OPTIMASI ALGORITMA *K-NEAREST NEIGHBOR* DALAM MENDETEKSI KOMENTAR *SPAM* BERBAHASA INDONESIA PADA INSTAGRAM MENGGUNAKAN *CONVERT NEGATION* DAN TF-IDF *(TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY)* PADA TAHAP *PREPROCESSING*

Skripsi

disusun sebagai salah satu syarat

untuk memperoleh gelar Sarjana Komputer

Program Studi Teknik Informatika

oleh

Nanang Arif Andriyani

4611415002

JURUSAN ILMU KOMPUTER

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS NEGERI SEMARANG

2019

PERNYATAAN

Saya menyatakan bahwa skripsi ini bebas plagiat, dan apabila dikemudian hari terbukti terdapat plagiat dalam skripsi ini, maka saya bersedia menerima sanksi ketentuan peraturan perundang-undangan.

Semarang, 1 November 2019

Nanang Arif Andriyani

4611415002

PENGESAHAN

Skripsi yang berjudul

Optimasi Algoritma *K-Nearest Neighbor* dalam Mendeteksi Komentar *Spam* Berbahasa Indonesia pada Instagram Menggunakan *Convert Negation* dan TF-IDF *(Term Frequency - Inverse Document Frequency)* pada Tahap *Preprocessing*

disusun oleh

Nanang Arif Andriyani

4611415002

telah dipertahankan di hadapan sidang panitia ujian skripsi FMIPA UNNES pada tanggal 1 November 2019.

Panitia:

Ketua

Sugianto, M.Si.

NIP 196102191993031001

Sekretaris

Alamsyah, S.Si., M.Kom.

NIP 197405172006041001

Penguji 1

Alamsyah, S.Si., M.Kom.

NIP 197405172006041001

Penguji 2

Budi Prasetiyo, S.Si., M.Kom.

NIP 198805012014041001

Anggota Penguji

Endang Sugiharti, S.Si., M.Kom.

NIP 197401071999032001

MOTTO

Try not to become a man of success, but rather try to become a man of value.
(Albert Einstein)
The patch to success is to take massive, determined act
(Anthony Robbins)
Expectations are a form of first-class truth. If people believe it, it's true.
(Bill Gates)
Never give up. Today is hard, tomorrow will be worse, but the day after tomorrow will be sunshine.
(Jack Ma)
My Success is only by Allah.
(Q.S Huud: 88)

PERSEMBAHAN

Untuk Ayah, Ibu, Kakak, Keluarga besar,
Dosen Jurusan Ilmu Komputer UNNES,
serta sahabat dan teman-teman.

PRAKATA

Puji dan syukur kami panjatkan ke hadirat Allah SWT, yang telah melimpahkan rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi dengan judul "Optimasi Algoritma *K-Nearest Neighbor* dalam Mendeteksi Komentar *Spam* Berbahasa Indonesia pada Instagram Menggunakan *Convert Negation* dan TF-IDF *(Term Frequency - Inverse Document Frequency)* pada Tahap *Preprocessing*". Skripsi ini disusun guna melengkapi salah satu syarat untuk menyelesaikan Program Studi Teknik Informatika, Jurusan Ilmu Komputer Universitas Negeri Semarang. Atas tersusunnya skripsi ini, penulis mengucapkan terima kasih yang sebesar besarnya kepada :

(1) Bapak Prof. Dr. Fathur Rokhman, M.Hum., Rektor Universitas Negeri Semarang.

(2) Bapak Dr. Sugianto, M.Si., Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Semarang.

(3) Bapak Alamsyah, S.Si., M.Kom., Ketua Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Semarang.

(4) Ibu Endang Sugiharti, S.Si., M.Kom., Dosen Pembimbing yang dengan sabar membimbing, mengarahkan, dan memotivasi penulis dalam penyusunan skripsi ini.

(5) Bapak dan Ibu Dosen Jurusan Ilmu Komputer Universitas Negeri Semarang, yang telah memberikan bekal ilmu yang bermanfaat kepada penulis.

(6) Kedua orang tua yang telah memberikan doa, dukungan dan segalanya kepada penulis baik selama penyusunan skripsi ataupun sepanjang hidup ini.

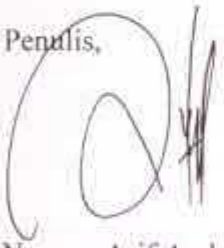
(7) Teman-teman Jurusan Ilmu Komputer UNNES, terutama angkatan 2015 yang telah memberikan bantuan, harapan, motivasi, doa, semangat dan saran-saran dalam penyusunan skripsi ini.

Semoga dengan membaca skripsi ini dapat memberi manfaat bagi kita semua, dalam hal ini dapat menambah wawasan yang bermanfaat.

Atas semua perhatian dari segala pihak yang telah membantu penulis dalam menyusun skripsi ini, penulis ucapkan terima kasih.

Semarang, 1 November 2019

Penulis,



Nanang Arif Andriyani

ABSTRAK

Andriyani, N. A. 2019. *Optimasi Algoritma K-Nearest Neighbor dalam Mendeteksi Komentar Spam Berbahasa Indonesia pada Instagram Menggunakan Convert Negation dan TF-IDF (Term Frequency - Inverse Document Frequency)* pada Tahap *Preprocessing*. Skripsi, Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Semarang Pembimbing Endang Sugiharti, S.Si., M.Kom.

Kata kunci: KNN, *Convert Negation,* TF-IDF, Komentar *Spam*.

Indonesia merupakan negara dengan pengguna Instagram terbesar nomor 3 di dunia. Instagram menyediakan ruang bebas dan terbuka dalam berinteraksi, kemudahan dalam mengunggah foto atau video serta dalam berkomentar. Permasalahannya, banyak sekali komentar *spam* yang ditulis pada Instagram dan sampai saat ini belum ada solusi penyelesaian yang efektif, terutama untuk *spam* berbahasa Indonesia. Dalam penelitian ini dilakukan pengumpulan *dataset* komentar Instagram dari 10 akun publik figur Indonesia dengan *follower* di atas 10 juta sejumlah 500 data, dimana data setelah diolah dapat dimanfaatkan menggunakan aplikasi Instablock untuk memblokir semua *username* yang terindikasi sebagai *spammer*. Pada penelitian ini penulis menggunakan metode *K-Nearest Neighbor,* karena metode ini mudah untuk diimplementasikan, dijalankan dan waktu yang dibutuhkan untuk menjalankan pembelajaran ini relatif cepat serta mudah dimodifikasi. Metode *K-Nearest Neighbor* memberikan tingkat akurasi yang lebih dapat dipercaya dalam klasifikasi dengan menentukan nilai k yang terbaik. Pada penelitian ini terdiri dari 3 tahapan proses analisis sentimen. Tahap pertama yaitu proses *prepocessing* yang terdiri dari *case folding*, *cleansing*, *convert negation*, *stopword removal*, *tokenizations* dan *stemming*, Selanjutnya pada tahap kedua yaitu proses perhitungan bobot pada setiap kata menggunakan metode TF-IDF (*Term Frequency – Inverse Document Frequency*). Tahap terakhir yaitu proses klasifikasi komentar menjadi 2 kategori yaitu *spam* dan *ham* menggunakan metode *K-Nearest Neighbor*. Hasil pengujian komentar Instagram berbahasa Indonesia dari 10 kali percobaan dengan nilai k adalah angka genap mulai dari 2 sampai 20 menggunakan metode *K-Nearest Neighbor* tanpa *Convert Negation* dan TF-IDF pada tahap *preprocessing* menghasilkan rata-rata akurasi sebesar 88,45%. Sedangkan hasil pengujian metode *K-Nearest Neighbor* menggunakan *Convert Negation* dan TF-IDF pada tahap *preprocessing* menghasilkan rata-rata akurasi sebesar 95,75%. Dari hasil penelitian tersebut dapat disimpulkan bahwa penambahan metode *Convert Negation* dan TF-IDF pada tahap *preprocessing* dapat meningkatkan akurasi sebesar 7,3%.

