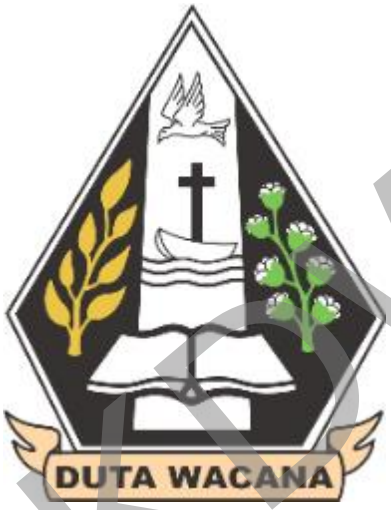
Sistem Analisis Gaya Penulisan untuk Pengelompokan Dokumen dengan Metode K-Means

Skripsi



oleh

LEONARDO SENDY DWI ATMOKO

71140098

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS KRISTEN DUTA WACANA

Sistem Analisis Gaya Penulisan untuk Pengelompokan Dokumen dengan Metode K-Means

Skripsi

Diajukan kepada Program Studi Informatika Fakultas Teknologi Informasi

Universitas Kristen Duta Wacana

Sebagai Salah Satu Syarat dalam Memperoleh Gelar

Sarjana Komputer

Disusun oleh

LEONARDO SENDY DWI ATMOKO

71140098

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

SISTEM ANALISIS GAYA PENULISAN UNTUK PENGELOMPOKAN DOKUMEN DENGAN METODE K-MEANS

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 21 Juni 2018

LEONARDO SENDY DWI ATMOKO
71140098

HALAMAN PERSETUJUAN

Judul Skripsi	:	SISTEM ANALISIS GAYA PENULISAN UNTUK PENGELOMPOKAN DOKUMEN DENGAN METODE K-MEANS
Nama Mahasiswa	:	LEONARDO SENDY DWI ATMOKO
N I M	:	71140098
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Genap
Tahun Akademik	:	2017/2018

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 21 Juni 2018

Dosen Pembimbing I	Dosen Pembimbing II
Lucia Dwi Krisnawati, Dr. Phil.	Danny Sebastian, S.Kom., M.M., M.T.

HALAMAN PENGESAHAN

SISTEM ANALISIS GAYA PENULISAN UNTUK PENGELOMPOKAN DOKUMEN DENGAN METODE K-MEANS

Oleh: LEONARDO SENDY DWI ATMOKO / 71140098

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 31 Mei 2018

Yogyakarta, 21 Juni 2018
Mengesahkan,

Dewan Penguji:

1. Lucia Dwi Krisnawati, Dr. Phil. .....
2. Danny Sebastian, S.Kom., M.M., M.T. .....
3. Antonius Rachmat C., S.Kom.,M.Cs. .....
4. R. Gunawan Santosa, Drs. M.Si. .....

Dekan

(Budi Susanto, S.Kom., M.T.)

Ketua Program Studi

(Gloria Virginia, Ph.D.)

UCAPAN TERIMAKASIH

Puji syukur dan terima kasih penulis panjatkan kepada Tuhan Yesus Kristus atas berkat, penyertaan dan anugerah-Nya, sehingga skripsi yang berjudul "Sistem Analisis Gaya Penulisan untuk Pengelompokan Dokumen dengan Metode K-Means" dapat selesai dengan baik dan tepat waktu.

Penelitian ini diajukan untuk melengkapi syarat kelulusan dan mencapai gelar strata satu (S1) di Fakultas Teknologi Informasi, prodi Informatika, Universitas Kristen Duta Wacana. Peneliti sadar walaupun telah berusaha semaksimal mungkin untuk menyajikan pembahasan dan analisis, namun masih banyak kekurangan dan kata yang kurang tepat pada tugas akhir ini. Hal ini dikarenakan masih terbatasnya kemampuan dan pengetahuan peneliti. Atas sebab itum peneliti mengharapkan kritik dan saran untuk membangun serta menyempurnakan tugas akhir ini.

