p-ISSN: 2406-7768
e-ISSN : 2581-2181

TOPIC MODELLING SKRIPSI MENGGUNAKAN METODE *LATENT DIRICLHET ALLOCATION*

Alif Iffan Alfanzar[1], Khalid[2], Indri Sudanawati Rozas[3]

[1,2,3] *Jurusan Sistem Informasi Fakultas Sains dan Teknologi Universitas Islam Negeri Sunan Ampel Surabaya*
Jl. Ahmad Yani No.117 Jemur Wonosari – Wonocolo Kota Surabaya

[1]alfanzar27@gmail.com
[2]khalid@uinsby.ac.id
[3]indrisrozas@gmail.com

Abstrak **-** Program Studi Sastra Inggris di Universitas Islam Negeri Sunan Ampel Surabaya (UINSA) telah ditemukan permasalahan bahwa belum ada yang melakukan clustering pada topik skripsi mahasiswa. Clustering tersebut digunakan dalam topic modelling untuk melihat tren dan kesuaian minat pada Program Studi Sastra Inggris UINSA. Metode Latent Dirichlet Allocation (LDA) merupakan salah satu metode topic modelling yang paling populer saat ini. Dalam penelitian ini mengambil sejumlah 584 abstact skirpsi dalam bahasa inggris sebagai dataset. Penggunaan dataset berbahasa inggris dikarenakan pada pre-processing data yang tersedia standartnya baru untuk bahasa inggris. Setelah melewati proses tersebut, setiap kata yang muncul akan dihitung menggunakan metode Bag of Word. Metode LDA mengklusterkan dengan menglihat jumlah kemunculan kata pada Bag of Word, kemudian menentukan jumlah cluster atau jumlah topik dan menentukan jumlah iterasi. LDA menandai setiap kata pada topik secara semi random distribution kemudian menghitung probabilitas topik pada dokumen dan menghitung probabilitas kata pada topik setiap iterasinya. Penelitian ini melakukan percobaan pemodelan topik sebanyak 5 kali uji iterasi dan jumlah topik yang berbeda. Berdasarkan percobaan tersebut telah didapatkan hasil kemudian dianalisis bahwa 3 adalah jumlah topik yang paling fit. Hasil tersebut diujikan secara kualitatif kepada pihak stakeholder Program Studi Sastra Inggris UINSA, dan dinyatakan sesuai dengan tren serta minat pada Program Studi Sastra Inggris UINSA.

Kata kunci : *Clustering, Iterasi*, LDA, Probabilitas, *Topic Modelling*.

I. PENDAHULUAN

Pada era saat ini perusahaan sudah banyak merasakan efektifitas dari penggunaan data mining untuk melihat segmentasi pasar [1]. Sebuah proses yang menggunakan teknk statistika, matematika, kecedarsan buatan dan machine learning untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dari berbagai database besar disebut Data mining [2]. Data mining diterapkan bukan hanya dalam menganalisa perusahaan saja. Ada banyak bidang yang menjadikan data mining sebagai solusi untuk memecahkan masalah yang ada. Data mining memiliki bidang khusus yang hampir sama, yakni text mining [3]. Text mining sendiri merupakan ilmu bagian dari data mining yang berupaya menemukan pola yang menarik dari sekumpulan data text dalam jumlah yang besar [4]. pencarian pola dalam text merupakan tujuan dari text mining melalui proses analisis text guna mencari informasi yang bermanfaat. Salah satu fungsi data mining maupun text mining adalah clustering. Clustering merupakan metode yang bersifat tanpa arahan (unsupervised). Tanpa arahan yang dimaksud dalam metode ini diterapkan tanpa adanya data latihan (data traning) serta tidak memererlukan target keluaran (output). Metode clustering dibagi menjadi dua jenis dalam ngelompokkan data, yaitu hierarchical clustering dan non-hierarchical clustering.

Permasalahan yang terjadi saat ini pada Program Studi Sastra Inggris UIN Sunan Ampel Surabaya (UINSA) adalah belum mengetahui jumlah cluster topik penelitian untuk

