Terbit online pada laman web jurnal : **http://jurnal.iaii.or.id**

JURNAL RESTI

(Rekayasa Sistem dan Teknologi Informasi)

Vol. 2 No. 3 (2018) 708 – 715 ISSN : 2580-0760 (media online)

Pengaruh Normalisasi Teks Dengan Text Expansion Dalam Deteksi Komentar Spam Pada Youtube

Imam Thoib[a], Arief Setyanto[b], Suwanto Raharjo[c]

[a]Magister Teknik Informatika, Universitas Amikom Yogyakarta, imam.thoib@students.amikom.ac.id

[b]Magister Teknik Informatika, Universitas Amikom Yogyakarta, arief_s@amikom.ac.id

[c]Teknik Informatika, Fakultas Teknologi Industri, Institut Sains & Teknologi AKPRIND Yogyakarta, wa2n@akprind.ac.id

Abstract

The popularity of Youtube as the largest video sharing website in the wolrd give spammers opportunities to get benefit from Youtube in illegal ways by putting spam comments on Youtube's videos. Spam comments are very troubling to channel owners. The variants of spam comments are becoming more difficult to detect. One of them is spam comments using abbreviations, symbols, terms or misspelled word to make detection difficult. This research evaluate some classification techniques and employ text normalization method called TextExpansion to deal with this problem. This research uses Youtube Spam Collections dataset from UCI Machine Learning Library composed by five different datasets, which each one contains text comments extracted from YouTube videos (Psy, Katty Perry, LMFAO, Eminem and Shakira). The evaluation results shows TextExpansion is able to produce the highest accuracy value of 90.23%. To determine the impact of applying the TextExpansion method, this research conducted t-test for each dataset. The results of t-test for each dataset shows P(T<=t) two-tail < 0.05 which indicates a significant impact after applying text normalization using TextExpansion.

Keywords: spam detection, text normalization, text expansion, youtube spam comments

Abstrak

Popularitas Youtube sebagai website berbagi video terbesar di dunia memberikan celah kepada *spammer* untuk mendapatkan keuntungan dari Youtube dengan cara yang ilegal. Salah satu cara yang dilakukan adalah melakukan komentar spam pada video-video yang ada di Youtube. Komentar spam menjadi hal yang sangat meresahkan pemilik channel. Varian komentar spam kian banyak dan menyulitkan untuk dideteksi oleh Youtube. Peneliti melakukan evaluasi beberapa metode klasifikasi dengan menerapkan normalisasi teks pada tahap praproses menggunakan metode *TextExpansion* untuk mengatasi masalah ini. Dataset yang diuji adalah dataset Youtube Spam Collections dari UCI Machine Learning Library yang terdiri dari 5 dataset dari komentar yang diekstrak dari video Psy, Katy Perry, LMFAO, Eminem dan Shakira. Dataset terdiri dari 1005 komentar spam dan 951 komentar bukan spam. Hasil evaluasi menunjukkan *TextExpansion* mampu menghasilkan nilai akurasi tertinggi 90.23%. Untuk mengetahui pengaruh penerapan metode *TextExpansion*, peneliti melakukan uji beda *t-test* pada setiap dataset. Hasil uji beda *t-test* pada setiap dataset menunjukkan nilai *P(T<=t) two-tail* < 0.05 yang menunjukkan adanya pengaruh yang signifikan setelah diterapkan normalisasi teks menggunkan metode *TextExpansion.*

Kata kunci: deteksi spam, normalisasi teks, *text expansion*, komentar spam youtube

1. Pendahuluan

Saat ini Youtube menempati urutan pertama website berbagi video di dunia. Pada tahun 2018 ini, pengguna Youtube mencapai lebih dari 1 milyar atau mencapai 1/3 dari seluruh pengguna Internet di dunia yang terdiri dari pengguna berusia 18-49 tahun[1]. Pada perkembangannya, lebih dari separuh penayangan di Youtube berasal dari perangkat mobile [1]. Hal ini menjadi sebuah pencapaian yang luar biasa bagi Google yang telah menyisihkan website berbagi video seperti Vimeo dan yang lainnya.

Namun, di balik pesatnya perkembangan dan banyaknya pengguna Youtube, terdapat tantangan yang dihadapi. Salah satunya adalah pemanfaatan Youtube yang tidak semestinya, yaitu spam. Salah satu jenis spam yang ada di Youtube adalah komentar spam. Ciri komentar spam di Youtube biasanya terdapat link ke situs pornografi, website *online dating* atau video yang tidak relevan yang biasanya dilakukan secara otomatis menggunakan mesin [2][3].

Spam bukanlah hal yang baru, sebuah perusahaan keamanan komputer bernama Nexgate merilis laporan

Diterima Redaksi : 18-10-2018 | Selesai Revisi : 30-10-2018 | Diterbitkan Online : 16-12-2018

yang menunjukkan adanya peningkatan volume spam di media sosial sebesar 355% di pertengahan tahun 2013. Facebook dan Youtube menempati peringkat teratas, dimana ditemukan 100 komentar spam di Youtube dan Facebook setiap satu komentar spam ditemukan di sosial media lain[4].