Banyak kendala dan hambatan yang di alami penulis dalam proses penyusunan tugas akhir ini, namun berkat bantuan, bimbingan, dan kerjasama dari berbagai pihak sehingga kendala dan hambatan tersebut dapat diatasi dengan baik. Oleh karena itu peneliti mengucapkan terima kasih kepada,

1. Tuhan Yesus Kristus sumber kehidupan yang senantiasa menyertai dan memberikan berkat serta menghadirkan inspirasi bagi penulis.
2. Keluarga yang terus memberikan dukungan hingga akhirnya skripsi ini dapat selesai tepat waktu.
3. Bapak Ir. Hendry Feriadi, M.Sc., Ph.D. selaku Rektor Universitas Kristen Duta Wacana.
4. Bapak Budi Susanto, S.Kom., M.T. selaku Dekan Fakultas Teknologi Informasi Universitas Kristen Duta Wacana.
5. Ibu Gloria Virgina, S.Kom., M.AI, Ph.D. selaku Kepala Program Studi Informatika Universitas Kristen Duta Wacana.

6. Ibu Lucia D. Krisnawati selaku dosen pembimbing I dan Bapak Danny Sebastian, S.Kom., M.M., M.T.. Selaku dosen pembimbing II yang telah mendukung, membimbing, dan bersedia meluangkan waktu, tenaga dan pikiran dalam memberikan arahan, motivasi, serta saran yang sangat membantu bagi peneliti dalam menyusun tugas akhir ini.
7. Segenap anggota grup A.H.I.E yang jumlah anggotanya cukup banyak dan cukup melelahkan untuk ditulis satu persatu. Yang mau terus berjuang bersama, berbagi ilmu dan saling memberi dukungan dengan bentuk sindiran dan canda tawa.
8. Irmeliana, Cyndy dan teman teman seperjuangan lainnya yang telah berjuang bersama selama ini.
9. Semua pihak yang tidak dapat disebutkan satu persatu yang telah ikut serta dalam memberikan dukungan baik secara langsung ataupun tidak langsung.

Peneliti menyadari bahwa masih banyak kekurangan dalam penelitian ini, baik dalam penulisan dan pembahasan. Akhir kata peneliti mengucapkan terima kasih kepada semua pihak yang telah berkontribusi dalam penelitian tugas akhir ini. Peneliti juga berharap semoga tugas akhir ini dapat bermanfaat bagi para pembaca.

INTISARI

Sistem Analisis Gaya Penulisan untuk Pengelompokan Dokumen dengan Metode K-Means

Mudahnya membuat identitas palsu atau menggunakan identitas orang lain di dunia maya. Membuat tindakan tidak bertanggung jawab hingga kriminal sering terjadi di dunia maya. Hal tersebut terjadi karena identitas palsu membuat pelaku merasa lebih aman dalam melakukan tindakannya. Namun meskipun pelaku menggunakan akun palsu, tindakan tidak bertanggung ataupun criminal yang terjadi dalam dunia maya sebagian besar masih menggunakan teks.

Setiap teks selalu terkadung gaya penulisan dari penulisnya. Gaya penulisan inilah yang dapat menjadi bahan penelitian dalam bidang teks forensik. Penelitian teks forensik dibidang teknologi pun semakin banyak digelar. Sehingga sistem untuk membantu menanggulangi masalah banyaknya teks tidak bertanggung jawab semakin memungkinkan untuk dibuat.

Dengan menggunakan metode pengelompokan K-Means dan memanfaatkan koefisien silhouette penulis mencoba membangun sistem yang dapat membantu menanggulangi masalah tersebut dengan mengelompokan dokumen berdasarkan gaya penulisannya. Dengan harapan dapat memberikan gambaran mengenai jumlah penulis yang tidak bertanggung jawab dari sejumlah dokumen yang diinput.

Dengan menggunakan fitur variasi dan frekuensi bigram trigram *stopword*, variasi dan frekuensi kata slang, prosentasi jumlah *stopword* per non *stopword*, jumlah kalimat langsung dalam dokumen, dan rata rata kalimat dalam paragraph. Hasil yang didapat untuk pengelompokan ini dinilai kurang baik dan perlu mencari fitur lain yang lebih baik agar dapat merepresentasikan dokumen yang di uji.

