

# Pemodelan Topik Latent Dirichlet Allocation Buat Informasi Saham

*(Latent Dirichlet Allocation Topic Modeling for Stock Information)*

**Ekka Pujo Ariesanto Akhmad<sup>1</sup>, Carlos Lazaro Prawirosastro<sup>2</sup>, Budi Priyono<sup>3</sup>**

**<sup>1,2,3</sup> Program Studi Manajemen Pelabuhan dan Logistik Maritim,  
Fakultas Vokasi Pelayaran, Universitas Hang Tuah**

**Abstrak:** Penanam modal mendapatkan data ataupun informasi saham melalui halaman web Bursa Efek Indonesia. Penanam modal kadangkala memperoleh data lain mengenai analisa saham serta perkiraan saham yang profitabel dari laman website informasi online. Tetapi, penanam modal membutuhkan durasi buat memastikan topik yang kerap timbul serta jadi pembicaraan hangat pada informasi saham. Oleh sebab itu, pemodelan topik dibutuhkan buat inti informasi saham online. Tujuan riset ini merupakan pemodelan topik dicoba buat menganalisa topik- topik yang lagi diulas pada laman website informasi saham online. Informasi riset digabungkan sebesar 181 informasi saham sepanjang bulan Februari sampai Juli 2021 dari web kontan.co.id. Pemodelan topik dicoba dengan tata cara Latent Dirichlet Allocation. Pemodelan topik menciptakan 3 kelompok penting serta penyaluran 10 kata pembuat masing- masing topik dari informasi saham online. Penggambaran pemodelan topik menciptakan catatan tiga puluh kata yang sering muncul dari masing- masing topik.

**Kata kunci:** pemodelan topik, latent dirichlet allocation

**Abstract:** Investors obtain data or share information through the Indonesia Stock Exchange web page. From online information portals, investors occasionally obtain more knowledge about stock analysis and profitable stock projections. However, it takes time for investors to identify subjects that frequently come up and become popular stock information themes. The core of online stock information therefore requires subject modeling. Topic modeling, which seeks to look at the issues being discussed on a page of an internet stock information website, is the focus of this study. Research information is combined with 181 stock information February to July 2021, web source: kontan.co.id. Utilizing the Latent Dirichlet Allocation approach, topic modeling was tested. The topic modeling created 3 important groups and the distribution of the 10 creator words of each topic from online stock information. The topic modeling description creates a thirty-word note that occurs frequently from each topic.

**Keywords:** topic modeling, latent dirichlet allocation

## Alamat Korespondensi:

Ekka Pujo Ariesanto Akhmad, Program Studi Manajemen Pelabuhan dan Logistik Maritim, Fakultas Vokasi Pelayaran, Universitas Hang Tuah, Jalan A. R. Hakim 150, Surabaya. e-mail: eka.pujo@hangtuah.ac.id

## PENDAHULUAN

Pergerakan harga di pasar saham berarti untuk penanam modal. Informasi tekstual menjadi pangkal data tentang saham. Biasanya penanam modal mendapatkan data ataupun informasi saham melalui [www.idx.co.id](http://www.idx.co.id). Kadangkala, penanam modal memperoleh data lain mengenai analisa saham serta perkiraan saham dari laman web informasi online. Informasi tekstual berhubungan dengan informasi pasar saham merupakan tata cara yang terkenal buat menekuni sikap harga.

Tetapi, penanam modal membutuhkan tempo buat memastikan topik yang kerap timbul serta jadi pembicaraan hangat mengenai saham. Sebab itu, riset ini hendak melaksanakan pemodelan topik informasi mengenai saham yang diterbitkan oleh laman website informasi online. Pemodelan topik ialah bentuk statistik buat memastikan bagian utama topik dari sekumpulan naskah (Bansal, 2020; Tong, 2016).

Pemodelan topik dipakai buat mengenali gaya topik saham yang jadi

atensi warga, sampai data saham jadi lebih singkat. Tata cara pemodelan topik yang sempat diusulkan antara lain, pengelompokan bacaan pustaka dengan memakai tata cara Support Vector Machine (SVM) (Kanan serta Fox, 2016) serta TF/IDF buat inti fitur (Dadgar, dkk., 2016). SVM bisa membongkar permasalahan overfitting, lama keadaan menuju satu titik temu, serta sekurang-kurangnya data pelatihan. Namun SVM mempunyai kelemahan pada sulitnya penentuan patokan SVM yang maksimal.