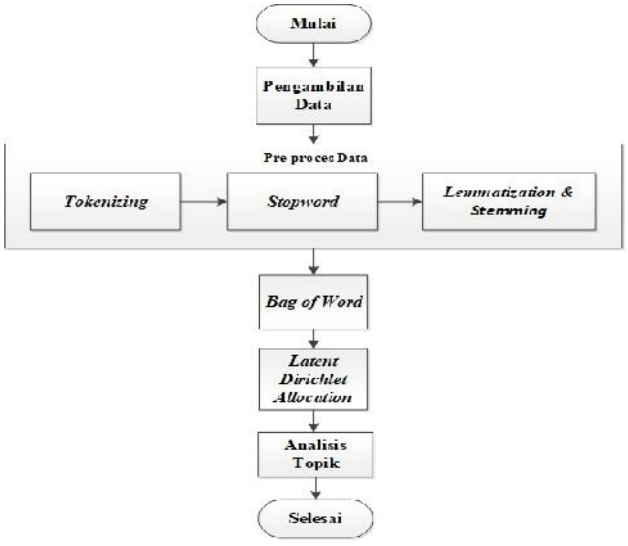
p-ISSN: 2406-7768
e-ISSN : 2581-2181

skripsi Program Studi Sastra Inggris UIN Sunan Ampel Surabaya (UINSA). Teknik pengklusteran pada topik penelitian dibutuhkan untuk melihat tren topik yang ada. Teknik mengklusterkan topik bisa dilakukan manual oleh manusia, akan tetapi dapat menghabiskan banyak waktu. Permasalahan waktu tersebut dapat diselesaikan dengan menggunakan bantuan computer dengan metode topic modelling. Topic modelling merupakan metode non-hierarchical clustering yang secara otomatis mengklusterkan kedalam topik yang muncul dari pemodelan sehingga didapatkan topik cluster yang sesuai. Solusi ini dapat mengatasi pemasalahan pada Program Studi Sastra Inggris UIN Sunan Ampel Surabaya saat ini.

Topic modelling mempunyai banyak metode yang dapat digunakan seperti Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA). LDA Menurut [5] merupakan peningkatan cara model campuran yang menangkap pertukaran dari kata-kata dan dokumen dari cara lama oleh PLSA dan LSA. Untuk saat ini metode yang paling populer dalam topic modelling adalah metode LDA. LDA mampu meringkas, mengklusterkan, menghubungkan, dan memproses data text yang sangat besar. sehingga dalam penelitian ini LDA dipilih sebagai metode topic modelling pada Program Studi Sastra Inggris (UINSA).

II. METODOLOGI PENELITIAN

Metodologi penelitian merupakan prosedur beserta langkah-langkah yang disusun secara sistematis untuk menyelesaikan permasalahan yang sedang diteliti dengan landasan ilmiah tertentu Kerangka metodologi penelitian dapat dilihat pada Gambar 1 berikut ini.



Gambar 1. Kerangka Metodologi Penelitian.

A. *Pengambilan Data*

Tahap pertama penelitian adalah pengambilan data dengan mengambil data dari website digilib uinsa. Proses pengambilan data menggunakan tool dari Google Chrome yang bernama web scrapper. Penelitian ini mempunyai batasan masalah yaitu mengambil data abstract pada penelitian yang dilakukan oleh program studi sastra inggris UINSA.

B. *Pre-Processing Data*

Pada tahap Pre-processing Merubah bentuk dokumen kedalam bentuk yang memudahkan dan mempercepat proses dalam dokumen yang relavan [6]. Setiap pre-processing ditugaskan untuk membangun index dari koleksi dokumen. Pengindeksan dilakukan untuk membedakan suatu dokumen terhadap dokumen yang lain. pembuatan index harus melibatkan konsep linguistic processing yang bertujuan untuk mengekstrak term-term penting dari setiap dokumen yang direpresentasikan sebagai bag of words. Konsep linguistic processing terdiri dari tokenizing, stopwords, lemmatization dan stemming. Berikut merupakan empat tahap pre-process.

C. *Bag of Words*

Setelah proses pre-processing dilakukan, matriks yang berisi kata-kata tersebut dimodelkan dengan model bag of words. Bag of words digunakan untuk memodelkan setiap dokumen dengan menghitung jumlah kemunculan setiap katanya. Model bag of words merepresentasikan setiap dokumen dengan mengabaikan urutan dari kata-kata dalam dokumen serta struktur sintakis dari dokumen dan kalimat. Nilai perhitungan jumlah kemunculan setiap kata tersebut digunakan dalam topic modelling [7].

D. *Latent Dirichlet Allocation (LDA)*

Proses pemodelan topik bertujuan untuk memperoleh distribusi kata yang membentuk suatu topik dan dokumen dengan topik tertentu. Pemodelan topik memiliki dua tahapan yang dilakukan. Tahap pertama adalah melakukan pemodelan topik berdasarkan penambahan dan pengurangan jumlah topik. Tahap kedua adalah melakukan pemodelan topik berdasarkan banyaknya iterasi. Hasil dari kedua pemodelan topik kemudian dilakukan analisa dengan cara membandingkan kata kata setiap klasternya dalam topik dan melihat visualisasi dari pemodelan LDA tersebut. Proses pemodelan topik dapat berulang selama rentang kandidat jumlah topik dan jumlah iterasi yang di tentukan [5].

