REPOSITOR, Vol. 2, No. 4, April 2020, Pp. 415-426
ISSN : 2714-7975
E-ISSN : 2716-1382

Deteksi Topik Tentang Tokoh Publik Politik Menggunakan Latent Dirichlet Allocation

Faizun Nuril Hikmah*[1], Setio Basuki[2], Yufis Azhar[3]
[1,2,3]Teknik Informatika/Universitas Muhammadiyah Malang
faizunnuril03@gmail.com[*1], setiobasuki.umm@gmail.com[2], yufis@umm.ac.id[3]

Abstrak

Twitter merupakan salah satu Social Networking yang memperbolehkan pengguna untuk mengirim dan membaca sebanyak 140 karakter. Berdasarkan survey sekitar 500 juta tweet tiap harinya yang dikirim melalui twitter. Data-data tersebut dapat berupa opini-opini publik mengenai politik, tokoh publik, makanan, dan lain sebagainya. Data tersebut akan diolah dengan teknik Topic Detection untuk menghasilkan suatu topik yang sedang marak dibicarakan masyarakat tentang tokoh publik politik. Permasalahan dalam penulisan ini yaitu, bagaimana mengekstraksi suatu tweet tentang tokoh publik politik dari pengguna Twitter. Data tweet yang diambil tentang tokoh publik politik diantaranya yaitu mengenai Joko Widodo, Basuki Tjahaja Purnama (Ahok), Anies Baswedan, Sandiaga Uno, dan Habib Rizieq Shihab. Dengan adanya data atau tweet tentang tokoh publik politik dapat diolah menggunakan metode Agglomerative untuk mengcluster tiap data yang akan digunakan sebagai topik acuan, LDA (Latent Dirichlet Allocation) yang akan berfungsi sebagai pemodelan topik dari tweet-tweet yang telah tercluster, serta TF-IDF untuk mengetahui tweet mana saja yang mengandung kata-kata dalam LDA yang akan dijadikan sebagai topik acuan. Sehingga akan menghasilkan deteksi topik yang relevan berdasarkan tweet mengenai tokoh publik politik.

Kata Kunci: *Deteksi Topic, Tokoh Publik Politik, Agglomerative Hierarchical Clustering, LDA, TF-IDF*

Abstract

Twitter is a Social Networking website that allows users to send and read 140 characters. Based on a survey of about 500 million tweets per day sent via twitter. These data can be public opinions about politics, public figures, food, and more. The problem in this paper is, how to extract a tweet about the public political character of Twitter users. Tweet data taken on public political figures include Joko Widodo, Basuki Tjahaja Purnama (Ahok), Anies Baswedan, Sandiaga Uno, and Habib Rizieq Shihab. Given the data or tweets about public political figures it will be processed by using Agglomerative method to cluster each data, LDA (Latent Dirichlet Allocation) which will serve as a topic modeling of twisted tweets, and TF-IDF for know which tweets contain the words in the LDA that will serve as a reference topic. This will result in the detection of relevant topics based on tweets about public political figures.

Keywords: *Topic Detection, Public Political Figure, Agglomerative Hierarchical Clustering, LDA, TF-IDF*

1. Pendahuluan

Twitter merupakan Social Networking website yang memperbolehkan pengguna untuk mengirim dan membaca 140 karakter, atau sering disebut *tweets* [1]. Berdasarkan survey yang telah dilakukan terdapat 304 juta pengguna tiap bulannya yang aktif menggunakan twitter. Sekitar 500 juta tweet tiap harinya yang dikirim melalui twitter. Dengan banyaknya tweet yang pengguna posting tiap harinya, terdapat banyak pula data-data yang didapat dari tweet-tweet tersebut. Data-data tersebut dapat berupa opini-opini publik mengenai politik, tokoh publik, makanan, dan lainnya. Pada penelitian ini, data yang diambil tentang opini atau postingan pengguna tentang tokoh publik khusunya tokoh publik yang bergelut dalam dunia politik. Data tersebut akan diolah dengan teknik *Topic Detection* untuk menghasilkan suatu topik yang sedang marak dibicarakan masyarakat tentang tokoh publik politik.

Untuk melakukan *Topic Detection* atau Deteksi Topik dapat diimplementasikan menggunakan metode *Clustering*, dimana pengertian clustering itu sendiri ialah suatu metode

pengelompokan berdasarkan ukuran kedekatan (kemiripan) suatu data. Pengelompokkan klaster berdasarkan nilai kedekatan suatu sample data yang ada dan pengelompokkan tidak harus mempunyai data yang sama. Secara garis besar ada beberapa kategori dalam clustering, seperti metode hirarki dimana pada metode hirarki terdapat dua jenis algoritma, salah satu algoritmanya yaitu, *Agglomerative.* Dimana algoritma Agglomerative yang akan diimplementasikan dalam penelitian ini untuk mengetahui cluster-cluster dari data yang telah terkumpul. Topik yang telah ter-*cluster* akan diolah menggunakan metode *LDA (Latent Dirichlet Allocation)* untuk menghasilkan pemodelan topik mengenai tokoh publik politik.

