TUGAS AKHIR – KS141501

RANCANG BANGUN PERANGKAT LUNAK TEENSTAGRAM UNTUK MENGELOMPOKKAN TOPIK CAPTION AKUN INSTAGRAM DENGAN PERMODELAN AUTHOR TOPIC MODELS (STUDI KASUS: SISWA SMA DI SURABAYA)

TEENSTAGRAM APPLICATION FOR TOPIC CAPTION CLASSIFICATION FROM THE INSTAGRAM ACCOUNTS WITH AUTHOR TOPIC MODELS (CASE STUDY: HIGH SCHOOL STUDENTS IN SURABAYA)

FAIZ NUR FITRAH INSANI
NRP 5214 100 131

Dosen Pembimbing
1. Nur Aini Rakhmawati, S.Kom, M.Sc.Eng, Ph.D
2. Irmasari Hafidz, S.Kom, M.Sc.

Departemen Sistem Informasi
Fakultas Teknologi Informasi dan Komunikasi
Institut Teknologi Sepuluh Nopember
Surabaya 2018

Halaman ini sengaja dikosongkan

TUGAS AKHIR – KS141501

RANCANG BANGUN PERANGKAT LUNAK TEENSTAGRAM UNTUK MENGELOMPOKKAN TOPIK CAPTION AKUN INSTAGRAM DENGAN PERMODELAN AUTHOR TOPIC MODELS (STUDI KASUS: SISWA SMA DI SURABAYA)

FAIZ NUR FITRAH INSANI
NRP 5214100131

Dosen Pembimbing
1. Nur Aini Rakhmawati, S.Kom, M.Sc.Eng, Ph.D
2. Irmasari Hafidz, S.Kom, M.Sc.

Departemen Sistem Informasi
Fakultas Teknologi Informasi dan Komunikasi
Institut Teknologi Sepuluh Nopember
Surabaya 2018

Halaman ini sengaja dikosongkan

UNDERGRADUATE THESIS – KS141501

TEENSTAGRAM APPLICATION FOR TOPIC CAPTION CLASSIFICATION FROM THE INSTAGRAM ACCOUNTS WITH AUTHOR TOPIC MODELS (CASE STUDY: HIGH SCHOOL STUDENTS IN SURABAYA)

FAIZ NUR FITRAH INSANI
NRP 5214100131

Supervisor
1. Nur Aini Rakhmawati, S.Kom, M.Sc.Eng, Ph.D
2. Irmasari Hafidz, S.Kom, M.Sc.

Departement of Information System
Faculty of Information Communication and Technology
Institut Teknologi Sepuluh Nopember
Surabaya 2018

Halaman ini sengaja dikosongkan

LEMBAR PENGESAHAN

RANCANG BANGUN PERANGKAT LUNAK TEENSTAGRAM UNTUK MENGELOMPOKKAN TOPIK CAPTION AKUN INSTAGRAM DENGAN PERMODELAN AUTHOR TOPIC MODELS (STUDI KASUS: SISWA SMA DI SURABAYA)

TUGAS AKHIR

Disusun Untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer
pada
Departemen Sistem Informasi
Fakultas Teknologi Informasi dan Komunikasi
Institut Teknologi Sepuluh Nopember

Oleh:

FAIZ NUR FITRAH INSANI
NRP. 52 14 100 131

Surabaya, 16 Januari 2018

Plh KEPALA
DEPARTEMEN SISTEM INFORMASI

Edwin Riksakomara, S.Kom, MT
NIP 196907252003121001

Halaman ini sengaja dikosongkan

LEMBAR PERSETUJUAN

RANCANG BANGUN PERANGKAT LUNAK TEENSTAGRAM UNTUK MENGELOMPOKKAN TOPIK CAPTION AKUN INSTAGRAM DENGAN PERMODELAN AUTHOR TOPIC MODELS (STUDI KASUS: SISWA SMA DI SURABAYA)

TUGAS AKHIR

Disusun Untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer
pada
Departemen Sistem Informasi
Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember

Oleh:

FAIZ NUR FITRAH INSANI
NRP. 5214 100 131

Disetujui Tim Penguji: Tanggal Ujian: 11 Januari 2018

Periode Wisuda: Maret 2018

Nur Aini Rakhmawati, S.Kom, M.Sc.Eng, Ph.D **(Pembimbing 1)**

Irmasari Hafidz, S.Kom, M.Sc. **(Pembimbing 2)**

Faizal Johan Atletiko, S.Kom., M.T **(Penguji 1)**

Radityo Prasetianto Wibowo, S.Kom, M.Kom **(Penguji 2)**

Halaman ini sengaja dikosongkan

RANCANG BANGUN PERANGKAT LUNAK TEENSTAGRAM UNTUK MENGELOMPOKKAN TOPIK CAPTION AKUN INSTAGRAM DENGAN PERMODELAN AUTHOR TOPIC MODELS (STUDI KASUS: SISWA SMA DI SURABAYA)

Nama Mahasiswa : Faiz Nur Fitrah Insani
NRP : 5214100131
Departemen : Sistem Informasi FTIK-ITS
Dosen Pembimbing :
1. Nur Aini Rakhmawati, S.Kom, M.Sc.Eng, Ph.D
2. Irmasari Hafidz, S.Kom, M.Sc

ABSTRAK

Menurut survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) [2] bahwa penetrasi pengguna internet di Indonesia adalah pelajar sebanyak 69.8 persen dari 132,7 juta pengguna, dan 97,4 persen menggunakannya untuk media sosial. Konten dari media sosial yang sering dikunjungi salah satunya adalah Instagram yaitu 15 persen dengan data sebanyak 19.9 juta. Hal tersebut menunjukkan bahwa pelajar yang memakai media sosial tidak lagi hanya untuk mengikuti tren, melainkan menjadi sebuah kebutuhan untuk menunjukkan eksistensi dirinya kepada publik. Tidak hanya menunjukkan foto ataupun lokasi kegiatan para pelajar, mereka juga menunjukkan beberapa ungkapan emosi dan perasaan yang dituang dalam caption Instagram. Dari fenomena tersebut dibutuhkan sebuah platform yang mampu memberikan informasi visual terhadap aktivitas pelajar dalam hal berekspresi di sosial media Instagram dengan melakukan analisis topic modeling terhadap perilaku dan kebiasaan pelajar ketika mengunggah gambar beserta caption menggunakan metode Author-Topic Models atau ATM. Penelitian ini dikhususkan untuk menganalisa data caption akun Instagram siswa SMA di Surabaya, setelah data

dididapatkan serta dianalisa menggunakan Author-Topic Models kemudian dilakukan visualisasi terhadap topik dan atau kategori siswa berdasarkan caption dari masing – masing sekolah. Melalui proses pembuatan model, telah didapatkan hasil terbaik berupa 6 topik. Adapun 6 topik tersebut dapat dikatakan baik karena memiliki nilai perplexity yang kecil setelah dilakukan percobaan 30 kali. 6 Topik yang terbentuk dianalisis dan diterjemahkan ke dalam label kategori, yaitu ***perasaan****,* ***fotografi****,* ***fotografi dan artis****,* ***event Surabaya****,* ***liburan****, dan* ***agama dan musik.*** *Masing – masing topik memiliki hasil topik* ***perasaan*** *dibahas oleh* ***3 sekolah****, topik* ***event surabaya*** *dibahas oleh* ***2 sekolah****, topik* ***fotografi*** *dibahas oleh* ***6 sekolah****, topik* ***fotografi dan artis*** *dibahas oleh* ***3 sekolah****, topik* ***liburan*** *dibahas oleh* ***3 sekolah****, dan topik* ***agama dan musik*** *dibahas oleh* ***1 sekolah*** *dengan jumlah data caption* ***3346*** *dari* ***18*** *sekolah.*

Kata Kunci: Instagram, Caption, Topic Modeling, SMA, Author-Topic Models, ATM.

RANCANG BANGUN PERANGKAT LUNAK TEENSTAGRAM UNTUK MENGELOMPOKKAN TOPIK CAPTION AKUN INSTAGRAM DENGAN PERMODELAN AUTHOR TOPIC MODELS (STUDI KASUS: SISWA SMA DI SURABAYA)

Nama Mahasiswa : Faiz Nur Fitrah Insani
NRP : 5214100131
Departemen : Sistem Informasi FTIK-ITS
Dosen Pembimbing :
1. Nur Aini Rakhmawati, S.Kom, M.Sc.Eng, Ph.D
2. Irmasari Hafidz, S.Kom, M.Sc

ABSTRACT

According to a survey conducted by the Association of Indonesian Internet Service Providers (APJII) [2] that internet user penetration in Indonesia is 69.8 percent of 132.7 million are students, and 97.4 percent use it for social media. Content from social media that often visited one of them is Instagram that have 15 percent with data as much as 19.9 million. It shows that students who use social media are no longer just to follow trends, but become a necessity to show their existence to the public. They not only show photos or locations of their activities, but also show their emotional expressions and feelings in Instagram captions. From the phenomenon requires a platform that is able to provide visual information on the activities of students in terms of expression in social media Instagram by doing topic modeling, analysis of student behavior and habits when uploading images and captions using the method of Author-Topic Models or ATM. This study was devoted to analyze data caption of Instagram account of high school students in Surabaya, after the data was obtained and analyzed using Author-Topic Models then visualization of the topic or categories of students based on the caption of each

*school. Through the modeling process, the best results have been obtained in the form of **6 topics**. The 6 topics can be said good because it has a small perplexity value after the experiment was done 30 times. 6 The topics formed are analyzed and translated into category labels, namely **feelings**, **photography**, **photography and celebrity**, **Surabaya events**, **holidays**, and **religion and music**. Each topic has the topic of **feelings** discussed by **3 schools**, the topic of the **Surabaya events** discussed by **2 schools**, **photography** topics discussed by **6 schools**, **photography topics and artists** discussed by **3 schools**, **holiday topics** discussed by **3 schools**, and **religious and music** topics discussed by **1 school** with **3346** data caption data from 18 schools.*

Keywords: Instagram, Caption, Topic Modeling, High School, Author-Topic Models, ATM.

KATA PENGANTAR

Puji syukur penulis haturkan kehadirat Allah SWT yang telah memberikan anugerah dan tuntunan kepada penulis sehingga penulis dapat menyelesaikan tugas akhir dengan judul "RANCANG BANGUN PERANGKAT LUNAK TEENSTAGRAM UNTUK MENGELOMPOKKAN TOPIK CAPTION AKUN INSTAGRAM DENGAN PERMODELAN AUTHOR TOPIC MODELS (STUDI KASUS: SISWA SMA DI SURABAYA)" sebagai salah satu syarat kelulusan pada Departemen Sistem Informasi, Fakultas Teknologi Informasi dan Komunikasi, Institut Teknologi Sepuluh Nopember Surabaya. Penyusunan tugas akhir ini senantiasa mendapatkan dukungan dari berbagai pihak baik dalam bentuk doa, motivasi, semangat, kritik, saran dan berbagai bantuan lainnya. Untuk itu, secara khusus penulis akan menyampaikan ucapan terima kasih yang sedalam-dalamnya kepada:

1. Allah SWT yang telah memberikan hidayah serta atas ijin-Nya pula saya mampu mengerjakan tugas akhir ini hingga selesai.
2. Segenap keluarga besar penulis, terutama kedua orang tua, Bapak Suwandi dan Ibu Dra. Siti Nur Hasanah yang selalu senantiasa mendoakan, memberikan motivasi, dan kebutuhan materiil maupun non-materiil sehingga penulis mampu untuk menyelesaikan pendidikan S1 ini dengan baik.
3. Bapak Dr. Ir. Aris Tjahyanto, M.Kom., selaku Ketua Jurusan Sistem Informasi ITS, Bapak Nisfu Asrul Sani, S.Kom, M.Sc selaku KaProdi S1 Sistem Informasi ITS serta seluruh dosen pengajar beserta staf dan karyawan di Jurusan Sistem Informasi, FTIF ITS Surabaya selama penulis menjalani kuliah.
4. Ibu Nur Aini Rakhmawati, S.Kom., M.Sc., Eng. Ph.D dan Ibu Irmasari Hafidz, S.Kom, M.Sc selaku dosen

pembimbing yang telah banyak meluangkan waktu untuk membimbing, mengarahkan, dan mendukung dengan memberikan ilmu, petunjuk, dan motivasi dalam penyelesaian Tugas Akhir dan selama menjalankan masa perkuliahan penulis.

5. Bapak Faizal Johan Atletiko, S.Kom., M.T dan Bapak Radityo Prasetianto Wibowo, S.Kom, M.Kom selaku dosen penguji yang telah memberikan kritik, saran, dan masukan yang dapat menyempurnakan Tugas Akhir ini.
6. Keluarga besar E-Home yang telah memberikan pelajaran, hitam dan putihnya dunia kampus ITS, dan mengajarkan bahwa kehidupan adalah dunia yang keras dan perlu kita perjuangkan. Canda, tawa, sedih, lelah, dan amarah selalu ada, namun kalian adalah yang terbaik.
7. Teruntuk anyung yang selalu sabar dan tabah, memberikan semangat yang tiada henti, memberikan motivasi, dan semangat untuk memulai kehidupan kembali, terima kasih atas segalanya.
8. Sesama pejuang 3.5 adit, nia, najib, azmi, UKM Cinta Lab dan seluruh stakeholder ADDI, yang telah memberikan sarana prasarana serta ilmu akademis.
9. Keluarga SRD terutama mbak ijah keluarga SRD 'Visioner' Ibu elva, hendro, tatan, kesha, enji, dan anak anak yang telah memberikan dukungan emotional dan doa selama ini.

Terima kasih atas segala bantuan, dukungan, serta doanya. Semoga Allah SWT senantiasa melimpahkan anugerah serta membalas kebaikan yang telah diberikan kepada penulis. Penulis menyadari bahwa masih terdapat kekurangan dalam penyusunan tugas akhir ini, oleh karena itu penulis mengharapkan saran dan kritik yang membangun demi kebaikan penulis dan tugas akhir ini. Akhir kata, penulis berharap bahwa tugas akhir ini dapat memberikan kebermanfaatan.

,

DAFTAR ISI

Halaman ini sengaja dikosongkan

DAFTAR GAMBAR

DAFTAR TABEL

Halaman ini sengaja dikosongkan

DAFTAR KODE

Halaman ini sengaja dikosongkan

BAB I
PENDAHULUAN

Pada bab ini, akan dijelaskan pengenai proses identifikasi masalah yang meliputi latar belakang masalah, perumusan masalah, batasan masalah, tujuan, manfaat, dan relevansi tugas akhir. Berdasarkan uraian pada bab ini, diharapkan gambaran umum permasalahan dan pemecahan masalah pada tugas akhir dapat dipahami.

1.1. Latar Belakang

Media sosial merupakan fenomena masyarakat yang sedang banyak dinikmati oleh banyak kalangan. Media sosial menyediakan cara berkomunikasi aktif bagi setiap penggunanya. Tidak hanya organisasi ataupun kelompok, media sosial juga bisa digunakan oleh individu tanpa batasan yang jelas. Meningkatnya jumlah pengguna media sosial juga dikarenakan faktor mudahnya akses bagi setiap orang dan semua kalangan dapat dengan mudah bergabung menjadi pengguna. Media sosial Instagram menjadi sangat populer akhir – akhir ini di seluruh dunia. Media sosial berbasis *mobile* dan *web* ini menghadirkan fitur untuk berbagi foto dan video secara mudah ditambah *filter* unik yang bisa digunakan secara instan pada foto yang akan di unggah. Pada *blog Instagram* sendiri, disebutkan bahwa pengguna *Instagram* mencapai 600 juta pada tahun 2016 dan akan terus bertambah setiap harinya [1]. Angka ini menunjukkan bahwa aplikasi *Instagram* merupakan media sosial yang cukup dibilang sangat sukses dibidangnya.

Menurut survey Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) [2] bahwa penetrasi pengguna internet di Indonesia adalah pelajar sebanyak 69.8 persen dari 132,7 juta pengguna, dan 97,4 persen menggunakannya untuk media sosial. Konten dari media sosial yang sering dikunjungi salah satunya adalah *Instagram* yaitu 15 persen dengan data sebanyak 19.9 juta. Hal

tersebut menunjukkan bahwa pelajar yang memakai media sosial tidak lagi hanya untuk mengikuti tren, melainkan menjadi sebuah kebutuhan untuk menunjukkan eksistensi dirinya kepada publik. Tidak hanya menunjukkan foto ataupun lokasi kegiatan para pelajar, mereka juga menunjukkan beberapa ungkapan emosi dan perasaan yang dituang dalam *caption Instagram*.

Seperti yang diungkapkan oleh Keke Mahardika [3] bahwa penggunaan media sosial *instagram* tentu membawa kemudahan bagi siswa untuk membangun komunikasi dan menampilkan dirinya kepada orang lain, akan tetapi instagram juga membawa dampak negatif seperti krisis percaya diri, persaingan kehidupan mewah, dan tidak mau menatap realita dan kenyataan. Hal tersebut menyebabkan pengaruh terhadap keadaan mental dalam pengunggahan foto dan caption dalam akun sosial media mereka.

Dari fenomena yang terjadi dikalangan remaja terhadap penggunaan media sosial, maka dibutuhkan sebuah *platform* yang mampu memberikan informasi visual mengenai aktivitas remaja dan cara mereka mengungkapkan ekspresi di media sosial Instagram, dengan cara melakukan analisis topic modeling terhadap perilaku atau kebiasaan remaja dalam upload gambar beserta caption tertentu, menggunakan metode *Author-Topic Models* yang merupakan perpanjangan dari *Latent Dirichlet Allocation*, yang biasa disebut ATM. Penelitian ini dikhususkan untuk menganalisa caption dari siwa sekolah menengah atas (SMA) dari 5 sekolah di Surabaya, yang menjadi target dari mata kuliah Etika Profesi di Jurusan Sistem Informasi Institut Teknologi Sepuluh Nopember Surabaya. Setelah akun dan caption didapatkan serta dianalisa menggunakan ATM, kemudian dilakukan visualisasi terhadap topik atau kategori aktivitas siswa berdasarkan captionnya. Melalui penelitian ini diharapkan memberikan informasi kepada masyarakat, khususnya orang tua dan guru untuk lebih mengetahui dan mengarahkan putra – putrinya agar dapat bertanggung jawab dalam hal berekspresi di media sosial,

sehingga pelanggaran dan dampak negatif dari yang dikhawatirkan dari media sosial dapat dihindari.

1.2. Rumusan Masalah

Berdasarkan latar belakang diatas, maka rumusan masalah yang akan diteliti pada tugas akhir ini adalah:

1. Bagaimana cara mengakuisisi data akun dan caption anak SMA di Surabaya melalui media sosial *instagram*?
2. Bagaimana cara merancang *platform* yang mampu melakukan visualisasi data topik pembicaraan setiap SMA di Surabaya ke dalam sebuah gambar atau *dashboard*?
3. Bagaimana cara mengelompokkan topik caption atau caption instagram dari setiap SMA di Surabaya?
4. Bagaimana melakukan pemodelan serta pelabelan data secara berkelanjutan pada periode tertentu?