Kata Kunci : clustering, gaya penulisan, k-means, Indonesia,

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

INTISARI

Sistem Analisis Gaya Penulisan untuk Pengelompokan Dokumen dengan Metode K-Means

Mudahnya membuat identitas palsu atau menggunakan identitas orang lain di dunia maya. Membuat tindakan tidak bertanggung jawab hingga kriminal sering terjadi di dunia maya. Hal tersebut terjadi karena identitas palsu membuat pelaku merasa lebih aman dalam melakukan tindakannya. Namun meskipun pelaku menggunakan akun palsu, tindakan tidak bertanggung ataupun criminal yang terjadi dalam dunia maya sebagian besar masih menggunakan teks.

Setiap teks selalu terkadung gaya penulisan dari penulisnya. Gaya penulisan inilah yang dapat menjadi bahan penelitian dalam bidang teks forensik. Penelitian teks forensik dibidang teknologi pun semakin banyak digelar. Sehingga sistem untuk membantu menanggulangi masalah banyaknya teks tidak bertanggung jawab semakin memungkinkan untuk dibuat.

Dengan menggunakan metode pengelompokan K-Means dan memanfaatkan koefisien silhouette penulis mencoba membangun sistem yang dapat membantu menanggulangi masalah tersebut dengan mengelompokan dokumen berdasarkan gaya penulisannya. Dengan harapan dapat memberikan gambaran mengenai jumlah penulis yang tidak bertanggung jawab dari sejumlah dokumen yang diinput.

Dengan menggunakan fitur variasi dan frekuensi bigram trigram *stopword*, variasi dan frekuensi kata slang, prosentasi jumlah *stopword* per non *stopword*, jumlah kalimat langsung dalam dokumen, dan rata rata kalimat dalam paragraph. Hasil yang didapat untuk pengelompokan ini dinilai kurang baik dan perlu mencari fitur lain yang lebih baik agar dapat merepresentasikan dokumen yang di uji.

Kata Kunci : clustering, gaya penulisan, k-means, Indonesia,

BAB 1
PENDAHULUAN

1.1. Latar Belakang

Mudahnya membuat identitas palsu atau menggunakan identitas orang lain di dunia maya. Membuat tindakan tidak bertanggung jawab hingga kriminal sering terjadi di dunia maya. Hal tersebut terjadi karena identitas palsu membuat pelaku merasa lebih aman dalam melakukan tindakannya. Salah satu contoh ialah Saracen. Berdasar BBC (2017) dan Sohuturon, M. (2017) Saracen adalah sebuah kelompok yang menerima bayaran untuk menyebarkan posting berisi ujaran kebencian di dunia maya. Dengan menggunakan website-website berita dan opini serta lebih dari 800.000 akun sosial media. Ketika ditangkap ternyata Saracen hanya di kelola oleh 3 orang. Salah seorang pelaku mengaku menggunakan beberapa akun palsu untuk mengelola sejumlah grup dan mengambil alih akun milik orang lain untuk menyebarkan ujaran kebencian.

Meskipun menggunakan banyak akun palsu, ujaran ujaran kebencian yang disebar tersebut berupa sebuah tulisan atau teks. Teks sendiri dapat menjadi sumber infromasi karena dalam teks dapat mengandung informasi infromasi diluar isi teks itu sendiri. Salah satu informasi yang ada dalam sebuah teks atau tulisan adalah gaya penulisan. Gaya penulisan adalah kecenderungan seseorang dalam menyampaikan opini atau gagasan kepada pembaca. Kecenderungan tersebut dapat berupa pemilihan kata, susunan kata, panjang kalimat atau hal hal lain seperti penggunaan tanda baca dan sebagainya. Hal tersebut dapat menjadi informasi yang terkandung dalam sebuah teks dengan atau tanpa disadari oleh penulisnya. Gaya penulisan inilah yang dapat menjadi bahan penelitan dalam tes teks forensik.

