EKSTRAKSI TF-IDF N-GRAM DARI KOMENTAR PELANGGAN PRODUK *SMARTPHONE* PADA *WEBSITE E-COMMERCE*

Sulis Mardianti[1)], **Muhammad Zidny Naf'an**[2)], **Indra Hidayatulloh**[3)]

[1,2,3)] *Fakultas Teknologi Industri dan Informatika, Institut Teknologi Telkom Purwokerto*
Jl. D.I Panjaitan No. 128, Purwokerto Kidul, Purwokerto Selatan, Purwokerto, Jawa Tengah 53147
Email : 14102084@st3telkom.ac.id[1)], zidny@ittelkom-pwt.ac.id[2)], indra@ittelkom-pwt.ac.id[3)]

Abstrak

Jumlah pengguna internet di Indonesia meningkat pada tahun 2017 dengan jumlah 132,7 juta. Hal tersebut juga meningkatkan belanja online masyarakat Indonesia. Tingginya jumlah belanja online smartphone pada e-commerce dapat disebabkan karena fitur-fitur yang ada pada e-commerce, diantaranya adalah fitur komentar. Sebagian besar menyatakan memerlukan ringkasan fitur komentar yang berisi kriteria-kriteria smartphone sehingga memudahkan pelanggan menilai positif dan negatif kualitas produk yang akan dibeli. Ringkasan komentar yang digunakan adalah ringkasan komentar bahasa Indonesia yang berisi spesifikasi dari smartphone. Untuk mendapatkan kualitas komentar yang baik maka dilakukan teknik preprocessing. Teknik preprosesing digunakan untuk menghilangkan noise data komentar yang sudah diperoleh. Cara untuk mengetahui jumlah term pada komentar adalah melalui proses perhitungan tf idf. Penggunaan fitur tf idf dapat memberikan informasi seberapa banyak kata tersebut muncul dalam dokumen yang sudah memiliki klasifikasi positif, negatif, dan netral pada masing-masing kriteria. Untuk memudahkan dalam melakukan perhitungan serta mengindari adanya kesalahan pada ekstrasi tf idf, maka digunakan N-Gram. Penggunaan N-Gram memberikan keuntungan karena hasil yang diperoleh menjadi lebih akurat dan lebih efektif.

Kata kunci: *Bahasa Indonesia, Komentar, N-Gram, Preprocessing, Smartphone, TF IDF*

1. Pendahuluan

Berdasarkan laporan perusahaan riset *We Are Social* tanggal 26 Januari 2017 menyebutkan bahwa "Indonesia sebagai jumlah pengguna internet terbesar urutan ketiga di dunia dengan jumlah 132,7 juta". Dari data tersebut diperoleh informasi bahwa jumlah pengguna internet di Indonesia naik 51% dari tahun 2016. Selain itu, perusahaan riset *We Are Social* juga menyebutkan jumlah pengguna internet yang berbelanja secara *online* tahun 2017 mencapai 24,74 juta orang dengan menghabiskan sekitar Rp 74,6 triliun [1]. Menurut data yang diperoleh Badan Statistika Kominfo, pada tahun 2015 sebanyak 12,2% adalah kegiatan belanja *smartphone* dari total belanja melalui *e-commerce* [2].

Tingginya angka belanja melalui *e-commerce* disebabkan oleh fitur yang ada pada *e-commerce* itu sendiri. Fitur yang ditawarkan diantaranya adalah fitur komentar, dimana pelanggan bebas memberikan pernilaian terhadap barang yang sudah dibeli. Berdasarkan survei yang dilakukan oleh peneliti sebanyak 93,5% atau 87 dari 93 responden menyebutkan selalu menggunakan fitur komentar sebelum melakukan transaksi dan sebanyak 38,7% selalu membaca ulasan komentar. Selain itu, responden juga memberikan pernyataan bahwa mereka membutuhkan ringkasan komentar dari setiap pelanggan yang melalukan transaksi dan mempertimbangkan komentar pelanggan yang berisi kriteria *smartphone* yaitu memori, jaringan, kamera, simcard, dan kemampuan konektifitas *smartphone.*

Salah satu cara meringkas komentar pelanggan adalah melalui *sentiment analysis. Sentiment Analysis* merupakan sebuah teknik yang digunakan untuk menentukan kalimat positif, negatif, dan netral. Tugas utama dari *sentiment analysis* adalah mengklasifikasian perbedaan pada satu kalimat atau dokumen yang menggambarkan adanya kata *sentiment* positif, negatif, dan netral [3].

