ANALISIS SENTIMEN, PENILAIAN PEMILIHAN DAN TINGKAH LAKU PENGGUNA E-DAGANG MENGGUNAKAN LDA DAN PEMILIHAN FITUR

SITI KHADIJAH BINTI JASNI

UNIVERSITI KEBANGSAAN MALAYSIA

ANALISIS SENTIMEN, PENILAIAN PEMILIHAN DAN TINGKAH LAKU PENGGUNA E-DAGANG MENGGUNAKAN LDA DAN PEMILIHAN FITUR

SITI KHADIJAH BINTI JASNI

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEHI
IJAZAH SARJANA SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2021

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

05 November 2021

SITI KHADIJAH BINTI JASNI
P102456

PENGHARGAAN

Terlebih dahulu saya ingin mengucapkan syukur Alhamdulillah ke hadrat Allah S.W.T kerana di atas limpah dan kurniaNya, maka dapatlah saya menyiapkan tesis ini dengan jayanya walaupun menempuhi pelbagai cabaran. Setinggi-tinggi penghargaan dan ucapan terima kasih ditujukan kepada Prof Madya Dr Zakree Ahmad Nazri, selaku penyelia yang telah banyak membantu dan memberikan tunjuk ajar, perhatian, semangat dan nasihat di dalam menjalankan kajian disertasi ini. Segala nasihat, semangat dan strategi beliau telah banyak mengajar saya untuk menjadi seorang penyelidik yang baik. Tidak dilupakan juga kepada pensyarah-pensyarah di FTSM yang turut membantu sedikit sebanyak dalam pencarian bahan-bahan dan maklumat-maklumat berkaitan kajian ini.

Ucapan terima kasih yang tidak terhingga saya dedikasikan buat ayah dan arwah ibu tercinta, Jasni bin Jaafar dan Lily Zakiah binti Mohd Daud serta seluruh ahli keluarga tersayang kerana memberi semangat dari segi sokongan fizikal serta mental. Tidak dilupakan juga kepada suami, Shazwan Hakim bin Muhamad kerana banyak memberikan sokongan moral dan galakkan dikala berputus asa menyiapkan penulisan ini. Tidak ketinggalan juga kepada rakan-rakan seperjuangan yang banyak memberikan sokongan dan peransang. Akhir kata, saya ingin merakamkan jutaan terima kasih kepada semua pihak yang terlibat secara langsung dan tidak langsung dalam membantu saya menghasilkan penulisan ini

ABSTRAK

Amazon.com mula mendapat permintaan tinggi dalam kategori peruncitan sejak dunia dilanda krisis pandemik koronavirus di mana ramai semakin bergantung kepada pembelian secara dalam talian berbanding sebelum ini. Maka, memahami bagaimana pengguna membuat keputusan ketika membeli-belah dalam talian telah menjadi subjek penting kepada industri e-dagang kerana keputusan pembelian pengguna secara langsung akan mempengaruhi penjualan barangan. Keadaan ini telah menjadikan satu keperluan kepada Amazon.com untuk mendapatkan maklumat terperinci mengenai kehendak pengguna dan membentuk produk yang dapat memenuhi permintaan mereka seperti produk kopi. Lantaran itu, objektif kajian ini adalah mengekstrak ciri-ciri produk, pola pilihan dan tingkah laku pengguna terhadap suatu produk dari teks dengan membangunkan kaedah permodelan topik berasaskan *Latent Dirichlet Allocation* (LDA). Selain itu, kajian ini juga turut membangunkan model fitur berasaskan jenis kata untuk meningkatkan ketepatan model LDA dan mengoptimumkan model ini dengan mencari jumlah topik yang paling relevan, *K*. Dengan menggunakan ulasan-ulasan pengguna di Amazon.com, kajian ini menerapkan pendekatan berdasarkan permodelan topik-sentimen iaitu LDA dan Analisis Sentimen VADER untuk mengekstrak ciri-ciri produk dan menemui pilihan dan tingkah laku pengguna terhadap produk kopi dengan menggunakan data ulasan pelanggan dan metadata produk yang diekstrak oleh Ni et al. (2019). Kajian ini melihat kepada pemilihan dan tingkah laku pengguna berdasarkan pandangan mereka terhadap lebih daripada 16,275 produk kopi yang dijual di laman e-dagang Amazon.com. Hasil penemuan kajian ini menunjukkan bahawa Kata Nama, Kata Kerja dan Kata Keterangan adalah kombinasi terbaik jenis kata dan bilangan topik *K* yang optimum ialah 6. Ciri-ciri kopi yang diekstrak daripada model ini adalah ‘Kualiti Kopi’, ‘Pembungkusan’, ‘Harga’, ‘Khidmate Pelanggan’, ‘Rasa Kopi’ dan ‘Biji Kopi’ dan semua ciri-ciri ini adalah ciri positif kerana sentimen positif mereka mempunyai bilangan ulasan tertinggi berbanding sentimen negatif dan neutral. Dengan hasil kajian ini, pengurus perniagaan yang besar mahupun yang kecil dapat memanfaatkan ulasan tekstual pelanggan untuk lebih memahami keperluan mereka dengan menyediakan perkhidmatan yang lebih baik serta dapat meningkatkan penjualan mereka dengan memperbaiki tajuk, deskripsi dan ciri-ciri produk dengan menggunakan kata kunci atau istilah carian yang lebih relevan.

SENTIMENT ANALYSIS, SELECTION EVALUATION AND BEHAVIOR OF E-COMMERCE USERS USING LDA AND FEATURE SELECTION

ABSTRACT

Amazon.com has started to get high demand in the retail category since the world was hit by the coronavirus pandemic crisis where many are increasingly relying on online purchases than ever before. Thus, understanding how consumers make decisions when shopping online has become an important subject to the e-commerce industry as consumers' purchasing decisions will directly influence the sale of goods. This situation has made it a necessity for Amazon.com to obtain detailed information on consumers' needs and formulate products that can meet their demands such as coffee products. Therefore, the objective of this study is to extract product characteristics, choice patterns and consumer behavior towards a product from the text by developing a topic modeling-based method, Latent Dirichlet Allocation (LDA). In addition, the study also developed Part of Speech tagging-feature model to improve the accuracy of the LDA model and optimize this model by finding the most relevant number of topics, K. This study applied a topic-sentiment modeling approach which is LDA and VADER Sentiment Analysis to extract product characteristics and discover consumer preferences and behaviors over 16,275 of coffee products using customer review data and product metadata from Amazon.com which was extracted by Ni et al. (2019). The results of this study showed that Nouns, Verbs and Adverb are the best combination of word types and the optimal number of topics, *K* is 6. The characteristics of coffee extracted from this model are 'Coffee Service', 'Coffee Quality', 'Price', ' Coffee Bean', 'Coffee Flavor' and 'Packaging' and all these features are positive features because their positive sentiments have the highest number of reviews compared to negative and neutral sentiments. With the results of this study, business managers and sellers can leverage their customers' reviews to better understand their needs by providing better services and to increase their sales by improving the product titles, descriptions and features by using keywords or search terms that are more relevant.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

BAB I

PENGENALAN

1.1 PENDAHULUAN

Penjualan barangan runcit dalam talian di Amazon.com melonjak tiga kali ganda pada suku kedua disebabkan oleh pesanan dari rumah dan penularan wabak COVID-19 (Browne 2020). Apabila wabak COVID-19 mula merebak menjadi lebih serius pada Mac 2020, permintaan barangan runcit semakin meningkat di Amazon.com di mana orang ramai semakin bergantung kepada pembelian dalam talian berbanding tahun sebelumnya. Dalam pendapatan suku kedua syarikat peruncitan gergasi ini, pihak eksekutif syarikat menjelaskan bahawa platform e-dagang, Amazon.com telah menyaksikan lonjakan kepada perdagangan dan peruncitan dalam talian berbanding suku sebelumnya (Browne 2020). Dengan peningkatan pembelian barangan basah dan makanan dalam talian, maklumat terperinci mengenai faktor-faktor yang mempengaruhi pilihan produk makanan dalam talian oleh pengguna menjadi keperluan penting kepada peniaga. Berbeza dengan kaji selidik yang kebiasaannya digunakan bagi mendapatkan maklumat mengenai pemilihan dan pengalaman pengguna yang dimanipulasi oleh penyelidik dan syarikat, pandangan dan komen yang jujur oleh pelanggan dalam talian juga menyajikan kita dengan maklumat yang tidak berat sebelah terhadap sesuatu produk. Oleh yang demikian, dengan menggunakan kepelbagaian pandangan dan kritikan yang ditulis di ruangan komen pelanggan di Amazon.com, kajian ini telah menerapkan pendekatan berdasarkan model sentimen-topik untuk mengetahui dan memahami pemilihan dan tingkah laku pengguna terhadap produk kopi yang terdapat di platform Amazon.com.

1.2 LATAR BELAKANG KAJIAN

Perniagaan dalam talian telah menjadi saluran penting untuk penjualan pelbagai produk. Pembelian barangan dalam talian seperti barangan runcit segar yang siap dibungkus kemas telah meningkat sebanyak 15 peratus sejak akhir dua tahun ini (Nielsen 2018). Menurut laporan oleh syarikat pengukuran global Nielsen, pembelian kategori yang lain juga turut meningkat ekoran daripada peningkatan keyakinan pengguna terhadap ekosistem membeli-belah dalam talian (Nielsen 2018). Di samping itu, penjualan e-dagang pada suku kedua tahun 2020 telah menyumbang sebanyak 16.1 peratus daripada jumlah keseluruhan penjualan di Amerika Syarikat (AS). Anggaran penjualan e-dagang runcit AS adalah berjumlah $200.7 bilion untuk suku kedua pada tahun 2020 di mana 37.0 peratus (± 1.2%) lebih daripada suku pertama 2020 (US Census Bureau 2020). Menjelang 2021, penjualan di laman web akan menjejaskan 46 peratus penjualan luar talian dan penjualan yang dipengaruhi laman web akan berkembang pada kadar pertumbuhan tahunan kompaun (CAGR) iaitu sebanyak 8.5% dari 2016 hingga 2021 (Forrester 2016).

Tambahan pula, dengan ancaman virus COVID-19 yang tinggi dan setiap negara mengarahkan semua penduduk untuk kekal berada di rumah, pembelian barangan keperluan secara dalam talian secara tidak langsung menjadi lebih menarik kepada banyak pengguna. Menurut kaji selidik runcit dalam talian oleh US Coresight Research, e-dagang runcit akan berkembang sebanyak 40 peratus pada tahun 2020 untuk mencapai $ 38 bilion dalam jualan, setara dengan 3.5 peratus dari keseluruhan pasaran (Research 2020). Di samping itu, pemerolehan *Whole Foods Market* baru-baru ini oleh Amazon.com adalah contoh terbaik mengenai potensi membeli makanan dalam talian dan membeli-belah di masa hadapan. Oleh yang demikian, memahami bagaimana pengguna membuat keputusan ketika membeli-belah dalam talian telah menjadi subjek penting kepada penyelidik dan syarikat e-dagang kerana keputusan pembelian pengguna secara langsung akan mempengaruhi penjualan barangan.

Justeru, matlamat kajian ini adalah untuk membangunkan kaedah atau alatan cerdas untuk pemain industri untuk meneliti dan memberi tumpuan kepada pelbagai topik yang dapat menafsirkan pandangan dan pendapat pengguna dalam talian mengenai suatu produk seperti kopi untuk mengetahui pilihan dan tingkah laku

pengguna kopi terhadap produk kopi yang terdapat di platform e-dagang Amazon.com Kajian ini juga akan mengeluarkan dan mentafsirkan maklumat yang diperlukan untuk menaikkan taraf platform pandangan dan pendapat pelanggan agar memberi manfaat kepada pembeli dalam talian malah mengurangkan masa mereka dalam mencari produk yang relevan untuk keperluan mereka. Di samping itu, bukan sahaja pelanggan mendapat keuntungan malah kajian diharapkan dapat memberi sumbangan serba sedikit memberi kefahaman kepada penjual dan syarikat tentang kepentingan maklum balas pengguna untuk memperbaiki sistem pasaran mereka, termasuklah memulihkan perkhidmatan dan produk, audit fungsi dalaman, dan meningkatkan kualiti produk.

1.2.1 Pasaran Kopi

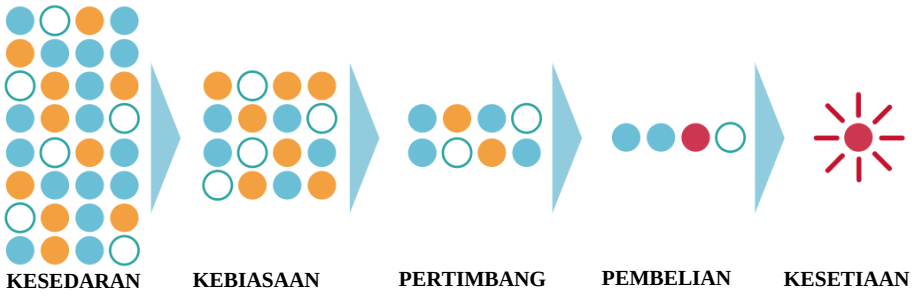
Wabak koronavirus telah mempengaruhi hampir setiap industri termasuklah industri kopi. Buat masa sekarang, kopi telah menjadi produk kedua terpenting dalam perdagangan antarabangsa selepas minyak dan menurut H.R. Neumann dari Kumpulan Neumann Kaffe, permintaan kopi di peringkat dunia akan meningkat daripada 144 juta beg pada tahun 2015 hingga ke 200 juta beg menjelang tahun 2030 (Wróblewski & Mokrysz 2017). Keadaan ini sedikit sebanyak mempengaruhi peningkatan persaingan di pasaran dan penawaran produk menjadi semakin luas. Kepelbagaian produk kopi yang ditawarkan di pasaran sering menyukarkan pengguna dalam membuat pilihan kopi semasa membeli. Keputusan dalam pembelian kopi dipengaruhi oleh beberapa faktor yang kompleks seperti kualiti produk, pasaran harga, kaedah penyediaan dan kemudahan mendapatkan kopi tersebut (Wróblewski & Mokrysz 2017).