Penelitian teks forensik dibidang teknologi semakin banyak digelar diberbagai negara. Oleh karena itu penelitian teks forensik dibidang teknologi dirasa perlu. Sehingga dapat membuat sebuah sistem yang dapat membantu test teks forensik yang berbahasa Indonesia. Dalam penelitian ini, penulis ingin mengembangkan sebuah sistem pengujian teks berbasis teks forensik. Sistem ini

diharap dapat membantu memecahkan permasalahan-permasalahan seperti kasus Saracen. Hal tersebut dilakukan dengan mengelompokan (*clustering*) dokumen dengan metode K-means berdasarkan gaya penulisannya (*stylometry)*. Dengan begitu meskipun teks yang tersebar di dunia maya dishare oleh ratusan ribu akun. Jika dikelompokan berdasar gaya penulisnya hanya terbentuk 1 atau beberapa kelompok saja. Dimungkinkan teks tersebut ditulis oleh segelintir orang saja.

1.2. Rumusan Masalah

Adapun rumusan masalah dalam penelitian ini sebagai berikut

- Fitur gaya penulisan (*stylometry)* apa saja yang dapat merepresentasikan gaya penulisan seseorang dalam dokumen berbahasa Indonesia ?
- Apakah K-means dapat digunakan untuk mengelompokan dokumen berdasar penulisnya?

1.3. Batasan Masalah

Dalam penelitian ini terdapat batasan batasan masalah sebagai berikut:

a. Dokumen yang di ambil berupa artikel artikel dari website opini yaitu Seword.com

b. Dokumen yang akan digunakan adalah dokumen berbahasa Indonesia

c. Setiap dokumen yang digunakan ditulis oleh 1 penulis. Hal tersebut di pastikan secara manual dengan pembuatan meta data dokumen

1.4. Tujuan Penelitian

Penelitian ini bertujuan mengetahui fitur yang tepat, cara pendeteksian berdasar fitur yang didapat pada dokumen berbahasa Indonesia. Sebab hal tersebut dapat membantu pengembangkan sebuah sistem berbasis teks forensik yang

berguna untuk membantu pengidentifikasian atau setidak tidaknya membantu pengelompokan dokumen berdasar gaya penulisan yang digunakan.

1.5. Manfaat Penelitian

Dalam penelitan ini adapun manfaat manfaat yang ingin dicapai ialah:

1.5.1. Manfaat Umum:

Dengan adanya penelitian ini diharap mampu mendorong terbentuknya. sebuah sistem berbasis teks forensik yang bertujuan untuk melawan kejahatan dunia maya dalam teks teks berbahasa Indonesia.

1.5.2. Manfaat Pengguna:

Dengan adanya penelitian ini akan dihasilkan sebuah korpus yang dapat digunakan untuk penelitian penelitian serupa. Sehingga dapat membantu pengguna untuk melakukan penelitian lebih lanjut.

1.6. Metodologi Penelitian

Dalam penelitian ini metode metode yang akan digunakan adalah sebagai berikut:

1.6.1. Metode Pengumpulan data

Metode pembangunan korpus yang akan dilakukan adalah secara manual menyalin artikel-artikel pada situs opini seword.com.

1.6.2. Metode Pembangunan Fitur

1. SPIMI

   Single Pass In Memori *Indexing* (SPIMI) digunakan untuk mengindeks dokumen dokumen yang akan di proses.

2. Pra-pemrosesan

   Pra-pemrosesan yang akan dilakukan meliputi normalisasi teks, penghapusan *stopword*, dan segmentasi dokumen per kalimat.

3. Frekuensi N-gram

Frekuensi N-gram akan digunakan untuk menentukan fitur fitur yang menjadi representasi dari gaya penulisan penulis dalam sebuah dokumen. Jumlah N yang akan digunakan adalah 2 dan gram yang digunakan adalah kata.

1.6.3. **Metode Pengelompokan**

1. K-means

Metode K-means digunakan untuk pengelompokan dokumen-dokumen. Jarak antar dokumen akan dihitung dengan *cosine similarity* fitur tiap dokumen. Untuk mengoptimalkan hasil dari pengelompokan k-means akan di gunakan koefisien silhouette.

1.6.4. **Metode Evaluasi**

1. Bcubed Precision

Metode penghitungan Bcubed Precision digunakan untuk mengukur tingkat ketepatan dokumen yang di ambil atau ditunjuk pada setiap cluster secara keseluruhan sistem.

