika. Gdyby analizowane było podobieństwo wszystkich wiadomości, nawet tylko z danego tagu na portalu Twitter, nadal byłoby ich o wiele za dużo. Taka operacja byłaby możliwa jedynie w warunkach analizy małego zbioru danych. W przypadku realnego działania jest to niemożliwe do zrealizowania.

4.2.9. Liczba wiadomości użytkownika podobnych semantycznie

Analiza operuje na tych samych danych jak w przypadku poprzedniej analizy dotyczącej podobieństwa wiadomości. Tak samo jak poprzednio, dla każdej wiadomości porównujemy jej podobieństwo z każdą inną tego samego autora. Zliczamy przypadki, gdy wartość podobieństwa przekracza określony próg podobieństwa. Analiza jest ograniczona wyłącznie do pojedynczego użytkownika. Zostanie przeprowadzona dla różnych wartości progu.

4.3. Analiza zachowań i statystyk użytkowników

Analizy z tej grupy skupiają się głównie na wyszukiwaniu anomalii związanych z częstością pisania wiadomości oraz cechach opartych na kategoryzowaniu danych: źródła publikacji czy rodzaju portalu z odnośnika. Sprawdzana jest również wiarygodność retweetowanych użytkowników.

4.3.1. Najdłuższa seria wiadomości z dopuszczalnym określonym oknem czasowym pomiędzy kolejnymi wiadomościami

We wszystkich wiadomościach użytkownika wyszukiwane są serie wiadomości, w których różnice czasu między kolejnymi wiadomościami są nie większe niż N minut (okno czasowe). Dla każdego użytkownika zwracamy najdłuższą serię oraz średnią K najdłuższych serii. Zostaną przebadane różne opcje parametrów N i K.

Analiza oprócz zmian parametrów zostanie przeprowadzona w dwóch wersjach:

- uwzględniając wszystkie wiadomości użytkownika,

- uwzględniając tylko wiadomości, które zawierają wkład treści od użytkownika publikującego - odrzucane są wszystkie retweety.

Użytkownik, aby uzyskać duży rozmiar serii musi napisać wiele wiadomości w bardzo krótkim czasie lub publikować je w rozsądnych odstępach (10 lub 15 minut), przez bardzo długi okres. Pierwszy przypadek jest trywialny. Nikt normalny nie opublikuje 50 wiadomości w ciągu 15 minut. Można wtedy również łatwo wychwycić prymitywne boty. Drugi przypadek, który może nie zostać uwzględniony przez proste analizy, wychwyci użytkowników piszących dużo i regularnie. Gdy ktoś bez przerwy jest aktywny na portalu przez kilka godzin, może być potencjalnym zawodowym trollem. Może być to również bot ustawiony w taki sposób, aby publikował wiadomości regularnie w różnych odstępach czasu.

4.3.2. Maksymalna liczba wiadomości, która została napisana w oknie czasowym

Jest to prosta cecha, która pozwoli poprzez porównanie osiąganych wyników, zweryfikować przydatność bardziej zaawansowanej, opisanej w 4.3.1. W niektórych przypadkach może dostarczać bardzo podobne rezultaty. Nie wychwyci jednak inteligentniej działających użytkowników, którzy nie piszą wiadomości bez przerw. Zwróci uwagę jedynie na użytkowników, którzy są bardzo oczywistymi spamerami.

4.3.3. Źródło publikacji wiadomości

Wiadomość może być publikowana przez użytkowników z wielu różnych źródeł np. z aplikacji na telefony z systemem Android lub przeglądarki internetowej. Z każdą wiadomością na Twitterze powiązane jest źródło jej publikacji. Jest ich jednak bardzo wiele. Każdy może stworzyć swoją własną aplikację na dowolną platformę będącą kolejnym źródłem publikacji. Wystarczy zarejestrować się jako developer na portalu udostępnianym przez Twittera i wyrobić specjalne klucze, które pozwolą na "opakowanie" funkcji udostępnianych przez Twittera w zewnętrznych aplikacjach. Pozwala to na łatwe stworzenie botów. Trzeba zatem dokonać podziału na oficjalne i nieoficjalne aplikacje.

Źródła publikacji powinno się skategoryzować ze względu na popularność oraz zwracając uwagę na to, czy są one oficjalne. Zwykły użytkownik nie będzie korzystać z nieoficjalnej, mało popularnej aplikacji. Dużą wagę może odegrać również urządzenie, z którego opublikowano wiadomość. Sporo osób korzysta z telefonów, które nie sprawdziłyby się w przypadku osób, które zawodowo zajmują się trollingiem lub spamem. Należy więc manualnie przeanalizować źródła oraz na podstawie tej analizy stworzyć stosowne kategorie, do których zostaną zaliczone wiadomości.

4.3.4. Ulubiony sposób publikacji wiadomości użytkownika

Użytkownicy mogą korzystać z portalu na różne sposoby. Nie są przywiązani wyłącznie do jednej aplikacji. Osoba, która publikuje swoje wiadomości z wykorzystaniem przeglądarki internetowej, może incydentalnie korzystać z klienta mobilnego dostępnego na telefonach. Kategoryzowanie pojedynczych wiadomości jak w analizie 4.3.3, nie patrząc na ich większą

ilość może być, w niektórych przypadkach krzywdzące. Przypisując każdemu użytkownikowi jego ulubioną kategorię źródła publikacji eliminujemy ten problem.

Wykorzystujemy te same kategorie, które zostaną stworzone w analizie 4.3.3. Muszą być one jednak uzupełnione o kilka nowych, które uwzględnią przypadki, w których liczba publikacji z różnych kategorii jest zbliżona lub równa.

4.3.5. Stosunek liczby obserwujących i obserwowanych retweetowanych użytkowników

Zwykle retweetowane są wiadomości z popularnych źródeł takich jak profile informacyjne, profile celebrytów lub autorytetów w danych dziedzinach. O wiele rzadziej zdarza się podanie dalej wiadomości napisanej przez mało popularną osobę. Może to zostać wykorzystane do wybrania wiadomości, które mogą być próbami retweetowania wiadomości innego trolla lub spamera, dla którego współczynnik ten nie będzie mieć dużej wartości. Stosunek może być interpretowany jako wiarygodność źródeł, z których wiadomości są podawane dalej. Wartość obliczana jest w ten sam sposób jak w analizie 4.4.5 z wykorzystaniem wzoru 6.

4.3.6. Kategoria portalu, do którego prowadzi link

Analizie należy poddać również portal, do którego prowadzi link. Z pewnością mniej podejrzany jest link prowadzący do YouTube czy znanego portalu informacyjnego niż do całkowicie nieznanej domeny, szczególnie pochodzącej z mało znanego kraju. System opracowany w artykule [42] jest bardzo skomplikowany. Nie damy rady sprawdzić treści portali, do których prowadzą linki, lecz możemy spróbować stworzyć kategorie pozwalające na choć minimalne rozróżnienie portali. Przeanalizowane zostanie kilkadziesiąt najpopularniejszych portali z każdego zbioru. Na tej podstawie zostaną stworzone kategorie, do których z pewnością zaliczą się: portal informacyjny, portal z multimediami lub portal do generowania skróconych linków.

4.4. Analiza zależności między użytkownikami

Praktycznie wszystkie cechy z kategorii są powiązane z mechanizmami dostępnymi wyłącznie na Twitterze i nie są możliwe do wykorzystania w tej samej formie na innym portalu. Zależności opierają się głównie na mechanizmach obserwowania użytkowników oraz retweetowania wiadomości.

4.4.1. Rozmiar maksymalnej kliki użytkownika w grafie retweetujących się użytkowników

Jest to cecha zainspirowana artykułem [36], w którym tworzony był graf współpracujących użytkowników. Analizowano tam jednak fora portali informacyjnych. Przekładając analizę grafową na portal Twitter możemy zwrócić uwagę na mechanizm retweetowania. Krawędziami łączymy użytkowników, z których jeden zretweetował drugiego minimum N razy. W grafie szukamy rozmiaru maksymalnej kliki, w której znajduje się użytkownik. Maksymalnej, czyli takiej, której nie da się rozszerzyć o kolejny wierzchołek. Analiza zostanie przeprowadzona dla różnych wartości parametru N.

Wszyscy użytkownicy, którzy są członkami klik o dużych rozmiarach są podejrzani. Istnienie dużej grupy użytkowników, w której wszyscy się wzajemnie często retweetują nie może być przypadkiem. Charakterystyka Twittera wyklucza sens przypadku, w którym użytkownicy ci są znajomymi. Zwykle podajemy dalej wiadomości od znanych użytkowników z bardzo dużą liczbą obserwujących. Są to informacje ze świata lub wyróżniające się opinie na różne tematy.

4.4.2. Liczba maksymalnych klik w grafie retweetów, w których znajduje się użytkownik

Wykorzystując ten sam graf jak w analizie 4.4.1 zliczamy w ilu maksymalnych klikach, dla różnych innych użytkowników znajduje się badany użytkownik. Pozwoli to na określenie stopnia integracji użytkownika w społeczności. Przeprowadzona zostanie analiza zliczająca wszystkie kliki, oraz wyłącznie te, których rozmiar jest większy lub równy 3.

Cecha 4.4.1 i opisywana w podpunkcie operujące na grafach mogą być skomplikowane do przeprowadzenia na dużych zestawach danych. Możemy w takim przypadku analizować zależności wyłącznie w określonych przedziałach czasowych lub ograniczając się do konkretnych tagów.

4.4.3. Liczba obserwujących

Jest to prosta cecha zwracająca uwagę na to ilu innych użytkowników obserwuje badany użytkownik. Żaden użytkownik nie ma wpływu na wartość liczby powiązanej z jego kontem. Od innych zależy to, czy zaobserwują profil na podstawie jego treści i tematyki. W przypadku niepożądanych użytkowników spodziewamy się, że liczba ta będzie mała.

4.4.4. Liczba obserwowanych

W przeciwieństwie do liczby obserwujących na wartość swojej omawianej cechy ma wpływ każdy użytkownik. Może on obserwować dowolną liczbę innych użytkowników. Bardzo duże liczby obserwowanych użytkowników mogą jednak być dziwne. Może być przykładowo próba zwiększenia zasięgu operacji bota, komentującego wpisy obserwowanych użytkowników.

4.4.5. Stosunek liczby obserwujących do liczby obserwowanych użytkowników

Stosunek wartości liczby obserwujących i obserwowanych użytkowników dla każdego konta obliczany jest z wykorzystaniem wzoru:

$$StosunekWartosci = \frac{LiczbaObserwujcych}{LiczbaObserwowanych + 1}. \tag{6}$$

Do mianownika wiadomości dodawana jest jedynka. Nie ma to dużego wpływu na osiągane wyniki, zwłaszcza przy większych liczbach, a pozwala uwzględnić użytkowników, którzy nie obserwują ani jednej osoby.

Im mniejszy jest stosunek tych wartości tym większe prawdopodobieństwo, że użytkownik jest niepożądany. Inni użytkownicy nie będą chcieli obserwować osoby, która jest trollem lub spamerem. Użytkownik ma wpływ tylko na liczbę obserwowanych użytkowników.

Na podstawie stosunku wartości można wyróżnić trzy grupy użytkowników:

- użytkownicy, których stosunek jest wartością bardzo małą. Można przyjąć, że są to wartości poniżej 0,5. Konta o takich wartościach współczynnika możemy podejrzewać;
- użytkownicy, których stosunek jest wartością bardzo dużą. Są to zazwyczaj znani ludzie, konta portali informacyjnych lub profile poświęcone konkretnym tematom. Publikowane przez nich wiadomości są często retweetowane;
- użytkownicy, których stosunek jest wartością rzędu jedności. Są to zwykli użytkownicy Twittera.

Należy jednak podkreślić, że powyższe grupy mają sens dla konkretnego użytkownika, gdy jego liczba obserwowanych lub obserwujących to minimum kilkadziesiąt osób.

4.4.6. Liczba polubień wiadomości

Rodzaj systemu ocen wypowiedzi oraz przyzwyczajenia użytkowników mają duży wpływ na wyniki analizy cechy. Wiele portali pozwala jedynie na ocenę poprzez polubienia wiadomości (do tej grupy należy Twitter), inne na wyrażanie emocji odpowiednim emotikonem, a jeszcze inne na dawanie plusów i minusów wiadomościom. Mechanizm polubień jest często ignorowany przez użytkowników. W dodatku może dochodzić do sytuacji, w których wiadomości trolli obrażające innych uzyskują duży poklask. W dodatku mechanizm, który używany jest na Twitterze, nie jest zbyt dobry - nie udostępnia on możliwości negatywnej oceny. Cecha oparta na polubieniach wiadomości będzie mieć raczej nikły wpływ na kategoryzację użytkowników.

4.5. Kryteria wyboru zbiorów danych

Analizowanym portalem jest portal Twitter. Zwracamy uwagę wyłącznie na wiadomości napisane w języku angielskim. Szukane zbiory danych powinny mieć tematykę poważną - muszą zawierać merytoryczne rozmowy. Przykładem są tematy polityczne, w których zawsze znajdziemy wiele wiadomości i różnych wątków w jednym temacie. Przykładem mogą być zbiory danych na temat wyborów prezydenckich ze Stanów Zjednoczonych, w których można wyróżnić wiele prób trollingu i spamu wśród bardziej poważnych dyskusji.

Znalezienie zbioru danych zawierającego wiadomości z Twittera jest utrudnione jego polityką prywatności. Nie jest zezwolone publikowanie w jakikolwiek sposób całych wiadomości. Można jednak publikować zbiory numerów identyfikujących tweety za pomocą, których mogą zostać one później pobrane z wykorzystaniem udostępnianego API.

Istnieje wysokie prawdopodobieństwo, że nie zostanie znaleziony zbiór zawierający jednocześnie normalnych i niepożądanych użytkowników. Konieczne jest więc odszukanie

dwóch zbiorów mocno zbliżonych do siebie, które ewentualnie będą mogły być uzupełnione o brakujące parametry. Niektóre zbiory trolli i spamerów udostępniane są do badań publicznie przez samego Twittera. Zbiór normalnych użytkowników można stworzyć poprzez wykorzystanie mechanizmu weryfikacji kont dostępnego na Twitterze i zbiory numerów identyfikujących tweety.

Należy wziąć pod uwagę to, że nie można stworzyć zbioru idealnego, który pozwoli na przebadanie wszystkich wymaganych kryteriów oceny wiadomości i użytkowników. Jest to tym bardziej niemożliwe, gdy używamy zbiorów, których nie tworzymy sami. Analizowane dane wyróżniają się jedynie niektórymi cechami, a stworzone na ich bazie klasyfikatory będą wyspecjalizowane głównie na najbardziej różniących się zjawiskach obecnych w tych zbiorach.

4.6. Metody analizy cech

Podstawowa analiza wszystkich cech będzie przeprowadzana w podobny sposób. Wyniki analiz zostaną zobrazowane poglądowymi diagramami rozkładów wartości zawierających konkretne liczby elementów w danych przedziałach wartości. Będzie to pozwalało na szybkie zestawienie ze sobą pierwszych różnic pomiędzy dwoma grupami użytkowników. Te same dane zostaną również zaprezentowane w postaci tabel, które posłużą do bardziej precyzyjnej analizy. Tabele będą już zawierać procentowe udziały poszczególnych kategorii w obu zbiorach, wraz z kolumną obrazującą różnicę. Aby ułatwić porównanie ze sobą rozkładów w obu zbiorach tabele będą posiadać również kolumny obrazujące procent pokrycia pełnego zbioru po dodaniu każdego nowego przedziału (wiersza).

W przypadku wyciągnięcia ciekawych spostrzeżeń z analiz podstawowych, zostaną przeprowadzone bardziej specyficzne dla zaobserwowanych zjawisk. Każda analiza zostanie podsumowana listą wyciągniętych z niej wniosków.

4.7. Sposób opracowania zbioru treningowego i testowego

Zbiory normalnych i niepożądanych użytkowników nie będą równoliczne. Zbiór normalnych użytkowników może być znacznie większy. Wszystkie analizy cech zostaną przeprowadzone na wszystkich dostępnych danych. Następnie, większy ze zbiorów zostanie poklastrowany z wykorzystaniem cech, a z każdego klastra zostanie wylosowana odpowiednia liczba użytkowników wchodząca w skład zredukowanego zbioru. Stosunek liczby reprezentantów każdego klastra w zredukowanym zbiorze zostanie zachowany. Taki sposób przekształcenia dużego zbioru pozwoli na zachowanie jego charakterystyk.

Zbiory treningowe i testowe zostaną stworzone z wykorzystaniem kroswalidacji. Liczby użytkowników normalnych i niepożądanych w zbiorach będą równe.

4.8. Porównanie cech pod względem skuteczności klasyfikacji

Wszystkie analizowane cechy zostaną przebadane pod kątem ich indywidualnych zdolności klasyfikacyjnych. Dla każdej cechy zostanie skonstruowany klasyfikator stworzony z wykorzystaniem takiego samego algorytmu. Skuteczność klasyfikacji oceniana będzie poprzez miary: F1, precyzję, czułość i dokładność.

Analizie poddane zostaną również grupy użytkowników zaklasyfikowanych przez pojedyncze cechy do tych samych grup. Zostanie zweryfikowane to, czy poszczególne cechy klasyfikują tych samych użytkowników do tych samych grup. Pozwoli to na określenie zgodności różnych cech w klasyfikacji.

Porównana zostanie skuteczność klasyfikacji różnych podzbiorów zbioru wszystkich analizowanych cech. Zostanie opracowany ranking najlepszych podzbiorów. Miarą oceny w rankingu będzie metryka F1. Zbiory będą wybierane na podstawie wcześniejszych obserwacji analiz cech oraz ich indywidualnych zdolności klasyfikacyjnych. Przeanalizowanie wszystkich możliwych podzbiorów jest niemożliwe w rozsądnym czasie. Wyróżnić można 3 grupy, które zostaną przebadane. Są to kategorie źródeł przedstawione na obrazku 2.

4.9. Opracowanie finalnego zbioru najlepszych cech i wybór algorytmu klasyfikacji

Zostanie przebadana skuteczność klasyfikacji wszystkich cech razem. Następnie poszczególne cechy bedą usuwane ze zbioru, na podstawie analiz ich korelacji z innymi cechami oraz wkładu jaki dają w kwestii poprawy skuteczności. Analizy niektórych czynników mogą być bardzo zbliżone do siebie i wprowadzać redundancję.

Porównane zostaną wyniki różnych algorytmów klasyfikacji dla tych samych zestawów cech co pozwoli na odniesienie się wcześniejszych badań i ewentualne potwierdzenie najlepszej skuteczności algorytmu lasów losowych w badanym problemie. Zostaną dobrane również optymalne konfiguracje wykorzystywanych algorytmów.

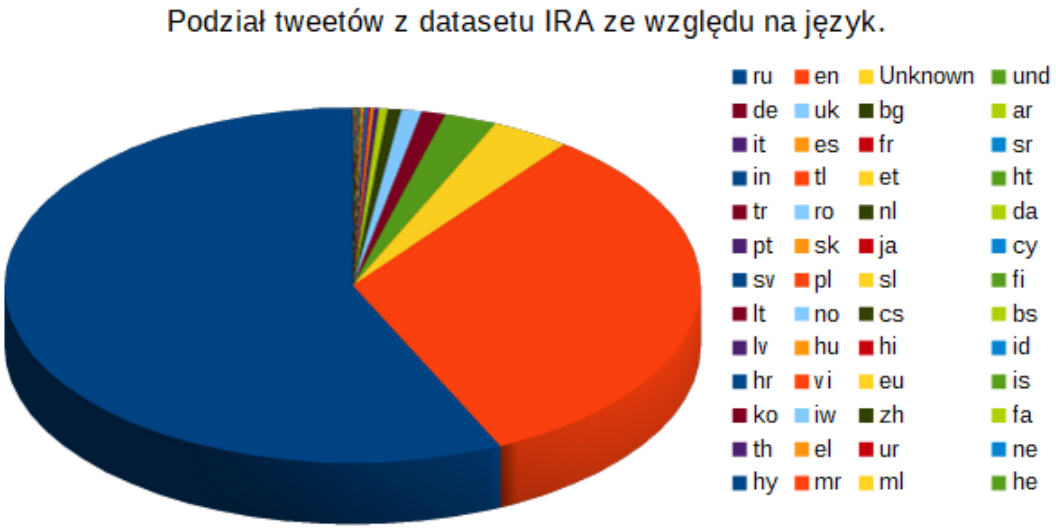
5. Opis zbiorów danych oraz środowiska ich przetwarzania i analizy

Rozdział zawiera informacje na temat wykorzystywanych zbiorów danych, statystyki o nich, metody ich stworzenia lub uzupełnienia. Przedstawiony jest przepływ danych w systemie oraz opisane są jego główne elementy.

5.1. Opis wybranych zestawów danych

5.1.1. IRA dataset

Jest to publicznie udostępniony przez Twittera zbiór wiadomości w formie pliku CSV, zawierający posty 3293 niepożądanych użytkowników usuniętych z portalu. Większość z nich stanowią trolle z Internet Research Agency (IRA), rosyjskiej agencji dbającej o polityczne i biznesowe interesy Rosji w internecie. Zbiór zawiera 1 825 877 tweetów w różnych językach. Większość wiadomości jest po rosyjsku, lecz język angielski również ma spory udział w zbiorze (595 823 tweetów, co stanowi prawie 1/3 zbioru). Udziały poszczególnych języków zostały przedstawione na diagramie 3.



Rysunek 3: Powyższy diagram kołowy przedstawia udział poszczególnych języków w wiadomościach zebranych w zbiorze IRA. Wyróżniamy 51 języków. Część wiadomości nie została przyporządkowana do żadnej z kategorii (poniżej 4% wiadomości). Widzimy, że ponad połowę stanowią tweety napisane w języku rosyjskim. Udział wiadomości napisanych po angielsku badanych w pracy jest również zadowalający i wynosi około 32.6%. (diagram jest twórczością własną)

Zebrane dane zawierają podstawowe informacje o tweetach (między innymi tekst wiadomości, data publikacji, hasztagi, odniesienia do innych użytkowników, linki, liczbę polubień, odpowiedzi i retweetów) oraz bardzo ograniczone informacje na temat użytkowników

(między innymi liczbę obserwujących i obserwowanych, lokalizacja, data założenia konta, opis w profilu użytkownika).

Uzupełnienie zbioru

Zbiór IRA został uzupełniony o dane powiązane z informacjami o użytkownikach, których wiadomości były retweetowane. Część danych pochodzi z tego samego zbioru (użytkownicy ze zbioru IRA wzajemnie się retweetowali), lecz znacząca większość została pobrana z wykorzystaniem Twitter API [17] opisanego w kolejnym punkcie. Nie udało się jednak zdobyć informacji o prawie 42% retweetowanych użytkownikach (292 525 z 697 230). Jest to spowodowane brakiem informacji o zawieszonych lub usuniętych kontach na Twitterze.

Zostały również dodane liczbowe kolumny: liczba hashtagów, liczba odniesień do innych użytkowników oraz liczba linków. Obliczono je na podstawie tekstowych danych z powiązanych kolumn, będącymi listami konkretnych linków, hashtagów czy użytkowników rozdzielonymi przecinkami.

5.1.2. Zbiór z bazy Harvard Dataverse

"2016 United States Presidential Election Tweet Ids" [37] to zbiór zawierający numery id około 280 milionów tweetów dotyczących wyborów prezydenckich w Stanach Zjednoczonych z 2016 roku. Dane podzielone są na 12 kolekcji. Zostały zebrane w 2016 roku i aktualnie około 30% tweetów jest już niedostępne. Z pewnością znaczna część usuniętych tweetów jest powiązana z kontami spamerów i trolli, którzy aktywnie działali w obrębie badanych tagów. Tematy powiązane z wyborami z 2016 roku były wielokrotnie badane pod tym kątem i większość niepożądanych wiadomości została już usunięta.

Do analizy w pracy została wykorzystana tylko część dostępnych danych. Wybrana została kolekcja "Candidates and key election hashtags". Jest ona podzielona na 6 plików. Każdy zawiera około 50 milionów id tweetów. Kolekcja zawiera wiadomości z tagów:

- election2016,
- election,
- clinton,
- kaine,
- trump,
- pence.

oraz z oficjalnych kont kandydatów:

- Hillary Clinton,
- Donald J. Trump,
- Mike Pence,
- Tim Kaine.

Kolejność tweetów, których id znajdują się w plikach jest uporządkowana w czasie. Dzięki temu można łatwo wybrać podzbiór całości, który zawiera wiadomości opublikowane w danych ramach czasowych. W zbiorze okazjonalnie występują jednak duplikaty, które należy odfiltrować. W 12 milionach wybranych id wiadomości znalazło się około 40 tysięcy duplikatów.

5.1.3. Własny zbiór danych na podstawie zbioru Harvard Dataverse

Zbiór został stworzony na podstawie numerów id tweetów zebranych w zbiorze [37] opisanym w poprzednim punkcie. Z wykorzystaniem Twitter API zostało pobrane 11 752 372 tweetów z kolekcji "Candidates and key election hashtags". Zbiór zawiera publikacje na obserwowanych tagach i kontach z kolekcji źródłowej na przestrzeni około 9 dni. Data publikacji najstarszego tweeta to 2016-07-13 13:50:23, a najmłodszego 2016-10-22 14:42:54. Zostały zapisane tylko wybrane informacje na temat tweetów i ich autorów. Zbiór danych zawiera 1 859 531 użytkowników, z których 34 277 jest zweryfikowanych. Jak można zaobserwować na rysunku 4 w zbiorze dominują wiadomości napisane w języku angielskim.

Rysunek 4: Diagram przedstawia stosunek ilości tweetów napisanych w różnych językach znajdujących się w stworzonym zestawie danych. 93,4% stanowią tweety w języku angielskim. Na kolejnych miejscach znajdują się odpowiednio języki hiszpański i francuski z udziałami 2% oraz 0,56%. Udział pozostałych języków jest marginalny. W języku rosyjskim zostało napisane zaledwie 0,03% wiadomości. 2,8% tweetów nie zostało sklasyfikowanych ("und").

5.2. Wykorzystywany stos technologiczny

Język programowania i główne biblioteki

Do stworzenia wszystkich części implementacyjnych został wybrany język Python [13]

w wersji 3. Wydaje się to być rozsądny wybór patrząc na liczbę dostępnych bibliotek oraz wsparcie społeczności. Kluczowe używane biblioteki to numpy [10], pandas [11] oraz scikit-learn [15] wykorzystywany do klasyfikacji i klastrowania. Można powiedzieć, że wybrany język i biblioteki zaliczają się do typowego stosu technologicznego w analizie danych i uczeniu maszynowym.

Wiadomości są pobierane poprzez udostępniane dla developerów Twitter API [17]. Za pomocą API można uzyskać wszystkie publiczne informacje o wiadomościach i użytkownikach z nimi powiązanymi. Szczegółowy opis wszystkich zwracanych informacji dostępny jest w publicznie dostępnej dokumentacji [17]. Zapytania do API wysyłane są z użyciem biblioteki tweepy [16], opakowującej wiele różnych zapytań w udostępnianych metodach, które wystarczy jedynie wywołać podając jako argument numery ID interesujących nas elementów. Mniej kluczowe biblioteki związane z realizacją konkretnych cech zostaną opisane przy okazji ich użycia w dalszej części pracy wraz z przykładami ich działania.

Baza danych

Do przechowywanie danych została wybrana baza relacyjna PostgreSQL [12]. Rozważana była również dokumentowa baza nierelacyjna MongoDB [8]. Format JSON, w którym przechowywane są przez nią dane, jest zgodny z tym, który jest zwracany z Twitter API używanym do pobierania danych. Nie jest to jednak bardzo kluczowe, zwracane dane z łatwością można sformatować, a jednorazowy nakład pracy na stworzenie struktury bazy relacyjnej nie jest duży. Minusem MongoDB jest rozmiar przechowywanych elementów, który w przypadku bazy relacyjnej będzie mniejszy. Największym argumentem za wyborem bazy relacyjnej jest jednak posiadanie o wiele większego doświadczenia w pracy z tego typu bazami. Pozwoli to na znaczne przyśpieszenie przetwarzania danych, co w przypadku bazy nierelacyjnej mogło by wymagać sporego nakładu pracy i uzupełnienia wiedzy.

Analiza wyników

Wyniki analizowane są z wykorzystaniem darmowego narzędzie Libre Office Calc [7]. Jest to program służący do pracy z arkuszami kalkulacyjnymi. Pliki wyjściowe zawierające wyniki analiz danych wykonywane w języku Python są zapisywane w formie, która pozwala na ich bezpośredni import do arkuszu. Dzięki temu w wygodny sposób można przeglądać oraz wizualizować otrzymane rezultaty. Większość wykresów i tabel została wykonana z wykorzystaniem opisywanego narzędzia.

5.3. Główne elementy systemu

5.3.1. Baza danych

Wszystkie wiadomości przechowywane są w dwóch osobnych tabelach. Jedna zawiera normalnych użytkowników, a druga niepożądanych. Nie zdecydowano się na rozdzielenie danych dotyczących jednego typu użytkowników do różnych tabel. Zepsuło by to możliwość przejrzystego przeglądania danych oraz zmusiło do budowy bardziej skomplikowanych zapytań. Zostały stworzone indeksy na kluczowych kolumnach (id wiadomości i użytkownika) znacznie przyśpieszające zapytania.

5.3.2. Moduł przygotowania danych

Moduł dostarczający dane opiera się głównie na opisanym już udostępnianym przez Twittera API oraz bibliotece tweepy [16] pozwalającej na komunikację z API w przyjaznej formie. Konieczne do jego działania jest wprowadzenie klucza i tokenu użytkownika, które można uzyskać poprzez rejestrację na stronie Twittera przeznaczonej dla developerów [17] oraz złożenie podania wraz z uzasadnieniem potrzeby posiadania dostępu. Zwykłe konto jest darmowe, lecz ma pewne limity. Możliwe jest wykonanie jedynie określonej liczby zapytań w oknie 15 minut. Stworzony moduł czyta z plików wejściowych id wiadomości do pobrania, wykorzystuje zapytania pozwalające uzyskać dane o 100 wiadomościach naraz, a w przypadku niepowodzeń wszystko loguje. Obsługiwane jest często występujące wyczerpanie limitu API poprzez oczekiwanie i sprawdzanie stanu podejmowane w stosownych odstępach czasu. Pobrane wiadomości zapisywane są do relacyjnej bazy danych. Moduł służy również do uzupełniania zbiorów o brakujące informacje potrzebne w analizach. Wyłącznie on dokonuje zmian w bazie danych.

5.3.3. Moduł udostępniający implementacje cech

Moduł skupia wszystkie implementacje cech określających użytkowników oraz ich wiadomości z wyjątkiem nielicznych wykonywanych wyłącznie po stronie bazy danych. Przesunięcie odpowiedzialności za obliczenia na bazę w niektórych przypadkach może zaowocować o wiele szybszym czasem wykonania.

5.3.4. Moduł przeprowadzania analiz

Moduł skupia w sobie metody zwracające dane wykorzystywane do analizy cech użytkowników. Skupia on w sobie wszystkie zapytania do bazy danych, które poddaje badaniu z wykorzystaniem modułu cech. Wyniki mogą być zliczone w kubełkach o ustawionych przedziałach wartości. Wykorzystane to może być przy bardzo szerokich zakresach wartości wyników, które z pewnością wystąpią. Jako wyjście z każdej analizy zwracany jest plik tekstowy, który łatwo może zostać wczytany do arkusza kalkulacyjnego pozwalającego na wygodną analizę rezultatów.

5.3.5. Moduł tworzenia zbiorów testowo-treningowych

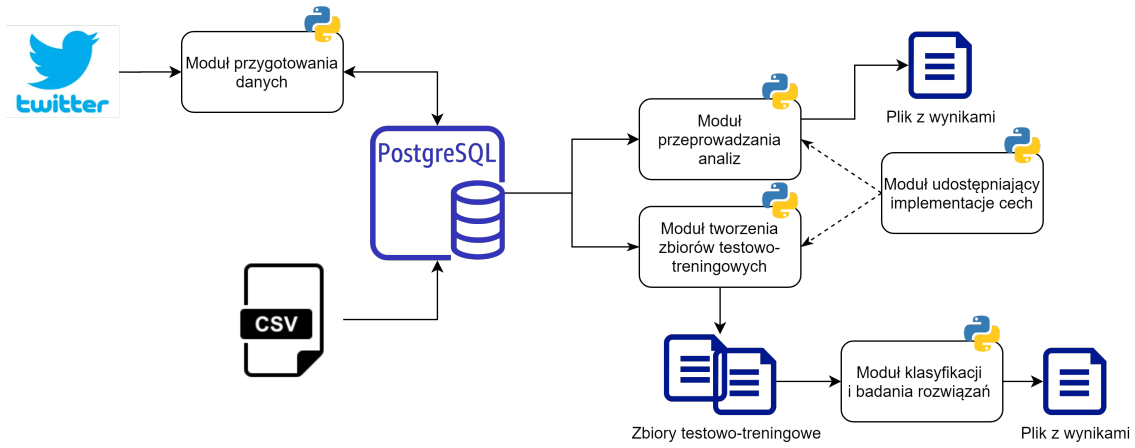
Moduł pozwala na obliczenie wszystkich dostępnych cech dla danych użytkowników i ich wiadomości. Zawiera również implementacje funkcji wykorzystywanych do filtrowania, klastrowania, skalowania oraz scalania zbiorów. Skalowanie powiązane jest klastrowaniem, tak aby zachować proporcje oryginalnego zbioru. Jest to najbardziej wymagający moduł potrzebujący bardzo dużej ilości pamięci operacyjnej komputera. Kolejne kroki przetwarzania zbioru z jego wykorzystaniem opierają się na przekazywaniu plików o dużych rozmiarach pomiędzy kolejnymi etapami.

5.3.6. Moduł klasyfikacji i badania rozwiązań

Moduł opiera się głównie na bibliotece scikit-learn [15]. Wykorzystywany jest do badania zbiorów cech oraz przeprowadzania klasyfikacji. Udostępnia analizy pozwalające na eliminację nadmiarowych cech. Z jego wykorzystaniem budowane i testowane są wszystkie klasyfikatory. Dostarcza różne warianty uruchamiania serii klasyfikacji, również poprzez pliki konfiguracyjne. Wyniki podobnie jak w przypadku modułu przeprowadzania analiz zwracane są w postaci plików tekstowych.

5.4. Przepływ danych w systemie

W celu lepszego przedstawienia systemu i roli jego elementów można zaprezentować diagram przepływu danych. Pozwoli on również na ogólne zapoznanie się z całym procesem przetwarzania i analizy danych.



Rysunek 5: Diagram prezentuje przepływ danych w stworzonym systemie ich przetwarzania i analizy. Można wyróżnić dwie części systemu: odpowiedzialną za stworzenie i przygotowanie danych zapisanych w bazie oraz drugą zajmującą się przetwarzaniem tych danych. Linią przerywaną na diagramie zaznaczono wykorzystanie innego modułu. Rysunek wykonano z wykorzystaniem narzędzia draw.io [2].

Na rysunku 5 został przedstawiony przepływ danych w stworzonym systemie. Przepływ możemy podzielić na dwie kluczowe części. Pierwsza część dotyczy stworzenia i uzupełniania zbiorów zapisywanych w relacyjnej bazie danych. Druga polega na odczycie, analizie i przetwarzaniu zawartości przygotowanej bazy.

Dane pozyskiwane są na dwa sposoby. Pierwszym i o wiele prostszym jest bezpośredni import do bazy danych pliku CSV zawierającego zbiór niepożądanych użytkowników. Drugi sposób dotyczy zbioru normalnych użytkowników, którzy pobierani są z wykorzystaniem modułu przygotowania danych bezpośrednio z Twittera. Oba zbiory są dodatkowo uzupełniane z wykorzystaniem tego modułu o brakujące dane. Jest to jedyny moduł dokonujący zmian w bazie danych.

Moduł przeprowadzania analiz oraz tworzenia zbiorów testowo-treningowych bezpośrednio odczytują dane z bazy. Oba elementy systemu korzystają z implementacji cech udostępnianych z osobnego modułu, co zostało przedstawione na diagramie linią przerywaną. Wnioski z analizy danych trafiają do plików tekstowych, które można łatwo

wykorzystać do wizualizacji wyników w arkuszu kalkulacyjnym. W przypadku modułu tworzenia zbiorów testowo-treningowych odczytane przez niego dane z bazy są przetwarzane i przekazywane w formie pliku, który jest gotowy do użycia w kolejnym module klasyfikacji i badania rozwiązań. Operuje on wyłącznie na dostarczonych plikach, nie odczytuje nic z bazy danych. Tak samo jak w przypadku innych elementów systemu zwraca on swoje wyniki w postaci pliku tekstowego.

6. Analiza zachowań użytkowników pozwalających na ich rozróżnianie w społeczności portalu

Rozdział prezentuje wyniki i obserwacje z wykonanych analiz. Uwaga jest zwracana szczególnie na rozkłady wartości badanych cech. W analizach tego wymagających dopisywane są szczegóły ich realizacji, w tym wykorzystane biblioteki wraz z przykładami ich działania.

Cechy podzielone są na trzy punkty odpowiadające grupom z zaproponowanego podziału ze względu na źródło pochodzenia cech 4.1. w punkcie 6.1 zebrane są cechy oparte wyłącznie na analizie tekstu pisanych wiadomości, punkt 6.2 zawiera cechy stworzone na bazie zachowań i statystyk użytkowników, a 6.3 na zależnościach między badanymi użytkownikami.

6.1. Analiza treści wiadomości

6.1.1. Liczba linków w wiadomościach

W analizie zostały uwzględnione wyłącznie wiadomości w języku angielskim. Język miał wyraźny wpływ na wyniki.

Niepożądani użytkownicy
Liczba wiadomości z linkami napisanych w dowolnym języku jest równa 917 007 - 50,22% tweetów posiada przynajmniej jeden link. Odnośniki w tweetach napisanych po angielsku występują 176 064 razy, co przekłada się na występowanie minimum jednego odnośnika w 29,55% wszystkich wiadomości napisanych po angielsku. Bardzo widoczne jest to, że język wypowiedzi jest mocno powiązany z liczbą linków.

Rysunek 6: Wykres przedstawia liczby wiadomości niepożądanych użytkowników zawierające określone liczby linków. Uwzględniane są wyłącznie wiadomości w języku angielskim.

Na histogramie (rysunek 6) widzimy dominację wiadomości bez linków lub tylko z jednym. Wiadomości bez odnośników jednak wyraźnie przeważają. Jest ich ponad dwa razy więcej niż z pojedynczym linkiem. Udział pozostałych kategorii jest praktycznie niezauważalny.

W przypadku analizy wiadomości we wszystkich językach rozkład wartości byłby zupełnie inny. Wiadomości bez linków miałyby udział 49,78%, a wiadomości z jednym linkiem 49,81%. Są to wartości całkowicie różne od tych przedstawionych na rysunku 6, co świadczy o dużym wpływie narodowości użytkowników na ich zachowania.

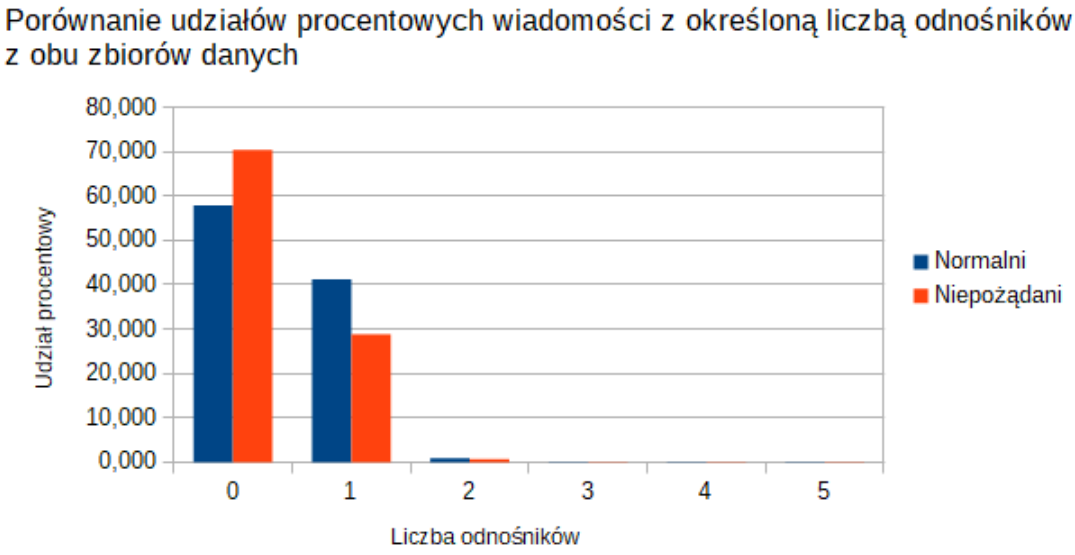
Zwykli użytkownicy

W zbiorze jest 5 139 514 wiadomości w dowolnym języku posiadających co najmniej jeden link. Przekłada się to na obecnosc odnośników w 43.73% wszystkich wiadomości. W przypadku odrzucenia wiadomości w innych językach niż angielski, których liczba jest bardzo mała udział procentowy spada do 42.05%.

Rysunek 7: Wykres przedstawia liczby wiadomości normalnych użytkowników zawierające określone liczby linków. Analizowane wiadomości są wyłącznie w języku angielskim.

Na histogramie 7 widzimy, że wiadomości bez linków lub tylko z jednym dominują tak samo jak w przypadku niepożądanych użytkowników. W tym przypadku również przeważają wiadomości bez odnośników, lecz widać, że normalni użytkownicy częściej używają linków.

Porównanie wyników



Rysunek 8: Zestawienie udziałów procentowych wiadomości z określoną liczbą linków z obu badanych zbiorów. Uwzględniane są tylko wiadomości w języku angielskim. Użytkownicy niepożądani piszą mniej wiadomości z odnośnikami.