Latent Dirichlet Allocation (LDA) merupakan salah satu model dari pemodelan topik. Model topik LDA merupakan unsupervised machine learning. Model tersebut berguna dalam mengidentifikasi informasi tersembunyi dalam kumpulan dokumen yang berukuran besar. Metode ini dapat diselesaikan menggunakan python, dengan terlebih dahulu

p-ISSN: 2406-7768
e-ISSN : 2581-2181

mengaktifkan package “LdaModel” dalam library gensim. Package “LdaModel” untuk memodelkan probabilitas kemunculan kata dalam dokumen. Menghasilkan data keluaran berupa grafik yang menunjukan topik pada data yang diteliti.

E. Analisis Data

Analisis topik dilakukan berdasarkan pada data keluaran dari tahap sebelumnya. Pada tahap sebelumnya diperoleh grafik data keluaran dari kumpulan penelitian dengan topik tertentu. Analisis topik dilakukan secara subjektif dengan melihat data keluaran. Data keluaran berupa kumpulan kata yang membentuk topik, kemudian setiap dokumen tersebut disesuaikan dengan data keluaran yang memuat dokumen dengan topik. Proses ini menghasilkan deskripsi topik yang bersifat informatif mengenai hal yang dapat mewakili isi dari masing-masing topik tersebut.

III. HASIL DAN PEMBAHASAN

A. Pengambilan Data

Tahap pertama yang dilakukan dalam penelitian ini adalah pengambilan data. Pada proses pengambilan data digunakan metode scrapping dengan menggunakan tools extension google chrome yang bernama web scraper. Data yang sudah terambil disimpan dalam bentuk csv. Jumlah dari keseluruhan total abstract yang diambil sebanyak 584 row data abstract. 585 row data yang telah tersimpan tersebut nantinya akan digunakan pada proses selanjutnya, yaitu tahap pre-processing data.

B. Pre-processing Data

Tahap yang bertujuan untuk mempersiapkan data sebelum dianalisis menggunakan LDA [6]. Pre-processing dilakukan dengan menggunakan aplikasi Jupytherlab. Langkah pada proses ini dimulai dengan mengunduh dan mengaktifkan beberapa library yang dibutuhkan. Tahap pada pre-processing data dibagi menjadi lima tahap. Tahapan tersebut diantaranya tokenizing, stopword, lemmatization dan stemming.

a. Tokenizing

Langkah pertama dari Pre-processing data adalah tokenizing. Tujuan dari proses ini adalah untuk memisahkan setiap kata ke dalam unit-unit kecil dalam suatu array atau term [8]. Tokenizing memisahkan setiap katanya oleh karakter spasi, sehingga pada proses ini mengandalkan karakter spasi pada dokumen untuk melakukan pemisahan. Pada proses ini bertujuan juga untuk menghilangkan mention, url, dan tanda baca yang ada pada teks. Tokenizing juga merupah setiap huruf dengan karakter uppercase menjadi karakter huruf lowercase menggunakan fungsi lower. Perbedaan sebelum tokenizing dan sesudah tokenizing dapat dilihat pada Tabel I.

TABEL 1
PERBEDAAN SEBELUM DAN SESUDAH TOKENIZING

Sebelum *Tokenizing*	**Sesudah *Tokenizing***
This thesis tries to analyze the drama script from Samuel Beckett entitled Endgame. This drama tells of an isolated Hamm in his own home with a saturating activity with his Clov aides and his parents Nagg and Nell, so they are trying to bring the side of the reality of life to be peeled. This thesis focuses on the analysis of Hamm's consciousness of all past and present events. The purpose of this thesis is to describe the absurdity of life of the main character and to reveal the meaning behind it all in the absurdity drama script.	["", 'this', 'thesis', 'tries', 'to', 'analyze', 'the', 'drama', 'script', 'from', 'samuel', 'beckett', 'entitled', 'endgame', '.', 'this', 'drama', 'tells', 'of', 'an', 'isolated', 'hamm', 'in', 'his', 'own', 'home', 'with', 'a', 'saturating', 'activity', 'with', 'his', 'clov', 'aides', 'and', 'his', 'parents', 'nagg', 'and', 'nell', ',', 'so', 'they', 'are', 'trying', 'to', 'bring', 'the', 'side', 'of', 'the', 'reality', 'of', 'life', 'to', 'be', 'peeled', '.', 'this', 'thesis', 'focuses', 'on', 'the', 'analysis', 'of', 'hamm', "'s", 'consciousness', 'of', 'all', 'past', 'and', 'present', 'events', '.', 'the', 'purpose', 'of', 'this', 'thesis', 'is', 'to', 'describe', 'the', 'absurdity', 'of', 'life', 'of', 'the', 'main', 'character', 'and', 'to', 'reveal', 'the', 'meaning', 'behind', 'it', 'all', 'in', 'the', 'absurdity', 'drama', 'script', '.',

b. Stopword

Setelah melewati tahap tokenizing, selanjutnya term dokumen diolah pada proses stopword. Proses stopword dilakukan untuk menghapus kata-kata yang tidak mempunyai informasi atau dengan kata lain hanya mengambil kata yang penting saja [9]. Dalam penelitian ini proses stopword terbagi menjadi 3 tahapan. Tahap pertama menggunakan library yang sudah tersedia di dalam python yaitu nltk.download (‘stopword’). Kemudian pada stopword ditambahkan nltk.corpus.stopword.words(‘english’) di dalam fungsi set() untuk mendownload kata-kata stopwords bahasa inggris yang sudah ditetapkan. Perbedaan sebelum dan sesudah proses stopword tahap pertama dapat dilihat Tabel II.