Untuk mendukung suatu penelitian agar lebih maksimal, terdapat beberapa fitur yang digunakan. Pada penelitian yang dilakukan oleh Socher, dkk [4] menggunakan fitur EP (*Experience Project*). Fitur yang digunakan pada penelitian ini tidak dapat secara maksimal mendukung kinerja metode sehingga metode tidak dapat mengenali kalimat sentiment pada dokumen. Selain itu terdapat penelitian yang dilakukan oleh Lu, dkk [5]menggunakan fitur N-Gram tetapi fitur tidak dapat mendukung pengklasifikasian dokumen. Penelitian lain yang dilakukan oleh Singh, dkk [6] menggunakan fitur *Feature-Based Heuristic* untuk klasifikasi kalimat sentimen. Terdapat masalah ketika menggunakan fitur ini yaitu sulitnya menentukan POS (*Part of Speech*) dan nilai *weightage* pada dokumen. Fitur *tf-idf* pada penelitian yang dilakukan oleh Hidayatullah, dkk [7] menghasilkan akurasi tinggi. Selain itu, fitur *tf-idf* juga dapat mendukung kinerja metode yang digunakan pada penelitian sehingga metode juga menunjukkan tingkat akurasi yang tinggi. Pada penelitian yang dilakukan oleh Indrayani dan Wahyudi [8] menggunakan N-Gram menunjukkan akurasi tinggi pada ekstrasi fitur *three-gram* dibanding dengan *uni-gram* dan *bi-gram.*

Dengan melihat penjabaran pada studi pustaka yang sudah dilakukan, peneliti menggunakan fitur *tf-idf* dengan pendukung N-Gram untuk melakukan pengolahan data komentar. N-Gram dipilih karena banyaknya komentar yang diperoleh menggunakan bahasa tidak baku.

2. Pembahasan

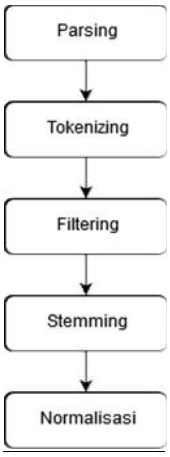
Penelitian yang dilakukan ini menggunakan objek komentar pelanggan terhadap pembelian *smartphone* pada *e-commerce* dengan kriteria tertentu. Berdasarkan kajian pustaka pada penelitian yang dilakukan oleh Hidayatulloh dan Naf'an, diperoleh objek *smartphone* yang digunakan juga sebagai objek penelitian ini [9].

Pengambilan data pada *e-commerce* tersebut dengan sesuai *smartphone* yang sudah dipilih dapat dilakukan dengan proses *crawling*. Berdasarkan studi pustaka yang dilakukan mengenai pengambilan data melalui *website* diperoleh tahap-tahap sebagai berikut. Tahap *crawling* dapat dilihat pada gambar 1.

Gambar 1. *Tahap Crawling*

Seperti yang sudah dijelaskan, *crawling* adalah proses pengumpulan data dalam jumlah banyak dengan mudah [10]. Proses *crawling* digunakan untuk mengambil seluruh data yang di inginkan dalam jumlah besar pada satu waktu. Pada proses pengambilan data komentar melalui *crawling* perlu dibangun sebuah algoritmanya terlebih dahulu. Algoritma tersebut bertujuan untuk mengetahui membaca struktur *website* agar data yang diambil sesuai dengan kebutuhan. Algoritma yang dibangun berisi proses pengambilan dan penyimpanan data komentar melalui *website e-commerce*. Selain itu keuntungan dibangunnya algoritma ini adalah pengambilan data dan penyimpanan data ke *database* menjadi lebih efektif. Proses *crawling* pada *e-commerce* membuat peneliti harus memahami struktur dari *e-commerce* tersebut sehingga komentar pada *e-commerce* dapat diambil. Data komentar yang sudah diambil, selanjutnya disimpan dalam *database* agar data dapat dengan mudah diolah untuk proses-proses berikutnya.