Banyaknya komentar spam di Youtube mengakibatkan beberapa channel mematikan fasilitas komentar pada videonya. Seperti yang diberitakan oleh situs berita The Guardian, pemilik channel PewDiePie yang telah memiliki 60 juta pengikut telah mematikan fasilitas komentar pada videonya. PewDiePie menyatakan bahwa mayoritas komentar di videonya adalah spam[5]. Jika tidak ditangani dengan serius, komentar spam dapat mengganggu pemilik channel dan pengguna Youtube lainnya.

Youtube sendiri sudah memiliki mekanisme untuk menangani komentar spam yaitu salah satunya dengan memfasilitasi pengguna dengan tombol untuk melaporkan sebuah komentar yang terindikasi spam. Namun cara ini dinilai masih kurang maksimal karena bisa saja laporan dari pengguna merupakan laporan palsu. Banyaknya komentar juga menjadi masalah yang besar jika harus menghapus satu per satu komentar. Sehingga masalah ini menarik para peneliti untuk melakukan penelitian untuk menangani komentar spam secara otomatis.

Ada beberapa peneliti yang telah melakukan penelitian untuk menangani masalah spam ini dengan berbagai macam teknik dan metode. Penanganan komentar spam yang banyak dilakukan adalah menggunakan teknik *machine learning* dengan metode klasifikasi yang jamak digunakan seperti *Support Vectore Machine* (SVM), *Naïve Bayes* (NB) dan *K-Nearest Neighbour*[6]. Metode-metode tersebut memberikan performa yang bagus untuk melakukan deteksi komentar spam di Youtube[7].

Namun ada masalah yang dihadapi dalam penanganan komentar spam, di mana komentar spam tidak hanya dilakukan oleh mesin, melainkan juga dilakukan oleh pengguna asli[8]. Biasanya digunakan kalimat yang sangat pendek, singkatan, simbol, istilah dan ejaan yang salah untuk mempersulit pendeteksian komentar spam tersebut secara otomatis[9]. Hal ini tentunya menarik untuk diteliti, diperlukan sebuah teknik normalisasi teks pada proses *pre-processing* untuk mengatasi masalah tersebut[9]. Diperlukan sebuah teknik normalisasi teks yang digunakan untuk mengubah teks asli yang berisi singkatan dan simbol menjadi teks yang standar[10].

Almeida, T. A. et al. merancang *TextExpansion* untuk melakukan normalisasi teks pada penanganan SMS spam. Teknik nomalisasi tersebut memanfaatkan *Freeling English* dan *NoSlang Dictionary* untuk mengkonversi teks asli menjadi teks yang standar dan tanpa singkatan[6]. Penggunaan *TextExpansion* tersebut menunjukkan adanya peningkatan akurasi dalam deteksi SMS spam dibanding tanpa menggunakan normalisasi teks.

Penggunaan teknik normalisasi teks juga dilakukan oleh beberapa peneliti untuk meningkatkan akurasi deteksi spam. Silva, R. M. et al. mengadopsi *TextExpansion* untuk melakukan normalisasi teks pada deteksi komentar spam di Youtube menggunakan metode *MDLText*. Penggunaan *TextExpansion* terbukti meningkatkan nilai *Spam Caught Rate* (SC) pada deteksi komentar spam. Nilai SC yang diperoleh dari metode tersebut adalah 0.937[9].

Pada penelitian ini akan dilakukan normalisasi teks menggunakan *TextExpansion*[10] pada deteksi komentar spam di Youtube menggunakan metode klasifikasi *Decision Tree, Logistic Regression, Naïve Bayes, K-Nearest Neighbour, Random Forest dan Support Vector Machine*[6] untuk mengetahui pengaruh dari normalisasi teks pada deteksi komentar spam di Youtube.

2. Tinjauan Pustaka

Masalah spam sudah ada sejak dahulu. Awalnya bentuk spam berupa email spam. Di mana email menjadi sasaran utama spamming. Dalam perkembangannya target spamming bertambah, seperti sms, online instant messaging, komentar blog dan media sosial. Penelitian tentang penanganan masalah spam sudah dimulai sejak beberapa tahun yang lalu. Penelitian tersebut berfokus untuk mengatasi masalah spam pada email [11][12][13][14].

Roy, K. et al. & Lee, C.-N. et al. mengklasifikasikan email spam menggunakan metode Naïve Bayes. Logika *Longest Common Sequence* (LCS) dikombinasikan dengan metode Naïve Bayes[13], sedangkan Lee, C.-N. et al. menggunakan metode Weighted Naïve Bayes. Keduanya menghasilkan nilai akurasi yang bagus untuk mendeteksi email spam. Perbedaannya, Lee, C.-N. et al. berfokus pada ekstraksi subjek email untuk melakukan deteksi dan menyarankan untuk meneliti header email pada penelitian selanjutnya.

Idris, I. et al. & Zavvar, M. et al. melakukan optimasi metode machine learning menggunakan *Particle Swarm Optimization* (PSO). *Negative Selection Algorithm* (NSA) dikombinasikan dengan PSO menghasilkan nilai akurasi lebih rendah dibandingkan dengan metode SVM. Dimana NSA-PSO mendapatkan nilai akurasi sebesar 83.2% sedangkan SVM 90%[11]. Sebaliknya performa SVM-PSO menunjukkan akurasi terbaik untuk mendeteksi email spam[14].