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

DAFTAR LAMPIRAN

Lampiran Halaman

BAB I

PENDAHULUAN

1.1 Latar Belakang

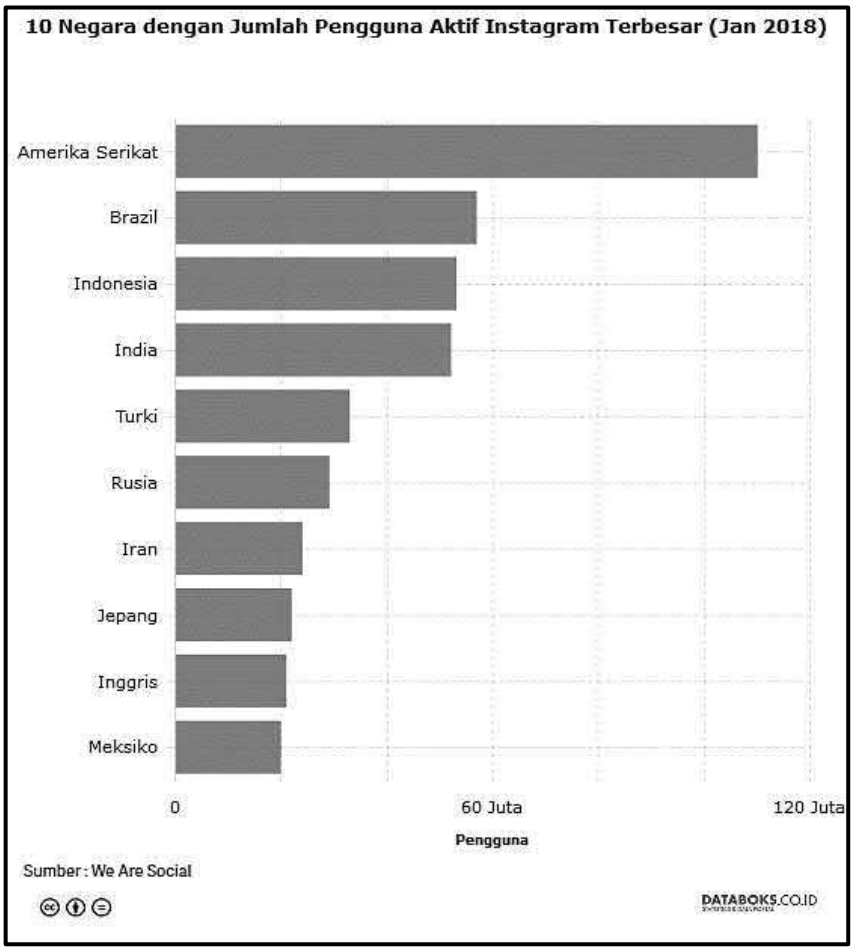
Media sosial saat ini telah menjadi *trend* dalam komunikasi pemasaran yang dapat digunakan oleh seluruh masyarakat di dunia. Media sosial sebagai "sebuah kelompok aplikasi berbasis internet yang membangun di atas dasar ideologi dan teknologi Web 2.0, dan yang memungkinkan penciptaan dan pertukaran *user-generated content*" (Kaplan & Haenlein, 2010).

Beberapa media sosial publik yang dapat digunakan pengguna supaya dapat diakui eksistensinya oleh masyarakat luas diantaranya, Instagram, Facebook, Line, atau Twitter. Karena sosial media ini menyediakan ruang bebas dan terbuka dalam berinteraksi. Sehingga banyaknya *update* status serta unggahan adalah salah satu bentuk pengguna media sosial supaya ingin dikenal secara luas.

Media sosial digunakan oleh pengguna internet untuk tetap eksis sekaligus bersosialisasi di dunia maya. Para publik figur, seperti politikus dan artis/aktor Indonesia banyak menggunakan media sosial seperti Facebook, Twitter, Instagram, Path, dan lain-lain. Jika Facebook dan Twitter lebih banyak menggunakan teks sebagai statusnya, Instagram dan Path menggunakan foto dan *caption* foto sebagai statusnya.

Menurut penelitian yang dilakukan We Are Social, perusahaan media asal Inggris yang bekerja sama dengan Hootsuite, rata-rata orang Indonesia

menghabiskan 3 jam 23 menit dalam sehari untuk mengakses media sosial. Dari laporan berjudul "*Essential Insights Into Internet, Social Media, Mobile, and E-Commerce Use Around The World*" yang diterbitkan tanggal 30 Januari 2018, dari total populasi Indonesia sebanyak 265,4 juta jiwa, pengguna aktif media sosialnya mencapai 130 juta dengan penetrasi 49 persen. Hasil survei menunjukkan bahwa Indonesia merupakan negara dengan pengguna Instagram terbesar nomor 3 di dunia. Data statistik pengguna aktif Instagram terbesar bulan Januari 2018 diperlihatkan pada Gambar 1.1.



Gambar 1.1 Daftar 10 Negara dengan Jumlah Pengguna Aktif Instagram

Salah satu hal yang menyebabkan Instagram banyak digunakan adalah kemudahannya untuk mengunggah foto langsung dari *smartphone*. Namun di samping kelebihan tersebut tentu terdapat kekurangan yang dapat mengganggu yaitu banyaknya komentar yang dapat dikategorikan sebagai komentar *spam* terhadap suatu unggahan foto yang diunggah pada IG. Komentar *spam* akan semakin banyak terhadap IG artis/orang terkenal karena *follower*-nya juga semakin banyak. Contoh komentar *spam* pada salah satu foto milik @ayutingting29 diperlihatkan pada Gambar 1.2.

