W obecnych czasach powszechnego dostępu do Internetu, każdy użytkownik może stać się kreatorem treści. Ogrom możliwości i zróżnicowanych platform z pozwala każdemu na swobodne wyrażanie poglądów. Złudzenie anonimowości często jednak powoduje, że część osób bezrefleksyjnie tworzy treści obraźliwe, atakujące oraz pełne nienawiści. Prawo dotyczące takich wpisów jest bardzo zróżnicowane, w zależności od kraju. Istnieją takie państwa, gdzie mowa nienawiści jest przestępstwem, jednak z drugiej strony, definicja takiej mowy nie jest precyzyjna. Okazuje się, że obraźliwość tekstu jest zjawiskiem dość subiektywnym i ciężko jest znaleźć taką granicę dopuszczalności swobody wypowiedzi, by usatysfakcjonować zarówno twórców opinii, jak i ich odbiorców. Mimo to, wielu właścicieli mediów społecznościowych jest zmuszona cenzurować treści, które uznają za obraźliwe. Niekiedy nosi to znamiona cenzury, gdyż proste algorytmy, bazujące na słowach kluczowych, są w stanie blokować także użytkowników, którzy w sposób kulturalny zabierają głos w tematach kontrowersyjnych. Często dochodzi też do zwyczajnej pomyłki. Frustrację części użytkowników rodzi też fakt, że ich potrzeby i uczucia w sferach poddawanych krytyce w przestrzeni publicznej, nie są prawnie chronione, podczas gdy inna grupa społeczna wywalczyła sobie specjalne traktowanie w tym aspekcie. Z uwagi na fakt, że każdego dnia powstaje bardzo dużo nowych treści, właściciele mediów nie są w stanie ręcznie moderować wszystkiego. Coraz bardziej powszechne jest zatem tworzenie i wykorzystywanie algorytmów, które w sposób automatyczny decydują, czy dany tekst jest np. obraźliwy. Większość prac z tej tematyki jest rozwijana w ramach przetwarzania języka naturalnego (ang. Natural Language Processing, NLP) i jest tam od dawna znana pod nazwą klasyfikacji tekstu. Istnieje dużo gotowych rozwiązań w tym temacie, jednak do tej pory zadania koncentrowały się przede wszystkim w obszarach o dużym stopniu obiektywności, np. rozpoznawanie języka tekstu, dziedziny, stylu funkcjonalnego, a także wydobywanie i klasyfikacja jednostek nazewniczych, wyrażeń temporalnych, relacji semantycznych, itp. W takich zadaniach praktycznie nie ma kontrowersji, a zbiory danych, ręcznie anotowanych przez ludzi, charakteryzują się wysokim poziomem zgodności decyzji pomiędzy ekspertami. Zupełnie inaczej wygląda sytuacja ze zjawiskami subiektywnymi, np. rozpoznawanie mowy nienawiści, obraźliwości, toksyczności, humoru, emocji, wydźwięku, itp. Obecnie dużo technik znanych historycznie z NLP, adaptuje się do tych problemów, stosując np. głosowanie osób na daną etykietę i modelując tym samym decyzję większościową. Drugi popularny nurt zakłada tworzenie wytycznych dla anotatorów, w których określa się, co osoby mają uznawać np. za treść obraźliwą. Metody te nie uwzględniają jednak wrażliwości konkretnych osób. Systemy budowane na podstawie tak przygotowanych zasobów i narzędzi, wykazują się stronniczością względem konkretnych grup społecznych i pojedynczych osób. Proponujemy zatem zupełnie nowe podejście do tego tematu, które bazuje na osiągnięciach znanych z systemów rekomendacyjnych. Rozwiązanie uwzględnia osobowość konkretnej osoby, zarówno w trakcie przygotowywania modelu, jak i podczas podejmowania decyzji w systemie produkcyjnym. Dodatkowo rozszerza istniejące rozwiązania o mechanizm rozumienia treści i budowania reprezentacji człowieka z wykorzystaniem reprezentacji zarówno treści, jak i potencjalnego kontekstu jej występowania. Opracowane metody będą w stanie odpowiedzieć np. **dla kogo** dana treść może być obraźliwa, a nie tylko **czy** dana treść jest obraźliwa. Jednocześnie w systemach mediów społecznościowych umożliwią użytkownikowi wybór, czy nie chce oglądać treści danego typu i pozwolą na filtrację z perspektywy konkretnego człowieka, która na podstawie jego historycznych decyzji będzie uczyć się jego preferencji. Rozwiązanie to będzie aplikowalne także dla wszystkich zjawisk subiektywnych, w których różnice w odbiorze między ludźmi są czymś naturalnym.