Antara faktor lain yang patut dipertimbangkan oleh pengguna kopi dalam proses membuat keputusan adalah jenis dan jenama kopi. Berdasarkan penulisan Wróblewski dan Mokrysz (2017), pemilihan pengguna kopi di Poland bergantung kepada jenis kopi (kopi tanah, kopi segera, biji kopi dan *cappuccino* segera) dan beberapa jenama pilihan biji kopi. Daripada hanya memberi tumpuan kepada pasaran kopi di negara tertentu sahaja, kajian ini meneliti dan mengkaji lebih mendalam mengenai pemilihan dan tingkah laku pengguna kopi di dunia melalui Amazon.com.

1.2.2 Definisi Dan Konsep Tingkah Laku Pengguna

Konsep tingkah laku pengguna muncul dalam sastera Barat secara besar-besaran pada pertengahan tahun 1960-an dan Kaufman (1995) mendefinisikan tingkah laku pengguna sebagai setiap tingkah laku manusia sama ada di rumah, di tempat kerja, di kedai mahupun di jalan, iaitu di mana sahaja mereka berada untuk melakukan aktiviti berkaitan dengan membeli sama ada memikir untuk membeli, melakukan pembelian, atau menggunakan produk yang dibeli (Kaufman 1995). Persatuan Pemasaran Amerika (AMA) mendefinisikan tingkah laku pengguna sebagai interaksi dinamik kognitif, tingkah laku dan keadaan persekitaran di mana manusia membuat pilihan dalam aspek kehidupan mereka dengan mencuba pelbagai pemboleh ubah sosial dan psikologi (Barmola & Srivastava 2010). Tingkah laku pengguna adalah reaksi-reaksi koheren yang berkaitan dengan membuat pilihan dalam proses memenuhi keperluan individu yang bergantung kepada keadaan ekonomi, sosial dan budaya tertentu (Solomon et al. 2012). Aktiviti-aktiviti tersebut yang memenuhi keperluan pengguna adalah pencarian, pembelian, penggunaan dan penilaian barang dan perkhidmatan (Priest et al. 2013).

Pemasaran mempunyai satu matlamat iaitu menjangkau pengguna pada saat "detik-detik yang penting" atau "titik sentuhan"di mana mereka senang dipengaruhi dalam membuat keputusan membeli (Stankevich 2017). Rajah 1.1 menunjukkan metafora corong tradisional iaitu ketika pengguna bermula dengan beberapa jenis jenama yang berpotensi dan mengurangkan jumlahnya secara teratur untuk membuat pembelian (Court et al. 2009). Walau bagaimanapun, konsep corong ini tidak lagi dapat menangkap semua titik sentuh dan faktor pembelian utama disebabkan oleh lambakan jenis produk dan saluran digital serta evolusi pengguna yang semakin mudah mendapat maklumat. Oleh yang demikian, pendekatan yang lebih berkesan diperlukan untuk membantu pemasar mengemudi persekitaran ini yang dikatakan lebih rumit daripada yang dicadangkan oleh corong ini (Stankevich 2017).



Rajah 1.1 Metafora corong tradisional

Berdasarkan kajian yang lepas, banyak kajian telah melakukan penyelidikan mendalam dan mengembangkan pelbagai teori dan model dalam membuat keputusan. Model tradisional proses pembuatan keputusan lima peringkat berfungsi sebagai asas untuk konsep moden seperti model McKinsey pada tahun 2009 (Court et al. 2009) yang telah menerima pelbagai kritikan. Walaupun begitu, tidak dinafikan kepentingan model tersebut. Dengan menggunakan model tradisional sebagai landasan atau rangka kerja yang mempengaruhi "detik-detik yang penting" dalam membuat keputusan dan faktornya dibangunkan dan dibuktikan. Perkara ini dapat membantu dalam pengembangan penyelidikan dengan lebih mendalam sama ada dari segi pengesahan atau penolakan mengenai hubungan ini.

Walau bagaimanapun, kajian-kajian tersebut masih mempunyai jurang yang perlu dikaji memandangkan pergerakan budaya globalisasi sentiasa berubah seiring dengan kemajuan teknologi. Oleh yang demikian, para pengkaji mula mengorak langkah dan berusaha mencari jawapan yang lebih kritis bagi memahami tugas mengawal maklumat mengenai keputusan pengguna kerana tugasan ini dapat membawa kepada peningkatan prestasi. Pandangan baru ini berpotensi menjadi sangat berguna dalam alaf digital ini di mana aktiviti mengawal aliran maklumat memberi pengaruh signifikan terhadap kualiti keputusan, ingatan, pengetahuan dan keyakinan pengguna. Penyelidikan struktur maklumat (jumlah maklumat dalam set pilihan) juga relevan di pasaran elektronik baru, di mana pengguna sering berhadapan dengan lebihan maklumat semasa membuat keputusan.

1.2.3 Peranan Pandangan Pengguna

Apabila berlakunya peningkatan dalam persaingan produk dan maklumat yang terlalu banyak berkemungkinan menyukarkan pengguna dalam talian untuk membuat pilihan yang diperlukan. Laman web peruncitan dalam talian seperti Amazon.com menyediakan satu platform untuk pengguna mengemukakan ulasan untuk berkongsi pendapat dan pengalaman mereka terhadap produk yang telah dibeli dan digunapakai. Kualiti kaji selidik dalam talian (faktor sistematik) yang bercirikan pandangan maklumat dan pandangan persuasif mempunyai kesan yang signifikan terhadap niat membeli sesuatu barangan (Zhang et al. 2014). Menurut Heng et al. (2018), walaupun pengguna pada amnya mempunyai segala maklumat terhadap satu produk yang ingin dibeli, namun paparan ratusan ribu ulasan daripada pengguna lain yang mengandungi pelbagai pendapat dan maklumat yang tidak konsisten akan menyukarkan pengguna untuk menggunakan maklumat tersebut untuk membuat pilihan terbaik. Oleh sebab itu, platform e-dagang seperti Amazon.com sering mengetengahkan maklumat berharga dengan memberikan penilaian, kedudukan produk dan meletakkan aspek "membantu" untuk memudahkan para pengguna menilai ulasan pengguna lain bagi mengurangkan masa mereka dalam mencari maklumat berguna. Setiap keseluruhan ulasan atau pandangan, kedudukan produk, penilaian bermanfaat dan pendapat pengguna dapat memberikan maklumat yang begitu banyak untuk memudahkan penyelidik mengkaji faktor-faktor yang dapat mempengaruhi keputusan pembelian pengguna ketika membeli-belah dalam talian.

Terdapat banyak kajian telah mengkaji tentang kepentingan pendapat dan pandangan dalam talian dalam mempengaruhi keputusan pembelian pengguna. Kajian oleh Heng et al. (2019), Mou et al. (2019), Heng et al. (2018), Chen et al. (2019) dan Jauhari et al.(2020) serta banyak lagi telah menggunakan pendekatan pemodelan topik untuk mengkaji dengan lebih terperinci mengenai topik tersembunyi yang disembunyikan daripada pendapat dan pandangan dalam talian bagi membantu pembeli yang berpotensi melakukan keputusan pembelian yang bijak.

1.3 PERMASALAHAN KAJIAN

Seiring dengan pertumbuhan Internet dan perdagangan elektronik, pandangan dan pendapat pengguna dalam talian menjadi sumber maklumat penting yang mempengaruhi keputusan pembelian pengguna. Maklumat teks yang terkandung dalam ulasan adalah lebih penting kepada pengguna daripada panjang ulasan atau penilaian keseluruhan produk (Stankevich 2017). Mengabaikan kandungan teks ulasan pengguna adalah kelemahan utama kajian-kajian sedia ada mengenai sistem cadangan (McAuley & Leskovec 2013). Selain itu, menurut Heng et al. (2018), setiap ulasan pengguna adalah maklumat teks yang berdimensi tinggi dan pelbagai topik yang dapat ditafsirkan secara laten dapat membantu pelanggan mencari produk yang paling relevan. Contohnya, Chen et al. (2019) menunjukkan bahawa pendapat dan pandangan dalam talian yang digunakan dalam kajiannya dengan menggunakan LDA dapat dikelompokkan menjadi tiga topik iaitu ekspresi subjektif dan penggunaan memasak, ciri produk, dan penggunaan kecantikan. Dalam hal ini, tinjauan dimensi tinggi dapat diproses dengan pendekatan perlombongan topik untuk mengekstrak topik laten yang berdimensi rendah.

LDA adalah teknik permodelan yang agak kompleks tetapi fleksibel. LDA dianggap mempunyai varians yang sangat tinggi dan bias yang rendah. Ini bermaksud LDA memerlukan banyak data untuk mempelajari sesuatu yang bermakna. Jika tidak mempunyai cukup kata dalam dokumen, maka tidak mempunyai cukup data untuk menyimpulkan sebaran topik yang dapat dipercayai untuk dokumen tersebut. Teks atau korpus ulasan pengguna yang digunakan dalam analisis menggunakan pembelajaran mesin secara umumnya adalah Bag of Word (BoW). Walau bagaimanapun, data input berasaskan BoW adalah tidak mencukupi meningkatkan prestasi pembelajaran mesin dalam menentukan polariti atau lain-lain analisis dokumen dalam talian. Oleh itu, teknik seperti LDA memerlukan input dalam fitur atau ciri yang lebih spesifik sehingga mampu memberikan keputusan yang lebih baik. Pemilihan atau padanan golongan kata (*Part-of-Speech (POS)*) yang diekstrak dari teks secara hipotetikal, boleh memberikan skor ketepatan lebih tinggi berbanding daripada dokumen tanpa proses pemilihan ciri POS. Dengan menggunakan beberapa set ciri berasaskan POS, maka kesan penggunaan set golongan kata seperti kata sifat atau kata kerja dapat dikaji. Antara objektif kajian

ini adalah untuk membangunkan teknik pemilihan fitur berasaskan POS untuk mendapatkan set golongan kata paling berkesan. Hasil proses fitur berasaskan POS akan menjadi input untuk proses analisis sentimen dengan menggunakan kaedah LDA.

1.4 PERSOALAN KAJIAN

1. Adakah *Latent Dirichlet Allocation* (LDA) berkesan dalam mengekstrak ciri produk, pola pilihan dan tingkah laku pengguna terhadap suatu produk dari teks?

2. Apakah set golongan kata yang dapat meningkatkan prestasi *Latent Dirichlet Allocation* (LDA)?

3. Apakah topik terpendam (tersembunyi) yang berdasarkan pandangan dan pendapat pelanggan dalam talian dan bagaimana topik ini dapat mempengaruhi keputusan mereka dalam talian?

4. Bagaimana mengoptimumkan model LDA?

1.5 OBJEKTIF KAJIAN

1. Membangunkan kaedah permodelan topik berasaskan *Latent Dirichlet Alloaction* (LDA) untuk mengekstrak ciri produk, pola pilihan dan tingkah laku pengguna terhadap suatu produk dari teks.

2. Untuk membangunkan model fitur berasaskan jenis kata untuk meningkatkan ketepatan model LDA.

3. Untuk mengoptimumkan model *Latent Dirichlet Allocation* (LDA) dengan mencari jumlah topik yang paling relevan, *K*.

1.6 SKOP KAJIAN

Kajian ini akan meneliti dan memberi tumpuan untuk mengenalpasti jumlah topik yang optimal, dikeluarkan daripada pandangan dan pendapat pelanggan dan menggunakan topik-topik ini seperti ciri-ciri kopi untuk mengetahui pilihan pengguna kopi dan tingkah laku mereka terhadap produk kopi yang terdapat di platform e-dagang Amazon.com. Selain itu, kajian ini turut menggunakan data daripada pandangan dan pendapat pengguna Amazon.com dan metadata produk ke atas barangan runcit dan ia akan menggunakan data ulasan pelanggan Amazon.com dan metadata produk pada barangan runcit dan produk makanan gourmet dari Februari 2007 sehingga Oktober 2018 yang dikumpulkan oleh Ni et al. (2019).

Terdapat lebih daripada 5,074,160 ulasan yang berbeza dan 287, 209 jenis produk tetapi kajian hanya memberi tumpuan kepada produk kopi sahaja. Selain itu, kajian ini akan menggunakan model *Latent Dirichlet Allocation* untuk meneroka topik yang paling penting dan menggunakan skor koheran (*Topic Coherence Score)* untuk menjustifikasikan pemilihan model secara kuantitatif di mana ia menggabungkan beberapa langkah ke dalam rangka kerja untuk menilai koheren antara topik yang disimpulkan oleh model. Seterusnya, selepas memperoleh topik, kajian ini turut membincangkan lebih terperinci mengenai kegunaan Analisis Sentimen untuk menentukan ciri-ciri kopi dengan cara yang signifikan dalam mengaruhi niat pembelian pelanggan.

1.7 SIGNIFIKAN KAJIAN

Memahami dan mengetahui kehendak pelanggan adalah kunci utama kejayaan setiap perniagaan, sama ada menjual secara langsung atau tidak langsung kepada individu atau perniagaan lain. Setelah para peniaga mempunyai pengetahuan sebegini, amat mudah bagi mereka menggunakannya kajian ini untuk meyakinkan pelanggan yang berpotensi dan sedia ada dalam menggunakan perkhidmatan mereka adalah keputusan yang terbaik bagi memenuhi kehendak pelanggan. Di samping itu, kajian ini juga berharap dapat akan menghubungkan ulasan tekstual pelanggan dalam talian dengan persepsi pelanggan untuk membantu pengurus perniagaan lebih memahami keperluan pelanggan dengan lebih baik. Dengan adanya maklumat sebegini, pengkaji berharap dapat

memberikan sumbangan yang berguna dalam meningkatkan perniagaan syarikat dari segi penambahbaikan produk.