Gambar 1.2 Contoh Komentar *Spam* di Instagram

Beberapa solusi menghadapi komentar *spam* sudah ada, namun semuanya dilakukan secara manual. Pengguna Instagram dapat menghapus secara manual komentar *spam* tersebut namun jelas-jelas membutuhkan waktu yang besar dan harus diperiksa satu persatu (D. Tamir, 2015). Selain dihapus secara manual Instagram juga menyediakan fitur untuk melaporkan semua komentar sebagai

spam secara manual juga, artinya harus dilakukan satu persatu. Hal berikutnya untuk meminimalisasi komentar *spam* adalah dengan mengubah akun Instagram menjadi privat. Hal ini tentu sulit dilakukan bagi akun publik figur, karena jika akun Instagram dibuat menjadi privat tidak bisa langsung di *follow* oleh akun lain. Hal terakhir yang dapat dilakukan adalah menggunakan pengaturan mengaktifkan fitur Instagram untuk menghapus komentar yang mengandung kata-kata tertentu yang dimasukkan sendiri oleh pengguna yang dianggap *spam*. Semua solusi tersebut hanya bisa digunakan dalam bahasa Inggris dan tidak dapat diterapkan dalam bahasa Indonesia.

Berdasarkan latar belakang tersebut pada penelitian ini akan dibangun suatu sistem yang dapat mengklasifikasikan komentar *spam* berbahasa Indonesia dengan mengambil *data training* komentar-komentar *spam* pada Instagram beberapa artis terkenal Indonesia. Terdapat beberapa metode untuk klasifikasi seperti *Naive Bayes, K-Nearest Neighbour, Decision Tree, Support Vector Machine,* atau *K-Means Clustering.* Metode klasifikasi yang digunakan dalam penelitian ini adalah *K-Nearest Neighbor*. *Metode K-Nearest Neighbor* menggunakan konsep meminimalkan variasi antar data yang ada di dalam suatu *cluster* dan memaksimalkan variasi dengan data yang ada di *cluster* lainnya. Metode ini dipilih karena mudah untuk diimplementasikan dan dijalankan dan waktu yang dibutuhkan untuk menjalankan pembelajaran ini relatif cepat serta mudah untuk diadaptasi. Metode *K-Nearest Neighbor* memberikan tingkat akurasi yang lebih dapat dipercaya dalam klasifikasi dengan memilih nilai *k* yang terbaik (Kavita Mittal, 2017).

Hal inilah yang menjadi latar belakang peneliti dalam melakukan penelitian pada skripsi yang berjudul "OPTIMASI ALGORITMA *K-NEAREST NEIGHBOR* DALAM MENDETEKSI KOMENTAR *SPAM* BERBAHASA INDONESIA PADA INSTAGRAM MENGGUNAKAN *CONVERT NEGATION* DAN TF-IDF *(TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY)* PADA TAHAP *PREPROCESSING)"*

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah:

1) Apakah algoritma *K-Nearest Neighbor* dapat mengidentifikasi komentar *spam* berbahasa Indonesia pada Instagram?
2) Bagaimana perbandingan dan peningkatan hasil akurasi pada algoritma *K-Nearest Neighbor* dengan algoritma *K-Nearest Neighbor* yang dioptimasi menggunakan *Convert Negation* dan TF-IDF pada tahap *preprocessing* dalam identifikasi komentar *spam* berbahasa Indonesia pada Instagram?
3) Bagaimana perbandingan akurasi algoritma *K-Nearest Neighbor* menggunakan *Convert Negation* dan TF-IDF pada tahap *preprocessing* dengan penelitian terkait?

1.3 Batasan Masalah

Pada penelitian ini diperlukan batasan-batasan agar tujuan penelitian dapat tercapai. Adapun batasan masalah yang dibahas pada penelitian ini adalah:

1) Data latih dan data uji teks komentar *spam* yang digunakan dalam sistem adalah berbahasa Indonesia.
2) Sistem yang dibangun mengidentifikasi komentar yang dinilai sebagai *ham* dan *spam*.
3) Data yang digunakan adalah *file* dokumen berekstensi *.txt* yang diambil dengan pengumpulan data komentar dari 10 akun Instagram publik figur di Indonesia.

1.4 Tujuan Penelitian

Tujuan perancangan dan pembangunan aplikasi ini adalah sebagai berikut:

1) Mengetahui apakah algoritma *K-Nearest Neighbor* dapat mengidentifikasi komentar *spam* berbahasa Indonesia pada Instagram.
2) Mengetahui tingkat akurasi algoritma *K-Nearest Neighbor* menggunakan *Convert Negation* dan TF-IDF pada tahap *preprocessing* dalam identifikasi komentar *spam* berbahasa Indonesia pada Instagram.

1.5 Manfaat Penelitian

Adapun manfaat penelitian ini adalah sebagai berikut.

1) Mengetahui proses kerja algoritma *K-Nearest Neighbor* dalam identifikasi komentar *spam* berbahasa Indonesia pada Instagram.
2) Dalam lingkungan akademis diperoleh pengetahuan terhadap akurasi algoritma *K-Nearest Neighbor* dalam melakukan identifikasi komentar *spam* bahasa Indonesia pada Instagram.

3) Membantu pengguna Instagram dalam menyaring *spam* serta memblokir akun yang terindikasi sebagai *spammer* dengan bantuan aplikasi Instablocks.

1.6 Sistematika Penulisan

Sistematika penulisan untuk memudahkan dalam memahami alur pemikiran secara keseluruhan skripsi. Penulisan skripsi ini secara garis besar dibagi menjadi tiga bagian yaitu sebagai berikut.

1.6.1 Bagian Awal Skripsi

Bagian awal skripsi terdiri dari halaman judul, halaman pengesahan, halaman pernyataan, halaman motto dan persembahan, abstrak, kata pengantar, daftar isi, daftar gambar, daftar tabel dan daftar lampiran.

1.6.2 Bagian Isi Skripsi

Bagian isi skripsi terdiri dari lima bab yaitu sebagai berikut.

1) BAB 1: PENDAHULUAN

   Bab ini terdiri atas latar belakang, rumusan masalah, batasan masalah, tujuan dan manfaat penelitian serta sistematika skripsi.

2) BAB 2: TINJAUAN PUSTAKA

   Bab ini terdiri atas landasan teori yang berhubungan dengan topik skripsi dan penelitian terkait.

3) BAB 3: METODE PENELITIAN

Bab ini terdiri atas studi pendahuluan, tahap pengumpulan dan pengumpulan data, studi pustaka, teknik analisis data, analisis kebutuhan, dan pengambilan kesimpulan.

4) BAB 4: HASIL DAN PEMBAHASAN

Bab ini terdiri atas hasil penelitian dan pembahasan penelitian.

5) BAB 5: PENUTUP

Bab ini terdiri atas simpulan dan saran.

1.6.3 Bagian Akhir Skripsi

Bagian akhir skripsi berisi daftar pustaka yang merupakan informasi mengenai buku-buku, sumber-sumber dan referensi yang digunakan penulis serta lampiran-lampiran yang mendukung dalam penulisan skripsi ini.

BAB II

TINJAUAN PUSTAKA

2.1 *Text Mining*

Text mining adalah proses mengambil informasi dari teks. Informasi biasanya diperoleh melalui peramalan pola dan kecenderungan pembelajaran pola statistik. *Text mining* yaitu *parsing*, bersama dengan penambahan beberapa fitur linguistik turunan dan penghilangan beberapa diantaranya, dan penyisipan *subsequent* ke dalam *database*, menentukan pola dalam data terstruktur, dan akhirnya mengevaluasi dan menginterpretasi *output*, *text mining* biasanya mengacu ke beberapa kombinasi relevansi, kebaruan, dan *interestingness*.