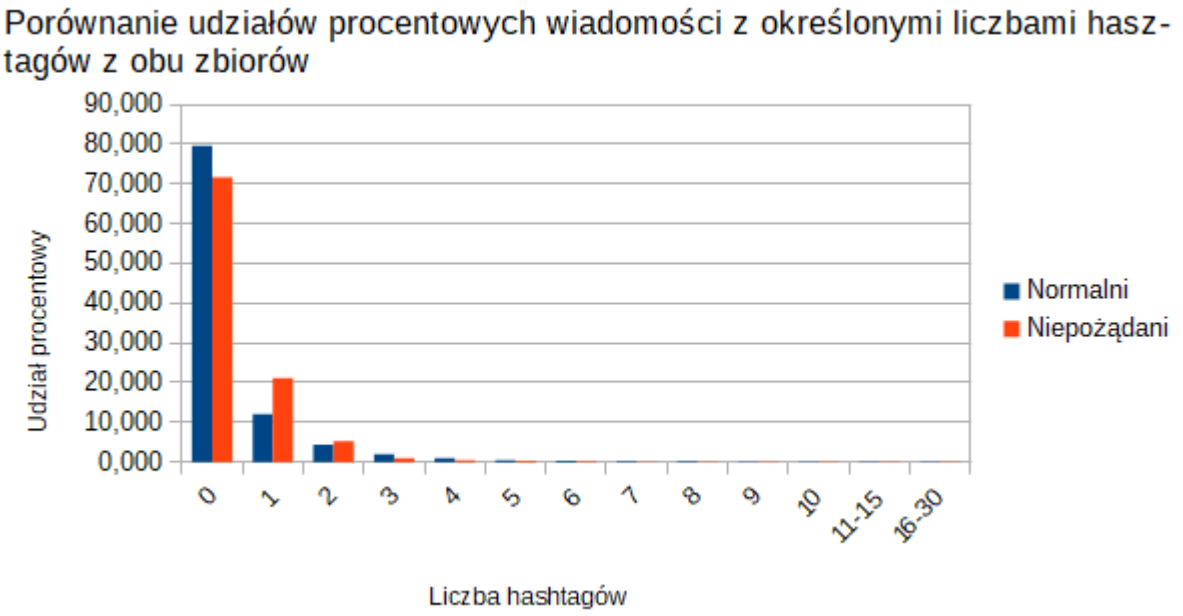
Patrząc na wykres 8 można stwierdzić, że w przypadku obu zbiorach widzimy zgodność w kwestii nie pisania wiadomości z wieloma linkami. Tabela 2 pokazuje, że w przypadku obu zbiorów wiadomości z 0 lub 1 linkiem stanowią ponad 99%. Można zaobserwować, że użytkownicy normalni częściej używają linków. Może być to spowodowane argumentacją wypowiedzi. Tweety są krótkimi wiadomościami, dlatego nie występuje w nich dużo odnośników, mimo tego, że nie wliczają się one do limitu znaków.

Tabela 2: Udziały procentowe wiadomości z określoną liczbą linków w obu badanych zbiorach

Liczba linków	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0	57,949	70,450	-12,501	57,949	70,450
1	41,216	28,808	12,409	99,166	99,258
2	0,817	0,708	0,109	99,983	99,966
3	0,015	0,032	-0,017	99,998	99,998
4	0,001	0,001	0,000	100,000	99,999
5	0,000	0,001	0,001	100,000	100,000

6.1.2. Liczba hasztagów

Uwzględniane są wiadomości we wszystkich językach. Liczba odniesień do innych użytkowników nie jest powiązana z językiem wypowiedzi.



Rysunek 9: Porównanie udziałów procentowych wiadomości z konkretną liczbą hasztagów z obu zbiorów danych. Widoczna jest wyraźna przewaga wiadomości bez hasztagów w obu zbiorach.

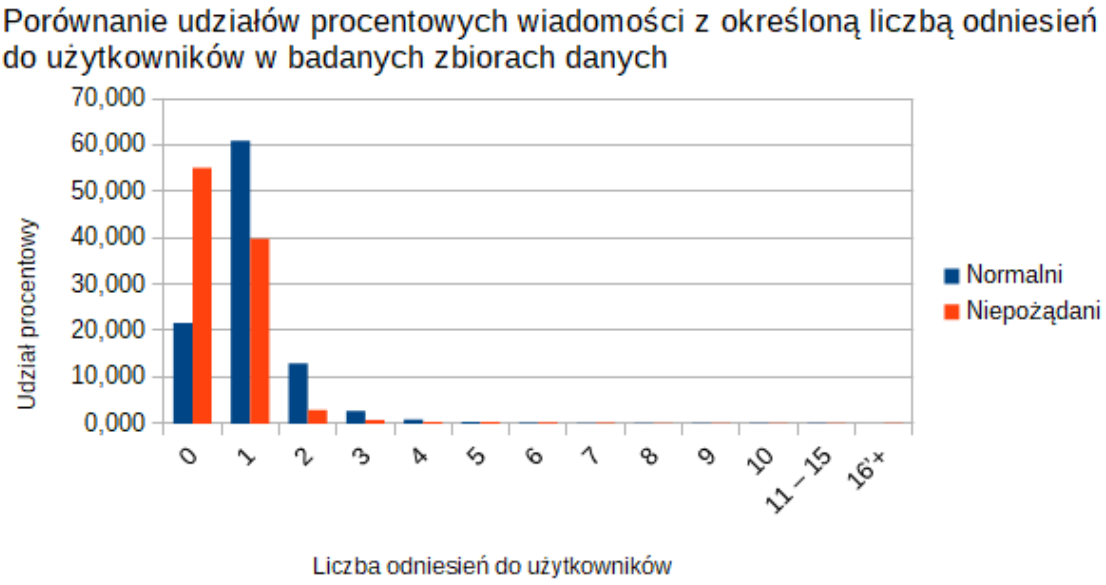
Wiadomości, w których autorzy nie zamieścili ani jednego hasztagu przeważają w przypadku obu zbiorów. Jest to wyraźnie widoczne na histogramie 9. Może to wynikać z charakterystyki badanych tematów politycznych, z których pobierane były wiadomości. Widzimy, że niepożądani użytkownicy mają większy udział wiadomości z hasztagami. Tabela 3 dostarczając bardziej szczegółowe dane z histogramu potwierdza obserwacje. Część niepożądanych użytkowników próbuje zwiększyć zasięg swoich wiadomości.

Tabela 3: Porównanie udziałów procentowych wiadomości z konkretną liczbą hasztagów z obu zbiorów danych

Liczba hasztagów	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0	79,642	71,645	7,996	79,642	71,645
1	12,063	21,118	-9,055	91,705	92,763
2	4,348	5,273	-0,925	96,052	98,036
3	1,900	0,922	0,978	97,953	98,958
4	0,979	0,418	0,561	98,932	99,376
5	0,460	0,262	0,197	99,392	99,638
6	0,252	0,123	0,129	99,644	99,761
7	0,124	0,079	0,046	99,768	99,840
8	0,123	0,055	0,068	99,891	99,895
9	0,046	0,040	0,006	99,937	99,935
10	0,025	0,030	-0,006	99,961	99,966
11-15	0,035	0,034	0,000	99,996	100,000
16-30	0,004	0,000	0,004	100,000	100,000

6.1.3. Liczba odniesień do innych użytkowników

Tak samo jak w przypadku hasztagów analizowane są wiadomości we wszystkich językach.



Rysunek 10: Wykres przedstawia zestawienie udziałów procentowych wiadomości z określoną liczbą odniesień do użytkowników w badanych zbiorach danych. Normalni użytkownicy częściej korzystają z tego mechanizmu.

Analizując wykres 10 widać, że normalni użytkownicy o wiele częściej odnoszą się do innych użytkowników. W przeciwieństwie do niepożądanych większość ich wiadomości zawiera odniesienie. Często wynika to z tego, że prowadzą oni merytoryczne dyskusje

i odnoszą się do argumentów pozostałych rozmówców. Wiadomości publikowane przez trolli są bardziej ogólne. W przypadku spamerów mechanizm odniesień do użytkowników nie ma dużego zastosowania. W tabeli 4 pokazane jest, że ponad 99% wiadomości w obu zbiorach odnosi się maksymalnie do 4 użytkowników. Odniesienia nie są tak skuteczne w kwestii zwiększenia zasięgów wiadomości jak hasztagi.

Tabela 4: Porównanie udziałów procentowych wiadomości z określoną liczbą odniesień do użytkowników w badanych zbiorach

Przedział	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0	21,724	55,060	-33,335	21,724	55,060
1	60,961	39,878	21,083	82,686	94,937
2	13,013	2,957	10,056	95,699	97,895
3	2,744	0,799	1,945	98,443	98,694
4	0,830	0,405	0,425	99,273	99,098
5	0,322	0,367	-0,045	99,595	99,466
6	0,183	0,284	-0,101	99,778	99,750
7	0,081	0,124	-0,043	99,859	99,874
8	0,061	0,048	0,013	99,920	99,922
9	0,045	0,032	0,013	99,965	99,953
10	0,020	0,027	-0,007	99,984	99,980
11-15	0,016	0,016	0,000	100,000	99,996
16+	0,000	0,004	-0,004	100,000	100,000

6.1.4. Długość wiadomości

Z tekstu tweetów podlegającego analizie zostały usunięte wszystkie linki oraz prefiksy oznaczające retweety, które też są analizowane. Poddawane badaniu tweety pochodzą z czasów, gdy maksymalna długość wiadomości wynosiła 140 znaków. Analizowana są wyłącznie wiadomości napisane w języku angielskim.

Liczby znaków

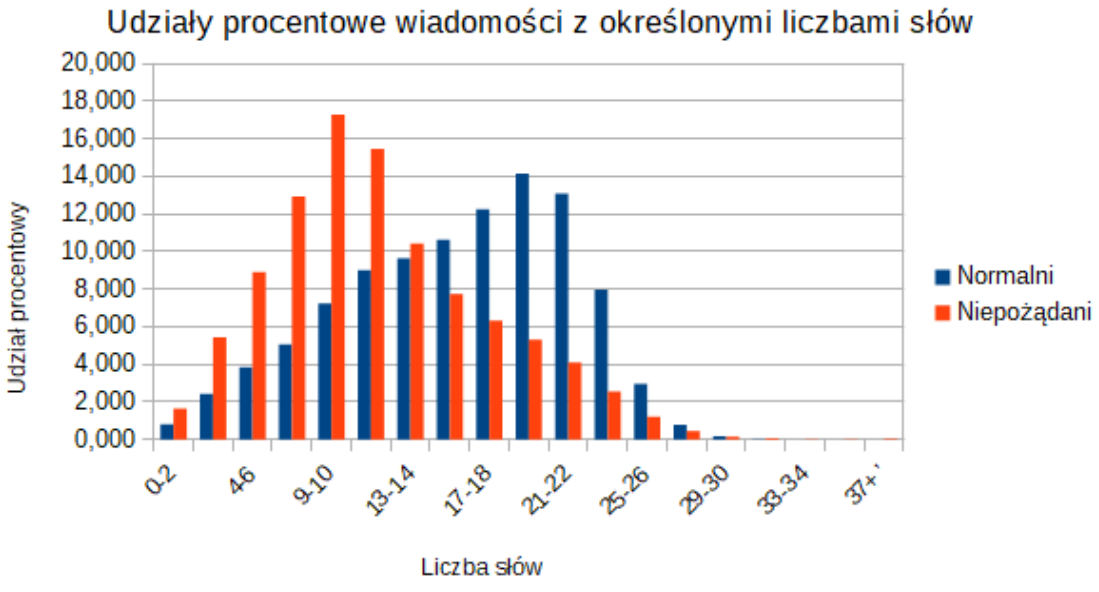
Rysunek 11: Porównanie udziałów procentowych wiadomości z obu zbiorów przypadających na określony przedział liczby znaków. Widocznie jest wyraźnie, że użytkownicy normalni publikują dłuższe wiadomości.

Na podstawie wyników z diagramu 11 i odpowiadającej tym samym danym tabeli 5 możemy powiedzieć, że użytkownicy normalni piszą dłuższe wiadomości wykorzystując pełny dostępny w tamtych czasach limit znaków (140). Najwięcej wiadomości znajduje się w przedziałach od 111 do 130 znaków w przypadku normalnych użytkowników oraz w przedziałach od 51 do 80 dla niepożądanych. Dziwić może duża przewaga 2 przedziałów normalnych użytkowników, która wynika z uwzględniania retweetów, które najczęściej trafiały do tych przedziałów.

Tabela 5: Porównanie udziałów procentowych wiadomości z obu zbiorów przypadających na określony przedział liczby znaków

Przedział	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0-10	0,415	0,584	-0,169	0,415	0,584
11-20	1,259	2,589	-1,330	1,674	3,173
21-30	2,324	4,848	-2,524	3,998	8,021
31-40	2,737	7,067	-4,329	6,735	15,088
41-50	4,001	9,666	-5,665	10,736	24,754
51-60	5,529	13,442	-7,913	16,265	38,196
61-70	6,650	14,753	-8,103	22,915	52,949
71-80	7,237	11,230	-3,993	30,152	64,179
81-90	7,814	8,136	-0,323	37,966	72,315
91-100	8,947	6,629	2,318	46,913	78,944
101-110	9,547	5,461	4,086	56,460	84,405
111-120	21,922	6,964	14,958	78,382	91,369
121-130	16,651	6,410	10,240	95,033	97,779
131-140	4,967	1,971	2,996	100,000	99,751
141+	0,000	0,249	-0,249	100,000	100,000

Liczba słów



Rysunek 12: Porównanie udziałów procentowych wiadomości z obu zbiorów przypadających na określony przedział liczby słów. Rozkłady wartości jeszcze wyraźniej pokazują, że normalni użytkownicy piszą dłuższe wypowiedzi.

Jak można było się spodziewać wyniki osiągane dla liczb słów są bardzo podobne do poprzednio uzyskanych. Rozkład wartości widoczny na histogramie 12 lepiej pokazuje

różnice między użytkownikami niż w przypadku analizy znaków. Ogólny trend jest taki sam - użytkownicy normalni piszą wyraźnie więcej. Szczegółowe dane przedstawione na wykresie można zobaczyć w tabeli 6.

W rozdziale poświęconym ewaluacji zostaną zestawione indywidualne wyniki klasyfikacji cech, które pozwolą na stwierdzenie, który sposób określania długości wiadomości jest lepszy. Można się jednak spodziewać bardzo zbliżonych wyników.

Tabela 6: Porównanie udziałów procentowych wiadomości z obu zbiorów przypadających na określony przedział liczby słów

Przedział	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0-2	0,808	1,643	-0,834	0,808	1,643
3-4	2,432	5,432	-3,001	3,240	7,075
4-6	3,839	8,906	-5,067	7,079	15,981
7-8	5,055	12,933	-7,877	12,134	28,914
9-10	7,226	17,271	-10,045	19,360	46,184
11-12	9,007	15,456	-6,449	28,367	61,641
13-14	9,648	10,435	-0,787	38,016	72,076
15-16	10,622	7,751	2,871	48,637	79,827
17-18	12,259	6,318	5,941	60,897	86,145
19-20	14,136	5,304	8,832	75,032	91,449
21-22	13,081	4,093	8,987	88,113	95,543
23-24	7,958	2,536	5,422	96,071	98,079
25-26	2,961	1,193	1,768	99,032	99,272
27-28	0,778	0,438	0,340	99,810	99,709
29-30	0,163	0,150	0,013	99,974	99,860
31-32	0,023	0,057	-0,034	99,996	99,916
33-34	0,003	0,020	-0,017	99,999	99,936
35-36	0,000	0,014	-0,013	100,000	99,950
37+	0,000	0,050	-0,050	100,000	100,000

6.1.5. Sentyment wypowiedzi

Sentyment wypowiedzi został wyliczony z wykorzystaniem biblioteki TextBlob. Analizowane są wyłącznie wiadomości w języku angielskim. Biblioteka dostępna jest w języku Python. Umożliwia dwa warianty analizy sentymentu:

- Domyślny **PatternAnalyzer** oparty o bibliotekę wzorców (linki prowadzące do źródła nie są już aktywne) zawierającą słowa z banku drzew Penn Treebank II. Wykorzystany został tutaj algorytm Brilla. Otrzymujemy jedną liczbę z zakresu -1.0 do +1.0, gdzie -1.0 to zdanie o sentymencie najbardziej negatywnym, a +1.0 o najbardziej pozytywnym.

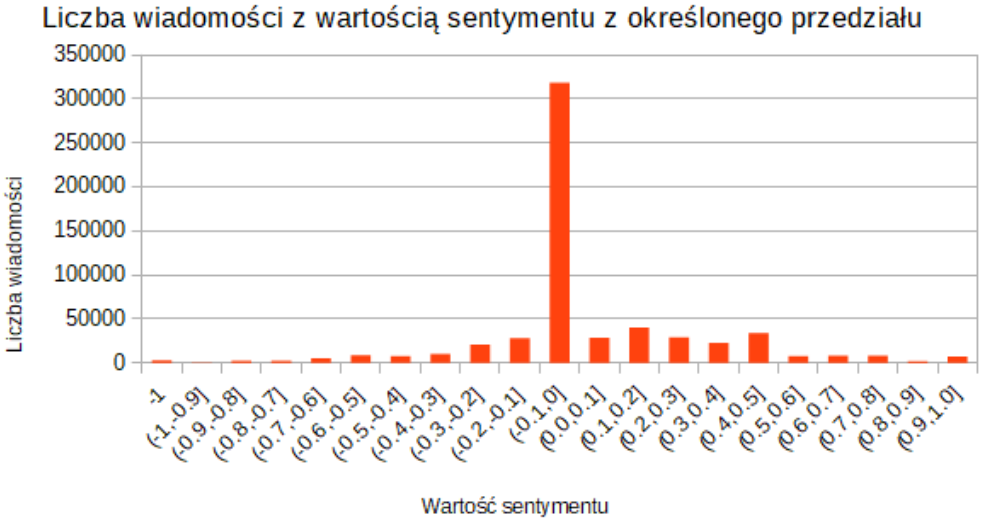
- Dużo bardziej zaawansowany **NaiveBayesAnalyzer** oparty o Stanford NLTK, w którym źródłem do nauki były recenzje filmów. Wszystkie słowa w zdaniu są

tagowane odpowiednim stosunkiem pozytywności/negatywności np. (70% do 30%). Następnie, na podstawie tych wartości wyliczane są wyniki dla całego wprowadzonego tekstu. Wynik przedstawiany jest w postaci dwóch dodatnich liczb sumujących się do jedynki, oznaczających stosunek pozytywności do negatywności wypowiedzi. Możemy dodać do siebie obie liczby przedstawiając je tak samo jak w przypadku poprzedniego wariantu. Zwracany jest również tag sentymentu (dwie możliwości "pos" i "neg"), który opiera się na na tym czy wynik sumy jest dodatni czy ujemny.

Przykładowe wyniki sentymentu obliczonego z wykorzystaniem NaiveBayesAnalyzer:

- "Media and Zombie Hillary More Angry at Trump for Calling a Bombing a Bombing Than They Are About the Bombing"
  Sentyment: -0.71;
- "Suspect Peter Selis 9.20.1967 is the one who police say is the gunman involved in the San Diego pool shooting, One fema..."
  Sentyment: -0.62;
- "Dont́ let me down, Georgia! No state with such great people and food and neighborhoods and fun should go any other way than..."
  Sentyment: 0.82;
- "CRRYYYIIINNNGGGGGG Damon officially thinks white people are crazy"
  Sentyment: 0.93.

Niepożądani użytkownicy



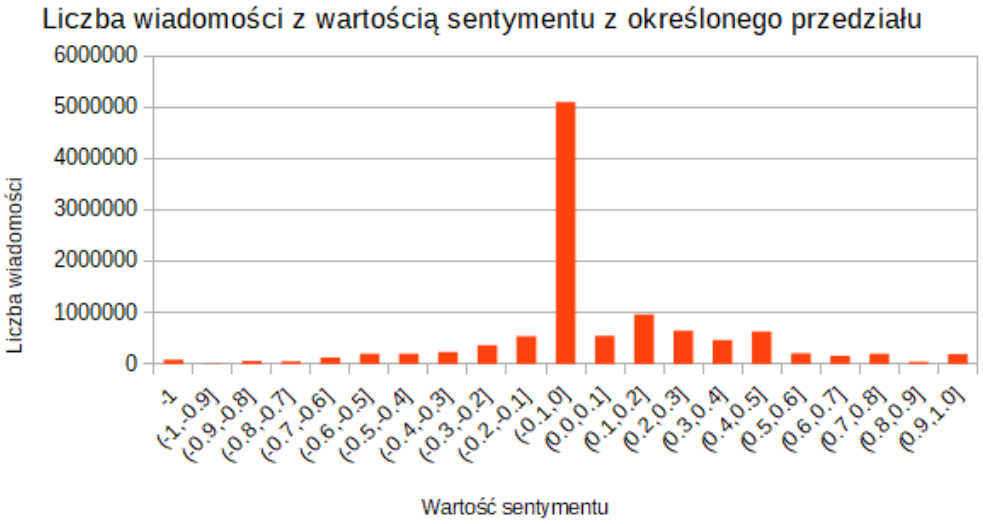
Rysunek 13: Wykres przedstawia rozkład wartości obliczonego sentymentu w przedziałach dla wiadomości ze zbioru IRA (niepożądani użytkownicy). Do obliczenia sentymentu wykorzystano PatternAnalyzer z biblioteki TextBlob.

Rysunek 14: Wykres przedstawia rozkład wartości obliczonego sentymentu w przedziałach dla wiadomości ze zbioru IRA (niepożądani użytkownicy). Obliczenia zostały wykonane z wykorzystaniem NeiveBayesAnalyzer opartego o Stanford NLTK.

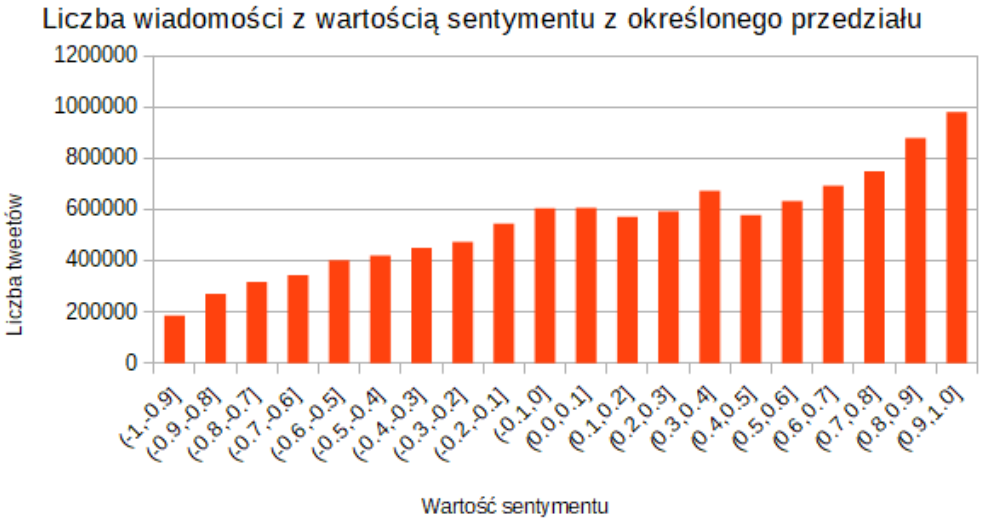
Porównując działanie biblioteki TextBlob w dwóch wariantach widać, że w przypadku prostszego rozwiązania, którego wyniki przedstawiono na wykresie 13 większość wiadomości ma sentyment z przedziału (-0.1, 0>. Prawie wszystkie wyniki z tego przedziału są równe 0. Wyrazy używane w klasyfikowanych wiadomościach nie są zapewne zawarte we wzorcach, z których korzysta TextBlob w swojej najprostszej wersji. Jest on przydatny do badania bardzo prostych zdań. Nie radzi on sobie w trudniejszych przypadkach, gdy zakres używanych słów jest rozległy.

Podejście zaawansowane, którego wyniki przedstawiono na wykresie 14 przynosi całkowicie inne rezultaty. O wiele więcej słów używanych w wiadomościach zostało otagowanych, co widać po mniejszym udziale wartości bliskich lub równych 0. Wyraźnie przeważa liczba wiadomości z pozytywnym sentymentem.

Zwykli użytkownicy



Rysunek 15: Wykres przedstawia rozkład wartości obliczonego sentymentu w przedziałach dla wiadomości stworzonego na bazie Harvard Dataverse. Do obliczenia sentymetu wykorzystano PatternAnalyzer z biblioteki TextBlob.

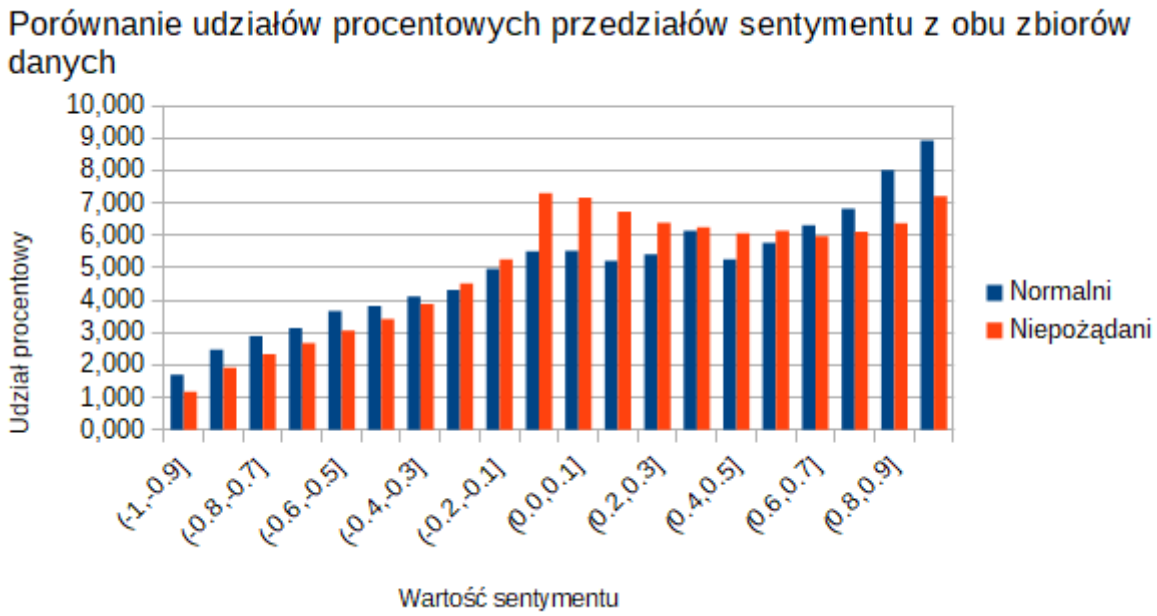


Rysunek 16: Wykres przedstawia rozkład wartości obliczonego sentymentu w przedziałach dla wiadomości stworzonego na bazie Harvard Dataverse. Obliczony został z wykorzystaniem dostępnego dla TextBlob NeiveBayesAnalyzer'a opartego o Stanford NLTK.

Tak samo jak w przypadku zbioru niepożądanych użytkowników większość wyników uzyskanych z użyciem PatternAnalyzer ma sentyment z przedziału (-0.1, 0>, a prawie wszystkie wyniki z tego przedziału to zera. Potwierdza to niezbyt rozległy zakres rozpoznawanych słów. Przez to rozkłady wartości dla niepożądanych i normalnych użytkowników są bardzo podobne (wykresy nr 13 i 15).

Wyniki osiągnięte z wykorzystaniem NaiveBayesAnalyzer widoczne na rysunku 16, tak samo jak w przypadku niepożądanych użytkowników wskazują znaczną przewagę wiadomości o sentymencie pozytywnym. Udział procentowy takich wiadomości jest jednak większy, tak samo jak osiągane przez nie wartości sentymentu.

Porównanie wyników



Rysunek 17: Porównanie udziałów procentowych przedziałów sentymentu wiadomości w obu badanych zbiorach. Wiadomości użytkowników normalnych przeważają w skrajnych przedziałach sentymentu. Duży udział wiadomości użytkowników niepożądanych w przedziale od -0,1 do 0,3.

Do porównania zobrazowanego na histogramie 17 użyto wyłącznie wyników wyliczanych z wykorzystaniem wariantu NaiveBayesAnalyzer. Pokazane zostało, że w przypadku obu zbiorów zdeklasował on mniej zaawansowane podejście. W dodatku mimo zwiększenia stopnia zaawansowania problemu czas przeprowadzenia analiz pozostał na rozsądnym poziomie. Widać, że zwykli użytkownicy przeważają w przedziałach skrajnych. Wartości osiągane przez użytkowników niepożądanych bardzo często są bliskie zera.

Tabela 7: Porównanie udziałów procentowych przedziałów sentymentu wiadomości w obu badanych zbiorach.

Przedział	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
(-1;-0,9]	1,696	1,183	0,513	1,696	1,183
(-0,9;-0,8]	2,469	1,917	0,552	4,166	3,101
(-0,8;-0,7]	2,893	2,341	0,552	7,059	5,441
(-0,7;-0,6]	3,139	2,678	0,462	10,198	8,119
(-0,6;-0,5]	3,661	3,063	0,597	13,858	11,182
(-0,5;-0,4]	3,830	3,417	0,414	17,689	14,599
(-0,4;-0,3]	4,111	3,885	0,227	21,800	18,483
(-0,3;-0,2]	4,317	4,522	-0,205	26,117	23,005
(-0,2;-0,1]	4,966	5,268	-0,301	31,083	28,273
(-0,1;0]	5,511	7,301	-1,790	36,594	35,574
(0,0;0,1]	5,523	7,164	-1,641	42,117	42,738
(0,1;0,2]	5,214	6,735	-1,521	47,331	49,473
(0,2;0,3]	5,412	6,383	-0,971	52,743	55,856
(0,3;0,4]	6,139	6,254	-0,114	58,882	62,109
(0,4;0,5]	5,262	6,073	-0,811	64,143	68,182
(0,5;0,6]	5,776	6,146	-0,370	69,919	74,328
(0,6;0,7]	6,314	5,978	0,335	76,233	80,307
(0,7;0,8]	6,822	6,109	0,713	83,055	86,416
(0,8;0,9]	8,016	6,377	1,639	91,072	92,793
(0,9;1,0]	8,928	7,207	1,721	100,000	100,000

Na podstawie tabeli 7 widzimy dokładnie, że udziały procentowe normalnych użytkowników przeważają w skrajnych przedziałach. Liczba ich wiadomości bardziej negatywnych i pozytywnych jest większa. Niepożądani użytkownicy mają przewagę w przedziałach pośrednich od -0,3 do 0,6. Różnice w każdym z przedziałów nie są jednak duże.

Osiągnięte wyniki sentymentu nie są takimi jakimi można byłoby się spodziewać. Wydawałoby się, że niepożądani użytkownicy, będą pisać więcej negatywnych wypowiedzi. Może to wynikać ze słabych zdolności do wykrywania negatywnych przypadków przez bibliotekę TextBlob. Zostało to sprawdzone w artykule [31], porównującym trzy podejścia. TextBlob miał wyraźne problemy z klasyfikacją negatywnych wiadomości. Inne podejścia, w tym jedno płatne, nie zaprezentowały się jednak dużo lepiej.

6.1.6. Subiektywność wypowiedzi

Obiektywność wypowiedzi jest wyliczana z wykorzystaniem biblioteki TextBlob, tak samo jak w przypadku wcześniej opisywanego sentymentu. Wykorzystywana jest tutaj biblioteka wzorców (linki prowadzące do źródła nie są już aktywne) zawierającą słowa z banku drzew Penn Treebank II. Na wyjściu otrzymujemy liczbę z zakresu od 0 do 1, gdzie 1 to całkowicie subiektywna wypowiedź. W przypadku gdy słowa są obiektywne lub nie zostaną odnalezione w słowniku otrzymamy wynik równy 0.

Przykłady zdań i wyniki obliczonej obiektywności:

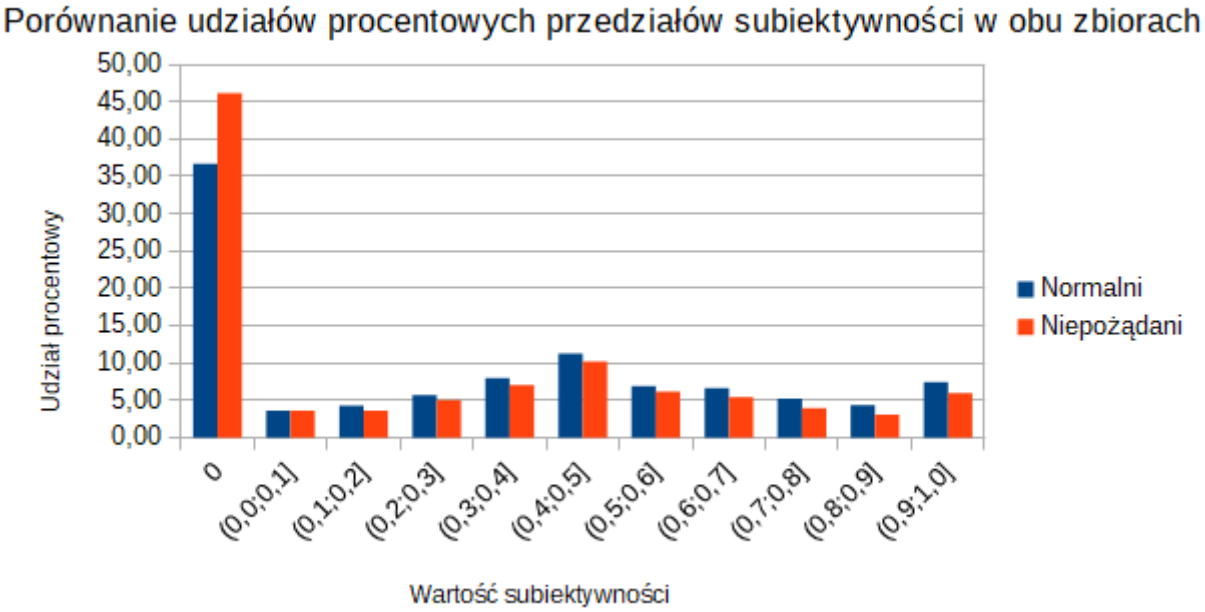
- “i made a trailer for american horror story: donald trump and it looks like the scariest show of all time . . . ”
  Obiektywność: 0.0;

- “Where is Obama? Stays silent on Chelsea explosion but talks election via @nypost #MAGA #AmericaFirst. . . ”
  Obiektywność: 0.1;

- “Lies and broken promises!”
  Obiektywność: 0.4;

- “great pic from the campaign trail, regardless of who you like.”
  Obiektywność: 0.75;

- “Warning: Being Famous Can Kill You.”
  Obiektywność: 1.0.



Rysunek 18: Porównanie udziałów procentowych przedziałów subiektywności wiadomości w obu zbiorach. W obu zbiorach przeważają wiadomości z zerową wartością cechy. Użytkownicy normalni piszą bardziej subiektywne wiadomości.

Analizując histogram 18 widać, że w przedziałach z niezerową wartością subiektywności nieznacznie przeważają wiadomości normalnych użytkowników. Piszą oni bardziej subiektywne wiadomości. Szczegółowy rozkład wartości widoczny jest w tabeli 8. Duży udział tweetów z zerową subiektywnością jest spowodowany tym, że wliczają się do niego również te wypowiedzi, w których słowa nie zostały znalezione w słowniku. Biblioteka TextBlob nie pozwala na zwrócenie informacji na temat tego, czy element z wartością równą zero został tak sklasyfikowany na podstawie dostępnej wiedzy, czy po prostu tej

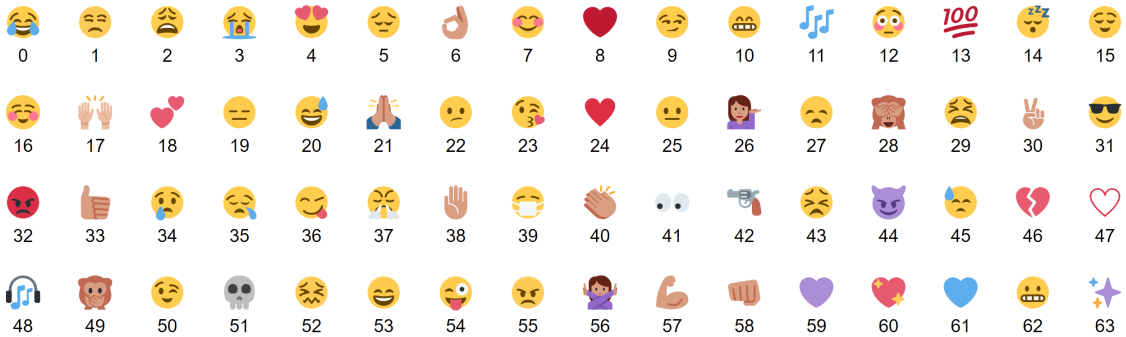
wiedzy nie było, a do słów zostały przypisane wagi równe 0. Nie można zatem wyciągnąć wniosków na temat obiektywności.

Tabela 8: Porównanie udziałów procentowych przedziałów subiektywności wiadomości w obu zbiorach

Przedział	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0	36,72	46,16	-9,44	36,72	46,16
(0,0;0,1]	3,58	3,60	-0,02	40,30	49,76
(0,1;0,2]	4,25	3,57	0,68	44,55	53,33
(0,2;0,3]	5,64	5,01	0,63	50,19	58,34
(0,3;0,4]	7,98	7,05	0,93	58,17	65,39
(0,4;0,5]	11,26	10,16	1,10	69,43	75,55
(0,5;0,6]	6,91	6,13	0,78	76,34	81,68
(0,6;0,7]	6,62	5,39	1,23	82,96	87,08
(0,7;0,8]	5,24	3,95	1,29	88,21	91,03
(0,8;0,9]	4,37	3,03	1,34	92,58	94,06
(0,9;1,0]	7,42	5,94	1,48	100,00	100,00

6.1.7. Emocje w wiadomościach

Do wykrywania emocji w wiadomościach został wykorzystany wytrenowany model o nazwie DeepMoji zaprezentowany w artykule [24]. Opracowano go z wykorzystaniem rekurencyjnych sieci neuronowych (dwukierunkowe LSTM) oraz danych uczących w większości otrzymanych z użyciem rozszerzonego zdalnego nadzoru (*distant supervision*). Zdalny nadzór [1] polega na tym, że na podstawie małej grupy otagowanych danych i opracowanych reguł dobierane są bez nadzoru człowieka dane z grupy nieotagowanych, co ostatecznie skutkuje powstaniem dużego zbioru, w którym część otagowanych elementów może być błędnie skategoryzowanych. Zakłada się jednak, że błędów będzie mało, a powstały przez nie szum nie wpłynie na końcowe wyniki. Z wykorzystaniem tego podejścia DeepMoji opiera się na bardzo dużym zbiorze uczącym składającym się z około 1246 milionów wiadomości. Wiadomości pochodzą z Twittera, więc ich forma jest podobna do analizowanych w tej pracy, co jest dużym plusem.



Rysunek 19: Obrazek przedstawia emocje wykrywane przez DeepMoji. Zbiór zawiera 64 emocje, zarówno te negatywne jak i pozytywne. Jako wynik analizy tekstu otrzymujemy numery, które możemy przełożyć na konkretną emocję zgodnie z obrazkiem. Źródło: [14].

DeepMoji pozwalana na identyfikację 64 emocji przedstawionych na obrazku 19. Są to emocje przedstawiane przez obrazki dostępne od wersji 6.0 Unicode. W repozytorium kodu projektu [14] dostępny jest plik mapujący numery emocji na ich kodowania w standardzie Unicode. Jako wynik analizy pojedynczej wiadomości otrzymujemy 5 różnych emocji, które najbardziej się wyróżniają, wraz z ich prawdopodobieństwami/udziałami w analizowanym tekście.

Przykładowe wyniki analiz (suma udziałów 5 emocji, 5 liczb naturalnych odpowiadających emocjom, 5 liczb rzeczywistych odpowiadających kolejno udziałom wykrytych emocji):

- "Why Trump? Because what the government has done is disgusting. America is in the fight of its life against."
  0.45203037932515144, 32, 55, 39, 37, 58, 0.17620103, 0.09780312, 0.07062909, 0.058798708, 0.048598427;

- "Hard To Believe, But Trump Could Win. May God have merci on our collective souls!"
  0.42710485123097897, 21, 34, 46, 17, 61, 0.27879155, 0.048960235, 0.037022427, 0.031843957, 0.030486686;

- "Hi America! Lifelong New Yorker here to let you know that i am more afraid of Donald Trump than Muslims."
  0.5854615271091461, 8, 47, 24, 61, 59, 0.21749094, 0.13908578, 0.08381305, 0.08028674, 0.06478501;

- "Thank you @FBCJAX for welcoming us in Jacksonville! It was a joy to worship with you. Appreciate your prayers"
  0.5247159004211426, 21, 17, 8, 47, 7, 0.2344473, 0.092608094, 0.07898577, 0.06861961, 0.050055124;

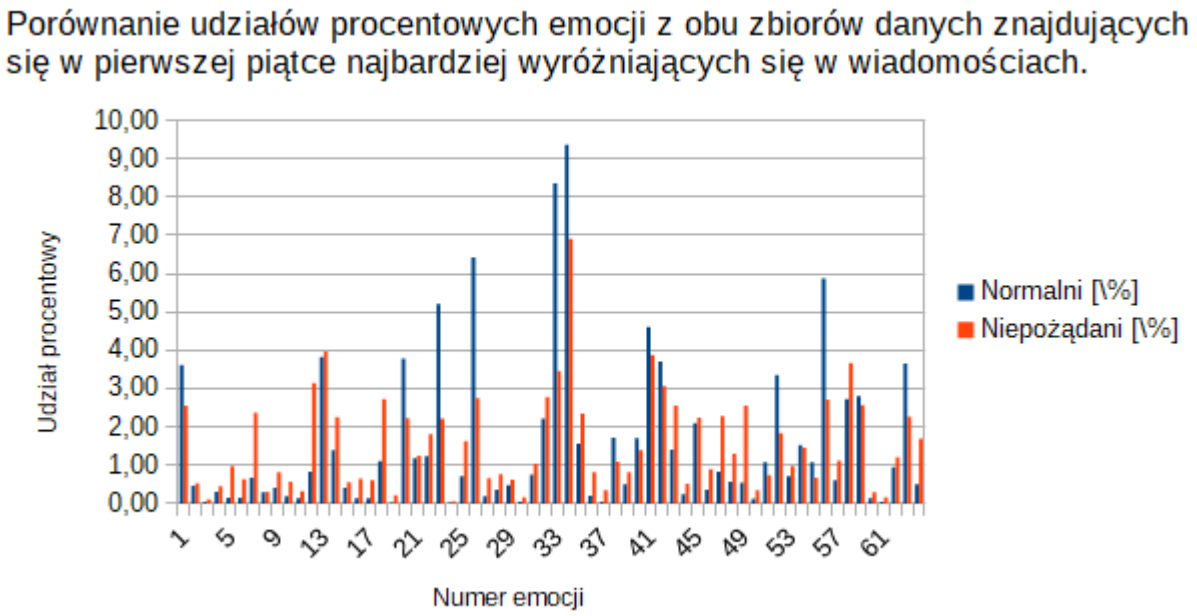
- "Illegal Immigrants is not a race, it's a CRIME."
  0.3702537715435028, 42, 56, 55, 44, 32, 0.12169931, 0.08720044, 0.055996165, 0.05320464, 0.052153215;

- "In 1999 Jesse Jackson Praised Donald Trump's Commitment To Minorities and Under-Served Communities"
  0.6273605339229107, 24, 16, 47, 48, 11, 0.28448877, 0.13667539, 0.1335447, 0.038982477, 0.033669204;

- "@mike_pence You're not deplorable your colossal lies to @MarthaRaddatz on @ThisWeekABC makes you DESPICABLE. Your family should be ashamed."
  0.3928871378302574, 55, 32, 38, 37, 22, 0.14391151, 0.13625593, 0.04082665, 0.03689356, 0.034999482.

Emocje najczęściej występujące w pierwszej piątce wyróżnionych z wiadomości

Zliczane są wystąpienia każdej emocji w pierwszej piątce najbardziej wyróżniających się emocji znalezionych w wiadomościach. Kolejność w piątce nie ma znaczenia.