Selain itu, kajian ini juga berharap dapat membantu perniagaan kecil atau sederhana seperti peniaga yang berdagang dalam Amazon.com untuk meningkatkan penjualan mereka dengan memperbaiki tajuk, deskripsi dan ciri-ciri dengan menggunakan kata kunci atau istilah carian yang lebih relevan dan menyediakan perkhidmatan yang lebih baik kepada pembeli yang sedia ada dan yang berpotensi. Akhir sekali, lebih ramai pengguna akan menggunakan platform membeli-belah dalam talian yang memiliki sistem mesra pengguna terutama bagi pengguna baru di mana mereka dapat mengakses dan menyelusuri platform dengan mudah dan menapis beribu-ribu produk dalam mencari produk relevan dalam masa yang singkat.

BAB II

TINJAUAN LITERATUR

2.1 PENGENALAN

Sejak kebelakangan ini, kemajuan teknologi Internet dan Web yang sedang bergerak laju telah mempercepat perkembangan perdagangan elektronik dan menempatkan perniagaan dan pelanggan dalam norma yang baru. Perusahaan telah membangunkan portal perniagaan baru dan menyediakan sejumlah besar maklumat produk untuk memperluas pasaran mereka dan mewujudkan lebih banyak peluang bisnes di mana pelanggan mempunyai lebih banyak peluang untuk memilih pelbagai produk bagi memenuhi keperluan mereka (Picard 2000).

Walaupun maklumat dalam pemasaran Internet dan perdagangan elektronik tidak menentu, maklumat daripadanya tetap mempunyai potensi dan memberi kesan kepada pengguna. Hal ini demikian kerana Internet dan perdagangan elektronik telah memperluaskan kemungkinan untuk penjenamaan, inovasi, harga, dan penjualan (Ahmad et al. 2018). Walau bagaimanapun, pertumbuhan maklumat yang eksponensial dan digabungkan dengan pengembangan laman web perniagaan yang pesat telah menyebabkan maklumat diedarkan secara berlebihan. Oleh yang demikian, pengguna telah menghabiskan terlalu banyak masa melayari internet untuk mencari maklumat yang mereka perlukan. Salah satu jalan penyelesaian untuk masalah yang disebutkan di atas adalah dengan mencipta satu sistem cadangan mudah iaitu menyediakan perkhidmatan maklumat diperibadikan yang mengumpul segala maklumat produk yang diperlukan oleh pengguna dalam membantu mereka membuat keputusan pembelian (Schafer et al. 2001). Tujuan perkhidmatan maklumat yang diperibadikan adalah untuk menyesuaikan strategi promosi dan iklan agar seiring dengan minat pelanggan (Cao & Li 2007).

Dengan menyediakan perkhidmatan pemperibadian pelanggan dan berkomunikasi atau berinteraksi dengan pelanggan, perniagaan baru boleh berurusan dengan pengalaman Web pelanggan khusus. Pemahaman pelanggan seperti itu dapat digunakan untuk mengubah maklumat pelanggan menjadi perkhidmatan atau produk yang berkualiti tinggi (Sung-Shun & Mei-Ju 2004). Pemasaran satu-antara-satu dianggap sebagai pendekatan yang paling berkesan untuk pengurusan perhubungan pelanggan dari segi meningkatkan kepuasan pelang12gan, kadar maklum balas, kesetiaan, jualan Web, dan reputasi. Namun begitu, bagaimana pula dengan perniagaan yang ingin mengenal pasti minat pelanggan mereka yang mempunyai jumlah pelanggan yang tinggi? Membina perkhidmatan Internet yang diperibadikan adalah jawapan kepada permasalahan ini. Matlamat pemperibadian adalah untuk menyesuaikan strategi promosi dan iklan dengan pilihan pelanggan (Cao & Li 2007). Sesuatu perniagaan perlulah memahami terlebih dahulu minat dan pilihan pelanggan sebelum memberikan produk atau perkhidmatan yang sesuai pada waktu yang bersesuaian. Kajian Cao & Li (2007) dapat meningkatkan jumlah orang yang mengunjungi kedai Web, peluang penjualan dan pendapatan iklan dan juga keuntungan laman web.

2.2 SISTEM PENGESYORAN PRODUK

Pelbagai sistem pengesyoran yang diperibadikan dapat dikembangkan untuk membantu pengguna melalui ruang ciri produk yang besar tetapi bergantung kepada jenis produk. Sistem cadangan untuk produk yang sering dibeli dapat dikembangkan dengan menganalisis maklumat peribadinya, sejarah penyemakan imbas, dan produk yang dibelinya (Gao et al. 2010). Namun untuk pengguna biasa yang kurang kerap membeli, perusahaan tidak mempunyai cukup maklumat mengenai pembelian masa lalu pelanggan dan keperluan spesifiknya untuk produk tertentu dan menyebabkan pilihan pelanggan menjadi sukar dan mustahil (Cao & Li 2007). Nasihat pakar domain sangat diperlukan dalam situasi ini. Oleh yang demikian, sistem cadangan perlu mempunyai pengetahuan domain tertentu serta kemampuan untuk berinteraksi dengan pengguna. Justeru itu, sistem dapat memperoleh dan menganalisis keperluan semasa pelanggan terhadap beberapa jenis produk yang telah dikenal pasti, dan kemudiannya menilai produk yang relevan untuk membantunya dalam membuat pemilihan yang terbaik.

Dengan sistem cadangan yang diperibadikan, pengguna dapat mengakses maklumat yang mereka inginkan dengan cepat sambil menjimatkan masa mereka membaca dokumen elektronik. Namun demikian, perusahaan dapat belajar tentang kebiasaan membeli pelanggan mereka dan kemudiannya mengembangkan strategi pemasaran yang paling tepat untuk menarik pelanggan lain dan menyampaikan maklumat yang mereka cari dengan tepat. Oleh yang demikian, kepuasan dan kesetiaan pelanggan dapat ditingkatkan dan peningkatan frekuensi kunjungan pelanggan dapat menciptakan lebih banyak peluang transaksi dan menguntungkan perusahaan Internet (Gummerus et al. 2012).

Kebiasaan membeli dan data demografi pelanggan masa lalu dapat digunakan untuk meramalkan kebiasaan pembelian masa depannya di laman web e-commerce. Oleh yang demikian, produk yang dipersonalisasi dapat disarankan kepada pelanggan ekoran penukaran daripada pengguna menjadi pelanggan dan peningkatan kesetiaan pelanggan dan penjualan silang (Grbovic et al. 2015). Selain itu, sistem cadangan boleh menyediakan perkhidmatan maklumat peribadi dalam pelbagai cara bergantung kepada sama ada sistem tersebut sebelumnya telah merekod dan menganalisis pilihan pelanggan (Le & Liaw 2017). Dalam jenis sistem cadangan peribadi yang pertama, maklumat peribadi pelanggan dikumpulkan terlebih dahulu, kemudiannya sistem menjelaskan pilihan pelanggan dengan menganalisis dan memodelkan maklumat peribadi yang ada (Grbovic et al. 2015). Setelah maklumat peribadi pengguna diperoleh, sistem cadangan kemudiannya dapat membina model komputerisasi untuk mencadangkan kepada pengguna pilihan item yang lain dari domain aplikasi yang sama. Malah, karya cadangan boleh dianggap sebagai klasifikasi dengan menggunakan maklumat yang sudah diketahui untuk menyusun model bagi meramalkan kejadian yang tidak diketahui (Cao & Li 2007).

Berbeza dengan sistem di atas yang berkaitan dengan pilihan pengguna sebelumnya, jenis sistem cadangan yang diperibadikan dirancang untuk produk yang kurang kerap dibeli dan keperluan khusus pengguna dalam satu transaksi (Hsu et al. 2012). Tambahan pula, pengguna memerlukan pengetahuan domain khusus untuk menilai kualiti produk yang sesuai (Cai & Zhu 2015). Daripada memodelkan pilihan masa lalu pelanggan, sistem cadangan mencari maklumat singkat yang diberikan oleh

pengguna ketika berunding dengan sistem untuk mendapatkan cadangan dan pengetahuan pakar yang terbaik mengenai produk tersebut (Kim et al. 2017). Terlebih dahulu, sistem cadangan mengambil beberapa produk dari pangkalan data dengan mengira persamaan antara produk di dalam pangkalan data dan ciri produk sasaran. Pengguna kemudiannya dapat menyesuaikan keperluannya berdasarkan ciri khas produk yang disarankan dan meminta agar sistem mencadangkan item yang baru. Dengan cara ini, pengguna dapat menemui produk terbaik yang memenuhi kehendaknya.

2.3 PEMODELAN TOPIK DAN ANALISIS TEKS DALAM PEMASARAN

Berdasarkan kajian lepas, soal selidik, panel pelanggan, dan pandangan pakar telah digunakan untuk menghasilkan pendapat pelanggan bagi tujuan perniagaan dan pemasaran. Kaedah ini dikatakan terlalu mahal dan tidak selalu mencapai bilangan responden yang besar. Pemantauan media dan penyelidikan manual media sosial, blog, dan sumber data luaran lain dapat mendedahkan perkembangan yang sedang meningkat atau masalah yang penting bagi khalayak ramai. Mencari isyarat dalam jumlah besar data teks tanpa memerlukan tenaga kerja manual dapat memberi manfaat kepada pengguna tentang perkembangan pasaran. Peningkatan sumber dan kemudahan mengakses perisian pengkomputeran yang lebih besar dan digabungkan juga dengan kaedah analisis teks menjadikannya lebih mudah dan senang untuk mengembangkan aplikasi (Chakraborty 2014).

Analisis teks adalah istilah luas yang merangkumi kaedah untuk menarik maklumat dari teks dan menukar teks menjadi data yang berguna untuk analisis (Sarkar 2016). Ini termasuk pengambilan dokumen, pendanaan bahagian ucapan, pemodelan topik dan analisis sentimen. Pemodelan topik adalah kaedah analisis teks utama yang dikaji dan digunakan dalam projek ini. Kaedah ini digunakan untuk klasifikasi tanpa pengawasan data teks bagi mengekstrak topik terpendam atau tersembunyi dari teks (Ibrahim & Wang 2019). Penguraian teks menjadi topik mengubah data teks tidak berstruktur menjadi format yang dapat digunakan untuk pemeriksaan yang lebih luas dari sekumpulan dokumen, pengindeksan dokumen, pengelompokan dokumen, dan analisis semantik. Kaedah pemodelan topik yang paling biasa digunakan adalah *Latent*

Dirichlet Allocation (LDA) yang dikembangkan oleh Blei, Ng dan Jordan (2003) yang akan menjadi fokus projek ini.

2.3.1 *Latent Dirichlet Allocation* (LDA) Model

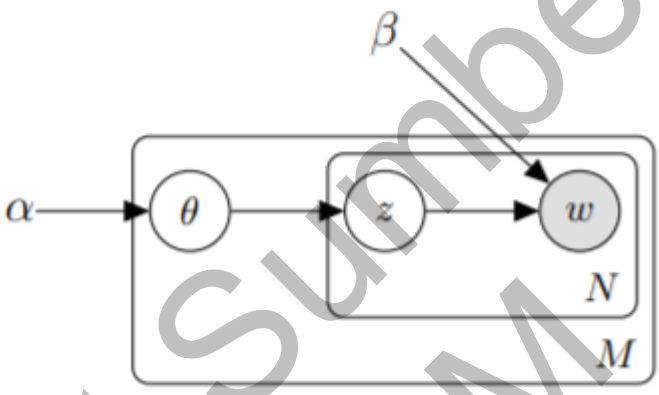
Latent Dirichlet Allocation (LDA) adalah model probabilistik generatif korpus dengan dua peringkat. LDA pada dasarnya adalah lanjutan dari Probabilistic Latent Semantic Indexing (pLSI) dengan model generatif untuk dokumen yang dikembangkan oleh Blei et al. (2003). Selain itu, LDA juga adalah teknik pemodelan topik yang standard industri dan paling banyak digunakan. Di samping itu, ia juga mengekalkan struktur pLSI di mana setiap dokumen diwakili oleh campuran topik dan topik tersebut diwakili oleh campuran istilah. Perwakilan topik laten dari corpus teks disediakan oleh model topik probabilistik. Setiap topik pada dasarnya adalah pengedaran kejadian perkataan dalam topik, dan setiap dokumen dalam korpus dimodelkan sebagai pengedaran topik.

Pembangunan LDA memfasilitasi pengembangan sekumpulan model topik yang lebih kompleks yang mempunyai struktur dokumen-topik-campuran dan istilah-topik-campuran yang sama. Pada waktu yang sama, LDA juga memperluaskan kemampuan model dengan cara yang lain. Seterusnya, LDA menggantikan distribusi topik tetap pLSI dengan proses generatif di mana campuran topik, θ dari dokumen diambil dari taburan Dirichlet yang mengikut parameter, α. Dirichlet adalah taburan *simplex*, yang bermaksud bahawa vektor yang dilukis berjumlah satu. Selain memodelkan pengedaran topik dokumen dengan Dirichlet dan bukan campuran topik statik, berdasarkan Rajah 2.1, LDA secara struktural serupa dengan pLSI dengan cara berikut:

1. Contoh bilangan istilah N dalam dokumen dari sebaran Poisson.
2. Contoh pengedaran topik multinomial θ dalam dokumen dari Dirichlet
3. Pengedaran berparameter oleh α.
4. Untuk setiap istilah w_n dengan N :
    a. Sampelkan topik z_n istilah dari taburan topik multinomial θ.

b. Sampelkan istilah w_n dengan kebarangkalian p (w_n | z_n, β), yang merupakan taburan multinomial yang ditentukan berdasarkan topik z_n.