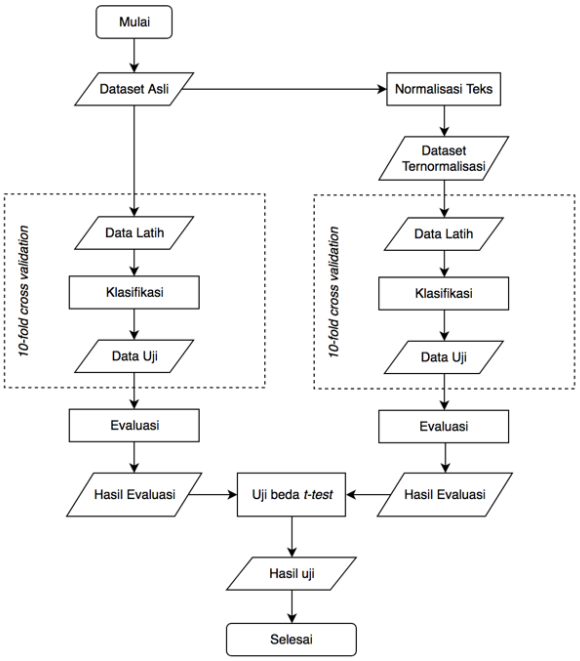
Pada era media sosial saat ini telah dilakukan penelitian untuk mendeteksi spam di media sosial [15][16][17]. Salah satu contohnnya adalah pendeteksian komentar spam pada Twitter menggunakan teknik machine learning. Metode *Hidden Markov Model* (HMM)

menunjukkan hasil akurasi yang baik dalam deteksi komentar spam di Twitter[15].

Deteksi komentar spam menghadapi masalah dimana spam dilakukan tidak hanya oleh mesin[8], melainkan juga dilakukan oleh pengguna asli. Mereka menggunakan simbol, singkatan, istilah dan menggunakan kata yang sengaja disalahkan ejaannya untuk mempersulit penyaringan. Diperlukan sebuah teknik untuk mengubah teks tersebut menjadi teks yang standar, sehingga mempermudah filterisasi. Teknik tersebut disebut teknik normalisasi teks. Pada penelitian ini akan ditambahkan proses normalisasi teks yang diadopsi dari metode *TextExpansion* untuk mengatasi masalah ini[10].

3. Metodologi Penelitian

Langkah-langkah yang dilakukan dalam penelitian disajikan dalam alur penelitian pada gambar 1.



Gambar 1. Alur Penelitian

3.1 Dataset

Untuk mendapatkan hasil yang akurat, peneliti menggunakan dataset yang terpercaya sebagai bahan penelitian. Dataset yang digunakan adalah dataset Youtube Spam Collections dari UCI Machine Learning Library yang terdiri dari 5 dataset dari komentar yang diekstrak dari video Psy, Katy Perry, LMFAO, Eminem and Shakira[6]. Pada setiap dataset terdapat komentar spam dan ham dengan jumlah yang berbeda. Rincian jumlah komentar spam dan ham pada setiap dataset dapat dilihat pada tabel 1.

Salah satu contoh komentar spam dan *ham* yang terdapat pada dataset yang akan diuji dapat dilihat pada tabel 2.

Tabel 1. Proporsi komentar spam dan ham

Dataset	Spam	Ham	Total
Psy	175	175	350
KattyPerry	175	175	350
LMFAO	236	202	436
Eminem	245	203	448
Shakira	174	196	370
Total	1005	951	1956

Tabel 2. Contoh komentar spam dan ham

Jenis	Komentar
Spam	Hey, check out my new website!! This site is about kids stuff. Kidsmediausa . com
Ham	This video will get to 2 billion just because of people checking if it has hit 2 billion yet.

Setiap dataset ini akan digunakan pada proses normalisasi teks dan klasifikasi, kemudian dibandingkan hasil pengujiannya antara dataset yang dinormalisasi dan yang tidak.

3.2 Metode Normalisasi Teks

Pada tahap ini dilakukan normalisasi teks pada setiap dataset untuk mengubah teks asli menjadi teks ternormalisasi. Metode yang digunakan adalah *TextExpansion.* Pada tahap ini akan diterapkan semua aturan yang ada pada metode *TextExpansion.* Terdapat 10 aturan pada metode *TextExpansion* yang disajikan pada tabel 3.

Tabel 3. Aturan TextExpansion

Aturan	Keterangan
Expansion 1	*Concept Generations*
Expansion 2	*Word Sense Disambiguation*
Expansion 3	*Text Normalization*
Expansion 4	*Text Normalization + Concept Generations*
Expansion 5	*Text Normalization + Word Sense Disambiguation*
Expansion 6	*Original Text + Concept Generations*
Expansion 7	*Original Text + Word Sense Disambiguation*
Expansion 8	*Original Text + Text Normalization*
Expansion 9	*Original Text + Text Normalization + Concept Generations*
Expansion 10	*Original Text + Text Normalization + Word Sense Disambiguation*

Pada tahap ini digunakan sebuah *library* bernama *TextExpansion Tool* pada python 2.76 untuk melakukan proses normalisasi teks. Setiap dataset akan menghasilkan 10 dataset ekspansi hasil normalisasi teks. Total dataset yang akan diuji sebanyak 55 dataset termasuk dataset asli. Dataset ekspansi akan diberi nama sesuai dengan aturan yang diterapkan pada tabel 3.