Rysunek 20: Histogram porównuje udziały procentowe emocji znajdujących się w pierwszej piątce najbardziej wyróżniających się dla wiadomości ze zbiorów normalnych i niepożądanych użytkowników

Wyraźnie widać, że wiele emocji jest popularnych w obu zbiorach. Zaproponowany w DeepMoji podział na aż 64 emocje może sugerować, że część z nich o bardzo małych udziałach w badanych zbiorach jest bardzo charakterystyczna i ciężko wykrywalna lub po prostu za bardzo zbliżona do innych. W artykule dotyczącym biblioteki [24] są zwizualizowane wyniki klastrowania hierarchicznego, pokazujące podobieństwa między emocjami.

W przypadku niektórych emocji widoczne są znaczące różnice. Szczegółowe udziały procentowe z wykresu 20 są przedstawione w tabeli 9.

Najczęściej emocje, które występowały w zbiorze normalnych użytkowników występują też często w zbiorze niepożądanych. Dokonano interpretacji najpopularniejszych emocji z wykorzystaniem dwóch encyklopedii emocji: [3] oraz [6].

Emocje występujące w obu zbiorach z udziałem ponad 5%:

- 34 - duży smutek, płacz.

Emocje, których udział w zbiorze niepożądanych użytkowników jest z przedziału 2%-5%, a w przypadku normalnych użytkowników wyraźnie ponad 5%:

- 23 - wysłanie buziaka;
- 26 - używany na końcu zdania ze znaczeniem: “co o tym myślisz”, “tak?”;
- 33 - poparcie;
- 56 - zatrzymaj się, robisz coś źle.

Emocje występujące w obu zbiorach o udziałach pomiędzy 2% i 5%:

- 1 - grymas i dezaprobata;
- 13 - doskonały wynik, wspaniałe osiągnięcie;
- 20 - radość, uśmiech, lekki stres;
- 32 - nadąsany, zdenerwowany;
- 41 - obserwowanie kogoś lub czegoś;
- 42 - przemoc, gorszy okres życia;
- 45 - grymas, duży stres;
- 58 - zaciśnięta pięść, agresja wobec innej osoby;
- 59 - wrażliwa, wyrozumiała i współczująca miłość;
- 63 - podekscytowanie, podziw, gratulacje, magia.

Emocje, których udział w zbiorze niepożądanych użytkowników wynosi od 2% do 5%, a w zbiorze normalnych ich udział jest o wiele mniejszy:

- 7 - radość, uśmiech;
- 12 - zawstydzenie;
- 14 - senność, znużenie;
- 18 - mocne wyrażenie miłości;
- 35 - senność w nieodpowiednim momencie;
- 43 - wytrwałość w podjętych trudnych decyzjach, czynnościach;
- 47 - szczere uczucia;
- 49 - cytując znaczenie: “nic nie widziałem, nic nie słyszałem, nie mów nic złego”.

Emocje występujące w zbiorze normalnych użytkowników z udziałem pomiędzy 2% i 5%, których udział w drugim zbiorze jest wyraźnie mniejszy:

- 52 - zatrzymanie w sobie wielu emocji np. irytacji, frustracji, wstrętu.

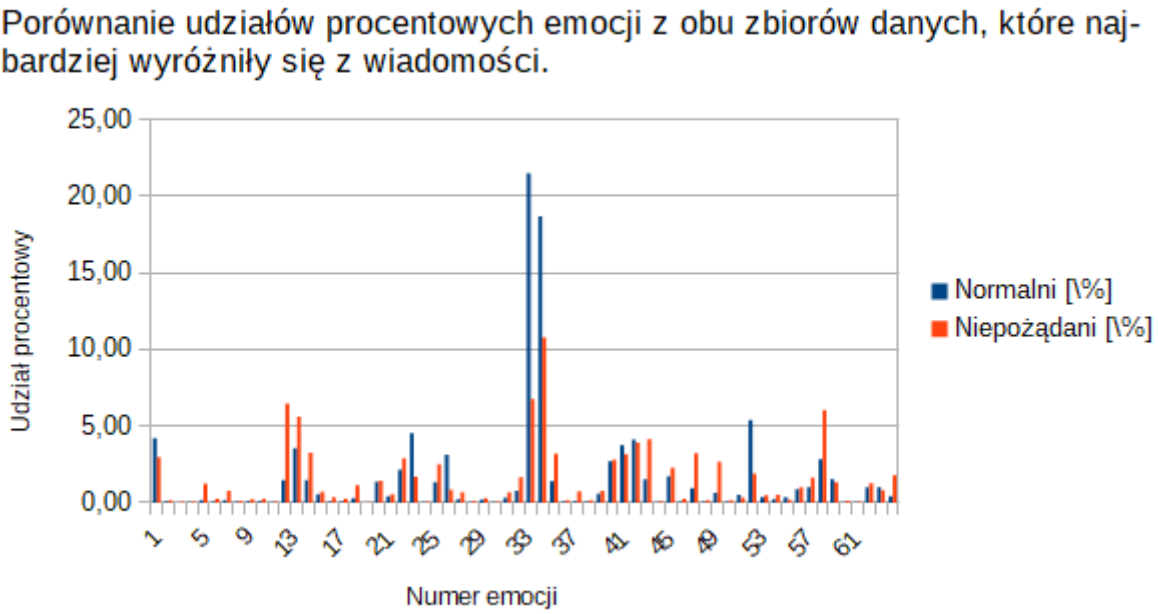
W zbiorze wiadomości niepożądanych użytkowników zostało wyróżnione o wiele więcej częściej używanych emocji. W pierwszej piątce z udziałami ponad 1% występuje 27 emocji w przypadku normalnych użytkowników oraz 35 w przypadku niepożądanych. Widać zatem, że użytkownicy normalni preferują bardziej ograniczony zbiór, przez co aż 5 emocji osiągnęło udział ponad 5%. Dla porównania w drugim zbiorze była tylko jedna taka emocja. Z analizy nie można powiedzieć, że któraś z grup wypowiada się bardziej negatywnie. Tweety obu grup zawierają zarówno pozytywne i negatywne emocje. Wydaje się, że użytkownicy normalni są bardziej konkretni, ponieważ częściej wyrażają poparcie, zwracają innym uwagę czy prowadzą dyskusję za czym przemawiają wykryte u nich najczęściej występujące emocje.

Tabela 9: Porównanie udziałów procentowych emocji znajdujących się w pierwszej piątce najbardziej wyróżniających się dla wiadomości ze zbiorów normalnych i niepożądanych użytkowników.

Nr emocji	Normalni, %	Niepożądani, %	Różnica, %
1	3,61	2,56	1,06
2	0,46	0,54	-0,07
3	0,05	0,12	-0,08
4	0,31	0,45	-0,13
5	0,15	0,98	-0,83
6	0,15	0,64	-0,48
7	0,68	2,38	-1,70
8	0,30	0,31	-0,01
9	0,42	0,82	-0,40
10	0,19	0,57	-0,37
11	0,15	0,32	-0,18
12	0,83	3,14	-2,31
13	3,82	3,97	-0,15
14	1,39	2,25	-0,86
15	0,42	0,55	-0,14
16	0,14	0,65	-0,50
17	0,14	0,61	-0,47
18	1,11	2,73	-1,62
19	0,02	0,22	-0,20
20	3,78	2,23	1,55
21	1,18	1,25	-0,07
22	1,24	1,82	-0,58
23	5,21	2,22	2,99
24	0,03	0,06	-0,04
25	0,72	1,63	-0,91
26	6,42	2,75	3,68
27	0,19	0,66	-0,47
28	0,36	0,77	-0,41
29	0,47	0,62	-0,14
30	0,06	0,17	-0,11
31	0,75	1,04	-0,29
32	2,21	2,78	-0,57
33	8,36	3,45	4,91
34	9,37	6,90	2,46
35	1,56	2,34	-0,78
36	0,21	0,82	-0,61
37	0,05	0,34	-0,29
38	1,72	1,09	0,63
39	0,50	0,82	-0,31
40	1,70	1,39	0,31
41	4,60	3,88	0,73
42	3,70	3,07	0,63
43	1,40	2,55	-1,15
44	0,24	0,51	-0,28
45	2,10	2,25	-0,15
46	0,36	0,90	-0,54
47	0,83	2,28	-1,45
48	0,57	1,30	-0,72
49	0,55	2,55	-2,00
50	0,13	0,35	-0,21
51	1,08	0,74	0,34
52	3,36	1,84	1,52
53	0,72	0,98	-0,26
54	1,52	1,45	0,07
55	1,08	0,68	0,40
56	5,87	2,72	3,16
57	0,61	1,12	-0,52
58	2,72	3,67	-0,94
59	2,81	2,57	0,25
60	0,14	0,30	-0,16
61	0,05	0,17	-0,12
62	0,95	1,21	-0,26
63	3,65	2,27	1,38
64	0,51	1,69	-1,19

Najbardziej wyróżniające się emocje w wiadomościach

Zliczane są wystąpienia każdej emocji wyłącznie na pierwszym miejscu listy najbardziej wyróżniających się emocji znalezionych w wiadomościach.



Rysunek 21: Histogram porównuje udziały procentowe emocji, które najbardziej wyróżniły się w wiadomościach ze zbiorów normalnych i niepożądanych użytkowników

Tak samo jak już wcześniej stwierdzono, użytkownicy normalni ograniczają się do wąskiego zestawu emocji. Udział procentowy powyżej 1% w gronie najbardziej wyróżniających się emocji mają 22 emocje ze zbioru normalnych użytkowników oraz 28 ze zbioru niepożądanych. Na wykresie 21 pokazano udziały najbardziej wyróżniających się emocji w obu zbiorach. Najbardziej wyraźne w wypowiedziach normalnych użytkowników są emocje o numerach: 33 - udzielenie poparcia oraz 34 - duży smutek, płacz. Obie emocje, zwłaszcza druga, wykrywane są bardzo często również w drugim zbiorze. Ich udziały są jednak o wiele mniejsze. Równie popularne w tweetach niepożądanych użytkowników są emocje nr 12 - zawstydzenie, 13 - doskonały wynik i wspaniałe osiągnięcie, 58 - zaciśnięta pięść, agresja wobec innej osoby. Ostatnia z emocji o numerze 58, która jest wyraźnie negatywna, występuje jako najbardziej rozpoznawalna w o wiele większej liczbie wiadomości ze zbiory niepożądanych niż normalnych użytkowników. Szczegółowe udziały procentowe emocji pokazanych na wykresie 21 pokazane są w tabeli 10.

Tabela 10: Porównanie udziałów procentowych emocji, które najbardziej wyróżniły się w wiadomościach ze zbiorów normalnych i niepożądanych użytkowników.

Nr emocji	Normalni, %	Niepożądani, %	Różnica, %
1	4,21	2,96	1,25
2	0,09	0,16	-0,08
3	0,00	0,04	-0,04
4	0,02	0,05	-0,03
5	0,15	1,24	-1,09
6	0,05	0,25	-0,21
7	0,16	0,76	-0,60
8	0,05	0,14	-0,09
9	0,08	0,22	-0,14
10	0,08	0,26	-0,19
11	0,01	0,06	-0,05
12	1,45	6,45	-5,00
13	3,55	5,62	-2,07
14	1,45	3,25	-1,79
15	0,53	0,70	-0,17
16	0,05	0,35	-0,30
17	0,04	0,26	-0,22
18	0,27	1,12	-0,85
19	0,00	0,02	-0,02
20	1,36	1,42	-0,06
21	0,41	0,55	-0,13
22	2,15	2,88	-0,73
23	4,54	1,67	2,87
24	0,03	0,05	-0,03
25	1,33	2,51	-1,18
26	3,12	0,83	2,29
27	0,22	0,67	-0,44
28	0,02	0,10	-0,08
29	0,20	0,28	-0,09
30	0,01	0,04	-0,04
31	0,31	0,67	-0,36
32	0,79	1,65	-0,86
33	21,50	6,77	14,73
34	18,68	10,78	7,90
35	1,39	3,20	-1,82
36	0,05	0,14	-0,09
37	0,08	0,75	-0,67
38	0,04	0,15	-0,11
39	0,59	0,79	-0,20
40	2,69	2,79	-0,11
41	3,73	3,15	0,58
42	4,12	3,90	0,21
43	1,52	4,15	-2,63
44	0,02	0,09	-0,07
45	1,71	2,27	-0,56
46	0,06	0,26	-0,20
47	0,94	3,21	-2,27
48	0,07	0,16	-0,09
49	0,63	2,65	-2,02
50	0,05	0,17	-0,12
51	0,49	0,33	0,17
52	5,37	1,88	3,49
53	0,35	0,47	-0,11
54	0,20	0,50	-0,29
55	0,33	0,20	0,13
56	0,86	1,01	-0,16
57	0,99	1,62	-0,63
58	2,84	6,04	-3,20
59	1,52	1,33	0,19
60	0,01	0,11	-0,10
61	0,01	0,05	-0,04
62	1,00	1,27	-0,27
63	1,01	0,79	0,22
64	0,43	1,80	-1,37

6.1.8. Analiza dwuwymiarowa sentymentu i subiektywności

Analiza opiera się na zbadaniu połączonych wyników analiz 6.1.5 i 6.1.6.

Niepożądani użytkownicy

Tabela 11: Dwuwymiarowa analiza sentymentu i subiektywności wiadomości niepożądanych użytkowników. Podane w tabeli wyniki to udziały procentowe poszczególnych komórek.

		Subiektywność												
		0,0	(0,0;0,1]	(0,1;0,2]	(0,2;0,3]	(0,3;0,4]	(0,4;0,5]	(0,5;0,6]	(0,6;0,7]	(0,7;0,8]	(0,8;0,9]	(0,9;1,0]	Ogółem	Suma
Sentyment	(-1;-0,9]	0,54	0,04	0,04	0,06	0,08	0,11	0,06	0,08	0,05	0,04	0,09	1,18	1,18
	(-0,9;-0,8]	0,88	0,07	0,06	0,09	0,13	0,19	0,11	0,13	0,07	0,06	0,12	1,92	3,10
	(-0,8;-0,7]	1,08	0,08	0,08	0,12	0,16	0,23	0,10	0,17	0,09	0,08	0,15	2,34	5,44
	(-0,7;-0,6]	1,32	0,09	0,09	0,11	0,17	0,24	0,15	0,15	0,09	0,08	0,18	2,68	8,12
	(-0,6;-0,5]	1,54	0,11	0,10	0,14	0,21	0,28	0,17	0,15	0,10	0,08	0,18	3,06	11,18
	(-0,5;-0,4]	1,64	0,11	0,11	0,15	0,24	0,33	0,20	0,20	0,12	0,10	0,22	3,42	14,60
	(-0,4;-0,3]	1,87	0,14	0,13	0,17	0,25	0,37	0,20	0,21	0,16	0,11	0,27	3,88	18,48
	(-0,3;-0,2]	2,24	0,15	0,15	0,19	0,28	0,41	0,27	0,24	0,17	0,13	0,28	4,52	23,00
	(-0,2;-0,1]	2,65	0,18	0,17	0,22	0,31	0,48	0,29	0,32	0,17	0,18	0,31	5,27	28,27
	(-0,1;0]	4,12	0,22	0,24	0,25	0,36	0,59	0,38	0,33	0,21	0,21	0,39	7,30	35,57
	(0,0;0,1]	3,52	0,22	0,22	0,33	0,44	0,67	0,53	0,33	0,27	0,22	0,42	7,16	42,74
	(0,1;0,2]	3,06	0,20	0,24	0,33	0,51	0,62	0,47	0,37	0,31	0,20	0,43	6,73	49,47
	(0,2;0,3]	2,89	0,20	0,24	0,33	0,44	0,64	0,42	0,36	0,28	0,23	0,37	6,38	55,86
	(0,3;0,4]	2,72	0,21	0,20	0,36	0,50	0,64	0,40	0,36	0,26	0,19	0,43	6,25	62,11
	(0,4;0,5]	2,60	0,21	0,23	0,33	0,46	0,62	0,40	0,34	0,26	0,20	0,42	6,07	68,18
	(0,5;0,6]	2,60	0,28	0,23	0,33	0,48	0,65	0,35	0,33	0,32	0,21	0,36	6,15	74,33
	(0,6;0,7]	2,59	0,23	0,23	0,34	0,47	0,70	0,36	0,30	0,25	0,18	0,33	5,98	80,31
	(0,7;0,8]	2,55	0,25	0,24	0,36	0,47	0,74	0,40	0,33	0,25	0,18	0,34	6,11	86,42
	(0,8;0,9]	2,75	0,28	0,27	0,38	0,51	0,77	0,36	0,34	0,24	0,18	0,31	6,38	92,79
	(0,9;1,0]	3,00	0,31	0,33	0,43	0,57	0,90	0,49	0,36	0,28	0,17	0,36	7,21	100,00
	Ogółem	46,17	3,60	3,57	5,01	7,05	10,16	6,13	5,39	3,95	3,03	5,94	100,00	
	Suma	46,17	49,76	53,34	58,35	65,40	75,56	81,69	87,08	91,03	94,06	100,00		

W tabeli 11 widzimy dużo większych wartości w pierwszej kolumnie wyników. Szczególnie duży udział mają wiadomości z sentymentem z przedziału od -0,1 do 0,2. Możemy zauważyć, że najczęściej wyróżniające się udziały, z pominięciem 1 kolumny, znajdują się w centrum tabeli. Dla kolumny z przedziałem subiektywności (0,4;0,5] udziały rosną ku dołowi wraz ze zwiększaniem sentymentu. Widać również, że udziały komórek tabeli dla subiektywności powyżej 0,3 i sentymentu powyżej 0,0 są większe, w porównaniu do pozostałych obszarów tabeli z pominięciem pierwszej kolumny wyników.

Zwykli użytkownicy

Tabela 12: Dwuwymiarowa analiza sentymentu i subiektywności wiadomości normalnych użytkowników. Podane w tabeli wyniki to udziały procentowe poszczególnych komórek.

		Subiektywność												
		0,0	(0,0;0,1]	(0,1;0,2]	(0,2;0,3]	(0,3;0,4]	(0,4;0,5]	(0,5;0,6]	(0,6;0,7]	(0,7;0,8]	(0,8;0,9]	(0,9;1,0]	Ogółem	Suma
Sentyment	(-1;-0,9]	0,57	0,04	0,04	0,07	0,11	0,18	0,11	0,13	0,10	0,13	0,22	1,69	1,69
	(-0,9;-0,8]	0,99	0,07	0,11	0,09	0,15	0,27	0,15	0,22	0,13	0,09	0,19	2,47	4,16
	(-0,8;-0,7]	0,92	0,09	0,21	0,15	0,24	0,30	0,18	0,20	0,17	0,16	0,27	2,89	7,05
	(-0,7;-0,6]	1,24	0,09	0,16	0,10	0,20	0,34	0,22	0,22	0,18	0,13	0,25	3,13	10,18
	(-0,6;-0,5]	1,34	0,11	0,12	0,29	0,27	0,31	0,27	0,30	0,21	0,13	0,30	3,66	13,84
	(-0,5;-0,4]	1,59	0,09	0,14	0,17	0,25	0,38	0,41	0,21	0,20	0,15	0,25	3,84	17,68
	(-0,4;-0,3]	1,61	0,21	0,19	0,21	0,30	0,41	0,25	0,26	0,19	0,18	0,33	4,12	21,80
	(-0,3;-0,2]	1,89	0,15	0,19	0,17	0,31	0,40	0,30	0,23	0,22	0,14	0,33	4,31	26,10
	(-0,2;-0,1]	2,00	0,16	0,19	0,26	0,30	0,65	0,33	0,29	0,26	0,24	0,32	4,98	31,09
	(-0,1;0]	2,51	0,15	0,21	0,19	0,31	0,60	0,35	0,31	0,20	0,23	0,45	5,50	36,59
	(0,0;0,1]	2,22	0,18	0,25	0,29	0,40	0,56	0,35	0,38	0,23	0,28	0,36	5,51	42,10
	(0,1;0,2]	2,04	0,19	0,21	0,31	0,51	0,51	0,29	0,36	0,28	0,22	0,31	5,22	47,32
	(0,2;0,3]	2,23	0,17	0,18	0,24	0,42	0,55	0,38	0,30	0,36	0,27	0,30	5,40	52,72
	(0,3;0,4]	2,09	0,20	0,29	0,75	0,46	0,58	0,32	0,27	0,27	0,37	0,54	6,13	58,85
	(0,4;0,5]	1,90	0,21	0,22	0,25	0,36	0,56	0,42	0,36	0,35	0,31	0,33	5,25	64,10
	(0,5;0,6]	1,92	0,23	0,23	0,38	0,50	0,71	0,39	0,41	0,29	0,26	0,45	5,78	69,88
	(0,6;0,7]	2,22	0,35	0,34	0,33	0,65	0,67	0,51	0,43	0,29	0,18	0,37	6,33	76,21
	(0,7;0,8]	2,16	0,22	0,26	0,42	0,60	1,02	0,43	0,45	0,39	0,30	0,60	6,83	83,04
	(0,8;0,9]	2,56	0,31	0,33	0,46	0,69	1,13	0,68	0,61	0,39	0,23	0,63	8,02	91,06
	(0,9;1,0]	2,67	0,36	0,39	0,54	1,00	1,16	0,61	0,70	0,53	0,37	0,62	8,94	100,00
	Ogółem	36,66	3,57	4,26	5,65	8,00	11,28	6,93	6,62	5,24	4,37	7,42	100,00	
	Suma	36,66	40,23	44,49	50,15	58,14	69,42	76,36	82,98	88,21	92,58	100,00		

Podobnie jak w przypadku niepożądanych użytkowników, pierwsza kolumna tabeli 12 ma bardzo dużo wartości z większymi udziałami procentowymi. Występuje w niej skupienie wartości wokół zerowego sentymentu, lecz jest mniejsze niż w poprzednim zbiorze. Tak samo w centrum tabeli obserwujemy większe wartości. Widoczna jest przewaga udziałów obszaru tabeli dla przedziałów subiektywności większych od 0,4 i sentymentu większego od -0,2.

Porównanie wyników

Tabela 13: Różnice udziałów procentowych kategorii dwuwymiarowej analizy dla obu zbiorów. Od wartości z tabeli normalnych użytkowników odjęto wartości niepożądanych użytkowników.

		Subiektywność											
		0,0	(0,0;0,1]	(0,1;0,2]	(0,2;0,3]	(0,3;0,4]	(0,4;0,5]	(0,5;0,6]	(0,6;0,7]	(0,7;0,8]	(0,8;0,9]	(0,9;1,0]	Ogółem
Sentyment	(-1;-0,9]	0,03	0,01	0,00	0,01	0,03	0,07	0,05	0,05	0,05	0,09	0,13	0,51
	(-0,9;-0,8]	0,11	0,00	0,05	0,00	0,01	0,08	0,04	0,09	0,06	0,03	0,07	0,55
	(-0,8;-0,7]	-0,17	0,02	0,13	0,04	0,07	0,07	0,08	0,04	0,07	0,08	0,12	0,55
	(-0,7;-0,6]	-0,08	0,00	0,07	-0,01	0,02	0,10	0,07	0,07	0,08	0,05	0,07	0,46
	(-0,6;-0,5]	-0,20	0,00	0,02	0,15	0,06	0,03	0,11	0,15	0,11	0,05	0,12	0,59
	(-0,5;-0,4]	-0,05	-0,02	0,03	0,02	0,02	0,04	0,21	0,01	0,08	0,06	0,03	0,42
	(-0,4;-0,3]	-0,26	0,06	0,06	0,04	0,04	0,04	0,04	0,05	0,03	0,06	0,05	0,23
	(-0,3;-0,2]	-0,35	-0,01	0,04	-0,02	0,03	-0,01	0,02	-0,01	0,05	0,01	0,05	-0,21
	(-0,2;-0,1]	-0,65	-0,02	0,03	0,04	-0,02	0,17	0,04	-0,03	0,09	0,05	0,01	-0,28
	(-0,1;0]	-1,61	-0,07	-0,03	-0,07	-0,04	0,01	-0,03	-0,02	-0,01	0,02	0,06	-1,80
	(0,0;0,1]	-1,30	-0,04	0,03	-0,04	-0,04	-0,11	-0,18	0,05	-0,04	0,05	-0,06	-1,65
	(0,1;0,2]	-1,02	-0,02	-0,03	-0,03	0,00	-0,11	-0,17	-0,01	-0,03	0,02	-0,12	-1,52
	(0,2;0,3]	-0,66	-0,03	-0,05	-0,09	-0,02	-0,09	-0,04	-0,06	0,08	0,04	-0,06	-0,98
	(0,3;0,4]	-0,63	-0,01	0,09	0,39	-0,04	-0,06	-0,09	-0,09	0,01	0,18	0,12	-0,12
	(0,4;0,5]	-0,70	-0,01	-0,01	-0,08	-0,10	-0,06	0,02	0,02	0,09	0,11	-0,09	-0,82
	(0,5;0,6]	-0,68	-0,05	0,00	0,05	0,02	0,06	0,04	0,08	-0,03	0,05	0,09	-0,36
	(0,6;0,7]	-0,37	0,11	0,11	-0,01	0,18	-0,02	0,14	0,13	0,04	0,00	0,04	0,35
	(0,7;0,8]	-0,39	-0,03	0,02	0,05	0,13	0,28	0,02	0,11	0,14	0,12	0,27	0,73
	(0,8;0,9]	-0,19	0,03	0,06	0,08	0,19	0,36	0,31	0,27	0,15	0,06	0,32	1,64
	(0,9;1,0]	-0,33	0,05	0,06	0,11	0,43	0,25	0,12	0,34	0,25	0,20	0,26	1,74
	Ogółem	-9,51	-0,02	0,69	0,64	0,95	1,12	0,80	1,23	1,28	1,34	1,48	0,00

Tabela 13 pokazuje różnice udziałów poszczególnych komórek pomiędzy zbiorami. Widoczna jest przewaga niepożądanych użytkowników w skrajnie lewej części tabeli oraz w jej centrum. Normalni użytkownicy przeważają wyraźnie w dwóch ostatnich przedziałach sentymentu w komórkach, dla których wartość subiektywność wiadomości jest większa od 0,3.

6.1.9. Średnie podobieństwo semantyczne wiadomości

Wykorzystywany jest moduł uczenia maszynowego dostępny w TensorFlow Hub, konkretnie Universal Sentence Encoder. Pozwala zamienić zdanie na jego reprezentację liczbową w postaci wektora liczb rzeczywistych (sentence embedding). Został opisany w artykule [21]. Udostępnia dwa wytrenowane modele, różniące się precyzją, stopniem skomplikowania oraz użyciem zasobów obliczeniowych:

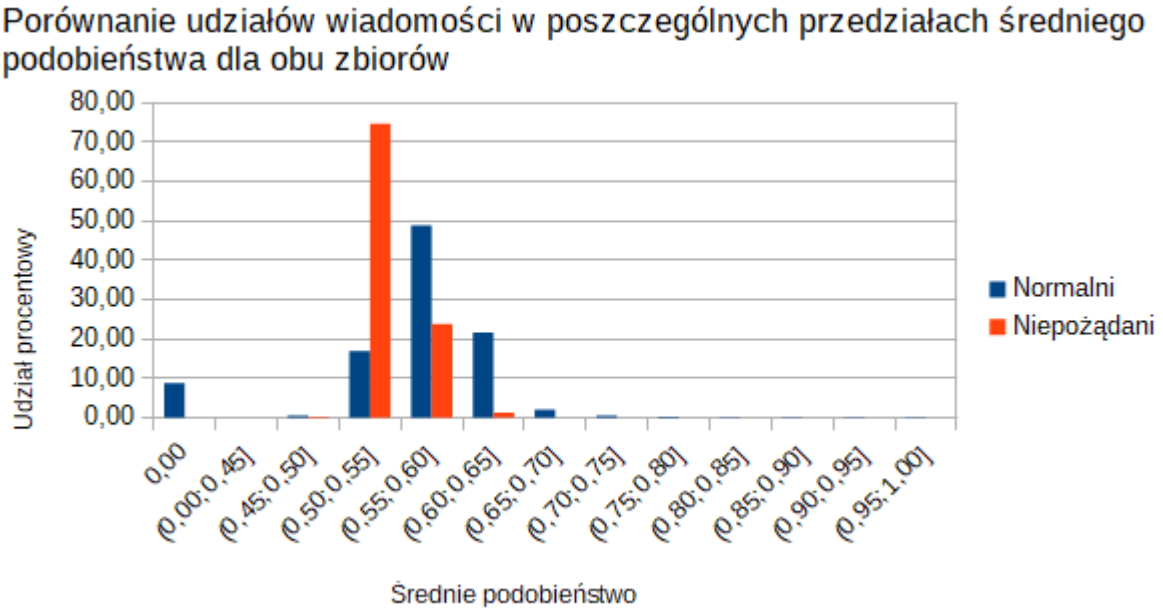
- Transformer encoder - większa precyzja, obarczona dużym użyciem zasobów. Złożoność pamięciowa i czasowa jest kwadratowa. Oparty na analizie podgrafu transformacji. W grafie w celu obliczenia reprezentacji kontekstowej słowa zwracana jest uwaga na inne słowa współwystępujące w zdaniu oraz ich kolejność. Reprezentacje kontekstowe słów są następnie konwerterowane na stałej długości wektory poprzez obliczanie elementarnej sumy reprezentacji słowa na każdej pozycji. Jako wyjście otrzymujemy 512-wymiarowy wektor. Jego długość jest stała bez względu na długość zdania.

- Deep averaging network (DAN) - gorsza jakość wyników, lecz o wiele lepsza wydajność. Liniowa złożoność pamięciowa i czasowa. Wejściowe wektory słów (word embeddings) i bi-gramów są razem uśredniane, a następnie podawane na wejście wyuczonej głebokiej sieci neuronowej, z której również otrzymamy 512-wymiarowy wektor reprezentujący zdanie dowolnej długości.

W celu zbadania użytkowników została wykorzystana mniej dokładna wersja wykorzystująca sieci neuronowe. Nie było możliwe przeprowadzenie bardziej dokładnej analizy dla całych zbiorów danych z powodu zbyt dużego użycia zasobów i czasu obliczeń. Porównując wyniki pomiędzy dwoma podejściami dla tych samych danych widać przewagę bardziej zaawansowanego rozwiązania w przypadku niektórych wyników. Dla pozostałych różnice są bardzo nieznaczne.

Dla otrzymanych wektorów obliczane jest podobieństwo cosinus na podstawie, którego obliczana jest odległość kątowa pomiędzy wektorami. Wartości funkcji cosinus dla dwóch wektorów mogą być bardzo bliskie siebie. Aby rozkład był bardziej przejrzysty wykorzystywane są odległości kątowe, co zostało zaproponowane przez autorów w artykule [21].

Do obliczeń został wykorzystany universal sentence encoder w najnowszej dostępnej wersji 4. Zostały również przetestowane poprzednie wersje biblioteki. Osiągały one bardzo zbliżone wyniki.



Rysunek 22: Wykres przedstawia porównanie udziałów procentowych wiadomości w poszczególnych przedziałach średniego podobieństwa wiadomości. Dziwne jest to, że normalni użytkownicy piszą bardziej podobne wiadomości. Bardzo możliwe jest, że wypowiadając się na temat używają podobnych słów, co sprawia, że wzrasta ogólne podobieństwo wypowiedzi.

Otrzymane wyniki zwizualizowane na wykresie 22 wskazują na to, że użytkownicy normalni piszą bardziej podobne wiadomości. Powodem tego, może być fakt badania wiado-

mości o treściach politycznych. Jeśli piszemy na temat i nie odbiegamy od głównego nurtu dyskusji, możemy użyć wiele takich samych słów. W zbiorze normalnych użytkowników wystąpiło również sporo krótkich cytatów wypowiedzi polityków, które mogły wpłynąć na taki wynik. W przypadku niepożądanych użytkowników o wiele częściej można w ich publikacjach wyszukać wiadomości na inny temat, a co za tym idzie mają większy zakres słów możliwych do wykorzystania.

Duża liczba wiadomości o wysokim średnim podobieństwie ze zbioru potencjalnie normalnych użytkowników może wynikać z faktu, że niektórzy użytkownicy napisali w badanym czasie mało wiadomości. W przypadku opublikowania dwóch identycznych wiadomości w odstępie kilku dni, ich średnie podobieństwo będzie równe 1,0. Patrząc na wysokie wyniki ostaniach przedziałów normalnych użytkowników w tabeli 14 możemy jednak powiedzieć, że w zbiorze potencjalnie normalnych użytkowników także możemy odszukać spamerów. Wiadomości zaliczające się do wysokich przedziałów podobieństwa zostały sprawdzone i były one często identyczne. W tej grupie można zwrócić uwagę na nieco mądrzej działających spamerów, którzy w celu obniżenia podobieństwa podmieniają różne fragmenty tej samej wypowiedzi. Wyróżniający się użytkownicy zostaną dokładniej opisani w analizie 6.1.10 oraz w punkcie poświęconym stworzeniu finalnego zbioru, w którym również zostaną odfiltrowani.

Tabela 14: Porównanie udziałów procentowych wiadomości z poszczególnych przedziałów średniego podobieństwa wiadomości

Przedział	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0,00	8,79	0,06	8,74	8,79	0,06
(0,00;0,45]	0,00	0,00	0,00	8,80	0,06
(0,45;0,50]	0,54	0,25	0,29	9,33	0,31
(0,50;0,55]	16,92	74,58	-57,66	26,25	74,89
(0,55;0,60]	48,77	23,71	25,06	75,02	98,60
(0,60;0,65]	21,56	1,24	20,31	96,58	99,84
(0,65;0,70]	2,01	0,07	1,94	98,59	99,91
(0,70;0,75]	0,52	0,03	0,49	99,10	99,94
(0,75;0,80]	0,25	0,00	0,24	99,35	99,94
(0,80;0,85]	0,14	0,04	0,11	99,49	99,98
(0,85;0,90]	0,15	0,02	0,14	99,65	100,00
(0,90;0,95]	0,19	0,00	0,18	99,84	100,00
(0,95;1,00]	0,16	0,00	0,16	100,00	100,00

6.1.10. Liczba wiadomości użytkownika podobnych semantycznie

Analiza jest podobna do opisanej w punkcie 6.1.9. Wykorzystywane są te same narzędzia. Jedyną różnicą jest to, że zamiast liczyć średnie podobieństwo wiadomości, zliczane są wiadomości użytkownika, których podobieństwo do jednej z jego badanych wiadomości przekracza określony próg. Wartości podobieństwa dwóch wiadomości wahają się od około 0,45 do 1,00. Dwa badane progi mają wartości 0,7 i 0,8 w tej samej skali.

Tabela 15: Porównanie udziałów procentowych wiadomości z określoną liczbą wiadomości podobnych do nich dla obu zbiorów. Próg podobieństwa równy 0,8.

Liczba wiadomości	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0	93,565	93,185	0,380	93,565	93,185
1	3,363	4,352	-0,989	96,928	97,537
2	0,711	1,070	-0,360	97,639	98,608
3	0,341	0,405	-0,064	97,980	99,013
4	0,181	0,180	0,001	98,160	99,192
5	0,140	0,102	0,038	98,301	99,294
6	0,100	0,057	0,043	98,401	99,351
7	0,081	0,043	0,037	98,482	99,395
8	0,064	0,041	0,023	98,546	99,436
9	0,061	0,036	0,025	98,606	99,472
10	0,045	0,028	0,017	98,652	99,500
11-20	0,293	0,138	0,155	98,945	99,637
21-30	0,140	0,071	0,069	99,084	99,708
31-40	0,102	0,082	0,020	99,186	99,790
41-50	0,204	0,034	0,170	99,391	99,824
51-60	0,047	0,027	0,019	99,437	99,851
61-70	0,036	0,037	-0,001	99,473	99,888
71-80	0,026	0,004	0,022	99,499	99,892
81-90	0,026	0,010	0,016	99,525	99,902
91-100	0,020	0,000	0,020	99,544	99,902
101-200	0,174	0,001	0,172	99,718	99,903
201-300	0,056	0,057	0,000	99,774	99,960
301-400	0,043	0,040	0,003	99,817	100,000
401-500	0,020	0,000	0,020	99,837	100,000
501-600	0,035	0,000	0,035	99,873	100,000
601-700	0,021	0,000	0,021	99,894	100,000
701-800	0,026	0,000	0,026	99,920	100,000
801-900	0,007	0,000	0,007	99,927	100,000
901-1000	0,001	0,000	0,001	99,928	100,000
1001-2000	0,052	0,000	0,052	99,981	100,000
2000+	0,019	0,000	0,019	100,000	100,000

Tabela 16: Porównanie udziałów procentowych wiadomości z określoną liczbą wiadomości podobnych do nich dla obu zbiorów. Próg podobieństwa równy 0,7.

Liczba wiadomości	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0	74,249	68,849	5,400	74,249	68,849
1	10,622	9,656	0,965	84,871	78,506
2	4,329	4,162	0,166	89,200	82,668
3	2,358	2,426	-0,068	91,557	85,094
4	1,477	1,610	-0,132	93,035	86,704
5	1,017	1,216	-0,199	94,052	87,920
6	0,748	0,953	-0,205	94,800	88,874
7	0,571	0,769	-0,198	95,371	89,642
8	0,440	0,644	-0,205	95,810	90,287
9	0,358	0,551	-0,192	96,169	90,837
10	0,287	0,500	-0,213	96,455	91,337
11-20	1,395	2,793	-1,398	97,850	94,130
21-30	0,494	1,313	-0,819	98,345	95,443
31-40	0,265	0,875	-0,611	98,609	96,319
41-50	0,292	0,588	-0,297	98,901	96,907
51-60	0,112	0,425	-0,314	99,013	97,332
61-70	0,077	0,366	-0,289	99,090	97,698
71-80	0,061	0,263	-0,202	99,151	97,961
81-90	0,054	0,233	-0,180	99,204	98,195
91-100	0,032	0,180	-0,148	99,236	98,375
101-200	0,258	0,931	-0,672	99,494	99,305
201-300	0,088	0,326	-0,238	99,582	99,631
301-400	0,061	0,203	-0,142	99,643	99,835
401-500	0,040	0,062	-0,022	99,684	99,897
501-600	0,035	0,072	-0,037	99,718	99,969
601-700	0,036	0,014	0,022	99,754	99,983
701-800	0,026	0,008	0,018	99,780	99,991
801-900	0,036	0,005	0,031	99,817	99,996
901-1000	0,003	0,002	0,001	99,819	99,998
1001-2000	0,096	0,002	0,094	99,916	100,000
2000+	0,084	0,000	0,084	100,000	100,000

Gdy zliczamy podobne wiadomości z wysokim progiem (tabela 15) widzimy, że jest ich bardzo mało. Różnice pomiędzy zbiorami są bardzo małe. Widać jednak, że w zbiorze potencjalnie normalnych użytkowników jest więcej wiadomości z dużym wynikiem. Liczby podobnych wiadomości rzędu setek przy bardzo wysokim współczynniku podobieństwa są już bardzo podejrzane i zostaną zweryfikowane.

Najprawdopodobniej zbiór użytkowników normalnych zawiera również takich, którzy nie powinni się tutaj znajdować. Tacy użytkownicy najczęściej publikują wiele takich samych wiadomości, które są dedykowanymi odpowiedziami do innych użytkowników biorących udział w rozmowie - są to przykładowo podziękowania. Występują też użytkownicy publikujący takie same wiadomości polityczne ze zmianą wyłącznie ich fragmentu, w celu

obniżenia podobieństwa. Wyróżniający się użytkownicy zostaną szczegółowo przeanalizowani i odrzuceni w fazie tworzenia zbioru testowo-treningowego do uczenia maszynowego.

Wiele podobnych wiadomości jest również publikowane przez konta zajmujące się tematyką sondaży wyborczych oraz prognozą pogody. Wykorzystują one szablony, w których zmieniają się tylko poszczególne fragmenty całej wypowiedzi. Takie konta nie są jednak szkodliwe.

W przypadku zmniejszenia progu podobieństwa do 0,7 wyniki otrzymane w tabeli 16 znacznie się różnią. Teraz to użytkownicy niepożądani przeważają w większych przedziałach. Różnica udziałów procentowych jest znacznie większa niż w przypadku poprzedniego progu, który ustawiony był wysoko. Pokrycie zbioru do 10 podobnych wiadomości włącznie w przypadku niepożądanych użytkowników jest mniejsze o około 5%.

6.2. Analiza zachowań i statystyk użytkowników

6.2.1. Maksymalna liczba wiadomości użytkownika napisanych w określonym oknie czasowym

W całym zbiorze wiadomości użytkowników z tematów politycznych poszukiwana jest maksymalna liczba wiadomości, które badany użytkownik napisał w oknie czasowym. Dokonano analizy dla okien czasowych równych:

- 5 minut,
- 15 minut.

Retweet wiadomości liczony jest jak zwykła wiadomość. Jest to również forma aktywności użytkownika. Przed retweetem musi poświecić trochę czasu w celu przeczytania podawanej dalej wiadomości.

Okno czasowe równe 5 minut

Tabela 17: Maksymalna liczba wiadomości w oknie czasowym równym 5 minut.

Liczba wiadomości	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
1	83,420	9,809	73,611	83,420	9,809
2	9,995	11,631	-1,636	93,414	21,439
3	2,952	10,841	-7,889	96,366	32,281
4	1,330	9,809	-8,479	97,696	42,089
5	0,724	9,141	-8,417	98,419	51,230
6	0,445	8,017	-7,572	98,865	59,247
7	0,288	7,683	-7,395	99,152	66,930
8	0,200	5,891	-5,691	99,352	72,821
9	0,146	4,677	-4,531	99,498	77,498
10	0,108	3,219	-3,111	99,606	80,717
11	0,081	2,551	-2,470	99,687	83,268
12	0,062	2,247	-2,185	99,749	85,515
13	0,049	2,308	-2,259	99,798	87,823
14	0,035	1,822	-1,787	99,833	89,645
15	0,028	1,458	-1,429	99,861	91,102
16	0,023	1,427	-1,404	99,884	92,530
17	0,017	1,154	-1,137	99,901	93,684
18	0,015	1,002	-0,987	99,916	94,686
19	0,014	1,002	-0,988	99,930	95,688
20	0,009	0,698	-0,689	99,939	96,386
21-30	0,040	3,189	-3,149	99,979	99,575
31-40	0,010	0,243	-0,232	99,990	99,818
41-50	0,004	0,121	-0,117	99,994	99,939
51-60	0,003	0,030	-0,028	99,996	99,970
61-70	0,002	0,000	0,002	99,998	99,970
71-80	0,001	0,030	-0,029	99,999	100,000
81+	0,001	0,000	0,001	100,000	100,000

Okno czasowe równe 15 minut

Tabela 18: Maksymalna liczba wiadomości w oknie czasowym równym 15 minut.