Kunci dari proses pada *text mining* adalah menggabungkan informasi yang berhasil diekstraksi dari berbagai sumber (Hearst, 2003). Sedangkan menurut (Harlian, 2006) *text mining* didefinisikan sebagai data yang berupa teks yang biasanya sumber data didapatkan dari dokumen, dengan tujuan adalah mencari kata-kata yang dapat mewakili isi dari dokumen tersebut yang nantinya dapat dilakukan analisa hubungan antar dokumen. Proses *text mining* yang khas meliputi kategorisasi teks, *text clustering*, ekstraksi konsep/entitas, produksi taksonomi *granular*, penyimpulan dokumen, identifikasi komentar *spam* dan pemodelan relasi entitas yaitu, pembelajaran hubungan antara entitas (Bridge, 2011).

Pendekatan manual *text mining* secara intensif dalam laboratorium pertama muncul pada pertengahan 1980-an, namun kemajuan teknologi telah

memungkinkan ranah tersebut untuk berkembang selama dekade terakhir. *Text mining* adalah bidang interdisipliner yang mengacu pada pencarian informasi, pertambangan data, pembelajaran mesin, statistik, dan komputasi linguistik. Dikarenakan kebanyakan informasi (perkiraan umum mengatakan lebih dari 80%) saat ini disimpan sebagai teks, *text mining* diyakini memiliki potensi nilai komersial tinggi (Bridge, 2011).

Saat ini, *text mining* telah mendapat perhatian dalam berbagai bidang (Sumartini, 2011):

1. Aplikasi Keamanan.

Banyak paket perangkat lunak *text mining* dipasarkan terhadap aplikasi keamanan, khususnya analisis *plaintext* seperti berita Internet. Hal ini juga mencakup studi enkripsi teks.

2. Aplikasi Biomedis

Berbagai aplikasi *text mining* dalam literatur biomedis telah disusun. Salah satu contohnya adalah PubGene yang mengkombinasikan *text mining* biomedis dengan visualisasi jaringan sebagai sebuah layanan Internet. Contoh lain *text mining* adalah GoPubMed.org. Kesamaan semantik juga telah digunakan oleh sistem *text mining*, yaitu, GOAnnotator.

3. Perangkat Lunak dan Aplikasi

Departemen riset dan pengembangan perusahaan besar, termasuk IBM dan Microsoft, sedang meneliti teknik *text mining* dan mengembangkan program untuk lebih mengotomatisasi proses pertambangan dan analisis. Perangkat lunak *text mining* juga sedang diteliti oleh perusahaan yang berbeda yang bekerja di bidang

pencarian dan pengindeksan secara umum sebagai cara untuk meningkatkan performansinya.

4. Aplikasi Media Online

Text mining sedang digunakan oleh perusahaan media besar, seperti perusahaan Tribune, untuk menghilangkan ambigu informasi dan untuk memberikan pembaca dengan pengalaman pencarian yang lebih baik, yang meningkatkan loyalitas pada situs dan pendapatan. Selain itu, editor diuntungkan dengan mampu berbagi, mengasosiasi dan properti paket berita, secara signifikan meningkatkan peluang untuk menguangkan konten.

5. Aplikasi Pemasaran

Text mining juga mulai digunakan dalam pemasaran, lebih spesifik dalam analisis manajemen hubungan pelanggan yang menerapkan model analisis prediksi untuk churn pelanggan (pengurangan pelanggan).

6. *Sentiment Analyst*

Sentimen analysis mungkin melibatkan analisis dari *review* film untuk memperkirakan berapa baik *review* untuk sebuah film. Analisis semacam ini mungkin memerlukan kumpulan data berlabel atau label dari efektifitas kata-kata. Sebuah sumber daya untuk efektivitas kata-kata telah dibuat untuk WordNet.

7. Aplikasi Akademik

Masalah *text mining* penting bagi penerbit yang memiliki *database* besar untuk mendapatkan informasi yang memerlukan pengindeksan untuk pencarian. Hal ini terutama berlaku dalam ilmu sains, di mana informasi yang sangat spesifik sering terkandung dalam teks tertulis. Oleh karena itu, inisiatif telah diambil seperti

Nature's proposal untuk Open Text Mining Interface (OTMI) dan *Health's common Journal Publishing* untuk *Document Type Definition* (DTD) yang akan memberikan isyarat semantik pada mesin untuk menjawab pertanyaan spesifik yang terkandung dalam teks.

2.2 Media Sosial

Media *online* adalah segala jenis atau format media yang hanya dapat diakses melalui internet, yang dapat berisi teks, foto, video, atau suara. Dalam pengertian umum ini, media *online* juga dapat diartikan sebagai sarana komunikasi *online* (Sugiharti *et al.,* 2018). Pengertian media sosial adalah media *online* yang dimanfaatkan sebagai sarana pergaulan sosial secara *online* di internet. Di media sosial, para penggunanya dapat saling berkomunikasi, berinteraksi, berbagi, *networking*, dan berbagai kegiatan lainnya. Media sosial menggunakan teknologi berbasis website atau aplikasi yang dapat mengubah suatu komunikasi ke dalam bentuk dialog interaktif. Beberapa contoh media sosial yang banyak digunakan adalah Youtube, Facebook, Blog, Twitter, Instagram, dan lain-lain.

Media sosial adalah media berbasis Internet yang memungkinkan pengguna berkesempatan untuk berinteraksi dan mempresentasikan diri, baik secara seketika ataupun tertunda, dengan khalayak luas maupun tidak yang mendorong nilai dari *user-generated content* dan persepsi interaksi dengan orang lain (Caleb T. Carr dan Rebecca A. Hayes (2015).

2.3 Instagram

Instagram adalah sebuah aplikasi berbagi foto dan video yang memungkinkan pengguna mengambil foto, mengambil video, menerapkan filter digital, dan membagikannya ke berbagai layanan jejaring sosial, termasuk milik Instagram sendiri. Satu fitur yang unik di Instagram adalah memotong foto menjadi bentuk persegi, sehingga terlihat seperti hasil kamera kodak instamatik dan polaroid. Hal ini berbeda dengan rasio aspek 4:3 atau 16:9 yang umum digunakan oleh kamera pada peranti bergerak.

Menurut Bambang, Instagram adalah sebuah aplikasi dari *smartphone* yang khusus untuk media sosial yang merupakan salah satu dari media digital yang mempunyai fungsi hampir sama dengan Twitter, namun perbedaannya terletak pada pengambilan foto dalam bentuk atau tempat untuk berbagi informasi terhadap penggunanya. Instagram juga dapat memberikan inspirasi bagi penggunanya dan juga dapat meningkatkan kreativitas, karena Instagram mempunyai fitur yang dapat membuat foto menjadi lebih indah, lebih artistik dan menjadi lebih bagus (Atmoko, 2012:10).