Rajah 2.1 menggambarkan model LDA sebagai gambarajah plat. Jumlah topik, *K* adalah satu-satunya hiperparameter model, seperti pLSI, dan ia ditetapkan sebelum dimuatkan ke dalam model. Oleh kerana LDA mempunyai model penjanaan lengkap, mengira kemungkinan model pada data yang dipegang atau menilai pentafsiran model yang diberikan *K* boleh digunakan untuk menentukan *K*.



Rajah 2.1 Gambarajah plat untuk proses penjanaan bagi LDA

Terdapat dokumen *M* dan setiap edaran topik dokumen, *θ* diambil sampel dari Dirichlet yang dipadan oleh α. Terdapat *N* terma dalam setiap dokumen yang dihasilkan dengan mengambil sampel topik *z* dari *θ* dan kemudian mengambil sampel istilah dari sebaran istilah multinomial topik yang dikondisikan pada *z*. Rajah 2.1 serupa dengan apa yang dikemukakan oleh Blei, Ng dan Jordan (2003).

2.3.2 Kajian yang Berkaitan Menggunakan *Latent Dirichlet Allocation* (LDA) Dalam Keputusan Pembelian Pengguna

Walaupun membeli-belah dalam talian telah menjadi saluran penting bagi pengguna untuk mendapatkan barang runcit, ulasan dalam talian menjadi sumber yang kaya untuk memahami tingkah laku pengguna. Sebilangan besar kajian dilakukan pada pelbagai jenis dokumen teks tidak berstruktur mengenai ulasan dalam talian melalui LDA untuk meningkatkan keputusan pembelian pengguna. Projek ini menggunakan data tinjauan

Amazon.com untuk meneroka faktor penting bagi pengguna dalam talian untuk membeli kopi pilihan mereka.

Jadual 2.1 Senarai Kajian yang Menggunakan Pemodelan Topik

Penulis	Tajuk	Dapatan Kajian
Wang et al. (2018)	*Topic analysis of online reviews for two competitive products using Latent Dirichlet Allocation (LDA)*	Kajian ini mengaplikasikan LDA untuk menganalisis kelebihan dan kekurangan persaingan dua produk kompetitif menggunakan ulasan dalam talian. Mereka menemui pelengkap ciri produk dalam ulasan positif dan negatif.
Heng et al. (2018)	*Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach*	Kajian ini menggunakan LDA untuk menemui empat faktor yang mempunyai kesan signifikan terhadap kesediaan ulasan pelanggan terhadap produk kopi: perkhidmatan Amazon, ciri fizikal, ciri perasa, dan ekspresi subjektif. Pembaca ulasan pelanggan menganggap ulasan objektif lebih bermanfaat daripada ulasan subjektif. Di samping itu, keselesaan ulasan pelanggan mempunyai hubungan cekung dengan panjang ulasan.
Heng et al. (2019)	*A Topic Mining Approach to Understand What Matters to Online Grocery Consumers: the Cases of Coconut Oil*	Kajian ini menggunakan LDA dan telah mengenal pasti tiga topik: ekspresi subjektif dan penggunaan memasak, ciri produk, dan penggunaan kecantikan. Hasil kajian mereka menunjukkan bahawa walaupun minyak kelapa dapat digunakan untuk memasak dan kecantikan, namun pengguna membezakan kegunaannya ketika mereka membeli dan memberikan ulasan.
Westerlund et al. (2019)	*Topic modelling analysis of online reviews: Indian restaurants at Amazon.com*	Kajian ini menggunakan LDA pada 3,302 ulasan dalam talian mengenai restoran berpusat di A.S. yang menjual makanan India melalui Restoran Amazon. Hasil kajian mereka menunjukkan lima topik yang diketahui sebelumnya dalam ulasan restoran dan dua topik baru yang berkaitan dengan pesanan dalam talian dan penghantaran ke rumah.
Mou et al. (2019)	*Understanding the topics of export cross-border e-commerce consumers feedback: an LDA approach*	Kajian ini menggunakan LDA dan menemui 35 topik utama yang paling banyak disebut oleh pembeli dan penjual. Berdasarkan penemuan mereka, penjual menganggap komisen, audit produk, komunikasi antara penjual dan pembeli, pengurusan pesanan dan lalu lintas sebagai paling penting. Sementara itu, pembeli pula menyebut pengembalian dan pengembalian dana, penjejakan produk, keterangan produk, masa penghantaran, dan prestasi penjual jauh lebih penting daripada topik yang lain.
Jauhari et al. (2020)	*Assessing Customer Needs Based on Online Reviews: A Topic Modeling Approach*	Kajian ini menggunakan LDA untuk meneroka pilihan dan penggunaan pelanggan terhadap trend fesyen. Hasilnya menemui empat topik umum yang dibincangkan: Aksesori, Pakaian, Kualiti, dan Penampilan yang akan memberi manfaat dan meningkatkan perniagaan fesyen dengan memperhitung pengembangan produk.
Lucini et al. (2020)	*Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews*	Kajian ini menggunakan LDA untuk mengenal pasti 27 dimensi kepuasan. Implikasi praktikal yang mereka dapati bagi pengurus perkhidmatan syarikat penerbangan dalam meningkatkan kepuasan pelanggan adalah layanan pelanggan kepada penumpang kelas pertama, keselesaan kepada penumpang ekonomi premium, dan pemeriksaan bagasi dan waktu tunggu untuk pelancong kelas ekonomi.

bersambung...

...sambungan

Sutherldan et al. (2020)	*Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation (LDA)*	Kajian ini menggunakan LDA dalam talian ulasan pelanggan penginapan Korea dan mereka mendapati kualiti lokasi dan perkhidmatan adalah penting. Mereka memperluaskan dimensi kepentingan yang sedia ada oleh aspek lokasi dan kualiti perkhidmatan dan mendedahkan perbezaan dalam topik yang penting antara ciri-ciri penginapan yang berbeza.
Ding et al. (2020)	*Employing structural topic modelling to explore perceived service quality attributes in Airbnb accommodation*	Kajian ini menggunakan LDA untuk mengekstrak atribut kualiti perkhidmatan dari 242,020 ulasan Airbnb di Malaysia. 22 topik berkaitan perkhidmatan diekstrak dari korpus dan empat topik belum muncul dalam kajian Airbnb sebelumnya. Hasil kajian mereka mendapati bahawa pengguna Airbnb Malaysia lebih mementingkan penampilan dan lokasi tempat penginapan berbanding dengan pengguna Airbnb antarabangsa yang lebih mementingkan sama ada tempat itu dapat menampung sekumpulan orang. Selanjutnya, komunikasi dengan tuan rumah memainkan peranan yang semakin penting dalam pengalaman penginapan pengguna Airbnb.
Luo et al.(2020)	*Topic modelling for theme park online reviews: analysis of Disneyland*	LDA digunakan untuk menganalisis ulasan dalam talian taman tema untuk menangkap dan mendedahkan pdanangan yang menyeluruh mengenai pengalaman, kebimbangan, dan kepuasan pengunjung. Pendekatan dan penemuan yang dicadangkan mereka bermanfaat untuk menolong pengurus taman tema dalam memahami persepsi pengunjung, di mana rancangan pemasaran dan penambahbaikan yang berkesan dapat dikembangkan untuk menarik dan mengekalkan pelanggan masa depan.
Li et al. (2021)	*Meal Kit Preferences during COVID-19 Pandemic: Exploring User-Generated Content with Natural Language Processing Techniques*	Kajian ini menggunakan LDA untuk menganalisis 51,497 ulasan pelanggan untuk sembilan syarikat kit makanan di Amerika Syarikat dari 2019 hingga 2020 dan memperoleh empat topik yang dibincangkan oleh pelanggan dalam komen, termasuk pengalaman, kualiti makanan, kemudahan dan perkhidmatan.

2.3.3 Kerja Berkaitan Penambahbaikan dan Mengoptimumkan *Latent Dirichlet Allocation* (LDA)

Corpus yang sederhana besar biasanya boleh mengandungi lebih daripada 100,000 kata unik dengan kebanyakan kata-kata ini hanya terdapat dalam beberapa dokumen. Menerapkan model LDA pada semua perkataan dalam korpus adalah mahal secara komputerisasi dan tidak begitu berguna, kerana kebanyakan perkataan mempunyai corak pengedaran yang tidak menyumbang kepada topik yang bermakna. Teknik yang berguna untuk menyaring perkataan yang terlalu jarang atau terlalu umum adalah menggunakan tf-idf (istilah frekuensi - kekerapan dokumen terbalik) yang memberikan skor rendah pada kata-kata yang sangat jarang atau sangat kerap (Jacobi et al. 2016). Pilihan lain adalah dengan memotong frekuensi minimum untuk menyaring kata-kata

yang jarang berlaku dan menggunakan senarai kata berhenti umum (dan atau membatasi kekerapan dokumen terbalik) untuk menyaring perkataan yang terlalu umum.

Untuk pemodelan topik, selalunya lebih baik hanya menggunakan bahagian pertuturan tertentu, terutamanya kata nama, kata nama yang betul dan bergantung pada tugas dan korpus, kata adjektif dan kata kerja. Model ini secara automatik menyaring kata berhenti yang paling umum, yang cenderung menjadi penentu atau preposisi (kecuali kata kerja umum seperti 'menjadi' ("*to be")* dan 'memiliki' ("*to have")*. Terdapat banyak kajian sebelumnya yang menggunakan bagian pendanaan ucapan dalam model LDA mereka seperti yang ditunjukkan pada Jadual 2.2. Salah satu kajian menunjukkan bahawa purata koherensi topik yang diperhatikan dan purata pengesanan pencerobohan perkataan meningkat berbanding dengan pemodelan korpus mentah (Martin & Johnson 2015).

Jadual 2.2 Senarai kajian menggunakan pemodelan topik dan kaedahnya untuk memperbaiki dan mengoptimumkan model LDA

Penulis	Tajuk	Metode
Darling et al. (2012)	*Unsupervised Part-of-Speech Tagging in Noisy and Esoteric Domains with a Syntactic-Semantic Bayesian HMM*	Kajian ini menggunakan *Part-of-Speech* LDA (POSLDA), sebuah model probabilistik generatif yang konsisten secara sintaksis dan semantik. Model ini menemui topik khusus POS dari korpus yang tidak berlabel dan menunjukkan bahawa model ini secara konsisten mencapai peningkatan dalam pendanaan POS tanpa pengawasan dan pemodelan bahasa melalui pendekatan Bayesian HMM dengan jumlah maklumat sampingan yang berbeza dalam domain Twitter yang bising dan esoterik. **Tag *Part of Speech*:** Kata Nama, Kata nama yang tepat, Kata Nama yang tepat + Berpengalaman, Kata Nama yang tepat + Kata Kerja, Kata Kerja, Kata Adjektif dan Kata Pepatah **Bilangan topik, *K*** = 5,10 *, 15,20,25,30
Hatami et al. (2013)	*N-gram Adaptation Using Dirichlet Class Language Model Based on Part-of-Speech for Speech Recognition*	Kajian ini mencadangkan *Dirichlet Class Language* Model (DCLM) berdasarkan *Part-of-Speech* untuk meningkatkan keadaan terkini LDA. Eksperimen mereka menunjukkan bahawa menggunakan maklumat POS bersama dengan kata-kata sejarah dan kelas kata sejarah meningkatkan model bahasa, dan mengurangkan kebingungan pada korpus kita. Mengeksploitasi maklumat POS bersama dengan DCLM, kadar kesalahan kata sistem ASR menurun sebanyak 1% berbanding dengan DCLM.

bersambung...

...sambungan

Darling & Song (2013)	*Probabilistic topic and syntax modeling with part-of-speech LDAs*	Model LDA gabungan topik dan model sintaksisnya, *Part-of-Speech* LDA atau POSLDA di mana output POSLDA boleh membawa kepada peningkatan kualiti yang kuat: bahagian yang tidak diawasi-pendanaan ucapan. Kajian mereka menunjukkan bahawa menggabungkan dua paksi makna kata iaitu sintaksis dan semantik menjadi model yang koheren dapat menghasilkan peningkatan pada kedua-dua pengedaran topik yang dipelajari dan tugas NLP seperti pemberian tag POS. Selain itu, mereka menunjukkan bahawa menggabungkan dua paksi kata mencapai kebingungan yang lebih rendah dan oleh itu kemampuan ramalan yang lebih baik. **Bahagian tag ucapan:** Kata Adjektif, Kata Kerja dan Kata Nama **Bilangan topik, *K*** = 30
Wang et al. (2014)	*Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology*	Kajian ini mengaplikasikan model LDA yang dipanjangkan dengan menentukan peraturan pengekstrakan frasa nama dan kata-kata teknologi yang telah diekstrak dari tajuk dan abstrak paten untuk mengkaji titik panas dan petunjuk penyelidikan dalam subkelas teknologi yang dipatenkan dalam bidang tertentu. Kajian empirikal ini mendedahkan titik panas teknologi LTE yang muncul dan mendapati bahawa syarikat-syarikat besar dalam bidang ini telah memfokuskan diri pada bidang teknologi yang berlainan dengan kedudukan kompetitif yang berbeza. **Bahagian tag ucapan:** Kata Nama **Bilangan topik, *K*** = Pengoptimuman (bilangan topik ditentukan berdasarkan hubungan antara semua topik)
Martin & Johnson (2015)	*More Efficient Topic Modelling Through a Kata Nama Only Approach*	Model yang dihasilkan dari corpus berita yang disekat, dikurangkan menjadi kata nama sahaja. Mereka mendapati bahawa membuang semua perkataan kecuali kata nama meningkatkan koheren semantik topik. Koherensi topik yang diperhatikan rata-rata meningkat sebanyak 6% dan rata-rata pengesanan pencerobohan perkataan meningkat 8% untuk kata nama satu-satunya korpus, berbdaning dengan pemodelan korpus mentah. Tambahan pula,model masa latihan menjadi lebih pantas apabila mengurangkan artikel menjadi kata nama sahaja. **Bahagian tag ucapan:** Kata Nama **Bilangan topik, *K*** = 20, 50, 100, 200* dan 500
Bhowmik et al. (2015)	*Leveraging topic modeling and part-of-speech tagging to support combinational creativity in requirements engineering*	Penulisan ini mengkaji kreativiti daripada perspektif gabungan dengan meletakkan hubungan yang tidak dikenali antara keperluan kemungkinan tidak asing di mana mereka mencadangkan kerangka baru yang mengekstrak idea-idea yang tidak asing daripada keperluan dan komen pihak berkepentingan dengan menggunakan pemodelan topik dan secara automatiknya menghasilkan keperluan dengan

bersambung...