Permasalahan dalam penulisan ini yaitu, bagaimana mengekstraksi suatu tweet tentang tokoh publik politik dari pengguna Twitter. Data tweet yang diambil tentang tokoh publik politik diantaranya yaitu mengenai *Joko Widodo, Basuki Tjahaja Purnama (Ahok), Anies Baswedan, Sandiaga Uno,* dan *Habib Rizieq.* Dengan adanya data atau tweet tentang tokoh publik politik tersebut akan diolah dengan menggunakan metode *Agglomerative* untuk mengcluster tiap data dan *LDA (Latent Dirichlet Allocation)* yang akan berfungsi sebagai pemodelan topik dari tweet-tweet yang telah ter*cluster*. Hal tersebut akan memudahkan pengguna untuk mengetahui suatu topik apakah yang sedang dibicarakan tentang tokoh publik politik. Namun, deteksi topik disini berbeda dengan *Trending Topic* pada Twitter. Deteksi topik yang diangkat dalam permasalahan tugas akhir ini yaitu mendeteksi topik tentang tokoh publik politik, dimana data-data yang digunakan untuk deteksi topik bukan mendeteksi topik yang sedang marak diperbincangkan pada hari itu seperti *Trending Topic*. Sedangkan pada *Trending Topic* yaitu mendeteksi topik apa yang sedang marak diperbincangkan saat itu juga dan pada hari itu.

Berbeda dengan penelitian-penelitian sebelumnya, peneliti *Saud Alashri* [2], mendistribusikan beberapa topik mengenai calon presiden AS, seperti topik tentang Iran, ISIS, Imigrasi, dan sebagainya. Dari topik-topik tersebut akan dideteksi *trending topik* dari komentar publik mengenai calon presiden AS yang menggunakan metode *LDA* dan juga menganalisa sentimen dari topik calon presiden AS. Serta pada penelitian *Suvarna D. Tembhurnikar* [3], menggunakan metode *BNgram* untuk melakukan deteksi topik dari sosial media, *Twitter*. Dimana pada metode *BNgram* untuk meng*klaster* data menggunakan perhitungan "*df-idft*", berdasarkan nilai *df-idft* yang telah dihitung untuk setiap *n-gram* maka peringkat *n-gram* akan dibuat dan dibentuk menjadi *cluster*. Proses *clustering* diulang sampai kesamaan antar *cluster* terdekat turun dibawah nilai ambang batas yang telah ditentukan. Semua *cluster* dibentuk dengan me*rangking* berdasarkan nilai *df-idft* tertinggi. Dan setiap klaster merupakan satu topik. Pada penelitian *Tahta Alfina* [4], melakukan analisa perbandingan antara metode *Hierarchical Clustering* dan *K-Means* dalam mengklaster data. Dengan didapatkan hasil bahwa algoritma *Hierarchical Cluster* menghasilkan pengelompokkan yang jauh lebih baik daripada *K-Means* dalam semua pengujian.

Dari penelitian-penelitian sebelumnya dan juga permasalahan yang ada. Pada tugas akhir ini akan membangun sebuah perangkat lunak Deteksi Topik Tentang Tokoh Publik Politik Menggunakan Twitter API. Dengan demikian, pengguna akan lebih mudah untuk membaca topik apakah yang sedang dibicarakan mengenai tokoh publik politik.

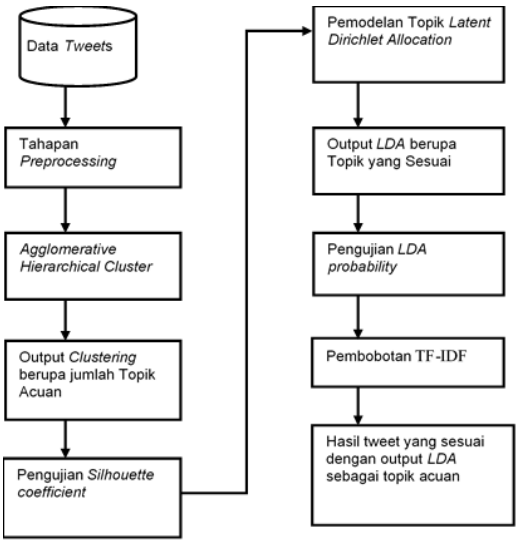
2. Metode Penelitian

2.1. Analisa Masalah

Dalam media sosial seperti *Twitter*, masyarakat banyak mengeluarkan opini-opini pendapat mereka tentang publik figur atau bahkan tokoh politik seperti, *Joko Widodo, Basuki Tjahaja Purnama (Ahok), Sandiaga Uno, Anies Baswedan* dan *Habib Rizieq*. Dengan adanya opini-opini masyarakat tersebut dapat diolah dengan menggunakan metode *Agglomerative Hierarchical Clustering* untuk *mengelompokkan* data *tweet* mengenai lima tokoh tersebut. Sehingga akan menghasilkan beberapa kelompok yang nantinya dapat diketahui jumlah topik acuannya. Selain itu pada penelitian ini juga menggunakan metode *Latent Dirichlet Allocation (LDA)* yang digunakan sebagai pengolahan topik model yang akan menghasilkan topik mengenai lima tokoh tersebut. Dikarenakan hasil pemodelan topik *LDA* berupa kata-kata yang tidak memiliki makna dan *user* tidak dapat memahami makna topik tersebut, maka menggunakan metode *TF-IDF* untuk mengetahui *tweet* mana saja yang banyak mengandung kata-kata dalam *LDA*. *Tweet* tersebut yang akan dijadikan sebagai topik acuan agar *user* mengetahui makna dari topik-topik yang telah dihasilkan menggunakan metode *LDA* dan akan dijadikan sebagai topik akhir dari tugas akhir ini. Sehingga tujuan pada penelitian ini yaitu mengetahui topik apakah yang sedang dibicarakan oleh masyarakat tentang lima tokoh yang menjadi acuan dalam penelitian ini dapat tercapai.