Liczba wiadomości	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
1	80,732	8,047	72,685	80,732	8,047
2	10,644	8,321	2,323	91,376	16,368
3	3,488	8,412	-4,924	94,864	24,780
4	1,646	7,106	-5,460	96,510	31,886
5	0,949	6,924	-5,975	97,459	38,810
6	0,606	5,527	-4,921	98,065	44,336
7	0,412	5,405	-4,993	98,477	49,742
8	0,298	4,130	-3,832	98,775	53,872
9	0,225	3,826	-3,601	99,001	57,698
10	0,180	4,009	-3,829	99,181	61,707
11	0,141	3,553	-3,412	99,321	65,260
12	0,115	3,887	-3,772	99,436	69,147
13	0,092	4,069	-3,977	99,528	73,216
14	0,068	3,432	-3,364	99,596	76,647
15	0,060	3,462	-3,402	99,656	80,109
16	0,052	2,369	-2,317	99,708	82,478
17	0,039	2,642	-2,602	99,747	85,120
18	0,034	2,156	-2,122	99,781	87,276
19	0,028	1,549	-1,521	99,809	88,825
20	0,024	1,609	-1,585	99,833	90,434
21-30	0,111	6,620	-6,509	99,944	97,054
31-40	0,032	1,579	-1,547	99,976	98,633
41-50	0,010	0,698	-0,689	99,986	99,332
51-60	0,005	0,486	-0,481	99,991	99,818
61-70	0,002	0,091	-0,089	99,993	99,909
71-80	0,002	0,091	-0,089	99,996	100,000
81+	0,004	0,000	0,004	100,000	100,000

Rozkłady wartości poszczególnych liczb wiadomości są podobne dla obu okien czasowych. Logicznym jest, że 5 minutowe okno czasowe (tabela 17) zmniejsza osiągane przez użytkowników liczby wiadomości. Wersja z wydłużonym czasem (tabela 18) jest lepsza, ponieważ mniej użytkowników wpada w bardzo liczny pierwszy przedział. Badając dłuższy, lecz nie przesadnie długi okres czasu możemy lepiej skategoryzować użytkowników.

W obu przypadkach ponad 80% użytkowników napisało wyłącznie jedną wiadomość w oknie czasowym. Wynika to z tego, że bardzo dużo normalnych użytkowników wypowiadało się sporadycznie w tematach politycznych. Często pisze tylko kilka lub kilkanaście wiadomości w dużych odstępach czasu sięgających nawet kilku dni. Bardzo możliwe, że piszą po jednej wiadomości podczas przerw w pracy lub jazdy autobusem. W przypadku użytkowników niepożądanych sprawa wygląda inaczej. W ich aktywności można wychwycić zachowania sugerujące ich mocniejsze zaangażowanie. Może być to tylko jednorazowa sytuacja, ponieważ uwzględniana jest tylko maksymalna liczba, lecz w sytuacji tak dużych dysproporcji rozkładów nie może być to przypadkiem. Należało by podać średnią pięciu lub dziesięciu największych liczb wiadomości w oknie czasowym. Zostanie to sprawdzone

w przypadku bardziej zaawansowanej analizy podobnego typu: 6.2.2.

6.2.2. Poszukiwanie najdłuższych serii wiadomości użytkownika

Wyszukiwana jest najdłuższa seria wiadomości w tematach politycznych, gdzie między kolejnymi wiadomościami w serii, różnica czasu ich publikacji nie przekracza określonego maksimum. Wyszukane zostały serie dla różnic czasu równych kolejno: 5 minut, 15 minut oraz 30 minut. W każdej tabeli uwzględnione są udziały procentowe przedziałów długości serii dla obu badanych zbiorów.

Wyniki analizy serii z uwzględnieniem wszystkich wiadomości

Uwzględniane są zarówno zwykłe wiadomości autorstwa użytkowników jak i opublikowane przez nich retweety.

Tabela 19: Udziały procentowe serii wiadomości o różnej długości, w których różnice czasowe pomiędzy kolejnymi wiadomościami są nie większe niż 15 minut

Liczba wiadomości	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
1	80,732	8,047	72,685	80,732	8,047
2	10,227	7,349	2,878	90,959	15,396
3	3,345	7,106	-3,761	94,304	22,502
4	1,619	5,436	-3,817	95,923	27,938
5	0,935	4,889	-3,954	96,858	32,827
6	0,613	4,221	-3,608	97,471	37,048
7	0,429	3,978	-3,549	97,900	41,026
8	0,326	3,158	-2,832	98,226	44,185
9	0,249	2,612	-2,363	98,474	46,796
10	0,205	2,612	-2,407	98,679	49,408
11-20	0,861	21,045	-20,184	99,540	70,452
21-30	0,237	13,878	-13,641	99,777	84,330
31-40	0,100	7,076	-6,976	99,877	91,406
41-50	0,048	2,369	-2,321	99,925	93,775
51-60	0,026	1,640	-1,614	99,951	95,415
61-70	0,014	1,093	-1,079	99,965	96,508
71-80	0,010	1,306	-1,295	99,975	97,814
81-90	0,007	0,729	-0,722	99,982	98,542
91-100	0,004	0,456	-0,451	99,987	98,998
101-200	0,011	0,820	-0,809	99,998	99,818
201-300	0,002	0,091	-0,090	100,000	99,909
301+	0,000	0,091	-0,091	100,000	100,000

Widać bardzo duży udział normalnych użytkowników z małymi seriami. Normalni użytkownicy piszą o wiele mniej, co zostało już zauważone w prostszej analizie 6.2.1. Gdy użytkownik pisze okazjonalnie kilka wiadomości w ciągu dnia, ciężko jest aby miał jakąś długą serię. Całkowicie odmienna sytuacja jest w przypadku niepożądanych użytkowników, którzy mogą często zawodowo zajmować się trollingiem lub po prostu być botami do spamu. W ich przypadku małe serie występują o wiele rzadziej. Około 50%

z nich ma w swojej historii wiadomości serię o rozmiarze co najmniej równym 11.

Porównanie z wynikami prostszej analizy z punktu 6.2.1

Wyniki z tabeli 19 można porównać z tabelą 18 przedstawiającą wynik prostszej analizy dla okna z takim samym czasem (15 minut), również wliczającej retweety. Widać, że dla mniejszych wartości udziały procentowe grup o tych samych liczbach wiadomości są podobne, a różnice stopniowo się zwiększają. Szczególnie widoczne to jest w przypadku zbioru niepożądanych użytkowników. W przypadku wyników tylko do 10 wiadomości pokrycie zbioru różni się o około 12.3%.

Wielu użytkowników, w prostszej analizie mogło zostać zakwalifikowanych do grup, które nie zwracają na nich żadnych podejrzeń. Są to użytkownicy, którzy piszą bardzo dużo i regularnie, często zawodowo trollują lub są botami publikującymi w regularnych odstępach czasu. Przykładowo prosta analiza dla użytkownika piszącego wiadomość regularnie co około 10 minut przez 5 godzin przydzieli tą samą kategorię co dla użytkownika, który napisał 2 wiadomości w ciągu 10 minut.

Zestawienie wyników analizy serii z różnymi dopuszczalnymi czasami między kolejnymi wiadomościami

Wszystkie wyniki w tabeli 20 uwzględniają retweety. Zbadano przypadki dla maksymalnych różnic czasowych między wiadomościami równych: 5 minut, 15 minut oraz 30 minut. Tabela uwzględnia serie o rozmiarze do 30 włącznie.

Tabela 20: Porównanie udziałów procentowych serii wiadomości dla różnych maksymalnych różnic czasowych między kolejnymi wiadomościami

	Normalni, %			Niepożądani, %		
Wielkość serii	5min.	15min.	30min.	5min.	15min.	30min.
1	83,42	80,73	79,41	9,81	8,05	7,23
2	9,35	10,23	10,60	10,23	7,35	6,32
3	2,85	3,35	3,57	7,96	7,11	6,26
4	1,33	1,62	1,75	7,96	5,44	4,83
5	0,77	0,94	1,03	5,89	4,89	4,74
6	0,49	0,61	0,68	4,59	4,22	4,43
7	0,34	0,43	0,48	4,68	3,98	3,40
8	0,25	0,33	0,36	3,10	3,16	3,04
9	0,19	0,25	0,28	3,16	2,61	2,43
10	0,15	0,20	0,24	3,13	2,61	2,28
11	0,12	0,16	0,19	2,43	2,49	2,19
12	0,10	0,14	0,16	2,46	2,25	2,25
13	0,08	0,11	0,13	2,64	2,43	2,28
14	0,07	0,09	0,11	2,22	2,34	2,55
15	0,06	0,08	0,09	1,97	1,91	2,03
16	0,05	0,07	0,08	1,97	2,28	2,46
17	0,04	0,06	0,07	2,82	2,03	2,03
18	0,03	0,05	0,06	2,16	1,94	2,00
19	0,03	0,05	0,06	1,43	1,64	1,82
20	0,03	0,04	0,05	1,55	1,73	1,82
21	0,02	0,04	0,05	1,40	1,40	1,61
22	0,02	0,03	0,04	1,37	1,67	1,88
23	0,02	0,03	0,03	1,18	1,76	1,70
24	0,02	0,03	0,03	1,21	1,70	1,82
25	0,01	0,02	0,03	0,82	1,06	1,21
26	0,01	0,02	0,03	0,76	1,09	1,28
27	0,01	0,02	0,02	0,85	1,12	1,15
28	0,01	0,02	0,02	0,79	1,37	1,40
29	0,01	0,02	0,02	0,97	1,40	1,18
30	0,01	0,01	0,02	0,88	1,31	1,37

Wyniki z tabeli 20 nie są zaskakujące. Wraz ze zwiększaniem dopuszczalnych różnic czasowych między kolejnymi wiadomościami w serii, coraz więcej użytkowników jest zaliczanych do serii o większych rozmiarach. Możemy zaobserwować różnice w rozkładach wartości. Pokrycie zbioru dla serii do 10 wiadomości włącznie jest następujące:

- dla 5 min 99,148% użytkowników normalnych i 60,492% użytkowników niepożądanych,

- dla 15 min 98,679% użytkowników normalnych i 49,408% użytkowników niepożądanych,

- dla 30 min 98,385% użytkowników normalnych i 44,944% użytkowników niepożądanych.

Serie z różnicą 15 minut nie uwzględniające retweetowanych wiadomości.

Wiadomości, które są retweetami nie są brane pod uwagę w procesie poszukiwania najdłuższej serii. W związku z tym liczby użytkowników poddawanych analizie zmniejszyły się następująco:

- liczba użytkowników ze zbioru IRA zmniejszyła się z 3293 do 3127,
- liczba użytkowników ze zbioru stworzonego na bazie Harvard Dataverse zmniejszyła się z 1859531 do 952189.

Przy okazji widoczna jest tutaj, spora liczba użytkowników z drugiego zbioru, którzy wyłącznie retweetowali wiadomości.

Tabela 21: Udziały procentowe serii o danych długościach zawierających wiadomości bez uwzględniania retweetów z oknem czasowym pomiędzy kolejnymi wiadomościami nie większym niż 15 minut

Liczba wiadomości	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
1	82,263	9,626	72,637	82,263	9,626
2	10,225	7,291	2,933	92,488	16,917
3	3,236	6,268	-3,032	95,724	23,185
4	1,535	4,989	-3,454	97,259	28,174
5	0,805	4,701	-3,896	98,064	32,875
6	0,502	4,189	-3,687	98,566	37,064
7	0,329	4,029	-3,700	98,895	41,094
8	0,227	2,942	-2,715	99,123	44,036
9	0,163	2,654	-2,491	99,286	46,690
10	0,136	2,622	-2,486	99,422	49,312
11-20	0,427	21,906	-21,479	99,849	71,218
21-30	0,082	14,359	-14,277	99,931	85,577
31-40	0,031	6,940	-6,908	99,962	92,517
41-50	0,012	2,047	-2,034	99,975	94,563
51-60	0,007	1,407	-1,400	99,982	95,971
61-70	0,003	0,768	-0,765	99,984	96,738
71-80	0,004	1,311	-1,307	99,989	98,049
81-90	0,003	0,640	-0,637	99,991	98,689
91-100	0,002	0,416	-0,414	99,993	99,105
101-200	0,005	0,704	-0,699	99,998	99,808
201-300	0,001	0,096	-0,095	99,999	99,904
301-400	0,000	0,032	-0,032	99,999	99,936
401-500	0,000	0,064	-0,064	99,999	100,000
501+	0,001	0,000	0,001	100,000	100,000

Rozkłady wartości w tabeli 21 są bardzo podobne do tych uzyskanych w wynikach uwzględniających retweety (tabela 19). Nawet bez ich uwzględniania serie są bardzo długie. Wyniki dla większych serii zmieniły się bardziej w przypadku użytkowników normalnych na co miało wpływ to, że aż około 64% ich wiadomości było retweetami. Dla

porównania niepożądani użytkownicy o wiele częściej piszą własne wiadomości, a udział retweetów w ich przypadku jest równy około 36%. Nieuwzględnianie retweetów w przypadku niepożądanych użytkowników nie przynosi dużych różnic. Pokrycie zbioru niepożądanych użytkowników przez serie o rozmiarze do 10 włącznie praktycznie się nie zmieniło, a użytkowników normalnych nieznacznie wzrosło. Szczegółowe różnice wyników z i bez uwzględniania retweetów widoczne są tabeli 22.

Tabela 22: Porównanie udziałów procentowych rozmiarów serii wiadomości uwzględniających i nie uzgledniających retweetów dla okna czasowego wynoszącego 15 minut.

	Normalni, %		Niepożądani, %	
Liczba wiadomości	z retweetami	Bez retweetów	z retweetami	Bez retweetów
1	80,73	82,26	8,05	9,63
2	10,23	10,23	7,35	7,29
3	3,35	3,24	7,11	6,27
4	1,62	1,54	5,44	4,99
5	0,94	0,81	4,89	4,70
6	0,61	0,50	4,22	4,19
7	0,43	0,33	3,98	4,03
8	0,33	0,23	3,16	2,94
9	0,25	0,16	2,61	2,65
10	0,21	0,14	2,61	2,62
11-20	0,86	0,43	21,05	21,91
21-30	0,24	0,08	13,88	14,36
31-40	0,10	0,03	7,08	6,94
41-50	0,05	0,01	2,37	2,05
51-60	0,03	0,01	1,64	1,41
61-70	0,01	0,00	1,09	0,77
71-80	0,01	0,00	1,31	1,31
81-90	0,01	0,00	0,73	0,64
91-100	0,00	0,00	0,46	0,42
101-200	0,01	0,01	0,82	0,70
201-300	0,00	0,00	0,09	0,10
301+	0,00	0,00	0,09	0,10

Średnia długość najdłuższych serii wiadomości użytkownika

Aby wyeliminować wpływ jednorazowych anomalii w aktywności użytkowników brane są pod uwagę średnie pięciu największych rozmiarów serii wiadomości wliczających retweety. Pokazuje to, kto pisze regularnie dużo wiadomości, a nie tylko jednorazowo. Liczba pięciu serii została wybrana ze względu na dużą liczbę normalnych użytkowników z małą liczbą wiadomości.

Tabela 23: Tabela przedstawia porównanie udziałów procentowych poszczególnych przedziałów wartości średniej długości pięciu najdłuższych serii użytkownika dla normalnych i niepożądanych użytkowników. Maksymalny czas pomiędzy kolejnymi wiadomościami w serii równy jest 15 minut.

Liczba wiadomości	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
1	80,732	8,047	72,685	80,732	8,047
2	12,968	10,295	2,673	93,700	18,342
3	2,698	10,446	-7,748	96,398	28,788
4	1,154	9,171	-8,017	97,552	37,959
5	0,627	7,227	-6,600	98,179	45,187
6	0,400	5,102	-4,702	98,579	50,288
7	0,280	4,798	-4,518	98,859	55,087
8	0,202	4,950	-4,748	99,061	60,036
9	0,147	4,343	-4,196	99,208	64,379
10	0,117	3,492	-3,375	99,325	67,871
11-20	0,476	20,832	-20,356	99,801	88,703
21-30	0,115	7,288	-7,173	99,915	95,991
31-40	0,043	2,186	-2,144	99,958	98,178
41-50	0,018	1,063	-1,045	99,976	99,241
51-60	0,009	0,304	-0,294	99,985	99,544
61-70	0,006	0,091	-0,085	99,991	99,636
71-80	0,003	0,121	-0,119	99,994	99,757
81-90	0,002	0,061	-0,059	99,996	99,818
91-100	0,001	0,030	-0,029	99,997	99,848
101-200	0,003	0,091	-0,089	99,999	99,939
201-300	0,000	0,061	-0,061	100,000	100,000

W tabeli 23 przedstawiono wyniki analizy średnich pięciu najdłuższych serii z maksymalnym czasem pomiędzy wiadomościami równym 15 minut. Największe zmiany widoczna jest w przypadku niepożądanych użytkowników. Wiele z ich serii było jednorazowe, a uwzględnienie średniej największych znacząco obniżyło wyniki. Udział serii o średniej do 10 wiadomości włącznie wynosi 67,8%, gdy dla pojedynczej wiadomości porównywalny udział wynosił 49,4% (tabela 19). Szczegółowe zestawienie udziałów procentowych przed i po uwzględnianiu średnich można zobaczyć w tabeli 24. W przypadku użytkowników normalnych nie ma tak wyraźnych zmian. Wynika to z tego, że przeciętny zwykły użytkownik praktycznie nie pisze wiadomości seriami.

Tabela 24: Porównanie wyników osiąganych przez obliczanie średniej pięciu największych serii z wynikami uwzględniającymi tylko największą serię. Okno czasowe równe 15 minut.

	Normalni, %		Niepożądani, %	
Liczba wiadomości	Tylko maksymalna	Średnia 5 największych	Tylko maksymalna	Średnia 5 największych
1	80,732	80,732	8,047	8,047
2	10,227	12,968	7,349	10,295
3	3,345	2,698	7,106	10,446
4	1,619	1,154	5,436	9,171
5	0,935	0,627	4,889	7,227
6	0,613	0,400	4,221	5,102
7	0,429	0,280	3,978	4,798
8	0,326	0,202	3,158	4,950
9	0,249	0,147	2,612	4,343
10	0,205	0,117	2,612	3,492
11-20	0,861	0,476	21,045	20,832
21-30	0,237	0,115	13,878	7,288
31-40	0,100	0,043	7,076	2,186
41-50	0,048	0,018	2,369	1,063
51-60	0,026	0,009	1,640	0,304
61-70	0,014	0,006	1,093	0,091
71-80	0,010	0,003	1,306	0,121
81-90	0,007	0,002	0,729	0,061
91-100	0,004	0,001	0,456	0,030
101-200	0,011	0,003	0,820	0,091
201-300	0,002	0,000	0,091	0,061
301+	0,000	0,000	0,091	0,000

6.2.3. Źródło publikacji wiadomości

Źródła publikacji tweetów możemy podzielić na dwie grupy:

- oficjalne aplikacje dostarczone przez Twittera, takie jak aplikacje mobilne na systemy iOS czy Android lub przeglądarkowa strona internetowa portalu,
- aplikacje stworzone przez inne podmioty wykorzystujące udostępniane przez Twittera API do wysyłania i odczytywania tweetów. W tej grupie mogą zawierać się programy stworzone do automatyzacji publikacji dużych ilości tych samych tweetów (spamu). Każdy może sam stworzyć aplikację tego typu.

Niepożądani użytkownicy

Tabela 25: Udział najpopularniejszych źródeł publikacji tweetów napisanych we wszystkich językach ze zbioru niepożądanych użytkowników.

Źródło	Liczba wystąpień	Udział, %
Twitter Web Client	523951	28,70
twitterfeed	310394	17,00
TweetDeck	121286	6,64
newtwittersky	85425	4,68
bronislav	67047	3,67
iziaslav	66356	3,63
rostislav	63071	3,45
generation	59573	3,26
Twibble.io	54934	3,00
Ohwee Messanger	50981	2,79
NovaPress Publisher	46808	2,56
tłum. Aplikacja dla ciebie	34052	1,87
vavilonX	30706	1,68
Twitter for Android	28208	1,55
LiveJournal.com	25851	1,42
slovoslav	25156	1,38
meceslav	24201	1,33
dlvr.it	21409	1,17
token_app	18312	1,00
IFTTT	17739	0,97
Twitter for iPhone	14121	0,77

Tabela 26: Udział najpopularniejszych źródeł publikacji tweetów napisanych wyłącznie w języku angielskim ze zbioru niepożądanych użytkowników.

Źródło	Liczba wystąpień	Udział, %
Twitter Web Client	300719	50,47
twitterfeed	135014	22,66
TweetDeck	42882	7,20
Twibble.io	38844	6,52
vavilonX	29828	5,00
Twitter for Android	17321	2,91
newtwittersky	9909	1,66
Mobile Web (M2)	3243	0,54
IFTTT	2751	0,46
Jerusalem	1303	0,22
masss post4	1271	0,21
POTUSADJT Bot	1241	0,21
Tweefilter	1096	0,18
Twitter for Android Tablets	794	0,13
masss post5	682	0,11
Uptwitter	616	0,10
erased8269673	504	0,09
erased8286905	466	0,08
Crowdfire - Go Big	440	0,07
erased8287492	387	0,07
T-Helper	375	0,06

Zostały przeanalizowane zarówno wiadomości napisane dowolnym językiem (tabela 25) jak i wyłącznie po angielsku (tabela 26). Podobnie jak w przypadku linków są widoczne zauważalne różnice. Większość źródeł na listach piętnastu najpopularniejszych pozostała taka sama, lecz doszło do przetasować kolejności i wyraźnych zmian w udziałach procentowych w całym zbiorze. Wyróżnia się tutaj ogromny skok procentowy udziału “Twitter Web Client”, czyli przeglądarkowego klienta Twittera w wiadomościach pisanych tylko po angielsku. To źródło publikacji jest najwygodniejsze dla zawodowego trolla.

Zwykli użytkownicy

Tabela 27: Udziały najpopularniejszych źródeł publikacji tweetów ze zbioru normalnych użytkowników.

Źródło	Liczba wystąpień	Udział, %
Twitter for iPhone	3677463	33,52
Twitter for Android	2496795	22,76
Twitter Web Client	2348897	21,41
Twitter for iPad	818258	7,46
Twitter Web App	216208	1,97
twitterfeed	153223	1,40
IFTTT	148437	1,35
TweetDeck	144515	1,32
Facebook	92984	0,85
dlvr.it	82680	0,75
Linkis: turn sharing into growth	75406	0,69
Mobile Web	54481	0,50
Tweetbot for iOS	44290	0,40
Twitter for Windows	43869	0,40
Hootsuite	42498	0,39
Mobile Web (M2)	37720	0,34
Twitter for Windows Phone	35070	0,32
Google	31219	0,29
TweetCaster for Android	23378	0,21
WordPress.com	22298	0,20
Twitter for BlackBerry	19139	0,17

Analizując tabelę 27 widzimy, że użytkownicy normalni najczęściej korzystają z oficjalnych źródeł mobilnych. Duży udział ma również popularny klient przeglądarkowy. Nie tylko niepożądani użytkownicy lubią z niego korzystać.

Porównanie wyników

Do porównania zostały wybrane wyłącznie wiadomości napisane w języku angielskim. Uwzględniono 10 najpopularniejszych źródeł z każdego zbioru. W przypadku pokrywania się pozycji nie były dobierane kolejne, dlatego liczba różnych źródeł w tabeli jest równa 15.

Tabela 28: Porównanie udziałów źródeł publikacji tweetów w języku angielskim z obu zbiorów.

Źródło	Normalni, %	Niepożądani, %	Różnica, %
Twitter for iPhone	33,52	0,03	33,48
Twitter for Android	22,76	2,91	19,85
Twitter Web Client	21,41	50,47	-29,06
Twitter for iPad	7,46	0,04	7,41
Twitter Web App	1,97	0,00	1,97
twitterfeed	1,40	22,66	-21,26
IFTTT	1,35	0,46	0,89
TweetDeck	1,32	7,20	-5,88
Facebook	0,85	0,00	0,85
dlvr.it	0,75	0,00	0,75
Twibble.io	0,04	6,52	-6,48
vavilonX	0,00	5,01	-5,01
newtwittersky	0,00	1,66	-1,66
Mobile Web (M2)	0,34	0,54	-0,20
Jerusalem	0,00	0,22	-0,22

Wyniki umieszczone w Tabeli 28 pokazują, że niepożądani użytkownicy używają do tweetowania głównie klienta przeglądarkowego. Suma udziałów publikacji ze źródeł mobilnych w ich przypadku nie przekracza 10%. Ponadto spory udział mają programy będące źródłami nieoficjalnymi. Wiele takich źródeł jest blokowanych przez Twittera, ponieważ mogą one być narzędziami do automatyzacji spamu.

Wnioski są całkowicie inne w przypadku użytkowników normalnych. Większość z nich korzysta z urządzeń mobilnych, choć popularny u złych użytkowników dostęp przez przeglądarkę również ma duży udział. Zwykli użytkownicy korzystają praktycznie wyłącznie z aplikacji pochodzących z oficjalnych źródeł.

Można zatem podsumować:

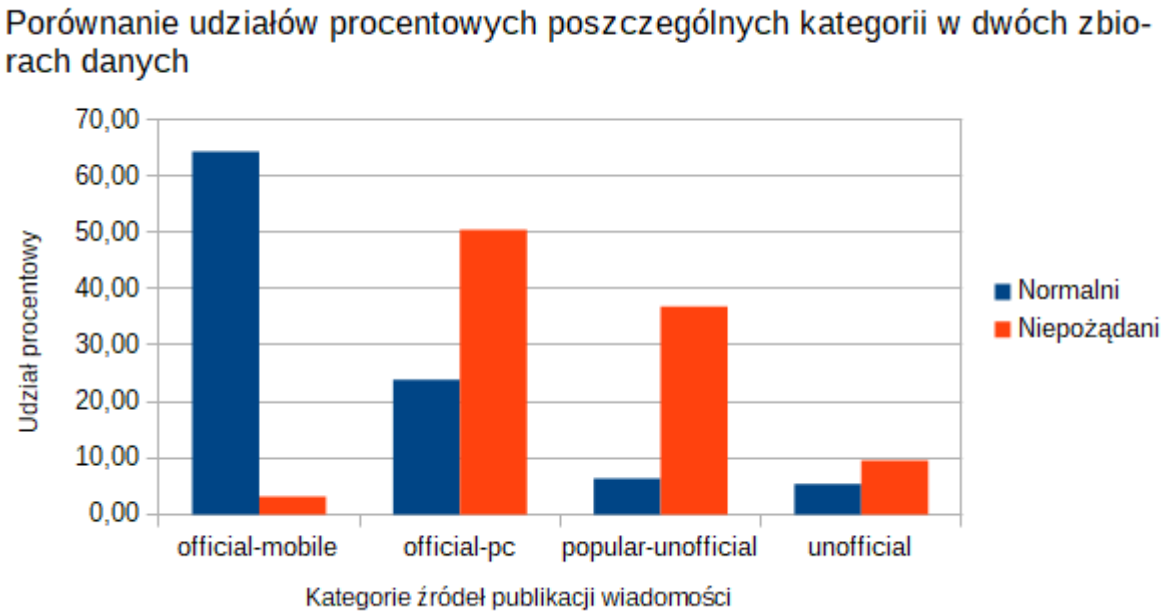
- typowy niepożądany użytkownik używa przeglądarki internetowej do publikacji swoich wiadomości;
- aplikacje do tweetowania z nieoficjalnych źródeł są praktycznie wykorzystywane wyłącznie przez niepożądanych użytkowników - mogą być to przykładowo programy do automatyzacji spamu;

- zwykły użytkownik najczęsciej wykorzystuje urządzenia mobilne do publikacji swoich wiadomości.

Podział na kategorie

Na podstawie wcześniejszych analiz wszystkie źródła publikacji z obu zbiorów zostały skategoryzowane do jednej z czterech kategorii, które najlepiej obrazują ich podział w badanych zbiorach. Do kategorii źródeł publikacji wliczają się:

- official-mobile - oficjalne aplikacje pochodzące od Twittera przeznaczone na urządzenia mobilne;
- official-pc - oficjalne aplikacje pochodzące od Twittera przeznaczone do użytku na komputerach;
- popular-unofficial - popularne aplikacje z nieoficjalnych źródeł, które w przynajmniej jedynym zbiorze miały udział co najmniej 0,5%. Należą do nich: twitterfeed, IFTTT, TweetDeck, Facebook, dlvr.it, Linkis: turn sharing into growth, Twibble.io;
- unofficial - aplikacje niezakwalifikowane do poprzednich grup.



Rysunek 23: Diagram zestawia udziały procentowe poszczególnych kategorii publikacji tweetów w badanych zbiorach. Widoczne jest, że zwykli użytkownicy najczęściej publikują wiadomości z urządzeń mobilnych, których praktycznie nie używają użytkownicy o złych zamiarach.

Na histogramie 23 Widoczna jest ogromna przewaga normalnych użytkowników w kwestii używania urządzeń mobilnych. Użytkownicy niepożądani mają wyraźną, lecz nie tak dużą przewagę w publikacjach z komputerów. Nie jest to zaskakujące. Na fizycznej klawiaturze o wiele łatwiej jest pisać dużo wiadomości niż na ekranie smartfona. Programy

do automatyzacji spamu, które mogą znaleźć się w kategorii popular-unofficial i unofficial, również są tworzone z przeznaczeniem na komputery, a nie urządzenia mobilne. Szczegółowe dane przedstawione na histogramie dostępne są w tabeli 29.

Ciekawy jest aspekt stworzenia nakładki do automatyzacji spamu na istniejącą aplikację, nawet oficjalną. Możliwe by to było nawet z urządzeniem mobilnym, co pozwoliłoby na całkowite odwrócenie uwagi w kontekście analizy źródła publikacji. W badanym zbiorze taka sytuacja nie występuje.

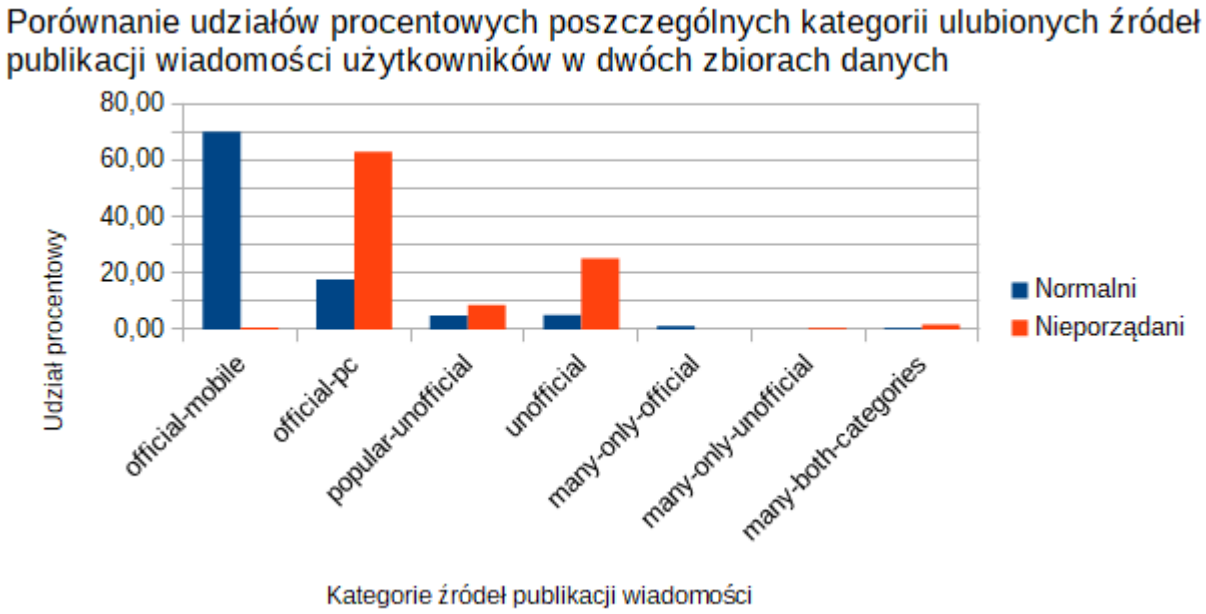
Tabela 29: Tabela obrazuje szczegółowe udziały procentowe poszczególnych kategorii zestawionych na diagramie 23.

Kategoria źródła	Normalni, %	Niepożądani, %	Różnica, %
official-mobile	64,38	3,12	61,27
official-pc	23,88	50,47	-26,59
popular-unofficial	6,39	36,86	-30,46
unofficial	5,35	9,55	-4,21

6.2.4. Ulubiona kategoria źródła publikacji wiadomości użytkownika

Na wzór analizy przedstawionej w punkcie 6.2.3, która badała to z źródła klienta została opublikowana konkretna wiadomość, analizowane są teraz ulubione źródła publikacji użytkowników. Badane są wyłącznie tweety użytkowników w języku angielskim - analizowanie wszystkich języków ma znaczny wpływ na rozkład wartości kategorii dla poszczególnych wiadomości co zostało już pokazane. Jako wynik otrzymujemy kategorię ulubionego źródła. Kategorie są rozszerzoną wersją 6.2.3, do której zaliczały się grupy: official-mobile, official-pc, popular-unofficial, unofficial. Zostały one już opisane. Wzbogacono je o grupy, które uwzględniają możliwość publikacji takiej samej liczby wiadomości z różnych kategorii:

- many-only-official - użytkownik opublikował równą liczbę wiadomości z kategorii official-mobile oraz official-pc,
- many-only-unofficial - użytkownik opublikował równą liczbę wiadomości z kategorii popular-unofficial oraz unofficial,
- many-both-categories - użytkownik opublikował równą liczbę wiadomości pochodzących ze źródeł oficjalnych i nieoficjalnych.



Rysunek 24: Histogram porównuje udziały procentowe kategorii ulubionych źródeł publikacji wiadomości dla użytkowników normalnych i niepożądanych. Widoczne są duże różnice w grupach oficjalnych źródeł.

Ogólny trend wyników w tabeli 24 jest podobny do uzyskanych w 6.2.3, gdzie analizowane były źródła pojedynczych wiadomości. Badając ulubione źródła dla użytkowników nie uwzględniane są ich kategorie drugiego wyboru lub okazjonalne publikacje z mniej znanych źródeł. Przez to różnice w poszczególnych kategoriach się pogłębiły. Jeszcze bardziej widoczna jest przewaga zwykłych użytkowników w kwestii urządzeń mobilnych, a niepożądanych w publikacjach z komputerów. Widzimy również, że wyklarowała się grupa użytkowników, która w większości korzysta z nieoficjalnych i mało popularnych źródeł. W przypadku poprzedniej analizy pojedynczych wiadomości grupy popular-unofficial i unofficial miały całkowicie inne udziały. Wynika to z tego, że użytkowników z pierwszej grupy jest mniej, lecz piszą o wiele więcej wiadomości. Szczegółowe wyniki przedstawione na histogramie 24 znajdują się w tabeli 30.

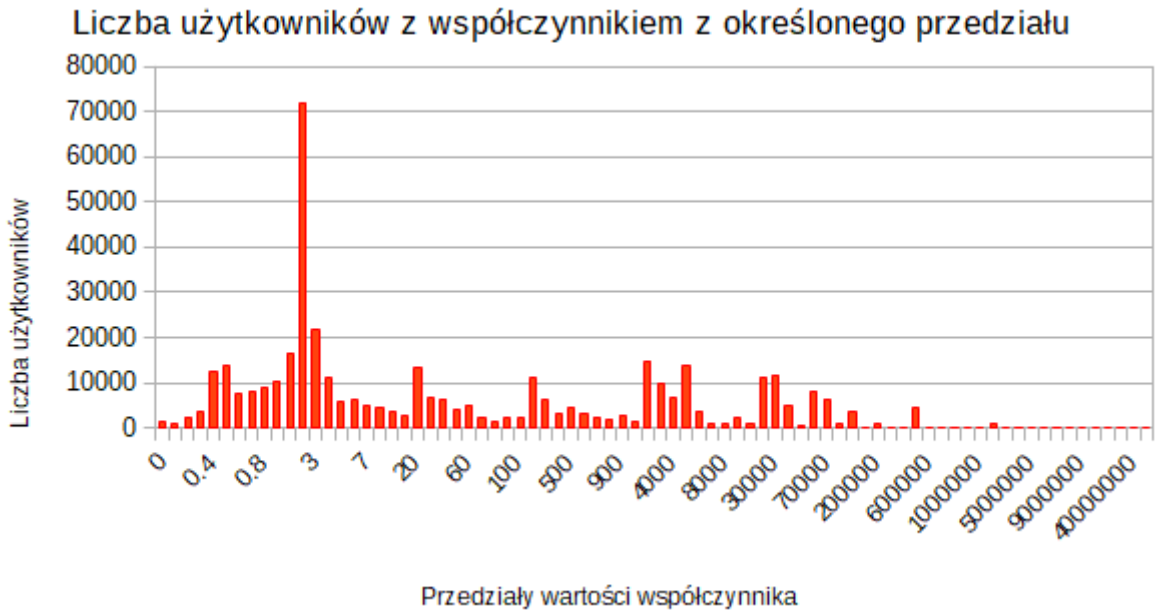
Tabela 30: Tabela obrazuje szczegółowe udziały procentowe poszczególnych kategorii ulubionych źródeł publikacji użytkowników zestawionych na diagramie 24.

Kategoria źródła	Normalni, %	Niepożądani, %	Różnica, %
official-mobile	69,94	0,78	69,15
official-pc	17,83	62,93	-45,10
popular-unofficial	5,06	8,64	-3,58
unofficial	5,20	25,19	-19,98
many-only-official	1,28	0,04	1,24
many-only-unofficial	0,06	0,59	-0,53
many-both-categories	0,63	1,84	-1,21

6.2.5. Stosunek liczby obserwujących i obserwowanych retweetowanych użytkowników

Stosunek wartości został wyliczony w taki sam sposób jak w przypadku zwykłych użytkowników 6.3.5. Wyliczoną wartość możemy traktować jako wiarygodność retweetowanego użytkownika. W większości przypadków spodziewamy się wysokiej wartości stosunku.

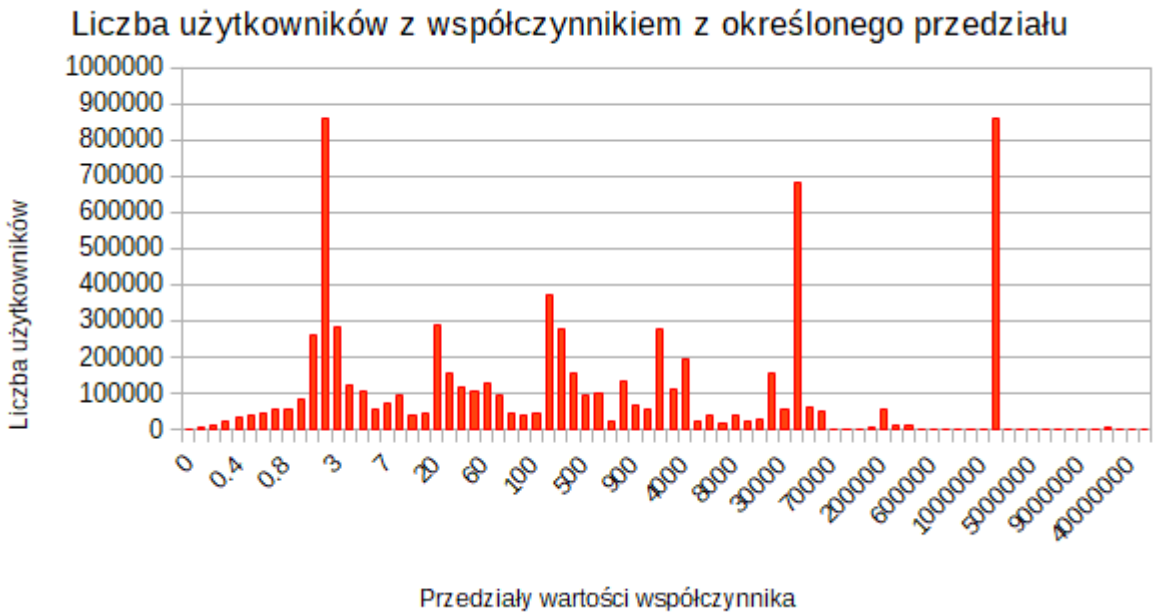
Niepożądani użytkownicy



Rysunek 25: Liczba retweetów wiadomości od użytkowników ze stosunkiem obserwujących i obserwowanych z określonego przedziału dla zbioru niepożądanych użytkowników. Widoczny jest wyraźny skok wartości w przedziale od 1,0 do 2,0 włącznie.

Na wykresie 25 widzimy jeden wyraźny skok wartości udziału procentowego dla przedziału stosunku od 1,0 do 2,0 włącznie. W takim przedziale możemy spodziewać się w szczególności zwykłych użytkowników, lecz występują czasami konta małych organizacji lub portali oraz mało popularnych osób publicznych takich jak pisarze lub publicyści.

Zwykli użytkownicy



Rysunek 26: Liczba retweetów wiadomości od użytkowników ze stosunkiem obserwujących i obserwowanych z określonego przedziału dla zbioru normalnych użytkowników. Widzimy trzy przedziały o wyraźnie większych wartościach. Użytkownicy normalni retweetują o wiele więcej bardzo popularnych osób.

Na wykresie 26 wyróżniają się trzy przedziały stosunku obserwowanych i obserwujących retweetowanych użytkowników. Pierwszy z nich to przedział (1, 2], który miał również bardzo duży udział w zbiorze niepożądanych użytkowników. Tak jak już stwierdzono są to normalni użytkownicy lub mało popularne organizacje lub osoby publiczne. Drugi przedział (30000, 40000] zawiera dużo retweetów ze względu na obecność w nim Hillary Clinton kandydującej w wyborach. Do retweetowanych użytkowników z tego przedziału należą również różne konta tematyczne lub lokalne popularnych portali informacyjnych czy znane osobistości, których przykładem może być aktor Trevor Noah. Należy jednak podkreślić, że głównie retweetowana jest tutaj kandydatka na prezydenta. Trzeci przedział (1000000, 2000000] współczynnika retweetowanych użytkowników zlicza podania dalej wiadomości prezydenta Donalda Trumpa, który w znacznej mierze odpowiada za popularność tego przedziału.

Porównanie wyników

Tabela 31: Udziały procentowe retweetów wiadomości od użytkowników o stosunku obserwujących i obserwowanych z określonego przedziału dla obu zbiorów użytkowników.