Tata cara Latent Dirichlet Allocation (LDA) digagas Blei buat pemodelan topik (Calheiros, 2017). Algoritma LDA ialah algoritma pemodelan topik dengan bentuk peluang generatif kumpulan naskah. Tujuannya supaya pemrosesan naskah dalam koleksi informasi besar jadi berdaya guna. Tata cara ini membuat naskah menciptakan luaran berbentuk bermacam tipe topik yang berlainan, alhasil tidak dengan cara khusus membagi naskah ke dalam suatu topik tersendiri. Eksploitasi Latent Dirichlet Allocation juga banyak diterapkan pada bidang lain, yaitu klasifikasi sentimen dokumen twitter (Usop, dkk., 2017), analisis trending topic twitter (Halima dan Chitrakala, 2016), analisis sentimen komentar mahasiswa terhadap universitas (Bashri, 2017), analisis sentimen industri pariwisata di Indonesia (Putri, 2017).

Penjelasan kerangka balik permasalahan itu bisa diturunkan dalil permasalahan yang hendak dicari tanggapannya lewat riset ini, ialah Bagaimana melaksanakan pemodelan topik buat menganalisa topik-topik yang lagi diulas pada laman website informasi saham online?

Tujuan riset selanjutnya ini ialah penjelasan hasil yang hendak digapai lewat riset, ialah melaksanakan

pemodelan topik buat menganalisa topik-topik yang lagi diulas pada laman website informasi saham online.

Penelitian terdahulu dilakukan beberapa peneliti, yaitu

Dalam pengelompokan opini publik, model LDA menghasilkan topik pembahasan umum tentang angkutan umum yang berbeda-beda untuk setiap jenis angkutan umum. Modelnya juga bisa menangkap komunikasi publik dengan akun resmi administrasi transportasi umum. Error model prediksi yang didapat paling rendah dengan rata-rata Persentase Error 8,79%. Dengan menambahkan variabel sentimen publik dapat meningkatkan akurasi model prediksi, tetapi variabel sentimen tweet tidak memiliki hubungan langsung dengan variabel jumlah pengguna angkutan umum (Rachman, dkk., 2021).

Lucini, dkk., (2020) menggunakan metode penambangan teks, mereka menjelajahi Online Customer Review (OCR) untuk memberikan pedoman bagi perusahaan penerbangan meningkatkan daya saing. Mereka menganalisis database lebih dari 55.000 OCR, mencakup lebih dari 400 maskapai penerbangan dan penumpang dari 170 negara. Menggunakan sebuah Model Alokasi Dirichlet Latent (LDA), mereka mengidentifikasi 27 dimensi kepuasan yang dijelaskan oleh 882 kata sifat. Ukuran dan kata sifat digunakan untuk memprediksi rekomendasi maskapai oleh pelanggan, menghasilkan akurasi 79,95%.

Setijohatmo, Urip T., dkk. (2020) memakai ekspansi PLSA asal rancangan lain yang dikatakan LDA, spesifiknya memakai prosedur pemecahan pengambilan sampel Gibbs, serta dicoba di riset pencarian naskah skripsi. Penelitian memakai kelompok skripsi yang telah dikasih identitas.

Berikutnya yang akan terjadi penelitian hendak ditakar strata kepentingannya, bila dibanding evaluasi orang pada wujud skripsi beridentitas.

Penelitian pemodelan topik merumuskan bila banyak topik terdapat pada catatan sosial media adalah empat topik. Luaran penelitian ini sudah dicoba dengan angka perplexity terbagus adalah 213, 41 serta dicoba kesederhanaannya buat diartikan orang lewat percobaan koherensi topik, yaitu Tugas Intrusi Kata serta Tugas Intrusi Topik. Ikhtisar dari percobaan keterkaitan topik melaporkan kalau bentuk yang diperoleh lewat tata cara LDA pada riset permasalahan ini bisa diterjemahkan orang dengan bagus. (Putra, I Made Kusnanta Bramantya, 2017).

Kusumaningrum, dkk., (2016) menerapkan salah satu metode Topic Modeling yang mewakili dokumen sebagai distribusi topik dan topik adalah diwakili oleh sekumpulan kata. Metodenya adalah Dirichlet Laten Alokasi. Studi eksperimental berdasarkan 10-fold cross validation dilakukan dengan menggunakan beberapa parameter mencakup sejumlah topik (5, 10, dan 15) dan kedua, LDA hyperparameter (0,001, 0,01, dan 0,1). Hasilnya membuktikan kalau ketepatan totalitas terbaik kurang lebih 70% untuk mengklasifikasikan dokumen Artikel berita Indonesia menjadi 5 kelas, yaitu ekonomi, pariwisata, kriminal, olahraga, dan politik.

Kebaruan riset ini membuktikan ukuran topik dan memvisualkan luaran topik, alhasil mempermudah pemakai mengidentifikasi edaran tutur serta frasa di masing- masing topik. Aturan LDA menangani permasalahan yang berhubungan dengan pemodelan topik. Aturan LDA jadi jalan keluar menangani pengelompokan term jadi topik khusus sambil mengindahkan

antrean tutur pada cara pembentukkannya lewat metode bentuk campuran.