2.4 *Spam*

Secara umum *spam* adalah cara pemanfaatan peralatan elektronik yang digunakan untuk mengirimkan informasi atau pesan berupa tulisan, gambar, video atau bentuk yang lainya kepada orang lain secara terus-menerus tanpa dimintai, diketahui, atau tanpa ijin oleh penerima pesan. *Spam* bisa terjadi jika ada

penyebabnya, seperti dua yang satu mengirim pesan, yang satunya lagi menerima pesan tanpa batas.

Spam memiliki beberapa bentuk diantaranya:

1. *Spam* Jenis Pesan Singkat (SMS)

Spam jenis ini dikirim melalui pengirim pesan kepada penerima berupa pesan singkat atau SMS. Mungkin ada yang pernah menerima hal semacam ini. *Spam* jenis ini biasanya berisi tentang tawaran iklan atau jasa atau apapun melalui telepon genggam.

2. *Spam* Jenis Email

Spam jenis ini hanya untuk orang yang aktif di email saja. Semakin kita sering aktif di email semakin banyak *spam* yang akan kita dapat, akan tetapi email memiliki fasilitas yang akan menyaring *spam* secara otomatis dan hanya berita penting yang akan diterima.

3. *Spam* Jenis *Mailing List* (Milis)

Jika anda menjadi anggota *mailing list* pasti pesan akan dikirim secara langsung, dan disitulah disusupi *spam*.

4. *Spam* Jenis *Search Engine* (Mesin Pencari)

Search engine merupakan situs yang sering dikunjungi oleh semua orang, seperti Yahoo, Google, Bing, dan lain-lain di mana para *spammers* dapat mengirim *spam* kemanapun dia inginkan.

5. *Spam* Jenis Blog

Spam jenis ini adalah sebuah situs web yang berisi berbagai informasi. *Spam* ini dianggap merugikan bagi orang lain karena biasanya orang akan terkecoh

dengan isinya. *Spam* ini tidak hanya melalui isi blog tetapi juga komentar-komentar yang ada. Biasanya banyak yang berkomentar dalam suatu artikel sampai melewati batas. Tapi meskipun begitu pemilik blog apakah komentar tersebut akan dihapus atau dipublikasikan.

6. *Spam* Jenis Iklan Baris

Jika anda membuka sebuah situs atau website pasti anda akan melihat iklan di situs tersebut. *Spam* inilah yang biasanya muncul. *Spam* ini berisi tentang produk-produk, jasa atau hal yang lainya yang disertai juga dengan biaya.

7. *Spam* Jenis Media Sosial

Maraknya pengguna Facebook, Twitter, Instagram, dan lain-lain, mengakibatkan banyak orang yang ingin mengirim pesan ke orang lain dalam bentuk *personal message* ataupun komentar pada sebuah *post*, baik yang dikenal maupun tidak. Kesempatan inilah yang membuat *spammers* gencar mengirim pesan, apalagi sekarang ini kita bisa mendapatkan uang melaui Facebook dengan mengirimkan pesan ke orang lain.

2.5 *Text Prepocessing*

Text Preprocessing adalah suatu proses pengubahan bentuk data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan untuk proses *mining* yang lebih lanjut (*sentiment analyst*, peringkasan, *clustering* dokumen, dsb.). Singkatnya, *preprocessing* adalah mengubah teks menjadi *term index*. Tujuannya adalah untuk menghasilkan sebuah set *term index* yang bisa mewakili dokumen.

Langkah-langkah dalam pemrosesan teks dalam penelitian ini antara lain *case folding, cleansing, convert negation, stopwords removal, tokenization* dan *stemming*, kemudian diberikan pembobotan dengan TF-IDF *(Term Frequency - Inverse Document Frequency).*

2.6 TF-IDF *(Term Frequency - Inverse Document Frequency)*

Metode TF-IDF merupakan metode untuk menghitung bobot dari kata yang digunakan pada *information retrieval*. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat. Metode ini akan menghitung nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada setiap token (kata) disetiap dokumen dalam korpus.

Term Frequency (TF) adalah jumlah kemunculan kata pada suatu dokumen. Semakin banyak suatu kata muncul pada dokumen, maka semakin besar kata tersebut berpengaruh pada dokumen tersebut. Sebaliknya, semakin sedikit suatu kata muncul pada dokumen, maka semakin kecil kata tersebut berpengaruh pada dokumen tersebut.

Inverse Document Frequency (IDF) adalah pembobotan kata yang didasarkan pada banyaknya dokumen yang mengandung kata tertentu. Semakin banyak dokumen yang mengandung suatu kata tertentu, semakin kecil pengaruh kata tersebut pada dokumen. Sebaliknya, semakin sedikit dokumen yang mengandung suatu kata tertentu, semakin besar pengaruh kata tersebut pada dokumen (Feldman & Sanger, 2007).

2.7 *Classification*

Klasifikasi merupakan suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasikan dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Aturan tersebut digunakan pada data - data baru untuk diklasifikasikan. Teknik ini menggunakan *supervised induction*, yang memanfaatkan kumpulan pengujian dari *record* yang terklasifikasi untuk menentukan kelas – kelas tambahan (Hafizh, 2019). Dalam penelitian ini teknik klasifikasi yang digunakan yaitu menggunakan algoritma *K-Nearest Neighbor.*

2.7.1 Definisi *K-Nearest Neighbor*

K-NN merupakan salah satu algoritma pembelajaran mesin sederhana. Hal ini hanya didasarkan pada gagasan bahwa suatu objek yang dekat satu sama lain juga akan memiliki karakteristik yang mirip. Ini berarti jika kita mengetahui ciri-ciri dari salah satu objek, maka kita juga dapat memprediksi objek lain berdasarkan tetangga terdekatnya. K-NN adalah improvisasi lanjutan dari teknik klasifikasi

Nearest Neighbor. Hal ini didasarkan pada gagasan bahwa setiap contoh baru dapat diklasifikasikan oleh suara mayoritas dari k tetangga, di mana k adalah bilangan bulat positif, dan biasanya dengan jumlah kecil (Khamis *et al.*, 2014). Algoritma klasifikasi K-NN memprediksi kategori tes sampel sesuai dengan sampel pelatihan k yang merupakan tetangga terdekat dengan sampel uji, dan memasukkan ke dalam kategori yang memiliki kategori probabilitas terbesar (Suguna dan Thanushkodi, 2010).

Dalam pengenalan pola, algoritma KNN adalah metode yang digunakan untuk mengklasifikasikan objek berdasarkan contoh pelatihan terdekat di ruang fitur. KNN adalah jenis *insctance-based learning*, atau *lazy learning* dimana fungsi ini hanya didekati secara lokal dan semua perhitungan ditangguhkan sampai klasifikasi (Imandoust dan Bolandraftar, 2013).