3.3 Metode Klasifikasi

Pada tahap ini dilakukan klasifikasi pada setiap dataset, baik dataset asli maupun dataset hasil normalisasi. Metode klasifikasi yang akan digunakan dapat dilihat pada tabel 4.

Tabel 4. Metode klasifikasi

Metode	Keterangan
DT	*Decision Trees*
LR	*Logistic Regression*
NB-B	*Bernoulli Naïve Bayes*
NB-G	*Gaussian Naïve Bayes*
NB-M	*Mulninomial Naïve Bayes*
1-NN	*1-Nearest Neighbour*
3-NN	*3-Nearest Neighbours*
5-NN	*5-Nearest Neighbours*
RF	*Random Forests*
SVM-L	*Support Vector Machine* dengan kernel linier
SVM-P	*Support Vector Machine* dengan kernel *polynomial*
SVM-R	*Support Vector Machine* dengan kernel *gaussian*

Parameter metode klasifikasi yang digunakan pada setiap dataset berbeda. Digunakan parameter yang sama dengan peneliti sebelumnya yang dapat dilihat pada tabel 5.

Tabel 5. Parameter metode klasifikasi

Metode	Parameter	Dataset				
		Psy	Katty Perry	LMFAO	Eminem	Shakira
LR	C	10	10^5	10^2	10^2	10^2
NB-B	α	1	10	10^{-2}	10^{-3}	10^{-5}
NB-M	α	1	10	10^{-5}	10^{-1}	10^{-2}
RF	*#trees*	80	40	90	30	30
SVM-L	C	10^{-1}	10^{-1}	1	10^{-1}	1
SVM-P	C	10^{-5}	10^{-4}	10^{-3}	10^{-5}	10^{-4}
	γ	10	10	10	10^2	10
SVM-R	C	1	10^3	10^3	10^3	10^2
	γ	10^{-2}	10^{-2}	10^{-3}	10^{-4}	10^{-2}

Selain yang disajikan pada tabel 5, digunakan parameter bawaan dari metode yang digunakan. *Library* yang digunakan untuk proses klasifikasi adalah *scikit-learn v.0.16.1* pada python 2.7.6[6].

3.4 Metode Evaluasi

Pada tahap ini akan dilakukan evaluasi terhadap hasil klasifikasi. Metode *k-fold cross validation* digunakan untuk melipat data dan mengulangi eksperimen sebanyak *k*. Pada penelitian ini akan digunakan *10-fold cross validation*, dimana ekseperimen akan dilakukan sebanyak 10 kali dengan proporsi data latih dibanding data uji 9:1.

Selanjutnya, untuk membandingkan performa dari setiap metode klasifikasi, digunakan beberapa metode pengukuran yaitu: akurasi (Acc), *Spam Caught Rate* (SC), *Blocked Ham Rate* (BH), *F-measure* (F1), *Matthews Correlation Coefficient* (MCC). Hasil evaluasi akan disajikan dalam bentuk tabel yang akan menampilkan nilai pengukuran klasifikasi pada setiap dataset.

3.5 Uji Beda *t-test*

Untuk mengetahui pengaruh dari metode *TextExpansion* pada deteksi komentar spam dilakukan uji beda *t-test*. Dengan uji beda *t-test* nantinya akan dapat diketahui signifikan atau tidaknya pengaruh dari penerapan metode *TextExpansion*. Nilai α yang digunakan adalah 0.05 dengan tingkat kepercayaan 95%. Apabila nilai *P(T<=t) two-tail* < α maka pengaruh penerapan metode *TextExpansion* adalah signifikan, sebaliknya jika nilai *P(T<=t) two-tail* > α maka penerapan metode *TextExpansion* dikatakan tidak signifikan.

4. Hasil dan Pembahasan

4.1 Hasil Normalisasi Teks

Pada proses ini dilakukan normalisasi teks untuk mengubah teks asli menjadi teks ternormalisasi. Metode yang digunakan pada proses ini adalah metode *TextExpansion*. Proses normalisasi teks pada masing-masing dataset membutuhkan waktu yang bervariasi. Peneliti mencatat waktu yang digunakan untuk normalisasi teks setiap dataset pada tabel 6.

Tabel 6. Waktu Normalisasi Teks

Dataset	**Total Data**	**Waktu (menit)**		
		Real	**User**	**Sistem**
Psy	350	163	964	24
Katty Perry	350	162	205	23
LMFAO	438	136	189	22
Eminem	453	208	271	30
Shakira	370	170	219	24

Proses normalisasi teks tercepat diperoleh dari dataset LMFAO yaitu selama 136 menit, sedangkan dataset Eminem membutuhkan waktu paling lama pada proses ini yaitu selama 208 menit.