2.2. Analisa Desain Sistem

Adapun gambaran dari sistem yang akan dibangun pada penelitian tugas akhir ini akan dijelaskan pada Gambar 1.



Gambar 1. Desain Sistem

2.2.1. Tahapan Preprocessing

Pada tahap ini bertujuan untuk menghilangkan *Stopword Removal* (kata-kata dasar yang tidak penting, seperti kata “ada”,”siapa”,”dari”, dan kata-kata dasar lainnya). Selain itu untuk *Normalisasi Fitur* (menghilangkan komponen dalam *tweet* seperti *@username, RT, #Hastag, URL*). Serta untuk merubah huruf menjadi huruf kecil. Berikut merupakan contoh data *tweet* dari kelima tokoh pada tahap *preprocessing* pada Tabel 1.

Tabel 1. Data Preprocessing

Data Sebelum *preprocessing*	Data Sesudah *preprocessing*
RT @singa_com: #KANDANGKANRIZIEQ #KandanginRizieq ARAB BIADAB @syihabrizieq PENGHINA PANCASILA, MENGKRITIK ALMARHUM BUNG KARNO, PENISTA	kngkanrizieq knginrizieq arab biadab penghina pancasila mengkritik almarhum bung karno penista
RT @DpaOdojCIREBON: 15 Gara-gara Ahok, semua dibikin ribet, eksekusi #AhokHarusDipenjara secepatnya	gara ahok dibikin ribet eksekusi ahokdipenjara secepat
RT @wakilgubernurKW: Presiden @jokowi Di Tuduh PKI Oleh @syihabrizieq Trus @DivHumasPolri @CCICPolri @BNPTRI Diam Aja? #JokowiTakutFPI ?	presiden di tuduh pki trus m aja jokowittfpi
RT @sandiuno: #MakasihAnies trending topic? Berarti kalian semua yakin kalau pendidikan merupakan kunci utama dalam pembangunan manusia.	sihanies trending topic pendidikan merupkunci utama pembann manusia
RT @gitamontolalu: https://t.co/4lHGsbrnIK Kok bisa anggaran berlebihan Rp.23.3T jmn Anies Baswedan ogah ah pilihdia aku Basuki Djarot No.2	anggaran rpt jmn anies baswedan ogah ah pilihbasuki djarot no

2.2.2. Tahapan Agglomerative Hierarchical Clustering

Tahapan kedua yang dilakukan setelah *preprocessing* yaitu tahapan *clustering*. metode *Agglomerative Hierarchical Clustering* sendiri ialah metode yang menggunakan strategi desain *Bottom-Up* yang dimulai dengan meletakkan setiap obyek sebagai sebuah *cluster* tersendiri dan selanjutnya menggabungkan objek-objek *cluster*. Objek-objek *cluster* tersebut menjadi *cluster* yang lebih besar dan lebih besar lagi sampai akhirnya semua obyek menyatu dalam sebuah *cluster* atau proses dapat pula berhenti jika telah mencapai batasan kondisi tertentu [5]. Hasil keseluruhan dari algoritma *Hierarchical Clustering* dapat digambarkan dalam sebuah grafik, yang disebut dengan *dendrogram*. Dimana *dendrogram* akan menggambarkan dari penggabungan klaster-klaster yang ada serta memperjelas proses hirarki tersebut [6].

Pada tahapan *clustering* pada penelitian ini menggunakan *library weka* untuk menentukan kelompok-kelompok data dan mengetahui jumlah topik acuan. Berikut parameter yang digunakan dalam tahapan *clustering* diantaranya:

- **Data Anies Baswedan**
  Debug : False
  distanceFunction : EuclidianDistance
  distanceIsBranchLength : False
  linkType : SINGLE
  numClusters : 4
  printNewick : True

- **Data Basuki Tjahaja Purnama**
  Debug : False
  distanceFunction : EuclidianDistance
  distanceIsBranchLength : False
  linkType : SINGLE
  numClusters : 4
  printNewick : True

- **Data Joko Widodo**
  Debug : False
  distanceFunction : EuclidianDistance
  distanceIsBranchLength : False
  linkType : SINGLE
  numClusters : 3
  printNewick : True