1.3. Batasan Masalah

Batasan pemasalahan dalam pengerjaan tugas akhir ini adalah :

1. Studi kasus yang digunakan pada penelitian ini hanya meliputi 18 Sekolah Menengah Atas di wilayah Surabaya.
2. Pengambilan data dalam tugas akhir ini dilakukan hanya dari data *caption* akun *instagram* pelajar SMA di Surabaya tanpa mengambil data di kolom komentar.
3. Proses pembuatan model dengan Author-Topic Models digunakan hanya sebatas untuk visualisasi dan mengetahui topik pembicaraan setiap SMA.

1.4. Tujuan

Berdasarkan hasil perumusan masalah dan batasan masalah yang telah disebutkan sebelumnya, maka tujuan yang dicapai dari tugas akhir ini adalah untuk menciptakan sebuah *platform* yang mampu mengklasifikasi dan memvisualkan perilaku atau

aktivitas pelajar SMA di media sosial instagram guna mendorong para orang tua atau guru untuk lebih memperhatikan putra-putrinya dalam hal berekspresi di media sosial. Sehingga dengan adanya *platform* ini diharapkan para pelajar SMA di Surabaya mampu berekspresi di media sosial dengan lebih bertanggung jawab.

1.5. Manfaat

Manfaat yang diharapkan dapat diperoleh dari tugas akhir ini adalah:

1.1. Memfasilitasi orang tua dan guru dalam mengawasi pergaulan pelajar SMA di Surabaya.
1.2. Memfasilitasi mahasiswa, khususnya Jurusan Sistem Informasi untuk mempelajari *social media analysis*.
1.3. Menyediakan data yang dapat digunakan sebagai acuan untuk menentukan tindakan lebih lanjut dalam hal kampanye penggunaan *smartphone* dan *social media* yang bertanggung jawab kepada pelajar di wilayah Surabaya.

1.6. Relevansi

Tugas akhir ini berkaitan dengan mata kuliah Pemrograman Berbasis Web, Analisa dan Desain Perangkat Lunak dan Konstruksi Pengembangan Perangkat Lunak, Penggalian Data Analitika Bisnis, Pemrograman Integratif, dan Etika Profesi.

BAB II
TINJAUAN PUSTAKA

Bab ini akan menjelaskan mengenai penelitian sebelumnya dan dasar teori yang dijadikan acuan atau landasan dalam pengerjaan tugas akhir ini. Landasan teori akan memberikan gambaran secara umum dari landasan penjabaran tugas akhir ini.

2.1. Studi Sebelumnya

Pada subbab ini dijelaskan tentang referensi penelitian yang berkaitan dengan tugas akhir. Pada bagian ini memaparkan acuan penelitian sebelumnya yang digunakan oleh penulis dalam melakukan penelitiannya.

1. Penelitian pertama mengenai *Topic Modeling in Twitter: Aggregating Tweets by Conversations* oleh David Alvarez dan Martin Saveski [12]. Pada penelitian ini, penulis meneliti tentang penggunaan Topic Modeling terhadap dokumen yang kecil / pendek dan susah untuk diinterpretasikan, yaitu tweet conversation dari media sosial *Twitter*. Metode yang digunakan adalah *Latent Dirichlet Allocation* (LDA) dan *Author-topic Model*s (ATM), sehingga dapat dibandingkan hasil perplexity dari keduanya ketika diuji menggunakan dataset tweet dan dibagikan dalam beberapa kategori. Kategorinya adalah berdasarkan *conversation*, atau percakapan dalam tweet, Hashtag yang digunakan, *users*, dan *tweet* itu sendiri. Hasilnya, pada *conversation*, penggunaan metode ATM dinilai lebih baik dibanding semua metode, untuk lama waktu pengerjaan ATM dinilai paling lama setelah dilakukannya iterasi berulang kali menggunakan *Gibbs Sampling iterations*.

2. Penelitian kedua mengenai *Rancang Bangun Perangkat Lunak Teenstagram untuk Mengelompokkan Topik Caption Akun Instagram Siswa Sekolah Menengah Pertama di Surabaya*, oleh Tetha Valianta [18]. Penulis meneliti tentang karakteristik siswa SMP di Surabaya dalam menggunakan Instagram dengan bentuk visualisasi dashboard, khususnya dalam pengunggahan foto beserta *caption*. Selanjutnya dilakukan *Topic Modeling* dari *caption* yang sudah di crawling menggunakan metode *Latent Dirichlet Allocation* (LDA). Hasil dari *Topic Modeling* berupa penerjemahan seluruh caption menjadi dua topik, yaitu edukasi dan interaksi karena dinilai memiliki perplexity yang paling kecil dibanding topik lainnya. Penggunaan crawler dengan menggunakan Integrasi PHP juga dinilai mampu menjawab kebutuhan visualisasi data caption akun siswa SMP.
3. Penelitian ketiga adalah *Discovering User Interest on Twitter with a Modified Author-Topic Model* oleh Zhiheng Xu, Rong Lu, Liang Xiang, Qing Yang pada tahun 2011[19]. Pada penelitian ini, dilakukan *Topic Modeling* menggunakan metode *Author-Topic Models* yang telah dimodifikasi. Tujuan dari penelitian ini adalah menghubungkan antara *tweet* dengan penggunanya pada *Twitter*. *Dataset* yang digunakan adalah *tweet* dan *user* pada *Twitter*, kemudian dilakukan perbandingan dengan model yang lain menggunakan pengukuran performa perplexity, model tersebut adalah *Latent Dirichlet Allocation* (LDA), *Author-Topic Models*, dan *Modified Author-Topic Models*. Hasilnya, penelitian ini sukses dengan mengungguli semua metode dengan *perplexity* paling rendah.

2.2. Dasar Teori

Berisi teori-teori yang mendukung serta berkaitan dengan tugas akhir yang sedang dikerjakan.

2.2.1. Instagram

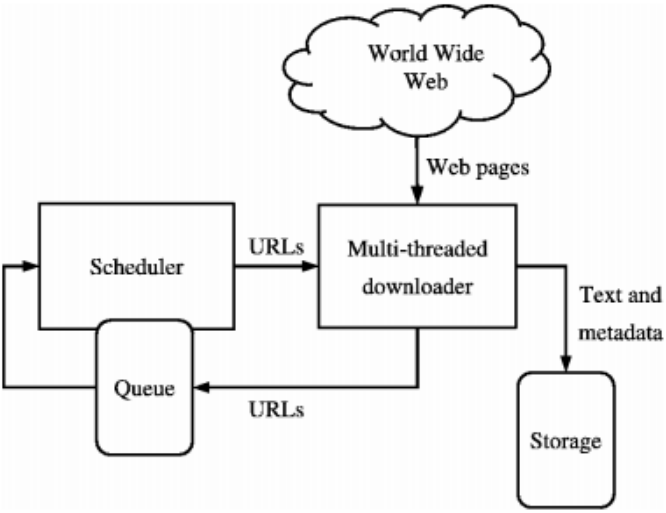
Instagram adalah cara menyenangkan dapat berbagi foto kepada orang lain melalui *mobile-based* and *web-based application*. *Instagram* memberikan fitur unggah dan edit foto secara instan menggunakan filter – filter menarik yang telah disediakan [4]. Setiap foto yang di unggah menggunakan *caption* layaknya *microblogging* lainnya dan tidak dibatasi oleh batasan karakter. *Instagram* telah mencapai lebih 600 juta pada tahun 2016 dan akan terus meningkat [1].

Instagram juga menyediakan *Application Programming Interface* (API) khusus untuk para *developer* yang ingin mengembangkan aplikasinya menggunakan *Instagram*, disamping membantu perseorangan berbagi kontennya sendiri, API ini juga berguna untuk advertising dan branding [6]. Dalam Instagram ada beberapa batasan yang perlu diperhatikan [7]:

1. Batas jumlah orang yang dapat diikuti *following* oleh satu akun adalah 7500.
2. Batas untuk melakukan like sebanyak 350 per hari.
3. Batas tagar (*hashtag* yang diperbolehkan adalah 30 *hashtag* per *post*.
4. Batas karakter untuk Biodata adalah 150 karakter dan
5. Batas karakter untuk *Caption* adalah 2200 karakter.

2.2.2. Web Crawler

Web Crawler atau biasa disebut *robots*, *spiders*, *worms*, *walkers*, dan *wanderers* [13] adalah suatu program atau *script* otomatis yang relatif sederhana dengan metode tertentu dapat melakukan proses pemindaian ke halaman – halaman web dan membuat indeks untuk mempermudah pencarian.

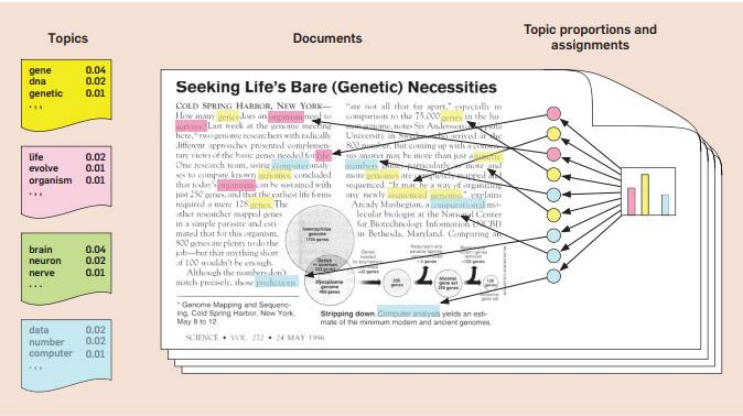


Gambar 2.1 Arsitektur Web Crawler [14]

Web Crawler sendiri memiliki arsitektur seperti gambar 2.1 [14], Informasi didapatkan melalui *World Wide Web* (WWW) atau halaman *web*, kemudian menggunakan *multi-threaded downloader* untuk mengunduh data yang dibutuhkan, dengan bantuan *scheduler* dan *URL Queue* data dapat dijadwalkan sesuai dengan kebutuhan. *Text* dan *metadata* yang sudah di *crawl* akan dimasukan *storage* untuk disimpan.

2.2.3. Topic Modeling

Topic Modeling merupakan algoritma untuk menemukan tema utama yang mencakup sekumpulan dokumen yang besar dan tidak terstruktur. *Topic Modeling* mengatur kumpulan dokumen sesuai dengan tema yang sudah ditermukan. Algoritmanya dapat diterapkan pada kumpulan dokumen yang masif. *Topic Modeling* juga dapat disesuaikan dengan berbagai jenis data, dalam beberapa aplikasinya algoritma ini digunakan untuk mencari pola dalam data genetik, gambar, dan jejaring sosial [15].

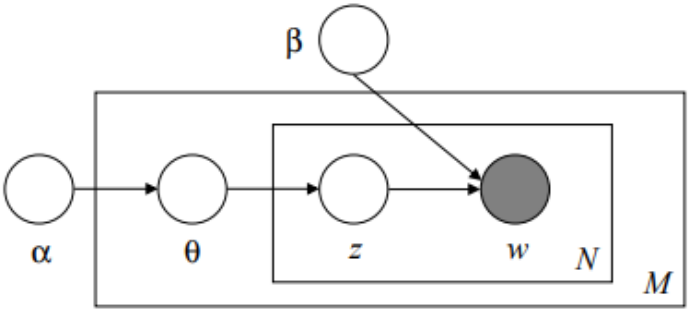


Gambar 2.2 Konsep Topic Modeling [15]

Konsep *Topic Modeling* menurut blei ditunjukan pada gambar 2.2 Algoritma dari *topic modeling* menganalisis kata – kata dari teks asli untuk menemukan topik atau tema yang berada diantara teks tersebut, bagaimana teks dapat saling terhubung satu sama lain, dan bagaimana tema tersebut dapat berubah seiring berjalannya waktu, sehingga dapat dikembangkan untuk dilakukannya pencarian ataupun peringkasan teks dalam dokumen [15].

2.2.4. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) adalah Teknik yang digunakan untuk mengidentifikasi topik tersembunyi dari kumpulan dokumen yang besar [10]. LDA juga merupakan sebuah metode statistika yang digunakan sebagai model dalam menganalisis sebuah dokumen. LDA dapat digunakan untuk meringkas, melakukan klasterisasi, menghubungkan maupun memroses data yang sangat besar karena LDA menghasilkan topik yang diberi bobot untuk masing – masing dokumen [8].



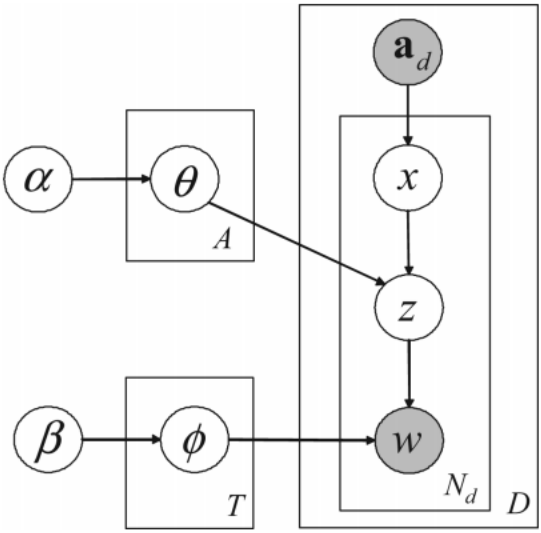
Gambar 2.3 Model Latent Dirichlet Allocation [8]

Blei merepresentasikan Model LDA seperti gambar 2.3 [8]. Pada gambar tersebut LDA merepresentasikan tiga tingkatan. Parameter α dan β adalah parameter untuk tingkatan korpus, diasumsikan sampel sekali setiap pembentukan korpus. Parameter α akan menentukan distribusi topik-topik yang ada dalam dokumen, sedangkan parameter β menentukan distribusi kata dalam tiap topiknya. variabel θ terletak pada tingkatan dokumen, memiliki sampel sekali per dokumennya. Variabel θ merupakan merupakan variable yang menggambarkan persebaran topik pada dokumen tertentu. Semakin tinggi nilai θ, maka semakin banyak topik yang ada di dalam dokumen, sedangkan semakin kecil nilai θ, maka dapat dikatakan dokumen tersebut semakin spesifik pada topik tertentu. Terakhir, variable z dan w berada pada tingkatan kata, dan dijadikan sampel sekali setiap kata dalam setiap dokumen. Variabel z menggambarkan topik yang terhubung dengan kata tertentu, sedangkan parameter w menggambarkan kata yang berhubungan dengan topik tertentu.

2.2.5. Author-Topic Models

Author-Topic Model (AT Model) adalah salah satu metode *topic modeling*, yang merupakan perpanjangan dari LDA dengan menggunakan author sebagai tambahan [9]. Dalam model ini, setiap kata dalam dokumen diasosiasikan dengan dua

variabel tersembunyi, yaitu *author* dan *topic*. Bedanya ATM dengan LDA adalah terletak pada variabel *author*-nya, sehingga prosesnya pun berubah, ATM akan memilih *author* dari daftar author, kemudian menentukan topik dari distribusi topik yang sudah terasosiasi dengan *author*. Dari topik yang sudah ada, maka dipilih kata – kata yang berhubungan dengan topik tersebut.



Gambar 2.4 Konsep Author-Topic Model [16]

Pada gambar 2.4 Merupakan konsep *Author-topic models* menurut Steyvers [16]. Kotak A merupakan kumpulan author yang di asosiasikan dengan distribusi multinomial melalui topiknya, ditunjukkan dengan simbol θ. Setiap topik diasosiasikan dengan distribusi multinomial melalui kata, ditunjukkan dengan simbol φ. Distribusi Multinomial θ dan φ memiliki parameter seperti halnya LDA, parameter yang digunakan adalah α dan β. Pada gambar φ berhubungan dengan variabel *w*, dimana w adalah kata pada dokumen. Pada setiap

kata dalam dokumen, dipilih author *x* dari a_d, lalu topik *z* ditentukan dari distribusi θ yang terhubung dengan author *x*. Kemudian kata w ditentukan dari ditribusi φ yang terhubung dengan topik *z*. Proses ini berulang selama *N* kali untuk membentuk sebuah dokumen *d*.