### **Pemodelan Topik**

Dalam machine learning serta pemrosesan bahasa natural, bentuk topik merupakan bentuk generatif, yang membagikan kerangka probabilistik( Grun serta Hopnik, 2011). Tata cara pemodelan topik biasanya dipakai buat menata, menguasai, mencari, serta meringkas arsip elektronik berdimensi besar dengan cara otomatis. Topik menunjukkan jalinan peubah yang tersembunyi, diperkirakan, mengaitkan kata-kata dalam suatu kosakata serta kemunculannya dalam naskah. Suatu naskah dianggap selaku kombinasi topik. Tema bentuk menciptakan tema tersembunyi di semua koleksi serta membubuhi penjelasan naskah cocok dengan tema- tema itu. Tiap tutur didapat dari salah satu topik itu. Suatu penyaluran jangkauan naskah topik diperoleh serta sediakan metode terkini buat menjelajahi informasi mengenai perspektif topik.

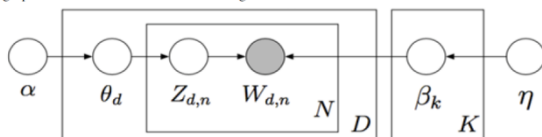
### **Latent Dirichlet Allocation**

Latent Dirichlet Allocation( LDA) merupakan bentuk generatif yang membolehkan berkas observasi dipaparkan oleh golongan yang tidak teramati yang menerangkan kenapa sebagian informasi seragam (Jing, 2015).

Secara intuitif dalam LDA, dokumen menunjukkan banyak topik (Blei, 2012). Dalam pra-pemrosesan teks, tanda baca dan kata henti (seperti, "jika", "itu", atau "pada", yang mengandung sedikit konten topikal) dikecualikan.

Oleh karena itu, tiap naskah dikira selaku kombinasi topik semua korpus. Suatu topik merupakan penyaluran atas kosa kata yang pasti. Topik- topik ini diperoleh dari berkas naskah (Blei,

2007; Yeh, 2016). Misalnya, topik olahraga mempunyai kata "sepak bola", "basket" dengan peluang tinggi dan topik komputer mempunyai kata "data", "jaringan" dengan kemungkinan tinggi. Kemudian, kelompok naskah memiliki penyebaran peluang atas topik, di mana setiap kata diambil dari salah satu dari topik itu. Dengan penyebaran kemungkinan naskah untuk setiap topik, peneliti akan tahu berapa banyak topik terlibat dalam sebuah naskah, artinya topik mana yang terutama diperbincangkan oleh naskah. Tampilan bentuk grafis LDA dapat dilihat pada Gambar 1.



Sumber: Tong (2016)

**Gambar 1. Bentuk grafis untuk LDA**

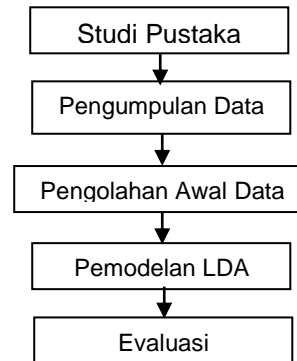
Serupa yang dipaparkan pada Gambar 1, LDA ditafsirkan dengan cara lebih resmi dengan notasi berikut. Awal,  $\alpha$  serta  $\eta$  merupakan patokan nisbah serta patokan topik. Topiknya  $\beta_{1:K}$ , dimana tiap  $\beta_k$  merupakan penyaluran kosakata. Nisbah topik buat akta ke-  $d$  merupakan  $\theta_d$ , dimana  $\theta_{d,k}$  merupakan nisbah topik buat topik  $k$  dalam naskah  $d$ . Kewajiban topik buat  $d$  naskah  $th$  merupakan  $Z_d$ , dimana  $Z_{d,n}$  merupakan penugasan topik buat tutur ke-  $n$  dalam naskah  $d$ . Kesimpulannya, kata-kata yang dicermati buat naskah  $d$  merupakan  $w_d$ , dimana  $w_{d,n}$  merupakan tutur ke-  $n$  dalam naskah  $d$ , bagian dari kosakata pasti. Dengan catatan ini, cara generatif buat LDA cocok dengan peluang bersama dari peubah tersembunyi serta yang dicermati pada rumus (1).

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D p(\theta_d) \left( \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right). \quad (1)$$

Sumber: Blei (2012)

## METODE PENELITIAN

Metode penelitian mencakup studi pustaka, pengumpulan data, pengolahan awal data, paradigma LDA, serta evaluasi. Alur penelitian bisa diamati dalam Gambar 2.