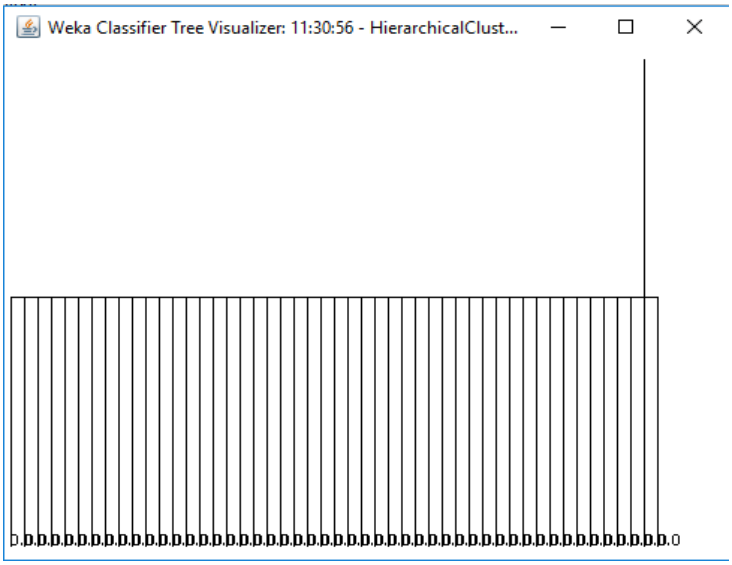
- **Data Habib Rizieq Shihab**
  Debug : False
  distanceFunction : EuclidianDistance
  distanceIsBranchLength : False
  linkType : SINGLE
  numClusters : 4
  printNewick : True

- **Data Sandiaga Uno**
  Debug : False
  distanceFunction : EuclidianDistance
  distanceIsBranchLength : False
  linkType : SINGLE
  numClusters : 3
  printNewick : True

Berikut hasil yang didapatkan dari tahapan *clustering* dapat dilihat pada Gambar 2 dan Gambar 3.

```
Cluster 0
((((((((((((((((((((((((((((0.0:1,0.0:1):0,0.0:1):0,(0.0:1,(0.0:1,0.0:1):0)

Cluster 1
((((((((((((((((((((((1.0:1,1.0:1):0,(((1.0:1,1.0:1):0,1.0:1):0,1.0:1):0)

Cluster 2
((((((((((((((((((((((((((((2.0:1,(2.0:1,2.0:1):0):0,2.0:1):0,(2.0:1,2.0

Cluster 3
((((((((((((((((((((((((((((((((((((((((((((((3.0:1,3.0:1):0,3.0:1):0,3.0:1):0,

Time taken to build model (full training data) : 0.14 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       49 ( 25%)
1       50 ( 25%)
2       50 ( 25%)
3       50 ( 25%)
```

Gambar 2. Hasil Agglomerative Hierarchical Clustering



Gambar 3 Dendrogram Agglomerative Hierarchical Clustering

2.2.3. Tahapan Pemodelan Topik LDA

Tahapan ketiga yang dilakukan setelah *clustering* yaitu tahapan pemodelan topik yang menggunakan metode *Latent Dirichlet Allocation (LDA)*. Pengertian *Latent Dirichlet Allocation (LDA)* merupakan sebuah metode statistika yang digunakan sebagai model untuk menganalisis suatu dokumen. *LDA* berusaha untuk melihat dokumen dengan cara mundur untuk menemukan satu set topik yang mungkin telah dikoleksi. *LDA* merepresentasikan dokumen dengan berbagai topik yang dibuat berdasarkan probabilitas tertentu. Probabilitas topik, merepresentasikan kejelasan dari suatu dokumen [7]. Tahapan pemodelan topik pada penelitian ini akan menghasilkan topik-topik yang sesuai dengan tokoh politik, *Basuki Tjahaja Purnama (Ahok), Joko Widodo, Anies Baswedan, Habib Rizieq, Sandiaga Uno.* Inisialisasi nilai parameter *LDA* yaitu jumlah topik (didapat dari hasil klaster), nilai α atau *Dirichlet* (nilai α yaitu 50 per topik), nilai β yaitu 0,1, jumlah iterasi 200. Berikut rumus *LDA* Persamaan 1.

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \quad (1)$$

Berikut hasil pemodelan topik *LDA* dapat dilihat pada Gambar 4.

```
0    12,5    anies ya dki a aja skriminasi bpk isu korupsi masyarakat presiden milik wakil calon beda indonesia bebas aniessanmemimpin bangun
1    12,5    pak mas jd dgn no debat an i warga lg hrs baik pas bang ng mo pemimpin mentri membangun
2    12,5    jakarta ahok utk cagub anak nya sbg maju terhp miskin u rakyat tp kota org bersih agama dr menghilangorang
3    12,5    yg baswedan tdk gubernur rja mah menteri ga hijabers jgn paham santun klo ja janji majelis ust bnyk islam
```