2. Bcubed Recall

Metode penghitungan Bcubed Recall digunakan untuk mengukur tingkat kesesuaian informasi dari dokumen yang diambil atau ditunjuk pada setiap cluster secara keseluruhan sistem.

1.7. Sistematika Penulisan

Untuk memudahkan dalam mendapatkan gambaran yang lengkap dan jelas mengenai penelitian yang akan dilakukan, penulis membagi laporan ini menjadi 5 (lima) bab yaitu Bab 1 Pendahuluan, Bab 2 Tinjauan Pustaka, Bab 3 Analisis dan Perancangan Sistem, Bab 4 Implementasi dan Analisis Sistem, dan Bab 5 Kesimpulan dan Saran.

BAB 1, bab ini berisi penjelasan mengenai pendahuluan dari penelitian yang meliputi latar belakang, tujuan, manfaat, batasan masalah, metode penelitian dan sistematika penulisan penyusunan laporan penelitian. Pada bab ini terangkum berbagai kebutuhan yang muncul sehingga menimbulkan alasan untuk membuat penelitian. Garis besar dan manfaat dari penelitian juga dicantumkan dalam bab ini.

BAB 2, bab ini berisi tentang tinjauan pustaka serta landasan teori yang diperlukan untuk memecahkan masalah dalam penelitian yang dilakukan. Teori yang diambil dari beberapa kutipan buku, yang berupa pengertian dan definisi. Bab ini juga menjelaskan konsep dasar sistem dan definisi lainnya yang berkaitan dengan sistem yang.akan dibuat.

BAB 3, bab ini berisi perancangan sistem yaitu tentang analisis teori yang digunakan dalam penelitian, uraian tentang variabel dan data yang akan dikumpulkan dan bagaimana menerapkannya ke dalam sistem yang akan dibuat..

BAB 4, bab ini berisi tentang hasil penelitian atau implementasi serta pembahasan/analisis dari penelitian yang telah dilakukan dan dijelaskan secara terpadu.

BAB 5, bab ini berisi kesimpulan dari sistem yang telah dibuat dan saran yang akan berguna untuk pengembangan sistem selanjutnya. Dengan adanya saran, diharapkan penelitian yang dilakukan selanjutnya akan menghasilkan hasil yang lebih baik.

Selain berisi bab-bab utama tersebut, penelitian ini dilengkapi juga dengan intisari, daftar isi, daftar gambar, daftar tabel, daftar pustaka, dan lampiran.

BAB 5
Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan penelitian ini dapat disimpulkan bahwa fitur-fitur yang telah dipilih kurang mampu mewakili gaya penulisan dari dokumen dokumen yang ada. Sehingga hasil dari sistem tidak dapat menunjukan kelompok berdasarkan gaya penulisannya. Namun dari Penelitian ini juga dapat di simpulkan bahwa fitur *stopword* n-gram berkontribusi besar dalam penentuan bentuk pengelompokan yang terjadi. Fitur kata slang dan ditur rata rata jumlah kalimat dapat dipertimbangkan sebagai fitur yang dapat digunakan untuk mengelompokan dokumen berdasarkan gaya penulisannya.

Pengelompokan dokumen dengan metode K-means yang memanfaatkan koefisien silhouette untuk menentukan nilai k dapat dikatakan berjalan dengan baik. dalam beberapa kondisi dapat menghasilkan jumlah cluster yang mendekati jumlah penulis sesungguhnya. Namun hasil dari pengelompokan ini masih jauh dari kata memuaskan sebab hasil pengelompokan sangat bergantung seberapa baik fitur yang dipilih. Sehingga perlu dilakukan penelitian lanjut dengan fitur yang lebih baik untuk dapat memutuskan apakah pengelompokan dokumen dengan metode K-means yang memanfaatkan koefisien silhouette ini benar benar baik untuk menentukan nilai k.