...sambungan

		mendapatkan kombinasi idea yang tidak dikenali dengan cara membalikkan bahagian pertuturan topik yang dikenal pasti. Hasil kajian mereka menunjukkan bahawa kreativiti keperluan yang dihasilkan oleh kerangka mereka dinilai oleh pakar manusia setdaning dengan keperluan yang dibuat secara manual. Tambahan pula, kos kerangka kerja mereka jauh lebih rendah daripada kerja manual, diukur dengan masa yang dihabiskan untuk menghasilkan keperluan. **Bahagian tag ucapan:** Kata Nama Am dan Kata Kerja
Jacobi et al. (2016)	*Quantitative analysis of large amounts of journalistic texts using topic modelling*	Kajian ini menggunakan lematisasi dan POS-tagger daripada *Stanford Core NLP Suites* dan memilih semua Kata Nama Am, Kata Nama Khas, Kata Kerja, dan Kata Adjektif pada model LDA, dan secara automatiknya akan mengatur arkib besar dokumen berdasarkan topik laten, diukur sebagai corak kata kejadian. Bagi melihat kegunaannya untuk tujuan penyelidikan kewartawanan, mereka melakukan kajian terhadap kes mengenai liputan New York Times untuk teknologi nuklear dari tahun 1945 hingga sekarang, yang mencipta semula sebahagian kajian Gamson dan Modigliani. **Bahagian tag ucapan:** Kata Nama Am, Kata Kerja, Kata Adjektif dan Kata Nama Khas **Bilangan topik, *K*** = 10, 25 (berdasarkan *Perplexity*)
Usop et al. (2017)	*Part of speech features for sentimen classification based on Peruntukan Latent Dirichlet(LDA)*	Kajian ini menggunakan teknik POS untuk menjalankan pemilihan ciri di mana Kata Adjektif dan negatif adalah tanda utama sentimen atau pendapat dalam dokumen dan digunakan pada model LDA. Hasil penyelidikan ini menunjukkan bahawa dokumen yang telah lulus proses ciri berasaskan POS boleh memberikan skor ketepatan yang lebih tinggi dengan perbezaan kira-kira 7.8% daripada dokumen tanpa proses pemilihan pos.. **Bahagian tag ucapan:** Kata Nama Am, Kata Kerja, Kata Keterangan, Kata Adjektif, Kata Nama Khas dan Negasi. **Eksperimen:** Kombinasi persampelan Gibbs iaitu α (0.1, 0.01, 0.001) dan β (0.1, 0.01, 0.001) dan tiga topik.
Chen (2021)	*Keyword Extraction for Privacy Policy Analysis Using Topic Modelling Approaches*	Kajian ini mengembangkan dua model topik: LDA dan POSLDA dengan maklumat sebelumnya iaitu mengenai amalan data yang berbeza dan *kelas Part-of-Speech* dan membandingkan prestasinya untuk pengambilan kata kunci dasar privasi. Hasil kajian mereka menunjukkan bahawa LDA dan POSLDA mampu mengekstrak kata kunci berkualiti daripada dasar privasi untuk pelbagai topik, dan POSLDA bukan sahaja dapat membezakan kelas kata kunci POS untuk topik yang berbeza tetapi juga

bersambung...

...sambungan

	dapat meningkatkan ketepatan pengekstrakan kata kunci dengan membuang kata berhenti yang disesuaikan daripada proses pemodelan yang sama. **Bahagian tag ucapan:** Kata Nama, Kata Kerja, Kata Adjektif dan Kata Keterangan

2.4 ANALISIS SENTIMEN DENGAN TEKNIK VADER

Analisis sentimen adalah cabang perlombongan teks digelar sebagai perlombongan pendapat sebagai kaedah analisis pemprosesan bahasa semula jadi yang mengenal pasti pendapat manusia sama ada daripada segi positif atau negatif yang disiarkan dalam data teks (Liu et al. 2017). Kandungan yang dihasilkan pengguna (UGC) seperti ulasan kebiasaannya mengandungi pendapat positif dan negatif. Analisis sentimen membantu pengekstrakan aspek sebagai ciri dan kata sentimen sebagai pilihan untuk sesuatu objek dari segi aspek. Kata-kata sentimen mempunyai sifat polariti iaitu positif, negatif, dan neutral dan mengubah perkataan yang mewakili aspek (Chakraborty 2014). Algoritma pembelajaran tanpa pengawasan digunakan untuk mengklasifikasikan ulasan seperti yang disyorkan atau tidak berdasarkan kata-kata sentimen (Usop et al. 2017). Polariti sesuatu dokumen ditentukan dengan mengira kekerapan frasa positif dan negatif (Bonta & Janardhan 2019). Jika dokumen tidak mempunyai kata-kata sentimen, dokumen itu dianggap neutral kerana sering mewakili pendapat objektif.

Terdapat beberapa Python algoritma untuk melaksanakan analisis sentimen dan salah satunya adalah teknik VADER. VADER (*Valence Conscious Dictionary and Sentiment Reasoner*) adalah platform analisis sentimen berdasarkan leksikon dan peraturan yang paling sesuai digunakan terutamanya pada sentimen media sosial (Bonta & Janardhan, 2019). Ini adalah sumber terbuka di bawah lesen MIT yang dikembangkan oleh George Berry, Ewan Klein, dan Pier Paolo. VADER menggunakan istilah leksikon yang berkaitan dengan sentimen sebagai asasnya untuk analisis sentimen. Dalam kaedah ini, setiap kata dalam leksikon dinilai untuk menentukan sama ada kata tersebut adalah positif atau negatif, serta berapa positif atau negatifnya dalam situasi tertentu (Nguyen et al. 2018). Setiap perkataan dalam leksikon VADER ditentukan sama ada positif atau negatif berdasarkan pada berapa peringkat positif atau negatif yang diterimanya (Nguyen et al. 2018).

Dalam domain media sosial, leksikon Vader menunjukkan prestasi yang sangat baik berbanding teknik sentimen lain (Bonta & Janardhan 2019). VADER mengekalkan kelebihan leksikon sentimen tradisional seperti LIWC (*Linguistic Inquiry and Word Count*) kerana lebih besar, lebih mudah disemak, difahami, cepat digunakan, dan mudah (Bonta & Janardhan 2019). Hasil leksikon sentimen VADER berkualiti dan telah disahkan oleh manusia (Bonta & Janardhan 2019). VADER meletakkan dirinya berbeza daripada LIWC kerana VADER lebih sensitif terhadap ekspresi sentimen dalam konteks media sosial dan pada masa yang sama lebih memihak kepada domain yang lain.

2.5 Kajian Analisis Sentimen Berasaskan *Latent Dirichlet Allocation* (LDA)

LDA dapat diterapkan pada analisis sentimen di mana Boyd-Graber dan Resnik (2010) telah memperluas model generatif ini untuk menyokong pelbagai bahasa untuk meramalkan sentimen tinjauan. Selain itu, Li et al. (2010) menggunakan LDA yang diperluas untuk menangkap kekutuban sentimen dan mengira kekutuban untuk topik. Terdapat banyak kajian sebelumnya bahawa model LDA mereka dengan analisis sentimen seperti dalam Jadual 2.3 di mana kaedah ini telah menemui topik yang memfokuskan pada kata-kata sentimen dan bukannya hanya perkataan sahaja.

Jadual 2.3 Senarai Kajian yang menggabungkan model LDA dan Analisis Sentimen dan tumpuan kajian mereka.

Penulis	Tajuk	Tumpuan Kajian
Lin & He (2009)	*Joint sentimen/topic model for sentimen analysis*	Kajian ini mencadangkan Sentimen Bersama Topik Model (JST) untuk aspek dan pengekstrakan sentimen dari ulasan pengguna. Untuk memodelkan sentimen dokumen, mereka menambahkan lapisan sentimen tambahan ke LDA antara dokumen dan lapisan topik. Di JST, sentimen dan topik diekstrak secara bersamaan dari teks. Set data dalam kajian ini adalah koleksi ulasan filem.
Jo & Oh (2011)	*Aspect dan sentimen unification model for online review analysis*	Kajian ini mencadangkan *Sentence*-LDA dan *Aspect Sentimen Unification Model* (ASUM), yang merupakan kaedah yang serupa dengan JST. Perbezaannya adalah bahawa di ASUM, setiap ayat dalam ulasan pengguna dianggap mengenai aspek produk tunggal yang tidak begitu dalam JST. Mereka menggunakan peranti elektronik dan set data ulasan restoran untuk menilai *Sentence*-LDA dan ASUM.
Xianghua et al. (2013)	*Multi-aspect sentimen analysis for Chinese online social reviews*	Kajian ini mengembangkan penyesuaian LDA untuk teks pendek yang mereka namakan sebagai MG-LDA di mana mereka menyatakan bahawa apabila algoritma

bersambung...

...sambungan

	based on topic modeling dan HowNet lexicon	LDA diterapkan pada ulasan pengguna, bukan sahaja ia mengetahui aspek produk sebagai topik, tetapi juga menemui topik selain aspek produk. Dalam MG-LDA, topik mengenai aspek produk ditemui oleh topik tempatan dan topik lain dijumpai oleh topik global. Untuk mengetahui topik global, algoritma LDA diterapkan dengan melibatkan semua ulasan. Untuk mengetahui topik tempatan (aspek produk), LDA diterapkan dengan menggunakan kaedah *sliding window*. Mereka menguji MG-LDA pada ulasan sosial dalam talian Cina.
Chen et al. (2015)	*Proposal of LDA-Based Sentimen Visualization of Hotel Reviews*	Kajian ini mencadangkan sistem interaktif yang memvisualisasikan pasangan sentimen yang diambil dari ulasan mengenai hotel. Di samping itu, kekutuban pasangan sentimen dapat diperbetulkan secara intuitif oleh pengguna. Sebelum memvisualisasikan data ulasan, ayat dikelaskan menjadi topik dengan menggunakan LDA. Dengan menggunakan leksikon sentimen, pasangan sentimen dapat dilihat dengan melukis graf. Sementara itu, antara muka melibatkan pengguna dalam interaksi untuk meningkatkan leksikon sentimen khusus untuk kawasan sasaran.
Xiong et al. (2018)	*A short text sentimen-topic model for product reviews*	Kajian ini mencadangkan WSTM (*Word-pair Sentimen-Topic Model)* untuk ulasan teks pendek (Xiong et al., 2018). Dalam WSTM, mereka memodelkan proses generatif dari aspek sentimen dan aspek secara serentak. Mereka menggunakan *sliding window* dan pada setiap langkah *sliding window*, mereka menghasilkan pasangan kata dan kemudian menggunakan pasangan kata ini dalam model topik mereka.
García et al. (2018)	*W2VLDA: almost unsupervised system for aspect based sentimen analysis*	Kajian ini mengembangkan kaedah W2VLDA untuk pengekstrakan aspek dan pengesanan polariti sentimen pada ulasan pengguna di mana mereka menggunakan algoritma LDA dalam kombinasi dengan penyisipan kata berterusan, word2vec dan pengklasifikasi entropi maksimum. Mereka kemudian menguji sistem mereka pada ulasan restoran dan peranti elektronik (komputer riba, digital-slr) dan pada set data tugas 5 SemEval-2016.
Tang et al. (2019)	*Aspect based fine-grained sentimen analysis for online reviews*	Kajian ini mengembangkan model topik sentimen berdasarkan aspek bersama (JABST) yang bersama-sama mengekstrak aspek dan pendapat berbilang butir melalui aspek pemodelan, pendapat, polariti sentimen dan butiran secara serentak. Model JABST mereka dan model MaxEnt – JABST mengatasi model sebelumnya yang lain pada ulasan dunia nyata untuk alat elektronik dan restoran yang ditunjukkan.
Irawan et al. (2019)	*Mining Tourist's Perception toward Indonesia Tourism Destination Using Sentimen Analysis dan Topic Modelling*	Kajian ini menggabungkan analisis LDA dan sentimen untuk meninjau persepsi pengunjung terhadap 10 laman web yang paling banyak dikunjungi di Indonesia. Emosi dan topik yang dibincangkan dalam komen adalah dua ciri yang akan diambil. Dengan menggunakan rangka kerja perlombongan data, lima jenis emosi dan topik yang berkaitan dengan pelancongan ditemui. Hasil kajian mereka menunjukkan bahawa *Joy* adalah emosi paling menonjol yang mengiringi pengalaman

bersambung...