Gambar 4. Hasil LDA Berupa Topik

2.2.4. Tahapan TF-IDF

Tahapan keempat yang dilakukan ialah menghitung bobot pada setiap kata yang terdapat dalam masing-masing klaster dengan menggunakan perhitungan *TF-IDF (Term Frequency – Inverse Document Frequency)*. Metode *TF-IDF (Term Frequncy Inverse Document Frequency)* merupakan suatu cara untuk memberikan bobot hubungan suatu kata *(term)* terhadap dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu, frekuensi kemunculan sebuah kata didalam sebuah dokumen tertentu dan *inverse* frekuensi dokumen yang mengandung kata tersebut [8]. Untuk melakukan pembobotan kata dengan *TF-IDF*, data hasil *clustering* digunakan sebagai acuan untuk menghitung pembobotan setiap kata pada masing-masing data klaster. Data dipecah menjadi masing-masing klaster, seperti pada data Ahok terdapat 4 klaster data utuh kemudian data dipecah menjadi tiap-tiap klaster sehingga data Ahok memiliki 4 data klister. Selanjutnya menghitung nilai *TF* setiap kata pada masing-masing data klister, serta menghitung nilai *IDF* setiap kata pada masing-masing data klaster. Dan menghitung nilai *TF-IDF* setiap kata pada masing-masing data klaster, serta menghitung nilai rata-rata *TF-IDF* setiap *tweet*. Kemudian mencari *tweet* mana saja yang mengandung kata-kata terbanyak pada *output LDA* untuk dijadikan topik acuan. Berikut rumus *tf-idf* Persamaan 2.

$$p(\theta \mid \alpha) = W = tfxIDF \quad (2)$$

Berikut merupakan hasil *TF, IDF,* rata-rata nilai *TF-IDF* tiap *tweet* dapat dilihat pada Gambar 5, Gambar 6, dan Gambar 7.

```
gsuka : <24> : biarin aja pa anies ya strategy gsuka a detin sigaja repot
kata : gsuka =  TF :1
dripmunafik : <31> : menngan kafir dripmunafik
kata : dripmunafik =  TF :1
japresiden : <38> : ya altuhan yme  negarajau anies japresiden
kata : japresiden =  TF :1
sanyuk : <36> : ren banget tweeps dukung anies sanyuk tweeps jakarta maju
kata : sanyuk =  TF :1
februari : <15> : februari bertarung dgn putaran
kata : februari =  TF :1
```

Gambar 5. Hasil TF Kata

```
nya : <1> : ayo coblos no  no  no nya bro hahaha
nya : <4> : anak muda jakarta nya berkari anak muda pilih
nya : <17> : bp spt nya tdk paham a steve job yg i garase rumah selayaknya jd ttinggal
nya : <47> : pak nya bijakseti mah sensitif loh rasa skriminasi hijabers nkri
nya=4
kata : nya =  IDF :1.0969100130080565
februari : <29> : ia liat  februari aja bro semoga
februari : <31> : februari bertarung dgn putaran ii
februari=2
kata : februari =  IDF :1.3979400086720377
sgt : <19> : pak knp bpk hrs melawan ahok knp gx nunggu  thn lg sy seh krn ahok bpk org yg sy sgt sukai
sgt=1
kata : sgt =  IDF :1.6989700043360187
orgnya : <20> : pak sy caya bpk orgnya baik sopan jujurshrsnya bpk jd wakil presiden bkn mamelawan ahokkrn ahok bagus
orgnya=1
kata : orgnya =  IDF :1.6989700043360187
biah : <46> : biah ity namanya tug cap ng capnya no  cap palsu
biah=1
kata : biah =  IDF :1.6989700043360187
```

Gambar 6. Hasil IDF Kata

pak knp bpk hrs melawan ahok knp gx nunggu thn lg sy seh krn ahok bpk org yg sy sgt sukai: 74.72939154903352
pak sy caya bpk orgnya baik sopan jujurshrsnya bpk jd wakil presiden bkn mamelawan ahokkrn ahok bagus: 56.25372603977551
ente tang jakarta ga jgn ng substansi beda ente yg ngga paham ama omongan org: 47.874440091787854
bp spt nya tdk paham a steve job yg i garase rumah selayaknya jd ttinggal: 46.402352032264055
gue gk interst jd mentri goblok kog malh cagub gayanya orator okopong penkn dah rusak skrg dki rusak: 45.47592205014354
mas jd cagub karna suruhan pepo sgkn mas gagal jd menteri: 44.24675539068267

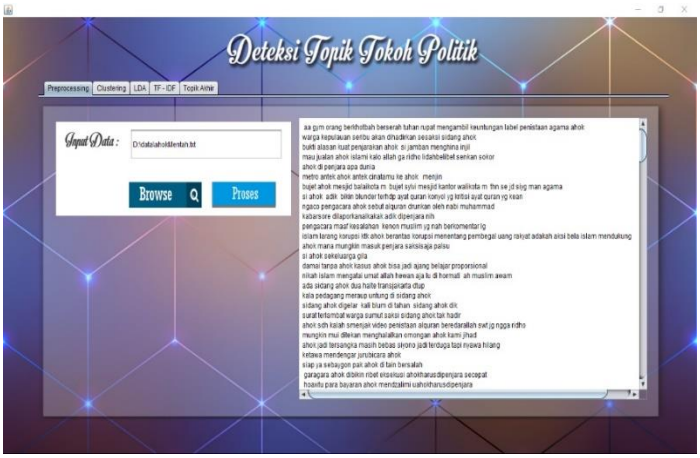
Gambar 7. Hasil TF-IDF Tweet

3. Hasil Penelitian dan Pembahasan

Berikut merupakan hasil dari penelitian *Deteksi Topik Tentang Tokoh Publik Politik Menggunakan Latent Dirichlet Allocation (LDA)* dapat dijeaskan sebagai berikut :