TOPIC 4	
WORD	PROB.
LIGHT	.0306
RESPONSE	.0282
INTENSITY	.0252
RETINA	.0241
OPTICAL	.0233
KOCH	.0190
BACKGROUND	.0162
CONTRAST	.0145
CENTER	.0124
FEEDBACK	.0118
AUTHOR	**PROB.**
Koch_C	.0903
Boahen_K	.0320
Skrzypek_J	.0283
Liu_S	.0250
Delbruck_T	.0232
Etienne-C._R	.0210
Bair_W	.0178
Bialek_W	.0133
Yasui_S	.0106
Hsu_K	.0103

TOPIC 13	
WORD	PROB.
RECOGNITION	.0500
CHARACTER	.0334
TANGENT	.0246
CHARACTERS	.0232
DISTANCE	.0197
HANDWRITTEN	.0166
DIGITS	.0154
SEGMENTATION	.0142
DIGIT	.0124
IMAGE	.0111
AUTHOR	**PROB.**
Simard_P	.0602
Martin_G	.0340
LeCun_Y	.0339
Henderson_D	.0289
Denker_J	.0245
Revow_M	.0206
Rashid_M	.0205
Rumelhart_D	.0185
Sackinger_E	.0181
Flann_N	.0142

TOPIC 28	
WORD	PROB.
KERNEL	.0547
VECTOR	.0293
SUPPORT	.0293
MARGIN	.0239
SVM	.0196
DATA	.0165
SPACE	.0161
KERNELS	.0160
SET	.0146
MACHINES	.0132
AUTHOR	**PROB.**
Scholkopf_B	.0774
Smola_A	.0685
Vapnik_V	.0487
Burges_C	.0411
Ratsch_G	.0296
Mason_L	.0232
Platt_J	.0225
Cristianini_N	.0179
Laskov_P	.0160
Chapelle_O	.0152

TOPIC 9	
WORD	PROB.
SOURCE	.0389
INDEPENDENT	.0376
SOURCES	.0344
SEPARATION	.0322
INFORMATION	.0319
ICA	.0276
BLIND	.0227
COMPONENT	.0226
SEJNOWSKI	.0224
NATURAL	.0183
AUTHOR	**PROB.**
Sejnowski_T	.0627
Bell_A	.0378
Yang_H	.0349
Lee_T	.0348
Attias_H	.0290
Parra_L	.0271
Cichocki_A	.0262
Hyvarinen_A	.0242
Amari_S	.0160
Oja_E	.0143

Gambar 2.5 Contoh Topik dari Model *Author-Topic* dengan *dataset Neural Information Processing Systems(NIPS) Conference* berdasarkan topik[11]

AUTHOR = Jordan_M		
PROB.	TOPIC	WORDS
.1389	37	MIXTURE, EM, LIKELIHOOD, EXPERTS, MIXTURES, EXPERT, GATING, PARAMETERS, LOG, JORDAN
.1221	60	BELIEF, FIELD, STATE, APPROXIMATION, MODELS, VARIABLES, FACTOR, JORDAN, NETWORKS, PARAMETERS
.0598	52	ALGORITHM, ALGORITHMS, PROBLEM, STEP, PROBLEMS, LINEAR, UPDATE, FIND, LINE, ITERATIONS
.0449	77	MOTOR, TRAJECTORY, ARM, INVERSE, HAND, CONTROL, MOVEMENT, JOINT, DYNAMICS, FORWARD

AUTHOR = Koch_C		
PROB.	TOPIC	WORDS
.2518	4	LIGHT, RESPONSE, INTENSITY, RETINA, OPTICAL, KOCH, BACKGROUND, CONTRAST, CENTER, FEEDBACK
.0992	45	VISUAL, STIMULUS, CORTEX, SPATIAL, ORIENTATION, RESPONSE, CORTICAL, RECEPTIVE, TUNING, STIMULI
.0882	84	SPIKE, FIRING, SYNAPTIC, SYNAPSES, MEMBRANE, POTENTIAL, CURRENT, SPIKES, RATE, SYNAPSE
.0504	64	CIRCUIT, CURRENT, VOLTAGE, ANALOG, CHIP, VLSI, CIRCUITS, SILICON, PULSE, MEAD

AUTHOR = LeCun_Y		
PROB.	TOPIC	WORDS
.2298	13	RECOGNITION, CHARACTER, TANGENT, CHARACTERS, DISTANCE, HANDWRITTEN, DIGITS, SEGMENTATION, DIGIT, IMAGE
.0930	53	GRADIENT, FUNCTION, DESCENT, ERROR, VECTOR, DERIVATIVE, DERIVATIVES, OPTIMIZATION, PARAMETERS, LOCAL
.0930	69	LAYER, WEIGHTS, PROPAGATION, BACK, OUTPUT, LAYERS, INPUT, NUMBER, WEIGHT, FORWARD
.0762	36	INPUT, OUTPUT, INPUTS, OUTPUTS, VALUES, ARCHITECTURE, SUM, ADAPTIVE, PREVIOUS, PROCESSING

Gambar 2.6 Contoh Topik dari Model Author-Topic dengan dataset Neural Information Processing Systems(NIPS) Conference berdasarkan author[11]

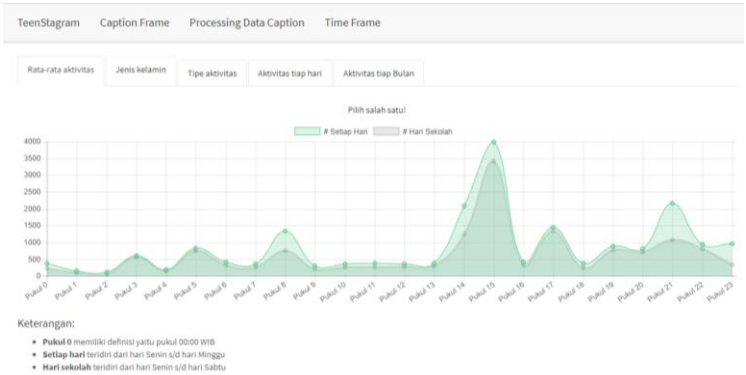
Pada gambar 2.5 dan 2.6 menunjukkan bahwa hasil penelitian dari *Rosen* dengan menggunakan *dataset* riset dari *Neural Information Processing System* (*NIPS*) *Conference* [11]. Disana dijelaskan bahwa setiap kata (*words*) dan *author* akan dihitung *probability* berdasarkan topik yang sesuai. Sehingga didapatkan bahwa topik terbentuk karena kumpulan kata dengan *probability* terbanyak dan *author* yang cenderung membahas tentang topik itu.

2.2.6. Validasi Model menggunakan Perplexity

Perplexity merupakan standar untuk melakukan penilaian terhadap performa dari suatu model probabilistik. Semakin rendah nilai *perplexity* suatu model, maka semakin baik pula performa dari model tersebut [9]. Rendahnya nilai *perplexity* pada topic modeling menunjukkan bahwa topik-topik yang ditemukan oleh model memiliki tingkat kesamaan yang rendah, sedangkan semakin tinggi nilai *perplexity*, topik yang ditemukan semakin mirip berdasarkan distribusi kata yang terdapat pada dokumen.

2.2.7. TeenStagram

TeenStagram adalah sebuah *platform* visualisasi *topic modeling* akun *Instagram* anak SMP di Surabaya [18]. Menggunakan LDA sebagai metode *Topic Modeling*. Gambar 2.7 merupakan *interface* dari *TeenStagram*. Aplikasi ini bertujuan untuk menganalisis bagaimana karakter seorang siswa SMP di Surabaya lewat *caption Instagram* yang sudah di unggah di media sosial *Instagram*.



Gambar 2.7 ***Interface Web TeenStagram[18]***

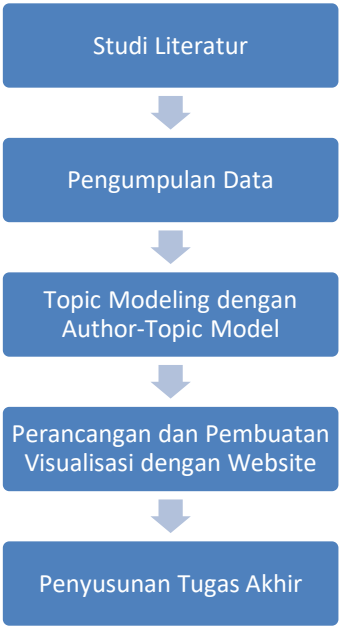
Fitur yang dihadirkan oleh aplikasi ini berupa pengelompokan topik yang biasa dibahas oleh siswa SMP, Analisa setiap region di Surabaya, Analisa berdasarkan waktu, dan aktivitas siswa SMP dalam menggunakan Instagram. Dibuat menggunakan *PHP* untuk keseluruhan *website* dan *Python* untuk pre-proses data sebelum di visualisasikan.

BAB III
METODOLOGI

Pada bab metode penelitian akan dijelaskan mengenai tahapan – tahapan apa saja yang dilakukan dalam pengerjaan tugas akhir ini beserta deskripsi dan penjelasan tiap tahapan tersebut. Lalu disertakan jadwal pengerjaan tiap tahapanan.

3.1. Tahapan Pelaksanaan Tugas Akhir

Pada sub bab ini akan menjelaskan mengenai metodologi dalam pelaksanaan tugas akhir. Metodologi ini dapat dilihat pada Gambar 3.1



Gambar 3.1 Metodologi Tugas Akhir

3.2. Uraian Metodologi

Pada bagian ini akan dijelaskan secara lebih rinci masing-masing tahapan yang dilakukan untuk penyelesaian tugas akhir ini.

3.2.1 Studi Literatur

Tahap studi literatur disini dilakukan dengan tujuan untuk dapat memahami konsep, metode, dan teknologi sesuai bahasan dan permasalahan sehingga dapat memberi solusi mengenai permasalahan yang akan digunakan dalam penyusunan tugas akhir. Adapun literatur yang digunakan dalam penelitian ini adalah terkait *Social Media Analysis* oleh Yuheng Hu, Lydia Manikonda, dan Subbarao Kambhampati [20] dan *Topic Modelling*, ATM mengacu pada penelitian Rosen dan Griffiths [11].

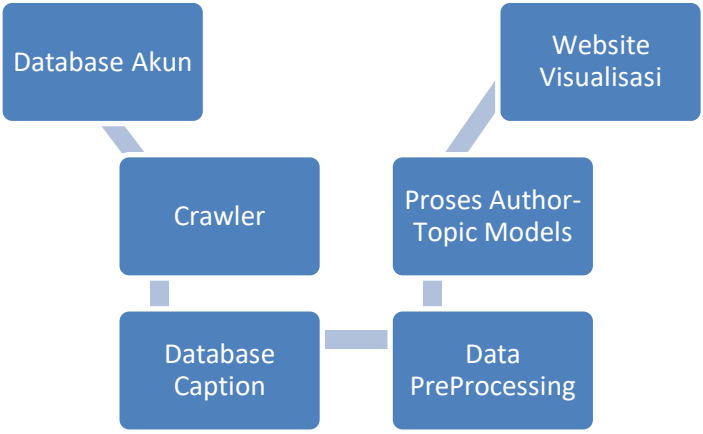
3.2.2 Pengumpulan Data

Tahap mempersiapkan data terdiri dari beberapa aktivitas, yakni, pengumpulan data akun Instagram, pemilihan akun *Instagram*, penambahan dan validasi akun Instagram, pengumpulan caption akun *Instagram* terpilih untuk selanjutnya dilakukan pemrosesan data *caption*, rangkaian tahapan ini dilakukan untuk mempersiapkan dokumen yang akan dianalisis menggunakan *topic modelling* ATM. Adapun pengambilan data dokumen yang akan dianalisis adalah data media sosial milik akun siswa dari 18 Sekolah Mengengah Atas di Surabaya dengan memanfaatkan *web crawler*.

1. Pengumpulan dan pemilihan akun Instagram sesuai dengan data berdasar hasil kuisioner mata kuliah Etika Profesi pada sekolah – sekolah menengah atas yang sudah ditentukan, kemudian dilakukan proses pemilahan untuk akun yang valid dan tidak valid, serta dilakukan pengelompokan sesuai dengan sekolah masing-masing. Periode dilakukannya pengumpulan data ini adalah dua tahun

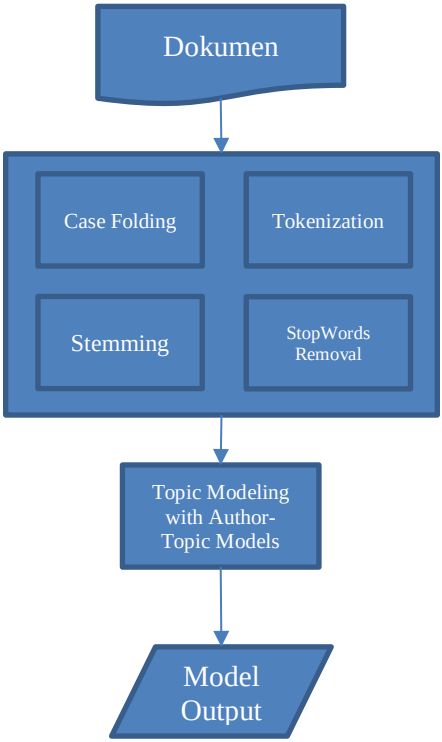
2. Proses validasi akun Instagram anak SMA dilakukan dengan cara mengambil username dari hasil kuisioner, kemudian dilakukan validasi apakah ada kesamaan *follow* antar akun, apabila ditemui kesamaan *follow* antar akun, kemudian dilakukan pengecekan jumlah *following*, jika *following* dalam batas wajar, dalam artian tidak melebihi 1000 maka akun tersebut bisa dikatakan valid akun siswa SMA, kemudian *username* akun tersebut *added* dalam *database*, sehingga caption dari akun tersebut dapat di-*crawling*.

3. Pengumpulan *caption* dilakukan dengan melakukan *crawling* terhadap akun Instagram berdasarkan username akun yang ada di dalam database kemudian data caption disimpan dalam database, secara skema seperti yang digambarkan pada gambar 3.2.



Gambar 3.2 Proses Pengumpulan data *caption* akun *Instagram*

3.2.3 Topic Modeling dengan Author-Topic Models



Gambar 3.3 Tahapan Topic Modeling dengan Author-Topic dan Pra-process terhadap corpus

3.2.3.1 Case Folding

Tahap pertama adalah *case folding*, dimana teks diubah menjadi bentuk *lowercase* atau *uppercase* dengan tujuan agar kata – kata yang diproses diseragamkan, sehingga tidak terjadi perbedaan penulisan huruf kapital maupun tidak. Contoh dari *case folding* dapat dilihat pada tabel 3.1, pada penelitian ini mengubah semua teks menjadi bentuk lowercase.

Tabel 3.1 Tabel sebelum dan sesudah *case folding*

Sebelum Case Folding	Setelah Case Folding
Hari ini hari dimana kita melihat masa lalu. Masa lalu yang merupakan sejarah. Sejarah bagaimana mereka para pejuang bangsa membangun dan mempertahankan Negeri ini. Sejarah yang akan kita ceritakan pada anak dan cucu kita kelak. Sejarah yang akan kita teruskan hingga bumi ini tak berbentuk. Sejarah yang akan mempertemukan aku dan dia.	hari ini hari dimana kita melihat masa lalu. masa lalu yang merupakan sejarah. sejarah bagaimana mereka para pejuang bangsa membangun dan mempertahankan negeri ini. sejarah yang akan kita ceritakan pada anak dan cucu kita kelak. sejarah yang akan kita teruskan hingga bumi ini tak berbentuk. sejarah yang akan mempertemukan aku dan dia.

3.2.3.2 Tokenization

Tahap selanjutnya adalah *tokenization*, dalam proses ini dilakukan pemisahan deretan kata dari kalimat atau paragraf menjadi kata tunggal, proses ini bertujuan untuk memperrsiapkan dokumen (*caption*) untuk proses berikutnya, yaitu stopwords removal dan stemming. Contoh *tokenization* dapat dilihat pada tabel 3.2.

Tabel 3.2 Tabel sebelum dan sesudah Tokenization

Sebelum Tokenization	Setelah Tokenization
hari ini hari dimana kita melihat masa lalu. masa lalu yang merupakan sejarah.	‘hari’ ‘ini’ ‘hari’ ‘dimana’ ‘kita’ ‘melihat’ ‘masa’ ‘lalu’ ‘masa’ ‘lalu’ ‘yang’

sejarah bagaimana mereka para pejuang bangsa membangun dan mempertahankan negeri ini. sejarah yang akan kita ceritakan pada anak dan cucu kita kelak. sejarah yang akan kita teruskan hingga bumi ini tak berbentuk. sejarah yang akan mempertemukan aku dan dia.	‘merupakan’ ‘sejarah’ ‘sejarah’ ‘bagaimana’ ‘mereka’ ‘para’ ‘pejuang’ ‘bangsa’ ‘membangun’ ‘dan’ ‘mempertahankan’ ‘negeri’ ‘ini’ ‘sejarah’ ‘yang’ ‘akan’ ‘kita’ ‘ceritakan’ ‘pada’ ‘anak’ ‘dan’ ‘cucu’ ‘kita’ ‘kelak’ ‘sejarah’ ‘yang’ ‘akan’ ‘kita’ ‘teruskan’ ‘hingga’ ‘bumi’ ‘ini’ ‘tak’ ‘berbentuk’ ‘sejarah’ ‘yang’ ‘akan’ ‘mempertemukan’ ‘aku’ ‘dan’ ‘dia’

3.2.3.3 Stopwords Removal

Tahap berikutnya adalah *stopwords removal*, *stopwords* atau kata umum (common words) adalah kata yang biasanya muncul dalam jumlah banyak dan dianggap tidak memiliki makna. Contoh kelas kata yang termasuk *stopwords* ditunjukkan oleh tabel 3.3. Adapun daftar *stopwords* yang digunakan adalah stopwords Bahasa Indonesia sesuai dengan penelitian yang disusun oleh Fadillah Z Tala [21]. Tahap menghilangkan stopwords merupakan tahapan yang penting, karena tingginya frekuensi kemunculan stopwords dalam setiap dokumen sehingga topik sulit diidentifikasi dan diinterpretasikan dengan baik. Contoh dari *stopwords removal* ditunjukkan pada tabel 3.3

Tabel 3.3 Tabel sebelum dan sesudah *Stopwords Removal*

Sebelum Stopwords Removal	Setelah Stopwords Removal
‘hari’ ‘ini’ ‘hari’ ‘dimana’ ‘kita’ ‘melihat’ ‘masa’ ‘lalu’ ‘masa’ ‘lalu’ ‘yang’ ‘merupakan’ ‘sejarah’	‘sejarah’ ‘pejuang’ ‘bangsa’ ‘membangun’ ‘mempertahankan’ ‘negeri’ ‘ceritakan’ ‘anak’ ‘cucu’

‘sejarah’ ‘bagaimana’ ‘mereka’ ‘para’ ‘pejuang’ ‘bangsa’ ‘membangun’ ‘dan’ ‘mempertahankan’ ‘negeri’ ‘ini’ ‘sejarah’ ‘yang’ ‘akan’ ‘kita’ ‘ceritakan’ ‘pada’ ‘anak’ ‘dan’ ‘cucu’ ‘kita’ ‘kelak’ ‘sejarah’ ‘yang’ ‘akan’ ‘kita’ ‘teruskan’ ‘hingga’ ‘bumi’ ‘ini’ ‘tak’ ‘berbentuk’ ‘sejarah’ ‘yang’ ‘akan’ ‘mempertemukan’ ‘aku’ ‘dan’ ‘dia’	‘kelak’ ‘teruskan’ ‘bumi’ ‘berbentuk’ ‘mempertemukan’

3.2.3.4 Stemming

Tahap selanjutnya adalah *stemming*, yaitu tahap untuk mengganti bentuk dari kata menjadi kata dasar, dalam arti lain menghilangkan imbuhan(*affix*) dari setiap kata tersebut, baik berupa awalan (*prefix*), sisipan (*infix*), akhiran (*suffix*) dan kombinasi awalan dan akhiran (*confix*). Pembentukan kata dasar bertujuan agar menyeragamkan kata-kata yang bermakna sama namun berbeda arti dikarenakan imbuhan, sehingga tidak salah interpretasi. Adapun proses *stemming* dilakukan dengan menggunakan *library* Sastrawi, yaitu *library stemmer* bahasa indonesia dengan lisensi MIT yang memanfaatkan kamus kata dasar dari Kateglo sebagai acuan. Contoh tahap *Stemming* ditampilkan pada tabel 3.4,

Tabel 3.4 Tabel sebelum dan sesudah Stemming

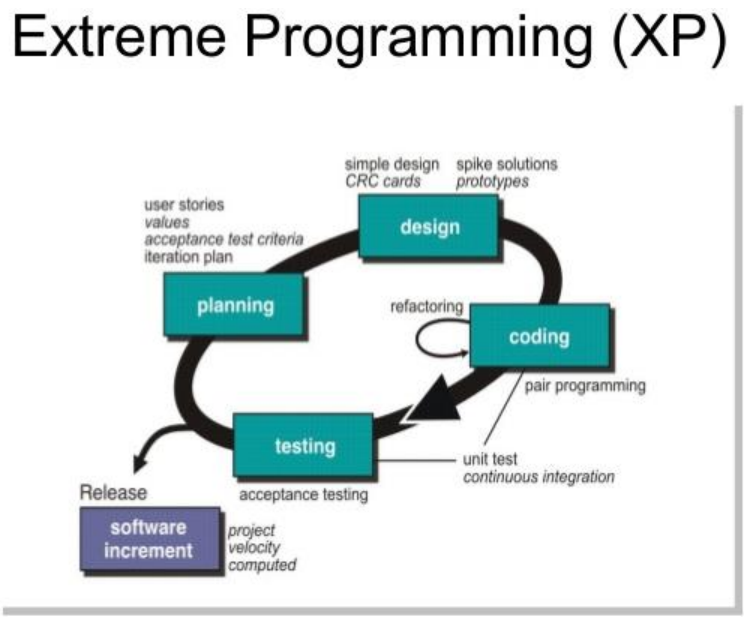
Sebelum Stemming	Setelah Stemming
‘sejarah’ ‘pejuang’ ‘bangsa’ ‘membangun’ ‘mempertahankan’ ‘negeri’ ‘ceritakan’ ‘anak’ ‘cucu’	‘sejarah’ ‘juang’ ‘bangsa’ ‘bangun’ ‘tahan’ ‘negeri’ ‘cerita’ ‘anak’ ‘cucu’ ‘kelak’ ‘bumi’ ‘bentuk’ ‘temu’

‘kelak’ ‘bumi’ ‘berbentuk’ ‘mempertemukan’	

Pembentukan model dilakukan dengan metode *Author-Topic Model*. Dimulai dari memetakan kata dari caption Instagram secara acak pada beberapa topik dan author dari kumpulan author dan topik yang didapat. Setelah itu dilakukan iterasi hingga jumlah tertentu sehingga mendapatkan beberapa hasil sampel pengelompokkan topik berdasarkan algoritma yang dilakukan. Hasil dari sampel tersebut diambil untuk membandingkan perbedaan setiap hasil, sehingga dapat ditemui pengelompokan topik.