...sambungan

		pengunjung.
Kwon et al. (2021)	*Topic Modeling dan Sentimen Analysis of Online Review for Airlines*	Kajian ini menggunakan analisis LDA dan sentimen untuk mengetahui jenis kata penting dalam ulasan dalam talian. Hasil dari pemodelan topik, 'tempat duduk', 'layanan', dan 'makanan' adalah masalah penting dalam penerbangan melalui analisis frekuensi. Selanjutnya, hasilnya menunjukkan bahawa kelewatan adalah masalah utama, yang dapat mempengaruhi ketidakpuasan pelanggan sementara ‘layanan staf’ dapat membuat pelanggan puas melalui analisis sentimen kerana hasilnya menunjukkan ‘layanan staf’ dengan makanan dan hidangan dalam pemodelan topik

BAB III

KAEDAH KAJIAN

3.1 PENGENALAN

Dalam bahagian ini, projek ini menerangkan rangka kerja yang dicadangkan untuk mencari sentimen berorientasikan ciri ulasan produk kopi yang dinyatakan oleh pelanggan. Ulasan tersebut kebanyakannya mengandungi pendapat bercampur-campur. Oleh itu, untuk mengetahui ciri-ciri laten, projek ini mengklasifikasikan ulasan ke dalam topik yang berbeza oleh *Latent Dirchlet Allocation* (LDA). Selepas klasifikasi, projek ini menggunakan analisis sentimen VADER untuk menyimpulkan skor sentimen ulasan.

Jadual 3.1 Set percubaan bahagian penandaan pertuturan ke atas model LDA

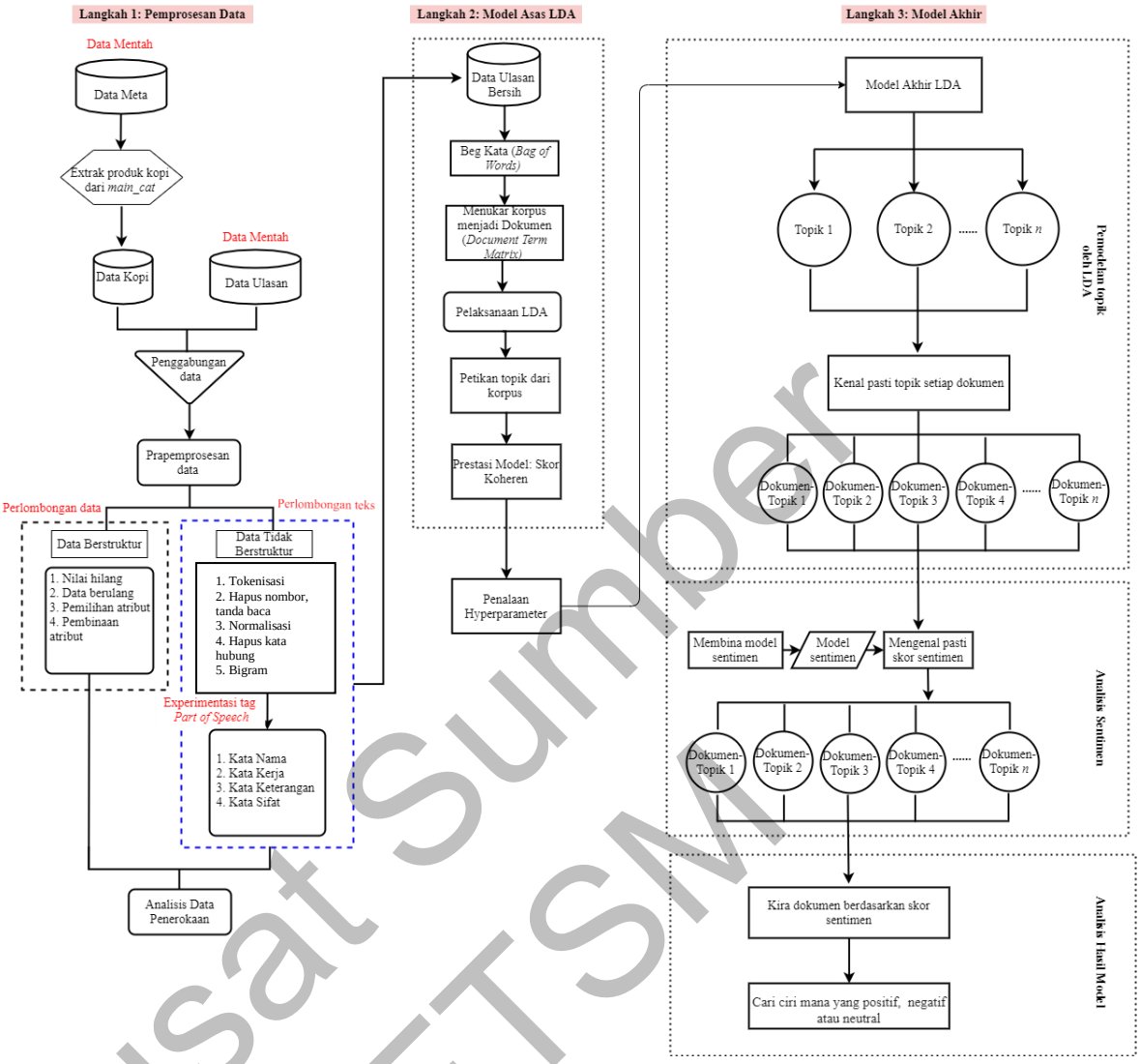
Percubaan	POSTAG
1	Kata nama
2	Kata kerja
3	Kata keterangan
4	Kata adjektif
5	Kata nama dan kata kerja
6	Kata nama dan kata keterangan
7	Kata nama dan kata adjektif
8	Kata kerja dan kata keterangan
9	Kata kerja dan kata kerja
10	Kata keterangan dan kata adjektif
11	Kata nama, kata kerja, dan kata keterangan
12	Kata nama, kata kerja dan kata adjektif

bersambung...

...sambungan

13	Kata nama, kata keterangan dan kata adjektif
14	Kata kerja, kata keterangan, dan kata adjektif
15	Kata nama, kata adjektif, kata kerja, dan kata keterangan

Aliran kerja keseluruhan pendekatan yang dicadangkan berdasarkan kedua-dua objektif teknikal dan perniagaan ditunjukkan dalam Rajah 3.1. Kajian ini menggunakan data ulasan pelanggan Amazon dan metadata produk yang terdiri daripada koleksi besar teks ulasan dan keterangan produk mengenai produk makanan runcit dan gourmet. Terdapat beberapa pra-pemprosesan yang dilakukan untuk menukar kedua-dua fail mentah ke dalam data berstruktur baik. Kemudian, memilih beberapa atribut yang berkaitan dengan objektif kajian. Seterusnya, menukar fail teks mentah kepada jujukan yang jelas bagi unit linguistik yang bermakna. Selepas itu, ekstrak perkataan daripada setiap ulasan menggunakan *Part of Speech Tagging* (POSTAG). Kemudian, algoritma LDA dilaksanakan di atas set data yang dibersihkan ini untuk menjana ciri produk. Kedua-dua langkah ini dijalankan lima belas kali berdasarkan gabungan POSTAG seperti yang ditunjukkan dalam Jadual 3.1 dan setiap kombinasi model dinilai dengan koheren topik dan model dengan skor koheren tertinggi dipilih sebagai model LDA akhir. Akhirnya, analisis sentimen VADER digunakan untuk mengetahui skor positivity, neutral dan negatif pada setiap ciri produk

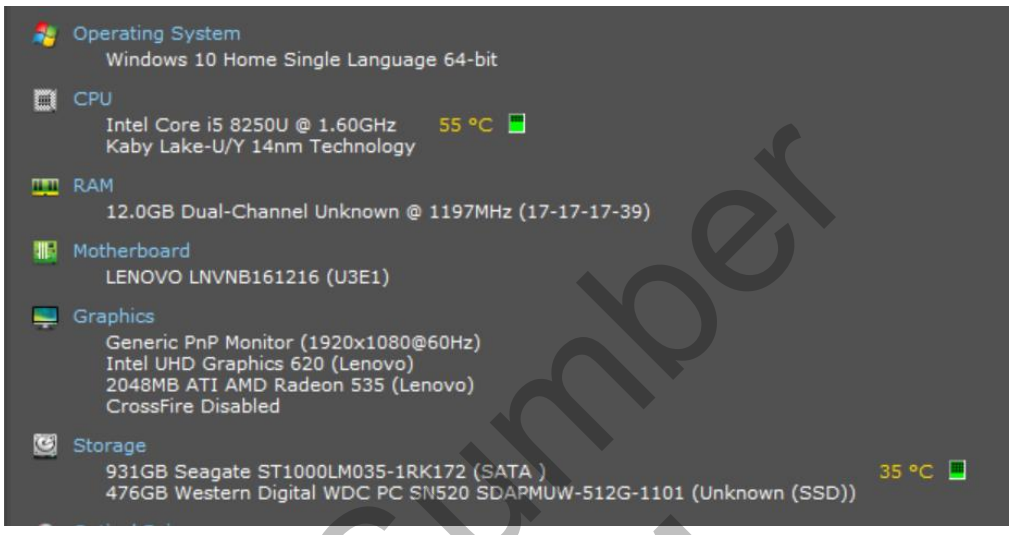


Rajah 3.1 Kerangka Kerja Sentimen Berasaskan LDA

3.2 MENANGANI BATASAN KOMPUTERISASI

Oleh kerana set data adalah besar yang lebih daripada 5 juta data dan pemprosesan dan pembinaan *Latent Dirichlet Allocation* yang dipaparkan berdasarkan model sentimen yang memerlukan kuasa pengiraan yang besar, kajian ini telah mencuba beberapa platform dengan konfigurasi sistem yang berbeza. Pada mulanya, kajian ini telah cuba melaksanakan keseluruhan rangka kerja dalam *Jupyter Notebook* di tempatan (*local)* yang mengambil masa kira-kira 12 jam untuk disiapkan. Selain itu, walaupun menggunakan *Google Colab* dengan mod GPU, ia masih mengambil masa yang lama untuk disiapkan dan had penggunaan untuk *Google Colab* hanyalah 12 jam sehari manakala di *Kaggle* hanya 30 jam seminggu. Pembinaan dan menganalisa adalah sangat

memenatkan dan memakan masa yang sangat lama seperti yang ditunjukkan dalam Jadual 3.2. Oleh yang demikian, kajian ini memutuskan untuk menggunakan platform *Google Cloud Compute Engine* untuk membina 15 model LDA yang berbeza kerana ia dapat memilih konfigurasi mesin yang boleh mengurangkan masa pelaksanaan.



Rajah 3.2 Ringkasan Konfigurasi Sistem Tempatan

Jadual 3.2 Perbandingan antara masa pelaksanaan merentasi platform

No.	Notebook, Konfigurasi Mesin	Anggaran Masa Pelaksanaan
1	Buku Nota Jupyter pada komputer peribadi	12 jam
2	Google Colab (standard)	10 jam
3	Google Colab (GPU)	6 jam
4	Buku Nota Kaggle	9 jam
5	Google Cloud Notebook Bersepadu dengan JupyterLab (16 Vcpu, 60GB RAM, NVDIA TESLA 4, No. daripada GPU:2)	45 minit
6.	Google Cloud Notebook Bersepadu dengan JupyterLab (16 Vcpu, 60GB RAM, NVDIA K80, No. daripada GPU:2)	40 minit
7.	Google Cloud Notebook Bersepadu dengan JupyterLab (16 Vcpu, 60GB RAM, NVDIA P100, No. daripada GPU:2)	35 minit

3.3 PRA-PEMPROSESAN DATA AWAL

Kajian ini akan menggunakan data ulasan pelanggan Amazon.com dan metadata produk terhadap produk makanan runcit dan gourmet dari Februari 2007-Oktober 2018 yang dikumpulkan oleh Ni et al. (2019). Dalam set data, terdapat sejumlah 5,074,160 ulasan yang berbeza dan 287,209 produk. Berdasarkan Rajah 3.1, kajian ini bermula dengan

menggabungkan kedua-dua set data dengan menggunakan ID produk. Oleh kerana kajian ini mempunyai minat dalam pasaran kopi, ia hanya memilih produk kopi dari atribut kategori utama. Selepas pemilihan, data yang digunakan untuk analisis termasuk 436,537 ulasan pelanggan kopi dan 18,341 produk seperti yang ditunjukkan dalam Jadual 3.3.

Jadual 3.3 Bilangan data untuk setiap set data: Ulasan Produk dan Metadata Produk

Set data	Ulasan	Metadata
Data penuh	5,074,160	287,209
Kopi	520,384	18,411

3.3.1 Pemilihan Atribut

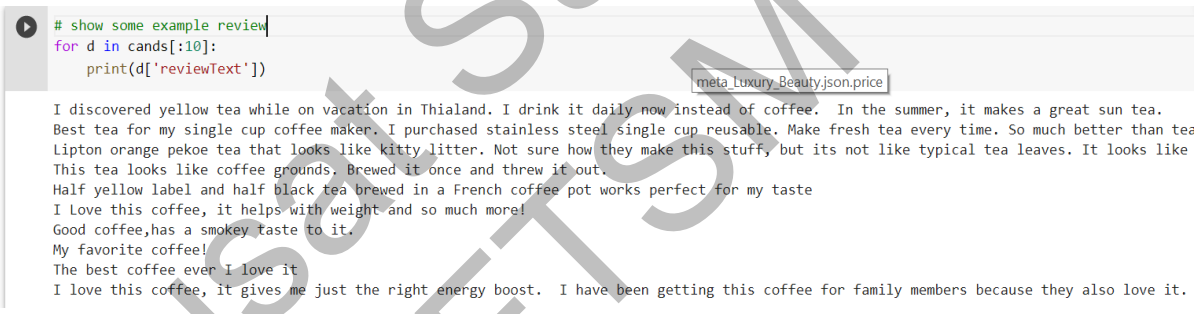
Terdapat banyak atribut untuk setiap set data: ulasan produk dan metadata produk seperti yang ditunjukkan dalam Jadual 3.4 dan Jadual 3.5. Berdasarkan objektif kajian, kajian ini hanya berminat dengan teks ulasan, ringkasan, penilaian produk dan masa ulasan dari set data, Ulasan Produk manakala set data, Metadata Produk, kajian ini memilih tajuk, ciri, perihalan dan jenama produk.