2.7.2 Algoritma *K-Nearest Neighbor*

Penentuan *k-values* yang tepat diperlukan agar didapatkan akurasi yang tinggi dalam proses kategorisasi dokumen uji. Algoritma *K-Nearest Neighbor* melakukan tahap dalam penentuan *k-values*. Dimana penetapan *k-values* tetap dilakukan, hanya saja tiap-tiap kategori memiliki *k-values* yang berbeda. Perbedaan *k-values* yang dimiliki pada setiap kategori disesuaikan dengan besar-kecilnya jumlah dokumen latih yang dimiliki kategori tersebut. Sehingga ketika *k-values* semakin tinggi, hasil kategori tidak terpengaruh pada kategori yang memiliki jumlah dokumen latih yang lebih besar. Untuk menghitung *similarity* dokumen menggunakan metode *Cosine Similarity* (CosSim). Dipandang sebagai pengukuran

(*similarity measure*) antara *vector document* (D) dengan *vector query* (Q). Semakin sama suatu *vector document* dengan *vector query* maka dokumen dapat dipandang semakin sesuai dengan *query*. Rumus yang digunakan untuk menghitung *cosine similarity* adalah sebagai berikut:

$$cosSim(x, dj) = \frac{\sum_{i=1}^{m} x_i . dj_i}{\sqrt{(\sum_{i=1}^{m} x_i)^2} . \sqrt{(\sum_{i=1}^{m} dj_i)^2}} \quad (2.1)$$

Keterangan:

x : dokumen uji

dj : dokumen latih

x_i dan dj_i : nilai bobot yang diberikan pada setiap *term* pada dokumen.

Kedekatan *query* dan dokumen diindikasikan dengan sudut yang dibentuk. Nilai *cosinus* yang cenderung besar mengindikasikan bahwa dokumen cenderung sesuai *query*. Dalam proses membandingkan dokumen yang sesuai dengan dokumen yang telah ada atau dokumen lainnya, maka digunakan perhitungan dengan rumus pada persamaan (2.1) untuk mengetahui angka similaritas dari dokumen tersebut.

Perhitungan penetapan *k-values* pada algoritma *K-Nearest Neighbor* dilakukan dengan menggunakan persamaan (2.2) dengan terlebih dahulu mengurutkan secara menurun hasil perhitungan similaritas pada setiap kategori. Selanjutnya pada algoritma *K-Nearest Neighbor*, *k-values* yang baru disebut dengan n. Persamaan (2.2) menjelaskan mengenai proporsi penetapan *k-values* (n) pada setiap kategori.

$$n = \left[\frac{k*N(cm)}{maks\{N(Cm)|j=1...Nc\}}\right] \quad (2.2)$$

Keterangan:

n : *k-values* baru

k : *k-values* yang ditetapkan

N(Cm) : jumlah dokumen latih di kategori/kategori m

maks{N(Cm)|j=1...Nc} : jumlah dokumen latih terbanyak pada semua kategori

Dalam menentukan kategori untuk dokumen uji menggunakan algoritma *K-Nearest Neighbors*, maka dilakukan perbandingan kemiripan dokumen uji dengan dokumen latih pada tiap kategori. Persamaan (2.3) menyatakan nilai maksimum perbandingan antara kemiripan dokumen *X* dengan dokumen latih *dj* sejumlah top n tetangga pada suatu kategori dengan kemiripan dokumen X dengan dokumen latih *dj* sejumlah top *n* tetangga pada *training set*.

$$p(x,c_m) = argmax_m \frac{\sum_{dj \in top\ n\ kNN(cm)} sim(x,dj)\ y(dj,cm)}{\sum_{dj \in top\ n\ kNN(cm)} sim(x,dj)} \quad (2.3)$$

Keterangan:

$p(x,c_m)$: probabilitas dokumen *X* menjadi anggota kategori c_m

$sim(x,d_j)$: kemiripan antara dokumen X dengan dokumen latih d_j

top n kNN : top n tetangga

$y(d_j,c_m)$: fungsi atribut dari kategori yang memenuhi persamaan

Adapun langkah-langkah untuk klasifikasi dokumen X menggunakan algoritma *K-Nearest Neighbor* adalah sebagai berikut:

1. Melakukan tahapan *pre-prosesing* sehingga didapatkan representasi dari dokumen X dan semua dokumen latih.
2. Hitung bobot masing-masing dokumen menggunakan TF-IDF.
3. Hitung nilai *cosine similarity* dokumen X dengan semua dokumen latih.
4. Urutkan hasil dari perhitungan nilai *cosine similarity* secara menurun. Nilai yang lebih tinggi menunjukkan bahwa di antara dokumen X dan dokumen latih tersebut memiliki kemiripan.
5. Kelompokkan hasil dari perhitungan nilai *cosine similarity* berdasarkan kategorinya.
6. Tentukan *k-values* kemudian melakukan perhitungan penetapan *k-values* baru (n) pada masing-masing kategori menggunakan persamaan (2.2)
7. Setelah didapatkan nilai *n* yang menyatakan sebagai top tetangga dari langkah 6, maka langkah selanjutnya adalah menentukan kategori dokumen X berdasarkan hasil perhitungan menggunakan persamaan (2.3).
8. Berdasarkan perhitungan pada persamaan (2.3), maka dokumen X akan dikategorikan ke dalam kategori yang memiliki *P(x,cm)* terbesar.

2.8 *Validation Method*

2.8.1 *Cross Validation*

Data *mining* merupakan proses analisis dan eksplorasi. Peta klasifikasi data menjadi kelompok-kelompok atau kelas yang telah ditetapkan. Pada penelitian ini digunakan algoritma *K-Nearest Neighbor classifier* untuk melakukan klasifikasi data pemasaran langsung. Untuk mengukur akurasi dari algoritma *k-nearest neihgbor*, digunakan metode *cross validation* yang melibatkan estimasi akurasi dengan baik. Hasil dari penelitian ini menunjukkan bahwa hasil akurasi klasifikasi dan prediksi data pemasaran langsung dengan algoritma *K-Nearest Neighbor* relatif tinggi (Govindrajan dan Chandrasekaran, 2010).

Cross validation merupakan pengujian standar yang dilakukan untuk memprediksi *error*. *Data training* dibagi secara *random* ke dalam beberapa bagian dengan perbandingan yang sama kemudian *error* dihitung bagian demi bagian, selanjutnya hitung rata-rata seluruh *error* untuk mendapatkan *error rate* secara keseluruhan (Sumarlin, 2015).

Evaluasi digunakan untuk mengukur kinerja metode klasifikasi, dalam penelitian ini digunakan untuk mengukur keakuratan metode klasifikasi yang diukur dengan akurasi, *precision* dan *recall*. *Recall* didefinisikan sebagai persentase antara data kelas data buruk yang dikelaskan dengan benar dan data kelas data buruk yang salah diprediksi ke kelas data baik. *Precision* adalah persentase dari kelas data buruk yang dikelaskan dengan benar dan kelas yang seharusnya termasuk kelas data baik tetapi dikelaskan sebagai kelas data buruk (Sumarlin, 2015).

Adapun perhitungan dalam memperoleh nilai akurasi dapat dilakukan dengan menggunakan persamaan (Hafizh, 2019):

$$\text{Akurasi} = \frac{Jumlah\ klasifikasi\ benar}{Jumlah\ data\ uji}\ x\ 100\% \quad (2.3)$$

2.9 Penelitian Terkait

Penelitian ini dikembangkan dari beberapa referensi yang mempunyai keterkaitan dengan metode dan objek penelitian. Penggunaan referensi ini ditujukan untuk memberikan batasan-batasan terhadap metode dan sistem yang nantinya akan dikembangkan lebih lanjut. Berikut adalah hasil dari penelitian sebelumnya.