**Gambar 2. Alur Penelitian**

Sumber: Sahria (2020)

### Studi Pustaka

Studi pustaka melakukan studi makalah, buku, dan halaman web yang berkaitan dengan pemodelan topik LDA. Penelitian ini mengulas lima penelitian terdahulu yang penting dengan kasus yang hendak diawasi mengenai pemodelan topik memakai tata cara LDA. Data yang didapat dari riset pustaka hendak digunakan buat referensi teori serta ulasan riset.

### Pengumpulan Data

Informasi riset yang dipakai berbentuk informasi file informasi yang didapat lewat cara pengumpulan informasi dengan cara scraping pada web kontan. co. id. Informasi riset yang digabungkan sebesar 181 data saham sepanjang Februari sampai Juli tahun 2021. Web kontan. co. id diseleksi, sebab data saham sudah bergabung dengan abstrak saham di Google untuk perusahaan yang terdaftar dalam BEI. Web kontan. co. id menampung kian besar data hal saham perusahaan Indonesia ketimbang situs web lain.

## Pengolahan Awal Data

Wahyuni, dkk (2020) menyatakan pengolahan awal data (*corpus*) memiliki tahapan yang terdiri dari

### a. Case Folding

Case Folding atau *case conversion* merupakan cara penyamaan case ataupun huruf yang ada dalam suatu naskah (Sarkar, 2016). Tidak seluruh naskah bacaan tidak berubah-ubah dalam pemakaian huruf kapital. Oleh sebab itu kedudukan cara case folding diperlukan dalam mengkonversi totalitas ciri abjad yang terdapat dalam naskah jadi huruf kecil. Dalam melaksanakan penyamaan huruf jadi huruf kecil memakai fungsi *lowercase*.

### b. Cleansing

Cleansing adalah proses buat membersihkan naskah dari karakter-karakter yang tidak dibutuhkan dengan tujuan buat kurangi *noise* (Sarkar, 2016). Proses cleansing umumnya menghilangkan URL situs lain, simbol angka, tanda baca, huruf yang berulang, dan huruf tunggal.

### c. Stopwords

Stopwords merupakan cara buat melenyapkan kata yang kerap timbul, tetapi tidak mempunyai arti khusus ataupun kedudukannya tidak dibutuhkan dalam analisa teks (Sarkar, 2016). Catatan stopwords yang dipakai merupakan stopwords untuk bahasa Indonesia dirangkai dari riset Fadillah Z. Tala( 2003).

### d. Stemming

Stemming merupakan sesuatu cara buat menciptakan kata dasar (Sarkar, 2016). Cara stemming berbeda untuk suatu bahasa. Ilustrasinya, dalam bahasa Indonesia diketahui terdapatnya awalan, akhiran dan juga sisipan, sedang untuk bahasa Inggris, hanya dikenal akhiran saja. Untuk bahasa Indonesia, stemming dilakukan dengan melenyapkan seluruh

imbuhan( affixes) berbentuk prefiks( prefixes), akhiran( suffixes), ataupun campuran dari prefiks serta akhiran( confixes) yang terdapat pada tiap kata dalam informasi. Bila imbuhan dibuang, lalu sebuah kata dasar hendak ditaruh pada bermacam wujud berlainan akan cocok dengan imbuhan yang melekatinya, jika tidak perihai itu akan memunculkan noise. Proses Stemming menggunakan *library* Sastrawi, yakni *library Stemmer* bahasa Indonesia.

## Pemodelan LDA

Pemodelan topik pada riset ini menciptakan topik dalam berita saham online. Pemodelan topik bisa melukiskan arti dari naskah dengan cara memberikan tanda yang tersembunyi dalam bacaan jumlahnya banyak serta bisa menciptakan informasi dari bacaan tidak tertata (Blei, et. al, 2003).

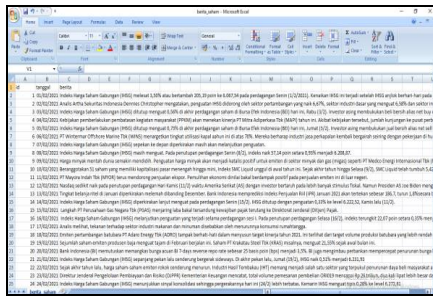
## Evaluasi

Penelitian ini menggunakan *pyLDavis* untuk memaknakan topik dalam bentuk topik yang cocok dengan berkas informasi bacaan. *Library python* ini mengambil sari data dari wujud topik LDA yang disetel buat memberikan penggambaran berplatform website saling aktif. Deskripsi ini dipakai pada notebook *ipython*, tetapi dapat ditaruh ke dokumen HTML tunggal buat mengampangkan beri data( *pypi. org*, 2021). *PyLDavis* sediakan 2 panel penggambaran selaku yang ditunjukkan pada panel bagian kiri membuktikan topik dengan cara totalitas. Pada panel bagian kiri bisa pula diamati ikatan ataupun hubungan antara topik satu dengan yang lain dengan memandang denah jarak antar topik. Ada pula bagian panel sisi kanan menampilkan penyaluran frekuensi tutur yang timbul pada tiap topik yang diperoleh.

