TOPIC MODELING DATA *TWITTER* TERHADAP CALON PRESIDEN REPUBLIK INDONESIA 2019 MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION (LDA)

Ahmad Fathan Hidayatullah, Anisa Miladya Hakim, Siwi Cahyaningtyas, Widya Puteri Aulia
Department of Informatics
Universitas Islam Indonesia, UII
Yogyakarta, Indonesia
e-mail: fathan@uii.ac.id, {15523198, 15523223, 15523235} @students.uii.ac.id

Abstract—Tahun 2019 menjadi tahun yang panas politik karena menjadi waktu pelaksanaan Pemilihan Umum secara serentak, mulai dari Pemilihan Legislator (Pileg) sampai Pemilihan Presiden (Pilpres). Menjelang pemilu, nama-nama calon mulai diperbincangkan. Terlebih nama-nama yang masuk ke dalam calon Presdien merupakan nama-nama yang juga masuk ke dalam Pilpres pada pemilu tahun 2015. Calon-calon tersebut ialah Joko Widodo dan Prabowo Subianto. Tingginya jumlah pengguna media sosial yakni sebesar 132,7 dan 27% di antaranya merupakan pengguna twitter. Banyaknya pengguna twitter inilah yang kemudian dijadikan sebagai salah satu alat kampanye bagi partai politik. Hal tersebut dapat dilihat dari mulai banyaknya akun yang membicarakan mengenai sosok calon Presiden RI yang seringkali masuk ke dalam topik yang sering dibicarakan di twitter. Untuk melihat topik apa saja yang dibicarakan mengenai kedua calon Presiden Republik Indonesia 2019-2024, digunakan metode LDA (Latent Dirichlet Allocation) dan word2vec untuk mengetahui seberapa erat keterikatan calon Presiden RI dengan beberapa topik yang telah ditemukan. Dari 1330 data tweet mentah yang membicarakan Joko Widodo, dan 1461 untuk Prabowo Subianto, dihasilkan data bersih 1307 untuk Joko Widodo, dan 1274 untuk Prabowo Subianto. Hasil dari penelitian ini dinilai cukup bagus, terlebih ketika dibandingkan dengan metode LSI.

Kata kunci : Calon Presiden, Joko Widodo, Prabowo Subianto, LDA

I. Pendahuluan

Pemilihan umum adalah salah satu metode yang digunakan untuk memilih orang-orang yang akan mengisi jabatan penting di pemerintahan, mulai dari kepala daerah, anggota legislatif hingga kepala negara. Pemilihan suara dipilih langsung oleh masyarakat, mengingat negara Indonesia merupakan negara yang menganut asa demokrasi.

Tahun 2019 menjadi tahun yang panas politik karena menjadi waktu pelaksanaan Pemilihan Umum secara serentak, mulai dari Pemilihan Legislator (Pileg) sampai Pemilihan Presiden (Pilpres). Menjelang pemilu, nama-nama calon mulai diperbincangkan. Terlebih nama-nama yang masuk ke dalam calon Presdien merupakan nama-nama yang juga masuk ke dalam Pilpres pada pemilu tahun 2015. Calon-calon tersebut ialah Joko Widodo dan Prabowo Subianto.

Hal ini kemudian menimbulkan berbagai persepsi di masyarakat. tak bisa dihindari bahwa media, baik itu media telekomunikasi bahkan media sosial sangat berpengaruh dalam mempromosikan salah satu calon dan menjatuhkan calon yang lain. Sehingga, muncul tren baru yang pada akhirnya menjadikan media sosial sebagai salah satu alat untuk melakukan kampanye.

Berdasarkan data yang didapat dari website resmi Kementrian Kominfo, (https://inet.detik.com/cyberlife/d-3912429/130-juta-orang-indonesia-tercatat-aktif-di-medsos) dapat diketahui bahwa hingga Januari 2018, pengguna sosial media mnecapai setengah dari jumlah warga negara Indonesia, yakni sebesar 132,7 juta jiwa dan 27% dari pengguna sosial media ini, adalah pengguna *twitter*.