Karakasli *et al.,* (2019) melakukan penelitian yang berjudul "*Dynamic Feature Selection for Spam Detection in Twitter"*. Penelitian ini menggunakan *dataset* yang diperoleh dari *CRAWLER Software*, dengan variabel komentar *spam* pada twitter yang diolah menggunakan algoritma *k-Nearest Neighboor*. Dalam penelitian ini diperoleh hasil akurasi sebesar 87.6%.

Kumar *et al.,* (2019) melakukan penelitian yang berjudul "*Empirical Analysis of Supervised Machine Learning Techniques for Cyberbullying Detection".* Penelitian ini menggunakan *dataset* yang diperoleh dari *Social Networking Website Youtube* menggunakan API, dengan variabel komentar *spam* pada Youtube yang diolah menggunakan algoritma *k-Nearest Neighboor* serta dikenai metode validasi *k-Fold Cross Validation* dengan nilai k=20. Dalam penelitian ini diperoleh hasil akurasi sebesar 83%.

Fitri Febriyani *et al.,* (2018) melakukan penelitian yang berjudul *"Sentiment Analysis on the Level of Customer Satisfaction to Data Cellular Services Using the*

Naive Bayes Classifier Algorithm". Penelitian ini menggunakan *dataset* yang diperoleh dari data servis telekomunikasi operator untuk akses internet di Indonesia, dengan variabel tingkat kepuasan pelanggan pada pelayanan data seluler yang diolah menggunakan algoritma *Naive Bayes Classifier* dan *Convert Negation.* Dalam penelitian ini diperoleh hasil akurasi sebesar 99,66%.

Goyal *et al.,* (2016) melakukan penelitian yang berjudul *"Spam Detection Using KNN and Decision Tree Mechanism in Social Network".* Penelitian ini menggunakan *dataset* yang diperoleh dari *Social Networking Website Twitter* menggunakan API, dengan variabel komentar *spam* pada Twitter yang diolah menggunakan algoritma *k-Nearest Neighboor* dan algoritma *Decision Tree*. Dalam penelitian ini dinyatakan bahwa dengan menggunakan algoritma *K-Nearest Neighboor* lebih menghasilkan hasil yang optimal dalam *mining text* komentar *spam* pada Twitter.

Surlakar *et al.,* (2016) melakukan penelitian yang berjudul *"Comparative Analysis of K-Means and K-Nearest Neighbor Image Segmentation Techniques".* Penelitian ini menggunakan segmentasi citra yang diolah menggunakan algoritma *K-Nearest Neighboor* dan algoritma *K-Means*. Dalam penelitian ini dinyatakan bahwa dengan menggunakan algoritma *k-Nearest Neighboor* lebih menghasilkan hasil yang optimal dalam segmentasi citra.

Chrismanto *et al.*, (2017) melakukan penelitian yang berjudul "Identifikasi Komentar *Spam* Pada Instagram". Penelitian ini menggunakan *dataset* yang diperoleh dari pengumpulan data 10 akun artis / aktor Indonesia yang memiliki *follower* lebih dari 1 juta dengan variabel komentar *spam* pada Instagram yang

diolah menggunakan algoritma *Support Vector Machine* serta dikenai metode validasi *k-Fold Cross Validation*. Dalam penelitian ini diperoleh hasil akurasi sebesar 78.49%.

Chrismanto *et al.,* (2017) melakukan penelitian yang berjudul "Deteksi Komentar *Spam* Bahasa Indonesia Pada Instagram Menggunakan *Naive Bayes*". Penelitian ini menggunakan *dataset* yang diperoleh dari pengumpulan data 10 akun artis / aktor Indonesia yang memiliki *follower* lebih dari 1 juta dengan variabel komentar *spam* pada Instagram yang diolah menggunakan algoritma *Naïve Bayes* serta dikenai metode validasi *k-Fold Cross Validation*. Dalam penelitian ini diperoleh hasil akurasi sebesar 77.25%.

Susanto *et al.,* (2018) melakukan penelitian yang berjudul *"A High Performace of Local Binary Pattern on Classify Javanese Character Classification"*. Penelitian ini menggunakan *dataset* yang diperoleh dari sebuah buku dengan variabel teks aksara jawa yang diolah menggunakan algoritma *k-Nearest Neighboor.* Dalam penelitian ini diperoleh hasil akurasi sebesar 82.5%.

Sugiharti *et al.,* (2017) melakukan penelitian yang berjudul *"Recognition Number of The Vehicle Plate Using Otsu Method and K-Nearest Neighbour Classification"*. Penelitian ini menggunakan *dataset* yang diperoleh dari data tempat parkir di Fakultas Matematika dan Ilmu Pengetahuan Alam UNNES yang diolah menggunakan algoritma *k-Nearest Neighboor*. Dalam penelitian ini diperoleh hasil akurasi sebesar 82%.

2.10 Kerangka Berpikir

Model kerangka pemikiran yang akan diaplikasikan pada penelitian ini yaitu, menambahkan *convert negation* pada tahap *preprocessing* dan pembobotan menggunakan TF-IDF *(Term Frequency - Inverse Document Frequency)* pada algoritma *K-Nearest Neighboor* dalam mendeteksi komentar *spam* berbahasa Indonesia pada Instagram.

Pada tahap awal setelah data siap diolah, dilakukan tahap *preprocessing* pada data. Langkah-langkah *preprocessing* dalam penelitian ini antara lain *case folding, cleansing, convert negation, stopwords removal, tokenization dan stemming,* kemudian diberikan pembobotan dengan TF-IDF *(Term Frequency - Inverse Document Frequency).* Penambahan *convert negation* pada tahap *preprocessing* dilakukan karena mudah diimplementasikan, menyaring lebih ketat pada pemrosesan teks sehingga data baru yang dihasilkan dapat lebih akurat (Fitri Febriyani, 2018). Adapun penambahan metode TF-IDF diberikan karena metode TF-IDF merupakan metode untuk menghitung bobot setiap kata yang paling umum digunakan pada *information retrieval*. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat.

Tahap selanjutnya adalah klasifikasi yang dilakukan dengan mengelompokkan data uji ke dalam kelas yang telah ditentukan dengan menggunakan algoritma *K-Nearest Neighbor* berdasarkan pada nilai data uji lingkungan dengan data pelatihan. Penerapan metode tersebut karena *K-Nearest Neighbor* adalah satu algoritma pembelajaran mesin yang paling popular, metode ini digunakan secara luas untuk operasi klasifikasi serta dapat digunakan untuk

analisis regresi, dengan langkah yang sederhana namun dapat memberikan hasil yang lebih akurat (Karakasli, 2019).

Untuk melakukan pengujian terhadap model yang dibangun, dilakukan dengan suatu skenario pengujian menggunakan metode *cross validation*. Dari pengujian yang dilakukan dengan metode *cross validation* dievaluasi untuk mengetahui tingkat akurasi dari setiap pengujian yang dilakukan.