Przedział	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0	0,01	0,33	-0,33	0,01	0,33
0,0-0,1	0,07	0,19	-0,12	0,07	0,52
0,1-0,2	0,19	0,60	-0,41	0,26	1,12
0,2-0,3	0,33	0,86	-0,53	0,59	1,98
0,3-0,4	0,45	3,07	-2,62	1,04	5,05
0,4-0,5	0,55	3,40	-2,85	1,59	8,45
0,5-0,6	0,64	1,85	-1,21	2,23	10,30
0,6-0,7	0,77	2,00	-1,23	3,00	12,30
0,7-0,8	0,76	2,18	-1,43	3,76	14,49
0,8-0,9	1,16	2,56	-1,40	4,91	17,04
0,9-1,0	3,55	4,11	-0,56	8,46	21,15
1-2	11,74	17,78	-6,04	20,20	38,93
2-3	3,87	5,37	-1,50	24,07	44,30
3-4	1,70	2,71	-1,01	25,77	47,01
4-5	1,43	1,39	0,04	27,20	48,40
5-6	0,76	1,49	-0,74	27,96	49,90
6-7	0,97	1,21	-0,23	28,93	51,10
7-8	1,29	1,14	0,15	30,22	52,24
8-9	0,53	0,87	-0,34	30,75	53,12
9-10	0,59	0,64	-0,05	31,35	53,76
10-20	3,93	3,32	0,62	35,28	57,08
20-30	2,09	1,61	0,48	37,37	58,69
30-40	1,62	1,60	0,02	38,99	60,29
40-50	1,48	0,99	0,50	40,47	61,27
50-60	1,76	1,25	0,51	42,23	62,52
60-70	1,27	0,55	0,72	43,50	63,08
70-80	0,65	0,30	0,34	44,15	63,38
80-90	0,55	0,53	0,01	44,69	63,91
90-100	0,60	0,51	0,09	45,30	64,42
100-200	5,09	2,73	2,35	50,39	67,16
200-300	3,83	1,52	2,31	54,21	68,68
300-400	2,12	0,77	1,35	56,34	69,45
400-500	1,27	1,12	0,16	57,61	70,56
500-600	1,36	0,76	0,60	58,97	71,32
600-700	0,33	0,59	-0,26	59,30	71,91
700-800	1,83	0,49	1,34	61,13	72,40
800-900	0,94	0,72	0,22	62,07	73,12
900-1000	0,75	0,36	0,39	62,82	73,49
1000-2000	3,77	3,67	0,10	66,59	77,16
2000-3000	1,53	2,43	-0,89	68,12	79,58
3000-4000	2,69	1,61	1,09	70,81	81,19
4000-5000	0,32	3,42	-3,10	71,13	84,60
5000-6000	0,52	0,91	-0,39	71,65	85,51
6000-7000	0,23	0,23	0,00	71,88	85,74
7000-8000	0,56	0,18	0,38	72,44	85,92
8000-9000	0,33	0,56	-0,23	72,77	86,48
9000-10000	0,35	0,19	0,16	73,12	86,67
10000-20000	2,11	2,76	-0,66	75,23	89,44
20000-30000	0,75	2,89	-2,14	75,98	92,32
30000-40000	9,33	1,18	8,15	85,31	93,51
40000-50000	0,87	0,14	0,74	86,18	93,64
50000-60000	0,69	2,00	-1,32	86,87	95,65
60000-70000	0,02	1,59	-1,57	86,88	97,23
70000-80000	0,04	0,18	-0,15	86,92	97,42
80000-90000	0,01	0,84	-0,83	86,93	98,25
90000-100000	0,05	0,03	0,02	86,98	98,28
100000-200000	0,75	0,18	0,57	87,73	98,46
200000-300000	0,16	0,02	0,14	87,88	98,48
300000-400000	0,20	0,04	0,16	88,08	98,52
400000-500000	0,05	1,09	-1,04	88,13	99,61
500000-1000000	0,06	0,11	-0,05	88,19	99,72
1000000-2000000	11,71	0,26	11,46	99,90	99,97
2000000+	0,10	0,03	0,08	100,00	100,00

Tabela 31 wyraźnie pokazuje różnice w zachowaniach użytkowników z obu zbiorów w kwestii dobierania źródeł swoich retweetów. Niepożądani użytkownicy częściej podają dalej wiadomości z mało popularnych źródeł. Mają wyraźną przewagę w retweetowaniu wiadomości publikowanych przez konta o współczynniku poniżej 0,9. Użytkownicy normalni bardzo często odwoływali się do wiadomości Donalda Trumpa i Hillary Clinton co jest wyraźnie widoczne w dużych skokach wartości, w których znajdują się retweety wiadomości od kandydatów. W drugim zbiorze retweety kandydatów były marginalne. Jedyna wyraźna zgodność występuje w kwestii podawania dalej wiadomości użytkowników ze stosunkiem wartości z przedziału (1,3], w którym często występują zwykłe konta. Osiągnięcie podobnego stosunku wartości przez trolla nie jest dużym wyczynem, dlatego, tacy użytkownicy nie mogą być traktowani poważnie wyłącznie na podstawie badanej cechy, w przeciwieństwie do użytkowników o wartości stosunku rzędu setek czy tysięcy.

6.2.6. Kategoryzacja portali, do których prowadzą linki

Kategoryzacja portali opiera się na analizie pierwszych kilkudziesięciu domen z każdego ze zbiorów. W tabelach przedstawiono wyłącznie po 15 najpopularniejszych.

Niepożądani użytkownicy

Tabela 32: Najczęściej występujące domeny z uwzględnieniem wiadomości we wszystkich językach pochodzących ze zbioru niepożądanych użytkowników

Domena	liczba linków	udział w wiadomościach z linkami
bit.ly	214759	23.42%
riafan.ru	102590	11.19%
livejournal.com	99212	10.82%
goo.gl	32654	3.56%
twitter.com	31065	3.39%
dlvr.it	27735	3.02%
gazeta.ru	27683	3.02%
j.mp	22350	2.44%
rt.com	17226	1.88%
ift.tt	15897	1.73%
youtu.be	12240	1.33%
vesti.ru	11217	1.22%
kiev-news.com	9483	1.03%
youtube.com	9182	1.00%
kievsmi.net	8932	0.97%

Tabela 33: Najczęściej występujące domeny z uwzględnieniem wiadomości wyłącznie w języku angielskim pochodzących ze zbioru niepożądanych użytkowników

Domena	liczba linków	udział w wiadomościach z linkami
bit.ly	22468	12.76%
twitter.com	19228	10.92%
goo.gl	7006	3.98%
youtu.be	5753	3.27%
youtube.com	4864	2.76%
vine.co	3863	2.19%
cbslocal.com	3333	1.89%
ow.ly	2395	1.36%
chicagotribune.com	2036	1.16%
cleveland19.com	1977	1.12%
instagram.com	1732	0.98%
ift.tt	1695	0.96%
mysanantonio.com	1670	0.95%
detroitnews.com	1659	0.94%
vimeo.com	1658	0.94%

Tabele 32 i 33 pokazujące udziały procentowe linków prowadzących do najpopularniejszych domen wyraźnie pokazują, że wiadomości pisane w języku angielskim zawierają znacznie inne linki. Jest to zrozumiałe, że w przypadku tweetów pisanych po rosyjsku znajdziemy więcej portali lokalnych, lecz zastanawiająca jest niska pozycja YouTube oraz bardzo wysoka portalu livejournal.com w tweetach pisanych we wszystkich językach. Portal livejournal jest portalem anglojęzycznym, a jest cytowany najczęściej w wiadomościach w innych językach. Dziwny jest również relatywnie wysoki udział bit.ly. Jest to serwis pozwalający na tworzenie skróconych linków do stron. Bardzo możliwe, że niepożądani użytkownicy za jego pomocą chcą ukryć docelowe źródło odnośnika.

Wyjaśnić należy, że w tabelach uwzględniano skróty linków jako osobne domeny. Przykładem takich skrótów są goo.gl i youtu.be. Nie są to próby podszycia się.

Zwykli użytkownicy

Tabela 34: Najczęściej występujące domeny w wiadomościach w języku angielskim pochodzących ze zbioru normalnych użytkowników

Domena	liczba linków	udział w wiadomościach z linkami
twitter.com	1403596	30.42%
bit.ly	203313	4.41%
ln.is	115145	2.50%
fb.me	99379	2.15%
youtu.be	95173	2.06%
cnn.it	83958	1.82%
nyti.ms	79027	1.71%
ift.tt	68608	1.49%
dlvr.it	63887	1.38%
hill.cm	57374	1.24%
ow.ly	54848	1.19%
breitbart.com	52956	1.15%
youtube.com	50890	1.10%
twimg.com	48963	1.06%
hrc.io	48654	1.05%

W przypadku badania wszystkich wiadomości, również w innych językach niż angielski, najbardziej popularne domeny nie zmieniają się. Nie obserwujemy tutaj takiego wpływu języka jak w przypadku zbioru niepożądanych użytkowników. Wynika to z bardzo małej liczby wiadomości nieanglojęzycznych w zbiorze.

Porównanie wyników

Analizując tabele 33 i 34 pokazujące najpopularniejsze domeny w linkach dla zbiorów, można potwierdzić wcześniejsze podejrzenia co do portalu bit.ly. W zbiorze trolli i spamerów jest on bardzo często używany. Normalni użytkownicy również z niego korzystają, lecz znajdując się na drugim miejscu popularności ma udział prawie 7 razy mniejszy niż twitter.com. W drugim zbiorze znajduje się on na pierwszym miejscu z małą przewagą nad drugim twitter.com. Portale zamieniły się więc miejscami w rankingach, a stosunki ich udziałów diametralnie się zmieniły.

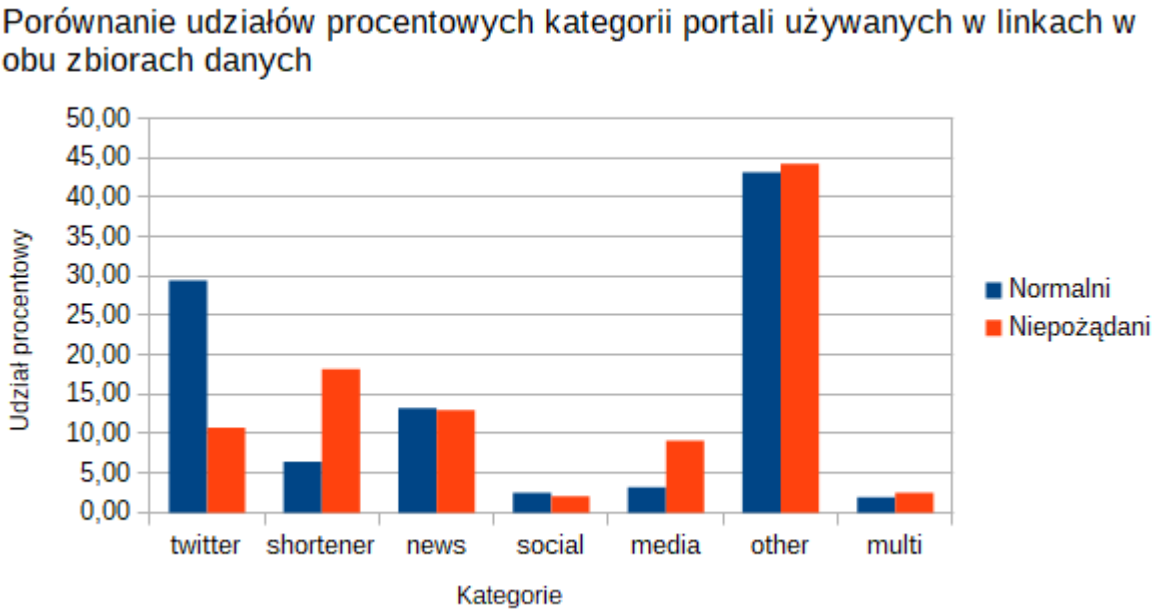
Większość odnośników z bit.ly prowadzi jednak do obrazków, które są wyświetlane w wiadomości. W innych przypadkach, gdy żaden nie wyświetla się w wiadomości, tak jak już wcześniej wspomniano, może być to próba zakamuflowania niebezpiecznych linków za

ich skrótami. Użytkownik, który nie jest tego świadomy może wejść w stronę z reklamami lub w najgorszym wypadku zainfekować swój komputer.

Kategoryzacja URL

Dokonano prostej kategoryzacji najpopularniejszych domen używanych w odnośnikach. Na podstawie ich analizy wyróżniono 7 kategorii:

- twitter - odnośniki prowadzące do innych zasobów portalu Twitter. Linki prowadzące do Twittera zostały wydzielone do osobnej kategorii z powodu ich znacznej popularności;
- shortener - najpopularniejsze serwisy pozwalające na generowanie skrótów linków, w tym bardzo popularny w zbiorach bit.ly. Należy podkreślić, że nie są tutaj wliczane oryginalnie skrócone linki popularnych portali informacyjnych. Lista portali: bit.ly, ow.ly, goo.gl, tinyurl.com;
- news - portale informacyjne znajdujące się w pierwszej pięćdziesiątce najbardziej popularnych w każdym ze zbiorów. Wiele portali ma własne generatory skrótów do swoich stron, powodujące, że linki nie są oczywiste. Lista portali: cnn.it, hill.cm, nyti.ms, washingtonpost.com, huffingtonpost.com, politi.co, washex.am, nytimes.com, thehill.com, fxn.ws, foxnews.com, politico.com, theguardian.com, dailycaller.com, nbcnews.to, cnn.com, chicagotribune.com, cbslocal.com, cleveland19.com, mysanantonio.com, detroitnews.com, cleveland.com, abc7news.com, abc7.com, dailym.ai, nbcchicago.com, seattletimes.com, nbcwashington.com, nydailynews.com;
- social - portale społecznościowe, z wyłączeniem Twittera, który jest traktowany jako osobna kategoria. Lista portali: fb.me, instagram.com, facebook.com;
- media - odnośniki prowadzące do stron z treścią multimedialną. Lista portali: youtu.be, youtube.com, vine.co, vimeo.com;
- multi - kategoria przypisywana wiadomościom, które zawierają wiele linków. Jest ich na tyle mało, że dalsza kategoryzacja w prostej analizie nie ma sensu;
- other - kategoria zawierająca elementy, które nie zostały zaliczone do każdej z poprzednich grup.



Rysunek 27: Histogram porównuje udziały procentowe poszczególnych kategorii portali, do których publikowane były odnośniki w przypadku obu typów użytkowników. Widać, że niepożądani użytkownicy częściej publikują skrócone linki i treści multimedialne.

Widoczny jest wyraźny wpływ najpopularniejszych używanych portali przedstawionych w tabelach 33 i 34 na udziały kategorii przedstawione na histogramie 27. Po kategoryzacji portali jeszcze bardziej wyróżnia się to, że portale do skracania linków są bardzo popularne u niepożądanych użytkowników. Sporą przewagę mają również w kategorii linków z treściami multimedialnymi. Patrząc od strony normalnych użytkowników widać, że w ich przypadku dominują linki z treściami prowadzącymi do Twittera. W pozostałych kategoriach obie grupy uzyskują podobne wyniki. Szczegółowe udziały procentowe kategorii znajdują się w tabeli 35.

Duży udział procentowy kategorii other wynika z małej liczby skategoryzowanych portali, których jest prawie 5000 w zbiorze niepożądanych użytkowników oraz ponad 20000 w zbiorze normalnych. Ich kategoryzacja byłaby zbyt długa i utrudniona przez brak znajomości portali popularnych w USA.

Tabela 35: Zestawienie udziałów procentowych kategorii portali, do których prowadzą linki publikowane w zbiorach danych.

Kategoria	Normalni, %	Niepożądani, %	Różnica, %
twitter	29,45	10,80	18,66
shortener	6,42	18,21	-11,79
news	13,24	13,02	0,22
social	2,52	2,08	0,44
media	3,22	9,12	-5,90
multi	1,96	2,51	-0,55
other	43,18	44,26	-1,08

6.3. Analiza zależności między użytkownikami

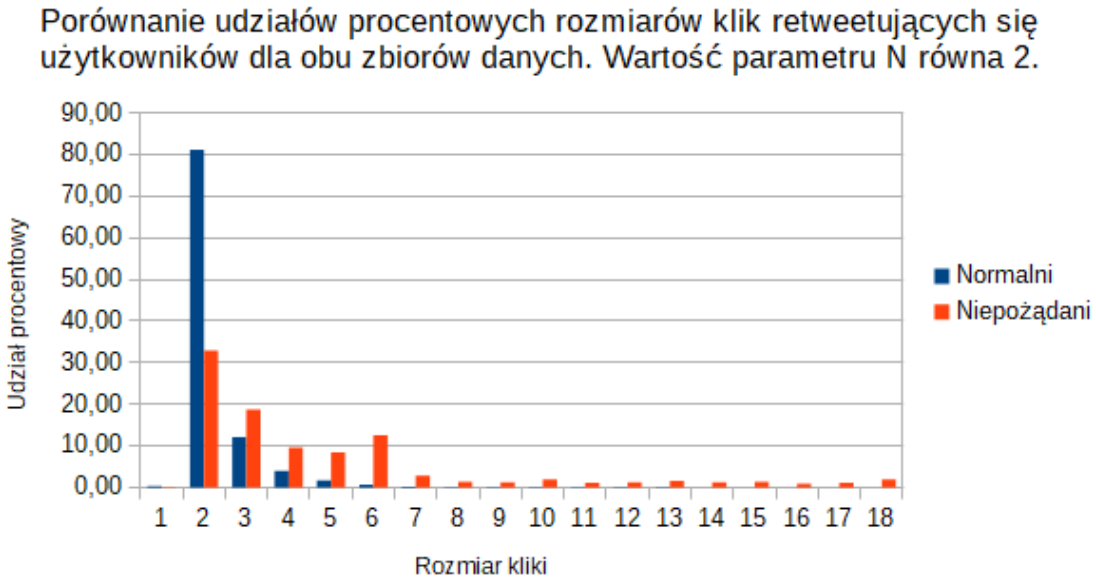
6.3.1. Rozmiar maksymalnej kliki użytkownika w grafie retweetujących się użytkowników.

Z wykorzystaniem biblioteki NetworkX [9] tworzony jest graf nieskierowany, w którym użytkownicy połączeni są krawędzią w przypadku zaszła między nimi następująca zależność: jeden z nich podawał dalej wiadomości drugiego minimum N razy (parametr N używany w dalszej części opisu). Następnie wyszukiwany jest rozmiar maksymalnej kliki w grafie dla każdego wierzchołka.

W zbiorze IRA zawierającym niepożądanych użytkowników jest 212 313 retweetów w języku angielskim na całkowitą liczbę wiadomości równą 595 823. Na 2557 użytkowników wypowiadających się po angielsku retweetowało 1735. 431 z nich podało dalej tylko jedną wiadomość.

Drugi zbiór, zawierający normalnych użytkowników zawiera 7 039 936 retweetów na 10 972 269 wiadomości. Wyraźnie widać, że normalni użytkownicy częściej korzystają z omawianego mechanizmu. Z grupy wszystkich 1 660 249 użytkowników retweetowało 1 118 165, w tym aż 682 268 z nich wyłącznie jeden raz. Duża liczba użytkowników podających wiadomość dalej tylko raz wynika z faktu ogólnej mniejszej aktywności użytkowników w zbiorze.

Wartość parametru N=2



Rysunek 28: Porównanie udziałów procentowych rozmiarów maksymalnych klik retweetujących się użytkowników w zbiorach normalnych i niepożądanych użytkowników dla wartości parametru N=2. Kliki użytkowników normalnych osiągają mniejsze rozmiary.

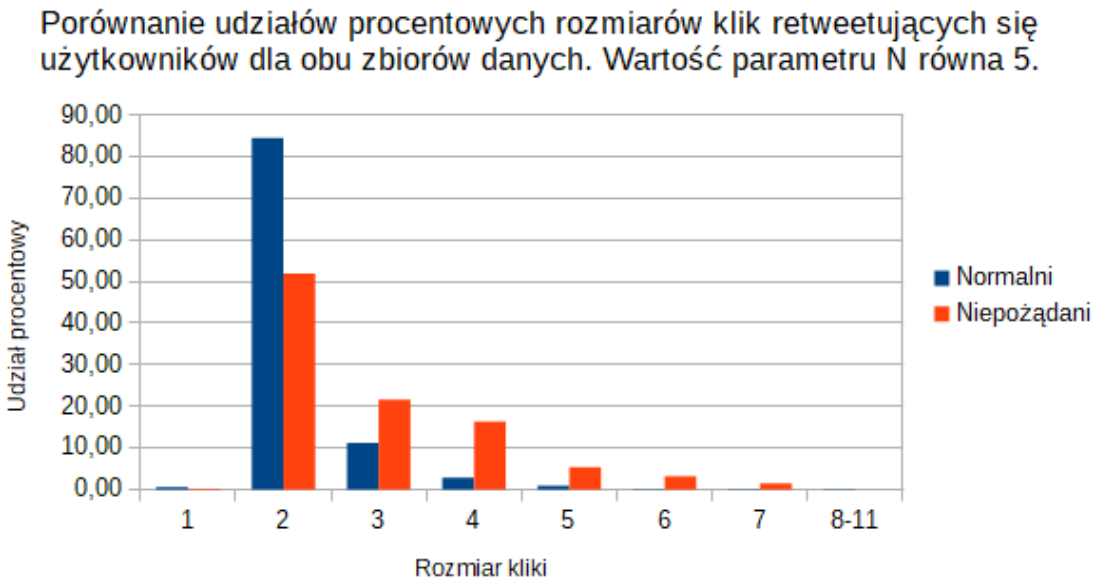
Histogram 28 pokazuje, że użytkownicy niepożądani o wiele częściej wchodzą w skład większych klik. Ich udziały procentowe przeważają we wszystkich grupach oprócz jednej,

w której dominują zwykli użytkownicy. Jak już stwierdzono w analizie 6.2.5 użytkownicy niepożądani podają dalej wiele wiadomości z mało wiarygodnych źródeł. Może to wskazywać na to, że często ze sobą współpracują. Wiele retweetów w zbiorze niepożądanych użytkowników dotyczyły wiadomości z tego samego zbioru. Szczegółowe wyniki przedstawione na wykresie 28 umieszczone są w tabeli 36. Jak widać po wynikach, możliwe jest na Twitterze retweetowanie swojej własnej wiadomości. Zostało to sprawdzone i nie jest to błąd.

Tabela 36: Porównanie udziałów procentowych rozmiarów maksymalnych klik retweetujących się użytkowników w zbiorach normalnych i niepożądanych użytkowników dla wartości parametru N=2.

Rozmiar kliki	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
1	0,34	0,11	0,22	0,34	0,11
2	81,13	32,90	48,24	81,47	33,01
3	12,07	18,73	-6,66	93,54	51,74
4	4,00	9,59	-5,60	97,54	61,34
5	1,64	8,40	-6,76	99,17	69,73
6	0,58	12,56	-11,98	99,75	82,30
7	0,14	2,86	-2,72	99,89	85,15
8	0,05	1,31	-1,27	99,94	86,46
9	0,02	1,20	-1,18	99,96	87,66
10	0,01	1,94	-1,93	99,97	89,61
11	0,01	1,14	-1,13	99,99	90,75
12	0,00	1,26	-1,25	99,99	92,00
13	0,01	1,54	-1,53	100,00	93,55
14	0,00	1,26	-1,26	100,00	94,80
15	0,00	1,37	-1,37	100,00	96,17
16	0,00	0,86	-0,86	100,00	97,03
17	0,00	1,09	-1,09	100,00	98,12
18	0,00	1,88	-1,88	100,00	100,00

Wartość parametru N=5



Rysunek 29: Porównanie udziałów procentowych rozmiarów maksymalnych klik retweetujących się użytkowników w zbiorach normalnych i niepożądanych użytkowników dla wartości parametru N=5. Większe kliki są domeną niepożądanych użytkowników. W porównaniu z wartością parametru N=2 kliki się zmniejszyły.

Na histogramie 29 widać zmniejszenie się maksymalnych rozmiarów klik po zwiększeniu parametru N do 5. Wynika z tego, że spora liczba użytkowników niepożądanych wchodzących w skład maksymalnych klik, miała od 2 do 5 retweetowanych wiadomości drugiego użytkownika. Samo jądro tych klik składa się z o wiele bardziej współpracujących kont.

Mimo zmniejszenia rozmiaru klik nadal występuje ich sporo o rozmiarze większym niż 3. Tak samo jak w poprzednim przypadku, we wszystkich kategoriach, oprócz rozmiaru kliki równego 1 i 2, przeważają niepożądani użytkownicy. Szczegółowe udziały rozmiarów klik dla obu zbiorów przedstawione są w tabeli 37.

Tabela 37: Porównanie udziałów procentowych rozmiarów maksymalnych klik retweetujących się użytkowników w zbiorach normalnych i niepożądanych użytkowników dla wartości parametru N=5.

Rozmiar kliki	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
1	0,55	0,23	0,32	0,55	0,23
2	84,50	51,94	32,56	85,05	52,17
3	11,18	21,55	-10,37	96,24	73,72
4	2,78	16,36	-13,57	99,02	90,08
5	0,81	5,35	-4,53	99,83	95,43
6	0,10	3,10	-3,00	99,93	98,53
7	0,02	1,47	-1,45	99,95	100,00
8-11	0,05	0,00	0,05	100,00	100,00

6.3.2. Liczba maksymalnych klik w grafie retweetów, w których znajduje się użytkownik.

Tak jak w analizie 6.3.1 tworzony jest graf nieskierowany, w którym użytkownicy połączeni są krawędzią w przypadku gdy zaszła między nimi zależność: jeden z nich retweetował drugiego minimum N razy (parametr N używany w dalszej części opisu). Dla każdego badanego wierzchołka zliczana jest liczba maksymalnych klik dla dowolnego wierzchołka, w których jest obecny badany. Liczone są wyłącznie kliki o rozmiarze nie mniejszym niż 3. Zliczanie mniejszych nie ma większego sensu w kwestii wykrywania niepożądanych zachowań.

Udziały procentowe, klik o rozmiarze równym minimum 3

Tabela 38: Rozkład wartości liczb maksymalnych klik dla dowolnego użytkownika o rozmiarze nie mniejszym niż 3, w których znajduje się badany użytkownik. Parametr N=2.

Liczba klik	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0	81,469	33,010	48,460	81,469	33,010
1	6,961	10,394	-3,433	88,430	43,404
2	2,867	6,853	-3,986	91,297	50,257
3	1,573	3,712	-2,139	92,870	53,969
4	1,138	3,370	-2,231	94,008	57,339
5	0,772	2,227	-1,455	94,780	59,566
6	0,606	1,256	-0,650	95,387	60,822
7	0,482	0,971	-0,489	95,869	61,793
8	0,355	1,542	-1,187	96,224	63,335
9	0,311	1,485	-1,174	96,535	64,820
10	0,250	0,800	-0,550	96,785	65,620
11-20	1,402	11,479	-10,077	98,187	77,099
21-30	0,558	5,254	-4,696	98,745	82,353
31-40	0,283	1,428	-1,145	99,028	83,781
41-50	0,189	1,085	-0,896	99,217	84,866
51-60	0,133	0,800	-0,667	99,350	85,665
61-70	0,088	0,171	-0,083	99,438	85,837
71-80	0,066	0,171	-0,105	99,504	86,008
81-90	0,069	0,171	-0,102	99,573	86,179
91-100	0,040	0,228	-0,189	99,613	86,408
101-200	0,179	1,028	-0,849	99,792	87,436
201-300	0,069	1,199	-1,130	99,861	88,635
301-400	0,033	0,628	-0,595	99,894	89,263
401-500	0,017	0,742	-0,725	99,911	90,006
501-600	0,011	0,343	-0,331	99,923	90,348
601-700	0,013	0,286	-0,272	99,936	90,634
701-800	0,004	0,286	-0,282	99,940	90,919
801-900	0,007	0,400	-0,393	99,947	91,319
901-1000	0,006	0,228	-0,222	99,953	91,548
1001-2000	0,023	0,857	-0,833	99,977	92,404
2001-3000	0,010	0,286	-0,276	99,987	92,690
3001-4000	0,004	0,628	-0,625	99,990	93,318
4001-5000	0,001	0,685	-0,684	99,991	94,003
5001-6000	0,003	0,171	-0,168	99,994	94,175
6001-7000	0,000	0,228	-0,228	99,994	94,403
7001-8000	0,001	0,171	-0,170	99,996	94,575
8001-9000	0,000	0,228	-0,228	99,996	94,803
9001-10000	0,001	0,228	-0,228	99,996	95,031
10001-20000	0,001	1,199	-1,198	99,998	96,231
20001-30000	0,001	0,742	-0,741	99,999	96,973
30001-40000	0,000	0,571	-0,571	99,999	97,544
40001-50000	0,001	0,228	-0,228	100,000	97,773
50001-60000	0,000	0,286	-0,286	100,000	98,058
60001-70000	0,000	0,057	-0,057	100,000	98,115
70001-80000	0,000	0,057	-0,057	100,000	98,172
80001-90000	0,000	0,114	-0,114	100,000	98,287
90001-100000	0,000	0,171	-0,171	100,000	98,458
10001-200000	0,000	1,142	-1,142	100,000	99,600
20001-300000	0,000	0,343	-0,343	100,000	99,943
30001-400000	0,000	0,057	-0,057	100,000	100,000

Tabela 39: Rozkład wartości liczb maksymalnych klik dla dowolnego użytkownika o rozmiarze nie mniejszym niż 3, w których znajduje się badany użytkownik. Parametr N=5.

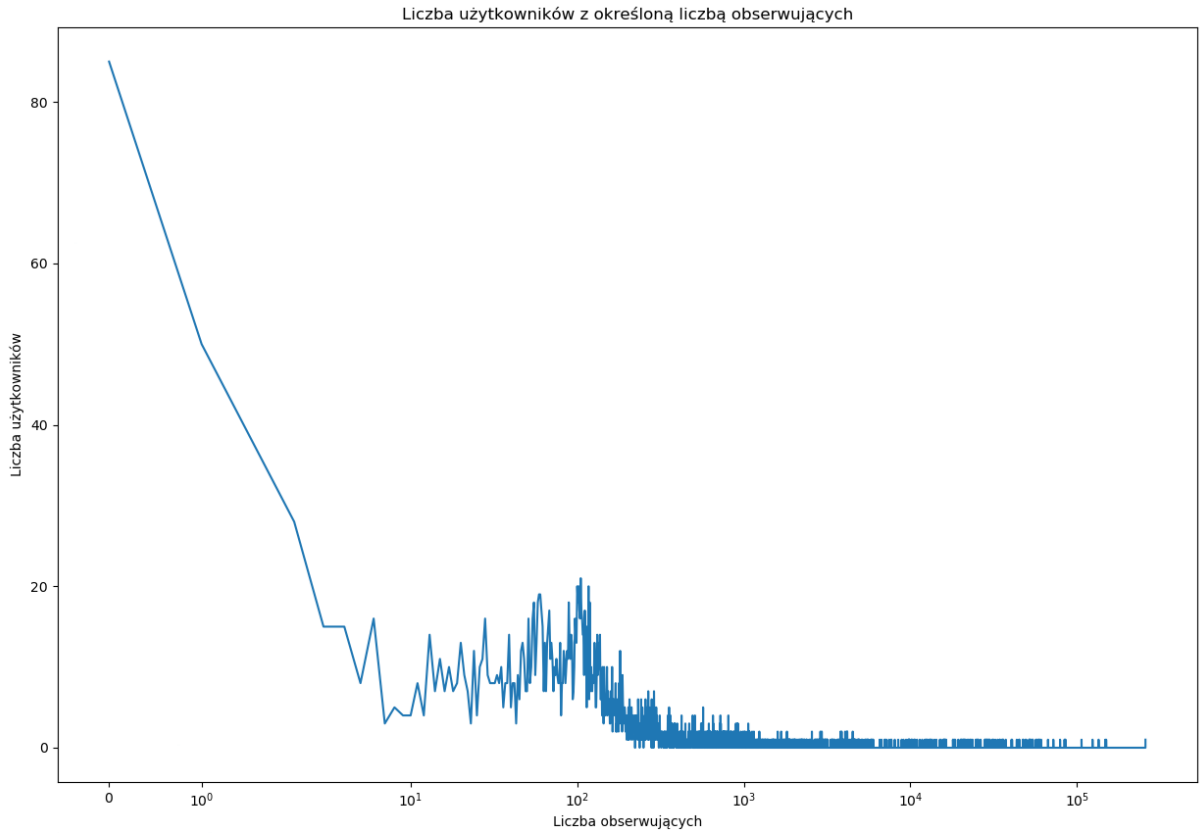
Liczba klik	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0	85,054	52,171	32,883	85,054	52,171
1	6,963	8,682	-1,719	92,017	60,853
2	2,641	4,264	-1,623	94,658	65,116
3	1,370	2,171	-0,801	96,027	67,287
4	0,852	2,636	-1,784	96,879	69,922
5	0,608	2,868	-2,260	97,487	72,791
6	0,418	2,791	-2,373	97,905	75,581
7	0,318	3,101	-2,783	98,222	78,682
8	0,211	2,946	-2,734	98,434	81,628
9	0,203	2,248	-2,045	98,637	83,876
10	0,144	1,550	-1,406	98,781	85,426
11-20	0,642	4,884	-4,242	99,423	90,310
21-30	0,247	0,853	-0,605	99,671	91,163
31-40	0,075	0,388	-0,312	99,746	91,550
41-50	0,048	0,853	-0,805	99,794	92,403
51-60	0,046	0,698	-0,652	99,839	93,101
61-70	0,021	0,543	-0,521	99,861	93,643
71-80	0,020	0,388	-0,368	99,880	94,031
81-90	0,015	0,078	-0,063	99,895	94,109
91-100	0,005	0,388	-0,383	99,900	94,496
101-200	0,048	2,326	-2,278	99,948	96,822
201-300	0,018	1,860	-1,842	99,966	98,682
301-400	0,007	0,853	-0,846	99,972	99,535
401-500	0,007	0,000	0,007	99,979	99,535
501-600	0,002	0,233	-0,231	99,980	99,767
601-700	0,002	0,000	0,002	99,982	99,767
701-800	0,000	0,000	0,000	99,982	99,767
801-900	0,003	0,078	-0,074	99,985	99,845
901-1000	0,002	0,000	0,002	99,987	99,845
1001-2000	0,003	0,078	-0,074	99,990	99,922
2001-3000	0,003	0,078	-0,074	99,993	100,000
3001+	0,007	0,000	0,007	100,000	100,000

Na podstawie wyników osiągniętych w tabelach 38 i 39 można stwierdzić, że normalni użytkownicy nie są członkami wielu maksymalnych klik. Nie współpracują z tak dużą liczbą innych użytkowników jak użytkownicy ze zbioru niepożądanych. Znacząca większość z nich nie znajduje się w ani jednej klice. W przypadku obu wartości parametru N, pokrycie podzbioru retweetujących się normalnych użytkowników dla liczby klik równej od 0 do 2 stanowi więcej niż 90%. W przypadku niepożądanych użytkowników rozkład wartości jest całkowicie inny. Tak duże liczby klik w ich przypadku spowodowane są tym, że współpracują między sobą. Użytkownicy niepożądani często wzajemnie się retweetują. Mają tez podobne źródła retweetów. Często nie są one wiarygodne, co było już wykazane w analizie 6.2.5.

6.3.3. Liczba obserwujących

Analiza została przeprowadzona dla wszystkich użytkowników, również tych piszących w innych językach niż angielski. Dotyczy to głównie zbioru niepożądanych użytkowników.

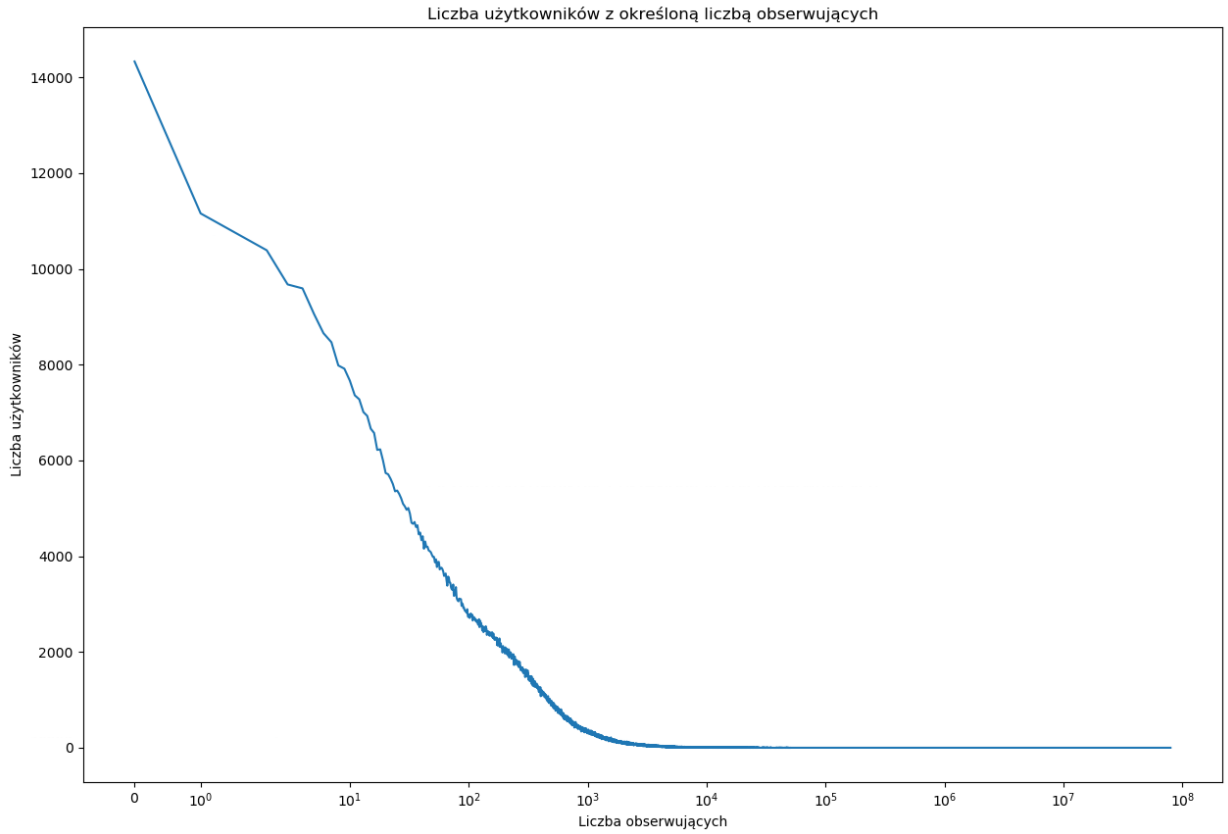
Niepożądani użytkownicy



Rysunek 30: Wykres przedstawia liczbę niepożądanych użytkowników obserwowanych przez konkretną liczbę innych użytkowników. Oś OX jest w skali logarytmicznej z powodu dużego zakresu jej wartości. Z powodu małej liczby użytkowników w zbiorze IRA (około 3300) wykres jest bardzo wyraźnie "poszarpany".

Wykres 30 pokazuje, że najwięcej użytkowników ma liczbę obserwujących bliską zera. Widoczny jest również wzrost w okolicach 100 obserwujących. Wartości rzędu tysięcy są bardzo rzadkie. Wykres jest bardzo poszarpany, ponieważ badany zbiór liczy nieznacznie ponad 3000 użytkowników.

Zwykli użytkownicy



Rysunek 31: Wykres przedstawia liczbę niepożądanych użytkowników obserwowanych przez konkretną liczbę innych użytkowników. Oś OX jest w skali logarytmicznej z powodu dużego zakresu jej wartości. W przeciwieństwie do niepożądanych użytkowników wykres jest dużo bardziej gładki, co spowodowane jest dużą dysproporcją liczby użytkowników pomiędzy zbiorami. Tutaj jest ich znacznie więcej.

Na wykresie 31 można zobaczyć podobne tendencje jak u użytkowników niepożądanych. Większość kont jest obserwowanych przez bardzo małe, bliskie zeru liczby innych użytkowników. Liczby obserwujących przekraczające 1000 są rzadkie. Należy jednak podkreślić, że najpopularniejsze konta obecne w zbiorze osiągają wartości przekraczające 10 milionów obserwujących.

Porównanie

Z powodu bardzo dużych dysproporcji w liczbie użytkowników i osiąganych przez nich wartości prezentowanych na wykresach 30 i 31 ich bezpośrednie porównywanie nie ma sensu. O wiele więcej można zaobserwować z histogramu 32 prezentującego udziały procentowe użytkowników z poszczególnych przedziałów liczby obserwujących.

Rysunek 32: Porównanie udziałów procentowych użytkowników z liczbą obserwujących z konkretnego przedziału.

Udział wiadomości użytkowników niepożądanych na histogramie 32 jest większy aż do przełomowego przedziału 101-200. Występuje w kulminacyjna przewaga wiadomości niepożądanych użytkowników - przedział zawiera 12,86% użytkowników normalnych i aż 23,44% niepożądanych. Poczynając od tego przedziału zmienia się tendencja. W kolejnych nieznacznie przeważają użytkownicy normalni. Szczegółowe dane przedstawione na histogramie znajdują się w tabeli 40. Użytkownicy normalni obserwowani są przez większe liczby osób, co wydaje się logicznym wynikiem. Wyróżniające się skoki wartości na wykresie spowodowane są wyłącznie przeskalowywaniem przedziałów o kolejne rzędy wartości.

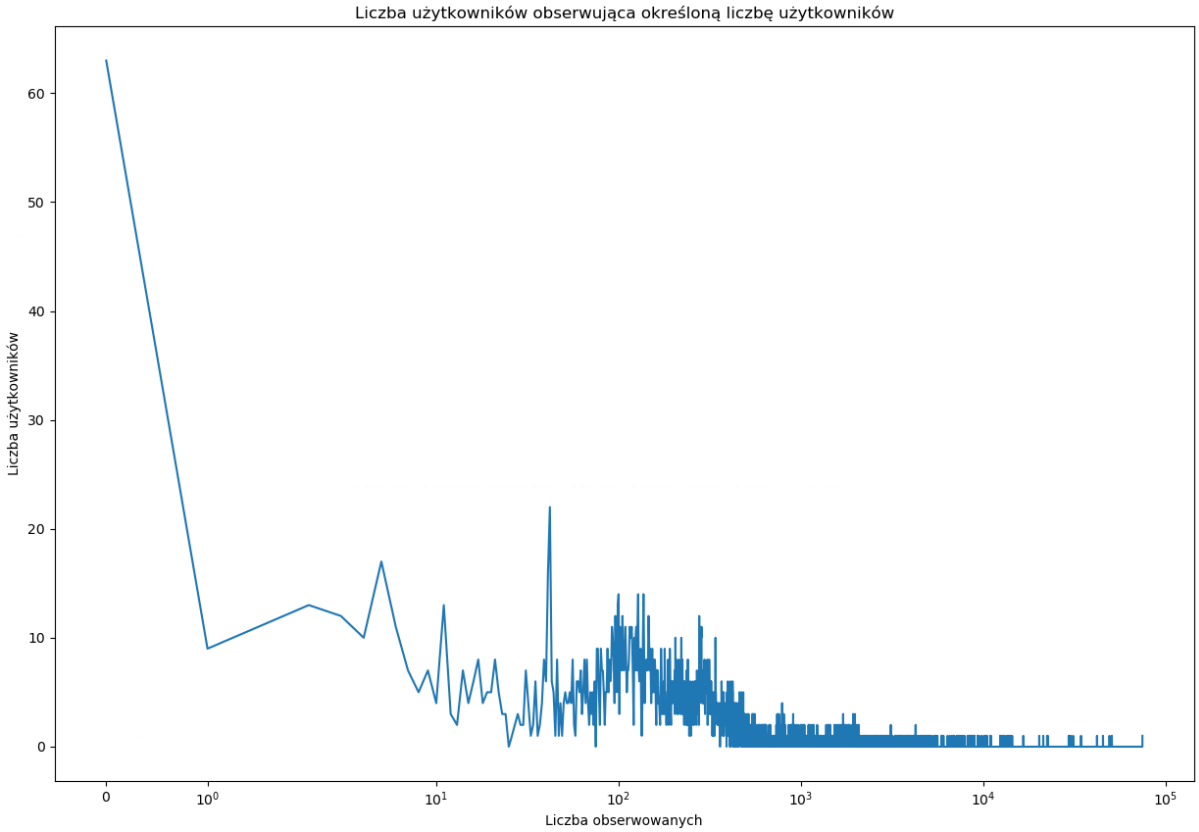
Tabela 40: Porównanie udziałów procentowych użytkowników z liczbą obserwujących z konkretnego przedziału.