TABEL 2
PERBEDAAN SEBELUM DAN SESUDAH STOPWORD TAHAP PERTAMA

Sebelum *Stopwords*	**Sesudah *Stopwords***
["", 'this', 'thesis', 'tries', 'to', 'analyze', 'the', 'drama', 'script', 'from', 'samuel', 'beckett',	["", 'thesis', 'tries', 'analyze', 'drama', 'script', 'samuel', 'beckett', 'entitled',

p-ISSN: 2406-7768
e-ISSN : 2581-2181

'entitled', 'endgame', '.', 'this', 'drama', 'tells', 'of', 'an', 'isolated', 'hamm', 'in', 'his', 'own', 'home', 'with', 'a', 'saturating', 'activity', 'with', 'his', 'clov', 'aides', 'and', 'his', 'parents', 'nagg', 'and', 'nell', ',', 'so', 'they', 'are', 'trying', 'to', 'bring', 'the', 'side', 'of', 'the', 'reality', 'of', 'life', 'to', 'be', 'peeled', '.', 'this', 'thesis', 'focuses', 'on', 'the', 'analysis', 'of', 'hamm', "'s", 'consciousness', 'of', 'all', 'past', 'and', 'present', 'events', '.', 'the', 'purpose', 'of', 'this', 'thesis', 'is', 'to', 'describe', 'the', 'absurdity', 'of', 'life', 'of', 'the', 'main', 'character', 'and', 'to', 'reveal', 'the', 'meaning', 'behind', 'it', 'all', 'in', 'the', 'absurdity', 'drama', 'script', '.',	'endgame', '.', 'drama', 'tells', 'isolated', 'hamm', 'home', 'saturating', 'activity', 'clov', 'aides', 'parents', 'nagg', 'nell', ',', 'trying', 'bring', 'side', 'reality', 'life', 'peeled', '.', 'thesis', 'focuses', 'analysis', 'hamm', "'s", 'consciousness', 'past', 'present', 'events', '.', 'purpose', 'thesis', 'describe', 'absurdity', 'life', 'main', 'character', 'reveal', 'meaning', 'behind', 'absurdity', 'drama', 'script', '.'

Berdasarkan tahap pertama stopwords yang telah dilakukan, hasilnya masih terdapat kata-kata yang masih bersifat tidak informatif. Kata-kata yang tidak informatif ini sering kali muncul dalam sebuah abstract penelitian. Untuk menghilangkan kata-kata yang tidak informatif, perlu dilakukan stopword tahap kedua. Pada Tahap kedua stopword dilakukan dengan cara menambahkan kata-kata secara manual sesuai dengan kebutuhan. Kata-kata yang ditambahkan didapat dari hasil wawancara terhadap pihak sastra inggris Hasil dari tahap kedua stopword ini dapat dilihat pada Tabel III.

TABEL 3
PERBEDAAN SEBELUM DAN SESUDAH STOPWORD TAHAP KEDUA

Sebelum *Stopwords*	Sesudah *Stopwords*
['"', 'thesis', 'tries', 'analyze', 'drama', 'script', 'samuel', 'beckett', 'entitled', 'endgame', '.', 'drama', 'tells', 'isolated', 'hamm', 'home', 'saturating', 'activity', 'clov', 'aides', 'parents', 'nagg', 'nell', ',', 'trying', 'bring', 'side', 'reality', 'life', 'peeled', '.', 'thesis', 'focuses', 'analysis', 'hamm', "'s", 'consciousness', 'past', 'present', 'events', '.', 'purpose', 'thesis', 'describe', 'absurdity', 'life', 'main', 'character', 'reveal', 'meaning', 'behind', 'absurdity', 'drama', 'script', '.'	['"', 'tries', 'drama', 'script', 'samuel', 'beckett', 'entitled', 'endgame', '.', 'drama', 'tells', 'isolated', 'hamm', 'home', 'saturating', 'activity', 'clov', 'aides', 'parents', 'nagg', 'nell', ',', 'trying', 'bring', 'side', 'reality', 'life', 'peeled', '.', 'focuses', 'hamm', "'s", 'consciousness', 'past', 'present', 'events', '.', 'purpose', 'absurdity', 'life', 'main', 'reveal', 'behind', 'absurdity', 'drama', 'script', '.'