## HASIL DAN PEMBAHASAN

Scraping data dari situs *kontan.co.id* menghasilkan 181 berita saham periode bulan Februari sampai

dengan Juli 2021. Gambar 3 menunjukkan hasil scraping berita saham yang disimpan dalam format \*.xls.



Gambar 3. berita\_saham.xls

Penerapan pengolahan awal data dilakukan saat sebelum menganalisis topik menggunakan LDA, yakni dengan menstrukturkan, berbenah, serta menaruh informasi sedia dianalisis. Pemrosesan awal berperan buat mensterilkan informasi yang dicoba buat menjauhi informasi yang tidak sempurna, informasi bermasalah, serta informasi yang berubah-ubah.

Case Folding Result :

```
0 indeks harga saham gabungan (ihsg) melesat :
1 analis artha sekuritas indonesia dennies chr
2 indeks harga saham gabungan (ihsg) ditutup #
3 kebijakan pemberlakuan pembatasan kegiatan
4 indeks harga saham gabungan (ihsg) ditutup #
Name: berita, dtype: object
```

Gambar 4. Case folding

Gambar 4. menjelaskan hasil pengolahan awal case folding. Selanjutnya, Gambar 5. menjelaskan hasil *tokenizing* atau menghapus karakter-karakter yang tidak penting dari teks.

Tokenizing Result :

```
0 [indeks, harga, saham, gabungan, ihsg, melesat...
1 [analis, artha, sekuritas, indonesia, dennies,...
2 [indeks, harga, saham, gabungan, ihsg, ditutup...
3 [kebijakan, pemberlakuan, pembatasan, kegiatan...
4 [indeks, harga, saham, gabungan, ihsg, ditutup...
Name: textdata_tokens, dtype: object
```

Gambar 5. Tokenizing

Gambar 6. menjelaskan stopwords atau membersihkan teks dari kata-kata yang termasuk ke dalam stopwords.

```
data8['textdata_tokens_8SH'] = data8['textdata_tokens'].apply(stopwords_remove)
print(data8['textdata_tokens_8SH'].head())
0 [indeks, harga, saham, gabungan, ihsg, melesat...
1 [analis, artha, sekuritas, indonesia, dennies, christophe...
2 [indeks, harga, saham, gabungan, ihsg, ditutup...
3 [kebijakan, pemberlakuan, pembatasan, kegiatan...
4 [indeks, harga, saham, gabungan, ihsg, ditutup...
Name: textdata_tokens_8SH, dtype: object
```

Gambar 6. Stopwords

Kemudian, Gambar 7. menjelaskan hasil stemming dari 181 berita saham.

```
0 [indeks, harga, saham, gabungan, ihsg, lesat, po...
1 [analis, artha, sekuritas, dennies, christophe...
2 [indeks, harga, saham, gabungan, ihsg, tutup, ku...
3 [bijak, laku, batas, giat, masyarakat, ppm, t...
4 [indeks, harga, saham, gabungan, ihsg, tutup, ku...
...
176 [susun, saham, indeks, konstituen, kerap, acu,...
177 [the, fed, tahan, suku, bunga, acu, pasar, nya...
...
178 [dagang, kemarin, kuat, indeks, topang, saham,...
179 [bank, central, asia, bbca, rencana, aksi, kor...
180 [sahamsaham, lapis, dominasi, daftar, saham, c...
Name: textdata_tokens_stemmed2, Length: 181, dtype: object
```

Gambar 7. Stemming

Setelah melakukan pengolahan awal teks, hasilnya dipakai untuk membuat pemodelan topik dengan LDA *gensim*. Sebelum pemodelan topik, terlebih dahulu dicoba pembuatan *dictionary* dan *corpus*. Proses pemodelan topik LDA dilakukan dengan menentukan *doc\_term\_matrix*, *word embedding* menggunakan library *doc2bow* untuk mengubah data string kata menjadi angka yang kemudian disimpan dalam *bag of word*. Langkah berikutnya mengatur jumlah topik yang menjadi klasterisasi kata, jumlah kata per topik, dan penentuan jumlah iterasi dalam pemodelan topik proses pembelajaran dari training model.

