INTISARI

IDENTIFIKASI KOMENTAR SPAM PADA SOSIAL MEDIA MENGGUNAKAN *COMPLEMENT NAÏVE BAYES*

Oleh

Syaiful Bachri Mustamin
17/418669/PPA05453

Permasalahan yang muncul ketika mengidentifikasi komentar *spam* dan *non spam* adalah ketika komentar tersebut ditulis menggunkan simbol-simbol unik yang berbeda dengan umumnya sering digunakan, sehingga mengakibatkan kata kata tersebut menjadi beragam dan sulit untuk diidentifikasi. Selain itu komentar *spam* lebih sedikit daripada komentar *non spam* sehingga mengarah pada masalah ketidakseimbangan data (*imbalance dataset*). *Imbalanced dataset* dapat memberikan pengaruh terhadap perfoma suatu metode klasifikasi.

Metode *Complement Naïve Bayes* (CNB) diketahui dapat menghitung *complement* kata atau fitur pada suatu kelas tertentu dengan mengidentifikasi bahwa kata tersebut berada di kelas lain. Solusi tersebut menjadi fokus penelitian terkait dengan pengembangan metode CNB dalam menangani *imbalance dataset* pada deteksi komentar *spam* pada sosisal media..

Berdasarkan hasil pengujian dengan data latih sebanyak 66 ribu dan data uji 1475 dan didalam praproses memasukkan normalisasi kata, maka didapatkan akurasi CNB sebesar 99 %, *precision* sebesar 96 % dan *f-measure* 96 %. Sedangkan *Gaussian Naïve Bayes* (GNB) menghasilkan akurasi sebesar 78 %, *precision* sebesar 87%, *f-measure* 79 %. Jika dalam praporses tanpa memasukkan proses normalisasi kata maka didapatkan akurasi CNB sebesar 96 %, precision sebesar 94 % dan f-measure 96 %. Sedangkan GNB menghasilkan akurasi sebesar 91 %, precision sebesar 88%, f-measure 82 %. Jika dalam praporses memasukkan proses normalisasi. Kesimpulannya normalisasi kata dapat meningkatkan akurasi dan metode CNB lebih cocok untuk mendeteksi komentar *spam* dengan dibandingkan dengan GNB.

Kata kunci: Sosial media, *Spam, Complement Naïve Bayes, Gaussian Naïve Bayes*.

ABSTRACT

IDENTIFICATION OF SPAM COMMENTS ON SOCIAL MEDIA USING COMPLEMENT NAÏVE BAYES

By:

Syaiful Bachri Mustamin
17/418669/PPA05453

The problem that arises when identifying spam and *non spam* comments is when the comments are written using unique symbols that are different from the commonly used ones, resulting in words that are diverse and difficult to identify. In addition spam comments are fewer than non-spam comments so it leads to the problem of imbalance dataset. Imbalanced dataset can influence the performance of a classification method.

Complement Naïve Bayes (CNB) method is known to be able to calculate the complement of words or features in a particular class by identifying that the word is in another class. The solution is the focus of research related to the development of CNB methods in handling dataset imbalance in the detection of spam comments on social media.

Based on the test results with 66 thousand training data and 1475 test data and in the pre-processing with word normalization, then obtained CNB accuracy of 99%, precision of 96% and f-measure 96%. The Gaussian Naïve Bayes (GNB) method produces an accuracy of 78%, precision of 87%, f-measure 79%. If the pre-process includes the word normalization process, the CNB accuracy is 96%, the precision is 94% and the f-measure is 96%. In addition, GNB produces an accuracy of 91%, precision of 88%, f-measure 82%. If in the pre-processing process, without word normalization process. In conclusion word normalization can improve accuracy and the CNB method is more suitable for detecting spam comments compared to GNB

Keywords: Social media, Spam, Complement Naïve Bayes, Gaussian Naïve Bayes.