3.2.4 Perancangan dan Pembuatan Visualisasi dengan Website

Pada tahap ini dilakukan pengembangan aplikasi yang merupakan hasil dari Analisa dan desain aplikasi. Aplikasi dirancang berdasarkan fungsionalitas yang telah di daftar dan artsitektur pada sistem yang sudah didesain. Pembuatan aplikasi menggunakan metode *Agile Software Development*, dengan model *Extreme Programming* (XP). XP ini merupakan model dengan berorientasi objek sebagai paradigm pengembangannya. XP memiliki 4 tahapan pengembangan sesuai dengan gambar 3.2, pertama yaitu perencanaan. Perencanaan aplikasi diharapkan memperhatikan kecepatan penyelesaiannya. Desain adalah tahapan kedua, desain semua arsitektur yang digunakan dalam mengembangkan aplikasi, dan diharapkan se-simpel mungkin. Rancang bangun atau kodifikasi adalah tahap pengkodean yang dilakukan setelah semua persiapan sudah dilakukan, dan yang terakhir adalah pengujian, apakah semua unit aplikasi diuji dan divalidasi apakah sudah berjalan seperti semestinya.



Gambar 3.4 Tahapan *Extreme Programming*

3.2.4.1 Perencanaan

Pada tahap ini, diawali dengan proses penggalian kebutuhan aplikasi, sehingga mempermudah mendefinisikan fungsionalitas dari aplikasi itu sendiri, mulai dari fitur yang akan diimplementasikan, timeline setiap pengerjaan dari awal hingga nantinya aplikasi akan diuji dan dipresentasikan.

3.2.4.2 Desain

Pada tahap desain, hasil perencanaan pada proses sebelumnya digunakan untuk membuat gambaran sesuai dengan aplikasi yang akan dibuat. Tahap desain merupakan rancangan awal pengembangan aplikasi sebelum melakukan pengkodean / coding. Desain yang akan dibuat meliputi:

1. Desain database
2. Desain *web-crawler*

3. Desain s*yste*m
4. Desain *User Inteface Web*
5. *Prototype* Aplikasi.

3.2.4.3 Rancang Bangun (*Coding*)

Pada tahap ini, dilakukan *coding* aplikasi secara keseluruhan berdasarkan desain yang sudah dibuat. Coding pada aplikasi meliputi:

1. Rancang bangun *Web Crawler*
   Web crawler mulai dikodifikasi sesuai fungsinya yaitu untuk mengakuisisi data yang digunakan untuk *Topic Modeling*, data yang diambil berupa *caption* dari foto yang telah diunggah kedalam Instagram siswa SMA, yang kemudian disimpan dalam database.
2. Rancang bangun *Topic Modeling*
   Kodifikasi pada *Topic Modeling* berupa preprocessing semua data caption yang sudah disimpan, kemudian dilakukan *Topic Modeling* menggunakan metode *Author-Topic* dengan Bahasa pemrograman *Python* dan *Framework Gensim*.
3. Rancang bangun Web Visualisasi
   Data yang sudah melewati proses topic modeling kemudian dibuat visualisasinya dalam bentuk *dashboard* pada *website*, menggunakan Bahasa PHP.

3.2.4.4 Pengujian

Pengujian aplikasi bertujuan untuk memeriksa apakah aplikasi berjalan sesuai dengan fungsionalitas yang sudah direncanakan. Pada tahap ini juga dilakukan pencatatan setiap *error* dan *bug* sehingga aplikasi dapat di track setiap kesalahannya. Pengujian pada tahap ini difokuskan pada pengujian ketepatan atau kesesuaian informasi dari dokumen dengan topik yang dihasilkan. Pada tahap ini dilakukan validasi terhadap hasil pengolahan data dengan ATM yang dilakukan untuk membuktikan bahwa distribusi topik yang dihasilkan memiliki

kesesuaian dengan dokumen caption instagram siswa SMA. Validasi dilakukan dengan menggunakan *perplexity.*

3.2.5 Penyusunan Laporan Tugas Akhir

Tahapan terakhir adalah penyusunan laporan tugas akhir sebagai bentuk dokumentasi atas terlaksananya tugas akhir ini. Laporan tugas akhir dibuat sesuai dengan format yang telah ditentukan. Tahapan penyusunan laporan tugas akhir dilakukan sejak awal hingga berakhirnya proses pengerjaan tugas akhir ini.

BAB IV
PERANCANGAN

Pada bab ini membahas terkait alur perancangan terkait beberapa hal yang diperlukan dalam proses pembuatan aplikasi sesuai dengan alur yang dijelaskan pada bab metodologi. Adapun perancangan ini diperlukan sebagai panduan dalam melakukan penelitian tugas akhir, yang dijelaskan sebagai berikut.

4.1 Pengambilan Data

Dalam pelaksanaan analisis dan visualisasi *topic modeling caption instagram* siswa SMA, data berupa *caption* merupakan objek utama yang analisis. Data yang dibutuhkan berupa data tipe teks hasil crawling media sosial instagram. Dalam melakukan pengumpulan data dibutuhkan *username* akun siswa SMA sebagai acuan lingkup data *caption* yang nantinya diambil. Adapun atribut pada tabel *username* ditunjukkan pada tabel 4.1 Proses penyimpanan data *caption* berdasarkan *username* akun, akan disimpan di *database caption* dimana attribut pada *database caption* dijelaskan melalui Tabel 4.2

Tabel 4.1 Keterangan atribut *Database Username*

Atribut	Tipe Data	Keterangan
username	Varchar (255)	Merupakan nama username dari setiap akun Instagram siswa SMA
link	Varchar (200)	Merupakan link menuju halaman profil akun Instagram siswa SMA
kode_sekolah	Varchar (30)	Merupakan kode dari sekolah yang digunakan untuk mendefinisikan Author.

gender	Varchar (5)	Merupakan jenis kelamin dari setiap siswa SMA
usia	Int (3)	Merupakan rentang usia dari siswa SMA

Tabel 4.2 Keterangan atribut *database Caption*

Atribut	Tipe Data	Keterangan
id	Varchar (255)	Merupakan id untuk setiap post yang diunggah oleh siswa SMA di Instagram
waktu	Int (11)	Merupakan waktu saat foto dan caption diunggah siswa SMA di Instagram
oleh	Varchar (255)	Merupakan nama username dari setiap akun Instagram siswa SMA
pesan	Text	Merupakan caption atau teks yang ada pada setiap foto yang diunggah di Instagram
foto	Int (3)	Merupakan *link* foto yang diunggah pada akun Instagram
timecrawl	timestamp	Merupakan waktu saat dilakukannya *crawling*
urutan	Int	Merupakan urutan unik setiap caption untuk memudahkan indeks dokumen.

4.2 Metodologi Implementasi

Metodologi implementasi dalam penelitian ini merupakan tahapan yang bertujuan untuk mencapai penelitian yang disesuaikan dengan komputasi secara otomatis. Komputasi yang dilakukan pada penelitian ini dilakukan dengan beberapa tahapan, yakni pemrosesan data, pencarian model hingga

analisis menggunakan bahasa python adapun proses visualisasi hasil pengolahan data akan dilakukan menggunakan PHP. Dalam penelitian ini ada 6 tahapan utama dalam tahap melakukan implementasi yaitu *crawling data, load data*, pra-proses data, pemrosesan data, analisa data, dan visualisasi data.

4.2.1 Crawling Data

Crawling data merupakan tahapan pengambilan data caption instagram yang kemudian akan digunakan dalam proses topic modeling. Tahap crawling data ini memanfaatkan nama akun setiap user yang sudah dicantumkan dalam database, tahap crawling ini menggunakan *PHP* dengan memanfaatkan *JSON* pada setiap halaman user instagram berbeda yang berisi informasi detail posting gambar untuk kemudian dimasukkan dalam bentuk array dan selanjutnya disimpan dalam database. Jumlah akun instagram yang digunakan adalah 236 akun aktif, dan jumlah caption yang berhasil didapatkan ada sebanyak 3272 data untuk periode Januari 2016 hingga Desember 2017. Gambaran proses *crawling* dijelaskan pada gambar 4.1:

Gambar 4.1 Alur proses *Crawling Data Caption*

Load data merupakan tahapan dimana data diproses untuk dapat dibaca ke dalam *tools* sebelum melakukan analisa pada penelitian. Data asli atau *raw data* yang berhasil diambil kemudian diexport ke dalam format .csv dari *database MySQL*. Setelah data siap dan terunduh, penggunaan python digunakan untuk *load data* dengan format csv dapat dilakukan dengan menggunakan modul csv. Adapun data yang dimuat adalah data caption instagram siswa SMA berupa kode sekolah sebagai author dan caption sebagai dokumen dari bulan Januari 2016

hingga bulan Desember 2017, untuk jelasnya digambarkan oleh gambar 4.2:

Gambar 4.2 Alur proses Crawling Data Caption

4.2.2 Pra-Proses Data

Tahap pra-proses data mencakup beberapa langkah utama pengerjaan yakni pembersihan data dari tag atau karakter-karakter tertentu, pengubahan data menjadi huruf kecil casefolding, stemming, stopword removal, serta tokenization. Untuk penjelasan secara lebih detail, dijelaskan pada Gambar 4.3:

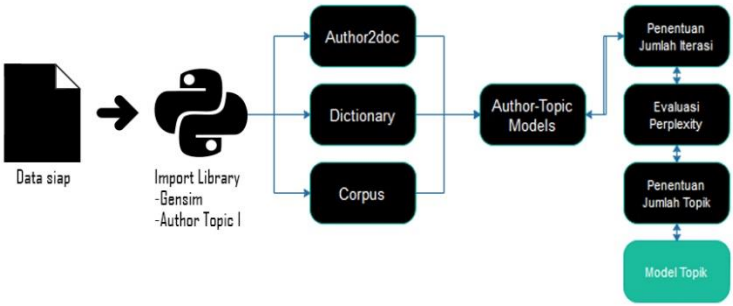
Gambar 4.3 Alur pra-proses data caption

1. *Cleansing* adalah tahap untuk menghilangkan karakter-karakter yang tidak terpakai pada data *caption*, seperti angka, simbol, dan menghapus beberapa *row* yang berisikan kosong karena emoji pada saat *crawling*.
2. *Casefolding* dilakukan untuk menyeragamkan struktur kata, pada penelitian ini menggunakan *lowercase* atau menjadi huruf kecil.

3. Stemming dilakukan untuk menghilangkan kata imbuhan serta mengubah kata-kata pada data menjadi kata dasar yang memanfaatkan modul Sastrawi.
4. *Tokenization* dilakukan untuk memecah atau memotong string pada kalimat menjadi tiap kata yang menyusunnya, dalam proses *tokenizing* ini digunakan modul *nltk.*
5. *Stopwords Removal* dilakukan dengan mendaftar kata-kata dalam bahasa indonesia, merujuk pada penelitian *Fadillah Z Tala* yaitu dengan memasukkan kata yang sering digunakan namun tidak memiliki nilai informasi, dan juga penggunaan stopword dari penelitian Tetha Valianta yang merujuk pada Bahasa anak SMP. Pada studi kasus penelitian ini dibuat stopword khusus untuk gaya bahasa siswa SMA yang mayoritas menggunakan bahasa kekinian yang berkembang di era-nya.

4.2.3 Proses Data

Pada tahapan proses data, langkah utama yang dilakukan adalah melakukan pencarian model dengan menggunakan modul gensim. Pada Gambar 4.4 menjelaskan tentang bagaimana alur pemrosesan data dijalankan.



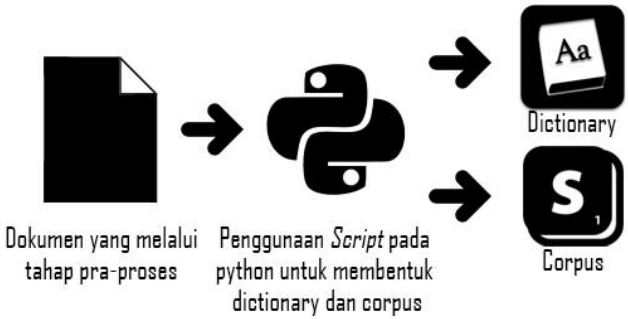
Gambar 4.4 Alur Proses Data

1. Pembentukan *author-to-doc*
   Dalam melakukan *Author-Topic Models*, data terlebih dahulu didefinisikan author setiap dokumennya. *Author-to-doc* adalah *list* dari dokumen yang akan dikumpulkan menjadi satu kesatuan sesuai dengan Author yang memiliki dokumen tersebut, lebih jelasnya seperti gambar 4.5. Data yang dipakai adalah author merupakan SMA di Surabaya dan documents adalah caption yang diambil hasil crawling.

Gambar 4.5 Pembentukan Author2doc List

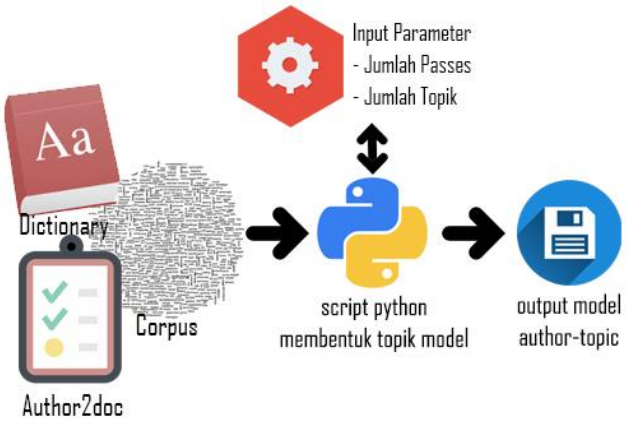
2. Pembentukan *Dictionary* dan *Corpus*
   Dalam melakukan topic modeling, data terlebih dahulu perlu dirubah ke dalam bentuk dictionary dan corpus. Pengertian *Dictionary* adalah format data yang mengandung himpunan kata unik yang memiliki indeks, sehingga dapat memudahkan dalam menampilkan kata yang termasuk dalam sebuah model, sedangkan *Corpus* merupakan format data yang

memiliki bentuk dokumen term-matrix, yang nantinya berguna untuk melakukan eksperimen pembentukan model. Adapun proses pembentukan dictionary dan corpus ditunjukkan dengan gambar 4.6



Gambar 4.6 Pembentukan Dictionary dan Corpus

3. Topic Modeling dengan *Author-Topic Model*
   Pada tahap topic modeling menggunakan ATM, langkah yang perlu dilakukan adalah membentuk model dengan menggunakan library *gensim*, kemudian model dievaluasi menggunakan *perplexity*. Dalam proses membentuk model percobaan input parameter sangat dibutuhkan. Hasil dari pencarian model akan digunakan untuk mendapatkan topik apa yang muncul dari analisis pada dokumen. Setelah model topik di dapatkan kemudian, dilakukan evaluasi *perplexity* menggunakan modul *logging*. Model dapat dikatakan terbaik apabila menunjukkan hasil *perplexity* yang lebih kecil dan stabil. Adapun skema proses *topic modeling* ditunjukkan dengan gambar 4.7



Gambar 4.7 Alur topic modelling dengan Author-Topic Models

4.2.4 Validasi Topik Model

Tahap validasi topik memiliki tujuan untuk memastikan model topik yang dihasilkan dari hasil topic modelling pada dokumen adalah sesuai, baik topik maupun kata-kata yang terkandung dalam topik tersebut. Adapun beberapa hal yang dianalisis dalam tahap validasi topik model adalah

1. Jumlah iterasi yang tepat untuk membentuk topik model.
2. Jumlah topik yang sesuai dengan rata – rata *perplexity* terkecil.

4.2.5 Analisa Topik

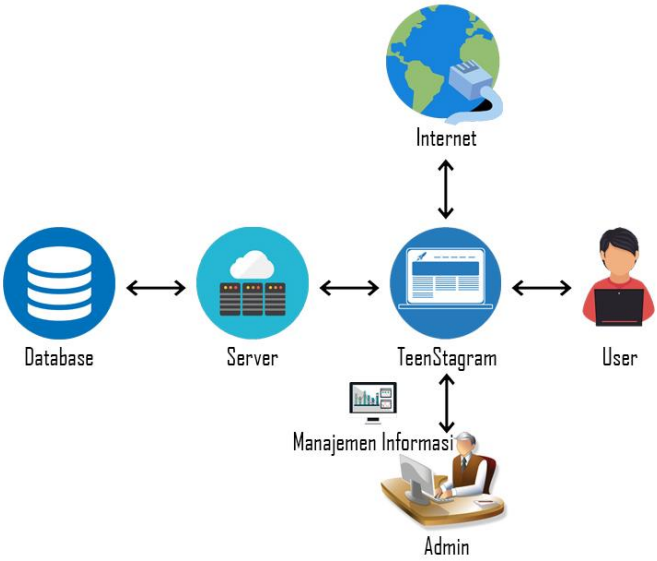
Analisis topik digunakan untuk mengidentifikasi hasil topik yang didapatkan dari proses permodelan *Author-Topic Models* terhadap jumlah topik yang dipilih. Caranya dengan cara menerjemahkan hasil kata-kata yang muncul dan membentuk sebuah topik, dari situ setiap kata-katanya disesuaikan dengan dokumen yang ada. Topik diterjemahkan kemudian dianalisis terhadap kesamaan satu topik dengan topik yang lainnya. Kemudian proses pendefinisian topik dilakukan dengan cara mengacu pada jurnal referensi topik *instagram* secara global,

selanjutnya akan terbentuk topik-topik yang dianggap sesuai dengan studi kasus dalam penelitian ini.