3.1. Preprocessing Data

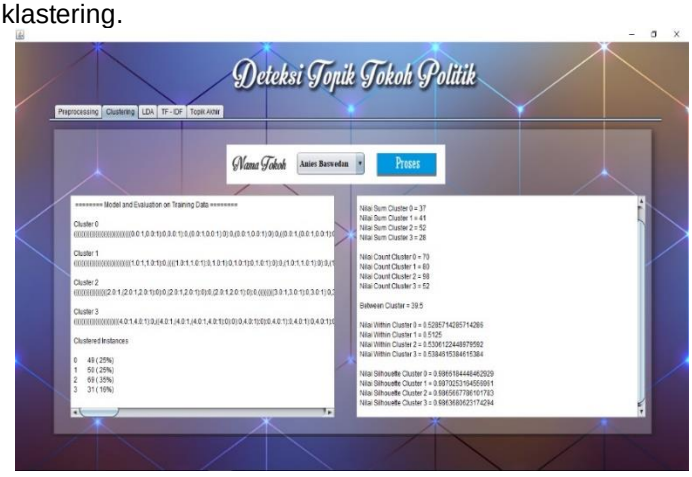
Pada tahap *preprocessing* data dilakukan dengan tujuan untuk menghilangkan kata-kata yang tidak penting yang terdapat dalam dokumen *(Stopword Removal)*, untuk menghilangkan komponen dalam *tweet* yang tidak digunakan seperti *@username, RT, #Hastag, URL (Normalisasi Fitur)*, serta merubah huruf dalam tweet menjadi huruf kecil *(Case Folding)*. Berikut hasil pada tahap *preprocessing* data ditunjukkan pada Gambar 7.



Gambar 8. Hasil Preprocessing Data

3.2. Agglomerative Hierarchical Clustering

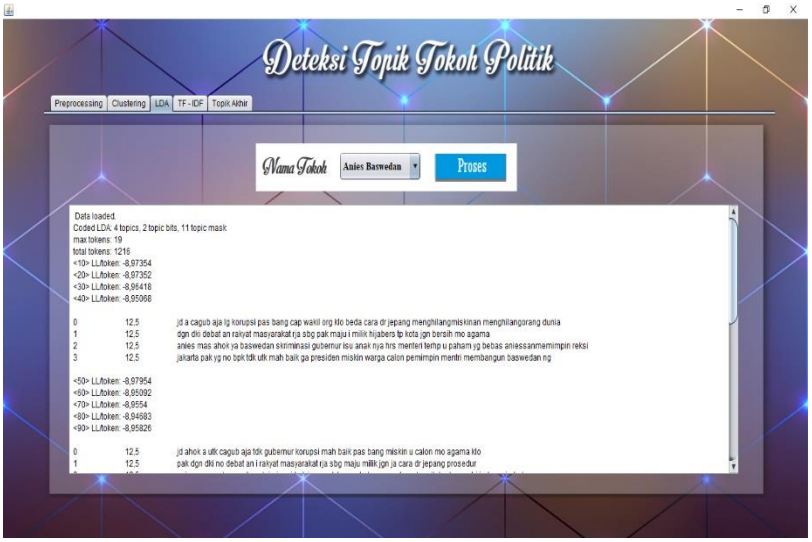
Pada tahap klastering menggunakan *Weka Library* dimana hasil *preprocessing* disimpan dalam format *.csv* yang nantinya akan diubah dalam format *.arff* untuk bisa diolah menggunakan *Weka Library*. Selanjutnya menentukan parameter untuk proses klastering, Gambar 9 berikut contoh parameter klastering.



Gambar 9. Hasil Agglomerative Hierarchical Clustering

3.3. Latent Dirichlet Allocation (LDA)

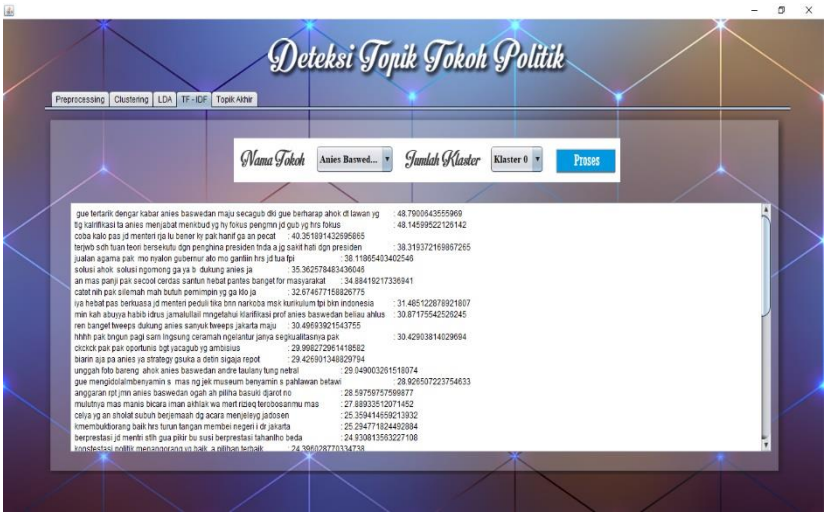
Tahapan selanjutnya yaitu pemodelan topik menggunakan *Latent Dirichlet Allocation (LDA)* dimana pada pemodelan topik ini menggunakan hasil dari *agglomerative clustering* yang digunakan sebagai topik acuan. Pada tahap pemodelan topik *LDA* data yang telah didapat dari hasil klaster dijadikan sebagai data *input*. Ketika data telah di*input*, inisialisasi nilai parameter *LDA* yaitu jumlah topik (didapat dari hasil klaster), nilai α atau *Dirichlet* (nilai α yaitu 50 per topik), nilai β yaitu 0,1, jumlah iterasi 200. Berikut hasil pemodelan topik menggunakan *LDA* ditunjukkan pada Gambar 10.