Gambar 8. menunjukkan hasil pemodelan topik LDA yang memiliki probabilitas tertinggi.

```
lda_model.show_topics(num_topics=total_topics, num_words=number_words)
([0,
 '0.022*saham' + 0.020*harga + 0.013*nilai + 0.012*pasar + 0.011*as + 0.01
 (1,
 '0.008*indeks' + 0.008*kerja + 0.008*rupiah + 0.008*analis + 0.007*investa
 (2,
 '0.042*saham' + 0.026*ihsg + 0.025*harga + 0.021*indeks + 0.018*kuat + 0.1
```

Gambar 8. Pemodelan topik LDA

Pemodelan topik LDA juga mengeluarkan kepentingan kata dan banyaknya kemunculan kata yang menyusun topik dapat dilihat pada Gambar 9.



|    | word      | topic_id | importance | word_count |
|----|-----------|----------|------------|------------|
| 0  | saham     | 0        | 0.021511   | 102        |
| 1  | harga     | 0        | 0.019804   | 69         |
| 2  | nilai     | 0        | 0.012675   | 17         |
| 3  | pasar     | 0        | 0.012111   | 27         |
| 4  | as        | 0        | 0.010822   | 16         |
| 5  | emas      | 0        | 0.010413   | 15         |
| 6  | rupiah    | 0        | 0.010239   | 25         |
| 7  | obligasi  | 0        | 0.007954   | 14         |
| 8  | bank      | 0        | 0.007645   | 14         |
| 9  | emiten    | 0        | 0.007118   | 41         |
| 10 | indeks    | 1        | 0.008287   | 49         |
| 11 | kerja     | 1        | 0.008192   | 29         |
| 12 | rupiah    | 1        | 0.007574   | 25         |
| 13 | analisis  | 1        | 0.007553   | 23         |
| 14 | investasi | 1        | 0.007278   | 13         |
| 15 | hasil     | 1        | 0.005825   | 17         |
| 16 | sekuritas | 1        | 0.005706   | 18         |
| 17 | utang     | 1        | 0.005582   | 9          |
| 18 | negeri    | 1        | 0.005554   | 9          |
| 19 | reksadana | 1        | 0.005521   | 19         |
| 20 | saham     | 2        | 0.041615   | 102        |
| 21 | ihsg      | 2        | 0.026454   | 53         |
| 22 | harga     | 2        | 0.024957   | 69         |
| 23 | indeks    | 2        | 0.020533   | 49         |
| 24 | kuat      | 2        | 0.017600   | 37         |
| 25 | emiten    | 2        | 0.016567   | 41         |
| 26 | sektor    | 2        | 0.014639   | 27         |
| 27 | dagang    | 2        | 0.013196   | 28         |
| 28 | investor  | 2        | 0.012651   | 30         |
| 29 | gabung    | 2        | 0.012520   | 23         |

Gambar 9. Importance dan word\_count

Gambar 10. menjelaskan topik yang dominan dan persentase kontribusi tiap topik dari sepuluh dokumen.

Berdasarkan hasil perolehan data pada pemodelan topik LDA dan topik yang dominan, hasil tersebut dapat divisualisasikan dengan pyLDavis.

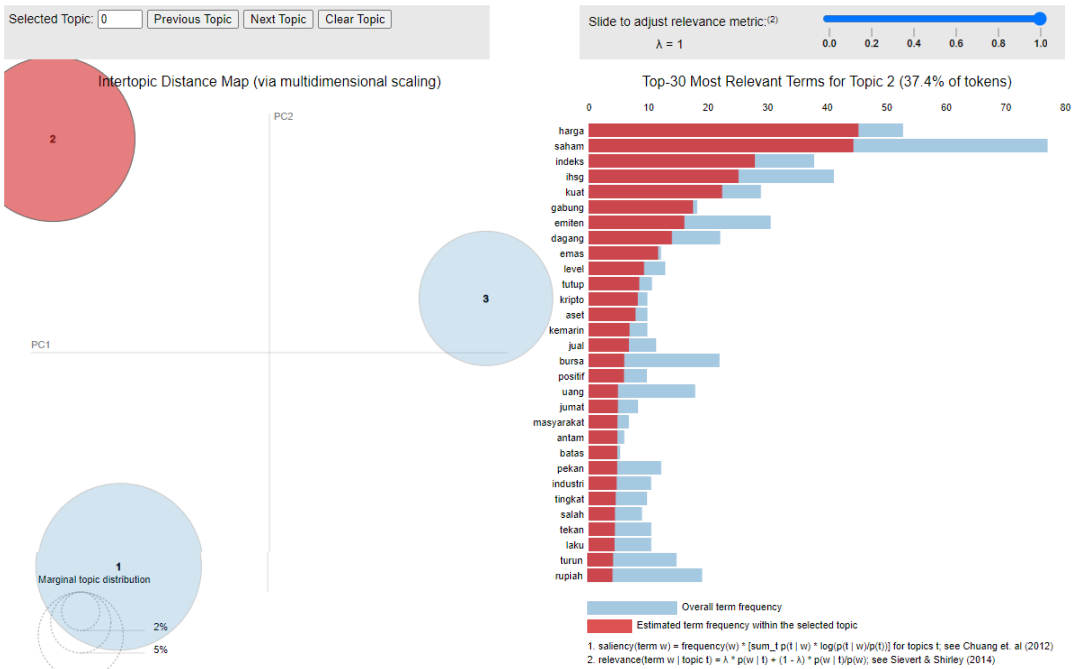
| Document_No | Dominant_Topic | Topic_Perc_Contrib |
|-------------|----------------|--------------------|
| 0           | 0              | 2.0                |
| 1           | 1              | 2.0                |
| 2           | 2              | 2.0                |
| 3           | 3              | 1.0                |
| 4           | 4              | 2.0                |
| 5           | 5              | 1.0                |
| 6           | 6              | 2.0                |
| 7           | 7              | 2.0                |
| 8           | 8              | 2.0                |
| 9           | 9              | 1.0                |