4.2.6 Konstruksi Perangkat Lunak

Berikut ini adalah perancangan perangkat lunak untuk visualisasi dashboard Aplikasi *TeenStagram*, meliputi arsitektur aplikasi, perancangan *database*, dan antarmuka aplikasi.

1. Arsitektur Aplikasi *TeenStagram* meliputi *database*, *server*, *web*, internet, admin, dan user yang dijelaskan pada gambar 4.8:

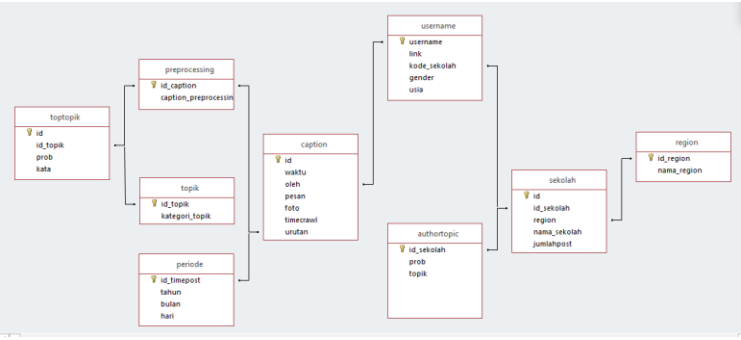


Gambar 4.8 Arsitektur Aplikasi *TeenStagram*

2. Perancangan database dalam penyusunan perangkat lunak TeenStagram menggunakan *database MySQL* sebagai database utama, dan menghasilkan database schema yang dijelaskan pada gambar 4.9
    - ***caption*** yang menampung informasi terkait caption siswa SMA, seperti caption, waktu

upload, link foto, nama pengguna, waktu crawling, dan urutan caption.

- ***username*** berisikan data dari masing – masing siswa SMA, seperti kode sekolah, usia, gender yang membantu proses pengambilan caption saat crawling.
- **sekolah** berisikan kode sekolah,
- **region** yang berisikan kode dan nama wilayah di Surabaya.
- **periode** berisikan tahun, bulan, dan hari yang dikonversi dari timepost pada tabel caption
- **topik** berisikan kategori topik setelah analisis model dan id topiknya.
- **preprocessing** berisikan hasil preprocessing data caption dan id setiap dokumennya.
- **toptopik** berisikan data topik setiap kata – katanya beserta probabilitas kata tersebut.
- **authortopic** berisikan id sekolah dan probabilitas topik terjadi pada sekolah tersebut.



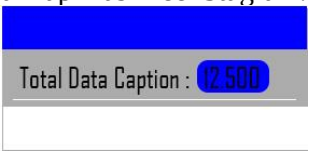
Gambar 4.9 Desain database TeenStagram

4.2.7 Integrasi PHP dan Python

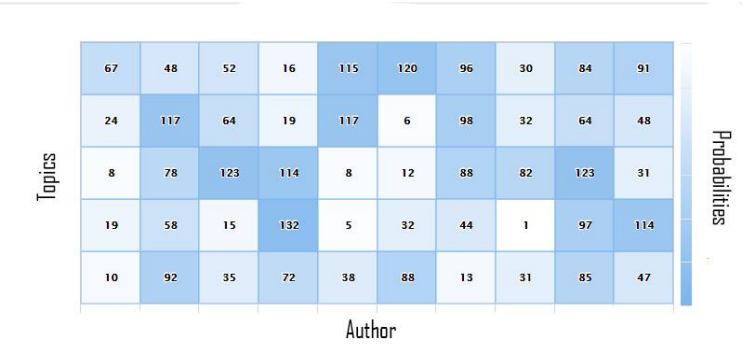
Pada tahapan ini dilakukan proses integrasi data melalui proses penambahan data akun dengan proses crawling, melakukan pra proses data dan proses data model menggunakan pengkodean *PHP* dengan menjalankan fungsi *Python* yang sebelumnya sudah dibuat, selanjutnya data akan disimpan ke dalam database, untuk kemudian divisualkan menggunakan PHP. User dapat melakukan tambah data melalui antarmuka aplikasi, kemudian ketika user menekan tombol "make model" merupakan *trigger* untuk menjalankan model *python* untuk topic modeling, hasil dari proses tersebut kemudian disimpan ke dalam database dan kemudian divisualisasikan menggunakan kode *PHP* dengan heatmap dan treemap.

4.2.8 Desain Antarmuka Aplikasi Visualisasi

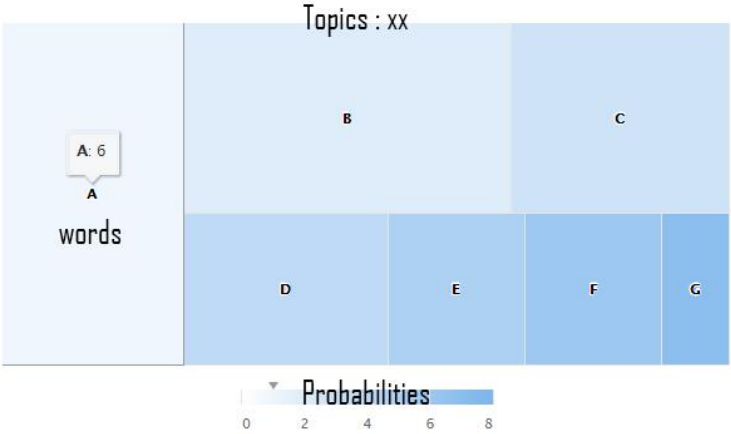
Melalui model yang sudah ada didapatkan hasil topic modeling yang siap divisualkan ke dalam beberapa kategori sesuai dengan topik yang dihasilkan. Visualisasi yang digunakan dalam penelitian ini adalah menggunakan bahasa pemrograman *PHP*, dimana user dapat melihatnya ke dalam sebuah dashboard. Rancangan dashboard yang akan ditampilkan dari analisa topic modeling, dibedakan menjadi beberapa konten, yakni, berdasarkan probabilitas author dengan topiknya, dan juga probabilitas setiap kata pada topik tersebut. Pada Gambar 4.10 menjelaskan tentang jumlah data caption yang berhasil terakuisisi, sedangkan Gambar 4.11 menjelaskan tentang diagram topik dan *author* beserta kemungkinannya yang terbentuk dari data caption, dan Gambar 4.12 menampilkan rancangan tampilan diagram topik beserta top 10 kemungkinan kata yang terkandung dalam topik tersebut, yang kemudian ditampilkan ke dalam aplikasi TeenStagram.



Gambar 4.10 Rancangan visualisasi jumlah data caption terambil.



Gambar 4.11 Rancangan visualisasi diagram author-topics



Gambar 4.12 Rancangan visualisasi top 10 kata per topik

BAB V
IMPLEMENTASI

Bab ini menjelaskan hasil dari implementasi perancangan studi kasus atau hasil dari proses pelaksanaan penelitian. Hasil yang akan dijabarkan adalah hasil eksperimen terhadap data yang digunakan sebagai acuan penelitian. Selain itu, akan dijelaskan juga mengenai tantangan dan kesulitfan dalam proses pelaksanaan penelitian.

5.1 Lingkungan Implementasi

Dalam pelaksanaan identifikasi dan visualisasi topik caption akun Instagram siswa SMA di Surabaya, dibutuhkan perangkat-perangkat untuk menunjang keberlangsungan penelitian. Adapun perangkat-perangkat yang dibutuhkan berupa perangkat keras dan spesifikasinya ditunjukkan dengan Tabel 5.1. Kemudian untuk perangkat lunak yang digunakan dalam implementasi model ditunjukkan dalam tabel 5.2. Selain menggunakan hardware dan software, dalam penelitian ini juga digunakan beberapa library untuk mendukung proses topic modeling menggunakan *python*, adapun beberapa library yang digunakan ditunjukkan dalam tabel 5.3.

Tabel 5.1 Spesifikasi perangkat keras

No	Hardware	Spesifikasi
1.	Jenis	Asus Notebook V550VX
2.	Processor	Intel Core I7-6700HQ
3.	RAM	DDR4 8 GB
4.	Hardisk	1TB HDD 5400/7200

Tabel 5.2 Perangkat lunak yang digunakan

No	Software	Penggunaan
1.	Windows10 Pro 64-bit	Sistem Operasi
2.	Xampp,3.2.2 & PHP7.1.7	Webserver
3.	Python 3.6 32-bit (IDE Spyder Anaconda 3.2.4)	Pemrosesan Data
4.	DBMS MySQL	Database Penyimpanan
5.	Ms. Excel 2016	Mengolah Angka

Tabel 5.3 Daftar library python yang digunakan

No	Library	Penggunaan
1.	nltk (3.2.4)	Preprocess Data
2.	sastrawi (1.0.1)	Stemming
3.	gensim (3.0.1)	Topic modeling (author-topic models)
4.	logging	Perplexity Evaluation
5.	pprint	print console
6.	collections	Menghitung kata (counter)

5.2 Pengambilan Data

Pengambilan data adalah tahap awal dimana data didapatkan dari hasil crawling halaman akun user Instagram yang namanya telah tercantum dalam database. Tahap ini dilakukan melalui pengumpulan data akun siswa terlebih dahulu.

5.2.1 Pengumpulan Data Akun

Proses pengumpulan data akun instagram didapatkan dari hasil rekap kuisioner sosialisasi mata kuliah etika profesi, kemudian dilakukan proses pemilihan untuk akun yang valid dan tidak

valid serta dilakukan pengelompokan sesuai dengan sekolah masing-masing.

- **Hasil pengumpulan data akun**
  Dalam rekap kuisioner sosialisasi mata kuliah etika profesi didapatkan 18 sekolah untuk nantinya dijadikan Author dalam permodelan. Sekolah tersebut antara lain dijelaskan pada Tabel 5.4

Tabel 5.4 Detail pengumpulan data

Kode Sekolah	Nama Sekolah
M01	MAN Surabaya
N02	SMA Negeri 2 Surabaya
N16	SMA Negeri 16 Surabaya
N20	SMA Negeri 20 Surabaya
N06	SMA Negeri 6 Surabaya
S01	SMA YAPITA
S02	SMA Muhammdiyah 2 Surabaya
N10	SMA Negeri 10 Surabaya
N13	SMA Negeri 13 Surabaya
K10	SMK Negeri 10 Surabaya
N17	SMA Negeri 17 Surabaya
N14	SMA Negeri 14 Surabaya
S03	SMA Luqman Hakim Surabaya
S04	SMA IPIEMS
N21	SMA Negeri 21 Surabaya
S05	SMA 10 Muhammadiyah Surabaya
N07	SMA Negeri 7 Surabaya
S06	SMA Ta'miriyah Surabaya

- **Validasi data akun**
  Validasi akun dilakukan dengan cara mengambil data username lalu dicek masing – masing halaman akun Instagram, kemudian dilihat apakah jumlah following lebih dari 1000 dan memiliki keterkaitan follow dengan akun lainnya, jika jumlah following kurang dari 1000 dan memiliki keterkaitan / saling me-follow satu sama lain, maka dapat dikatakan bahwa akun tersebut *valid*. Metode ini merupakan adopsi dari metode *tracking actor* [17] sehingga didapatkan data anak SMA sebanyak 464 akun, dengan rincian 235 akun publik (51%) dan 229 akun private (49%) setelah proses validasi, maka data akan disimpan dalam database untuk dilakukan crawling. Detail data ditunjukkan oleh tabel 5.5

Tabel 5.5 Detail pengumpulan data

Jumlah sekolah	18
Total Responden	619
Total Pengguna Aktif Instagram	464
Akun publik	235
Akun private	229
Akun publik dengan jenis kelamin laki - laki	102

Akun publik dengan jenis kelamin perempuan	133

5.2.2 Crawling Data

Proses *crawling* data caption dilakukan menggunakan *library CURL*, dan menggunakan *API* yang berguna untuk menguraikan *JSON* ke dalam *array*, adapun Kode 5.1 menunjukkan tentang pembuatan *API* yang digunakan untuk proses *crawling*

```
if ( $_csc->uri[2] == 'instasearch' || $_csc->uri[2] == 'instatag' ) {
        $c->cookiejar = 'views/faiznfi.txt';
        $_url = 'https://www.instagram.com';
if ( ! $_csc->uri[4] ) {
        $np = $_GET['np'] ?: null;
        $url = $_csc->uri[2] == 'instasearch'  ? '/':
                '/explore/tags/';
        $c->bc = $c->get($_url .$url.$_csc->uri[3].
                '/?max_id='.$np );
        $x = 1;
        $json = json_decode( $c->xp( '<script
                type="text/javascript">window._sharedData = ',
                ';</script>' ) );
        $tag = $_csc->uri[2] == 'instasearch'  ? $json-
                >entry_data->ProfilePage[0]->user->media-
                >nodes : $json->entry_data->TagPage[0]->tag-
                >media->nodes;
        $Arr['id'] = (int) str_replace( 'profilePage_', '', $json-
                >entry_data->ProfilePage[0]->logging_page_id
                );
        =foreach ( $tag as $r ) {
                  $y = $x-1;
                  $Arr['photo'][$y]['code'] = $r->code;
                  $Arr['photo'][$y]['waktu'] = $r->date;
                  $Arr['photo'][$y]['caption'] = $r->caption;
                  $Arr['photo'][$y]['display_src'] = $r->
                  display_src;
                  $x++;
                  }
```

Kode 5.1 Potongan *script* pembuatan *API* untuk proses *crawling data*

Tahap berikutnya setelah pembuatan *API*, dilakukan pembuatan *script* untuk melakukan *crawling* berdasarkan username yang telah tersimpan sebelumnya. Adapun Kode 5.2 menunjukkan fungsi untuk *crawling data caption*.

```
function masukkin ($i, $username) {
      global $db;
      $f = $db->f( 'caption_bot', 'id', 'WHERE id=?', $i->code );
      if ( ! $f && $i->caption)
            {
            $arr = array( $i->code, $i->waktu, $username,
            removeEmoji($i->caption), $i->display_src);
            var_dump($arr);
            $db->i( 'caption_bot', 'id, waktu, oleh, pesan, foto',
                  $arr );
            } else return true;
}
function ulangin( $i1, $username, $n=0 ) {
            $j = apiJson
            ("http://localhost/teenstagram/api/instasearch/$username
            /? np=$i1");
            if (is_array($j->photo)) foreach ( $j->photo as $t ) {
                  $habis = masukkin( $t, $username );
                  if ( $habis ) break;
            }
            if ( ! $habis && $j->next )
            return ulangin( $j->beda, $username, $n+1 );
            else return $n;}
$db->pfx="";
$f = $db->r( 'username', '*' );
foreach ( $f as $r ) {
            $username = $r['username'];
            $j = apiJson(
            'http://localhost/teenstagram/api/instasearch/' .
            $username );
            echo "data terambil". '<br>';
            if (is_array($j->photo))  foreach ( $j->photo as $t ) {
                  $habis = masukkin( $t, $username );
                  if ( $habis ) break;
            }
            if ( ! $habis ) $n = ulangin( $j->beda, $username );
            echo 'Done repeating ' . $n . ' times.';}
```

Kode 5.2 Potongan *script* fungsi *crawling* berdasarkan akun *username instagram*

5.3 Memuat Data

Data caption yang telah didapatkan dari hasil crawling dan tersimpan di dalam database kemudian akan dimuat ke dalam bentuk csv untuk dan diolah ke tahap selanjutnya. Langkah pengkodean yang dilakukan untuk memuat data adalah dengan menggunakan *library csv*, Kode 5.3 menunjukkan proses memuat data menggunakan *library* csv

```
import csv
# Get all text from each row
with open(dhv + 'dokumen - paling fix hehe.csv',
errors='ignore',encoding="utf8") as data:
    readCSV = csv.reader(data, delimiter=',')
    docs = []
    doc_ids = []
    for row in readCSV:
        doc_id = row[0]
        doc = row[0] + ' ' + row[1]
        doc = re.sub('\s', ' ', doc)
        doc_ids.append(doc_id)
        docs.append(doc)
# Get all author
author2doc = dict()
i = 0
with open(dhv + 'author (1).csv',
errors='ignore',encoding="utf8") as author:
            readCSV = csv.reader(author, delimiter=',')
            for contents in readCSV:
                        authorname = contents[0]
                        authorname = re.sub('\s', ' ', authorname)
                        ids = contents[1:]
                        if not author2doc.get(authorname):
                                    author2doc[authorname] = []
                        author2doc[authorname].extend(ids)
```

Kode 5.3 Potongan script untuk memuat data menggunakan library csv

5.4 Pra-Proses Data

Dalam melakukan analisa model topik caption instagram siswa SMA di Surabaya, pra-proses data merupakan salah satu tahapan yang penting dalam penelitian, agar data dapat diolah ke proses berikutnya. Pra-proses data meliputi beberapa tahapan, yaitu case folding, stemming, stopword removal, dan tokenization.

5.4.1 Case Folding

Tahapan case folding adalah proses dimana data diubah ke dalam bentuk huruf kecil, dengan tujuan untuk menyamaratakan format. Adapun pengkodean yang digunakan untuk melakukan case folding adalah dengan menggunakan Kode 5.4.

```
lowerdocs = str(i.lower())
```

Kode 5.4 *method* untuk *case folding* menjadi *lowercase*

5.4.2 Stemming

Stemming merupakan tahapan untuk mengubah sebuah kalimat menjadi kata dasar. Dalam penelitian ini, stemming dilakukan dengan menggunakan library Sastrawi dengan menggunakan fungsi stem. Adapun Untuk melakukan proses stemming data, dapat dilakukan dengan menggunakan Kode 5.5.