Hasil yang didapatkan berdasarkan tahap kedua stopword masih terdapat tanda baca yang terpisahkan dari proses tokenize. Tanda baca tersebut perlu dihilangkan. Untuk menghilangkan tanda baca, perlu dilakukan tahap ketiga stopword. Tahap ketiga stopword ini dengan cara menambahkan fungsi len(). Fungsi tersebut digunakan untuk mengembalikan panjang (jumlah anggota) dari suatu objek. Hasil dari tahap ketiga stopword ini dapat dilihat pada Tabel IV.

TABEL 4
PERBEDAAN SEBELUM DAN SESUDAH STOPWORD TAHAP KETIGA

Sebelum *Stopwords*	Sesudah *Stopwords*
['"', 'tries', 'drama', 'script', 'samuel', 'beckett', 'entitled', 'endgame', '.', 'drama', 'tells', 'isolated', 'hamm', 'home', 'saturating', 'activity', 'clov', 'aides', 'parents', 'nagg', 'nell', ',', 'trying', 'bring', 'side', 'reality', 'life', 'peeled', '.', 'focuses', 'hamm', "'s", 'consciousness', 'past', 'present', 'events', '.', 'purpose', 'absurdity', 'life', 'main', 'reveal', 'behind', 'absurdity', 'drama', 'script', '.'	['tries', 'drama', 'script', 'samuel', 'beckett', 'entitled', 'endgame', 'drama', 'tells', 'isolated', 'saturating', 'activity', 'aides', 'parents', 'trying', 'bring', 'reality', 'peeled', 'focuses', 'consciousness', 'present', 'events', 'purpose', 'absurdity', 'reveal', 'behind', 'absurdity', 'drama', 'script'

c. Lemmatization & Stemming

Hasil dari stopwords dilanjut ke proses Pre-processing teks mencakup proses stemming dan lemmatization. Lemmatization adalah sebuah proses pengelompokan kata yang berbeda dengan melalui tahap analisis sebagai satu kata yang sama. Berbeda dengan stemming, stemming merupakan sebuah proses menemukan sebuah kata dasar dari kata yang mempunyai awalan dan akhiran tanpa menganalisis apakah kata itu mempunyai arti yang sama dengan kata yang lain [10]. Oleh karena itu dalam proses penelitian ini digabungkan lemmatization dan stemming untuk mendapatkan hasil yang lebih baik dari pada menggunakan satu metode saja. Hasil dari tahap ketiga stopword ini dapat dilihat pada Tabel V.

p-ISSN: 2406-7768
e-ISSN : 2581-2181

TABEL 5
PERBEDAAN SEBELUM DAN SESUDAH LEMMATIZATION & STEMMING

Sebelum *Lemmatization* & *Stemming*	Sebelum *Lemmatization* & *Stemming*
['tries', 'drama', 'script', 'samuel', 'beckett', 'entitled', 'endgame', 'drama', 'tells', 'isolated', 'saturating', 'activity', 'aides', 'parents', 'trying', 'bring', 'reality', 'peeled', 'focuses', 'consciousness', 'present', 'events', 'purpose', 'absurdity', 'reveal', 'behind', 'absurdity', 'drama', 'script'	['try', 'drama', 'script', 'samuel', 'beckett', 'entitle', 'endgame', 'drama', 'tell', 'isolate', 'saturate', 'activity', 'aides', 'parent', 'try', 'bring', 'reality', 'peel', 'focus', 'consciousness', 'present', 'event', 'purpose', 'absurdity', 'reveal', 'behind', 'absurdity', 'drama', 'script'