Banyaknya pengguna *twitter* inilah yang kemudian dijadikan sebagai salah satu alat kampanye bagi partai politik. Terlebih di tahun pergantian presiden pada periode 2019-2024 ini. Hal tersebut dapat dilihat dari mulai banyaknya akun yang membicarakan mengenai sosok calon Presiden RI, yaitu Joko Widodo dan Prabowo Subianto. Munculnya berbagai akun yang membicarakan kedua sosok tersebut, membuat kedua calon tersebut beberapa kali masuk ke dalam *trending topics.* Di antaranya ialah #JokowiBanyakCinta dan #CeritaPrabowo.

Dengan mengetahui topik apa saja yang sering dibicarakan mengenai calon Presiden Republik Indonesia periode 2019-2024, dapat menjadi gambaran bagi masyarakat terhadap setiap calon. Dalam menentukan topik apa saja yang dibahas untuk *tweet* kedua calon tersebut, digunakan metode LDA (*Latent Dirichlet Allocation*. Selain itu, penulis juga mengimplementasikan *word2vec* untuk mengetahui seberapa erat keterikatan calon Presiden RI dengan beberapa topik yang telah ditemukan.

II. Penelitian Terkait

Hidayatullah dkk.(2017) telah melakukan penelitian mengenai topik-topik yang sering

dibahas terkait sepakbola di Indonesia sepanjang tahun 2017. Penelitian ini menggunakan metode LDA *(Latent Dirichlet Allocation).* Data ini menggunakan data twitter dari akun *official* sepak bola. Berdasarkan penelitian ini, dapat dilihat bahwa metode LDA cukup baik dalam menemukan topik dari data twitter sepak bola. Penelitian ini juga menyarankan untuk menghasilkan topik yang lebih relevan dengan pemahaman manusia perlu adanya pencarian jumlah topik yang disesuaikan dengan kondisi data yang dimiliki.

Penelitian selanjutnya masih dilakukan oleh Hidayatullah dkk.(2017) melakukan penelitian bertujuan untuk mendapatkan informasi tentang peristiwa-peristiwa penting yang terjadi pada lalu lintas di pulau Jawa, Indonesia. Data yang diperoleh melalui akun *official Traffic Management Centre* (TMC) di twitter. Dalam pencarian topik perlu dicari jumlah topik yang sesuai, penelitian ini membandingkan antara nilai perplexity dengan nilai *perplexity* perkata. Hasil yang diperoleh dari penelitian ini dengan menggunakan metode LDA adalah bahwa dapat diketahui bahwa metode LDA sudah dapat menemukan peristiwa penting dari data yang akun *official* TMC. Namun hasil yang didapatkan untuk setiap segmen tidak selalu sempurna. Hal ini dikarenakan masih adanya topik yang sama untuk segmen yang berbeda. Sehingga, penelitian ini menyarankan untuk menggunakan metode lain untuk dalam menentukan jumlah topik yang dibahas.

Penelitian selanjutnya dilakukan oleh Henning M dkk.(2016). Penelitian ini meneliti tentang cara untuk mendeteksi topik pada tweet berita terbaru menggunakan metode LDA dan kemudian dibandingkan menggunakan metode HDP. Hasil dari penelitian menunjukkan bahwa metode LDA menghasilkan ketelitian yang lebih unggul dibandingkan dengan HDP namun, untuk hasil dari penelitian ini menunjukkan bahwa berita terbaru tidak dapat dilakukan karena karakteristik kata yang pendek dan kata yang ambigu.

Sebuah penelitian yang dilakukan oleh Dong dkk.(2005) mendeteksi topik yang dibicarakan di Microsoft MSN message. Penelitian ini menggunakan metode *Naive Bayes, associative classification*, dan *support vector machines.* Penelitian ini juga menerapkan preprocessing untuk mengekstrak pesan yang berisi icon dan pesan yang berisi URL. Dari penelitian ini, ditemukan bahwa metode SVM menghasilkan keakuratan yang lebih tinggi dibandingkan metode NB dan AC.