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
factory = StemmerFactory()
stemmer = factory.create_stemmer()
        stemmeddocs = stemmer.stem(lowerdocs)
```

Kode 5.5 Potongan script untuk melakukan stemming menggunakan library sastrawi

5.4.3 Pendefinisian Stopword

Menurut Tim Depdikbud RI yang dipelopori oleh Hasan Alwi dkk. [21] Stopword didefinisikan ke dalam beberapa kategori kelas kata. Kategori kelas kata yang tergolong ke dalam stopword adalah kata ganti, kata keterangan, kata bilangan, kata depan, kata sambung, makna kata sambung, kata seru, kata sandang dan partikel penegas. Proses pembuatan stopword untuk penelitian ini sedikit unik karena studi kasus penelitian menyangkut caption instagram remaja yang memang memiliki kosakata baru dan luas. Oleh karena itu proses pembuatan stopword dibuat melalui proses penghitungan kata yang sering muncul. Kemudian dari kata yang sering muncul tersebut, akan didefinisikan apakah kata tersebut memiliki makna atau tidak, sesuai dengan justifikasi yang mengacu pada konteks kelas kata [21]. Apabila suatu kata dinilai tidak memiliki makna, maka kata tersebut akan masuk ke dalam daftar kata yang dibuang stopword. Adapun pengkodean yang digunakan untuk mengetahui kata yang sering muncul adalah menggunakan library *collections*, ditunjukkan oleh Kode 5.6

```
import collections
with open(dhv+'jumlahwords.csv','w') as tulisFile:
   tulisFileWriter =
   csv.writer(tulisFile,lineterminator='\n')
   hasil=[]
   i=0
   for items in docs:
      b=0
      for items in docs[i]:
         tulisFileWriter.writerow(docs[i][b])
         b=b+1
      i=i+1
      i=0
   tulisFile.close()
```

Kode 5.6 script untuk menghitung kata

Semua daftar kata stopword kemudian disimpan ke dalam sebuah file.txt, dan selanjutnya digunakan untuk proses stopword removal. Kode 5.7 menunjukkan pengkodean untuk melakukan proses stopword removal

```
# Preprocess
list_stopword = []
with open(dhv + 'stopwords.txt') as stopword:
 for line in stopword:
  list_stopword.append(line.strip())
stopped_tokens = [i for i in tokenWithoutInt if not i in
list_stopword]
```

Kode 5.7 Script untuk stopword removal

5.4.4 Tokenization

Tokenization merupakan tahap memecah teks yang dapat berupa kalimat, paragraf atau dokumen, menjadi token-token atau sekumpulan kata. *Tokenization* memisahkan teks berdasarkan spasi dan dijadikan kata dalam korpus. Untuk melakukan proses *tokenization*, dalam penelitian ini menggunakan library *NLTK* (*Natural Language Toolkit*) meliputi fungsi *tokenization* dan *regular expresion*. Berikut adalah pengkodean untuk melakukan *tokenization* menggunakan *NLTK* ditunjukkan oleh Kode 5.8:

```
from nltk.tokenize import RegexpTokenizer
tokenlist_data = []
tokenizer = RegexpTokenizer(r'\w+')
tokens = tokenizer.tokenize(stemmeddocs)
 tokenWithoutInt = [item for item in tokens if not
item.isdigit()]
```

Kode 5.8 *Script* untuk *Tokenization*

Setelah melalui pra-proses data, hasil proses tersebut akan disimpan menjadi file csv untuk kemudian melalui tahap proses berikutnya, secara keseluruhan pengkodean Pra-proses data dapat dilihat pada Kode 5.9:

```
from nltk.tokenize import RegexpTokenizer
from Sastrawi.Stemmer.StemmerFactory import
StemmerFactory
# Preprocess
list_stopword = []
with open(dhv + 'stopwords.txt') as stopword:
 for line in stopword:
  list_stopword.append(line.strip())
# create stemmer
factory = StemmerFactory()
stemmer = factory.create_stemmer()
tokenlist_data = []
tokenizer = RegexpTokenizer(r'\w+')
#mulaipreproses
for i in docs:
 lowerdocs = str(i.lower())
 stemmeddocs = stemmer.stem(lowerdocs) # stemming
process
 tokens = tokenizer.tokenize(stemmeddocs)
 tokenWithoutInt = [item for item in tokens if not
item.isdigit()]
 stopped_tokens = [i for i in tokenWithoutInt if not i in
list_stopword]
 tokenlist_data.append(stopped_tokens)
with open(dhv+'datastemfixbanget.csv','w') as tulisFile:
    tulisFileWriter =
csv.writer(tulisFile,lineterminator='\n')
    for row in tokenlist_data:
        tulisFileWriter.writerow([row])
    tulisFile.close()
```

Kode 5.9 Script keseluruhan pra-proses dokumen

5.5 Proses Data

Tahapan dalam pemrosesan data diawali dengan melakukan penyimpanan dictionary yang didapatkan dari tahap pra-proses data sebelumnya. *Dictionary* bisa diatur berdasarkan minimal kata yang muncul dan maksimum frekuensi kata tersebut muncul, pembuatan dictionary dalam penelitian ini menggunakan minimal satu kata yang muncul, dengan maksimum frekuensi 0.9 dikarenakan caption dalam Instagram cenderung sedikit dan hanya menggunakan satu kata saja. Fungsi *dictionary*() sendiri adalah untuk memberikan nilai unik atau sebuah index berupa integer pada setiap kata, yang bertujuan untuk mempermudah proses pada tahapan selanjutnya. Hasil dari dictionary yang telah diproses kemudian disimpan ke dalam file bernama dictionary.dict. Kode untuk melakukan pembuatan dictionary ditunjukkan pada Kode 5.10

```
dictionary = Dictionary(docs)
max_freq = 0.9
min_wordcount = 1
dictionary.filter_extremes(no_below=min_wordcount,
no_above=max_freq)
_ = dictionary[0]
```

Kode 5.10 *Script* pembuatan *dictionary*

Melalui tahap pembuatan *dictionary*, langkah yang harus dilakukan selanjutnya adalah membuat korpus atau kumpulan vektor kata dari dokumen. Proses membuat kata menjadi korpus dapat dilakukan dengan menggunakan fungsi *doc2bow*. Kode 5.11 menjelaskan cara pembuatan korpus

```
corpus = [dictionary.doc2bow(doc) for doc in docs]
```

Kode 5.11 *Script* pembuatan *corpus*

Hasil pembuatan korpus perlu disimpan untuk memudahkan proses pembuatan dan penggunaan model lebih lanjut. Korpus kemudian di proses untuk digunakan membuat model ATM

yang paling sesuai. Model dianggap memiliki kesesuaian yang besar atau baik, dilihat dari nilai *perplexity* yang dimilikinya. Semakin kecil semakin baik model tersebut.

5.6 Permodelan Topik dengan Author-Topic Models

5.6.1 Alur Permodelan Topik dengan Author-Topic Models

1. Memuat *Dictionary* dan *Corpus*
   Merupakan tahapan memuat data dictionary dan corpus yang sebelumnya telah disimpan sebagai file. Untuk memudahkan proses memuat data, dictionary dan corpus disimpan menggunakan variabel penamaan tertentu. Pengkodean yang digunakan untuk melakukan proses memuat dictionary dan Corpus ditunjukkan pada Kode 5.12

```
dictionary = Dictionary(docs)
rarely.
max_freq = 0.9
min_wordcount = 1
dictionary.filter_extremes(no_below=min_wordcount,
no_above=max_freq)
_ = dictionary[0]
corpus = [dictionary.doc2bow(doc) for doc in docs]
```

Kode 5.12 *Script load dictionary* dan *corpus*

2. Pembentukan Model Topik
   Pada tahapan pembentukan model topik, digunakan library gensim dengan modul AuthorTopicModel dan Dictionary. Dalam pembentukan model topik, diperlukan sebuah parameter, yaitu berupa jumlah topik, dan passes. Passes merupakan jumlah iterasi yang dilakukan pada proses pembentukan model topik. Penentuan parameter disini dimaksudkan untuk mencari nilai perplexity yang optimal. Semakin kecil nilai perplexity menunjukkan bahwa model yang

dibentuk semakin baik. Adapun kode untuk melakukan uji coba input parameter ATM ditunjukkan pada Kode 5.13

```
from gensim.models import Phrases
from gensim.corpora import Dictionary
from gensim.models import AuthorTopicModel
model = AuthorTopicModel(corpus=corpus,
                         num_topics=2,
                         id2word=dictionary.id2token,
                         author2doc=author2doc,
                         alpha='auto',
                         passes=6,
                         eval_every=5,
                         iterations=100)
```

Kode 5.13 *Script* untuk mengatur parameter model

3. Dokumentasi menggunakan Logging
   Dalam melakukan uji coba, diperlukan proses pencatatan atau logging. Proses pencatatan dianggap penting karena berguna untuk mengetahui rekaman catatan terkait proses yang terjadi dalam proses pembentukan model topik. Informasi penting yang perlu diperhatikan adalah nilai perplexity. Nilai perplexity yang muncul merupakan hasil akumulasi secara otomatis oleh fitur modul genism. Dalam melakukan pencatatan hasil uji coba dibutuhkan sebuah library yang bernama logging, kemudian file logging disimpan dengan format .csv untuk dianalisa. Pengkodean untuk melakukan logging ditunjukkan dengan Kode 5.14

```
import logging
logger = logging.getLogger()
fhandler =
logging.FileHandler(filename=dhv+'rekapmodel.csv',
mode='a')
formatter = logging.Formatter('%(asctime)s - %(name)s
- %(levelname)s - %(message)s')
fhandler.setFormatter(formatter)
logger.addHandler(fhandler)
logger.setLevel(logging.DEBUG)
```

Kode 5.14 *Script* untuk melakukan dokumentasi *logging*

5.6.2 Uji Coba Pemodelan Topik menggunakan ATM

Tahap uji coba pemodelan topik menggunakan Author-Topic Models merupakan tahapan yang dilakukan untuk membentuk model topik terbaik dengan melakukan uji coba sesuai parameter, yakni passes atau jumlah iterasi dan jumlah topik.

1. Penentuan Jumlah Iterasi

Pada metode Author-Topic Models atau ATM, passes merupakan iterasi dilakukannya model. Penentuan iterasi merupakan tahap yang penting dalam menentukan model, guna menghasilkan model yang terbaik. Apabila jumlah iterasi terlalu sedikit, akan menghasilkan model yang belum stabil atau dapat dikatakan *underfitting*, sementara iterasi yang terlalu banyak akan menghasilkan model yang *overfitting*. Penentuan jumlah passes diawali dengan memberikan nilai sebesar 100, kemudian jumlah topik ditentukan pada rentang 1 hingga 12 topik. Berdasarkan hasil uji coba jumlah iterasi, nilai perplexity yang muncul akan dicatat untuk dianalisis grafiknya secara visual, dan dilakukan penghitungan standar deviasinya. Nilai passes yang akan digunakan adalah nilai passes paling awal setelah menunjukkan tren yang stabil. Adapun pengkodean yang digunakan untuk penentuan jumlah iterasi ditunjukkan dengan Kode 5.15

```
model = AuthorTopicModel(corpus=corpus,
                         num_topics=7  ,
                         id2word=dictionary.id2token,
                         author2doc=author2doc,
                         alpha = 'auto',
                         passes=100,
                         eval_every=10,
                         iterations=100)
```

Kode 5.15 *Script* untuk mengatur jumlah iterasi

2. Penentuan Jumlah Topik

Setelah menentukan jumlah iterasi, uji coba dilakukan dengan penentuan jumlah topik. Uji coba jumlah topik merupakan tahap yang penting dalam menentukan model, hal ini untuk menghasilkan model yang terbaik, model dapat dikatakan terbaik apabila model memiliki nilai perplexity yang rendah, semakin rendah nilai perplexity, menunjukkan akurasi model yang semakin baik. Penentuan jumlah topik dilakukan dengan melakukan uji coba terhadap jumlah topik, penentuan jumlah topik ditentukan pada rentang 1 hingga 12 topik. Berdasarkan eksperimen jumlah topik, nilai perplexity yang muncul akan dicatat untuk dianalisis tren nilainya secara visual dan dilakukan penghitungan standar deviasinya. Sehingga pada akhirnya jumlah topik yang dipilih adalah jumlah topik yang memiliki nilai rata-rata perplexity paling rendah dengan standar deviasi minimum. Adapun pengkodean yang digunakan untuk penentuan jumlah topik ditunjukkan dengan Kode 5.16

```
model = AuthorTopicModel(corpus=corpus,
                         num_topics=7  ,
                         id2word=dictionary.id2token,
                         author2doc=author2doc,
                         alpha = 'auto',
                         passes=100,
                         eval_every=10,
                         iterations=100)
```

Kode 5.16 *Script* untuk mengatur jumlah topik

5.6.3 Menyimpan Model

Model dengan jumlah iterasi dan jumlah topik yang dianggap terbaik perlu disimpan agar dapat digunakan kembali dengan cepat. Model disimpan dalam format .model. Adapun cara menyimpan model ditunjukkan dengan Kode 5.17

```
model.save(dhv+'model.atmodel')
```

Kode 5.17 Script untuk menyimpan model

5.6.3 Validasi Topik

Validasi model topik disini mengacu pada nilai perplexity yang didapatkan dari hasil uji coba jumlah iterasi atau passes dan jumlah topik. Model dapat dikatakan terbaik jika memiliki nilai perplexity yang rendah. Nilai perplexity akan dicatat untuk dianalisis tren nilainya secara visual dan dilakukan penghitungan standar deviasinya. Sehingga pada akhirnya jumlah iterasi dan jumlah topik yang dipilih adalah yang memiliki nilai rata-rata paling rendah dengan standar deviasi minimum[11].

5.7 Analisa Topik

Analisis topik merupakan tahapan yang dilakukan berdasarkan luaran dari LDA Model yang telah dipilih. Analisis topik dilakukan dengan mengeluarkan semua kemungkinan topik serta distribusi kata-kata dalam topik tersebut. Untuk

menunjukkan hasil topik teratas hingga akhir, dapat menggunakan Kode 5.18

```
from pprint import pprint
top_topics = model.top_topics(model.corpus)
pprint(top_topics)
```

Kode 5.18 *Script* untuk melihat hasil topik teratas

Untuk melihat topik secara spesifik, kita hanya perlu menggani nomor topik yang akan kita lihat dengan menggunakan Kode 5.19

```
model.show_topic(0)
```

Kode 5.19 *Script* untuk melihat hasil topik spesifik

Analisa topik ini ditunjukkan dengan tabel berupa daftar kata untuk masing-masing topik, dalam menentukan kategori topik yang digunakan acuan studi literatur [20]. Setelah menganalisa masing – masing kata, kita bisa melakukan pelabelan pada topik dengan menggunakan Kode 5.20

```
topic_labels = ['1','2','3','4', '5']
for topic in model.show_topics(num_topics=6):
    print('Label: ' + topic_labels[topic[0]])
    words = ''
    for word, prob in model.show_topic(topic[0]):
        words += word + ' '
    print('Words: ' + words)
    print()
```

Kode 5.20 *Script* untuk memberi label sebuah topik

Kita juga bisa melihat daftar author beserta probabilitas topik yang sedang dibahas oleh author tersebut, menggunakan show_author, dengan Kode 5.21

```
        def show_author(name):
            print('\n%s' % name)
            #print('Docs:', model.author2doc[name])
            print('Topics:')
        pprint([(topic_labels[topic[0]], topic[1]) for topic in
model[name]])
        for row in author2doc:
            namaauthor = row[0]+row[1]+row[2]
            show_author(namaauthor)
```

Kode 5.21 *Script* untuk mencetak author beserta topiknya

5.8 Integrasi PHP dan Python

Sebelum melakukan visualisasi data, tahapan yang perlu dilakukan adalah menghubungkan PHP dengan model python yang telah dibuat. Hal ini bertujuan untuk proses kelanjutan program berikutnya, jika terdapat masukan berupa tambah caption program dapat secara otomatis melakukan pemodelan dan mengklasifikasi data. Kode PHP dibuat untuk menjalankan script Python untuk melakukan pemodelan topik dan melakukan klasifikasi topik yang akan kembali disimpan ke dalam database. Untuk mengintegrasikan PHP dengan Python dapat menggunakan Kode 5.22