Gambar 10. Topik yang dominan dan persentase kontribusi tiap topik

Gambar 11. menunjukkan keterkaitan antar kata yang ditampilkan dalam bentuk diagram lingkaran dan grafik batang mendatar.

Pembahasan

Tabel 1. menjelaskan kekerapan dan kepentingan kata dari ketiga topik hasil pemodelan topik LDA seperti Gambar 9. Kata saham menduduki peringkat pertama pada topik 1 dan 3, diikuti kata harga di peringkat kedua juga pada topik 1 dan 3, kemudian kata ihsg di peringkat ketiga pada topik 3, kata indeks di peringkat keempat pada topik 2 dan 3, serta kata emiten di peringkat kelima pada topik 1 dan 3.



Gambar 11. Visualisasi pyLDavis

**Tabel 1. Peringkat menurut frekuensi dan kepentingan kata dari 3 topik**

| no | kata   | importance | word_count | topik |
|----|--------|------------|------------|-------|
| 1  | saham  | 0.021511   | 102        | 1     |
|    |        | 0.041615   | 102        | 3     |
| 2  | harga  | 0.019804   | 69         | 1     |
|    |        | 0.024957   | 69         | 3     |
| 3  | ihsg   | 0.026454   | 53         | 3     |
| 4  | indeks | 0.008287   | 49         | 2     |
|    |        | 0.020533   | 49         | 3     |
| 5  | emiten | 0.007118   | 41         | 1     |
|    |        | 0.016567   | 41         | 3     |

Kata saham menempati urutan pertama dari berita saham online, karena saham merupakan salah satu wujud investasi yang banyak dipilih para penanam modal untuk memperoleh keuntungan yang menarik.

Kata harga menempati urutan kedua dalam berita saham online, karena harga menjadi estimasi penting untuk penanam modal ataupun warga, kala akan membeli saham di perusahaan tertentu.

Kata indeks harga saham gabungan (ihsg) menduduki ranking ketiga dalam berita saham online, karena Indeks harga saham gabungan ialah gambaran dari dorongan naik serta turunnya saham dengan cara totalitas yang terdaftar dalam bursa efek Indonesia pada satuan waktu tertentu.

Kata indeks menduduki peringkat keempat dalam berita saham online, sebab indeks mempunyai arti catatan harga saat ini dibanding dengan harga lebih dahulu sesuai persentase buat mengenali perubahan harga saham.

## KESIMPULAN

Pemodelan topik menghasilkan 3 topik berita saham online dari situs web kontan.co.id yang terdiri dari saham, harga, dan indeks harga saham gabungan (ihsg). Visualisasi pemodelan topik dengan pyLDavis menampilkan tiga kelompok topik dan 30 kata yang relevan untuk tiap topik.

Informasi yang dipakai dalam riset ini berasal dari laman website informasi saham kontan. co. id yang memakai bahasa informasi alat online. Pada pengembangan riset bisa menggunakan kumpulan data dari situs online lain. Kata luaran pemodelan topik, ada yang berupa julukan perusahaan, berbentuk kependekan yang mempunyai maksud serupa, tetapi dalam pemodelan topik jadi 2 entitas yang berlainan, semacam kata indeks harga saham gabungan buat ihsg serta bei bagi Bursa Efek Indonesia. Oleh karena itu, proses pemodelan topik membutuhkan cara normalisasi.