Data yang diperoleh dari hasil *crawling* untuk selanjutnya dilakukan proses *preprocessing* untuk mendapatkan kalimat yang lebih baik. Tujuan dilakukannya *preprocessing* adalah untuk menghilangkan *noise* pada kalimat komentar sehingga lebih mudah untuk dilakukan klasifikasi komentar positif, negatif, atau netral. Untuk memudahkan dalah proses *preprocessing*, digunakan algoritma dari Sastrawi Master. Tahap *preprocessing* dapat dilihat pada gambar 2 dibawah ini [11] :



Gambar 2. *Tahap Teknik Preprocessing*

Parsing adalah proses memecah sebuah data menjadi beberapa dokumen. Proses *parsing* berkaitan dengan proses menyimpan hasil *crawling* pada *database*. Proses setelah dilakukan *parsing* adalah proses *tokenizing.*

Proses *tokenizing* dilakukan untuk memisahkan atau menghapus seluruh tanda baca, karakter, dan angka yang terdapat dalam data komentar. Proses ini Hasil dari proses ini adalah berupa kumpulan kata-kata tanpa tanda baca, karakter, dan angka. Logika pengolahan *tokenizing* dijelaskan pada *pseudocode* dibawah ini.

if karakter = angka then

hapus karakter

end if

if karakter = tanda baca then

hapus karakter

end if

Contoh hasil *tokenizing* dapat dilihat pada tabel 1.

Tabel 1. *Hasil Tokenizing*

Sebelum *Tokenizing*	**Setelah *Tokenizing***
Untuk produk SAMSUNG Galaxy J7 Prime [SM-G610] - Pink/Gold Pink (Merchant) sendiri sih saya belum memiliki produk tersebut tetapi saya sangat terkesan saat menggunakan produk SAMSUNG Galaxy J7 Prime [SM-G610] punya kawan saya dan semua fitur serta spesifikasinya	"untuk", "produk", "samsung", "galaxy", "j" "prim", "smg", "pinkgold" , "pink", "merchant", "sediri", "sih", "saya", "belum", "memiliki", "produk", "tersebut", "tetapi", "saya", "sangat", "terkesan", "saat", "menggunakan", "produk" "samsung", "galaxy", "j", "prime", "smg", "punya", "kawan", "saya", "dan",

luar biasa, goodluck terus untung #SAMSUNG #GALAXYJ7	"semua", "fitur", "serta", "spesifikasinya", "luar", "biasa", "goodluck", "terus", "untung", "samsung", "galaxyj"
Pengalaman saya menggunakan samsung J 7 prime yg saya kagumi adalah kameranya yg bagus dan kualitas bentuk nya yang mewah dengan paduan warna gold dan putih.Intinya top banget deh utk samsung j 7 prime.	"pengalaman", "saya", "menggunakan", "samsung", "j", "prime", "yg", "saya", "kagumi", "adalah", "kameranya", "yg", "bagus", "dan", "kualitas", "bentuk", "nya", "yang", "mewah", "dengan", "paduan", "warna", "gold", "dan", "putih", "intinya", "top", "banget", "deh", "utk", "samsung", "j", "prime"

Proses *filtering* biasa disebut *stop-word-removal*. Proses *filtering* ini adalah untuk menghapus setiap kata-kata yang tidak memiliki arti atau kata-kata yang sering muncul pada data komentar. Contoh kata tersebut adalah yang, di, dari, dan sebagainya. Akan tetapi, proses *filtering* tidak bekerja dengan baik sehingga masih banyak kata *stop-word* pada dokumen meskipun sudah dilakukan proses filtering. Hal tersebut karena banyak kalimat tidak baku pada komentar, sedangkan kamus terdiri dari kata baku sehingga hanya kata baku saja yang dapat terhapus. Kamus yang digunakan mengacu pada ID-Stopwords. Berikut adalah *pseudocode filtering* untuk memberikan penjelasan proses *filtering.*