Jadual 3.4 Atribut dalam Ulasan Produk

Atribut	Deskripsi
reviewerID	ID pengulas, cth. A2SUAM1J3GNN3B
asin	ID produk, cth. 0000013714
reviewerName	Nama pengulas
vote	Ulasan undi yang berguna
style	Kamus bagi metadata produk, cth., "Format" ialah "Hardcover"
reviewText	Teks ulasan
overall	Penarafan produk
summary	Ringkasan ulasan
unixReviewTime	Masa kajian semula (masa unix)
reviewTime	Masa kajian semula (mentah)
image	Imej yang disiarkan oleh pengguna selepas mereka menerima produk

Jadual 3.5 Atribut dalam Metadata Produk

Atribut	Deskripsi
asin	ID produk, contohnya 0000031852
title	nama produk
feature	ciri format titik peluru produk
description	keterangan produk
price	harga dalam dolar AS (pada masa merangkak)
image	url imej produk
Related	produk berkaitan (juga dibeli, juga dilihat, dibeli bersama- sama, membeli selepas melihat)
salesRank	maklumat peringkat jualan
brand	nama jenama
categories	senarai kategori produk kepunyaan
tech1	jadual butiran teknikal pertama produk
tech2	jadual butiran teknikal kedua produk
similar	jadual produk yang serupa



Rajah 3.3 Contoh- contoh ulasan pelanggan

```
# show some example products
for d in cands[:10]:
    print(d['title'])
```

```
IKEA - BRYGGKAFFE MELLANROST Decaffeinated Coffee (X1)
IKEA - KAFFE HELA B&Ouml;NOR M&Ouml;RKROST Coffee Whole Beans, Dark Roast (8.8 oz)
Vacu Vin Coffee Saver Refill Container
Espressione 100% Arabica Coffee, 150-Count Pods
Coffee-mate 35170BX French Vanilla Creamer, 0.375oz (Box of 50)
Coffee-mate 35110BX Original Creamer, 0.375oz (Box of 50)
Douwe Egberts Aroma Rood Ground Coffee, 8.8-Ounce Package
Starbucks Decaf House Blend Ground Coffee, 12 Ounce (Pack of 6)
Starbucks Espresso Roast Dark Roast Ground Coffee, 12-Ounce Bag
Pilon Gourmet Whole Bean Restaurant Blend Espresso Coffee, 16 Ounce
```

Rajah 3.4 Contoh produk

3.3.2 Kejuruteraan Ciri

Dalam kajian ini, beberapa atribut baru dijana untuk memenuhi objektif dan penggambaran. Atribut "*reviewText*" dan "*summary"* dicantumkan dalam satu atribut kepada "*review_text"* manakala "*reviewTime*" dibahagikan kepada tiga atribut iaitu hari, bulan dan tahun. Selain itu, kajian ini juga mengklasifikasikan penilaian produk kepada dua kelas di mana penilaian lebih daripada 3 adalah produk yang baik manakala penilaian kurang daripada 3 adalah produk yang buruk. Jadual 3.6 di bawah ialah ringkasan ciri yang akan digunakan dalam kajian:

Jadual 3.6. Ringkasan Penggunaan Atribut dalam Kajian

Atribut	Deskripsi
reviewerID	ID pengulas, cth. A2SUAM1J3GNN3B
asin	ID produk, cth. 0000013714
overall	Penilaian produk
title	Tajuk produk
feature	Ciri produk
description	Deskripsi produk
brand	Jenama produk
review_text	Gabungkan ulasan Teks dan Ringkasan
day	Hari
month	Bulan
year	Tahun
rating_class	Kelas Penarafan Produk (produk yang baik atau buruk)

3.3.3 Nilai tidak lengkap atau hilang

Dalam kajian ini, membuang nilai yang hilang adalah langkah seterusnya selepas pemilihan atribut dan kejuruteraan ciri. Berdasarkan Jadual 3.7, terdapat beberapa atribut yang mempunyai nilai yang hilang. Walau bagaimanapun, disebabkan data besar iaitu dengan 520,384 ulasan atau baris, kajian akan membuang mana-mana baris yang nilainya hilang.

Jadual 3.7. Nilai Hilang untuk Setiap Atribut

Ciri	Nilai Hilang
reviewerID	1
asin	0
overall	1

bersambung...

...sambungan

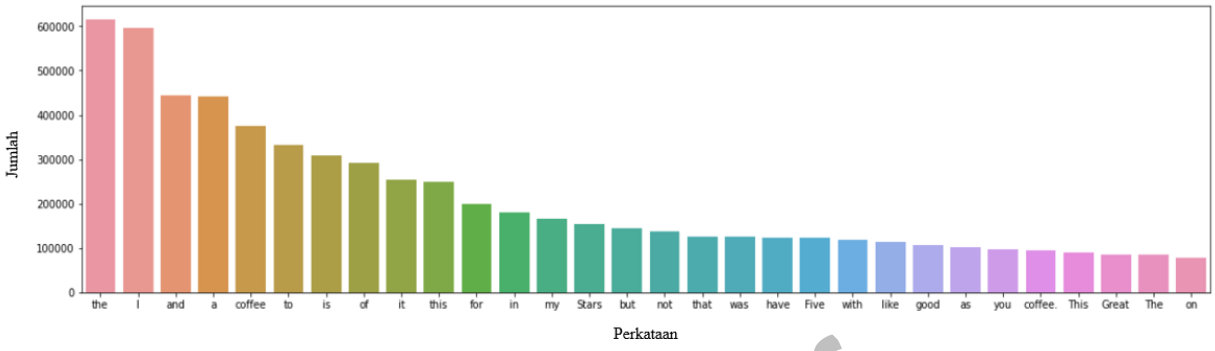
title	0
feature	0
description	0
brand	0
review_text	245
time	0
day	0
month	0
year	0
Rating_class	1

Jadual 3.8 Bilangan Data Sebelum dan Selepas Pembersihan Data

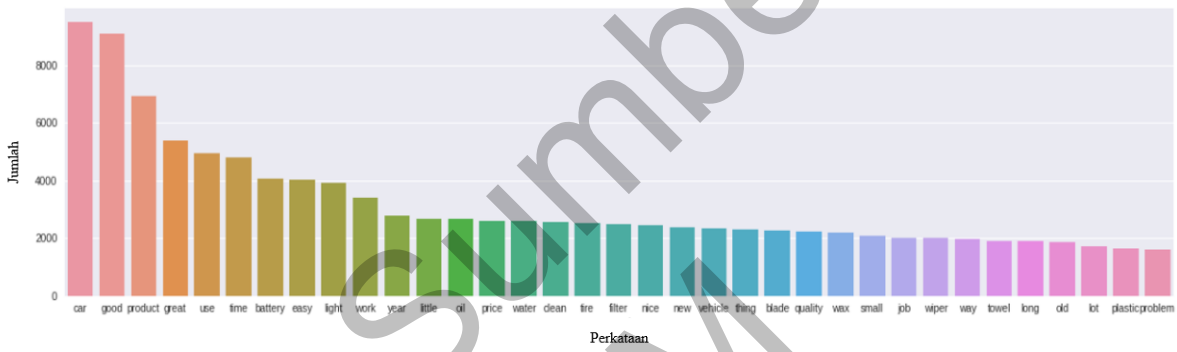
Data	**Bilangan Baris**
Gabungkan data (ulasan produk dan metadata)	520,384
Selepas mengalih keluar nilai yang hilang	520,038

3.4 PRA PEMPROSESAN TEKS

Proses keseluruhan kajian ini adalah untuk mengenal pasti bilangan topik yang optimum dari ulasan pelanggan dan menggunakan topik ini bersama dengan atribut lain untuk menentukan faktor-faktor yang mempengaruhi keputusan pembelian pengguna. Pertama, data teks telah diproses terlebih dahulu untuk membersihkannya daripada perkataan yang tidak perlu, tanda baca, dan aksara khas dalam setiap ulasan. Tujuannya adalah untuk membuang maklumat yang tidak diingini yang boleh mengganggu proses latihan yang mungkin menjejaskan hasilnya. Mengikuti proses ini adalah untuk mengubah setiap ayat menjadi format perkataan tunggal. Kemudian, setiap perkataan akan diubah balik ke bentuk perkataan asal atau kamus bergantung kepada jenis perkataan, sama ada Kata Nama, Kata Adjektif, Kata Kerja, atau Kata Keterangan melalui proses *lemmatization*. Perkataan-perkataan ini disimpan ke dalam model beg perkataan (*bag-of*-words) untuk tujuan pembinaan model. Oleh yang demikian, hasilnya hanya akan memberi tumpuan kepada perkataan yang bermakna dan relevan kepada topik yang dihasilkan. Kajian ini menggunakan NLTK (*Natural Language Toolkit*) dan *Spacy* untuk pemprosesan pra-teks. Rajah 3.5 menunjukkan taburan token mentah dalam korpus sebelum pemprosesan teks dan Rajah 3.6 menunjukkan taburan token selepas pemprosesan teks.



Rajah 3.5 Token Mentah Sebelum Pemprosesan Teks



Rajah 3.6 Token Selepas Teks Pemprosesan

3.4.1 Tokenisasi

Sebelum menganalisis kandungan teks, kajian ini perlu merungkaikan teks mentah menjadi beberapa bahagian kecil. Tokenisasi merungkaikan teks mentah kepada perkataan yang dipanggil sebagai token. Token ini membantu dalam memahami konteks atau membangunkan model. Tokenisasi membantu dalam mentafsir makna teks dengan menganalisis urutan perkataan.

3.4.2 Nombor, Aksara Khas dan Tanda Baca

Untuk pengelasan teks dan pengekstrakan kata kunci, kajian ini lebih bergantung kepada kandungan perkataan. Oleh yang demikian, selain daripada perkataan, nombor, simbol dan tanda baca yang tidak berkenaan dikeluarkan dari kajian. Ini membantu kajian ini untuk mengurangkan lagi bilangan token dalam perwakilan teks juga.

3.4.3 Normalisasi

Untuk mengurangkan bilangan jenis perkataan, kajian ini menggunakan teknik normalisasi untuk menukar perkataan seperti "COFFEE", "Coffee" dan"coffee" menjadi bentuk atau jenis perkataan yang sama. Teknik ini dilakukan dengan menukarkannya kepada sama ada huruf kecil atau huruf besar supaya mereka semua kelihatan sama dalam representasi. Dengan membuat seperti sedemikian, bukan sahaja menjadikan perkataan sepadan lebih mudah tetapi juga meningkatkan ketersambungan di antara kata-kata ini dan mengurangkan bilangan jenis perkataan dalam perbendaharaan kata.

3.4.4 Penghapusan Kata Hubung

Kata hubung (*stopwords*) sering digunakan dalam teks tetapi jarang mempunyai banyak makna berbanding dengan perkataan kandungan seperti kata nama, kata kerja, kata adjektif, dan kata keterangan. Mereka juga dipanggil sebagai perkataan berfungsi yang sepadan dengan bahagian lain pertuturan seperti penentu (‘*the*’, ‘*a*’, ‘*an’* , ‘*some*’ dan ‘*any*’), penyambung (‘*and*’, ‘*or*’, dan ‘*but’*), dan preposisi (contohnya, ‘*of’*, ‘*in’*, dan ‘*at’*). Untuk tugas klasifikasi, kata hubung di dalam teks lebih mirip kepada bunyi bising. Oleh yang demikian, ia perlu dikeluarkan kerana ia terdapat di setiap dokumen dan tidak membantu dalam pembinaan model. Selain itu, beberapa huruf dan perkataan tunggal yang biasa dengan set data juga boleh dianggap sebagai kata hubung. Dalam bahasa Inggeris, kata hubung biasanya dibina secara manual, dalam lingkungan beberapa ratus, bergantung kepada penyelidik yang mahu menghapuskannya dari set data. Untuk kajian ini, selain daripada senarai kata kunci dalam NLTK, senarai kata hubung yang digunakan disertakan dalam Rajah 3.7.

```
#Remove Stopwords, Make Bigrams and Lemmatize

# NLTK Stop words
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords

stop_words = stopwords.words('english')
stop_words.extend(['from', 'really', 're', 'would', 'use', 'go','be','ever','coffee'])
```

Rajah 3.7 Senarai sambungan kata hubung bagi kajian

3.4.5 Pemodelan Frasa: Model Bigram

Kajian ini menggunakan *Bigram* iaitu dua perkataan yang kerap berlaku bersama dalam dokumen dengan menggunakan model Frasa. Kemudiannya, diserahkan kepada *Phraser()* untuk kecekapan dalam kelajuan pelaksanaan.

Jadual 3.9 Senarai Pemprosesan Pra Teks dan saiz teks selepas setiap pemprosesan

Bil	Aktiviti Pra-Pemprosesan Teks	Saiz Teks (Bilangan Perkataan)
1	Jumlah token mentah	143,076,375
2	Selepas alih keluar tanda baca	126,593,928
3	Selepas henti-henti dipakai	90,768,652
4	Selepas *bigram*	90,297,052
5	Lematisasi: kata nama, kata kerja, kata keterangan dan kata adjektif	73,382,118
6	Lematisasi: kata nama, kata kerja dan kata keterangan	55,403,763
7	Lematisasi: kata nama, kata kerja dan kata adjektif	67,328,453
8	Lematisasi: kata nama, kata keterangan dan kata adjektif	58,859,046
9	Lematisasi: kata kerja, kata keterangan dan kata adjektif	39,622,992
10	Lematisasi: kata nama dan kata kerja	49,350,940
11	Lematisasi: kata nama dan kata keterangan	40,882,701
12	Lematisasi: kata nama dan kata adjektif	52,806,159
13	Lematisasi: kata kerja dan kata keterangan	21,843,969
14	Lematisasi: kata kerja dan kata adjektif	33,576,069
15	Lematisasi: kata keterangan dan kata adjektif	25,159,278
16	Lematisasi: kata nama	34,832,172
17	Lematisasi: kata kerja	15,830,648
18	Lematisasi: kata keterangan	7,636,467
19	Lematisasi: kata adjektif	19,560,913

3.5 PEMBANGUNAN PEMODELAN TOPIK DENGAN *LATENT DIRICHLET ALLOCATION*

Dalam kajian ini, kaedah pemodelan topik yang digunakan secara meluas iaitu, *Latent Dirichlet Allocation* (LDA) digunakan untuk mencari topik yang berdasarkan teks ulasan. Model LDA boleh mencungkil maklumat utama dan hubungan statistik yang penting dan dalam pada masa yang sama mengurangkan kerumitan korpus teks. Matlamat model ini adalah untuk mencari sekumpulan topik yang paling dekat menggambarkan kata-kata yang diperhatikan dalam semua dokumen. Maklumat yang dihasilkan daripada model LDA termasuk kata kunci yang berkaitan dengan setiap topik dan kebarangkalian bahawa setiap ulasan teks dikaitkan dengan setiap topik. Kajian ini akan menggunakan pakej *Gensim Python* untuk menganalisis ulasan pelanggan Amazon. Kajian sebelum ini telah menunjukkan beberapa cara untuk mencipta pemodelan topik: ada yang hanya mendekati dengan menggunakan semua ulasan Heng et al. (2018) dan Heng et al. (2019) dan penggunaan lain adalah hanya menggunakan Kata Nama untuk mencipta topik (Joung & Kim 2020).