Dari proses normalisasi teks ini dihasilkan 10 variasi dataset yang disebut dataset *expansion* yang menerapkan aturan pada tabel 3. Hal ini mengakibatkan waktu komputasi yang dibutuhkan pada tahap ini relatif lama. Hasil dari proses normalisasi teks dapat dilihat pada gambar 2.

```
[root@s94738 files]# ls
Eminem.txt_conc.txt                LMFAO.txt_orig_conc.txt
Eminem.txt_disa.txt                LMFAO.txt_orig_disa.txt
Eminem.txt_norm.txt                LMFAO.txt_orig_norm.txt
Eminem.txt_norm_conc.txt           LMFAO.txt_orig_norm_conc.txt
Eminem.txt_norm_disa.txt           LMFAO.txt_orig_norm_disa.txt
Eminem.txt_orig.txt                Shakira.txt_conc.txt
Eminem.txt_orig_conc.txt           Shakira.txt_disa.txt
Eminem.txt_orig_disa.txt           Shakira.txt_norm.txt
Eminem.txt_orig_norm.txt           Shakira.txt_norm_conc.txt
Eminem.txt_orig_norm_conc.txt      Shakira.txt_norm_disa.txt
Eminem.txt_orig_norm_disa.txt      Shakira.txt_orig.txt
KatyPerry.txt_conc.txt             Shakira.txt_orig_conc.txt
KatyPerry.txt_disa.txt             Shakira.txt_orig_disa.txt
KatyPerry.txt_norm.txt             Shakira.txt_orig_norm.txt
KatyPerry.txt_norm_conc.txt        Shakira.txt_orig_norm_conc.txt
KatyPerry.txt_norm_disa.txt        Shakira.txt_orig_norm_disa.txt
KatyPerry.txt_orig.txt             psy.txt_conc.txt
KatyPerry.txt_orig_conc.txt        psy.txt_disa.txt
KatyPerry.txt_orig_disa.txt        psy.txt_norm.txt
KatyPerry.txt_orig_norm.txt        psy.txt_norm_conc.txt
KatyPerry.txt_orig_norm_conc.txt   psy.txt_norm_disa.txt
KatyPerry.txt_orig_norm_disa.txt   psy.txt_orig.txt
LMFAO.txt_conc.txt                 psy.txt_orig_conc.txt
LMFAO.txt_disa.txt                 psy.txt_orig_disa.txt
LMFAO.txt_norm.txt                 psy.txt_orig_norm.txt
LMFAO.txt_norm_conc.txt            psy.txt_orig_norm_conc.txt
LMFAO.txt_norm_disa.txt            psy.txt_orig_norm_disa.txt
LMFAO.txt_orig.txt
```

Gambar 2. Output Hasil Normalisasi Teks

Hasil normalisasi teks disimpan dalam file berekstensi .txt. Bubuhan pada setiap file menandakan hasil dari rule yang digunakan. Contoh hasil dari normalisasi teks dapat dilihat pada tabel 7.

Tabel 7. Hasil Normalisasi Teks

Original	+447935454150 lovely girl talk to me xxxï»¿
Expansion 1	+447935454150 lovely pin-up female_child fille girl girlfriend lady_friend little_girl miss missy young_lady young_woman public_lecture talk talk_of_the_town talking to me xxxï»¿
Expansion 2	+447935454150 adorable female_child talk to me xxxï»¿
Expansion 3	+447935454150 lovely girl talk to me xxxï»¿
Expansion 4	+447935454150 lovely lovely pin-up girl female_child fille girl girlfriend lady_friend little_girl miss missy young_lady young_woman talk public_lecture talk talk_of_the_town talking to me xxxï»¿
Expansion 5	+447935454150 lovely adorable girl female_child talk to me xxxï»¿
Expansion 6	+447935454150 lovely lovely pin-up girl female_child fille girl girlfriend lady_friend little_girl miss missy young_lady young_woman talk public_lecture talk talk_of_the_town talking to me xxxï»¿
Expansion 7	+447935454150 lovely adorable girl female_child talk to me xxxï»¿
Expansion 8	+447935454150 lovely girl talk to me xxxï»¿
Expansion 9	+447935454150 lovely lovely pin-up girl female_child fille girl girlfriend lady_friend little_girl miss missy young_lady young_woman talk public_lecture talk talk_of_the_town talking to me xxxï»¿
Expansion 10	+447935454150 lovely adorable girl female_child talk to me xxxï»¿

4.2 Hasil Klasifikasi

Pada proses ini dilakukan klasifikasi terhadap 55 dataset yang terdiri dari dataset *original* dan dataset *expansion*. Pada proses klasifikasi ini dataset dibagi menjadi data latih dan data uji yang dibagi secara otomatis menggunakan metode *k-Fold Cross Validation*. Pada penelitian ini digunakan k=10. Pada proses klasifikasi ini akan diperoleh nilai *Accuracy* (Acc), *Spam Caught Rate* (SC), *Blocked Ham Rate* (BH), *F-Measure* (F1) dan *Matthews Correlation Coefficient* (MCC). Pada tabel 8-12 disajikan hasil pengukuran nilai MCC pada setiap dataset, yang mana nilai MCC akan menjadi tolak ukur untuk mengetahui kualitas klasifikasi *binary*.

Nilai MCC yang ditampilkan adalah nilai MCC yang diperoleh dari dataset *original* dan dataset *expansion* dengan nilai MCC terbaik. Nilai MCC yang dihasilkan berkisar antara -1 s.d. 1. Nilai 1 menunjukkan nilai prediksi yang sempurna, 0 menunjukkan kegagalan dalam prediksi dan nilai -1 menunjukkan kegagalan total dalam prediksi [18].