Berdasarkan penelitian yang ada penelitian ini mencoba untuk melakukan pencarian topik menggunakan metode LDA, dan nantinya akan dibandingkan dengan metode LSI terhadap data twitter terkait calon presiden nomor urut 1 dan calon presiden nomor urut 2. Selain itu, pada penelitian ini diimplementasikan pula *word2vec* untuk menentukan seberapa erat keterikatan seorang calon presiden terhadap topik tertentu.

III. Metodologi

Pada Gambar 1 memperlihatkan alur dari penelitian yang akan dilakukan.

Gambar 1 Alur Penelitian

A. *Data Retrieval*

Data yang dikumpulkan berasal dari *Twitter API* dengan memanfaatkan *library tweepy* untuk pengaksesan data *twitter*. Data yang digunakan dalam penelitian ini terdiri atas 2 data, yakni berupa *tweet* terkait Joko Widodo (Calon Presiden No.1) dan Prabowo Subianto (Calon Presiden No.2). Jumlah dari masing-masing data tersebut sebanyak 1330 data *tweet*.

B. *Preprocessing*

Gambar 2 *Preprocessing*

Preprocessing merupakan tahapan awal untuk memperoleh data yang terstruktur (Socrates, dkk, 2016). Pada penelitian ini, *preprocessing* yang dilakukan antara lain adalah menghapus URL, menghapus karakter khusus *Twitter* (seperti : #hashtag, RT, cc, @username/mention), mengganti kata tidak baku (*slangword)*, menghapus *stopword*, menghapus non-ASCII, menghapus *white space*, menghapus *punctuation*, mengahapus digit, menghapus karakter berulang sebanyak lebih dari 2 kali, menghapus kalimat yang mengandung 1 kata saja, dan *case*

folding. Gambar 2 memperlihatkan proses secara urut dalam melakukan *preprocessing.*

C. *Feature Extraction*

Proses pemilihan dan ekstraksi fitur yang digunakan dalam penelitian ini adalah *N-gram* dan TF-IDF (*Term Frequency-Inverse Document Frequency)*. Metode *N-gram* yang digunakan adalah *Bi-gram* dan *Tri-gram* dengan tujuan untuk mengidentifikasi frasa dari data *tweet*. Pembobotan TF-IDF adalah pembobotan terhadap setiap kata pada dokumen. Tujuan dari pembobotan TF-IDF pada penelitian ini adalah untuk meningkatkan keakuratan dan ketelitian hasil dari *topic modeling*.

D. *Topic Modeling* dengan LDA

LDA merupakan algoritma sederhana untuk *topic modeling* (Blei, dkk, 2003). Tujuan dari LDA adalah untuk menemukan topik dari kumpulan dokumen, dimana masing-masing topik merupakan distribusi kata-kata. Proses distribusi kata-kata dilakukan untuk menentukan kata tersebut berasal dari topik yang sama atau tidak. Proses inilah yang dinamakan distribusi *dirichlet*. Dalam LDA, dokumen merupakan objek yang dapat diamati, sedangkan distribusi topik per-dokumen merupakan struktur tersembunyi, maka oleh karena itu metode ini dinamakan *Latent Dirichlet Allocation* (LDA) (Blei, 2012).

E. *Visualization*

Dalam penelitian ini, kami menggunakan *library wordcloud* sebagai visualisasi hasil dari *topic modeling*.

IV. Hasil dan Pembahasan

A. *Topic Analysis*

Visualisasi data menggunakan *word cloud*. *Word cloud* adalah sebuah visualisasi yang menyusun kata-kata yang utama dalam sebuah text (Hidayatullah, 2017). Dimana, ukuran dari masing-masing kata ini mengindikasikan seberapa sering kata tersebut muncul dalam sebuah topik. Kata yang mendominasi dalam sebuah *word cloud*, menunjukkan hubungannya terhadap topik.