```
<?php
        include( 'views/header2.php' );
        $preprocessing = shell_exec('python
        C:/xampp/htdocs/teenstagram/preprocessing.py');
        echo $preprocessing;
?>
```

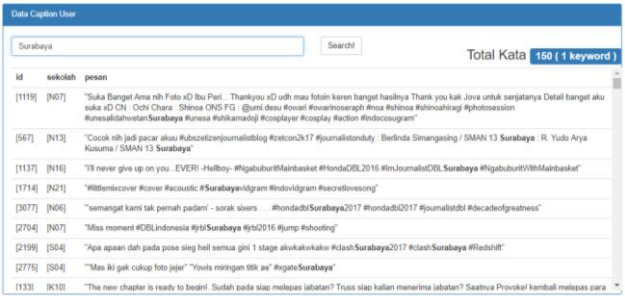
Kode 5.22 Menggunakan PHP untuk memanggil Python

5.9 Visualisasi Data

Visualisasi data dalam penelitian ini merupakan pembuatan dashboard yang dihasilkan dari data caption user siswa SMA di Surabaya yang masuk melalui aplikasi menggunakan proses crawling. Dashboard dibuat untuk menunjukkan trend topic

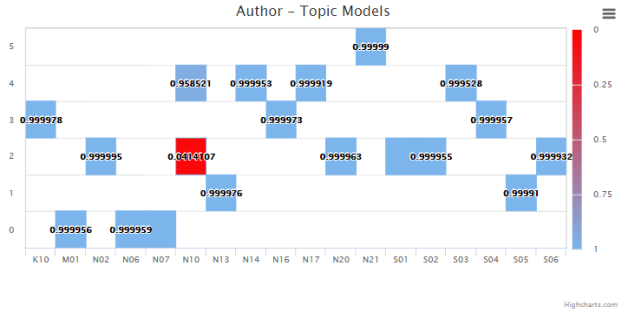
yang sering dibicarakan oleh siswa SMA di Surabaya berdasarkan sekolahnya. Visualisasi data ke dalam dashboard dibuat dengan menggunakan bahasa pemrograman PHP. Adapun bentuk visualisasi data menggunakan grafik dan tabel. Grafik ditampilkan dengan menggunakan heatmap dan treemap dengan menggunakan library highchart dan visualisasi berupa tabel menggunakan template bootstrap. Berikut ini adalah beberapa tampilan antarmuka apliasi dan visualisasi data:

1. *Interface* pencarian *caption* berdasarkan *keyword* tertentu. Fitur pencarian data ini berfungsi untuk mencari sesuai dengan *keyword* yang kita inputkan. Interface *pencarian* caption dapat dilihat pada gambar 5.1



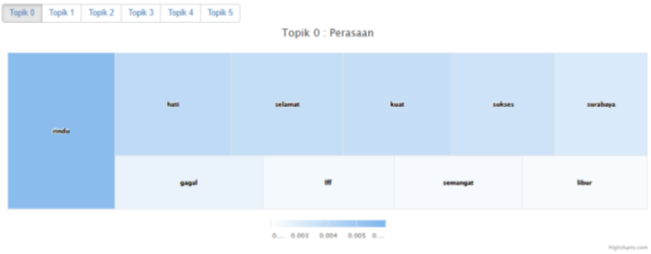
Gambar 5.1 *Interface* pencarian *caption*

2. Interface hasil permodelan author-topic untuk melihat bagaimana sekolah sedang membahas topik tertentu berdasarkan probabilitasnya Tampilan author-topic Models sesuai probability dapat dilihat pada gambar 5.2



Gambar 5.2 Interface author-topic models

3. Interface hasil permodelan author-topic untuk melihat top 10 kata yang terkandung topik tersebut, sehingga kita tahu bagaimana topik dan kata yang paling dominan dibahas oleh siswa SMA berdasarkan sekolah di Surabaya. Tampilan words per topic dapat dilihat pada gambar 5.3



Gambar 5.3 Interface words-per-topic

4. Interface Word Cloud dari jumlah kata yang sering muncul atau dipakai oleh siswa SMA di Surabaya setelah melalui tahap pra-proses data. Tampilan dapat dilihat pada gambar 5.4

Gambar 5.4 Interface word cloud

BAB VI
HASIL DAN PEMBAHASAN

Pada bab ini akan dijelaskan terkait analisa dan pengujian yang meliputi tiga hal, yaitu analisa hasil pemodelan, pengujian fungsional dan non fungsional

6.1 Analisa Hasil Permodelan

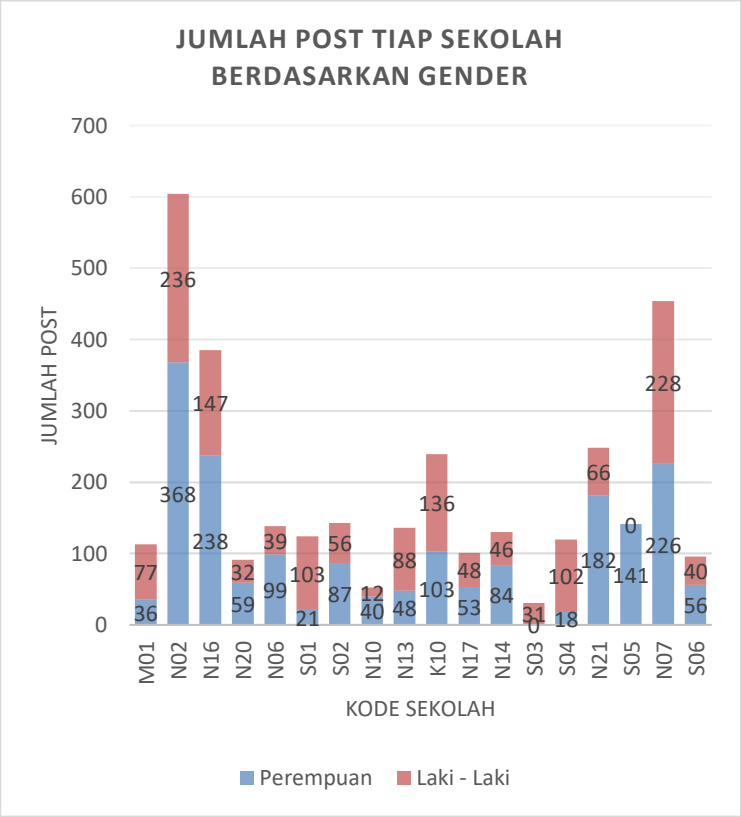
Pada Analisa hasil permodelan akan membahas mengenai hasil pengujian model menggunakan perhitungan *perplexity*. Adapun langkah – langkah yang dilakukan dijelaskan pada hal berikut.

6.1.1 Memuat Data

Total data yang berhasil didapatkan dari hasil crawling berjumlah 3346 data dari rentang januari 2016 hingga desember 2017. Adapun detail data yang berhasil terakuisisi ditunjukkan dengan Tabel 6.1 berdasarkan rentang waktu, hal tersebut dapat dilihat bahwa setiap bulan maka semakin meningkat jumlah intensitas upload pada media sosial Instagram.

Tabel 6.1 Detail data yang berhasil terakuisisi

Bulan / Tahun	2016	2017
Januari	19	115
Februari	18	129
Maret	36	102
April	67	135
Mei	51	166
Juni	81	166
Juli	88	201
Agustus	98	191
September	74	212
Oktober	98	275
November	114	385
Desember	132	229
Total Jumlah Caption		3346



Gambar 6.1 Jumlah Post tiap sekolah berdasarkan gender

6.1.2 Pra-Proses Data

Pendefinisian Data Training

Dalam penelitian ini untuk membuat model berdasarkan data caption instagram siswa SMA di Surabaya, digunakan data sebanyak (88 persen) dari total data 3346, yakni 2951 data.

Pendefinisian Stopword

Stopwords merupakan kata umum yang biasa digunakan dan tidak memiliki makna, pada penelitian ini didefinisikan berdasarkan dua landasan asumsi, yang pertama mengacu pada

pemaknaan kata sesuai dengan sistem tata Bahasa Indonesia baku[21] dan berdasarkan Analisa frekuensi munculnya kata dalam caption yang paling tinggi.

1. Berdasarkan tata Bahasa Indonesia yang baku [1], kelas kata merupakan kelas atau golongan (kategori) kata berdasarkan bentuk, fungsi, atau maknanya (KBBI). Berikut adalah kelas kata yang yang dianggap tidak memiliki makna dan digunakan dalam pembentukan stopwords dilihat pada tabel berikut:

Tabel 6.2 Kata Sandang

Kata Sandang			
Jumlah Tunggal	Jamak / Kelompok	Kata ganti orang	Berimbuhan
Sang	Jangan - jangan	Dia	Sebelum
Dan	Para	si	selama

Tabel 6.3 Kata Depan

Kata Depan			
Tempat	Maksud	Sebab	Waktu
ke	untuk	demi	hingga
dari	guna	karena	nyaris

Tabel 6.4 Kata Sambung

Kata Sambung			
Asal	Majemuk	Berimbuhan	Bentukan
sedang	lagipula	selama	Kalau-kalau
maka	karena itu	sehingga	seakan-akan

Tabel 6.5 Kata Seru

Kata Seru (interjeksi)		
Singkat	Wah	cih
Biasa	Aduh	celaka

Tabel 6.6 Partikel Penegas

Partikel Penegas		
Jumlah tunggal	kah	lah

2. Berdasarkan Analisa frekuensi munculnya kata dalam caption yang paling tinggi. Dengan menggunakan library collections pada python, maka terbentuklah sebuah kata dengan frekuensi kemunculannya, dan diambil semua kata diatas satu. Hasil dapat dilihat pada tabel 6.7, dari hasil tersebut, jika ada kata yang terindikasi tidak memiliki makna, maka kita tambahkan kedalam list stopwords.

Tabel 6.7 Hasil perhitungan frekuensi kata

Kata	Frekuensi
Surabaya	112
Allah	81
Global	81
Thecreatorclass	79
Letsgosomewhere	68
Juang	52
Zonafotografi	51
Selamat	50
Sman	50

Berdasarkan data Tabel 6.8, pengurangan data awal dan akhir disebabkan oleh terhapusnya sebuah caption yang pendek pada pra-proses data dan mengakibatkan kosongnya arraylist pada python, sehingga harus dilakukan penghapusan.

Tabel 6.8 Hasil pra-proses data

	Sebelum Stopwords Removal	Sesudah Stopwords Removal
Jumlah Dokumen	3346	2951
Jumlah Kata	28786	16028
Kata unik	19441	7501
Jumlah author (Sekolah)		18

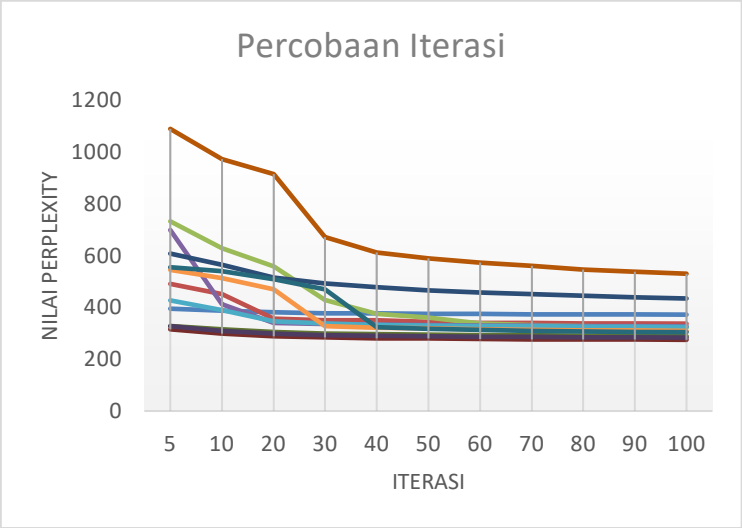
6.1.3 Pembentukan Model Topik dan Validasi

Dalam melakukan pembentukan model ATM, model melalui tahap validasi dengan cara pengujian jumlah iterasi menggunakan nilai perplexity.

Penentuan jumlah iterasi.
Jumlah iterasi ditentukan oleh pengujian passes setiap topik dan menganalisis saat titik iterasi tertentu akan mengalami kestabilan nilai perplexity ditandai dengan selisih nilai iterasi yang semakin kecil dan stabil. Percobaan analisis nilai perplexity menggunakan jumlah topik 1 hingga 12, dan 100 passes. Hasil dari percobaan tersebut kemudian dicatat dan divisualkan kedalam bentuk grafik. Gambar 6.2 merupakan hasil dari percobaan, kemudian dibuat sebuah grafik yang memudahkan kita mengetahui di titik iterasi mana kestabilan itu terjadi. Gambar 6.3 merupakan selisih setiap 10 passes, sehingga ditemukan pada titik iterasi 90, nilai perplexity mulai stabil.

	Number of Topics											
	1	2	3	4	5	6	7	8	9	10	11	12
5	394.3	490	731.1	697.5	426.3	543.9	606.7	314.9	327.5	327	554.1	1087.3
10	387.4	451	628.1	412.1	388.3	512	562.7	298.7	315.4	309.1	538.9	970.6
20	380.4	356.9	557.9	340.3	345.3	469.5	514.3	288	303.8	297.6	507.4	914
30	377.4	349.7	427	335.5	337.1	327.7	492.2	283.4	298.8	292.7	468.8	669.5
40	375.6	348.9	375.3	332.7	331.1	321.9	477.5	280.7	295.8	289.8	323.5	611.4
Passes 50	374.3	342.6	359.7	330.9	328.8	318.3	465.4	278.9	293.8	287.8	316.4	588.3
60	373.4	340	337.7	329.5	327.4	315.5	457.2	277.5	292.4	286.4	311.8	572.7
70	372.8	338.5	331	328.4	326.3	313	449.7	276.5	291.3	285.3	308.9	558.6
80	372.2	337.8	327.7	327.5	325.3	311.2	444	275.7	290.4	284.4	306.8	546
90	371.7	337.2	324.2	326.8	324.5	310.1	438.7	275	289.7	283.6	305.3	536
100	371.4	336.7	321.3	326.1	323.8	308.8	433.6	274.4	289	283	304	529.1

Gambar 6.2 Hasil percobaan menggunakan 100 passes dari 1 hingga 12 jumlah topik



Gambar 6.3 Visualisasi grafik nilai perplexity sesuai percobaan iterasinya

		Number of Topics											
		1	2	3	4	5	6	7	8	9	10	11	12
Passes	5												
	10	6.9	39	103	285.4	38	31.9	44	16.2	12.1	17.9	15.2	116.7
	20	7	94.1	70.2	71.8	43	42.5	48.4	10.7	11.6	11.5	31.5	56.6
	30	3	7.2	130.9	4.8	8.2	141.8	22.1	4.6	5	4.9	38.6	244.5
	40	1.8	0.8	51.7	2.8	6	5.8	14.7	2.7	3	2.9	145.3	58.1
	50	1.3	6.3	15.6	1.8	2.3	3.6	12.1	1.8	2	2	7.1	23.1
	60	0.9	2.6	22	1.4	1.4	2.8	8.2	1.4	1.4	1.4	4.6	15.6
	70	0.6	1.5	6.7	1.1	1.1	2.5	7.5	1	1.1	1.1	2.9	14.1
	80	0.6	0.7	3.3	0.9	1	1.8	5.7	0.8	0.9	0.9	2.1	12.6
	90	0.5	0.6	3.5	0.7	0.8	1.1	5.3	0.7	0.7	0.8	1.5	10
	100	0.3	0.5	2.9	0.7	0.7	1.3	5.1	0.6	0.7	0.6	1.3	6.9

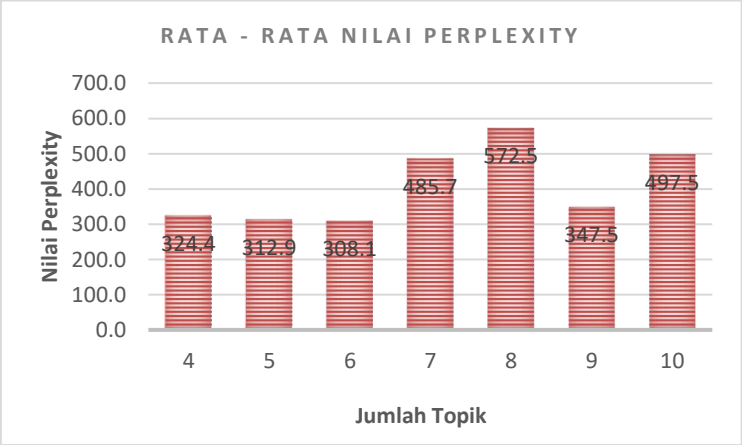
Gambar 6.4 Selisih setiap nilai perplexity berdasarkan passes dan jumlah topik

Penentuan jumlah topik

Pada tahap penentuan jumlah topik, nilai perplexity dijadikan acuan untuk analisa topik. Nilai perplexity terkecil nantinya akan digunakan sebagai landasan untuk memilih jumlah topik terbaik. Untuk mencari nilai tersebut, kita perlu melakukan percobaan dengan menjalankan model sebanyak 30 kali berdasarkan jumlah topik yang dipilih. Kemudian dihitung rata-rata dan standar deviasi setiap jumlah topik, topik terbaik akan dijadikan model dan hasilnya akan divisualisasikan pada aplikasi TeenStagram.

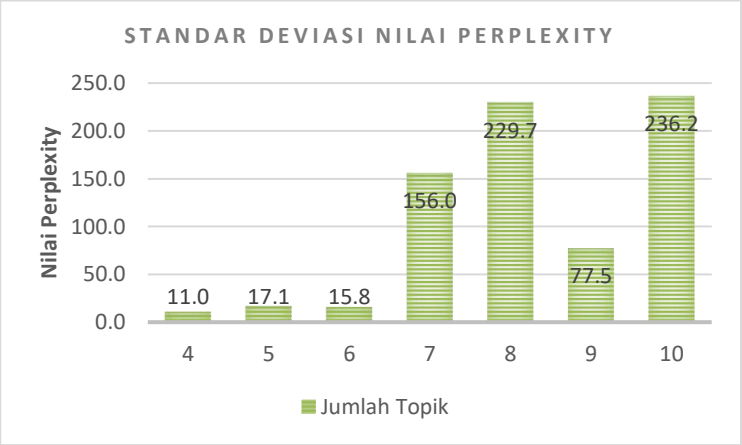
Menghitung rata-rata nilai perplexity

Berdasarkan Gambar 6.2 topik yang akan dilakukan percobaan adalah rentang antara 4 hingga 10 topik karena memiliki nilai perplexity terkecil. Setelah melakukan uji coba dan running 30 kali, didapatkan hasil pada Grafik 6.4 yang menunjukkan bahwa jumlah 6 topik memiliki rata – rata terkecil sebesar 308.1 dibanding lainnya.



Gambar 6.5 Hasil percobaan rata-rata nilai perplexity tiap jumlah topik

Setelah itu perlu dihitung standar deviasi untuk mengetahui bagaimana rentang nilai dari rata – rata nilai perplexity setelah di uji coba. Hasil dari perhitungan dapat dilihat pada Grafik 6.6 dan dipatkan bahwa standar deviasi terkecil adalah 11.0 dengan jumlah topik 4, setelah dikaitkan oleh Gambar 6.2 nilai perplexity kecil kadang muncul di beberapa jumlah topik, namun tidak selalu optimal karena memiliki standar deviasi yang besar, sebagai contoh jumlah topik 8 memiliki perplexity terbaik saat itu, namun setelah diuji 30 kali, hasilnya menunjukkan bahwa terdapat kesenjangan yang tinggi dengan standar deviasi 229.7 yang mengakibatkan topik tidak optimal.



Gambar 6.6 Standar Deviasi hasil percobaan nilai perplexity

Dengan melakukan uji tersebut, maka penelitian ini menentukan membuat model dengan jumlah topik 4 dan 6 karena keduanya memiliki nilai rata – rata dan standar deviasi perplexity terkecil, yaitu 4 topik dengan rata-rata 324.4 dan 11.0 standar deviasi, dan 6 topik dengan 308.1 rata-rata dan 15.8 standar deviasi perplexity-nya.

Pembentukan Author-topic Models

Pembentukan model dengan ATM berdasarkan nilai perplexity terkecil sesuai dengan percobaan running 30 kali, dan mendapatkan hasil 4 dan 6 sebagai topik yang akan dijadikan acuan model author-topic, dan hasilnya adalah sebagai berikut:

Percobaan jumlah 4 Topik

Setelah melakukan running model, dan didapatkan hasil 4 topik, penentuan jumlah topik 4 dinilai tidak ideal karena persebaran kata yang masih sulit untuk diklasifikasikan berdasarkan topik masing – masing, bisa dilihat pada Tabel 6.9 dan Tabel 6.10. Topik terbaik apabila kata yang muncul dalam satu topik berkaitan satu sama lain, sehingga membentuk satu topik dan dapat diklasifikasikan labelnya.

Tabel 6.9 Topik 0 dan Topik 1 hasil percobaan 4 jumlah topik

Topik 0		Topik 1	
Kata	**Prob.**	**Kata**	**Prob.**
libur	0.005432	juang	0.00391
surabaya	0.003469	asus	0.00371
sukses	0.003234	friendships	0.00350
juang	0.002774	surabaya	0.00349
mimpi	0.002715	photo	0.00328
hati	0.002602	insecta2k17	0.00301
selamat	0.002484	goodphotos	0.00265
allah	0.002480	hati	0.00238
kenang	0.002429	nurussalam	0.00224
indonesia	0.002197	holiday	0.00217
Sekolah	**Prob.**	**Sekolah**	**Prob.**
N02	0.999997	N13	0.999984
N07	0.999995	N16	0.999983
N14	0.999980	N21	0.999995
N17	0.999966	S03	0.999740
S06	0.999963		

Tabel 6.10 Topik 2 dan Topik 3 Hasil percobaan 4 jumlah topik

Topik 2		Topik 3	
Kata	**Prob.**	**Kata**	**Prob.**
global	0.00662	surabaya	0.00950
thecreatorclass	0.00645	sman	0.00751
letsgosomewhere	0.00556	zetcon2k17	0.00627
smadaf2017	0.00470	allah	0.00614
exploretocreate	0.00401	sman_surabaya	0.00611
global_explore	0.00401	zonafotografi	0.00552

tocreate			
surabaya	0.00400	alaihi	0.00423
thecreatorclass_ letsgosomewhere	0.00377	shallallahu	0.00401
film	0.00361	rasulullah	0.00379
summer	0.00316	abu	0.00359
Sekolah	**Prob.**	**Sekolah**	**Prob.**
M01	0.999971	K10	0.999986
S01	0.999982	N06	0.999974
S04	0.999971	N10	0.999940
S05	0.999946	N20	0.999972
		S02	0.999967

Percobaan jumlah 6 Topik

Berdasarkan hasil percobaan jumlah 6 topik, pada Tabel 6.11, 6.12 dan 6.13 dapat dilihat bahwa kata – kata setiap topik mulai terlihat mirip satu dengan lainnya. Sehingga dapat diklasifikasikan setiap topik memiliki label tertentu, seperti contoh untuk topik 5 dapat diklasifikasikan ke dalam label agama dan musik, karena probabilitas kata – kata yang paling tinggi membahas akan hal itu.

Tabel 6.11 Topik 0 dan Topik 1 Hasil percobaan 6 jumlah topik

Topik 0		Topik 1	
Kata	**Prob.**	**Kata**	**Prob.**
rindu	0.00556	sman_surabaya	0.06266
hati	0.00398	sman_surabaya_ sman_surabaya	0.02331
selamat	0.00380	surabaya	0.01842