C. Bag of Words

Hasil dari pre-processing data berupa term atau matriks yang berisi kata-kata yang muncul berulang-ulang. Model Bag of words digunakan menghitung jumlah kemunculan setiap kata pada term atau matriks kata-kata tersebut. Hasil perhitungan jumlah setiap kata tersebut digunakan dalam dalam perhitungan distribusi yang ada pada LDA. Langkah pertama pada proses ini yaitu menyimpan kata yang berbeda atau unique yang ada dalam term atau matriks kata-kata. Dalam proses tersebut dihasilkan 5993 unique kata yang berbeda. Langkah kedua dalam proses ini yaitu menghitung kata yang muncul. Dalam menghitung kata tersebut, Bag of Words mengindekskan setiap kata yang ada beserta menghitung kemunculan kata. Bag of words menghitung berdasarkan unique kata yang telah ditentukan pada proses sebelumnya. Hasil dari proses bag of words ditunjukan pada Gambar 2.

```
[[(0, 2), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, 2), (11, 1),
1), (20, 3), (21, 1), (22, 1), (23, 1), (24, 1), (25, 1), (26, 1), (27, 1), (28, 1), (29, 1), (30,
(38, 1), (39, 1), (40, 1), (41, 1), (42, 1), (43, 1), (44, 1), (45, 1), (46, 1), (47, 1), (48, 1),
1), (57, 1), (58, 2), (59, 1), (60, 1), (61, 2), (62, 1), (63, 1), (64, 1), (65, 1), (66, 1), (67,
(66, 1), (68, 1), (72, 1), (73, 5), (74, 1), (75, 1), (76, 1), (77, 1), (78, 1), (79, 2), (80, 1),
1), (89, 1), (96, 5), (91, 7), (92, 2), (93, 1), (94, 1), (95, 1), (96, 1), (97, 1), (98, 1), (99,
6, 1), (107, 1), (108, 1)], [(7, 2), (15, 1), (33, 1), (39, 2), (51, 1), (63, 4), (74, 3), (75, 2),
(114, 1), (115, 1), (116, 1), (117, 1), (118, 1), (119, 1), (120, 2), (121, 1), (122, 1), (123, 1),
1), (131, 1), (132, 1), (133, 5), (134, 1), (135, 1), (136, 2), (137, 1), (138, 5), (139, 1), (140,
7, 1), (148, 2), (149, 1), (150, 1), (151, 1), (152, 1), (153, 1), (154, 1), (155, 1), (156, 1), (1
(164, 1), (165, 1), (166, 2), (167, 14), (168, 1), (169, 1), (170, 3), (171, 4)], [(5, 1), (7, 1),
3), (76, 1), (77, 3), (78, 1), (81, 4), (172, 1), (173, 1), (174, 5), (175, 1), (176, 1), (177, 1),
1), (185, 1), (186, 3), (187, 3), (188, 1), (189, 1), (190, 1), (191, 1), (192, 1), (193, 1), (194,
1, 1), (202, 1), (203, 1), (204, 1), (205, 1), (206, 1), (207, 1), (208, 1), (209, 2), (210, 1), (2
(213, 1), (219, 1)], [(49, 1), (52, 1), (55, 1), (58, 1), (66, 1), (82, 1), (94, 1), (98, 1), (100,
1, 1), (220, 1), (221, 1), (222, 1), (223, 1), (224, 1), (225, 7), (226, 1), (227, 1), (228, 1), (2
(236, 1), (237, 1), (238, 7), (239, 1), (240, 1), (241, 3), (242, 1), (243, 1), (244, 1), (245, 1),
1), (253, 3), (254, 1), (255, 1), (256, 2), (257, 1), (258, 1), (259, 1), (260, 1)], [(55, 1), (61,
```

Gambar 2. Hasil Proses pada Bag of Words.

D. Latent Dirichlet Allocation (LDA)

Setelah serangkaian pre-processing yang telah dilakukan dan dimasukkan ke dalam bag of words tahap selanjunya adalah memodelkan topik menggunakan LDA. Sebelumnya pada tahap bag of words telah muncul token yang berasal dari banyaknya kata yang muncul dalam suatu dokumen. Token berfungsi sebagai ukuran dalam LDA agar dapat dimodelkan. Dalam pemodelan LDA perlu diadakan beberapa kali percobaan, supaya dapat menentukan jumlah topik yang sesuai. Percobaan yang dilakukan dengan mengubah jumlah topik dan percobaan dalam jumlah iterasi. Pada penelitian ini dilakukan percobaan sebanyak 5 uji iterasi dengan iterasi berbeda yakni: 100, 500, 1000, dan 5000. Sedangkan terhadap setiap uji iterasi dimasukkan jumlah topik yang berbeda yaitu: 2, 3, 4, dan 5.

E. Analisis Topik

Kemudian dilakukan analisis keseluruhan model LDA dari berbagai percobaan jumlah topik dan jumlah iterasi. Pada iterasi ke-100 menghasilkan 2 topik yang berbeda. Hasil yang sama juga didapat pada iterasi ke-500 dengan mengahasilkan 2 topik yang berbeda. Begitupun pada iterasi ke-1000 dan 5000 dimana masing-masing menghasilkan 2 topik yang berbeda. Berdasarkan pada percobaan dengan perbedaan iterasi yang digunakan, kedua topik mempunyai jarak yang berjauhan. Dari 100, 500, 1000, dan 5000 iterasi menunjukan bahwasanya topik yang terbentuk berjumlah 2 topik dari 2 jumlah topik yang ditentukan. 2 topik yang terbentuk memang sudah dapat dikategorikan sebagai 2 cluster yang berbeda. Akan tetapi untuk mengukur apakah 2 topik tersebut benar-benar 2 cluster yang berbeda, maka dilakukan pemodelan dengan jumlah topik 3 dan iterasi yang sama.