Gambar 10. Hasil Pemodelan Topik LDA

3.4. TF-IDF

Tahap selanjutnya yaitu *TF-IDF*, tahapan ini dilakukan dengan alasan pada tahap pemodelan topik *LDA* kata pada topik-topik yang dihasilkan tidak mempunyai makna sehingga *user* akan sulit memahami apa arti dari topik-topik tersebut. Dengan menggunakan *TF-IDF* dapat mengetahui *tweet* mana saja yang memiliki kata-kata terbanyak pada hasil topik *LDA*, yang nantinya *tweet* dengan kata-kata terbanyak pada hasil topik *LDA* yang akan digunakan sebagai topik acuan dan dengan tujuan agar *user* dapat memahami dan mengetahui makna dari topik. Berikut hasil *TF-IDF* ditunjukkan pada Gambar 11.



Gambar 11. Hasil TF-IDF

Dan berikut ini merupakan hasil *tweet* yang memiliki kata-kata terbanyak pada hasil topik *LDA* ditunjukkan pada Gambar 12.

Gambar 12. Hasil Topik Akhir

3.5. Silhouette Coefficient

Metode yang digunakan dalam evaluasi hasil klastering yaitu *Silhouette Coeffisient*. Metode *Silhouette Coeffisient* berguna untuk menguji kualitas dari hasil klastering [9]. Pengujian hasil yang dilakukan pada tahapan klastering yaitu menggunakan metode pengujian *Silhouette Index*. Dengan menggunakan metode pengujian tersebut dapat diukur kevalidasian suatu data klaster tunggal ataupun keseluruhan klaster. Berikut hasil dari pengujian klastering menggunakan *Silhouette Index* seperti pada Tabel 3.

Tabel 3. Hasil Silhouette Coefficient Anies Baswedan

Cluster ke-	Nilai Within (a)	Nilai Between (b)	Hasil Silhouette
0	0.5285714285714286	39.5	0.9866184448462929
1	0.5125		0.9870253164556961
2	0.5306122448979592		0.9865667786101783
3	0.5384615384615384		0.9863680623174294

Pada data *Anies Baswedan* didapatkan hasil *Silhouette* klaster-0 dengan nilai 0.9866184448462929, klaster-1 dengan nilai 0.9870253164556961, klaster-2 dengan nilai 0.9865667786101783, klaster-3 dengan nilai 0.9863680623174294. Hasil yang diperoleh pada data *Anies Baswedan* memiliki nilai diatas 0,7 < S*i* <= 1 dengan struktur *silhouette* yang kuat, dengan kata lain data berada pada klaster yang tepat.

3.6. Probability

Pada *LDA* sebuah topik digambarkan sebagai distribusi *probability* dari kata-kata. Artinya setiap topik memiliki probabilitas tertentu pada kata-kata yang ada dalam sebuah dokumen, topik yang berbeda akan memiliki nilai probabilitas yang berbeda pada kata yang sama. Misalnya pada sebuah topik yang menggambarkan tentang *sains dan artikel teknologi* yang lebih banyak menempatkan kemungkinan pada kata *universitas* daripada topik yang menggambarkan olahraga atau politik [10]. Untuk menganalisa topik yang telah dihasilkan pada tahapan *LDA* menggunakan *probability*. Analisa topik ini bertujuan untuk mengetahui probabilitas distribusi kata yang terdapat dalam topik di suatu dokumen. Dokumen yang digunakan dalam analisa topik ini berjumlah lima dokumen, diantaranya dokumen dari kelima tokoh politik yang digunakan dalam tugas akhir ini. Berikut hasil *probability* dapat dilihat pada Tabel 4.