kuat	0.00374	sman	0.01801
sukses	0.00350	zetcon2k17	0.01657
surabaya	0.00315	simangasing	0.00891

gagal	0.00263	ubszetizenjournalistblog	0.00832
lff	0.00233	ubsjournalistblog	0.00537
semangat	0.00227	sims	0.00478
libur	0.00221	juang	0.00267
Sekolah	**Prob.**	**Sekolah**	**Prob.**
N07	0.999988	N13	0.999976
N06	0.999959	S05	0.999910
M01	0.999956		

Tabel 6.12 Topik 2 dan Topik 3 Hasil percobaan 6 jumlah topik

Topik 2		Topik 3	
Kata	**Prob.**	**Kata**	**Prob.**
global_exploretocreate_ thecreatorclass_ letsgosomewhere	0.01691	zonafotografi	0.01030
kodak_film	0.01598	surabaya	0.00627
global_exploretocreate	0.01531	fotograferjomblo	0.00573
global_createexplore	0.01513	sman_surabaya	0.00562
thecreatorclass_ letsgosomewhere	0.01334	randomnesia	0.00561
thecreatorclass_ citybestviews	0.00790	fotograferpetakilan	0.00538
global	0.00762	cabello	0.00425
thecreatorclass	0.00743	camila	0.00425
letsgosomewhere	0.00640	sukses	0.00372
global_ createexplore_ thecreatorclass_ citybestviews	0.00593	selamat	0.00325
Sekolah	**Prob.**	**Sekolah**	**Prob.**
N02	0.99999	K10	0.999978
S01	0.99997	N16	0.999973
N20	0.99996	S04	0.999957

S02	0.99995
S06	0.99993
N10	0.04141

Tabel 6.13 Topik 4 dan Topik 5 Hasil percobaan 6 jumlah topik

Topik 4		Topik 5	
Kata	**Prob.**	**Kata**	**Prob.**
summer	0.01025	allah	0.010554
libur	0.008552	alaihi	0.008718
rodaduasampetua	0.006194	shallallahu	0.008258
cosmicv	0.004028	rasulullah	0.007799
surabaya	0.003993	abu	0.007401
cosmic	0.003368	imam	0.007339
handlettering	0.003368	wasallam	0.006486
onedayinsummer	0.003368	sabda	0.006222
smantass31	0.003368	ibnu	0.006088
mimpi	0.003074	music	0.005946
Sekolah	**Prob.**	**Sekolah**	**Prob.**
N14	0.999953	N21	0.999990
N17	0.999919		
S03	0.999528		
N10	0.958521		

Jika dibandingkan keduanya, penentuan jumlah topik 4 dikarenakan memiliki standar deviasi perplexity terkecil, dan jumlah topik 6 memiliki rata – rata nilai perplexity terkecil, ketika di analisa lebih lanjut, pemilihan topik terbaik berdasarkan rata – rata nilai perplexity terkecil, maka dari itu, jumlah topik 6 dapat dikatakan topik terbaik untuk menjadi hasil pada penelitian ini sesuai dengan nilai perplexity-nya.

Analisa hasil model

Berdasarkan hasil permodelan pada Tabel 6.11, 6.12 dan 6.13 didapatkan 10 kata dengan probabilitas terbesar, dan author (kode sekolah) serta probabilitasnya membahas topik tersebut.

1. **Kata dalam setiap topik**

   Penyebaran kata – kata dibilang belum sempurna dikarenakan caption yang digunakan oleh siswa SMA cenderung pendek dan memiliki kosa kata yang berbeda – beda setiap anak dan setiap sekolah. Probabilitas yang tinggi didominasi oleh kata yang sering muncul, dan ditemukan beberapa kata menarik setiap topiknya yang menjadi acuan dalam pelabelan topik. Topik 0 dapat diklasifikasikan menjadi Topik "perasaan" karena terdapat beberapa kata – kata yang menyangkut masalah perasaan, seperti contoh: rindu, hati, kuat, gagal, dll. Topik 1 dapat diklasifikasikan menjadi topik "event Surabaya" atau sebuah /*-+acara yang melibatkan siswa SMA di Surabaya, kata +yan++g dominan seperti contoh: sman Surabaya, zetcon2k17, simangasing, ubsjournalistblog, daf, dll. Topik 2 dapat diklasifikasikan menjadi topik "fotografi" karena mengandung banyak sekali sebuah hashtag dari fotografi yang biasa di upload oleh kebanyakan orang di Instagram, seperti contoh kata: global, exploretocreate, thecreatorclass letsgosomewhere, kodak film, dll. Topik 3 lebih membahas mengenai "fotografi dan artis", bedanya dengan topik 2, kata – kata dalam topik 3 lebih cenderung tercampur dengan beberapa nama artis yang sedang popular belakangan ini, contoh katanya adalah zona fotografi, fotografer jomblo, camila cabelo, dll. Pada topik 4, banyak siswa yang upload dengan caption seperti: libur, summer, rodaduasampetua, cosmicv, dll yang berarti label yang tepat untuk topik ini adalah "liburan", dan yang terakhir adalah topik 5 yang membahas mengenai agama, karena kata – kata didalamnya terindikasi seorang siswa yang sedang berdoa, ataupun mengutip sebuah dalil Al Quran, seperti contoh: Allah, alaihi, shalallahu, rasulullah, abu, imam, dll.

2. Author pada setiap topik

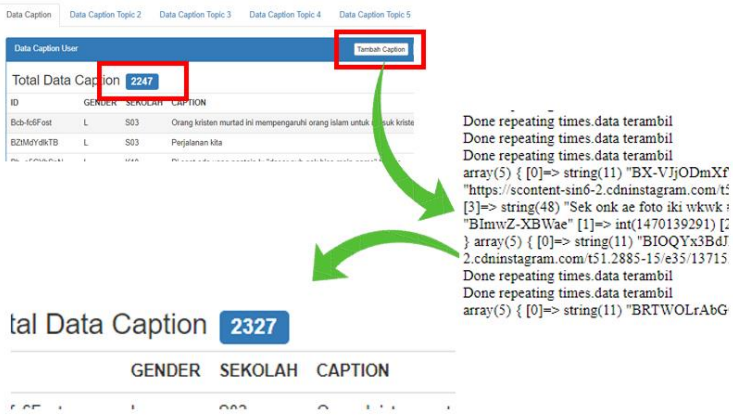
Pada author (sekolah) didapatkan probabilitas yang tinggi dikarenakan kecenderungan pada sekolah tertentu memiliki anak yang selalu upload Instagram dengan caption yang sama dan berulang dengan jumlah yang besar, sehingga memengaruhi perhitungan karena kata – kata yang paling sering muncul memiliki perhitungan yang sangat besar dan berdampak pada angka probabilitas. Semua author memiliki masing – masing topik kecuali pada author N10 yang memiliki topik di topik 2 sebesar 0.04141 (4%) dan pada topik 4 sebesar 0.9585 (96%), setelah dianalisis lebih lanjut pada kata yang sering muncul ditemukan bahwa kata yang termasuk dalam masing – masing topik terdapat lebih banyak pada topik 4, dan menyisakan kata "allah" dan juang" pada topik 2, hal itu menyebabkan author memiliki 2 topik yang berbeda.

6.2 Pengujian Fungsional

Pengujian Fungsionalitas dari aplikasi ini dengan cara mencoba berbagai scenario penggunaan aplikasi dimana setiap skenario menguji fungsionalitas berbeda dari aplikasi. Adapun pengujian yang akan dilakukan yaitu melakukan pengujian pada fitur penambahan data caption melalui proses crawling yang disediakan pada tombol tambah caption dan melakukan pengujian pada fitur pembuatan model dengan cara mengeksekusi script python melalui PHP.

Fitur tambah data

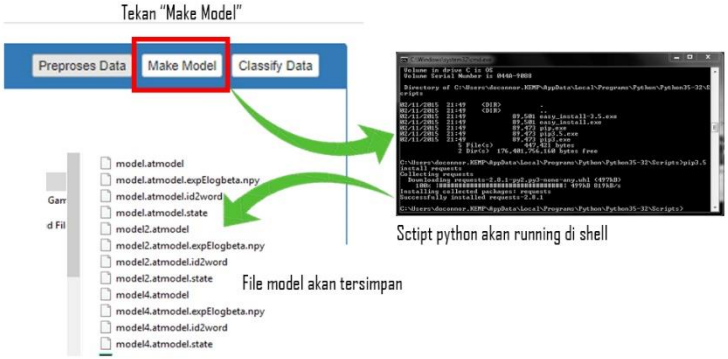
Terdapat fitur tambah data untuk melakukan crawling berdasarkan username didalam database, pengujian ini dilakukan untuk mengecek apakah data bertambah sesudah dan sebelum dijalankannya fitur. Lebih jelas dilihat pada Gambar 6.3



Gambar 6.7 Pengujian fungsional: tambah data

Fitur Make Model

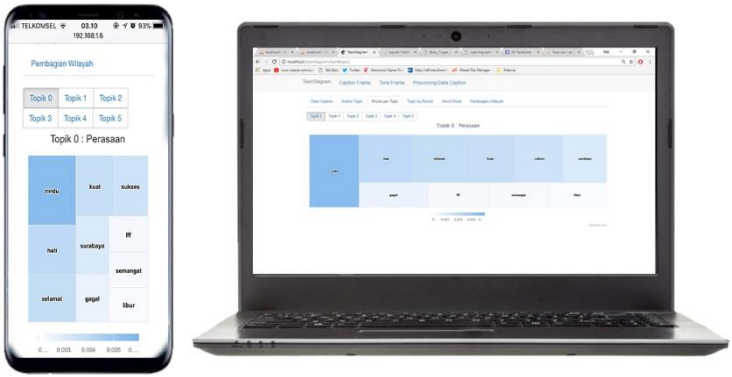
Fitur ini untuk membuat model dari *script PHP* untuk menjalankan *script Python*, dan menyimpan model, pengujian ini meninjau apakah model tersimpan pada komputer atau tidak. Bisa dilihat pada Gambar 6.4



Gambar 6.8 Pengujian fungsional: make model

6.3 Pengujian Non Fungsional

Pengujian non fungsional dilakukan dengan cara membandingkan hasil tampilan aplikasi *TeenStagram* pada *platform* yang berbeda. *Platform* yang digunakan adalah tampilan web pada browser *computer* dan tampilan *web* pada *mobile* (*handphone*).



Gambar 6.9 Tampilan web dan tampilan mobile pada pengujian non fungsional

BAB VII
KESIMPULAN DAN SARAN

Pada bab ini akan dijelaskan kesimpulan dan saran dalam pengerjaan tugas akhir.

7.1 Kesimpulan

Berdasarkan proses-proses yang telah dilakukan dalam pengerjaan tugas akhir dengan judul "Rancang Bangun Perangkat Lunak Teenstagram untuk Mengelompokkan Topik Caption Akun Instagram dengan Permodelan *Author Topic Models* (Studi Kasus: Siswa SMA di Surabaya)" yang telah dilakukan, dapat disimpulkan sebagai berikut:

1. Penelitian ini membuktikan bahwa metode crawling data dengan memanfaatkan Instagram API mampu menjalankan proses pengumpulan data caption milik akun siswa SMA yang telah sudah menjalankan proses validasi terlebih dahulu.
2. Hasil dari pengumpulan data dengan cara crawling data caption periode Januari 2016 hingga Desember 2017 menghasilkan 235 akun, dengan rincian 133 siswa perempuan dan 102 siswa laki – laki dari 18 sekolah.
3. Pendefinisian *stop words* yang dihasilkan pada penelitian dapat digunakan untuk tahap pra-proses data pada penelitian selanjutnya dalam lingkup sosial media.
4. Permodelan *Author-Topic Models* dapat digunakan untuk mengelompokkan sekolah – sekolah dengan topik tertentu sesuai dengan hasil permodelan. Hasil terbaik adalah penentuan 6 topik berdasarkan nilai perplexity terendah. 6 Topik tersebut diberi label sesuai dengan kata – kata yang termasuk dalam topik, hasilnya adalah perasaan, event Surabaya, fotografi, fotografi dan artis, liburan, dan agama dan music.
5. Perancangan aplikasi *TeenStagram* dapat mengintegrasikan Bahasa *PHP* untuk kebutuhan

crawling dan visualisasi, dan Bahasa *Python* untuk pra-proses data, hingga pembuatan model mampu menjawab kebutuhan visualisasi caption SMA di Surabaya.

7.2 Saran

Saran penulis untuk penelitian selanjutnya sebagai berikut:

1. Adanya permasalahan mengenai kosa kata yang digunakan oleh siswa SMA yang menggunakan Bahasa tidak baku, sehingga perlu dilakukannya normalisasi kata sebelum dilakukan Topic Modeling sehingga menghasilkan model yang optimal.
2. Belum diimplementasikannya fitur untuk mengklasifikasikan dokumen dengan topik tertentu, sehingga belum mendapatkan analisis lebih dalam tentang perilaku anak SMA di Surabaya.
3. Untuk penelitian kedepan, diharapkan adanya tambahan untuk mengidentifikasi kalimat yang berkonotasi negatif dan positif sehingga dapat dilakukannya tindakan untuk caption yang tidak bertanggung jawab.
4. Untuk aplikasi *TeenStagram* kedepannya perlu dilakukan penelitian lebih lanjut mengenai waktu (*time*), gambar (*image recognition*), ataupun cara *crawling* yang lebih optimal, sehingga memperkaya fitur aplikasi dalam analisa sosial media, terutama *Instagram*.

DAFTAR PUSTAKA

[1] Blog Instagram (2016, Desember) [Online]. http://blog.instagram.com/post/154506585127/161215-600million

[2] Survey Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) 2016.11

[3] Keke Mahardika. 2015. *Pengaruh Instagram terhadap Kehidupan Remaja*. Artikel.

[4] FAQ Instagram (2017, September) [Online]. https://www.instagram.com/about/faq/

[5] Berita CNN Indonesia (2016, Juni) [Online]. http://www.cnnindonesia.com/teknologi/20160623112758-185-140353/ada-22-juta-pengguna-aktif-instagram-dari-indonesia/

[6] Developer Instagram (2017, September) [Online]. https://www.instagram.com/developer/

[7] Developer Instagram. *What are your limits on instagram?* [Online] , February 2017.

[8] Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. *Latent dirichlet allocation. Journal of Machine Learning Research*.

[9] Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; and Smyth, P. 2004. *The author-topic model for authors and documents*. UAI.

10] Liangjie Hong; Brian D. Davison. 2016. *Empirical Study of Topic Modeling in Twitter*.

11] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers. Learning author-topic models from text corpora. ACM Transactions on Information Systems, 28(1):1–38, 2010.

12] David Alvarez-Melis; Martin Saveski. 2016. *Topic Modeling in Twitter: Aggregating Tweets by Conversations.* ICWSM
13] Allan Heydon and Marc Najork. Mercator: *A scalable, extensible web crawler*. World Wide Web, 2(4):219–229, 1999.
14] Carlos Castillo. *Effective web crawling. In ACM SIGIR Forum*, volume 39, pages 55–56. Acm, 2005
15] D. M. Blei, "Probabilistic Topic Models," pp. 77–84.
16] M. Steyvers, P. Smyth, and M. Rosen-zvi, *Probabilistic Author-Topic Models for Information Discovery*, no. 1990, pp. 306–315, 1999.
17] Stefan Stieglitz and Linh Dang-Xuan. *Social media and political communication: a social media analytics framework. Social Network Analysis and Mining*, 3(4):1277–1291, 2013.
18] Tetha Valianta. *Rancang Bangun Perangkat Lunak Teenstagram untuk Mengelompokkan Topik Caption Akun Instagram Siswa Sekolah Menengah Pertama di Surabaya*. 2017
19] Z. Xu, R. Lu, L. Xiang, and Q. Yang, *Discovering User Interest on Twitter with a Modified Author-Topic Model*, 2011.
20] Yuheng Hu, Lydia Manikonda, Subbarao Kambhampati, et al. *What we instagram: A first analysis of instagram photo content and user types*. In ICWSM, 2014.
21] Hasan Alwi. dkk. 1998. Tata Bahasa Baku Bahasa Indonesia

BIODATA PENULIS

Penulis lahir di kota industri, Gresik pada tanggal 25 Februari 2016. Anak kedua dari tiga bersaudara yang telah menempuh pendidikan formal yaitu; SD Negeri Sukomulyo 2 Gresik, SMP Negeri 1 Manyar, dan SMA Negeri 1 Gresik.

Pada tahun 2014, penulis melanjutkan pendidikan ke jenjang yang lebih tinggi, yaitu di Institut Teknologi Sepuluh Nopember (ITS) Surabaya, sebagai mahasiswa departemen Sistem Informasi, Fakultas Tekonologi Informasi dan Komunikasi (FTIK). Terdaftar sebagai pemilik NRP 5214100131. Selama menjadi mahasiswa, penulis banyak mengikuti kegiatan kemahasiswaan, antara lain seminar, organisasi dan kajian. Penulis diberikan tanggung jawab oleh organisasi BEM Fakultas Teknologi Informasi dan Komunikasi (BEM FTIK) sebagai Kepala Departemen Student Resource Development (SRD) pada tahun kepengurusan 2016/2017. Disamping aktif dalam kegiatan kemahasiswaan, penulis juga pernah menjadi Asisten Praktikum mata kuliah Perencanaan Sumber Daya Perusahaan.

Pada tahun ke-4, penulis tertarik dengan bidang Social Media Analysis, sehingga mengambil bidang minat laboratorium Akuisisi Data dan Diseminasi Informasi (ADDI) dan lulus dalam waktu 3.5 tahun atau 7 semester. Penulis dapat dihubungi melalui email faiznfi@gmail.com

Dataset yang dihasilkan

No	Dataset
1	Pendefinisian Stopwords anak SMP + SMA
2	Dataset akun Instagram siswa SMA di Surabaya
3	Hasil Crawling caption akun Instagram siswa SMA
4	Hasil Pra-proses caption.