if kata = stopword then

hapus kata

end if

Tabel 2 dibawah ini adalah hasil setelah dilakukan *filtering.*

Tabel 2. *Hasil Filtering*

Sebelum Filtering	**Setelah Filtering**
"untuk", "produk", "samsung", "galaxy", "j" "prim", "smg", "pinkgold" , "pink", "merchant", "sediri", "sih", "saya", "belum", "memiliki", "produk", "tersebut", "tetapi", "saya", "sangat", "terkesan", "saat", "menggunakan", "produk" "samsung", "galaxy", "j", "prime", "smg", "punya", "kawan", "saya", "dan", "semua", "fitur", "serta", "spesifikasinya", "luar", "biasa", "goodluck", "terus", "untung", "samsung", "galaxyj"	"untuk", "produk", "samsung", "galaxy", "j", "prime", "smg", "pinkgold", "pink", "merchant", "sih", "memiliki", "produk", "terkesan","produk", "samsung", "galaxy", "j", "prime", "smg", "kawan", "fitur","spesifikasi", "goodluck", "untung", "Samsung", "galaxy"
"pengalaman", "saya", "menggunakan", "samsung", "j", "prime", "yg", "saya", "kagumi", "adalah", "kameranya", "yg", "bagus", "dan", "kualitas", "bentuk", "nya", "yang", "mewah", "dengan", "paduan", "warna", "gold", "dan", "putih", "intinya", "top", "banget", "deh", "utk", "samsung", "j", "prime"	"pengalaman", "Samsung", "j", "prime", "yg", "kagumi", "kameranya", "yg", "bagus", "kualitas", "bentuknya", "nya","mewah", "paduan", "warna", "gold", "putih", "intinya", "top", "banget", "deh", "utk", "samsung", "j", "prime"

Setelah melakukan *filtering*, yang dilakukan selanjutnya adalah proses *stemming.* Proses *stemming* digunakan untuk mengubah setiap kata pada data komentar menjadi kata dasarnya. Proses *stemming* ditunjukkan pada *pseudocode* dibawah ini.

if kata = stopword then

ganti kata dengan kamus

end if

Tabel 3 dibawah ini adalah contoh hasil *stemming.*

Tabel 3. *Hasil Stemming*

Sebelum Stemming	**Setelah Stemming**
"untuk", "produk", "samsung", "galaxy", "j", "prime", "smg", "pinkgold", "pink", "merchant", "sih", "memiliki", "produk", "terkesan","produk", "samsung", "galaxy", "j", "prime", "smg", "kawan", "fitur","spesifikasi", "goodluck", "untung", "Samsung", "galaxy"	"untuk", "produk", "samsung", "galaxy", "j", "prime", "smg", "pinkgold", "pink", "merchant", "sih", milik", "produk", "kes","produk", "samsung", "galaxy", "j", "prime", "smg", "kawan", "fitur","spesifikasi", "goodluck", "untung", "Samsung", "galaxy"
"pengalaman", "Samsung", "j", "prime", "yg", "kagumi", "kameranya", "yg", "bagus", "kualitas", "bentuknya", "nya","mewah", "paduan", "warna", "gold", "putih", "intinya", "top", "banget", "deh", "utk", "samsung", "j", "prime"	"alam", "Samsung", "j", "prime", "yg", "kagum", "kamera", "yg", "bagus", "kualitas", "bentuk", "nya","mewah", "padu", "warna", "gold", "putihintinya", "top", "banget", "deh", "utk", "Samsung", "j", "prime"

Proses normalisasi dilakukan untuk mengubah kata tersebut menjadi kata baku. Akan tetapi, kata hasil akhir setelah dilakukan normalisasi tidak merepresentasikan suatu kata. Hal tersebut karena kata-kata pada komentar *e-commerce* di Indonesia menggunakan kata santai dan banyak menggunakan singkatan. Oleh karena itu, hasil untuk proses normalisasi tidak digunakan dan hanya menampilkan hasil sampai proses *stemming* saja. Hasil teknik preprosesing data komentar dapat dilihat pada tabel 4 dibawah ini.