Pada tabel 8 dapat diketahui nilai MCC yang diperoleh dari dataset Psy. Nilai MCC tertinggi sebesar 0.949 pada dataset *expansion* diperoleh dari metode *Logistic Regression.* Hal ini menunjukkan performa yang bagus dihasilkan dari metode *Logistic Regression.*

Tabel 8. Nilai MCC pada Dataset Psy

Metode	Original	Expansion
DT	0.846	0.863
LR	0.932	**0.949**
NB-B	0.9	0.912
NB-G	0.669	0.697
NB-M	0.903	0.899
1-NN	0.458	0.823
3-NN	0.754	0.909
5-NN	0.824	0.937
RF	0.91	0.934
SVM-L	0.91	0.928
SVM-P	0.359	0.585
SVM-R	0.865	0.892

Tabel 9. Nilai MCC pada Dataset KattyPerry

Metode	Original	Expansion
DT	0.876	0.87
LR	0.909	0.898
NB-B	0.469	0.544
NB-G	0.655	0.682
NB-M	0.837	0.877
1-NN	0.404	0.744
3-NN	0.705	0.825
5-NN	0.741	0.841
RF	0.9	0.895
SVM-L	0.838	0.861
SVM-P	0.226	0.514
SVM-R	0.886	**0.914**

Pada tabel 9 diperoleh nilai MCC tertinggi 0.914 pada dataset dari dataset KattyPerry. Nilai tersebut diperoleh dari dataset *expansion* menggunakan metode SVM dengan kernel *gaussian.*

Tabel 10. Nilai MCC pada Dataset LMFAO

Metode	Original	Expansion
DT	0.881	0.891
LR	0.89	**0.909**
NB-B	0.881	0.904
NB-G	0.779	0.789
NB-M	0.723	0.812
1-NN	0.644	0.769
3-NN	0.71	0.803
5-NN	0.748	0.85
RF	0.887	0.903
SVM-L	0.876	0.905
SVM-P	0.693	0.734
SVM-R	0.871	0.904

Pada tabel 10 dihasilkan nilai MCC pada dataset LMFAO. Performa metode klasifikasi terbaik dengan nilai MCC 0.909 diperoleh dari dataset *expansion* dengan metode *Logistic Regression.*

Pada tabel 11 diketahui nilai MCC tertinggi pada dataset Eminem sebesar 0.932. Nilai tersebut diperoleh dari dataset *expansion* menggunakan metode klasifikasi SVM dengan kernel *gaussian.*

Tabel 11. Nilai MCC pada Dataset Eminem

Metode	Original	Expansion
DT	0.876	0.915
LR	0.892	0.928
NB-B	0.847	0.878
NB-G	0.726	0.776
NB-M	0.76	0.854
1-NN	0.487	0.767
3-NN	0.406	0.799
5-NN	0.426	0.83
RF	0.866	0.929
SVM-L	0.851	0.906
SVM-P	0.77	0.847
SVM-R	0.849	**0.932**

Tabel 12. Nilai MCC pada Dataset Shakira

Metode	Original	Expansion
DT	0.797	0.851
LR	0.875	**0.903**
NB-B	0.837	0.826
NB-G	0.741	0.75
NB-M	0.835	0.844
1-NN	0.564	0.594
3-NN	0.462	0.525
5-NN	0.7	0.747
RF	0.851	0.889
SVM-L	0.881	0.893
SVM-P	0.793	0.827
SVM-R	0.875	**0.903**

Pada tabel 12 diperoleh nilai MCC pada dataset Shakira. Nilai MCC tertinggi sebesar 0.903 diperoleh dari dataset *expansion* menggunakan metode SVM-R.

Dari keseluruhan pengukuran nilai MCC pada setiap dataset, dapat diketahui 2 metode yang menghasilkan nilai MCC tertinggi dengan dataset *expansion* yaitu metode SVM dengan kernel *gaussian* dan *Logistic Regression.*

4.3 Hasil Uji Beda *t-test*

Pada tahap ini dilakukan uji beda *t-test* pada nilai MCC yang diperoleh untuk mengetahui adanya pengaruh yang signifikan atau tidak setelah diterapkan metode *TextExpansion.* Hasil uji beda *t-test* pada setiap dataset disajikan pada tabel 13-17.

Tabel 13. Uji Beda *t-test* pada Dataset Psy

	Original	*Expansion*
Mean	0.7775	0.860666667
Variance	0.035799727	0.012287879
Observations	12	12
Pearson Correlation	0.840262467	
Hypothesized Mean Difference	0	
df	11	
t Stat	-2.542428185	
P(T<=t) one-tail	0.013680728	
t Critical one-tail	1.795884819	
P(T<=t) two-tail	0.027361456	
t Critical two-tail	2.20098516	

Tabel 14. Uji Beda t-test pada Dataset KattyPerry

	Original	*Expansion*
Mean	0.703833333	0.78875
Variance	0.050570333	0.019214023
Observations	12	12
Pearson Correlation	0.906664335	
Hypothesized Mean Difference	0	
df	11	
t Stat	-2.554509116	
P(T<=t) one-tail	0.013389603	
t Critical one-tail	1.795884819	
P(T<=t) two-tail	0.026779206	
t Critical two-tail	2.20098516	