Przedział	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0	0,77	2,58	-1,81	0,77	2,58
1	0,60	1,52	-0,92	1,37	4,10
2	0,56	0,85	-0,29	1,93	4,95
3	0,52	0,46	0,06	2,45	5,41
4	0,52	0,46	0,06	2,97	5,86
5	0,49	0,24	0,24	3,45	6,10
6	0,47	0,49	-0,02	3,92	6,59
7	0,46	0,09	0,36	4,37	6,68
8	0,43	0,15	0,28	4,80	6,83
9	0,43	0,12	0,30	5,23	6,95
10	0,41	0,12	0,29	5,64	7,08
11-20	3,55	2,70	0,85	9,19	9,78
21-30	2,86	2,70	0,16	12,05	12,48
31-40	2,50	2,52	-0,02	14,56	15,00
41-50	2,24	2,55	-0,31	16,79	17,55
51-60	2,06	4,43	-2,37	18,85	21,99
61-70	1,91	3,89	-1,98	20,76	25,87
71-80	1,75	2,73	-0,98	22,51	28,61
81-90	1,64	3,37	-1,73	24,15	31,98
91-100	1,53	3,77	-2,24	25,68	35,74
101-200	12,86	23,44	-10,58	38,53	59,19
201-300	9,69	8,14	1,55	48,22	67,32
301-400	7,36	4,16	3,20	55,58	71,48
401-500	5,65	2,92	2,74	61,23	74,40
501-600	4,39	2,73	1,66	65,62	77,13
601-700	3,46	2,34	1,12	69,08	79,47
701-800	2,78	2,19	0,59	71,86	81,66
801-900	2,29	1,88	0,40	74,15	83,54
901-1000	1,95	1,06	0,89	76,10	84,60
1001-2000	10,40	4,80	5,60	86,50	89,40
2001-3000	4,00	2,31	1,69	90,50	91,71
3001-4000	2,18	1,55	0,63	92,68	93,26
4001-5000	1,32	1,70	-0,38	94,00	94,96
5001-6000	0,88	0,70	0,18	94,88	95,66
6001-7000	0,64	0,06	0,58	95,52	95,72
7001-8000	0,49	0,18	0,31	96,01	95,90
8001-9000	0,38	0,09	0,29	96,39	95,99
9001-10000	0,32	0,27	0,04	96,71	96,26
10001-20000	1,55	1,12	0,43	98,26	97,39
20001-30000	0,53	1,06	-0,53	98,79	98,45
30001-40000	0,27	0,39	-0,13	99,06	98,85
40001-50000	0,17	0,30	-0,14	99,22	99,15
50001–100000	0,35	0,67	-0,32	99,57	99,82
100001-10000000	0,37	0,18	0,19	99,95	100,00
10000001-80000000	0,05	0,00	0,05	100,00	100,00

6.3.4. Liczba obserwowanych

Analiza została przeprowadzona dla wszystkich użytkowników, również tych piszących w innych językach niż angielski. Dotyczy to głównie zbioru niepożądanych użytkowników.

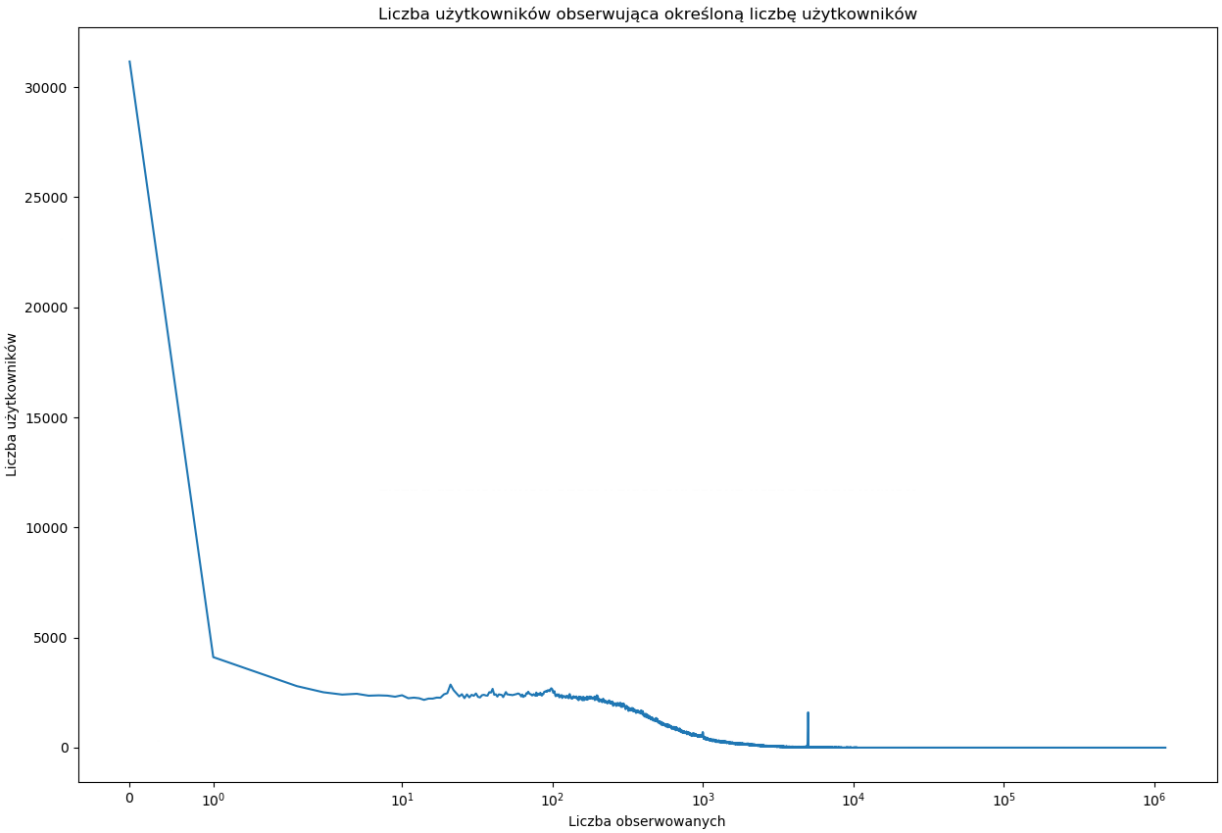
Niepożądani użytkownicy



Rysunek 33: Wykres przedstawia liczby użytkowników niepożądanych z konkretnymi liczbami obserwowanych. Oś OX jest w skali logarytmicznej z powodu dużego zakresu jej wartości. Użytkowników niepożądanych jest około 3300, co przekłada się na wyraźnie poszarpany wygląd wykresu.

Wykres 33 pokazuje, że najwięcej użytkowników niepożądanych posiada 0 obserwowanych. Sporo użytkowników obserwuje do 10 kont. Wzrost widoczny jest również w okolicach 100 obserwowanych. Wyraźne poszarpanie wykresu wynika z małej liczby użytkowników niepożądanych w zbiorze (nieznacznie ponad 3000).

Zwykli użytkownicy

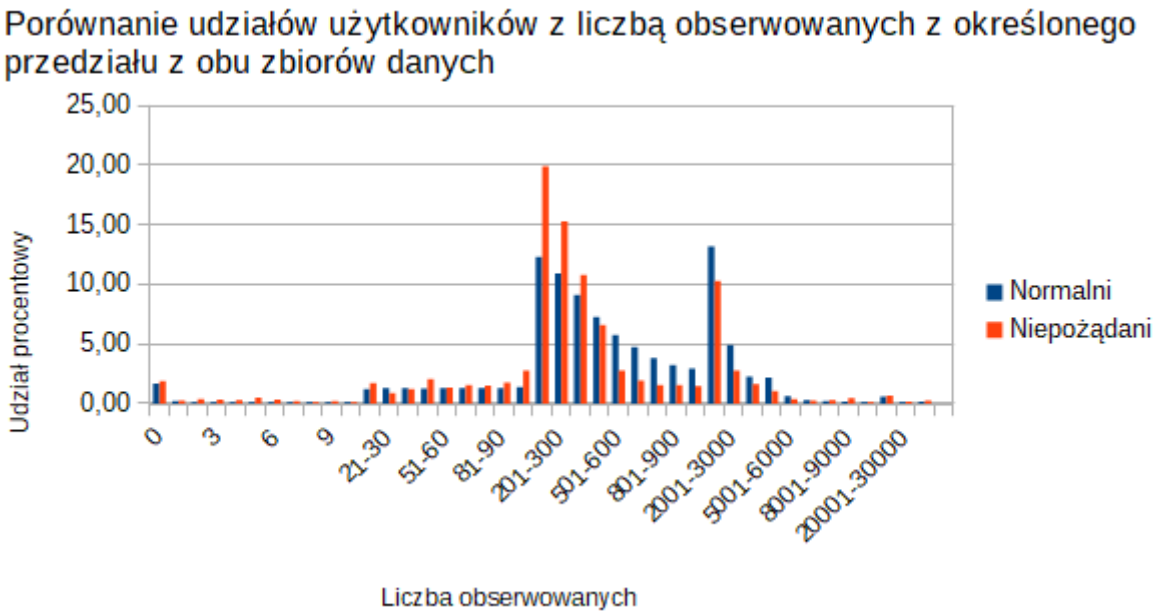


Rysunek 34: Wykres przedstawia liczby użytkowników normalnych z konkretnymi liczbami obserwowanych. Oś OX jest w skali logarytmicznej z powodu dużego zakresu jej wartości. Widoczny jest wyraźny spadek liczby użytkowników obserwujących ponad 300 kont.

Patrząc na wykres 34 można zaobserwować, że podobnie jak w przypadku niepożądanych użytkowników najwięcej kont nie obserwuje innych. Użytkownicy obserwują zwykle od 1 do 400 innych użytkowników. Po przekroczeniu granicy 400 obserwowanych widoczny jest już stopniowy spadek liczby kont.

Porównanie

Tak samo jak w przypadku analizy obserwujących występują duże dysproporcje w liczbie użytkowników prezentowanych na wykresach 33 i 34. Porównanie dokonano tak samo jak poprzednio poprzez obliczenie udziałów procentowych użytkowników z poszczególnych przedziałów liczb obserwowanych.



Rysunek 35: Porównanie udziałów procentowych użytkowników z liczbą obserwowanych z konkretnego przedziału. Wyraźna przewaga użytkowników niepożądanych w zakresie od 101 do 400 obserwowanych. Skoki wartości na wykresie są spowodowane skalowaniem przedziałów.

Analizując wykres 35 można zauważyć, że do liczby 400 obserwowanych wyraźnie przeważają użytkownicy niepożądani. Pokrycie ich zbioru do tej wartości wynosi 65,72%, a dla użytkowników normalnych tylko 47,20%, co można odczytać z tabeli 41. Podobnie jak w przypadku analizy obserwujących po pewnym przedziale następuje zmiana trendu. W prawie wszystkich kolejnych przedziałach nieznacznie przeważają użytkownicy normalni. Wyraźnie obserwują oni o wiele więcej kont. Wyróżniające się skoki wartości na wykresie, tak samo jak poprzednio, spowodowane są wyłącznie przeskalowywaniem przedziałów o kolejne rzędy wartości.

Tabela 41: Porównanie udziałów procentowych użytkowników z liczbą obserwowanych z konkretnego przedziału.

Przedział	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0	1,68	1,91	-0,24	1,68	1,91
1	0,22	0,27	-0,05	1,90	2,19
2	0,15	0,39	-0,24	2,05	2,58
3	0,14	0,36	-0,23	2,18	2,95
4	0,13	0,30	-0,17	2,31	3,25
5	0,13	0,52	-0,38	2,44	3,77
6	0,13	0,33	-0,21	2,57	4,10
7	0,13	0,21	-0,08	2,70	4,31
8	0,13	0,15	-0,02	2,83	4,46
9	0,12	0,21	-0,09	2,95	4,68
10	0,13	0,12	0,01	3,08	4,80
11-20	1,23	1,73	-0,50	4,31	6,53
21-30	1,31	0,88	0,43	5,62	7,41
31-40	1,30	1,25	0,06	6,92	8,65
41-50	1,29	2,06	-0,77	8,22	10,72
51-60	1,30	1,37	-0,07	9,52	12,09
61-70	1,29	1,55	-0,25	10,81	13,63
71-80	1,30	1,52	-0,22	12,11	15,15
81-90	1,32	1,79	-0,47	13,43	16,95
91-100	1,41	2,79	-1,39	14,84	19,74
101-200	12,31	19,89	-7,58	27,15	39,63
201-300	10,92	15,27	-4,36	38,07	54,90
301-400	9,14	10,81	-1,68	47,20	65,72
401-500	7,28	6,59	0,69	54,49	72,30
501-600	5,77	2,76	3,00	60,25	75,07
601-700	4,74	1,94	2,79	64,99	77,01
701-800	3,84	1,55	2,29	68,83	78,56
801-900	3,23	1,55	1,68	72,05	80,11
901-1000	2,93	1,46	1,47	74,98	81,57
1001-2000	13,21	10,29	2,91	88,19	91,86
2001-3000	4,91	2,76	2,15	93,10	94,62
3001-4000	2,27	1,67	0,60	95,37	96,30
4001-5000	2,21	1,06	1,15	97,58	97,36
5001-6000	0,63	0,39	0,24	98,21	97,75
6001-7000	0,31	0,24	0,07	98,52	98,00
7001-8000	0,23	0,30	-0,08	98,75	98,30
8001-9000	0,17	0,46	-0,29	98,92	98,75
9001-10000	0,14	0,12	0,02	99,06	98,88
10001-20000	0,59	0,70	-0,11	99,64	99,57
20001-30000	0,15	0,18	-0,03	99,80	99,76
30001-100000	0,17	0,24	-0,07	99,97	100,00
100001-2000000	0,03	0,00	0,03	100,00	100,00

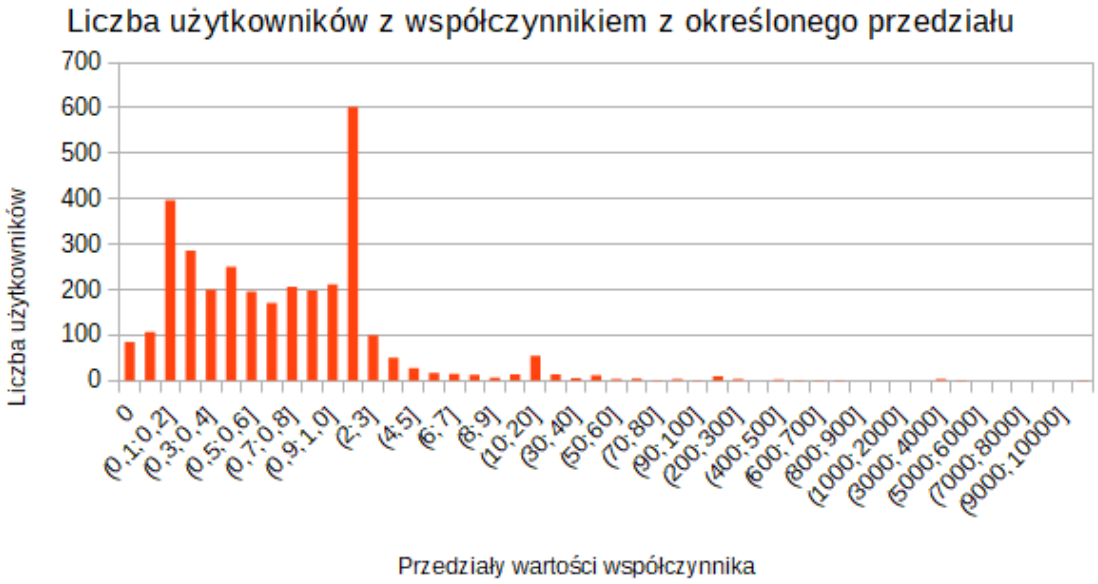
6.3.5. Stosunek liczb obserwowanych i obserwujących

Stosunek wartości został obliczony z wykorzystaniem wzoru 6. Uwzględniani w analizie są wszyscy użytkownicy z obu zbiorów.

Niepożądani użytkownicy

Średnia wartość stosunku dla niepożądanych użytkowników jest równa 8.02. Uwzględniając konta wypowiadające się w języku angielskim współczynnik ten spada do wartości 7.64. Wartość współczynnika jest nieznacznie podnoszona przez obecność w zbiorze IRA oficjalnych kont organizacji oraz znanych osobistości, z których część może być fałszywymi profilami. Dla 38 użytkowników stosunek jest większy od 50. Do zablokowanych kont o największych współczynnikach (uwzględniane wszystkie języki) należą między innymi:

- konto byłego premiera i prezydenta Czeczenii Ramzana Kadyrowa
  (followers: 123 989, following: 10, ratio: 12 398.90);
- konto ministera obrony Federacji Rosyjskiej Siergieja Szojgu
  (followers: 4176, following: 5, ratio: 835.20);
- konto Departamentu Polityki Gospodarczej i Rozwoju w Moskwie
  (followers: 31 010, following: 39, ratio: 795.13);
- konta informacyjne większych miast rosyjskich: NovostiPermi, NovostiYarsl, NovostiMsk, NovostiCrimea, NovostiKrsnrsk, NovostiOmska, NovostiRnD, NovostiChel, NovostiSPb, NovostiArkh, NovostiNN, NovostiUf, NovostiKzn, NovostiVolga, LuganskNovosti, NovostiVrnezh.
  (ratio od 55.42 do 588.60);
- konto o nazwie “TrueNavalny”, rosyjskiego prawnika, publicysty i działacza politycznego Aleksieja Nawalnego. Możliwe, że jest to jedynie próba podszycia się pod jego osobę
  (followers: 43 445, following: 99, ratio: 438.84).



Rysunek 36: Wykres przedstawia liczby niepożądanych użytkowników przypadające na konkretne przedziały wartości współczynnika popularności obliczanego na podstawie atrybutów każdego konta na Twitterze. Widzimy, że wartość dla największej grupy użytkowników znajduje się w przedziale od 1 do 2.

Na diagramie 36 wyraźnie widoczne jest, że stosunek wartości dla większości użytkowników jest nie większy niż 2. W przedziale od 1 do 2 znajduje się zdecydowanie najwięcej użytkowników.

Zwykli użytkownicy

Dla całego zbioru normalnych użytkowników średni współczynnik jest równy 65.36. Jest nieznacznie zawyżany, ponieważ w zbiorze znajduje się zbyt wiele kont celebrytów, polityków lub portali, w przypadku osiągane wartości są skrajnie inne niż dla zwykłego użytkownika. Dla zobrazowania można wyróżnić kilka kont:

- konto amerykańskiego komika i scenarzysty telewizyjnego Conan'a O'Brien (followers: 28 602 116, following: 1, ratio: 28 602 116);
- konto o nazwie UberFacts publikujące tweety będące ciekawostkami ze świata (followers: 13 956 388, following: 1, ratio: 13 956 388);
- konto publikujące newsy ze świata: BBC Breaking News (followers: 41 622 306, following: 3, ratio: 13 874 102);
- konto prezydenta Stanów Zjednoczonych Donalda Trumpa (followers: 73 496 164, following: 47, ratio: 1 563 748.17).

Rysunek 37: Wykres przedstawia liczby normalnych użytkowników przypadające na konkretne przedziały wartości współczynnika popularności obliczanego na podstawie atrybutów każdego konta na Twitterze. Tak samo jak w przypadku niepożądanych użytkowników najbardziej liczna jest grupa przypadająca na przedział od 1 do 2.

Wyniki stosunku obserwujących i obserwowanych dla normalnych użytkowników na wykresie 37 są bardzo podobne do uzyskanych w przypadku niepożądanych użytkowników. Stosunek najczęściej nie przekracza wartości 3.

Porównanie wyników

Tabela 42: Udział procentowy użytkowników ze stosunkiem liczb obserwujących i obserwowanych z danego przedziału.

Przedział	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0,0	0,771	2,581	-1,810	0,771	2,581
(0,0;0,1]	5,053	3,249	1,803	5,823	5,831
(0,1;0,2]	9,146	12,056	-2,910	14,969	17,886
(0,2;0,3]	8,871	8,685	0,186	23,840	26,572
(0,3;0,4]	7,870	6,073	1,797	31,710	32,645
(0,4;0,5]	6,758	7,622	-0,865	38,467	40,267
(0,5;0,6]	5,726	5,952	-0,226	44,194	46,219
(0,6;0,7]	5,026	5,193	-0,167	49,220	51,412
(0,7;0,8]	4,549	6,286	-1,737	53,769	57,698
(0,8;0,9]	4,272	6,013	-1,741	58,040	63,711
(0,9;1,0]	5,094	6,438	-1,344	63,134	70,149
(1;2]	21,073	18,281	2,792	84,207	88,430
(2;3]	5,041	3,037	2,004	89,248	91,467
(3;4]	2,230	1,549	0,681	91,478	93,015
(4;5]	1,268	0,850	0,418	92,746	93,866
(5;6]	0,853	0,516	0,337	93,599	94,382
(6;7]	0,619	0,456	0,164	94,218	94,838
(7;8]	0,477	0,395	0,082	94,696	95,232
(8;9]	0,377	0,213	0,164	95,072	95,445
(9;10]	0,313	0,425	-0,112	95,385	95,870
(10;20]	1,578	1,670	-0,093	96,963	97,540
(20;30]	0,615	0,425	0,190	97,578	97,965
(30;40]	0,338	0,182	0,156	97,917	98,148
(40;50]	0,231	0,364	-0,134	98,147	98,512
(50;60]	0,162	0,121	0,041	98,310	98,633
(60;70]	0,129	0,152	-0,022	98,439	98,785
(70;80]	0,104	0,061	0,044	98,543	98,846
(80;90]	0,088	0,121	-0,034	98,631	98,968
(90;100]	0,071	0,030	0,041	98,702	98,998
(100;200]	0,415	0,304	0,111	99,117	99,302
(200;300]	0,191	0,121	0,070	99,308	99,423
(300;400]	0,120	0,000	0,120	99,428	99,423
(400;500]	0,083	0,091	-0,008	99,511	99,514
(500;600]	0,057	0,061	-0,003	99,569	99,575
(600;700]	0,044	0,030	0,014	99,613	99,605
(700;800]	0,037	0,061	-0,023	99,650	99,666
(800;900]	0,031	0,000	0,031	99,681	99,666
(900;1000]	0,032	0,000	0,032	99,714	99,666
(1000;2000]	0,126	0,000	0,126	99,839	99,666
(2000;3000]	0,046	0,000	0,046	99,886	99,666
(3000;4000]	0,023	0,121	-0,098	99,909	99,787
(4000;5000]	0,013	0,061	-0,047	99,922	99,848
(5000;6000]	0,012	0,000	0,012	99,935	99,848
(6000;7000]	0,008	0,000	0,008	99,943	99,848
(7000;8000]	0,006	0,000	0,006	99,949	99,848
(8000;9000]	0,005	0,000	0,005	99,954	99,848
(9000;10000]	0,004	0,000	0,004	99,958	99,848
(10000;20000]	0,017	0,030	-0,013	99,976	99,879
(20000;30000]	0,008	0,091	-0,083	99,983	99,970
(30000;40000]	0,004	0,000	0,004	99,987	99,970
(40000;50000]	0,002	0,000	0,002	99,989	99,970
(50000;60000]	0,002	0,030	-0,029	99,991	100,000
(60000,20000000]	0,009	0,000	0,009	100,000	100,000

Najwięcej użytkowników w obu zbiorach posiada stosunek obserwujących i obserwowanych z przedziału od 1 do 2. W przypadku użytkowników niepożądanych możemy zaobserwować nieznaczną przewagę w przedziałach mniejszych od 1 - pokrycie zbioru jest 7% większe dla niepożądanych użytkowników co widać w tabeli 42. Użytkownicy normalni mają nieznaczną przewagę przy najwyższych wartościach omawianego stosunku. Wynika to z obecności wielu kont polityków wypowiadających się w badanych tematach.

Tabela 43: Średnia liczba obserwujących dla użytkowników znajdujących się w danym przedziale wartości współczynnika popularności. Ograniczono się do analizy przedziałów nie większych niż 100.

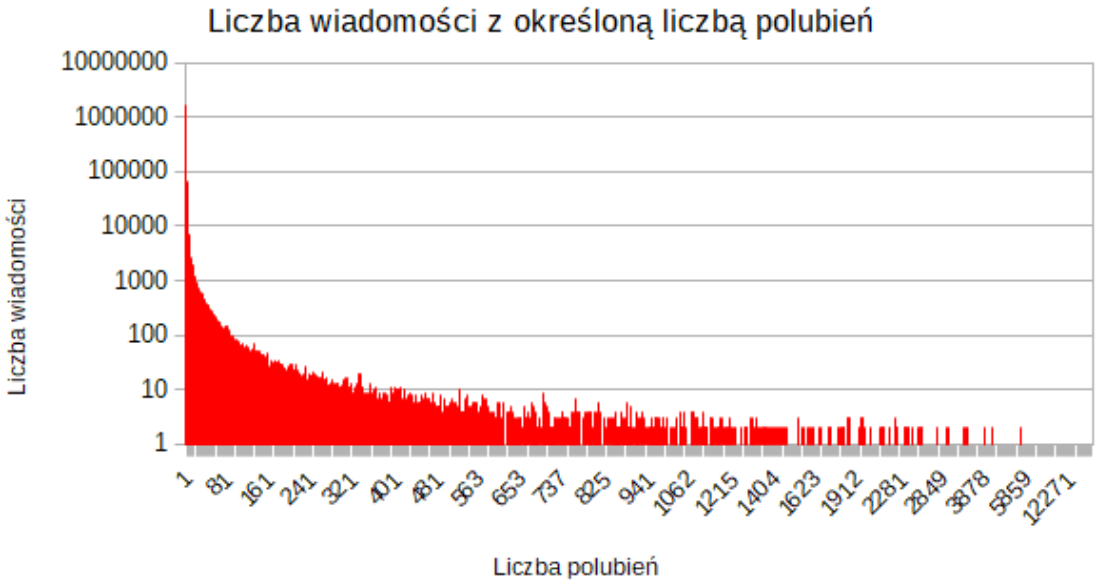
Przedział	Normalni	Niepożądani
(0,0;0,1]	36,68	21,03
(0,1;0,2]	110,44	102,91
(0,2;0,3]	203,61	155,08
(0,3;0,4]	303,73	252,71
(0,4;0,5]	413,65	276,23
(0,5;0,6]	539,34	255,09
(0,6;0,7]	646,68	246,17
(0,7;0,8]	789,30	343,93
(0,8;0,9]	1 116,05	423,57
(0,9;1,0]	2 815,11	1 157,22
(1;2]	1 997,49	2 490,83
(2;3]	2 320,64	6 379,82
(3;4]	3 147,25	6 782,22
(4;5]	4 106,83	7 136,97
(5;6]	4 691,56	734,69
(6;7]	5 562,86	9 015,86
(7;8]	6 382,35	5 383,00
(8;9]	7 363,78	4 533,14
(9;10]	8 074,60	7 339,75
(10;20]	11 360,64	5 125,91
(20;30]	17 838,53	7 645,67
(30;40]	25 988,94	13 090,43
(40;50]	33 360,35	28 591,64
(50;60]	33 841,22	5 322,75
(60;70]	41 708,53	21 659,40
(70;80]	47 778,25	72,00
(80;90]	56 623,63	19 732,00
(90;100]	57 608,81	6 744,00

Porównanie stosunków obserwujących i obserwowanych nie pokazuje między nimi wyraźnych różnic. Współczynnik przydziela taką samą wartość dla użytkownika, który posiada 2 obserwujących i 1 obserwowanego oraz dla innego użytkownika, który posiada 1000 obserwujących i 500 obserwowanych. Wiarygodność tego drugiego jest o wiele większa. W tabeli 43 potwierdza się to, że normalni użytkownicy mają więcej obserwujących niż niepożądani zaliczeni do tej samej grupy ze względu na obliczony stosunek. Połączenie obliczonych stosunków wraz z liczbą obserwowanych może dać o wiele lepsze wyniki.

6.3.6. Liczba polubień wiadomości

Analizowane są wszystkie wiadomości w obu zbiorach. Język wypowiedzi nie ma wpływu na liczbę polubień. Dzięki temu możemy przeanalizować więcej wiadomości w zbiorze niepożądanych użytkowników (zbiór IRA).

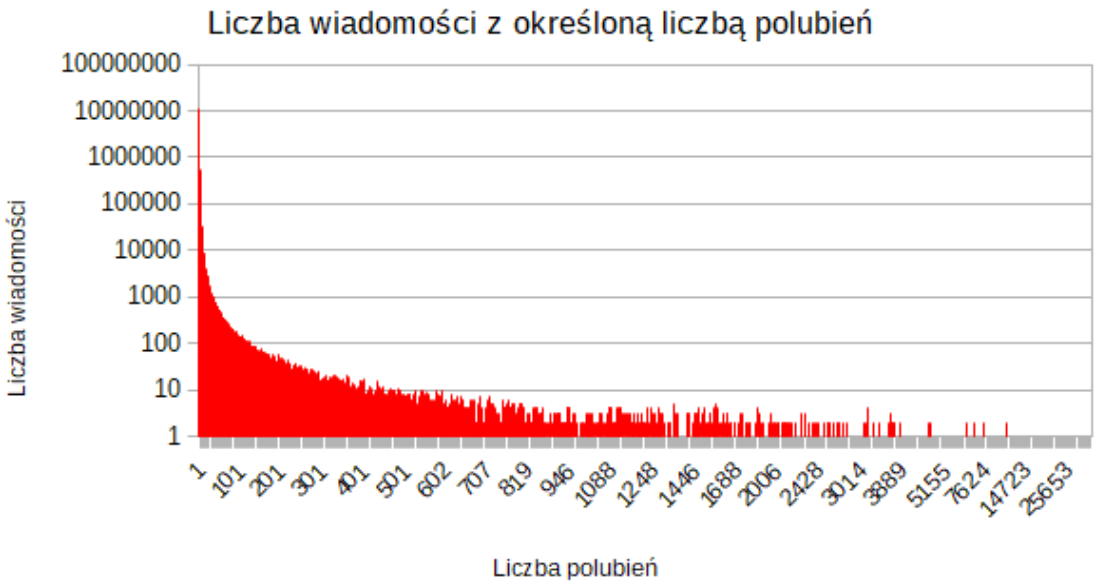
Niepożądani użytkownicy



Rysunek 38: Diagram przedstawia liczbę wiadomości w zbiorze niepożądanych użytkowników o określonych liczbach polubień tych wiadomości. Widzimy, że najwięcej wiadomości ma zero lub kilka polubień. Po przekroczeniu granicy około 100 polubień wykres powoli opada. Ma on jednak już wtedy bardzo małe wartości patrząc na liczbę wszystkich analizowanych wiadomości.

Wykres 38 pokazuje, że większość wiadomości nie posiada polubień lub ich wartość jest bliska zeru. Polubienia wiadomości nie są często używanym mechanizmem.

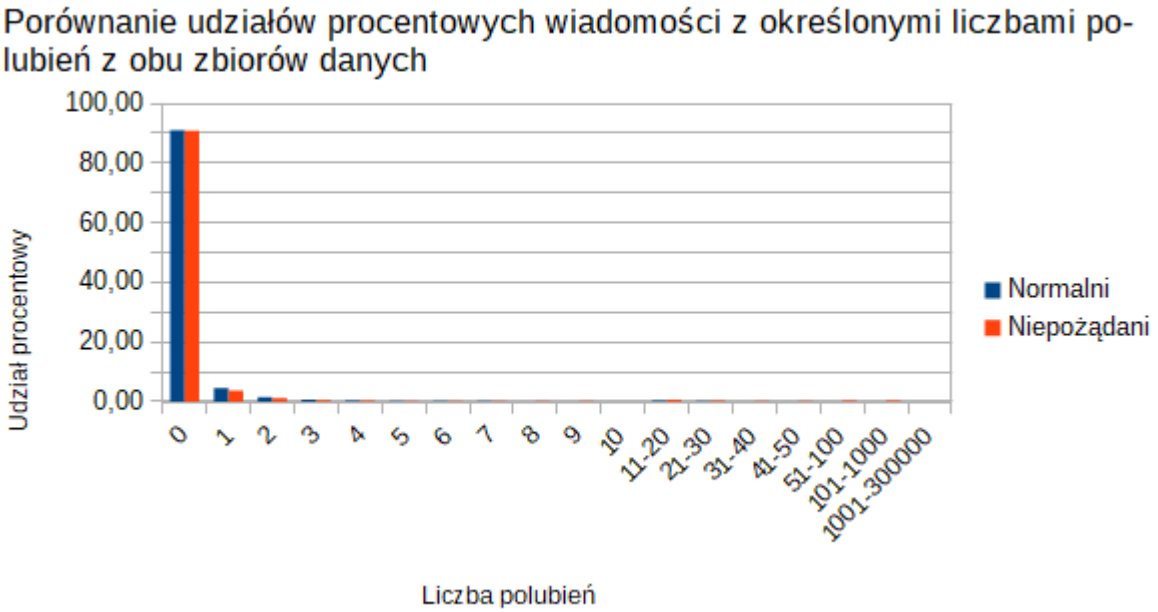
Zwykli użytkownicy



Rysunek 39: Diagram przedstawia liczbę wiadomości w zbiorze normalnych użytkowników o określonych liczbach polubień tych wiadomości. Tak samo jak w przypadku niepożądanych użytkowników, licznik polubień wiadomości ma w większości przypadków wartości bliskie zeru. Tak samo, od około 100 polubień widzimy powolny spadek liczby wiadomości.

Sytuacja widoczna na wykresie 39 jest analogiczna jak w przypadku niepożądanych użytkowników. Liczba polubień znacznej części wiadomości jest bliska lub równa zeru.

Porównanie wyników



Rysunek 40: Histogram zestawia ze sobą udziały procentowe wiadomości z konkretnymi liczbami polubień w zbiorach danych. Widzimy, że rozkłady wyglądają praktycznie tak samo. W obu zbiorach zgodnie około 90% stanowią wiadomości bez polubień.

Zgodnie z wcześniejszymi wykresami prezentującymi konkretne liczby wiadomości, na diagramie 40 widzimy, że wiadomości nie posiadające polubień dominują w przypadku obu zbiorów. W obu przypadkach stanowią aż 90% całości. Wyraźnie pokazuje to, że system polubień jest ignorowany przez użytkowników Twittera. Szczegółowe dane zaprezentowane na histogramie przedstawione są w tabeli 44.

W obu zbiorach rozkład wartości jest bardzo podobny. Drobne różnice występują jedynie dla 10% wiadomości w każdym zbiorze, które otrzymały jakiekolwiek polubienia. W tabeli 44 można zobaczyć, że niepożądani użytkownicy mają nieznacznie więcej wiadomości, których licznik polubień prezentuje bardziej okazałe wartości. Są to jednak liczby pomijalnie małe, na podstawie których nie możemy wyciągać dalej idących wniosków.

Analiza pokazała, że w przypadku Twittera polubienia nie są mechanizmem popularnym. Prawie wszyscy użytkownicy je ignorują, przez co możemy spodziewać się, że przeanalizowana cecha nie sprawdzi się dobrze w klasyfikacji.

Tabela 44: Tabela prezentuje szczegółowe udziały procentowe wiadomości z konkretnymi liczbami polubień w obu zbiorach danych.

Przedział	Normalni, %	Niepożądani, %	Różnica, %	Suma norm., %	Suma niep., %
0	91,10	90,94	0,16	91,10	90,94
1	4,59	3,65	0,94	95,70	94,59
2	1,48	1,16	0,32	97,18	95,75
3	0,70	0,59	0,11	97,88	96,35
4	0,41	0,39	0,02	98,29	96,74
5	0,27	0,29	-0,02	98,56	97,03
6	0,19	0,22	-0,03	98,76	97,25
7	0,14	0,18	-0,03	98,90	97,43
8	0,11	0,15	-0,04	99,01	97,58
9	0,09	0,13	-0,04	99,10	97,70
10	0,07	0,11	-0,04	99,17	97,81
11-20	0,35	0,67	-0,32	99,53	98,48
21-30	0,14	0,35	-0,22	99,66	98,83
31-40	0,07	0,22	-0,15	99,74	99,06
41-50	0,05	0,15	-0,10	99,78	99,21
51-100	0,10	0,34	-0,24	99,88	99,55
101-1000	0,10	0,40	-0,30	99,99	99,95
1001-300000	0,01	0,05	-0,04	100,00	100,00

7. Ewaluacja skuteczności opracowanych cech w klasyfikacji użytkowników portalu społecznościowego

W rozdziale zostanie omówiony proces stworzenia zbioru testowo-treningowego (7.1), na którym przeprowadzone zostaną wszystkie dalsze badania. Przedstawione zostaną wyniki indywidualne wszystkich stworzonych cech (7.2) oraz ważniejszych podgrup tych cech (7.3). Z wykorzystaniem różnych metod zostaną opracowane finalne zbiory pozwalające osiągać najlepsze wyniki (7.4). Zostanie również porównana przydatność różnych algorytmów klasyfikacji w badanym problemie (7.5). Rozdział kończy się przeanalizowaniem użytkowników, na których myli się najlepsze osiągnięte rozwiązanie (7.6) oraz całościowym podsumowaniem i wyciągnięciem wniosków z wszystkich wyników, które udało się uzyskać (7.7).

7.1. Utworzenia zbioru treningowo-testowego

7.1.1. Usunięcie części użytkowników

W analizach 6.1.9, 6.1.10 oraz w dużo mniejszym stopniu w 6.2.2 zauważone zostały podejrzane zachowania użytkowników, którzy powinni być normalni. Tak jak już wcześniej omówiono, zbiór może zawierać małe liczby niepożądanych użytkowników. Kosztem posiadania większej ilości danych jest w nich mały "szum".

Dokonano klastrowania algorytmem k-means z wykorzystaniem wszystkich cech związanych z analizami, w których wyraźnie zaobserwowano występowanie oczywistych niepożądanych użytkowników. Do grupy cech zaliczyły się:

- maksymalna długość serii z uwzględnieniem retweetów,
- średnia 5 najdłuższych serii z uwzględnieniem retweetów,
- średnie podobieństwo wiadomości,
- liczba podobnych wiadomości (próg podobieństwa równy 0,7).

Klastrowanie zostało wykonane na grupie 10 972 264 wiadomości w języku angielskim ze stworzonego zbioru normalnych użytkowników. Dokonano podziału na 5 klastrów algorytmem k-means z euklidesową funkcja obliczania odległości. Liczba 5 klastrów została wybrana eksperymentalnie. Dodatkowe klastry były bardzo nieliczne. Wyniki udziałów procentowych poszczególnych grup widoczne są w tabeli 45. Większość wiadomości zalicza się do klastra nr 2 i 3. Udział pozostałych jest marginalny.

Tabela 45: Udziały procentowe klastrów po przeprowadzeniu klastrowania w oparciu o analizę serii i podobieństwa wiadomości zbioru normalnych użytkowników.

Nr klastra	Udział, %
1	8,03
2	0,10
3	0,19
4	91,64
5	0,04

Tabela 46: Środki klastrów po przeprowadzeniu klastrowania w oparciu o analizę serii i podobieństwa wiadomości zbioru normalnych użytkowników.

	Numer klastra				
Cecha	1	2	3	4	5
Maksymalna długość serii	86,21	161,11	788,67	10,13	2 140,00
Średnia 5 najdłuższych serii	57,94	116,50	608,37	6,98	973,75
Średnie podobieństwo wiadomości	0,59	0,88	0,64	0,53	0,79
Liczba podobnych wiadomości	10,67	1 434,21	63,51	0,40	1 027,83

Na podstawie wyników z tabeli 46 możemy wyróżnić 5 grup użytkowników. Użytkownicy zaklasyfikowani do grup zostali w ramach możliwości manualnie sprawdzeni, zwłaszcza podejrzane grupy.

Grupa 4 to przeciętni normalni użytkownicy użytkujący okazjonalnie Twittera. Grupa nr 1 to normalni użytkownicy, którzy bardzo aktywnie użytkują portal. Ich długie serie wynikają z dużej liczby retweetów. W większości korzystają z urządzeń mobilnych co dodaje im wiarygodności. Pozostałe grupy to niepożądani użytkownicy. Grupy 2 i 5 piszą prawie identyczne wiadomości, które są masowo publikowane. W grupie nr 3 wiadomości są mało do siebie podobne, lecz nadal są zautomatyzowane co widać po analizie serii wiadomości. Możemy w tej grupie odszukać wiadomości botów publikowane co równe 5 minut, ogłoszenia sklepu z naklejkami na zderzaki samochodowe o tematyce politycznej czy mało wiarygodne konta informacyjne publikujące wiadomość co kilka sekund.

Ze zbioru zostały odfiltrowane wszystkie wiadomości użytkowników z klastrów o numerach: 2, 3 i 5. Zostało usunięte 42 838 wiadomości należących do 15 użytkowników, którzy zostali manualnie przeanalizowani i zakwalifikowani jako niepożądani. W zbiorze nie pozostała ani jedna wiadomość należąca do usuniętych użytkowników.

7.1.2. Przeskalowanie jednego zbioru

Występuje duża dysproporcja w liczebnościach zbioru IRA zawierającego niepożądanych użytkowników oraz zbioru stworzonego na bazie id dostępnych w bazie Harvard Dataverse, czyli zbioru normalnych użytkowników. Większy zbiór został więc odpowiednio przeskalowany. Z pierwszego wybrano wszystkie wiadomości w języku angielskim jakie były dostępne. W przypadku drugiego najpierw dokonano klastrowania wiadomości z wykorzystaniem algorytmu k-means (4 klastry, 500 iteracji, euklidesowa funkcja obliczania odległości), a następnie z klastrów losowo wybrano wiadomości tak aby zachować pomiędzy nimi początkowe proporcje. Liczba klastrów została tak samo jak w punkcie poświęconym filtrowaniu zbioru wybrana eksperymentalnie. Zwiększanie liczby klastrów prowadziło do podziałów na bardzo mało liczne grupy. W klastrowaniu wzięto pod uwagę wszystkie cechy z wykluczeniem dziesięciu cech z analizy emocji. Wzięto pod uwagę jedynie najbardziej wyróżniającą się emocję. Jest to konsekwencją zbyt małej liczby pamięci RAM komputera, który nie był w stanie wczytać tak dużej liczby cech dotyczących ponad 11 milionów wiadomości.

Tabela 47: Udziały procentowe klastrów po przeprowadzeniu klastrowania w oparciu o wszystkie cechy na zbiorze normalnych użytkowników.

Nr klastra	Udział, %
1	7,71
2	92,22
3	0,01
4	0,06

Udział procentowy poszczególnych klastrów zbioru normalnych użytkowników przedstawiony jest w tabeli 47. Wszystkie cechy związane z seriami wiadomości uwzględniają retweety. Widoczna jest wyraźna przewaga udziałów 2 klastrów.

Tabela 48: Środki klastrów po przeprowadzeniu klastrowania w oparciu o wszystkie badane cechy na zbiorze normalnych użytkowników.

