Penelitian terkait penggunaan pendekatan *machine learning* dan *natural language processing* untuk melakukan penyaringan terhadap tulisan-tulisan negatif di komunitas *online* telah banyak dilakukan. Salah satunya tertuang pada paper yang ditulis oleh Theodora Chu dan kawan-kawan [1]. Pada penelitian tersebut digunakan pendekatan machine learning untuk melakukan klasifikasi *abusive comment* dengan membandingkan tiga metode deep learning yang berbeda, yaitu recurrent neural network (RNN), convolutional neural network (CNN) dengan word embedding, dan convolutional neural network dengan *character embedding*. Dari penelitian yang mengguanakan 150.000 sample data yang diambil dari Wikipedia tersebut diperoleh kesimpulan bahwa penggunaan CNN dengan character embedding memiliki performa terbaik di antara ketiga metode yang dibandingkan. Akurasi dari model CNN yang dihasilkan pada penelitian tersebut dapat mencapai lebih dari 90%.

Penelitian lain juga pernah dilakukan oleh Chikashi Nobata dan kawan-kawan yang tertulis dalam paper berjudul “Abusive Language Detection in Online User Content” [2]. Pada penelitian tersebut, secara keseluruhan peneliti menggunakan pendekatan NLP untuk melakukan deteksi abusive content. Untuk merancang model, peneliti menggunakan empat kelas fitur terkait NLP, yaitu fitur n-gram, fitur linguistik, fitur sintaktik, dan fitur semantik. Penelitian tersebut mengguanakan data komentar-komentar yang diekstrak dari Yahoo! Finance and News. Dari penelitian tersebut dihasilkan model dengan akurasi yang sangat tinggi. Bahkan dalam paper tersebut dinyatakan bahwa akurasi dari model yang didasari oleh fitur-fitur NLP tersebut memiliki performa yang lebih baik dari model-model deep learning.

Selain kedua penelitian tersebut, masih ada beberapa penelitian lain terkait deteksi konten tulisan pada komunitas *online*, misalnya penelitian pada paper berjudul “*Detecting Offensive Language in Social Media to Protect Adolescent Online Safety*” yang ditulis oleh Ying Chen dan kawan-kawan [3] dan penelitian pada paper berjudul “*Offensive Language Detection Using Multi-level Classification*” yang ditulis oleh Amir H. Razavi dan kawan-kawan [4]. Penelitian-penelitian di atas telah dilakukan menggunakan metode yang berbeda-beda. Artinya, memang terdapat cukup banyak cara untuk melakukan deteksi konten terhadap tulisan-tulisan pada komunitas *online*. Selain itu, model yang diperoleh dari pengunaan metode-metode tersebut juga cukup akurat untuk melakukan deteksi konten negatif.