Tabel 15. Uji Beda t-test pada Dataset LMFAO

	Original	*Expansion*
Mean	0.798583333	0.84775
Variance	0.008412992	0.004014386
Observations	12	12
Pearson Correlation	0.924142784	
Hypothesized Mean Difference	0	
df	11	
t Stat	-4.147754948	
P(T<=t) one-tail	0.000811302	
t Critical one-tail	1.795884819	
P(T<=t) two-tail	0.001622604	
t Critical two-tail	2.20098516	

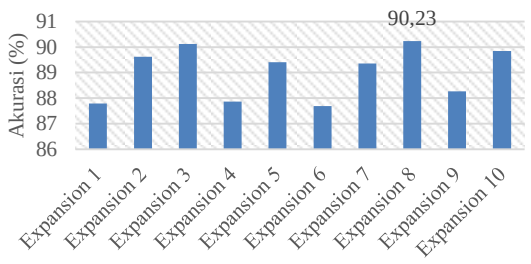
Tabel 16. Uji Beda t-test pada Dataset Eminem

	Original	*Expansion*
Mean	0.729666667	0.863416667
Variance	0.033409333	0.003674992
Observations	12	12
Pearson Correlation	0.78600043	
Hypothesized Mean Difference	0	
df	11	
t Stat	-3.303924851	
P(T<=t) one-tail	0.003514436	
t Critical one-tail	1.795884819	
P(T<=t) two-tail	0.007028871	
t Critical two-tail	2.20098516	

Tabel 17. Uji Beda t-test pada Dataset Shakira

	Original	*Expansion*
Mean	0.767583333	0.796
Variance	0.017624992	0.015173455
Observations	12	12
Pearson Correlation	0.989188546	
Hypothesized Mean Difference	0	
df	11	
t Stat	-4.664553436	
P(T<=t) one-tail	0.000344302	
t Critical one-tail	1.795884819	
P(T<=t) two-tail	0.000688604	
t Critical two-tail	2.20098516	

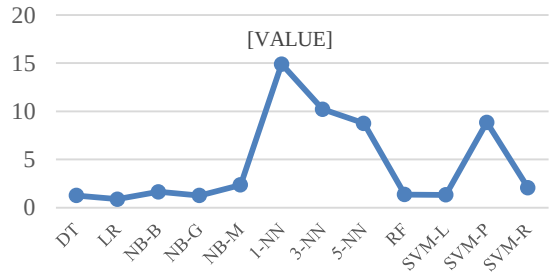
Dari keseluruhan hasil uji beda *t-test* pada setiap dataset diperoleh nilai *P(T<=t) two-tail* < 0.05. Hal ini menunjukkan bahwa adanya pengaruh yang signifikan setelah diterapkan normalisasi teks menggunakan metode *TextExpansion.* Untuk mengetahui aturan *expansion* terbaik, disajikan grafik perbandingan akurasi dari setiap dataset *expansion* pada gambar 3.



Gambar 3. Perbandingan Akurasi Dataset *Expansion*

Dari gambar 3 dapat diketahui nilai akurasi tertinggi 90.23% diperoleh dari dataset *Expansion 8*. Dimana merupakan representasi dari aturan *Original Text + Text Normalization + Concept Generations*. Sehingga dapat disimpulkan aturan *Expansion 8* adalah aturan terbaik pada kasus ini.

Selanjutnya untuk mengetahui metode klasifikasi dengan akurasi terbaik dilakukan penghitungan nilai gap akurasi dari dataset asli dan dataset *expansion*. Nilai gap akurasi dapat dilihat pada gambar 4.



Gambar 4. Grafik gap akurasi

Dari grafik pada gambar 4 dapat diketahui metode KNN dengan *k=1* menunjukkan nilai gap tertinggi sebesar 14.89. Hal ini menunjukkan adanya peningkatan akurasi yang signifikan setelah diterapkan metode *TextExpansion*, dimana nilai akurasi dataset asli sebesar 71.68% dan dataset *expansion* sebesar 86.58%.

5. Kesimpulan

5.1 Simpulan

Berdasarkan hasil penelitian dan serangkaian uji coba yang dilakukan dapat disimpulkan bahwa penerapan normalisasi teks menggunakan metode *TextExpansion* untuk deteksi komentar spam pada Youtube menunjukkan hasil yang baik. Keseluruhan hasil uji beda *t-test* menunjukkan adanya pengaruh yang signifikan setelah diterapkan metode *TextExpansion*. Hal ini ditunjukkan dengan nilai nilai *P(T<=t) two-tail* < 0.05 pada seluruh dataset. Nilai akurasi tertinggi yang dihasilkan adalah 90.23% diperoleh dari dataset *Expansion 8*, dimana aturan tersebut merupakan kombinasi dari *original text + text normalization + concept generations*. Dari sini dapat disimpulkan bahwa aturan *Expansion 8* merupakan aturan terbaik untuk deteksi komentar spam pada Youtube. Sedangkan jika ditinjau dari nilai MCC, metode SMV-R dan *Logistic Regression* menunjukkan kualitas klasifikasi terbaik. Selanjutnya jika ditinjau dari gap akurasi, metode K-NN dengan *k=1* menunjukkan adanya peningkatan yang pesat sebesar 14.89% setelah menerapkan metode *TextExpansion.*

Dari hasil penelitian ini diharapkan dapat menjadi acuan untuk pengembangan tool untuk mendeteksi komentar spam secara otomatis pada Youtube yang lebih baik. Pada penelitian ini belum fokus pada satu metode klasifikasi, pada penelitian selanjutnya dapat dilakukan penelitian yang lebih mendalam pada metode klasifikasi terbaik dikombinasikan dengan metode *TextExpansion* untuk mendapatkan hasil yang lebih baik.