	Numer klastra			
Cecha	1	2	3	4
liczba znaków	98,97	86,81	86,60	98,19
liczba słów	18,01	15,49	15,45	17,96
sentyment	0,37	0,19	0,28	-0,02
subiektywność	0,46	0,34	0,32	0,38
maks. długość serii wiadomości	12,36	16,50	12,00	3,02
liczba linków	0,09	0,45	0,68	0,16
liczba hasztagów	0,31	0,35	0,16	0,11
liczba odniesień do użytkowników	1,20	1,02	0,37	1,18
liczba polubień	0,00	0,96	5 590,42	0,00
kategoria źródła publikacji	0,22	0,55	2,00	0,24
liczba obserwujących	4 005,89	10 465,14	38 461 910,25	3 767,97
liczba obserwowanych	2 795,90	3 538,59	738,42	1 287,72
stosunek obserwujących i obserwowanych	23,40	54,22	318 097,13	24,79
stosunek obserwu. i obserwo. retweetowanych	1 514 459,99	5 120,10	21 216,28	10 952 130,81
rozmiar maks. kliki w grafie retweetów	2,57	2,48	6,80	1,25
liczba obecności w maksymalnych klikach	17,38	25,75	16 794,47	1,44
średnia 5 serii wiadomości	7,94	11,27	9,40	2,27
maks. liczba wiadomości w czasie 15 minut	7,28	5,99	3,53	1,87
średnie podobieństwo wiadomości	0,50	0,53	0,58	0,31
liczba podobnych wiadomości (próg 0,7)	0,03	1,17	0,34	0,00
ulubione źródło publikacji użytkownika	3,20	3,49	5,10	3,20
kategoria portalu z odnośnika	3,10	3,70	3,23	3,16
najbardziej wyróżniająca się emocja	34,35	32,94	33,41	27,29

Pierwszy klaster zawiera użytkowników, którzy piszą dłuższe wiadomości o pozytywnym sentymencie, które są najbardziej subiektywne. Pomimo pozytywnego sentymentu, najbardziej wyróżniającą się emocją jest smutek. Retweetowane przez nich wiadomości pochodzą z bardzo wiarygodnych źródeł.

Drugi klaster zawierający najwięcej użytkowników wyróżnia się większą aktywnością na portalu, w porównaniu do pozostałych grup. Użytkownicy częściej używają hasztagów i zdarza im się opublikować podobną wiadomość, która może być formą jakiejś akcji. W pozostałych aspektach grupa się nie wyróżnia.

Najmniej liczna trzecia grupa zawiera polityków, celebrytów i portale informacyjne. Wiadomości z tej grupy są bardzo często podawane dalej co widoczne jest w analizie grafu retweetów oraz otrzymują bardzo duże liczby polubień. Konta posiadają bardzo duże liczby obserwujących, co przekłada się na bardzo dużą wartość stosunku obserwujących do obserwowanych. Użytkownicy są aktywni i nie odbiegają znacznie w kwestiach analizy serii od drugiego klastra.

Czwarta grupa wyróżnia się negatywnym sentymentem publikowanych wypowiedzi. Najczęściej wiadomości wyrażają smutek i rozczarowanie. Użytkownicy z tej grupy piszą bardzo rzadko, znacząco odbiegając od wyników osiąganych w poprzednich klastrach.

Sprawiają wrażenie użytkowników logujących się tylko raz na jakiś czas w celu podania dalej lub skomentowania bardzo ważnej dla nich wiadomości.

7.1.3. Połączenie obu zbiorów w finalny zbiór

Zbiór finalny składa się po połowie z reprezentantów obu zbiorów: niepożądanych i normalnych użytkowników. Zbiór normalnych użytkowników został przeskalowany zgodnie z klastrami i ich stosunkami opisanymi w sekcji 7.1.2. Równy stosunek klas w zbiorze treningowo-testowym pozwoli na bardziej dokładne nauczenie i późniejszą weryfikację zdolności klasyfikatora. W późniejszym etapie zbudowany klasyfikator zostanie również zastosowany na zbiorze składającym się z około 3% niepożądanych użytkowników.

7.2. Klasyfikacja z wykorzystaniem pojedynczych cech

Do zbadania zdolności klasyfikacyjnych każdej cechy opisanej w podpunkcie wykorzystano algorytm lasów losowych z domyślnymi parametrami dostępnymi w bibliotece scikit-learn oraz z 10-krotną kros-walidacją. Zbiór pozytywny w przypadku metryk oznacza niepożądanych użytkowników.

7.2.1. Wyniki klasyfikacji pojedynczych cech

Dla każdej dostępnej w zbiorze cechy zbudowano i przebadano klasyfikator. W tabeli 49 zamieszczone są indywidualne wyniki osiągnięte przez cechy.

Tabela 49: Wyniki klasyfikacji wykonanych z wykorzystaniem pojedyńczych cech.

Nr	Cecha	F1	Precyzja	Czułość	Dokładność	ROC AUC
1	liczba znaków	0,68	0,66	0,71	0,67	0,72
2	liczba słów	0,68	0,66	0,69	0,67	0,71
3	sentyment	0,74	0,75	0,73	0,74	0,80
4	subiektywność	0,58	0,55	0,62	0,56	0,58
5	tag sentymentu	0,57	0,50	0,66	0,51	0,51
6	maks. długość serii wiadomości	0,73	0,68	0,79	0,71	0,79
7	liczba linków	0,62	0,55	0,70	0,56	0,56
8	liczba hasztagów	0,58	0,74	0,47	0,66	0,66
9	liczba odniesień do użytkowników	0,67	0,74	0,61	0,70	0,70
10	liczba polubień wiadomości	0,23	0,62	0,14	0,53	0,53
11	kategoria źródła publikacji	0,83	0,73	0,97	0,81	0,84
12	liczba obserwujących	0,90	0,89	0,92	0,90	0,97
13	liczba obserwowanych	0,90	0,88	0,91	0,89	0,97
14	stosunek obserwujących i obserwowanych	0,98	0,98	0,99	0,98	1,00
15	stosunek obserwujących i obserwowanych retweetowanych	0,82	0,71	0,97	0,79	0,83
16	rozmiar maksymalnej kliki w grafie retweetów	0,64	0,63	0,66	0,63	0,67
17	liczba obecności w maksymalnych klikach	0,48	0,70	0,36	0,60	0,66
18	średnia 5 serii wiadomości	0,78	0,74	0,81	0,76	0,85
19	maks. liczba wiadomości w czasie 15 minut	0,71	0,61	0,86	0,66	0,71
20	średnie podobieństwo wiadomości	0,77	0,75	0,80	0,77	0,82
21	liczba podobnych wiadomości (próg 0,7)	0,11	0,58	0,06	0,51	0,51
22	ulubione źródło publikacji użytkownika	0,84	0,74	0,96	0,81	0,85
23	kategoria portalu z odnośnika	0,65	0,56	0,79	0,58	0,60
24	analiza emocji	0,81	0,82	0,80	0,81	0,90

Indywidualnie najbardziej wyróżniają się cechy oparte o liczby obserwujących i obserwowanych o miarach F1 mniejszych lub równych 0,9. Bardzo dobre wyniki indywidualne uzyskują również stworzone kategorie źródeł publikacji w obu wariantach: dotyczącym poszczególnych wiadomości oraz ulubionego źródła publikacji użytkownika. Bardzo dobry wynik uzyskuje analiza emocji, która odnosi wyraźnie lepsze miary od sentymentu. Wyniki opracowanych cech dotyczących częstości pisania wiadomości oraz średniego podobieństwa wiadomości osiągają poprawne wyniki zbliżające się do 0,8 F1.

Na drugim biegunie znajduje się liczba podobnych wiadomości osiągająca diametralnie inny wynik od średniego podobieństwa wiadomości, które zostało stworzone z użyciem tego samego narzędzia. Niewiele więcej wnosi liczba polubień wiadomości. W zbiorze znajduje się 12 cech o wartości miary F1 poniżej 0,7, w tym połowa z nich nie przekracza wartości 0,6.

7.2.2. Użytkownicy klasyfikowani do tych samych grup przez różne cechy

Numery w tabelach 50 i 51 odpowiadają numeracji cech przydzielonej w tabeli wyników indywidualnych o numerze 49. Na przekątnej możemy odczytać jaki % użytkowników normalnych lub niepożądanych ze zbioru wykryła dana cecha.

Tabela 50: Udział procentowy normalnych użytkowników obecnych w zbiorze, którzy zostali wykryci przez obie z pary cech.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	63,6	60,1	50,0	36,3	20,8	39,1	17,9	52,7	52,8	58,9	42,4	56,3	56,1	62,2	43,6	37,4	53,1	44,9	27,7	48,4	61,2	43,3	15,7	55,5
2	60,1	65,2	50,9	37,3	21,7	40,3	18,2	54,8	54,0	60,2	43,6	57,6	57,5	63,8	44,2	38,4	54,5	46,3	28,8	49,4	62,8	44,5	16,2	56,7
3	50,0	50,9	75,6	38,9	26,6	47,0	30,0	63,4	63,3	71,3	50,9	66,7	66,7	74,0	55,5	44,2	63,0	53,8	33,4	56,4	72,2	52,0	27,1	67,9
4	36,3	37,3	38,9	50,5	17,0	31,6	18,1	42,7	40,6	46,3	33,5	44,5	44,4	49,4	32,6	29,9	42,5	36,1	22,8	36,8	48,5	34,3	16,5	42,5
5	20,8	21,7	26,6	17,0	35,7	22,6	15,6	29,8	27,2	32,2	23,1	31,5	31,4	34,9	20,6	21,9	30,2	25,7	16,6	27,2	34,1	23,7	14,2	28,7
6	39,1	40,3	47,0	31,6	22,6	62,8	26,0	52,9	48,2	57,0	40,3	54,9	54,7	61,2	36,8	41,7	55,5	52,3	38,0	46,6	60,3	41,4	23,3	51,0
7	17,9	18,2	30,0	18,1	15,6	26,0	41,9	35,5	25,5	36,9	22,0	37,3	37,2	41,0	19,4	26,1	35,7	29,6	18,9	30,9	38,8	23,2	37,8	31,9
8	52,7	54,8	63,4	42,7	29,8	52,9	35,5	83,6	65,6	76,6	54,1	73,8	73,6	81,8	50,9	51,0	70,9	60,2	38,8	61,5	80,0	55,4	32,2	68,9
9	52,8	54,0	63,3	40,6	27,2	48,2	25,5	65,6	78,3	74,8	56,0	69,0	69,0	76,5	59,4	45,3	64,9	55,6	34,1	56,7	75,5	56,9	23,0	67,8
10	58,9	60,2	71,3	46,3	32,2	57,0	36,9	76,6	74,8	91,4	60,4	80,4	80,4	89,3	61,2	55,1	76,9	65,2	41,0	67,2	87,4	61,6	33,2	76,7
11	42,4	43,6	50,9	33,5	23,1	40,3	22,0	54,1	56,0	60,4	64,7	56,8	56,9	63,3	46,8	37,7	53,5	46,4	28,9	46,3	62,6	62,7	20,5	54,5
12	56,3	57,6	66,7	44,5	31,5	54,9	37,3	73,8	69,0	80,4	56,8	88,4	78,6	86,7	54,2	53,3	74,3	62,8	39,3	65,1	84,2	58,2	33,7	72,6
13	56,1	57,5	66,7	44,4	31,4	54,7	37,2	73,6	69,0	80,4	56,9	78,6	88,1	86,4	54,4	53,0	74,1	62,6	39,0	64,9	84,0	58,4	33,7	72,5
14	62,2	63,8	74,0	49,4	34,9	61,2	41,0	81,8	76,5	89,3	63,3	86,7	86,4	97,8	60,1	59,1	82,4	69,9	44,4	71,9	93,3	64,9	37,1	80,4
15	43,6	44,2	55,5	32,6	20,6	36,8	19,4	50,9	59,4	61,2	46,8	54,2	54,4	60,1	61,2	33,5	49,5	42,8	24,8	44,8	59,4	47,2	17,8	57,4
16	37,4	38,4	44,2	29,9	21,9	41,7	26,1	51,0	45,3	55,1	37,7	53,3	53,0	59,1	33,5	60,6	55,2	46,9	34,0	46,4	57,6	38,7	23,3	48,5
17	53,1	54,5	63,0	42,5	30,2	55,5	35,7	70,9	64,9	76,9	53,5	74,3	74,1	82,4	49,5	55,2	84,5	62,7	42,7	62,5	80,6	54,9	32,1	68,7
18	44,9	46,3	53,8	36,1	25,7	52,3	29,6	60,2	55,6	65,2	46,4	62,8	62,6	69,9	42,8	46,9	62,7	71,6	40,4	53,0	68,6	47,6	26,7	58,4
19	27,7	28,8	33,4	22,8	16,6	38,0	18,9	38,8	34,1	41,0	28,9	39,3	39,0	44,4	24,8	34,0	42,7	40,4	45,7	34,3	44,2	29,7	16,8	36,5
20	48,4	49,4	56,4	36,8	27,2	46,6	30,9	61,5	56,7	67,2	46,3	65,1	64,9	71,9	44,8	46,4	62,5	53,0	34,3	73,5	69,5	47,5	27,4	62,2
21	61,2	62,8	72,2	48,5	34,1	60,3	38,8	80,0	75,5	87,4	62,6	84,2	84,0	93,3	59,4	57,6	80,6	68,6	44,2	69,5	95,4	64,1	35,2	78,5
22	43,3	44,5	52,0	34,3	23,7	41,4	23,2	55,4	56,9	61,6	62,7	58,2	58,4	64,9	47,2	38,7	54,9	47,6	29,7	47,5	64,1	66,3	21,6	55,8
23	15,7	16,2	27,1	16,5	14,2	23,3	37,8	32,2	23,0	33,2	20,5	33,7	33,7	37,1	17,8	23,3	32,1	26,7	16,8	27,4	35,2	21,6	37,8	28,8
24	55,5	56,7	67,9	42,5	28,7	51,0	31,9	68,9	67,8	76,7	54,5	72,6	72,5	80,4	57,4	48,5	68,7	58,4	36,5	62,2	78,5	55,8	28,8	82,2

Tabela 51: Udział procentowy niepożądanych użytkowników obecnych w zbiorze, którzy zostali wykryci przez obie z pary cech.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	71,0	65,0	52,4	47,1	45,2	55,0	48,4	35,0	48,1	10,6	69,6	65,2	64,6	70,4	69,2	48,4	24,2	57,9	60,1	56,2	4,8	69,2	54,4	59,6
2	65,0	69,0	50,6	46,3	44,0	53,7	46,6	35,3	46,5	10,6	67,7	63,6	63,1	68,4	67,4	47,6	23,8	56,6	58,5	54,5	4,6	67,3	52,7	57,7
3	52,4	50,6	72,9	44,9	48,6	57,0	52,8	33,4	44,5	9,2	70,9	65,6	64,9	72,2	70,8	47,0	26,2	57,9	62,2	59,6	5,2	70,3	58,7	60,4
4	47,1	46,3	44,9	61,6	39,6	48,5	42,9	30,6	38,6	9,1	59,9	56,8	56,3	61,1	59,8	42,0	22,0	50,5	52,7	48,4	4,1	59,4	47,8	49,7
5	45,2	44,0	48,6	39,6	65,6	51,7	46,4	30,9	40,0	9,5	63,5	60,2	59,5	65,0	63,4	43,8	23,6	53,2	56,3	52,6	4,2	62,8	51,7	51,9
6	55,0	53,7	57,0	48,5	51,7	78,9	55,6	38,1	44,8	11,2	75,8	73,9	73,4	78,4	75,9	54,0	32,8	69,7	72,1	63,0	5,1	74,8	62,1	62,3
7	48,4	46,6	52,8	42,9	46,4	55,6	70,5	41,4	44,0	9,5	68,2	63,8	63,1	69,8	68,2	46,1	28,3	56,1	60,2	57,1	4,7	67,6	70,5	56,6
8	35,0	35,3	33,4	30,6	30,9	38,1	41,4	47,4	32,2	7,7	46,6	44,7	44,4	47,1	46,4	36,1	18,6	40,7	42,5	35,8	3,1	46,4	44,0	37,4
9	48,1	46,5	44,5	38,6	40,0	44,8	44,0	32,2	61,0	13,1	60,5	55,4	54,3	60,3	60,8	37,2	14,3	48,5	50,2	44,7	4,9	60,4	50,0	49,7
10	10,6	10,6	9,2	9,1	9,5	11,2	9,5	7,7	13,1	14,3	14,2	14,0	14,0	14,3	14,3	10,2	4,9	12,0	11,0	10,2	0,8	14,2	10,8	10,6
11	69,6	67,7	70,9	59,9	63,5	75,8	68,2	46,6	60,5	14,2	96,9	88,7	87,7	96,0	93,8	66,1	33,1	78,3	82,6	77,1	6,1	95,8	76,2	77,8
12	65,2	63,6	65,6	56,8	60,2	73,9	63,8	44,7	55,4	14,0	88,7	91,7	86,9	91,5	88,5	63,4	34,6	77,4	79,4	72,6	5,8	87,8	71,1	72,1
13	64,6	63,1	64,9	56,3	59,5	73,4	63,1	44,4	54,3	14,0	87,7	86,9	90,7	90,4	87,5	64,2	35,1	76,6	78,3	71,9	5,9	86,7	70,2	71,3
14	70,4	68,4	72,2	61,1	65,0	78,4	69,8	47,1	60,3	14,3	96,0	91,5	90,4	99,1	95,8	66,1	36,0	80,9	85,0	79,1	6,2	95,1	77,8	78,8
15	69,2	67,4	70,8	59,8	63,4	75,9	68,2	46,4	60,8	14,3	93,8	88,5	87,5	95,8	96,7	63,8	34,3	78,5	82,5	76,7	6,2	93,0	76,1	77,3
16	48,4	47,6	47,0	42,0	43,8	54,0	46,1	36,1	37,2	10,2	66,1	63,4	64,2	66,1	63,8	66,3	25,6	57,7	60,9	53,1	3,7	66,2	51,3	52,6
17	24,2	23,8	26,2	22,0	23,6	32,8	28,3	18,6	14,3	4,9	33,1	34,6	35,1	36,0	34,3	25,6	36,1	32,1	31,7	30,6	2,3	32,1	30,1	28,2
18	57,9	56,6	57,9	50,5	53,2	69,7	56,1	40,7	48,5	12,0	78,3	77,4	76,6	80,9	78,5	57,7	32,1	81,3	73,7	63,8	4,9	77,3	62,8	63,8
19	60,1	58,5	62,2	52,7	56,3	72,1	60,2	42,5	50,2	11,0	82,6	79,4	78,3	85,0	82,5	60,9	31,7	73,7	85,6	68,3	5,3	81,6	67,5	67,8
20	56,2	54,5	59,6	48,4	52,6	63,0	57,1	35,8	44,7	10,2	77,1	72,6	71,9	79,1	76,7	53,1	30,6	63,8	68,3	79,7	4,6	76,4	63,2	64,6
21	4,8	4,6	5,2	4,1	4,2	5,1	4,7	3,1	4,9	0,8	6,1	5,8	5,9	6,2	6,2	3,7	2,3	4,9	5,3	4,6	6,3	6,0	5,2	5,5
22	69,2	67,3	70,3	59,4	62,8	74,8	67,6	46,4	60,4	14,2	95,8	87,8	86,7	95,1	93,0	66,2	32,1	77,3	81,6	76,4	6,0	96,0	75,6	77,2
23	54,4	52,7	58,7	47,8	51,7	62,1	70,5	44,0	50,0	10,8	76,2	71,1	70,2	77,8	76,1	51,3	30,1	62,8	67,5	63,2	5,2	75,6	78,5	63,0
24	59,6	57,7	60,4	49,7	51,9	62,3	56,6	37,4	49,7	10,6	77,8	72,1	71,3	78,8	77,3	52,6	28,2	63,8	67,8	64,6	5,5	77,2	63,0	79,6

Bardzo dobrze wypadające cechy związane z liczbami obserwujących i obserwowanych, a zwłaszcza ich stosunek, pokrywają znaczną część zbioru. Widoczne to jest w obu tabelach, gdzie użytkownicy wykryci przez prawie wszystkie inne cechy są również wykryci przez wcześniej wymienione zarówno w przypadku normalnych i niepożądanych użytkowników. Analizując wyniki cech związanych z częstością pisania wiadomości: maksymalna seria wiadomości, maksymalna liczba wiadomości w czasie 15 minut oraz średnia pięciu najdłuższych serii można zauważyć, że wszystkie trzy podejścia mają swoje grupy użytkowników klasyfikowanych tylko przez siebie. W kwestii kategoryzacji źródeł publikacji poszczególnych wiadomości oraz ulubionych źródeł publikacji użytkowników zgodność jest bardzo wysoka. Zwłaszcza w przypadku wykrywania niepożądanych użytkowników, których obie metody wykrywają o wiele lepiej niż normalnych.

7.3. Porównanie zdolności klasyfikacyjnych podzbiorów opracowanych cech

Do zbadania wszystkich grup cech w podpunkcie wykorzystano algorytm lasów losowych z domyślnymi parametrami dostępnymi w bibliotece scikit-learn oraz z 10-krotną kroswalidacją. Zbiór pozytywny w przypadku metryk oznacza niepożądanych użytkowników.

7.3.1. Cechy pochodzące z różnych źródeł

Jest to porównanie wyników osiąganych przez grupy cech pochodzące z różnych źródeł ich pozyskiwania zaproponowanych w punkcie 4.1.

Tabela 52: Wyniki osiągane przez grupy cech pochodzące z różnych kategorii zaproponowanych w 4.1.

	Analiza treści	Analiza zachowań i statystyk	Analiza zależności
F1	0,924	0,984	0,997
Precyzja	0,921	0,982	0,996
Czułość	0,927	0,986	0,998
Dokładność	0,924	0,984	0,997
ROC AUC	0,975	0,997	1,000

Tabela 52 porównuje wyniki grup cech pochodzących z różnych źródeł. Cechy z grupy dotyczącej analizy zależności między użytkownikami osiągają najlepsze wyniki. Wynika to z tego, że do grupy zaliczają się cechy związane z liczbą obserwujących i obserwowanych, które jak zostało pokazane w tabeli 49 osiągają indywidualnie bardzo dobre wyniki.

Najgorzej wypadły cechy dotyczące analizy treści pisanych wiadomości. Proste cechy związane z tekstem nie pozwalają na osiągnięcie wysokich wyników, a te bardziej zaawansowane obarczone są pewnym błędem i często są zawodne w trudniejszych przypadkach. Przykładem może być analiza emocji, sentymentu lub subiektywności.

7.3.2. Porównanie wyników emocji, sentymentu, subiektywności

Zestawione są ze sobą różne warianty wyboru cech analizy emocji, która składa się z maksymalnie 11 cech. Sentyment i subiektywność wyrażane są przez pojedyncze cechy.

Tabela 53: Wyniki osiągane przez cechy badające emocje, sentyment i subiektywność tekstu.

	F1	Precyzja	Czułość	Dokładność	ROC AUC
wszystkie cechy dotyczące analizy emocji	0,81	0,82	0,80	0,81	0,90
5 najbardziej wyróżniających emocji i suma ich udziałów	0,81	0,81	0,80	0,81	0,89
5 najbardziej wyróżniających emocji	0,79	0,80	0,78	0,79	0,87
suma udziału 5, emocja nr 1 i jej udział	0,77	0,79	0,76	0,78	0,87
sentyment i subiektywność	0,75	0,76	0,74	0,75	0,82
sentyment	0,74	0,75	0,73	0,74	0,80
subiektywność	0,58	0,55	0,62	0,56	0,58

W wynikach prezentowanych w tabeli 53 widać, że analiza emocji w każdym badanym wariancie przynosi lepsze rezultaty niż sentyment i subiektywność. Może to być spowodowane tym, że biblioteka wykorzystana w analizie emocji była uczona z wykorzystaniem wiadomości z Twittera. Dzięki temu może być bardziej przystosowana do charakterystycznych cech tweetów. Subiektywność osiąga znacznie gorsze wyniki niż sentyment.

7.3.3. Analizy częstości pisania wiadomości

Tabela 54: Wyniki osiągane przez cechy dotyczące częstości pisania wiadomości.

	F1	Precyzja	Czułość	Dokładność	ROC AUC
wszystkie poniższe cechy	0,94	0,93	0,96	0,94	0,99
maks. seria i maks. liczba wiad. w oknie	0,82	0,78	0,85	0,81	0,90
średnia 5 serii i maks. liczba wiad. w oknie	0,87	0,84	0,90	0,86	0,95
średnia 5 serii wiadomości	0,78	0,74	0,81	0,76	0,85
maks. długość serii wiadomości	0,73	0,68	0,79	0,71	0,79
maks. liczba wiadomości w oknie	0,71	0,61	0,86	0,66	0,71

Połączenie wszystkich dostępnych cech analizujących częstość pisania wiadomości daje wyniki wszystkich metryk przekraczające 0,9, co pokazane jest w tabeli 54. Indywidualnie najlepsza jest cecha uwzględniająca średnią 5 najdłuższych serii. Porównując wyniki analizy pojedynczej serii i prostej analizy liczby wiadomości w oknie czasowym widać małą przewagę bardziej zaawansowanej analizy serii. Prostsza analiza wygrywa w kwestii czułości, lecz robi to kosztem wielu błędnych klasyfikacji.

7.3.4. Analizy źródeł publikacji

Tabela 55: Wyniki osiągane przez cechy dotyczące kategorii publikacji wiadomości.

Cechy	F1	Precyzja	Czułość	Dokładność	ROC AUC
kategoria wiadomości i ulubiona użytkownika	0,841	0,751	0,956	0,820	0,859
kategoria źródła pub. wiadomości	0,834	0,733	0,969	0,808	0,844
ulubiona kategoria źródła pub. użytkownika	0,835	0,740	0,959	0,811	0,846

Analizy źródeł publikacji pokazywały wyraźne różnice pomiędzy normalnymi i niepożądanymi użytkownikami. Przekłada się to na osiągane wyniki klasyfikacji zaprezentowane w tabeli 55. Widać jednak, że uwzględnianie wyłącznie źródła publikacji badanej wiadomości nie ustępuje określaniu ulubionej kategorii publikacji dla autora uwzględniając jego wszystkie wiadomości. Pozwalało to na eliminację jednorazowych publikacji z mniej preferowanych przez użytkownika źródeł. Połączenie obu cech daje nieznacznie lepsze wyniki. Są to bardzo skorelowane cechy.

7.3.5. Obserwowani i obserwujący

Tabela 56: Wyniki osiągane z wykorzystaniem liczb obserwujących i obserwowanych.

Cechy	F1	Precyzja	Czułość	Dokładność	ROC AUC
liczba obserwujących, obserwowanych, stosunek	0,997	0,995	0,998	0,996	0,999
liczba obserwujących i obserwowanych	0,996	0,994	0,998	0,996	0,999
stosunek obserwujących i obserwowanych	0,985	0,978	0,991	0,985	0,998
liczba obserwujących	0,902	0,888	0,917	0,901	0,972
liczba obserwowanych	0,896	0,884	0,908	0,894	0,968

Liczby obserwujących, obserwowanych oraz ich stosunek dają indywidualnie najlepsze wyniki spośród wszystkich badanych cech. Praktycznie wszystkie zestawy w tabeli 56

przekraczają lub są bardzo bliskie wartości 0,9 miary F1. W przypadku połączenia liczby obserwowanych i obserwujących osiągane wyniki są lepsze niż w przypadku wybierania wyłącznie ich stosunku. Widać, że pewne informacje, które niosą te cechy są tracone przez obliczenie stosunku. Tak dobre wyniki tych cech mogą być konsekwencją wyraźnie mniejszej liczby kont niepożądanych użytkowników. Stworzony zbiór testowo-treningowy jest zbalansowany pod względem liczby wiadomości, a nie użytkowników.

7.3.6. Analizy cech dotyczących retweetów

Wyniki zostaną przedstawione dla dwóch zbiorów:

- zwykłego zbioru, na którym były badane wcześniejsze wiadomości,
- zbioru zawierającego wyłączenie retweetowane wiadomości, co pozwoli przetestować cechy na danych, do których są przeznaczone.

Zbiór retweetów został stworzony w dokładnie taki sam sposób jak normalny zbiór. Zostało to opisane w punkcie 7.1.

Zbiór zwykły

Tabela 57: Wyniki osiągane przez cechy opisujące retweetowane wiadomości na normalnym zbiorze.

Cecha	F1	Precyzja	Czułość	Dokładność	ROC AUC
stosunek ret. użytk., liczba obec. w maks. klikach, rozmiar maks. kliki	0,81	0,85	0,78	0,82	0,90
stosunek ret. użytk., rozmiar maks. kliki	0,79	0,81	0,77	0,80	0,88
stosunek ret. użytk., liczba obecności w maks. Klik	0,82	0,72	0,95	0,79	0,85
liczba maks. klik, rozmiar maks. kliki	0,72	0,69	0,76	0,71	0,77
stosunek obserwujących i obserwowanych retweetowanych	0,82	0,71	0,97	0,79	0,83
rozmiar maksymalnej kliki dla użytkownika	0,64	0,63	0,66	0,63	0,67
liczba obecności w maksymalnych klikach	0,48	0,70	0,36	0,60	0,66

W tabeli 57 widać, że najlepszy wynik miary F1 uzyskiwany jest przez stosunek liczby obserwujących i obserwowanych retweetowanych użytkowników. Cecha ma bardzo dużą czułość, lecz jej precyzja jest relatywnie niska. Podobnie jest w przypadku połączenia tej cechy z najgorszą z grupy liczbą obecności w maksymalnych klikach. Indywidualnie stosunek wypada najlepiej. Zwracając uwagę na wyniki grup widać, że połączenie wszystkich cech ma wyraźnie niższą czułość niż stosunek, nieznacznie mniejszy wynik F1, lecz pozostałe kryteria oceny, w tym wyraźnie precyzja, osiągają lepsze rezultaty.

Zbiór retweetów

Tabela 58: Wyniki osiągane przez cechy opisujące retweetowane wiadomości na zbiorze składającym się wyłącznie z wiadomości retweetowanych.

Cecha	F1	Precyzja	Czułość	Dokładność	ROC AUC
stosunek ret. użytk., liczba obec. w maks. klikach, rozmiar maks. kliki	0,92	0,92	0,91	0,92	0,95
stosunek ret. użytk., rozmiar maks. kliki	0,90	0,90	0,90	0,90	0,94
stosunek ret. użytk., liczba obecności w maks. klik	0,88	0,89	0,88	0,88	0,93
liczba maks. klik, rozmiar maks. kliki	0,81	0,70	0,95	0,77	0,87
stosunek obserwujących i obserwowanych retweetowanych	0,91	0,91	0,91	0,91	0,95
rozmiar maksymalnej kliki dla użytkownika	0,75	0,62	0,95	0,68	0,76
liczba obecności w maksymalnych klikach	0,66	0,75	0,59	0,70	0,76

Jak można się było spodziewać wyniki pokazane w tabeli 58 we wszystkich przypadkach są o wiele lepsze. Różnice między precyzją i czułością są znacznie zniwelowane. Optymalnym wyborem jest stosunek obserwujących i obserwowanych, ponieważ osiąga praktycznie takie same wyniki jak zestaw wszystkich trzech cech.

7.3.7. Analiza prostych liczb dotyczących tekstu wiadomości

Tabela 59: Wyniki osiągane przez proste cechy liczbowe z tekstu wiadomości.

Cechy	F1	Precyzja	Czułość	Dokładność	ROC AUC
liczby: znaków, słów, linków, hasztagów, odniesień do użytk.	0,78	0,82	0,74	0,79	0,87
liczby: znaków, linków, hasztagów, odniesień do użytk.	0,76	0,82	0,72	0,78	0,85
liczby: słów, linków, hasztagów, odniesień do użytkowników	0,75	0,81	0,71	0,77	0,84
liczby: linków, hasztagów, odniesień do użytkowników	0,69	0,82	0,59	0,73	0,79
liczby: hasztagów, odniesień do użytkowników	0,72	0,72	0,72	0,72	0,76

Wyniki z powyższej tabeli 59 pokazują, że nawet prosta analiza tekstu oparta o trywialne cechy może przynieść całkiem dobre wyniki. W kwestii nakładu pracy do osiąganych wyników wymienione cechy bardzo się wyróżniają. Trzy zestawy osiągnęły wartość miary F1 równą lub większą 0,75.

7.3.8. Analizy podobieństwa wiadomości

Tabela 60: Wyniki osiągane przez cechy zajmujące się podobieństwem tekstu wiadomości.

Cechy	F1	Precyzja	Czułość	Dokładność	ROC AUC
średnie podobieństwo i liczba podobnych (próg 0,7)	0,775	0,754	0,797	0,768	0,822
średnie podobieństwo wiadomości	0,773	0,751	0,797	0,766	0,818
liczba podobnych wiadomości (próg 0,7)	0,113	0,577	0,063	0,508	0,513

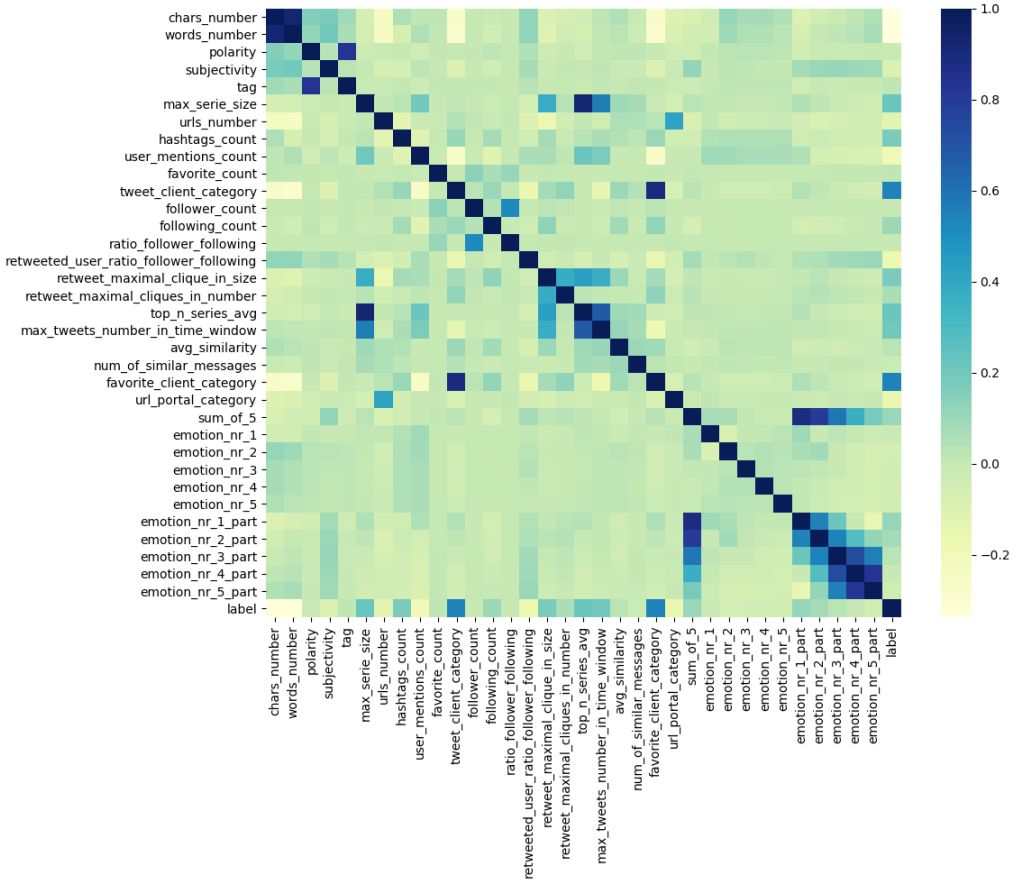
Analizy podobieństwa przynoszą bardzo skrajne wyniki widoczne w tabeli60. Jedna z nich osiąga całkiem przyzwoite miary, a z kolei druga jest pod względem indywidualnym najgorszą ze wszystkich stworzonych cech, co zostało pokazane w tabeli 49. Jej czułość jest prawie zerowa. Wynika to zapewne z małych różnic pomiędzy zbiorami oraz braku w nich bardzo oczywistych spamerów. Klasyfikacja w oparciu o obie cechy jednocześnie praktycznie nie poprawia wyników względem wyboru jedynie średniego podobieństwa wiadomości.

7.4. Wybór podzbioru cech

W podrozdziale na różne sposoby zostaną wybrane możliwie najlepsze zestawy cech. Wykorzystana zostanie korelacja Pearsona, metoda Lasso oraz zachłanna metoda eliminacji cech w kolejnych krokach. Na samym końcu wyniki stworzonych zbiorów zostaną porównane.

7.4.1. Korelacja Pearsona

Obliczenie korelacji pozwoli na odrzucenie cech, które są ze sobą ze sobą za bardzo powiązane statystycznie. Eliminacja takich cech nie dość, że zmniejsza nakład obliczeń to w dodatku może pozwolić na podwyższenie wyników klasyfikacji.



Rysunek 41: Wizualizacja wyników obliczonej korelacji Pearsona dla wszystkich cech w zbiorze oraz etykiety/kategorii. Wysokie wartości dla cech związanych z analizą emocji, długości wiadomości oraz częstości pisania wiadomości.

Na podstawie wizualizacji przedstawionej na rysunku 41 można wyróżnić wysokie korelacje zachodzące pomiędzy liczbą słów i znaków, analizami częstości pisania wiadomości (serie wiadomości, średnia pięciu serii i wiadomości w oknie czasowym 15 minut), kategorią źródła publikacji wiadomości oraz ulubioną kategorią użytkownika, cechami związanymi z analizą emocji. Są one oczywiste, ponieważ wyróżnione cechy są do siebie bardzo zbliżone. Mniejsze, lecz również wysokie korelacje zachodzą między współczynnikiem ob-

serwowanych i obserwujących, a samą liczbą obserwujących oraz w przypadku analiz grafu retweetów oraz częstości pisania wiadomości.

Wszystkie pary cech, dla których korelacja przekroczyła wartość 0,5 wyróżnione są w tabeli 61. Podane są w niej również korelacje cech z par z kategorią przypisywaną do danych (normalny lub niepożądany użytkownik).

Tabela 61: Pary cech, których korelacja Pearsona przekracza wartość 0,5. Dwie ostatnie kolumny odnoszą się do korelacji cech z kategorią przypisywaną do danych.

Cecha nr 1	Cecha nr 2	Korelacja cech (abs)	Korel. nr 1 z kategorią	Korel. nr 2 z kategorią
chars_number	words_number	0,94	-0,33	-0,34
max_serie_size	top_n_series_avg	0,93	0,22	0,22
tweet_client_category	favorite_client_category	0,89	0,55	0,54
sum_of_5	emotion_nr_1_part	0,88	0,10	0,11
polarity	tag	0,83	-0,01	0,01
emotion_nr_4_part	emotion_nr_5_part	0,83	-0,01	-0,04
sum_of_5	emotion_nr_2_part	0,80	0,10	0,08
emotion_nr_3_part	emotion_nr_4_part	0,73	0,03	-0,01
top_n_series_avg	max_tweets_number_in_time_window	0,68	0,22	0,20
sum_of_5	emotion_nr_3_part	0,58	0,10	0,03
max_serie_size	max_tweets_number_in_time_window	0,57	0,22	0,20
emotion_nr_3_part	emotion_nr_5_part	0,55	0,03	-0,04
emotion_nr_1_part	emotion_nr_2_part	0,55	0,11	0,08
emotion_nr_2_part	emotion_nr_3_part	0,54	0,08	0,03
follower_count	ratio_follower_following	0,52	0,00	-0,01

Tabela 61 posłużyła do eliminacji cech. Cechą odrzucaną z pary w większości przypadków jest ta, której wartość bezwzględna korelacji z kategorią jest mniejsza. Wartości są jednak najczęściej bardzo do siebie zbliżone i mało pomocne. Przydatne są również wcześniejsze tabele 50 i 51 pokazujące % wspólnych wyborów par cech w kwestii klasyfikacji obu typów użytkowników.

Do cech odrzuconych na podstawie wyników obliczonych korelacji należą:

- liczba słów,
- maks. długość serii wiadomości,
- maks. liczba wiadomości w czasie 15 minut,
- suma udziałów 5 najbardziej wyróżniających się emocji,
- kategoria źródła publikacji,
- tag sentymentu,
- liczba obserwujących.
- cechy dotyczące udziałów poszczególnych emocji: udział emocji nr 1, udział emocji nr 2, udział emocji nr 3, udział emocji nr 4 i udział emocji nr 5.

Zdolności klasyfikacyjne zbioru po usunięciu wymienionych cech zostaną przebadane i zestawione z innymi grupami w dalszym podpunkcie.

7.4.2. Metoda Lasso

Lasso [25] (*Least Absolute Shrinkage and Selection Operator*) jest metodą pozwalającą na regularyzację oraz dobór cech. Nakłada ona ograniczenie na sumę wartości bezwzględnych parametrów modelu, która musi być nie większa niż ustalona granica. Metoda przeprowadza proces regularyzacji, w którym karze współczynniki zmiennych regresji poprzez zmniejszanie ich do zera. Dobór cech polega w tej sytuacji na wybraniu zmiennych, których współczynniki mają niezerowane wartości.

Cechy zostały poddane analizie metodą Lasso kilkukrotnie. Wejściem do kolejnego kroku jest grupa cech odrzuconych w poprzednim. Kolejność cech w tabelach nie jest znacząca. Jedyną ważną informacją jest to czy współczynnik cechy w analizie jest różny od zera. W takim przypadku cecha jest wybierana.

Krok pierwszy - wszystkie cechy

Tabela 62: Cechy wybrane i odrzucone metodą Lasso w pierwszym kroku. Zostały wybrane jedynie 4 cechy.

	Wybrane	Odrzucone
1	liczba obserwowanych	liczba znaków
2	liczba obecności w maksymalnych klikach	liczba słów
3	stosunek obserwujących i obserwowanych retweetowanych	sentyment
4	liczba obserwujących	subiektywność
5		tag sentymentu
6		maks. długość serii wiadomości
7		liczba linków
8		liczba hasztagów
9		liczba odniesień do użytkowników
10		liczba polubień
11		kategoria źródła publikacji
12		stosunek obserwujących i obserwowanych
13		rozmiar maksymalnej kliki w grafie retweetów
14		średnia 5 serii wiadomości
15		maks. liczba wiadomości w czasie 15 minut
16		średnie podobieństwo wiadomości
17		liczba podobnych wiadomości (próg 0,7)
18		ulubione źródło publikacji użytkownika
19		kategoria portalu z odnośnika
20		suma udziałów 5 najbardziej wyróżniających się emocji
21		emocja nr 1
22		emocja nr 2
23		emocja nr 3
24		emocja nr 4
25		emocja nr 5
26		udział emocji nr 1
27		udział emocji nr 2
28		udział emocji nr 3
29		udział emocji nr 4
30		udział emocji nr 5

Z puli wszystkich dostępnych cech zostały wybrane wyłącznie cztery widoczne w tabeli 62. Dwie z nich to cechy, które osiągały jedne z najlepszych wyników indywidualnych ze wszystkich osiągając wynik miary F1 w okolicach 0,9. Dopełniają je dwie cechy związane

z retweetami wiadomości. Tak jak w tabeli 56 stosunek obserwujących i obserwowanych przegrał z wyborem obu cech. Wyniki kolejny raz pokazują, że badane zbiory mogą zostać dobrze rozróżnione wyłącznie z pomocą prostych parametrów posiadanych przez wszystkie konta na Twitterze.

Skuteczność klasyfikacji zbiorów stworzonych w kolejnych krokach analizy zostanie przedstawiona w kolejnym podpunkcie.