Pemodelan dengan jumlah topik 3 pada iterasi ke-100 menghasilkan 3 topik yang berbeda. Hasil yang sama didapat pada iterasi ke-500 dengan mengahasilkan 3 topik yang berbeda. Begitupun pada iterasi ke-1000 dan 5000 dimana masing-masing menghasilkan 3 topik yang berbeda. Berdasarkan pada percobaan dengan perbedaan iterasi yang digunakan, ketiga topik mempunyai jarak yang berjauhan. Dari 100, 500, 1000, dan 5000 iterasi menunjukan bahwasanya topik yang terbentuk berjumlah 3 topik dari 3 jumlah topik. 3 topik yang terbentuk memang sudah dapat dikategorikan sebagai 3 cluster yang berbeda. Akan tetapi untuk mengukur apakah 3 topik tersebut benar-benar 3 cluster yang berbeda, maka dilakukan pemodelan dengan jumlah topik 4 dan iterasi yang sama.

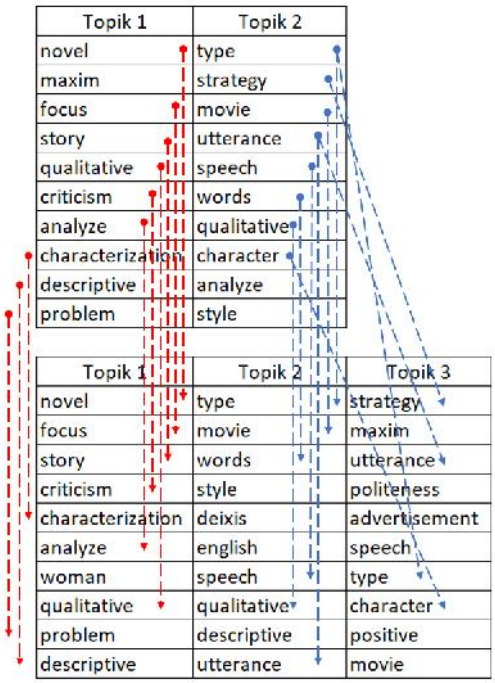
Untuk melihat apakah jumlah topik 3 merupakan topik yang cocok, maka penelitian ini mencoba melakukan pemodelan dengan jumlah topik berjumlah 4. Iterasi yang dilakukan pun sama dengan pemodelan sebelumnya. Hasil pemodelan untuk jumlah topik 4 untuk iterasi ke-100 menghasilkan 4 topik yang berbeda. Pada percobaan tersebut tidak ada topik yang bergabung. Begitu juga pada iterasi ke-500 dihasilkan 4 topik berbeda, namun ada 2 topik yang

p-ISSN: 2406-7768
e-ISSN : 2581-2181

berdekatan tetapi tidak beririsan. Selanjutnya dilakukan pemodelan pada iterasi ke-1000 dan didapatkan hasil 4 topik, dimana ada 2 topik yang beririsan. Sehingga dapat dikatakan 2 topik yang beririsan tersebut merupakan satu cluster. Sama halnya pada iterasi ke-5000 yang menghasilkan 4 topik yang berbeda dan 2 topik diantaranya beririsan. Dan untuk mengukur apakah 4 topik tersebut benar-benar 3 cluster yang berbeda, maka dilakukan pemodelan dengan jumlah topik 5 dan iterasi yang sama.

Iterasi yang dilakukan pun sama dengan pemodelan sebelumnya. Hasil pemodelan pada iterasi ke-100 menunjukkan 5 topik berbeda dengan 2 topik beririsan, namun ada 1 topik yang cenderung berdekatan terhadap topik yang beririsan. 3 topik tersebut mempunyai kemungkinan sebagai satu cluster yang sama. Untuk iterasi ke-500 dihasilkan 5 topik yang terpisah dengan 2 topik yang cenderung berdekatan. 2 topik yang cenderung berdekatan ini mempunyai kemungkinan berada dalam cluster yang sama. Untuk iterasi ke-1000 dihasilkan 5 topik yang berbeda. Dan untuk iterasi ke-5000 dihasilkan 5 topik yang berbeda. dimana terdapat 2 topik yang saling beririsan. 2 topik tersebut dapat dikatakan sebagai satu cluster yang sama. Untuk 2 topik yang berdekatan mempunyai kemungkinan berada dalam satu cluster yang sama.

Kesimpulan sementara yang dihasilkan pada pemodelan dengan jumlah topik 2, 3, 4, dan 5 dengan iterasi 100, 500, 1000, dan 5000 mengerucut bahwa jumlah topik 3 merupakan pemodelan topik yang fit. Namun untuk melihat apakah jumlah topik 3 merupakan jumlah yang fit, penelitian ini melakukan verifikasi terhadap stakeholder program studi sastra inggris uinsa. Hasil verifikasi dari stakeholder mengatakan bahwasanya jumlah topik 3 belum sesuai dengan pembagian topik sebenarnya. Pihak stakeholder mengatakan bahwa seharusnya ada 2 pembagian topik pada program studi uinsa. Berdasarkan hasil verifikasi pertama, dilakukan penambahan pemodelan dengan jumlah topik 2. Secara visual dapat dilihat bahwa isi dari topik 2 dan 3 memiliki pola yang dapat dilihat pada Gambar 3.