- Topic LDA-JOKOWI-#0 :Presiden Dicintai Rakyat Indonesia

  Topic LDA-JOKOWI-#0 membahas tentang jokowi adalah presiden yang dicintai rakyatnya. Hal ini dikarenakan beberapa kata yang mendominasi seperti ‘presiden’, ‘dicintai’, ‘rakyat’,’Indonesia’. Sehingga dapat disimpulkan Topic LDA-JOKOWI-#0 berbicara tentang Presiden Dicintai Rakyat Indonesia

Gambar 3. *Word Cloud* Topic LDA-JOKOWI-#0

- Topic LDA-JOKOWI-#1 : Optimis Jokowi Menjadi Pemimpin

  Topic LDA-JOKOWI-#1 membahas tentang jokowi adalah Optimis Jokowi Menjadi Pemimpin. Hal ini dikarenakan beberapa kata yang mendominasi seperti ‘Optimis’, ‘Pemimpin’, ‘percaya’,’pemimpin_opti mis’,’rakyat’. Sehingga dapat disimpulkan Topic LDA-JOKOWI-#1 berbicara tentang Optimis Jokowi Menjadi Pemimpin

Gambar 4. *Word Cloud* Topic LDA-JOKOWI-#1

- Topic LDA-JOKOWI-#2 : Memimpin dengan Penuh Cinta

  Topic LDA-JOKOWI-#2 membahas tentang jokowi adalah Memimpin dengan Penuh Cinta. Hal ini dikarenakan beberapa kata yang mendominasi seperti ‘penuh_cinta’, ‘cinta’, ’rakyat’. Sehingga dapat disimpulkan Topic LDA-JOKOWI-#2 berbicara tentang Memimpin dengan Penuh Cinta.

Gambar 4. *Word Cloud* Topic LDA-JOKOWI-#2

- Topic LDA-JOKOWI-#3 : Semangat Dihati

  Topic LDA-JOKOWI-#3 membahas tentang Semangat Dihati. Hal ini dikarenakan beberapa kata yang mendominasi seperti ‘semangat’, ‘dihati’. Sehingga dapat disimpulkan Topic LDA-JOKOWI-#3 berbicara tentang Semangat Dihati.

Gambar 5. *Word Cloud* Topic LDA-JOKOWI-#3

- Topic LDA-JOKOWI-#4 :Isu Bohong

  Topic LDA-JOKOWI-#4 membahas tentang Isu Bohong. Hal ini dikarenakan beberapa kata yang mendominasi seperti ‘isu_bohong’, ‘bohong’. Sehingga dapat disimpulkan Topic LDA-JOKOWI-#4 berbicara tentang Isu Bohong.

Gambar 5. *Word Cloud* Topic LDA-JOKOWI-#4

- Topic LDA-PRABOWO-#0 :Isu Bohong

  Topic LDA-PRABOWO-#0 membahas tentang Isu Bohong. Hal ini dikarenakan beberapa kata yang mendominasi seperti ‘isu_bohong’, ‘bohong’. Sehingga dapat disimpulkan Topic LDA-PRABOWO-#0 berbicara tentang Isu Bohong.

Gambar 6. *Word Cloud* Topic LDA-PRABOWO-#0

- Topic LDA-PRABOWO-#1 : Prabowo-Sandi Pemimpin Indonesia

  Topic LDA-PRABOWO-#1 membahas tentang Prabowo-Sandi Pemimpin Indonesia. Hal ini dikarenakan beberapa kata yang mendominasi seperti ‘prabowo-sandi’, ‘pemimpin’,’menang’. Sehingga dapat disimpulkan Topic LDA-PRABOWO-#1 berbicara tentang Prabowo-Sandi Pemimpin Indonesia

  .

Gambar 7. *Word Cloud* Topic LDA-PRABOWO-#1

- Topic LDA-PRABOWO-#2 : Pemilihan presiden

  Topic LDA-PRABOWO-#2 membahas tentang Pemilihan presiden. Hal ini dikarenakan beberapa kata yang mendominasi seperti ‘pemilihan_presiden’, ‘presiden’. Sehingga dapat disimpulkan Topic LDA-PRABOWO-#2 berbicara tentang Pemilihan presiden.