Krok drugi

Tabela 63: Cechy wybrane i odrzucone metodą Lasso w drugim kroku. Wejściem analizy są cechy odrzucone w poprzednim.

	Wybrane	Odrzucone
1	ulubione źródło publikacji użytkownika	liczba słów
2	kategoria źródła publikacji	sentyment
3	liczba linków	subiektywność
4	liczba odniesień do użytkowników	tag sentymentu
5	liczba hasztagów	rozmiar maksymalnej kliki w grafie retweetów
6	kategoria portalu z odnośnika	średnie podobieństwo wiadomości
7	maks. liczba wiadomości w czasie 15 minut	suma udziałów 5 najbardziej wyróżniających się emocji
8	liczba znaków	udział emocji nr 1
9	maks. długość serii wiadomości	udział emocji nr 2
10	liczba podobnych wiadomości (próg 0,7)	udział emocji nr 3
11	średnia 5 serii wiadomości	udział emocji nr 4
12	emocja nr 4	udział emocji nr 5
13	emocja nr 5	
14	emocja nr 2	
15	emocja nr 3	
16	emocja nr 1	
17	liczba polubień	
18	stosunek obserwujących i obserwowanych	

Drugi krok, w którym analizowane są cechy odrzucone w poprzednim prezentuje wyniki o całkowicie innych proporcjach. Teraz została wybrana wyraźna większość dostępnych cech pokazanych w tabeli 63. Podkreśla to duże znaczenie grupy wybranej w kroku pierwszym.

Odrzucone zostały subiektywność i sentyment, a wybrane cechy dotyczące analizy emocji. Indywidualnie wyniki były takie same. Widać również, że liczba znaków dostarcza więcej informacji od liczby słów. Zostały wybrane wszystkie analizy dotyczące częstości pisania wiadomości, które były ze sobą mocno skorelowane. Wiele odrzuconych cech w tym kroku jest jednak zgodne z tymi, które zostały usunięte na podstawie analizy korelacji.

Krok trzeci

Tabela 64: Cechy wybrane i odrzucone metodą Lasso w trzecim kroku. Wejściem analizy są cechy odrzucone w poprzednim.

	Wybrane	Odrzucone
1	suma udziałów 5 najbardziej wyróżniających się emocji	udział emocji nr 2
2	subiektywność	udział emocji nr 3
3	udział emocji nr 1	udział emocji nr 4
4	rozmiar maksymalnej kliki w grafie retweetów	udział emocji nr 5
5	tag sentymentu	
6	liczba słów	
7	średnie podobieństwo wiadomości	
8	sentyment	

Cechy poddawane analizie w trzecim i ostatnim kroku pokazane w tabeli 64 można traktować jako grupę najmniej wnoszących. Obecność w nich sentymentu może dziwić, lecz mógł on zostać wyparty przez analizę emocji. Z 11 cech związanych z emocjami 6 znalazło się w ostatnim kroku, a 4 z nich zostały odrzucone. Udziały poszczególnych emocji są nadmiarowe, co potwierdza już wcześniejsze obserwacje.

7.4.3. Zbiór stworzony eliminacją kolejnych cech

Zaczynamy od zbioru wszystkich wybranych do analizy cech przedstawionych w tabeli 65. W każdym kroku analizy tworzymy dokładnie tyle zbiorów ile mamy cech, różniących się tym, że z każdego z nich została usunięta inna cecha. Następnie dla każdego zbioru trenujemy i badamy klasyfikator. Jest to metoda zachłanna. Ze zbioru analizowanych cech zostaje wyeliminowana ta, której usunięcie najbardziej poprawia wynik miary F1. Kroki są powtarzane, aż do chwili, gdy dalsza eliminacja cech nie przynosi poprawy wyników.

Tabela 65: Zbiór cech wybrany do analizy zachłanną metodą eliminacji

	Cecha
1	liczba znaków
2	sentyment
3	subiektywność
4	maks. długość serii wiadomości
5	liczba linków
6	liczba hasztagów
7	liczba odniesień do użytkowników
8	liczba polubień
9	kategoria źródła publikacji
10	liczba obserwujących
11	liczba obserwowanych
12	stosunek obserwujących i obserwowanych
13	stosunek obserwujących i obserwowanych retweetowanych
14	rozmiar maksymalnej kliki w grafie retweetów
15	liczba obecności w maksymalnych klikach
16	średnia 5 serii wiadomości
17	maks. liczba wiadomości w czasie 15 minut
18	średnie podobieństwo wiadomości
19	liczba podobnych wiadomości (próg 0,7)
20	ulubione źródło publikacji użytkownika
21	kategoria portalu z odnośnika
22	emocja nr 1
23	emocja nr 2
24	emocja nr 3
25	emocja nr 4
26	emocja nr 5

Zbiór cech będących wejściem do analizy jest bardzo podobny do zbioru stworzonego w oparciu o korelację Pearsona. Różni się jedynie tym, że dodano do niego wcześniej usunięte trzy cechy: liczbę obserwujących, maksymalną liczba wiadomości w czasie 15 minut oraz źródło publikacji pojedynczych wiadomości. Zbiór początkowy liczy 26 cech.

Kolejność eliminacji cech

Tabela 66: Kolejność odrzucania cech zachłanną metodą eliminacji.

Kolejność eliminacji	Usunięta cecha
1	sentyment
2	emocja nr 1
3	liczba znaków
4	emocja nr 3
5	emocja nr 2
6	subiektywność
7	emocja nr 5
8	średnie podobieństwo wiadomości
9	stosunek obserwujących i obserwowanych retweetowanych
10	emocja nr 4
11	liczba hasztagów
12	liczba odniesień do użytkowników
13	liczba linków
14	kategoria portalu z odnośnika
15	kategoria źródła publikacji
16	liczba polubień
17	liczba podobnych wiadomości (próg 0,7)
18	liczba obecności w maksymalnych klikach

W tabeli 66 przedstawione są cechy eliminowane w kolejnych krokach metody. Bardzo szybko zostały usunięte cechy związane z analizą sentymentu oraz emocji. Wyeliminowane zostały też obie cechy związane z analizą podobieństwa wiadomości. Odrzucono również wszystkie proste cechy dotyczące prostego zliczania linków, hasztagów, odniesień do użytkowników oraz polubień. Ostatnią usuniętą cechą jest jedną z dwóch dotyczących analizy grafów.

Zbiór otrzymany po zakończeniu eliminacji

Tabela 67: Zbiór cech otrzymanych po analizie zachłanną metodą eliminacji

	Cecha
1	maks. długość serii wiadomości
2	liczba obserwujących
3	stosunek obserwujących i obserwowanych
4	liczba obserwowanych
5	maks. liczba wiadomości w czasie 15 minut
6	rozmiar maksymalnej kliki w grafie retweetów
7	średnia 5 serii wiadomości
8	ulubione źródło publikacji użytkownika

Wykonano 18 kroków, z których w każdym została odrzucona jedna z cech, której eliminacja najbardziej poprawiała wynik. Z otrzymanej grupy 8 cech usunięcie jakiejkolwiek z nich nie przynosi już dalszego efektu. Grupa zawiera sporo cech o dużej korelacji. Analiza była jednak nastawiona na zachłanne osiągnięcie maksymalnego możliwego wyniku. Podane w tabeli 67 cechy pozwalają na osiągniecie wyniku miary F1 równego 0,99965.

7.4.4. Zbiory stworzone z najgorszych cech

Zbióry składa się z cech, które w analizie indywidualnych zdolności znajdującej się w punkcie 7.2 osiągnęły najgorsze wyniki. Zbiór pozwoli na sprawdzenie przydatności takich cech w grupie.

Cechy z wynikiem miary F1 poniżej 0,7

W puli wszystkich cech znajduje się 12, których wartość miary F1 w indywidualnej klasyfikacji nie przekroczyła 0,7. Zaliczają się do nich:

- liczba znaków,
- liczba słów,
- subiektywność,
- tag sentymentu,
- liczba linków,
- liczba hasztagów,
- liczba odniesień do użytkowników,

- liczba polubień wiadomości,
- rozmiar maksymalnej kliki w grafie retweetów,
- liczba obecności w maksymalnych klikach,
- liczba podobnych wiadomości (próg 0,7),
- kategoria portalu z odnośnika.

Zbiór 12 cech został poddany eliminacji opisanej w punkcie 7.4.3. Nie przyniosło to efektów. Usunięcie każdej z cech pogarsza wynik.

Cechy z wynikiem miary F1 poniżej 0,6
Do grupy cech, którym idywidualnie nie udało się przekroczyć wyniku F1 równego 0,6 należą:

- subiektywność,
- tag sentymentu,
- liczba hasztagów,
- liczba polubień wiadomości,
- liczba obecności w maksymalnych klikach,
- liczba podobnych wiadomości (próg 0,7).

Zbiór 6 cech został poddany eliminacji opisanej w punkcie 7.4.3. Tak samo jak w przypadku wcześniejszej grupy nie doprowadziło to do wyeliminowania żadnej z cech.

7.4.5. Finalne zbiory cech

Do zbadania wszystkich zestawów cech wykorzystano algorytm lasów losowych z domyślnymi parametrami dostępnymi w bibliotece scikit-learn oraz z 10-krotną kros-walidacją. Zbiór pozytywny w przypadku metryk oznacza niepożądanych użytkowników.

Tabela 68: Porównanie wyników osiąganych przez różne zbiory danych

Zestaw cech	F1	Precyzja	Czułość	Dokładność
wszystkie dostępne cechy	0,99579	0,99755	0,99404	0,99580
bez obserwujących, obserwowanych oraz ich stosunku	0,99216	0,99326	0,99106	0,99217
po usunięciu skorelowanych, Pearson (7.4.1)	0,99472	0,99646	0,99299	0,99473
Pearson + obserwujący	0,99584	0,99746	0,99423	0,99585
krok 1, metoda Lasso (7.4.2)	0,99508	0,99507	0,99509	0,99508
krok 2, metoda Lasso (7.4.2)	0,98569	0,98624	0,98513	0,98569
krok 1 i 2, metoda Lasso	0,99548	0,99736	0,99360	0,99549
cechy przed metoda eliminacji (7.4.3)	0,99688	0,99827	0,99550	0,99689
cechy wybrane metodą eliminacji (7.4.3)	0,99964	0,99981	0,99948	0,99964
najgorsze 12 cech (7.4.4)	0,88268	0,89743	0,86841	0,88457
najgorsze 6 cech (7.4.4)	0,68627	0,74906	0,63321	0,71053

W tabeli 68 zostały zestawione wyniki wszystkich stworzonych zbiorów danych. Klasyfikatory zostały zbudowane z wykorzystaniem algorytmu lasów losowych.

Wszystkie dostępne cechy osiągają wynik miary F1 równy 0,9958. Gdy usuniemy cechy dotyczące obserwujących i obserwowanych, które wypadają najlepiej indywidualnie miara F1 nadal przekracza 0,99.

Po usunięciu najbardziej skorelowanych cech (korelacja Pearsona powyżej 0,5) wyniki są nieznacznie gorsze niż dla zbioru początkowego. Po dodaniu liczby obserwujących, która została wcześniej usunięta w oparciu o jej korelację wyniki wyraźnie się zwiększają i są nieznacznie większe niż dla zbioru wszystkich cech. W przypadku zbioru wybranego w 1 kroku metodą Lasso widać, że za pomocą wyłącznie 4 cech można osiągnąć prawie taki sam wynik. O wiele bardziej liczne cechy z 2 kroku osiągają wyraźnie gorszy wynik. Połączenie cech wybranych w obu krokach nie ma dużego wpływu na wyniki.

Zestaw cech wybranych do metody eliminacji osiągnął wynik równy 0,997. Zachłannie eliminując 18 cech, które w danym momencie dawały najlepszą poprawę wyników został osiągnięty wynik F1 równy ponad 0,999 przez zestaw składający się z 8 cech.

Dwa najgorsze zbiory zawierają odpowiednio 12 i 6 cech. Pierwszy z nich osiągnął wynik bardzo zbliżający się do 0,9, co można uznać za sukces, uwzględniając indywidualne wyniki klasyfikacji zawierających się w nim elementów. Drugi ze zbiorów wypadł wyraźnie gorzej. Nie dość, że zawierające się w nim cechy były gorsze, to w dodatku dużą rolę odegrała ich o wiele mniejsza liczba.

7.5. Porównanie wyników różnych klasyfikatorów

Użyte algorytmy klasyfikacji zostały krótko opisane w punkcie 2.5. Klasyfikatory porównywane są na dwóch zbiorach cech. Pierwszy z nich 7.4.3 składa się z 8 cech, a drugi 7.4.1 rozszerzony o liczbę obserwowanych liczy 23 cechy. Wszystkie algorytmy działają z wykorzystaniem domyślnych parametrów ustawionych w bibliotece scikit-learn. Zbiór pozytywny w przypadku metryk oznacza niepożądanych użytkowników.

Tabela 69: Porównanie wyników różnych klasyfikatorów na zbiorze otrzymanym w punkcie 7.4.3

Klasyfikator	F1	Precyzja	Czułość	Dokładność	ROC AUC
Lasy losowe	0,99964	0,99981	0,99948	0,99964	0,99987
CART	0,99927	0,99897	0,99957	0,99927	0,99931
Extra Tree	0,99926	0,99894	0,99958	0,99926	0,99929
K-najbliższych sąsiad. (k=3)	0,99844	0,99795	0,99893	0,99844	0,99936
K-najbliższych sąsiad. (k=5)	0,99774	0,99687	0,99861	0,99774	0,99940
K-najbliższych sąsiad. (k=7)	0,99703	0,99581	0,99826	0,99703	0,99943
AdaBoost	0,91554	0,87310	0,96232	0,91122	0,96979
MLP	0,83575	0,85045	0,83616	0,84195	0,90850
Regresja logistyczna	0,83568	0,78492	0,89907	0,82245	0,87084
Bernoulli Naive Bayes	0,66810	0,50384	0,99126	0,50755	0,50755
Gaussian Naive Bayes	0,66287	0,50088	0,97973	0,50173	0,64606
Multinomial Naive Bayes	0,61497	0,48382	0,84368	0,47178	0,48483

Analizując wyniki osiągnięte na mniejszym ze zbiorów przedstawione w tabeli 69, najlepszy wynik klasyfikacji uzyskiwany jest przez algorytm lasów losowych. Podobnie jak w wielu innych pracach przynosi on bardzo dobre rezultaty. Nie można powiedzieć tego samego o również polecanym naiwnym algorytmie bayesowskim, w różnych wersjach. Osiąga on kompletnie odmienne, niesatysfakcjonujące wyniki. Widoczne jest, że wszystkie 3 najlepsze klasyfikatory oparte są o drzewa decyzyjne. Dziwić może wynik algorytmu CART, który wypada bardzo dobrze w porównaniu do pozostałych dwóch, które w przeciwieństwie do niego budują wiele drzew decyzyjnych. Wyniki powyżej 0,9 miary F1 uzyskał również algorytm K-najbliższych sąsiadów w różnych konfiguracjach. Można zaobserwować, że zwiększanie liczby analizowanych sąsiadów negatywnie wpływało na osiągane wyniki.

Tabela 70: Porównanie wyników różnych klasyfikatorów na zbiorze otrzymanym w punkcie 7.4.1 uzupełnionym o wcześniej usuniętą liczbę obserwujących

Klasyfikator	F1	Precyzja	Czułość	Dokładność	ROC AUC
Lasy losowe	0,99568	0,99721	0,99416	0,99569	0,99967
CART	0,99353	0,99277	0,99428	0,99352	0,99352
AdaBoost	0,96594	0,95763	0,97439	0,96564	0,99418
K-najbliższych sąsiad. (k=3)	0,96155	0,94986	0,97354	0,96108	0,98261
Extra Tree	0,95682	0,95363	0,96004	0,95668	0,95668
MLP	0,88675	0,90429	0,87190	0,88898	0,94896
Bernoulli Naive Bayes	0,71270	0,73593	0,69090	0,72149	0,79940
Regresja logistyczna	0,70247	0,71747	0,68816	0,70857	0,77242
Gaussian Naive Bayes	0,67665	0,52030	0,96733	0,53773	0,71667

W tabeli 70 przedstawiono porównanie wyników klasyfikatorów działających z wykorzystaniem większego zbioru cech. Z racji większej liczby cech część klasyfikacji obecnych w poprzedniej tabeli nie została przeprowadzona. Wiąże się to ze znaczną komplikacją obliczeń, a co za tym idzie czasem analizy. Nastąpiły spore zmiany w przypadku dwóch dobrze wypadających klasyfikatorów. Wyniki algorytmu CART oraz K-najbliższych sąsiadów mocno się obniżyły. Najwięcej po zwiększeniu zestawu cech zyskał algorytm AdaBoost oraz MLP (sieć neuronowa).

7.6. Analiza użytkowników klasyfikowanych do złych grup

W przypadku klasyfikowania algorytmem lasów losowych z wykorzystaniem zbioru 8 cech 7.4.3 osiągającego najlepsze wyniki, można wyróżnić dwa typy użytkowników normalnych klasyfikowanych niepoprawnie. Pierwszy typ to użytkownicy, których ulubionym źródłem publikacji są oficjalne metody dostępne na komputerach, a serie ich wiadomości są krótkie. Drugi typ użytkowników również pisze z wykorzystaniem komputerów, ma często więcej obserwujących niż obserwowanych oraz serie wiadomości o rozmiarach rzędu kilkudziesięciu. Są to bardzo często mniej popularne konta informacyjne. Praktycznie wszyscy normalni użytkownicy, którzy zostali sklasyfikowani jako niepożądani publikowali ze źródeł komputerowych. Zwrócili przez to na siebie większą uwagę. W przypadku użytkowników niepożądanych pomyłki występują przy użytkownikach piszących bardzo mało, co za tym idzie z małymi rozmiarami serii wiadomości, często publikujący z innych źródeł niż oficjalne dostępne na komputerach. Mylący są zwłaszcza sporadycznie występujący trolle piszący z urządzeń mobilnych.

7.7. Wnioski z przeprowadzonych eksperymentów

Wszystkie grupy z zaproponowanego podziału ze względu na źródło pochodzenia osiągnęły zadowalające wyniki przekraczające wartość 0,9 miary F1. Najlepiej wypadła analiza zależności między użytkownikami (0,997 F1), niewiele gorsze wyniki osiągnęła analiza zachowań i statystyk (0,984 F1) wykorzystująca taką samą liczbę cech. Najgorsza okazała

się analiza tekstu (0,924 F1) skupiająca najwięcej cech. Wynik osiągany przez analizę zależności nie jest zaskakujący, ponieważ w jej skład wchodzą cechy powiązane z liczbą obserwujących i obserwowanych, które uzyskały najlepsze wyniki indywidualne. Niższy wynik analizy tekstu można wyjaśnić tym, że wiele cech wchodzących w jej skład zajmuje się obiecującymi, lecz bardzo trudnymi do określenia zagadnieniami.

Analizując wyniki cech z grupy analizy tekstu widać, że analiza emocji okazała się najlepsza indywidualnie przekraczając 0,8 F1. Średnie podobieństwo wiadomości oraz sentyment osiągnęły gorsze wyniki, odpowiednio 0,77 i 0,74. Przewaga analizy emocji nad analizą sentymentu może wynikać z tego, że została wykonana z wykorzystaniem biblioteki stworzonej na bazie wiadomości z Twittera. Dzięki temu była lepiej przystosowana do badania charakterystycznych dla niego wypowiedzi. Przewaga analizy emocji widoczna była również w wyborze cech metodą Lasso w 2 kroku oraz zachłanną metodą eliminacji. Analiza subiektywności osiągnęła słaby wynik gorszy od prostych cech. Wyniki cechy opartej o zliczanie podobnych wiadomości, która jest bardzo podobne do cechy dotyczącej średniego podobieństwa są najgorsze ze wszystkich zbadanych. Liczba podobnych wiadomości mimo tragicznego wyniku indywidualnego była jednak często odrzucana o wiele później od pozostałych cech. Proste cechy polegające na zliczaniu różnych elementów treści wiadomości pozwalają razem na uzyskanie wyraźnie powyżej 0,7 F1 małym nakładem pracy.

Analiza zachowań i statystyk użytkowników była nieznacznie gorsza od analizy zależności. Na jej sukces w głównej mierze zapracowały cechy związane z kategoryzacją źródeł publikacji. Zarówno kategoryzacja źródeł poszczególnych wiadomości jak i ulubionych źródeł publikacji użytkowników przyniosły wyniki powyżej 0,83 F1. Dalsze rozdrabnianie kategorii źródeł nie powinno już przynieść znaczącej poprawy. Wiarygodność retweetowanych użytkowników w postaci stosunku obserwujących i obserwowanych jest również bardzo przydatna. W kwestii analizy częstości pisania wiadomości bardziej zaawansowane wyszukiwanie serii wiadomości użytkownika osiągnęło lepsze rezultaty niż prosta cecha badająca maksymalną liczbę wiadomości w oknie 15 minut. Można było jednak liczyć na nieco większą różnicę. Uwzględnienie średniej pięciu serii znacząco podwyższyło osiągane wyniki. Wszystkie analizy częstości przekroczyły 0,7 F1, a średnia pięciu serii użytkownika zbliżyła się do wartości 0,8. W przeciwieństwie do kategoryzacji źródeł słabo wypada kategoryzacja portali z linków. Wprowadzony podział nie jest najprawdopodobniej wystarczający. Cechy z tej grupy są wraz z cechami dotyczącymi podobieństwa wiadomości są najlepsze do wykrywania spamerów, co zostało pokazane przy okazji odfiltrowywania małej grupy kont ze zbioru normalnych użytkowników.

Analiza zależności między użytkownikami osiągnęła najlepsze wyniki. Składają się na nią, tak jak już wspomniano, cechy związane z liczbą obserwujących i obserwowanych osiągające indywidualnie 0,9 F1 nieosiągalne dla innych cech. Analizy związane z grafami indywidualnie nie osiągają bardzo wysokich wyników, ponieważ odnoszą się do podgrupy wiadomości będącej retweetami, lecz razem pozwalają na przekroczenie 0,7 F1 na zwykłym zbiorze. Swoją przydatność pokazują o wiele wyraźniej na zbiorze składającym się z samych retweetów. Liczba polubień wiadomości również znajdująca się w tej grupie jest drugą z najgorszych cech, czego można było się spodziewać po zauważeniu, że mechanizm polubień jest ignorowany przez użytkowników Twittera.

Najlepsze wyniki osiągane są przez bardzo proste cechy dotyczące liczb obserwujących i obserwowanych. Zostały one wybrane w większości zbiorów cech, w tym do zbioru składającego się z 4 cech wybranych metodą Lasso dającego wynik F1 równy 0,995. W przypadku odrzucenia liczb obserwujących, obserwowanych oraz ich stosunku nadal jesteśmy w stanie otrzymać wynik powyżej 0,992 miary F1. Zachłannie dążąc do jak największego wyniku eliminując kolejne najmniej wnoszące cechy można osiągnąć wynik w okolicach 0,999 z wykorzystaniem 8 cech. Osiągane wyniki klasyfikacji stworzonych zbiorów są bardzo wysokie. Można powiedzieć, że jest to podejrzane. Wygląda na to, że w zbiorze testowo-treningowym znajduje się mała liczba niepożądanych użytkowników trudnych do wykrycia. Można powiedzieć, że trudni do wykrycia niechciani użytkownicy działający z dobrze przemyślanymi taktykami są wyjątkowi, a co za tym idzie bardzo rzadcy.

Najlepiej sprawdzającymi się algorytmami klasyfikacji w innych pracach z badanej dziedziny były lasy losowe i naiwny algorytm bayesowski. W przypadku drzew decyzyjnych zalazło to potwierdzenie w uzyskanych wynikach, gdzie pośród różnych dobrze wypadających algorytmów opartych o drzewa, lasy losowe osiągają najlepsze rezultaty. Na kolejnych miejscach znajduje się klasyfikacja z wykorzystaniem algorytmu k-najbliższych sąsiadów w różnych konfiguracjach. Wynik powyżej 0,9 F1 uzyskał również algorytm AdaBoost, który po zwiększeniu liczby cech wyprzedził algorytm ExtraTree i k-najbliższych sąsiadów i znalazł się na trzecim miejscu najlepszych klasyfikatorów. Wszystkie naiwne algorytmy bayesowskie wypadły słabo. W niektórych pracach o podobnej tematyce algorytmy bayesowskiego przynosiły najlepsze wyniki lub były bardzo blisko drzew decyzyjnych. W badanym przypadku to się nie powtórzyło.

8. Podsumowanie pracy

Rozdział podsumowuje dokonane prace oraz dokonuje oceny realizacji założonych celów. Zostaje również podjęty temat dalszych możliwych prac, które mogą rozwinąć lub usprawnić działanie stworzonego rozwiązania.

8.1. Wnioski wyciągnięte z analizy i klasyfikacji użytkowników

Do celów pracy należała analiza zachowań użytkowników portalu społecznościowego i na ich podstawie opracowanie zestawu cech, który pozwoli na stworzenie klasyfikatora identyfikującego niepożądanych użytkowników, czyli trolli i spamerów. Pracę można więc podzielić na dwa etapy: zaobserwowanie zachowań oraz implementacja cech pozwalających na ich opisanie oraz późniejszą ewaluację zdolności klasyfikacyjnych cech i wybranie z nich podzbiorów najlepiej rozróżniających użytkowników.

Do analizy został wybrany portal społecznościowy Twitter. Zdobył on dużą popularność na świecie, szczególnie w Stanach Zjednoczonych, co było dużą zaletą, ponieważ analizowanym językiem był angielski. Dodatkowo można znaleźć na nim wiele tematów politycznych, które idealnie wpasowały się w wymogi pracy. Niestety polityka udostępniania danych Twittera wpłynęła na wyniki. Nie było możliwe znalezienie idealnego zbioru danych, a tym bardziej własnoręczne manualne otagowanie wiadomości. Stworzony zbiór potencjalnie normalnych użytkowników na bazie id wiadomości udostępnianych w bazie Harvard Dataverse, nawet po odfiltrowaniu wprowadził pewien nieznaczny szum do rozwiązania. Dodatkowo zbiór niepożądanych użytkowników miał wpływ na ograniczenie innowacyjności pracy, ponieważ zawierał o wiele mniej parametrów, które udostępniane były w drugim ze zbiorów.

Cechy możliwe do uzyskania zostały podzielone na trzy różne kategorie ze względu na źródło ich pochodzenia: pochodzące z analizy treści wiadomości, analizy zachowań i statystyk użytkowników oraz analizy zależności między użytkownikami. W sumie ze wszystkich kategorii zostało opracowane ponad 20 cech, które zostały przeanalizowane w różnych wariantach. Nie skupiano się wyraźnie na konkretnej kategorii, lecz starano zachować równowagę między wszystkimi, aby podejść do problemu z różnych stron. Pozwoliło to na opracowanie klasyfikatorów, które nie są ograniczone do konkretnych typów użytkowników, lecz mają bardziej rozległe pojęcie o badanym środowisku i funkcjonujących w nim użytkownikach.

W pracy podobnie jak w innych rozwiązaniach skorzystano z grupy wspólnych cech nazywanych “bazowymi”, stanowiących filar do dalszego rozwoju. Również w przypadku tej pracy potwierdziła się zasadność użycia wielu cech z tej grupy, między innymi liczb obserwujących i obserwowanych, które przynosiły bardzo dobre wyniki i znalazły się w większości najlepszych zestawów. Cechy bazowe zostały uzupełnione przez własne, autorskie cechy. Niektóre z nich pozwoliły na osiągnięcie satysfakcjonujących wyników. Stworzone finalne grupy cech osiągnęły wysokie wyniki klasyfikacji. Nawet w przypadku odrzucenia najwięcej wnoszących liczb obserwujących, obserwowanych oraz ich stosunku klasyfikator osiągał wynik miary F1 przekraczający 0,99. Potwierdziła się stosowność zastosowania lasów losowych, które okazały się najlepszym algorytmem klasyfikacji. Również wszystkie

pozostałe zastosowane algorytmy oparte o drzewa decyzyjne odniosły bardzo dobre wyniki, tak samo jak algorytm k-najbliższych sąsiadów oraz AdaBoost. Naiwne algorytmy bayesowskie przynoszące w innych pracach jedne z lepszych wyników, bardzo zbliżonych do lasów losowych, w tym przypadku nie odniosły sukcesu osiągając słabe wyniki.

Patrząc na osiągnięte wyniki poprzez pryzmat zaproponowanego podziału ze względu na pochodzenie cech, wszystkie trzy grupy przekroczyły wartość miary F1 równą 0,9. Najgorzej wypadły cechy dotyczące analizy tekstu, która zarazem stanowiła największe wyzwanie. Można wyróżnić analizę emocji wykorzystującą bibliotekę stworzoną na bazie analizy ponad miliarda wiadomości z Twittera, czyli badanego portalu, której zastosowanie pozwoliło na osiągnięcie najlepszych wyników indywidualnych w tej kategorii. Analiza sentymentu w obu wariantach z wykorzystaniem innej biblioteki osiągnęła wyraźnie gorsze wyniki. Stworzone cechy kategoryzujące źródła publikacji wiadomości przyniosły całkiem przyzwoite wyniki przekraczające indywidualnie wartość 0,8 F1. Wręcz przeciwnie wypadła kategoryzacja linków wiadomości, w której zaproponowany podział nie okazał się pomocny. Zaproponowana metoda badania częstości użytkowników osiągnęła lepsze wyniki niż proste podejście. Analiza wiarygodności źródeł wiadomości podawanych danych, w postaci stosunku obserwujących i obserwowanych retweetowanych kont oraz zaadaptowana do Twittera na wzór innej pracy analiza grafowa retweetujących się użytkowników przynoszą dobre wyniki, będąc ważnymi czynnikami przy klasyfikacji retweetów.

Założone na początku cele pracy zostały osiągnięte. Przeanalizowano zachowania użytkowników na portalu społecznościowym, dokonano implementacji cech oraz zbadano ich zdolności do rozróżniania użytkowników, a następnie zaproponowano różne zestawy cech pozwalające na stworzenie klasyfikatorów, które osiągnęły bardzo zadowalające wyniki. Nie znaczy to jednak, że nie ma możliwości dalszego rozwoju, które zostaną opisane w kolejnym punkcie.

8.2. Możliwości rozwoju

Możliwości rozwoju pracy są mocno uzależnione od posiadanego zbioru danych. Ciężko już w nim wyróżnić nowe kryteria charakteryzujące użytkowników. Aby to zmienić należało by to zrobić na własnoręcznie otagowanym zbiorze danych, który pozwoliłby na przeanalizowanie wielu odrzuconych w tej pracy cech. Stworzenie takiego zbioru jest jednak mało realnie w przypadku pracy wyłącznie jednej osoby. Należało by się więc skupić głównie na udoskonaleniu aktualnego rozwiązania w obrębie przebadanych cech. Należało by również przeanalizować opracowanego cechy na innym zbiorze danych. Osiągnięte wyniki są bardzo dobre, co może budzić pewne podejrzenia co do liczby trudniejszych do wykrycia niepożądanych użytkowników obecnych w zbiorze.

W kwestii analizy sentymentu i subiektywności z pewnością można dokonać dalszych postępów. Wybrana biblioteka TextBlob przejawia lepsze predyspozycje w wykrywaniu pozytywnych wypowiedzi. Bardziej przydatna byłaby sytuacja odwrotna. Przeanalizowane biblioteki badające sentyment nie osiągały znacząco lepszych wyników, lecz niektóre z nich radziły sobie o wiele lepiej w wykrywaniu negatywnych zdań. Były one jednak płatne, tak jak przykładowo Google sentiment API. Dużo wniosłoby zestawienie ich skuteczności z wykorzystaną w pracy biblioteką, lecz musiałoby się to opierać na po-

mniejszonym zbiorze danych, ze względu na koszty rosnące z każdą kolejną analizowaną wiadomością. Lepsze wyniki analizy emocji, która wykorzystywała bibliotekę stworzoną w oparciu o wiadomości z Twittera, mogłoby również sugerować próbę znalezienia innych bibliotek opartych na podobnym zestawie danych.

Możliwe byłoby również uogólnienie badanych cech wiadomości użytkowników poprzez uśrednienie wszystkich badanych cech, następnie przypisanie każdemu użytkownikowi modelu jego typowej wiadomości. Należałoby jednak wziąć pod uwagę to, że zwykli użytkownicy najczęściej piszą wyraźnie mniej wiadomości, przez co ich ocena mogłaby być mniej znacząca.

Jak już wcześniej wspomniano, kategoryzacja portali w linkach publikowanych w wiadomościach nie przyniosła dobrego efektu. Oparta była ona na kategoriach treści portali oraz ich popularności. Bardzo możliwe, że została ona wykonana z uwzględnieniem zbyt małej liczby portali, ponieważ analizowane były jedynie pierwsze pięćdziesiątki popularności w obu zbiorach.

Materiały źródłowe

[1] Distant supervision definition. http://deepdive.stanford.edu/distant_supervision. Data dostępu: 2020-08-19.

[2] draw.io. https://app.diagrams.net/. Data dostępu: 2020-08-10.

[3] Emojipedia - biblioteka symboli emocji. https://emojipedia.org/. Data dostępu: 2020-07-26.

[4] Encyklopedia PWN. https://encyklopedia.pwn.pl. Data dostępu: 2020-01-10.

[5] Global Digital Report 2019. https://wearesocial.com/global-digital-report-2019. Data dostępu: 2019-12-5.

[6] iemoji - biblioteka symboli emocji. https://www.iemoji.com/. Data dostępu: 2020-07-25.

[7] Libre office calc. https://pl.libreoffice.org/poznaj/calc/. Data dostępu: 2020-05-16.

[8] Mongodb. https://www.mongodb.com/. Data dostępu: 2020-08-10.

[9] Networkx. https://networkx.github.io/. Data dostępu: 2020-06-10.

[10] Numpy. https://numpy.org/. Data dostępu: 2020-08-10.

[11] pandas. https://pandas.pydata.org/. Data dostępu: 2020-08-10.

[12] PostgreSQL. https://www.postgresql.org/. Data dostępu: 2020-08-10.

[13] Python. https://www.python.org/. Data dostępu: 2020-08-19.

[14] Repozytorium kodu deepmoji na portalu github. https://github.com/bfelbo/DeepMoji. Data dostępu: 2020-06-25.

[15] scikit-learn. https://scikit-learn.org/stable/. Data dostępu: 2020-08-10.

[16] tweepy. https://www.tweepy.org/. Data dostępu: 2020-08-10.

[17] Twitter API. https://developer.twitter.com/en/docs. Data dostępu: 2020-01-16.

[18] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference, CEAS 2010*, 2010. Cited By :436.

[19] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. Cited By :2739.

[20] J. Brownlee. Logistic regression for machine learning. https://machinelearningmastery.com/logistic-regression-for-machine-learning/. Data dostępu: 2020-04-11.

[21] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil. Universal sentence encoder. *CoRR*, abs/1803.11175, 2018.

[22] J. De-La-Peña-Sordo, I. Santos, I. Pastor-López, and P. G. Bringas. *Filtering trolling comments through collective classification*, volume 7873 LNCS of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2013. Cited By :3.

[23] L. M. de La Vega. Determining trolling in textual comments. 2017.

[24] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[25] V. Fonti. Feature selection using lasso. page 4, 2017.

[26] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003. Cited By :1487.

[27] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas. *Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying*, volume 239 of *Advances in Intelligent Systems and Computing*. 2014. Cited By :29.

[28] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. Cited By :1723.

[29] B. Ghanem, D. Buscaldi, and P. Rosso. Textrolls: Identifying russian trolls on twitter from a textual perspective, 2019.

[30] C. Hardaker. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research*, 6(2):215–242, 2010. Cited By :196.

[31] Intellica.AI. Vader, ibm watson or textblob: Which is better for unsupervised sentiment analysis? https://medium.com/@Intellica.AI/vader-ibm-watson-or-textblob-which-is-better-for-unsupervised-sentiment-analysis-db4143a39445, 2019.

[32] M. Iqbal. Twitter revenue and usage statistics (2019). 2019.

[33] H. Jalonen, J. Paavola, T. Helo, M. Sartonen, and A.-M. Huhtinen. Understanding the trolling phenomenon: The automated detection of bots and cyborgs in the social media. *Journal of Information Warfare*, 15:100–111, 12 2016.

[34] N. K. Kain. Understanding of multilayer perceptron (mlp). https://medium.com/@AI_with_Kain/understanding-of-multilayer-perceptron-mlp-8f179c4a135f. Data dostępu: 2020-04-11.

[35] A. M. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010. Cited By :5303.

[36] T. Li, J. Gharibshah, E. E. Papalexakis, and M. Faloutsos. Trollspot: Detecting misbehavior in commenting platforms. *CoRR*, abs/1806.01997, 2018.

[37] J. Littman, L. Wrubel, and D. Kerchner. 2016 United States Presidential Election Tweet Ids, 2016.

[38] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009. Cited By :1345.

[39] A. Spruds, A. Rožukalne, K. Sedlenieks, M. Daugulis, D. Potjomkina, B. Tölgyesi, and I. Bruge. Internet trolling as a hybrid warfare tool: the case of latvia. 2015.

[40] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *Proceedings - Annual Computer Security Applications Conference, ACSAC*, pages 1–9, 2010. Cited By :410.

[41] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R. P. Sheridan, and B. P. Feuston. Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1948, 2003. Cited By :1272.

[42] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. Design and evaluation of a real-time url spam filtering service. pages 447–462, 2011. Cited By :275.

[43] A. H. Wang. Don't follow me - spam detection in twitter. In *SECRYPT 2010 - Proceedings of the International Conference on Security and Cryptography*, pages 142–151, 2010. Cited By :233.

[44] X. Wu, V. Kumar, Q. J. Ross, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. . Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):20, 22–23, 24–26, 27–28, 2008. Cited By :2683.

Załączniki

A. Organizacja repozytorium kodu

Repozytorium kodu składa się z czterech katalogów zawierających kolejno: kod napisany w języku Python, skrypty SQL, arkusze kalkulacyjne na podstawie, których stworzono tabele i wykresy, kod źródłowy dokumentu pracy.

Katalog "code" zawierający kod posiada pięć pakietów, które skupiają powiązane ze sobą moduły napisane w języku Python, wykorzystywane w kolejnych etapach badania. Nazwy pakietów jednoznacznie wskazują na funkcjonalności zawartego w nich kodu i odpowiadają przedstawionej w pracy strukturze systemu. Pakiet "prepare_data" zawiera skrypt (TweetsGetterPostgres) pozwalający na pobieranie twitterowych wiadomości, których id zawarte są w plikach wejściowych i zapisywanie ich do bazy oraz skrypt (DatasetUpdater) pozwalający na dokonywanie drobnych zmian w już pobranych danych. Pakiet "features_impl" w jednym skrypcie skupia implementacje cech. w pakiecie "analyze" znajdują się skrypty pozwalające na pojedyncze przeprowadzenie analiz w oparciu o zaimplementowane cechy i uzyskanie wyników w postaci plików tekstowych. Pakiet "create_testtrain_dataset" zawiera skrypty pozwalające na utworzenie zbioru testowo-treningowego. Jest on bardziej złożony od innych. Z wykorzystaniem klasy DatasetCreator możemy obliczyć wartości wszystkich cech dla wiadomości dla każdego zbioru z osobna i zapisać je w odpowiednim formacie, akceptowanym w kolejnych krokach. Skrypt o nazwie AppendEmotionResults pozwala na dołączenie do wcześniej stworzonego pliku wyników analizy emocji, która poprzez duże wymagania sprzętowe często musi być przeprowadzana oddzielnie, a potem scalana z głównym plikiem. Skrypty FilterMsgs, ClusterMsgs oraz CreateDatasetSample służą między innymi do odfiltrowywania wiadomości i procesu przeskalowywania zbiorów. Ostatni moduł "classify_examine" zawiera klasę, w której znajdują się metody analizujące przydatność klasyfikacyjną cech oraz metody dokonujące klasyfikacji.

W katalogu "sql_scripts" znajdują się dwa skrypty SQL. Pierwszy, który jest o wiele ważniejszy, pozwala na stworzenie odpowiedniej struktury relacyjnej bazy danych. Drugi zawiera sporo przydatnych w analizie zapytań do bazy danych.

Arkusze kalkulacyjne, które agregują wszystkie wyniki analiz i badań dostępne są w katalogu "csv_sheets". Arkusze zawierają tabele i diagramy zamieszczone w pracy. Można w nich jednak znaleźć również wyniki, który nie zostały uwzględnione w dokumencie.

Katalog "document_sources" zawiera kompletny kod źródłowy dokumentu pracy, napisany z wykorzystaniem narzędzia LaTeX.

Do repozytorium pracy na GitHub prowadzi poniższy link:
`https://github.com/WookaszU/TrollSpamTwitter`.

B. Przygotowanie i obsługa środowiska

Pierwszym krokiem przygotowującym jest instalacja bazy danych PostgreSQL w domyślnej konfiguracji. w przypadku systemu operacyjnego Windows powinna ona się uruchomić w tle jako serwis. Po podłączeniu do bazy należy stworzyć osobne tabele dla normalnych i niepożądanych użytkowników z wykorzystaniem dołączonego skryptu SQL. Zbiór niepożądanych użytkowników importowany jest bezpośrednio do tabeli z pliku CSV. Normalni użytkownicy muszą zostać pobrani z wykorzystaniem skryptu TweetsGetterPostgres dostępnego w module "prepare_data". Przed rozpoczęciem pobierania należy w nim wprowadzić klucz i token dostępowy do Twitter API oraz przygotować pliki wejściowe zawierające w kolejnych liniach numery id wiadomości do pobrania. w przypadku opisywanego badania czas pobrania ponad 11 milionów wiadomości wyniósł prawie 7 dni. Zbiór niepożądanych użytkowników musi zostać uzupełniony o brakujące kolumny obecne w zbiorze normalnych użytkowników. Możliwe jest to z wykorzystaniem metod dostępnych w skrypcie DatasetUpdater z tego samego pakietu.

Pliki wyjściowe uruchamianych analiz przeprowadzanych z wykorzystaniem modułu "analyze" są zwracane w formacie tekstowym. Zwykle zawierają pojedynczą kolumnę danych, której wartości znajdują się w kolejnych linijkach, przez co można je łatwo przekopiować do arkusza kalkulacyjnego. Z racji dużych zakresów wartości badanych cech, większość analiz zlicza wystąpienia wartości cech z danych przedziałów. Niestety, zmiana wartości tych przedziałów musi być wykonana poprzez modyfikację małego fragmentu kodu samej analizy.

Najbardziej skomplikowane tworzenie zbioru testowo-treningowego opiera się na uruchomieniu skryptów z modułu "create_testtrain_dataset". Cały proces tworzenia zbioru opiera się na przekazywaniu plików tekstowych. Wszystkie pliki wyjściowe mają zgodny format, co pozwala na przekazywanie wyjścia z jednego skryptu na wejście drugiego. w przypadku dodania nowej cechy należy dokonać aktualizacji parsera plików w kodzie. Skrypty obecne w pakiecie są bardzo wymagające sprzętowo. w przypadku pracy, 16 GB pamięci RAM okazało się niewystarczające do przeprowadzenia klastrowania ponad 11 milionów wiadomości w oparciu o wszystkie cechy.