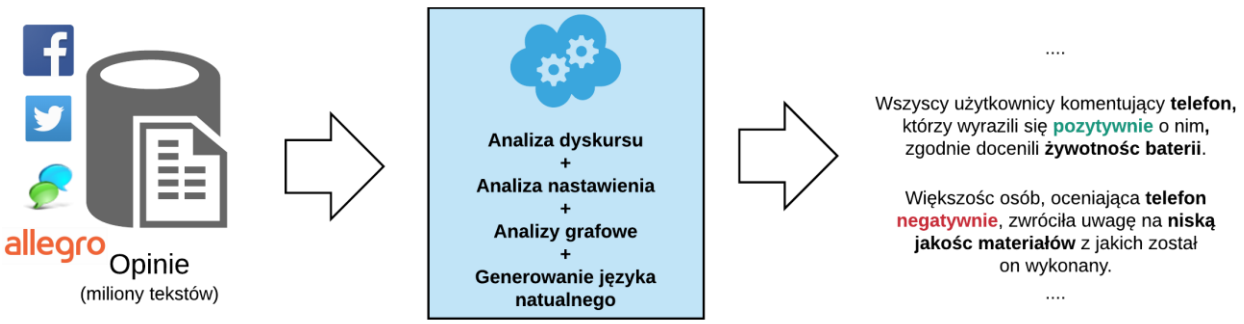
Współczesne społeczeństwo to społeczeństwo na wskroś informacyjne, bombardowane ze wszystkich stron ciągle rosnącą ilością różnych informacji. XXI wiek przyniósł nam dynamiczny rozwój mediów takich jak Internet oraz spowodował przeniesienie wielu dziedzin naszego życia do świata wirtualnego. W ten sposób wykreowane zostały nowe formy naszej komunikacji, a wraz z ich rozwojem powstało zapotrzebowanie na analizy tak powstałych danych. Owa, z reguły, nieustrukturyzowana otchłań informacyjna dostępna jest w postaci cyfrowej, jednak w jaki sposób analizować, podsumowywać miliardy nowych tekstów pojawiających się codziennie w Internecie? W tym miejscu z pomocą przychodzą techniki analizy języka naturalnego, statystyka oraz uczenie maszynowe. Co prawda są one rozwijane od wielu lat, jednak wciąż są potrzebne lepsze techniki umożliwiające analizę na niższym i bardziej złożonym poziomie. W ostatnich latach szczególnym zainteresowaniem cieszy się m.in. **analiza nastawienia lub wydźwięku**. Dziedzina ta jest definiowana jako badania na temat wyrażanych przez ludzi opinii, nastawienia, ocen czy emocji w stosunku do określonego tematu, przedmiotu, wydarzeni czy też innej osoby. Analiza nastawienia polega na określeniu polaryzacji tekstu, tzn. odpowiada na pytanie czy zadany fragment tekstu ma charakter **pozytywny, negatywny, neutralny**.

Ocena nastawienia może być dokonywana na poziomie (1) całego dokumentu, (2) poszczególnych zdań czy też (co jest najbardziej interesującym mnie zagadnieniem i zarazem niezwykle trudnym zagadnieniem naukowym) (3) na poziomie poszczególnych fragmentów tekstu. Trzeci z przedstawionych poziomów jest dopiero rozwijającym się działem analizy nastawienia i ma na celu przewidywanie nastawienia użytkowników do określonych tematów, cech produktów, usług, osób, tego typu podejście najczęściej nazywane jest analizą sentymentu na poziomie aspektów, a jego dużo częściej używanym odpowiednikiem angielskim jest *aspect-based sentiment analysis*. Dzięki temu jesteśmy w stanie agregować opinie nie tylko na poziomie całych tekstów, ale także określać, że np. wyświetlacz danego telefonu ma pozytywne opinie, ale bateria jest często oceniana negatywnie. Tego typu cechują się dużo wyższym skomplikowaniem, gdyż wymagają bardziej zaawansowanych reprezentacji wiedzy aniżeli jedynie na poziomie całych dokumentów. Dodatkowo, często dokumenty składają się z wielu zdań, a powiedzenie iż dokument jest pozytywny to przekazanie jedynie częściowej informacji opisującej dane zjawisko. Aktualnie w literaturze występują początkowe prace związane z analizą treści dotykające tego problemu. Temat jest niezwykle złożony z uwagi na problemy związane tym jak połączyć określone części wypowiedzi z produktami i ich aspektami (cechami, atrybutami). Brak w owych badaniach tworzenia modelu, który jest w stanie w sposób automatyczny określać na tak niskim poziomie nastawienia występującego w tekście, szczególnie na przekroju wielu domen - domen rozumianych w tym miejscu jako tematycznie powiązanych zbiorów danych, np. tekstów związanych z muzyką, elektroniką, książkami, polityką, ubraniami, hotelami itd.



Cel projektu

Tematyka moich badań wypełnia lukę przestawioną powyżej i ma na celu budowę kompleksowego zbioru technik umożliwiających, przygotowanie, analizę tekstów opinii oraz generowanie przyjaznych dla użytkownika raportów, także w postaci opisowej w postaci języka naturalnego. Aktualnie istniejące rozwiązania oferują analizy na poziomie całych dokumentów, a jeśli przechodzą już na poziom poszczególnych cech produktów są one jedynie powierzchowne i słabo przygotowane do analizy wielkich wolumenów danych – co widać szczególnie w artykułach naukowych gdzie analizy dokonuje się jedynie na kilkuset recenzjach. Wartym wspomnienie jest fakt, iż zadanie to jest bardzo problematyczne z uwagi na różnorodność języków i trudność w zbudowaniu jednego rozwiązania, które może pokryć wszystkie języki wykorzystywane na świecie. Analiza języka naturalnego często wymaga dodatkowych kroków szczególnie na etapie przygotowania danych do analizy, kroków specyficznych dla języka tekstu. Duże różnice widać w analizie języka polskiego (wiele odmian, język silnie fleksyjny) oraz języka angielskiego. Proponowane przeze mnie rozwiązanie pokrywać będzie co najmniej języki: **polski i angielski**. Przeprowadzenie tego typu analiz wymaga przeprowadzenia szeregu badań dla obu języków i wybraniu najwłaściwszych rozwiązań dla każdego z nich.