Gambar 3. Hasil analisis output kata-kata antara jumlah topik 2 dan 3

Dari hasil analisis tersebut dapat dilihat bahwa dalam pemodelan dengan jumlah topik 3, terdapat 2 topik yang merupakan pecahan bagian dari salah satu topik tersebut. Kemudian untuk membuktikan apakah benar ada 2 topik yang merupakan hasil pecahan dari satu toik, penelitian ini melakukan verifikasi tahap dua. Hasil verifikasi tersebut, pihak stakeholder mengatakan bahwa memang benar 2 topik tersebut adalah pecahan dari suatu topik. Jadi pemodelan dengan jumlah topik 3 merupakan pemodelan cluster yang terbaik diantara pemodelan dengan jumlah topik lainya. Hal tersebut dikarenakan tidak adanya topik yang beririsan dan saling berjauhan.

IV. KESIMPULAN

Proses implementasi topic modelling menggunakan metode Latent Dirichlet Allocation (LDA) pada data abstract skripsi Program Studi Sastra Inggris Universitas Islam Negeri Sunan Ampel Surabaya (UINSA) dimulai dari tahap pengambilan data. Data yang diperoleh berjumlah 584 abstract skripsi, data tersebut dipersiapkan melalui tahap pre-processing untuk mempermudah dalam topic modelling. Hasil dari pre-processing kemudian dihitung jumlah kemunculan setiap kata dengan model bag of words. Jumlah kemunculan setiap kata tersebut menjadi ukuran dalam metode Latent Dirichlet Allocation (LDA) untuk dimodelkan. Dalam metode LDA jumlah topik cluster dan jumlah iterasi ditentukan diawal. Percobaan yang dilakukan dengan mengubah jumlah topik

p-ISSN: 2406-7768
e-ISSN : 2581-2181

dan jumlah iterasi. Hasil dari pemodelan topik tersebut kemudian dilakukan analisa untuk melihat seberapa lazim kata tersebut dalam suatu topik. Pada penelitian ini dilakukan percobaan sebanyak 5 uji iterasi dengan iterasi berbeda yakni: 100, 500, 1000, dan 5000. Sedangkan terhadap setiap uji iterasi dimasukkan jumlah topik yang berbeda yaitu: 2, 3, 4, dan 5. Hasil cluster topik terbaik didapat pada jumlah topik 3. Hasil cluster tersebut telah diverifikasi oleh pihak stakeholder Program Studi Sastra Inggris (UINSA) bahwa kata-kata yang ada pada topik cluster sesuai dengan pembagian topik menurut konsentrasi pada Program Studi Sastra Inggris (UINSA).

REFERENSI

[1] Albert Verasius Dian Sano, (2019). "Cara Kerja Data Mining – Seri Data Mining For BusinesS Intelligence (3)," *Binus University*, 2019. [Online]. Available: https://binus.ac.id/malang/2019/01/cara-kerja-data-mining-seri-data-mining-for-business-intelligence-3/. [Accessed: 18-Jan-2020].

[2] E. Turban, J. E. Aronson, and T.-P. Liang, (2004). *Decision Support Systems and Intelligent Systems (7th Edition).*

[3] R. Diaz, "Pengertian Data Mining,Teks Mining,dan Web Mining.," (2013). [Online]. Available: http://yosephoriolryandiaz.blogspot.com/2013/03/pengertian-data-miningteks-miningdan.html. [Accessed: 18-Jan-2020].

[4] F. Ronen and J. Sanger, (2007). *The Text Mining Handbook: Advance Approaches in Analyzing Unstructured Data*. United States of America: Cambridge University Press.

[5] M. I. J. David M. Blei, Andrew Y. Ng, (2003). "Machine Learning Research 3," *Latent Dirichlet Alloc.*, pp. 993–1022.

[6] A. T. Jaka, (2015). "Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining," *J. Inform. UPGRIS*.

[7] D. S, P. Raj, and S.Rajaraajeswari, (2016). "A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction," *Int. J. Adv. Netw. Appl.*, pp. 320–323.

[8] K. P. Utami, (2017). "Analisis topik data media sosial twitter menggunakan model topik latent dirichlet allocation keke putri utami," .

[9] J. Kaur and P. K. Buttar, (2018). "A Systematic Review on Stopword Removal Algorithms," *Int. J. Futur. Revolut. Comput. Sci. Commun. Eng.*, vol. 4, no. 4.

[10] A. K. Ingason, S. Helgadóttir, H. Loftsson, and E. Rögnvaldsson, (2018). "A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI)," *Lect. Notes Artif. Intell.*, pp. 205–216.