5.2 Saran

Pada penelitian ini tentunya masih banyak kekurangan, maka dari itu peneliti merangkum saran-saran untuk penelitian selanjutnya sebagai berikut:

1. Optimasi tool untuk *TextExpansion* karena masih dibutuhkan waktu yang lama dalam proses *TextExpansion*
2. Optimasi paramater pada setiap dataset Expansion menggunakan *grid search* untuk mendapatkan parameter terbaik pada setiap dataset
3. Mencari metode yang tepat untuk memilih dataset *Expansion* terbaik

Daftar Rujukan

[1] Youtube, “Press - Youtube,” 2018. [Online]. Available: https://www.youtube.com/yt/about/ press/%0D. [Accessed: 02-Mar-2018].

[2] M. Chakraborty, S. Pal, R. Pramanik, and C. Ravindranath Chowdary, “Recent developments in social spam detection and combating techniques: A survey,” *Inf. Process. Manag.*, vol. 52, no. 6, pp. 1053–1073, Nov. 2016.

[3] A. Mehmood, B.-W. On, I. Lee, I. Ashraf, and G. Sang Choi, “Spam comments prediction using stacking with ensemble learning,” *J. Phys. Conf. Ser.*, vol. 933, p. 012012, Jan. 2018.

[4] H. Nguyen, “Research Report 2013 State of Social Media Spam,” 2013.

[5] K. Stuart, “PewDiePie switches off YouTube comments: ‘It’s mainly spam,’” *The Guardian*, 2014. [Online]. Available: https://www.theguardian.com/technology/2014/sep/03/pewdiepie-switches-off-youtube-comments-its-mainly-spam. [Accessed: 02-Mar-2018].

[6] T. C. Alberto, J. V. Lochter, and T. A. Almeida, “TubeSpam: Comment Spam Filtering on YouTube,” in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015, pp. 138–143.

[7] A. Pinandito, R. S. Perdana, M. C. Saputra, and H. M. Az-zahra, “Spam detection framework for Android Twitter application using Naïve Bayes and K-Nearest Neighbor classifiers,” in *Proceedings of the 6th International Conference on Software and Computer Applications - ICSCA ’17*, 2017, pp. 77–82.

[8] M. Alsaleh, A. Alarifi, F. Al-Quayed, and A. Al-Salman, “Combating Comment Spam with Machine Learning Approaches,” in *2015 IEEE 14th International Conference on*

Machine Learning and Applications (ICMLA), 2015, pp. 295–300.

[9] R. M. Silva, T. C. Alberto, T. A. Almeida, and A. Yamakami, "Towards filtering undesired short text messages using an online learning approach with semantic indexing," *Expert Syst. Appl.*, vol. 83, pp. 314–325, Oct. 2017.

[10] T. A. Almeida, T. P. Silva, I. Santos, and J. M. Gómez Hidalgo, "Text normalization and semantic indexing to enhance Instant Messaging and SMS spam filtering," *Knowledge-Based Syst.*, vol. 108, pp. 25–32, Sep. 2016.

[11] I. Idris *et al.*, "A combined negative selection algorithm–particle swarm optimization for an email spam detection system," *Eng. Appl. Artif. Intell.*, vol. 39, pp. 33–44, Mar. 2015.

[12] C.-N. Lee, Y.-R. Chen, and W.-G. Tzeng, "An online subject-based spam filter using natural language features," in *2017 IEEE Conference on Dependable and Secure Computing*, 2017, pp. 479–487.

[13] K. Roy, S. Keshari, and S. Giri, "Enhanced Bayesian spam filter technique employing LCS," in *2016 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, 2016, pp. 1–6.

[14] M. Zavvar, M. Rezaei, and S. Garavand, "Email Spam Detection Using Combination of Particle Swarm Optimization and Artificial Neural Network and Support Vector Machine," *Int. J. Mod. Educ. Comput. Sci.*, vol. 8, no. 7, pp. 68–74, Jul. 2016.

[15] Q. Dang, F. Gao, and Y. Zhou, "Spammer detection based on Hidden Markov Model in micro-blogging," in *2016 12th World Congress on Intelligent Control and Automation (WCICA)*, 2016, pp. 407–412.

[16] S. Sedhai and A. Sun, "Semi-Supervised Spam Detection in Twitter Stream," *IEEE Trans. Comput. Soc. Syst.*, pp. 1–7, 2017.

[17] T. Wu, S. Liu, J. Zhang, and Y. Xiang, "Twitter spam detection based on deep learning," in *Proceedings of the Australasian Computer Science Week Multiconference on - ACSW '17*, 2017, pp. 1–8.

[18] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric," *PLoS One*, vol. 12, no. 6, p. e0177678, Jun. 2017.