5.2. Saran

Dalam Penelitian ini terdapat beberapa hal yang perlu dikembangkan lebih lanjut. Terutama pada pemilihan fitur yang akan digunakan untuk mewakili dokumen. Untuk fitur *stopword* N-gram, kata slang dan rata rata panjang kalimat dalam paragraph dapat digunakan. Namun perlu percobaan lebih lanjut untuk mengetahui penerapan yang lebih baik seperti menambah batasan berupa jarak antar *stopword* dalam dokumen tidak lebih dari jumlah kata tertentu (Contoh jarak antar kata tidak 5 – 8 kata).

Diharap pada penelitian kedepannya, dapat digunakan fitur lain yang mungkin lebih merepresentasikan gaya penulisan dalam sebuah dokumen seperti jumlah dan variansi kata ganti orang pertama, jumlah dan variansi kata ganti orang kedua, atau rata rata jumlah kata dalam setiap paragraph. terutama jika sudah memungkinkan untuk membentuk fitur dengan stylometry jenis fitur semantik.

DAFTAR PUSTAKA

Bagnall, D. (2016). Authorship clustering using multi-headed recurrent neural networks. *arXiv preprint arXiv:1608.04485*.

BBC. (2017, September 13). *Kasus Saracen: Pesan kebencian dan hoax di media sosial 'memang terorganisir'* . Retrieved from BBC Indonesia: http://www.bbc.com/indonesia/trensosial-41022914

Bird, S., & Loper, E. (2004). NLTK: the natural language toolkit. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. *Association for Computational Linguistics*, p. 31.

Cali. K. & Kim Bowen. (2017, September 15). *Five features of effective writing.* Retrieved from LEARN NC: http://www.learnnc.org/lp/editions/few/684

Canales, O. et al. (2011). A stylometry system for authenticating students taking online tests. P. of Student-Faculty Research Day, Ed. *CSIS. Pace University*.

Cherny. (2017, 08 21). *Cosine Distance as Similarity Measure in KMeans.* Retrieved from stackexchange: https://stats.stackexchange.com/q/299016

Krisnawati, L. D., & Schulz, K. U. (2013, Desember). Plagiarism detection for Indonesian texts. In Proceedings of International Conference on Information Integration and Web-based Applications & Services. *ACM*, p. 595.

Kuznetsov, M., Motrenko, A., Kuznetsova, R., & Strijov, V. (2016). Methods for Intrinsic Plagiarism Detection and Author Diarization. *In CLEF*, pp. 912-919.

Luyckx, K., & Daelemans, W. (2008, Agustus). Authorship attribution and verification with many authors and limited data. *In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pp. 513-520.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to information retrieval.* Cambridge: Cambridge University Press.

Miao, Y., Kešelj, V., & Milios, E. (2005). Document clustering using character N-grams: a comparative evaluation with term-based and word-based clustering. *ACM* (pp. 357-358). ACM.

Pedregosa, F. et al. (2011, Oktober 12). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, pp. 2825-2830.

Potthast M., Stein B., Barr´on-Cede˜no A., & Rosso P. (2010). An Evaluation Framework for Plagiarism Detection. *Coling 2010*, pp. 997–1005.

Sari, Y., & Stevenson, M. (2016). Exploring Word Embeddings and Character N-Grams for Author Clustering. *In CLEF*, pp. 984-991.

Sohuturon, M. (2017, Agustus 23). *Polisi Tangkap Pengelola Grup 'Saracen' Penyebar Kebencian.* Retrieved from CNN Indonesia: http://www.cnnindonesia.com/nasional/20170823141007-12-236690/polisi-tangkap-pengelola-grup-saracen-penyebar-kebencian/

Stamatatos, E., et al. (2016). Clustering by Authorship Within and Across Documents. *In CLEF*, pp. 691-715.

Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., & Potthast, M. (2017). Overview of the Author Identification Task at PAN 2017: Style Breach Detection and Author Clustering. *Working Notes Papers of the CLEF*.

Vartapetiance, A., & Gillam, L. (2016, September 5-8). A Big Increase in Known Unknowns: from Author Verification to Author Clustering-Notebook for PAN at CLEF 2016. *In Working Notes of CLEF*, pp. 1008-1013.

Wu, J. (2016). *Advances in K-means clustering: a data mining thinking.* London: Springer Science & Business Media.

Yendra, S. S. (2016). *Mengenal Ilmu Bahasa (Linguistik)* . Deepublish.