Tabel 4. *Tabel Hasil Preprosesing*

Sebelum Preprosesing	**Setelah Preprosesing**
Untuk produk SAMSUNG Galaxy J7 Prime [SM-G610] - Pink/Gold Pink (Merchant) sendiri sih saya belum memiliki produk tersebut tetapi saya sangat terkesan saat menggunakan produk SAMSUNG Galaxy J7 Prime [SM-G610] punya kawan saya dan semua fitur serta spesifikasinya luar biasa, goodluck terus untung #SAMSUNG #GALAXYJ7	"untuk", "produk", "samsung", "galaxy", "j", "prime", "smg", "pinkgold", "pink", "merchant", "sih", milik", "produk", "kes","produk", "samsung", "galaxy", "j", "prime", "smg", "kawan", "fitur","spesifikasi", "goodluck", "untung", "Samsung", "galaxy"
Pengalaman saya menggunakan samsung J 7 prime yg saya kagumi adalah kameranya yg bagus dan kualitas bentuk nya yang mewah dengan paduan warna gold dan putih.Intinya top banget deh utk samsung j 7 prime.	"alam", "Samsung", "j", "prime", "yg", "kagum", "kamera", "yg", "bagus", "kualitas", "bentuk", "nya","mewah", "padu", "warna", "gold", "putihintinya", "top", "banget", "deh", "utk", "Samsung", "j", "prime"

Tahap-tahap *preprocessing* diatas menggunakan *sample* data komentar yang diperoleh dari *website e-commerce* yang disimpan dalam *database.*

Hasil proses *preprocessing* selanjutnya akan diberikan klasifikasi berdasarkan kalimat *sentiment* yaitu positif, negatif, dan netral. Untuk menentukan klasifikasi komentar berdasarkan positif, negatif, dan netral terhadap suatu komentar dapat dilakukan secara manual. Proses menentukan klasifikasi komentar terhadap kata dengan melihat pada kriteria memori, jaringan, kamera, simcard, dan konektifitas pada komentar. Untuk memudahkan dalam memahami klasifikasi maka memori diganti sebagai "M", jaringan sebagai "J", kamera sebagai "K", simcard sebagai "S", konenktifitas sebagai "C". Nilai klasifikasi pada proses ini adalah 1 untuk positif, -1 untuk negatif, dan 0 untuk netral berdasarkan kriteria. Apabila dalam komentar tidak berisi kriteria sesuai kriteria ketentuan maka nilai pada seluruh kolom kriteria adalah 0 atau netral. Hasil klasifikasi disimpan dalam *database* agar dapat dilakukan proses perhitungan *tf,df,* dan *idf.* Contoh dapat dilihat pada tabel 5 dibawah ini.

Tabel 5. *Penentuan Kelas Pada Data Latih*

Komentar	**M**	**J**	**K**
untuk produk samsung galaxy j prime smg pinkgold pink merchant sih milik produk kes produk samsung galaxy j prime smg kawan fitur spesifikasi goodluck untung samsung galaxy	0	0	0
alam samsung j prime yg kagum kamera yg bagus kualitas bentuk nya mewah padu warna gold putihintinya top banget deh utk samsung j prime	0	0	1

Selanjutya, data hasil klasifikasi yang masih memiliki nilai kualitatif akan diolah menjadi bernilai kuantitatif melalui perhitungan *tf, df,* dan *idf.*

Penilaian komentar berdasarkan positif, negatif, dan netral suatu komentar dengan melihat pada soesifikasi seperti tabel 2 akan digunakan untuk menghitung *term frequency* (*tf*) adalah kemunculan suatu term terhadap dokumen atau dalam hal ini seluruh komentar, *document frequency* (*df*) adalah banyaknya dokumen suatu term, dan *invers document frequency* (*idf*).

Untuk memudahkan penghitungan *tf, df,* dan *idf* peneliti mengekstrasi seluruh dokumen menjadi N-Gram. Pada ekstrasi fitur N-Gram tidak boleh adanya duplikasi kata pada hasil N-Gram karena hal tersebut dapat memengaruhi hasil perhitungan. Pada penelitian ini, penghitungan *tf* dilakukan berdasarkan kemunculan suatu term pada dokumen yang memiliki nilai pada masing-masing spesifikasi. Untuk lebih jelasnya dapat melihat tabel 6 yang hasilnya diperoleh dari tabel 5.