## DAFTAR PUSTAKA

- Bansal, H. (2020). Latent Dirichlet Allocation. <https://iq.opengenus.org/latent-dirichlet-allocation>, 12 November 2020.
- B. S. Halima & S. Chitrakala. (2016). Trending topic analysis using novel sub topic detection model. 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). IEEE. pp. 157-161.
- Bashri, M. F. A. dan Retno Kusumaningrum. (2017). Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity Wordcloud Visualization. Fifth International Conference on Information and Communication Technology (ICoICT)
- Blei, David M. (2012). Probabilistic Topic Models, *Communications of the ACM Vol. 55, No. 4*, pp77-84.
- Blei, David M., Andrew Y. Ng, Michel I. Jordan. (2003). Latent Dirichlet Allocation. *Journal of*



- Machine Learning Research* 3 (2003) 993-1022
- Blei, David M. and John D. Lafferty. (2007). A Correlated Topic Model of Science, *Annals of Applied Statistics Vol. 1, No. 1*, pp17-35.
- Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing and Management*, 26(7), 675–693.  
<https://doi.org/10.1080/19368623.2017.1310075>
- Dadgar, S. M. H., et al. (2016). A novel text mining approach based on TF-IDF and Support Vector Machine for news classification. IEEE International Conference on Engineering and Technology (ICETECH). IEEE, pp.112-116.
- E. S. Usop, R. R. Isnanto, and R. Kusumaningrum, (2017). Part of Speech Features for Sentiment Classification based on Latent Dirichlet Allocation, in Proc. of 2017 4th Int. Conf. on Information Tech., Computer, and Electrical Engineering (ICITACEE), 2017, pp. 27–30.
- Fadillah Z. Tala. (2003). A study of stemming effects on Information retrieval in Bahasa Indonesia. Institute for Logic, Language and Computation, Universiteit van Amsterdam, Netherlands.
- Grun, Bettina and Kurt Hornik. (2011). topicmodels: An R Package for Fitting Topic Model, *Journal of Statistical Software Vol. 40, No. 13*.
- I R Putri dan R Kusumaningrum. (2017). Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia. *J. Phys.: Conf. Ser.* 801 0120
- J. F. Yeh, Y. S. Tan & C. H. Lee. (2016). Topic detection and tracking for conversational content by using conceptual dynamic Latent Dirichlet Allocation, *Neurocomputing*.
- Jing, Qi. (2015). Searching for Economic Effects of User Specified Event Based on Topic Modelling and Event Reference, Jordery School of Computer Science, Acadia University.
- Kusumaningrum, Retno, dkk. (2016). Classification of Indonesian News Articles based on Latent Dirichlet Allocation. International Conference on Data and Software Engineering (ICoDSE), Udayana, Bali.
- Lucini, F. R., Tonetto, L. M., Fogliatto, F. S., & Anzanello, M. J. (2020). “Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews”. *Journal of Air Transport Management*, 83(December 2019), 101760.  
<https://doi.org/10.1016/j.jairtraman.2019.101760>
- Putra, I Made Kusnanta Bramantya. (2017). Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA). *Jurnal Teknik ITS Vol. 6, No. 2* (2017), 2337-3520 (2301-928X Print)
- Pypi.org. (2021). pyLDAvis 3.3.1. <https://pypi.org/project/pyLDAvis/>, 24/03/2021
- Rachman, F. F., Nooraeni, R., & Yuliana, L. (2021). Public Opinion of Transportation integrated (Jak Lingko), in DKI Jakarta, Indonesia. *Procedia*

- Computer Science*, 179(2020), 696–703.  
<https://doi.org/10.1016/j.procs.2021.01.057>
- Sahria, Yoga dan Dhomas Hatta Fudholi. (2020). Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode *Topic Modeling LDA (Latent Dirichlet Allocation)*. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)* Vol. 4 No. 2 (2020) 336 – 344
- Sarkar, Dipanjan. (2016). *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data*. APress. Bangalore, Karnataka, India.
- Setijohatmo, U. T., dkk. (2020). Analisis metoda Latent Dirichlet Allocation untuk klasifikasi dokumen laporan tugas akhir berdasarkan pemodelan topik. Prosiding The 11th Industrial Research Workshop and National Seminar Bandung, 26-27 Agustus 2020.
- T. Kanan dan E. A. Fox. (2016). Automated Arabic text classification with P-Stemmer, machine learning, and a tailored news article taxonomy. *J. Assoc. Inf. Sci. Technol.*
- Tong, Zhou & Haiyi. Zhang (2016). A text mining research based on LDA topic modelling, *Jodrey School of Computer Science, Acadia University, Wolfville, NS, Canada*, vol. 10.5121/csit.2016.60616, p. 201–210.
- Wahyuni, Eka Dyar, dkk. (2020). *Klasifikasi Teks dengan Python*. Sidoarjo: Indomedia Pustaka.