Tabel 4. Hasil Probability Data Anies Baswedan

#doc	name	topic	proportion	...			
0	bentuk	0	0.2636363636363636	3	0.24545454545454545	2	0.24545454545454545
	1	0.24545454545454545					
1	gue	2	0.2627118644067797	3	0.2457627118644068	1	0.2457627118644068
	0	0.2457627118644068					
2	setahun	3	0.2672413793103448	2	0.2672413793103448	0	0.25
	1	0.21551724137931033					

Sebagai contoh pada data *Anies Baswedan* memiliki nilai *probability* sebagai berikut, pada dokumen 0 dengan kata *bentuk* pada topik 0 memiliki nilai *probability* 0.2636363636363636, pada topik 1 memiliki nilai *probability* 0.24545454545454545, pada topik 2 memiliki nilai *probability* 0.24545454545454545, pada topik 3 memiliki nilai *probability* 0.24545454545454545. Dapat dilihat nilai *probability* terbesar didapatkan dengan kata *bentuk* dalam topik 0 itulah yang sesuai sebagai topik 0. Pada dokumen 1 dengan kata *gue* pada topik 0 memiliki nilai *probability* 0.2457627118644068, pada topik 1 memiliki nilai *probability* 0.2457627118644068, pada topik 2 memiliki nilai *probability* 0.2627118644067797, pada topik 3 memiliki nilai *probability* 0.2457627118644068. Dapat dilihat nilai *probability* terbesar didapatkan dengan kata *gue* dalam topik 2 itulah yang sesuai sebagai topik 2. Pada dokumen 2 dengan kata *setahun* pada topik 0 memiliki nilai *probability* 0.25, pada topik 1 memiliki nilai *probability* 0.21551724137931033, pada topik 2 memiliki nilai *probability* 0.2672413793103448, pada topik 3 memiliki nilai *probability* 0.2672413793103448. Dapat dilihat nilai *probability* terbesar didapatkan dengan kata *gue* dalam topik 2 itulah yang sesuai sebagai topik 2. Jika nilai *probability* pada masing-masing topik terdapat nilai *probability* yang paling besar dari nilai *probability* topik yang lainnya, maka nilai *probability* terbesar itulah yang sesuai dengan topik tersebut seperti penjelasan diatas.

4. Kesimpulan

Pada penelitian ini hasil evaluasi klastering sudah berada di struktur kuat yaitu pada rentang antara 0,7 < S*i* <= 1. Hal tersebut telah membuktikan bahwa anggota-anggota dalam masing-masing klaser sudah berada dalam klaster yang tepat. Hasil pemodelan topik pada *LDA* yang memiliki nilai *probability* yang lebih tinggi dari yang lainnya menunjukkan bahwa kata dalam sebuah dokumen tersebut merupakan kata yang tepat pada topik. Dengan hasil evaluasi pengujian *silhouette coefficient* pada tahap klaster, menunjukkan bahwa algoritma *agglomerative hierarchical clustering* merupakan algoritma yang sesuai dan bagus untuk digunakan pada tugas akhir ini. Pada tahapan pemodelan topik *LDA* memiliki *output* topik-topik mengenai tokoh politik. Topik-topik tersebut berisi kata-kata dan tidak ada makna, sehingga *user* akan sulit memahami topik-topik tersebut. Oleh karena itu, pada tugas akhir ini menggunakan metode *TF-IDF* untuk mengetahui nilai rata-rata setiap *tweet* dan mengetahui *tweet* mana saja yang mengandung kata-kata terbanyak dalam *output LDA*. Dengan menggunakan metode *TF-IDF*, *user* akan lebih mudah mengetahui makna dari *output* yang dihasilkan pada tahapan *LDA*.

Daftar Notasi

Berikut daftar notasi yang terdapat pada jurnal sebagai berikut :

D : dokumen
DF : dokumen frekuensi
W : tf-idf
θ : distribusi topik dalam dokumen
α : parameter Dirichlet sebelumnya pada distribusi topik per dokumen
k : jumlah topik

Referensi

[1] Jain AP. Sentiments Analysis Of Twitter Data Using Data Mining. 2015;807–10.
[2] Alashri S, Ravi R, Smith KL, Desouza KC. An Analysis of Sentiments on Facebook during the 2016 U.S. Presidential Election. 2016;795–802.
[3] Tembhurnikar SD, Patil NN. Topic Detection using BNgram Method and Sentiment Analysis on Twitter Dataset. 2015;
[4] Alfina T, Santosa B, Barakbah R. Analisa Perbandingan Metode Hierarchical Clustering , K-means dan Gabungan Keduanya dalam Cluster Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS). 2012;1.
[5] Februariyanti H. Klastering Dokumen Menggunakan Hierarchical Agglomerative Clustering. 2009;
[6] PTIIK-UB. Pengenalan Pola Hierarchical Clustering. 2014;
[7] Agustina A. Analisis dan Visualisasi Suara Pelanggan Pada Pusat Layanan Pelanggan Dengan Pemodelan Topik Menggunakan Latent Dirichlet Allocation (LDA) Studi Kasus: PT. PETROKIMIA Gresik. 2017;
[8] Putung KD, Lumenta A, Jacobus A, Informatika T, Sam U, Manado R. Penerapan Sistem Temu Kembali Informasi Pada Kumpulan Dokumen SkripsI. 2016;8(1).

[9] Kurniawan AA. Implementasi Algoritma Agglomerative Hierarchical Clustering Untuk Mengelompokkan Capaian Belajar Siswa SD. 2017;
[10] Hansen J V. Inside Latent Dirichlet Allocation : An Empirical Exploration Inside Latent Dirichlet Allocation : An Empirical Exploration.

ISSN: 2714-7975; E-ISSN: 2716-1382