Tabel 6. *Contoh Hasil tf*

Kata	M+	M-	M0	K+	K-	K0	S+	S-	S0
Untuk	0	0	1	0	0	1	0	0	1
Produk	0	0	1	0	0	1	0	0	1
Samsung	0	0	2	1	0	1	0	0	1
galaxy	0	0	1	0	0	1	0	0	1
J	0	0	2	1	0	1	0	0	2
Prime	0	0	2	1	0	1	0	0	2
Smg	0	0	1	0	0	1	0	0	1
Pink	0	0	1	0	0	1	0	0	1

Saya	0	0	2	1	0	1	0	0	2
sendiri	0	0	1	0	0	1	0	0	1
Belum	0	0	1	0	0	1	0	0	1
kagum	0	0	1	1	0	0	0	0	0

Melihat pada tabel 7 dapat jelaskan jika suatu kata termasuk dalam komentar yang memiliki nilai pada lebih dari satu spesifikasi, maka nilai dari spesifikasi lain juga ditampilkan pada term kata tersebut. Begitu pula dengan kesamaan nilai. Apabila terdapat suatu kata "oke" dimana kata tersebut termasuk pada dua atau lebih komentar yang memiliki nilai sama misalkan pada memori positif, maka nilai yang ditampilkan pada hasil tf kata "oke" pada memori positif adalah sebanyak komentar yang mengandung kata "oke". Pada tabel 6 dan 7 diatas adalah sebagai contoh cara mengklasifikasikan komentar dan menghitung *tf* pada komentar, sehingga hanya digunakan dua komentar. Data sebenarnya yang diperoleh adalah 213 komentar. Komentar yang telah dilakukan *preprocessing* kemudian di ekstraksi dalam N-Gram.

Hasil masing-masing term terhadap satu kata dapat digunakan untuk mencari nilai *df*. Untuk memeroleh nilai *idf* digunakan persamaan (1) dibawah ini [12] :

$$idf = \log 10(doc / df) \quad .....(1)$$

doc pada merupakan jumlah seluruh dokumen dalam hal ini adalah jumlah data komentar yang diperoleh. *df* merupakan jumlah keseluruhan term pada satu dokumen atau dalam hal ini adalah kata. Contoh pada tabel 7 adalah hasil *df* dan *idf*.

Tabel 7. *Hasil df dan idf*

Kata	M+	M-	M0	K+	K-	K0	S+	S-	S0	*df*	*Idf*
Untuk	0	0	1	0	0	1	0	0	1	3	1.85126
Produk	0	0	1	0	0	1	0	0	1	3	1.85126
Samsung	0	0	2	1	0	1	0	0	2	6	1.55023
galaxy	0	0	1	0	0	1	0	0	1	3	1.85126
J	0	0	2	1	0	1	0	0	2	6	1.55023
Prime	0	0	2	1	0	1	0	0	2	6	1.55023
Smg	0	0	1	0	0	1	0	0	1	3	1.85126
Pink	0	0	1	0	0	1	0	0	1	3	1.85126
Saya	0	0	2	1	0	1	0	0	2	6	1.55023
sendiri	0	0	1	0	0	1	0	0	1	3	1.85126
Belum	0	0	1	0	0	1	0	0	1	3	1.85126
kagum	0	0	1	1	0	0	0	0	1	3	1.85126

Dari tabel 8 diketahui bahwa semakin sedikit jumlah term pada dokumen maka nilai *idf* akan semakin besar. Sedangkan semakin banyak jumlah term pada dokumen makan nilai *idf* akan semakin besar. Sebagai buktinya, pada kata "bagus" memiliki term kemunculan sebanyak 9 sehingga hasil *idf*nya lebih kecil dari kata "beli" yang hanya memiliki term kemunculan sebanyak 6.

