Abstrak

Product review banyak digunakan oleh calon pembeli untuk melihat opini pembeli lainnya, sebelum memutuskan untuk membeli suatu produk. Kumpulan review produk tersebut juga digunakan oleh perusahaan untuk mengidentifikasi permasalahan produk yang dipasarkannya serta menemukan informasi strategi pemasaran perusahaan pesaingnya.

Akan tetapi karena tidak ada *quality control* tentang penulisan review yang dapat dipercaya, maka setiap orang bebas untuk memberikan opininya sehingga tidak jarang kesempatan tersebut dimanfaatkan oleh *spammer*. Ada tiga jenis spam pada product review yaitu *untruthfull opinions, reviews on brand only*, dan *non-review*.

Metode yang dipilih untuk menyelesaikan kasus *review spam* ini adalah *Logistic Regression* karena dapat menghasilkan estimasi probabilitas suatu review merupakan spam atau bukan seperti yang dibutuhkan. Dalam tugas akhir ini keseluruhan sistem dibagi menjadi dua, yaitu proses klasifikasi untuk spam 1 dan proses klasifikasi untuk spam 2 dan 3. Untuk klasifikasi spam 2 dan 3 akan dilakukan pelabelan manual terlebih dahulu, kemudian dilakukan proses preprocessing untuk menemukan variabel prediksi yang diinginkan. Karena pada proses klasifikasi spam 1 tidak dilakukan pelabelan manual, maka terlebih dahulu akan dipisahkan *review* duplikat dan non duplikatnya dengan metode shingling dimana nanti *review* duplikat akan dianggap sebagai spam dan sisanya sebagai non spam untuk kemudian akan dianalisis karakteristik spam tipe 1. Dari hasil analisis tersebut diperoleh karakteristik spam 1 yang banyak menjadi target *spammer* adalah yang memiliki rating baik dan rata-rata (rating 3-5), untuk selanjutnya akan diberi *review spam* negatif untuk menjatuhkan produk tersebut

Kata kunci : Logistic regression, klasifikasi, duplikat, spam