Gambar 8. *Word Cloud* Topic LDA-PRABOWO-#2

word2vec Jokowi

```
w1 = ["jokowi"]
model_j.wv.most_similar (positive=w1,topn=10)
```

```
C:\Users\ANNISA\Anaconda3\lib\site-packages\gensim\matutils.py:737: FutureWarning: Conversion of the second argument of issubdt
ype from `int` to `np.signedinteger` is deprecated. In future, it will be treated as `np.int32 == np.dtype(int).type`.
  if np.issubdtype(vec.dtype, np.int):
```

```
[('masyarakat', 0.999805212020874),
 ('kepemimpinan', 0.9997609260711167),
 ('maruf_amin', 0.9997678995132446),
 ('warga', 0.9997577667236328),
 ('presiden', 0.9997571110725403),
 ('fakta', 0.9997552037239075),
 ('pns', 0.9997391700744629),
 ('menteri', 0.9997365474700928),
 ('pemerintah', 0.9997317700985107),
 ('pkh', 0.9997305274009705)]
```

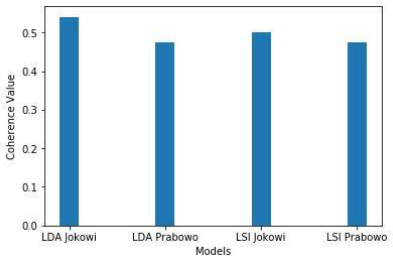
word2vec Prabowo

```
w1 = ["prabowo"]
model_p.wv.most_similar (positive=w1,topn=10)
```

```
2019-01-04 11:46:18,449 : INFO : precomputing L2-norms of word weight vectors
C:\Users\ANNISA\Anaconda3\lib\site-packages\gensim\matutils.py:737: FutureWarning: Conversion of the second argument of issubdt
ype from `int` to `np.signedinteger` is deprecated. In future, it will be treated as `np.int32 == np.dtype(int).type`.
  if np.issubdtype(vec.dtype, np.int):
```

```
[('nkri', 0.9993458986282349),
 ('program', 0.9993394613265991),
 ('keluarga', 0.9993193745613098),
 ('semangat', 0.9993059635162354),
 ('indonesia', 0.9993013143539429),
 ('pemerintahan', 0.9993000567260742),
 ('presiden', 0.9993001222610474),
 ('daerah', 0.9992990493774414),
 ('fakta', 0.9992954730987549),
 ('sosok', 0.9992837905883789)]
```

B. Perbandingan Metode



Berdasarkan grafik diatas diperoleh bahwa metode LDA Jokowi memiliki nilai *coherence* yang lebih tinggi dibandingkan LSI Jokowi, begitupun dengan nilai *coherence* LDA Prabowo dengan LSI Prabowo. Hal ini menunjukkan

bahwa algoritma LDA lebih baik dibandingkan dengan LSI dalam melakukan *topic modeling*.

V. Kesimpulan dan Saran

Penelitian ini telah berhasil membangun model untuk melakukan *topic modeling* menggunakan metode LDA terhadap data *twitter* tentang calon presiden nomor urut 1 dan 2. Metode LDA yang digunakan dinilai sudah cukup menampilkan hasil yang bagus, di mana untuk topik pada Jokowi dihasilkan 5 topik, dan untuk Prabowo 3 topik. Selain itu telah berhasil juga menggunakan *word2vec* untuk mencari tahu seberapa dekat calon presiden terhadap data yang dihasilkan. Untuk penelitian selanjutnya perlu dicoba dikembangkan juga *Topic Modelling* untuk data twitter dari calon wakil presiden. Selain itu melakukan *Text Classification* untuk mengetahui sentimen terhadap calon presiden dan wakil presiden.

REFERENSI

Blei, D. M. (2012). Probabilistic Topic Models.

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., & Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Dong, Haichao, Hui, Siu Cheung, He, Yulan. *Structural Analysis of Chat Message for Topic Detection.*

Wold, Henning.M, Vikre, Linn, Gulla, Jon Atle, Ozgobek, Ozlem, Su, Xiaomeng.(2016). *Topic Modeling for Breaking News Detection.*

Hidayatullah, A. F., Kurniawan, W., Pembrani, E. C., Akbar, G., & Pranata, R. (2018). Twitter Topic Modeling on Football News. *Computer Science*.

Hidayatullah, A. F., & Maarif, M. R. (2017). Road Traffic Topic Modeling on Twitter using Latent Dirichlet Allocation, 47–52.