3. Kesimpulan

Ekstrasi *tf-idf* dapat digunakan untuk mengetahui nilai term kemunculan suatu kata terhadap dokumen. Penerapan N-Gram pada ekstrasi *tf-idf* menjadikan algoritma perhitungan *tf-idf* berdasarkan kriteria *smartphone* lebih efektif. Selain itu penggunaan N-Gram pada ekstrasi fitur *tf idf* menjadi lebih akurat dan terhindar dari kesalahan perhitungan. Pada penelitian ini menghasilkan sebuah nilai *tf-idf* dari seluruh dokumen. Untuk memeroleh hasil terbaik pada penelitian ini, direncanakan untuk menggunakan sebuah metode yang dapat mengklasifikasikan data komentar menjadi lebih spesifik sesuai dengan kriteria. Hal tersebut untuk memeroleh kalimat sentimen pada masing-masing komentar berdasarkan kriteria.

Daftar Pustaka

[1] A. H. Pratama, "Pertumbuhan Pengguna Internet di Indonesia Tahun 2016," 2017. [Online]. Available: https://id.techinasia.com/pertumbuhan-pengguna-internet-di-indonesia-tahun-2016. [Accessed: 02-May-2017].

[2] Kominfo, "klasifikasi produk yang dibeli secara online," 2015. [Online]. Available: https://statistik.kominfo.go.id/site/data?idtree=430&iddoc=1457. [Accessed: 01-Jan-2017].

[3] A. Buche, D. M. B. Chandak, and A. Zadgaonkar, "Opinion Mining and Analysis: A survey," *Int. J. Nat. Lang. Comput.*, vol. 2, no. 3, pp. 39–48, 2013.

[4] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions," no. ii, pp. 151–161, 2011.

[5] B. Lu *et al.*, "Multi-aspect Sentiment Analysis with Topic Models," *Proc. - IEEE Int. Conf. Data Min.*, pp. 81–88, 2011.

[6] W. P. Singh V, Piryani R, Uddin A, "Sentiment analysis of movie reviews : A new feature-based heuristic for aspect-level sentiment classification Sentiment Analysis of Movie Reviews," *Proc. - 2013 IEEE Int. Multi Conf. Autom. Comput. Control. Commun. Compress. Sens.*, no. March, pp. 712–717, 2013.

[7] A. F. Hidayatullah and A. Sn, "Analisis Sentimen dan Klasifikasi Kategori Terhadap Tokoh Publik Pada Twitter," *Semin. Nas. Inform. 2014*, vol. 2014, no. August 2013, pp. 0–8, 2014.

[8] E. Indrayuni, M. Wahyudi, S. Informasi, J. Selatan, I. Komputer, and J. Selatan, "Penerapan Character N-Gram Untuk Sentiment Analysis Review Hotel Menggunakan Algoritma Naive Bayes," pp. 88–93, 2015.

[9] I. Hidayatulloh and M. Z. Naf'an, "Metode MOORA dengan Pendekatan Price-Quality Ratio untuk Rekomendasi Pemilihan Smartphone," *Semin. Nas. Teknol. Inf. dan Apl. Komput.*, pp. 62–68, 2017.

[10] W. Crawler, "Pentingnya Web Crawling sebagai Cara

Pengumpulan Data di Era Big Data."
[11] A. Fauzi, "Text-Pre-Processing-v2," 2016. [Online]. Available: malifauzi.lecture.ub.ac.id/files/2016/02/Text-Mining-and-IR.pptx. [Accessed: 08-May-2017].
[12] Y. V. Yosnaningsih, "Klasifikasi Dokumen Bahasa Jawa Menggunakan Metode Naive Bayes," Sanata Dharma University, 2015.

Biodata Penulis

Sulis Mardianti, lulus Sekolah Menengah Atas pada tahun 2014 dan saat ini menjadi Mahasiswa Jurusan S1 Informatika semester tujuh di Institut Teknologi Telkom Purwokerto.

Muhammad Zidny Naf'an , memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika UIN Syarif Hidayatullah, lulus tahun 2012. Memperoleh gelar Magister Komputer (M.Kom) Program Pasca Sarjana Magister Ilmu Komputer Universitas Indonesia, lulus tahun 2015. Saat ini menjadi Dosen di IT Telkom Purwokero.

Indra Hidayatulloh , memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika STMIK AmikBandung, lulus tahun 2010. Memperoleh gelar Magister Teknik Elektro (M.T) Program Pasca Sarjana Magister Teknik Elektro Institut Teknologi Bandung, lulus tahun 2015. Saat ini menjadi Dosen di IT Telkom Purwokero.