BAB V

PENUTUP

5.1 Simpulan

Berdasarkan hasil penelitian, maka dapat ditarik kesimpulan sebagai berikut.

1) Algoritma *K-Nearest Neighboor* dalam mendeteksi komentar *spam* pada Instagram dapat dibangun dengan baik dengan memanfaatkan *framework Django* yang berbasis bahasa *Python*.
2) Perbandingan rata-rata hasil akurasi dalam 10 kali percobaan yang didapatkan pada klasifikasi algoritma *K-Nearest Neighboor* dalam mendeteksi komentar *spam* pada Instagram yaitu sebesar 88,45%, sedangkan klasifikasi algoritma *K-Nearest Neighboor* yang dioptimasi dengan *Convert Negation* dan TF-IDF *(Term Frequency – Inverse Document Frequency)* menghasilkan akurasi rata-rata sebesar 97,75%. Dari hasil penelitian tersebut dapat disimpulkan bahwa klasifikasi algoritma *K-Nearest Neighboor* yang dioptimasi dengan *Convert Negation* dan TF-IDF mampu meningkatkan hasil akurasi sebesar 7,3%.
3) Hasil akurasi algoritme *K-Nearest Neighbor* menggunakan *Convert Negation* dan TF-IDF pada tahap *preprocessing* dibanding penelitian terkait menghasilkan akurasi yang lebih baik, yaitu sebesar 97,75%.

5.2 Saran

Adapun saran dari penelitian ini adalah sebagai berikut.

1) Mengombinasikan Algoritma *K-Nearest Neighboor* dengan metode yang lain, atau menerapkan metode yang sama pada algoritma klasifikasi lainnya.
2) Melakukan pengembangan Algoritma *K-Nearest Neighboor* untuk mendapatkan hasil akurasi lebih tinggi.

DAFTAR PUSTAKA

Baoli, L., Y. Shiwen, & L. Qin. 2003. An Improved k-Nearest Neighbors for Text Categorization. *To appear in the Proceedings of the 20th International Conference of Computer Processing of Oriental Language.*

Carr, C. T. & R. A. Hayes. 2015. Social Media Defining, Developing, and Divining. *Journal Atlantic Journal of Communication*, 23(1): 46-65.

Chrismanto, A. R. & Y. Lukito. 2017. Deteksi Komentar Spam Bahasa Indonesia Pada Instagram Menggunakan Naive Bayes. *ULTIMATICS*, 9(1): 57-58.

Chrismanto, A. R. & Y. Lukito. 2017. Identifikasi Komentar Spam Pada Instagram. *Jurnal Ilmiah Teknologi Informasi*, 8(3): 219-231.

Dwi, A. dan Bambang. 2012. *Instagram Handbook Tips Fotografi Ponsel*. Jakarta: Media Kita.

Febriyani, S. Fitri, Muhammad Nasrun dan Casi Setianingsih. 2018. Sentiment Analysis on the Level of Customer Satisfaction to Data Cellular Services Using the Naive Bayes Classifier Algorithm. *IEEE International Conference on Internet of Things and Intelligence System (IoTaIS),* 201-206.

Feldman, R & Sanger, J. 2007. *The Text Mining Handbook: Advanced pproaches in Analyzing Unstructured Data.* Cambridge University Press: New York.

Goyal, S., R. K. Chauhan, & S. Parveen. 2016. Spam Detection using KNN and Decision Tree Mechanism in Social Network. *Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE,* 522-526.

Govindrajan, M. & R.M. Chandrasekaran. 2010. Evaluation of K-Nearest Neighbor classifier performance for direct marketing. *Expert Systems with Applications,* 37(1): 253–258

Harlian, Milka. 2006. *Machine Learning Text Kategorization*. Austin: University of Texas.

Imandoust, S.B. & M. Bolandraftar. 2013. Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoritical Background. *International Journal of Engineering Research and Applications*, 3(5): 605-610.

Kaplan, A M. & M. Haenlein. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Kelley School of Business,* 53(1): 59–68.

Karakasli, M. S., M. A. Aydin, S. Yarkan, & A. Boyaci. 2019. Dynamic Feature Selection for Spam Detection in Twitter. *International Telecommunications Conference*, 15(1): 239-250.

Khamis, H. S., K. W. Cheruiyot & S. Kimani. 2014. Application of k-Nearest Neighbor Classification in Medical Data Mining. *International Journal of Information and Communication Technology Research*, 4(4): 121-126.

Kohavi, R. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence (IJCAI).*

Kumar, A., S. Nayak, & N. Chandra. 2019. Empirical Analysis of Supervised Machine Learning Techniques for Cyberbullying Detection. *International Conference on Innovative Computing and Communications*, 2(1): 223-230.

Maulidia R. H., E. Sugiharti, & I. Akhlis. 2017. Recognition Number of The Vehicle Plate Using Otsu Method and K-Nearest Neighbour Classification. *Scientifics Journal of Informatics*, 4(1): 67-73.

Mittal, Kavita & P. Mahajan. 2017. Performances Analysis of K-Nearest Neighbor and K-Means Clustering To Predict The Diagnostic Accuracy. *Proceedings of International Conference on: Information, Communication and Computing Technology (ICICCT 2017),* 26-36.

Pramesti, R. P. A. 2013. Identifikasi Karakter Plat Nomor Kendaraan Menggunakan Ekstraksi Fitur ICZ dan ZCZ dengan Metode Klasifikasi KNN. *Scientific Repository of Bogor Agricultural University.*

Read, Jonathon. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification In ACL. *The Association for Computer Linguistics.*

Hafizh, A. 2019. *Implementation of k-nearest neighbor method for nutrition status classification web mobile based.* Skripsi. Yogyakarta: STIMIK AKAKOM.

Sugiharti, E., R. Arifudin, & A. T. Putra. 2018. C-Means And Fuzzy Tahani As Base Of Cattle Data Collection From Manual Card System To Online Information System. *Journal of Theoretical and Applied Information Technology*, 96(21): 7176 –7186.

Suguna, N. & K. Thanushkodi. 2010. An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. *International Journal of Computer Science Issues*, 7(2): 18-21.

Sumarlin. 2015. Implementasi Algoritma K-Nearest Neighbor Sebagai Pendukung Keputusan Klasifikasi Penerima Beasiswa PPA dan BBM. *Jurnal Sistem Informasi Bisnis*, 1(1): 52-61.

Surlakar, P., S. Araujo, & K. M. Sundaram. 2016. Comparative Analysis of K-Means and K-Nearest Neighbor Image Segmentation Techniques. *IEEE 6th International Conference on Advanced Computing (IACC)*, 96-99.

Susanto, A., D. Sinaga, C. A. Sari, & E. H. Rachmawanto. 2018. A High Performace of Local Binary Pattern on Classify Javanese Character Classification. *Scientifics Journal of Informatics*, 5(1): 1-7.

Tala, F. Z. 2003. *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. Institute for Logic, Language and Computation. Universiteti van Amsterdam The Netherlands.

Vijayarani, S., Ilamathi, & J. Nithya. 2015. Preprocessing Techniques for Text Mining - An Overview. *International Journal of Computer Science & Communication Networks*, 5(1): 7-16.