3.5.1 Memilih gabungan Bahagian Penandaan Pertuturan (POSTAG) yang terbaik

Salah satu objektif kajian ini adalah untuk melakukan penambahbaikan model LDA dengan mencari bahagian penandaan pertuturan (*Part of Speech Tagging*) yang terbaik. Kajian ini memilih kata nama, kata kerja, kata keterangan dan kata adjektif untuk mencipta banyak kumpulan topik yang berbeza berdasarkan Jadual 3.1.

3.5.2 Memilih bilangan topik yang optimum, *K*

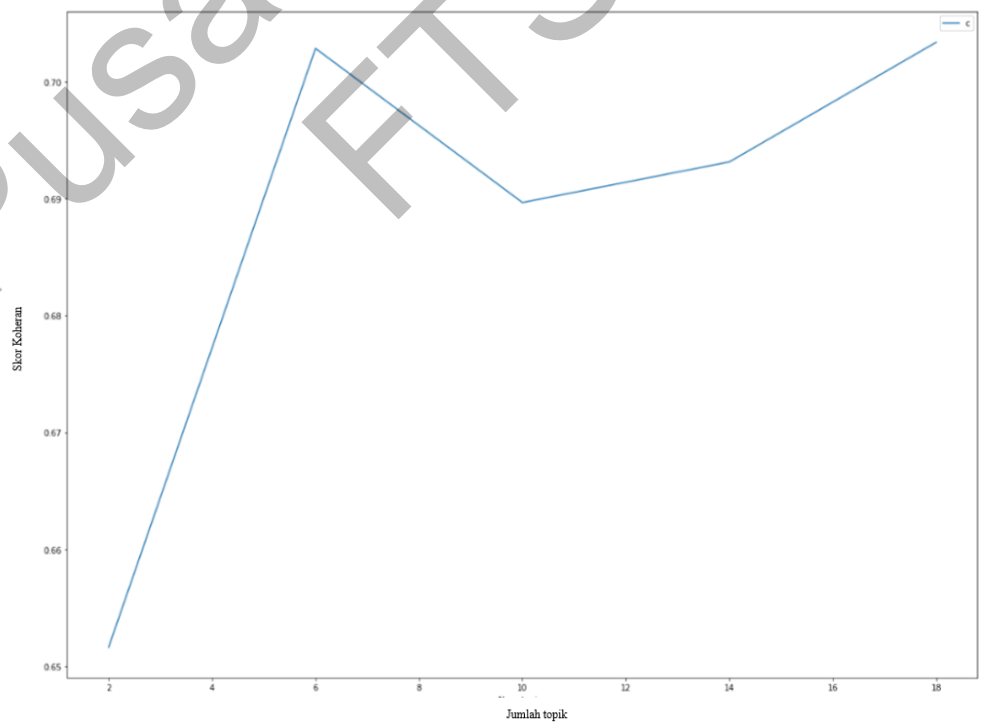
Objektif seterusnya adalah untuk mengoptimumkan model LDA dengan mencari bilangan topik yang optimum. Kajian ini membina banyak model LDA dengan 15 kombinasi POSTAG dan mengambil model yang mempunyai nilai koheren yang paling tinggi sebagai model akhir kajian ini. Memilih K yang betul menandakan berakhirnya perkembangan koheren topik di mana topik yang dihasilkan dengan koheran yang tinggi dapat menawarkan topik yang bermakna dan boleh ditafsirkan. Memilih nilai yang lebih tinggi kadang kala boleh menyediakan lebih banyak sub-topik. Walau bagaimanapun,

jika kata kunci yang sama diulang dalam pelbagai topik, ia mungkin tanda bahawa *K* terlalu besar atau berlakunya berlebihan (*overfitting*).

Kajian ini menggunakan *compute_coherence_values()* seperti yang ditunjukkan dalam Rajah 3.8 yang melatihkan pelbagai model LDA dan menyediakan model serta skor koheren mereka yang sepadan. Rajah 3.9 dan Jadual 3.10 menunjukkan bilangan topik yang berbeza, *K* dan skor yang sepadan dengan mereka.

```
def compute_coherence_values(dictionary,corpus,texts,start,limit,step):
    coherence_vals=[]
    model_list=[]

    for num_topics in range(start,limit,step):
        # building LDA Model
        model=gensim.models.LdaMulticore(corpus=corpus,id2word=dictionary,
                                         num_topics=num_topics,random_state=100,
                                         chunksize=100,passes=10,per_word_topics=True)

        model_list.append(model)

        coherencemodel=CoherenceModel(model=model,texts=texts,dictionary=dictionary,coherence='c_v')

        coherence_vals.append(coherencemodel.get_coherence())
    return model_list,coherence_vals

model_list,coherence_vals=compute_coherence_values(dictionary=id2word,
                                                   corpus=corpus,texts=data_lemmatized,
                                                   start=2,limit=20,step=4)
```

Rajah 3.8 Fungsi *compute_coherence_values()* untuk mencari bilangan *K* yang optimum



Rajah 3.9 Plot pada bilangan *K* dan markah sepadan mereka

Jadual 3.10 Bilangan Topik, *K* dan Markah Sepadan mereka

Bilangan Topik, *K*	**Markah Koheren**
2	0.6516
6	0.7029
10	0.6897
14	0.6931
18	0.7034

3.5.3 Memilih Teras Tunggal LDA Vs. LDA Pelbagai Teras

Gensim secara khususnya mensasarkan bidang pemodelan topik dan pemfaktoran matriks. Ia mempunyai dua jenis pelaksanaan LDA iaitu *LdaModel* dan *LdaMulticore*. Di *LdaModel*, semua kelompok diproses secara berurutan dengan serentak dan berlaku sepenuhnya di dalam *NumPy*. Manakala, di *LdaMulticore*, beberapa kelompok diproses secara serentak, dan terdapat utas agregasi tunggal yang menggabungkan hasilnya secara segerak (Vorontsov et al. 2015). Dalam kajian mereka, terdapat perbezaan antara *LdaMode*l dan *LdaMulticore* dari segi kelajuan dan nilai kebingungan (*perplexity*). *LdaMulticore* menunjukkan prestasi yang lebih baik di mana kelajuannya adalah 22 minit lebih cepat dan kadar kebingungan yang lebih tinggi berbanding *LdaModel* (Vorontsov et al. 2015).

Berurusan dengan korpus besar adalah memenatkan dan memakan masa. Selain daripada memilih konfigurasi sistem yang berbeza untuk mengurangkan masa komputerisasi, kajian ini mengguna pakai *LdaMulticore* yang menggunakan semua teras CPU untuk menyelarikan dan mempercepatkan latihan model. Berdasarkan jadual 3.11, terdapat sedikit berbeza skor koheren antara *LdaMode*l dan *LdaMulticore*.

Jadual 3.11 Perbandingan antara Prestasi LdaModel dan LdaMulticore bagi Setiap model LDA

No.	**POSTAG**	***LdaModel***		***LdaMulticore***	
		K	**Skor Koheran**	***K***	**Skor Koheran**
1	Kata nama	10	0.595	10	0.600
2	Kata kerja	10	0.536	10	0.524
3	Kata keterangan	10	0.333	10	0.323
4	Kata adjektif	10	0.479	10	0.465
5	Kata nama dan kata kerja	10	0.629	10	0.678

bersambung...

...sambungan

6	Kata nama and kata keterangan	10	0.623	10	0.642
7	Kata nama and kata adjektif	10	0.603	10	0.643
8	Kata kerja and kata keterangan	10	0.595	10	0.593
9	Kata kerja and kata adjektif	10	0.637	10	0.629
10	Kata keterangan and kata adjektif	10	0.547	10	0.556
11	Kata nama, kata kerja, kata keterangan	10	0.667	10	0.714
12	Kata nama, kata kerja and kata adjektif	10	0.589	10	0.653
13	Kata nama, kata keterangan and kata adjektif	10	0.612	10	0.626
14	Kata kerja, kata keterangan, kata adjektif	10	0.669	10	0.672
15	Kata nama, kata adjektif, kata kerja, kata keterangan	10	0.571	10	0.614

Selain itu, berdasarkan kajian oleh Sahin (2020) untuk menilai perbezaan prestasi antara dua model, kajian ini menggunakan ujian *Wilcoxon Signed-Rank* untuk melihat adakah perbezaan prestasi antara dua model signikan. Jika nilai statistik ujian adalah kurang daripada nilai kritikal (p-nilai kurang daripada 0.05), perbezaan prestasi dikatakan signifikan secara statistik (Sahin 2020). Oleh kerana nilai *p* adalah 0.062 lebih daripada 0.05, tidak ada perbezaan penting dalam prestasi antara *LdaModel* dan *LdaMulticore.*

H_0: Terdapat perbezaan signifikan dalam prestasi antara LdaModel dan LdaMulticore

H_1: Tiada perbezaan signifikan dalam prestasi antara LdaModel dan LdaMulticore

Jadual 3.12 Hasil nilai *P* ujian *Wilcoxon Signed-Rank*

	LdaModel-LdaMulticore
Z	-1.87
Asymp. Sig (2-tailed)	0.062

3.5.4 Analisis Sentimen Kaedah Vader

Seperti yang dinyatakan sebelum ini, VADER adalah alat analisis sentimen berasaskan peraturan dan leksikon. Ia menggunakan gabungan sentimen, senarai ciri leksikal yang biasanya dilabelkan mengikut orientasi semantik mereka sama ada positif atau negatif. VADER boleh dikatakan berjaya apabila berurusan dengan teks media sosial, ulasan

filem, dan ulasan produk. Ini kerana VADER bukan sahaja memberitahu tentang skor positiviti dan negatif tetapi juga memberitahu mengenai betapa positifnya atau negatifnya sentimen tersebut. Terdapat beberapa kelebihan VADER iaitu ia berfungsi dengan sempurna pada teks jenis media sosial dan tidak memerlukan data latihan tetapi dibina daripada leksikon piawaian emas berasaskan manusia, berasaskan valensi, yang boleh digunakan secara umum. Di samping itu, ia dapat menyokong emoji dan tidak mengalami pertukaran yang rugi dengan prestasi yang pantas.

3.6 Pengujian dan Pengesahan

Penilaian model dalam aliran data adalah asas supaya model yang berprestasi rendah dapat dikenalpasti dan ditingkatkan atau digantikan oleh model yang berprestasi lebih baik. Penyelidik perlulah menilai prestasi kaedah yang mereka gunakan dengan berkesan. Di dalam kajian ini, terdapat beberapa pengujian dan pengesahan yang digunakan sebagai tanda ukuran baik atau buruk bagi sesuatu model atau ujian eksperimen.

3.6.1 Menggunakan Koheren Topik sebagai Ukuran Prestasi

Bahasa semula jadi tidak kemas, tidak jelas dan penuh dengan penafsiran subjektif, dan kadang-kadang berusaha membersihkan kesamaran dengan mengurangkan bahasa menjadi bentuk yang tidak wajar. Kajian ini akan meneroka koheren topik, metrik penilaian intrinsik, dan bagaimana ia boleh membenarkan pemilihan model secara kuantitatif.

Koheren topik adalah suatu pendekatan untuk menilai satu topik dengan menilai persamaan semantik di antara perkataan pemarkahan tinggi dalam topik (Stevens et al. 2012). Ini akan membantu dalam membezakan topik-topik yang boleh ditafsirkan secara semantik daripada yang digunakan sebagai artifak inferens statistik (Jauhari et al. 2020). Terdapat pelbagai teknik untuk metrik koheren, walau bagaimanapun, kajian ini akan melaksanakan langkah *Cv* seperti yang dicadangkan (Röder et al. 2015). *Cv* telah terbukti menunjukkan prestasi yang lebih baik daripada *Pointwise Mutual Information* (PMI) yang lebih baik dan menunjukkan korelasi yang lebih baik

berhubung dengan topik yang dinilai oleh manusia (Bouma 2009). Terdapat empat bahagian pengiraan *Cv*:

1. Data dibahagikan kepada pasangan perkataan.
2. Bagi setiap pasangan perkataan atau perkataan tunggal, kebarangkalian mereka dikira.
3. Langkah pengesahan dikira untuk memeriksa kekuatan sokongan pada set perkataan lain.
4. Skor koheren keseluruhan dikira.

3.6.2 Pengesahan Tafsiran Topik

Hasil keluaran model LDA seperti Rajah 4.7 dan 4.10, tidak mempunya spesifik nama bagi setiap topik. Oleh yang demikian, penyelidik perlu secara manual menamakan topik satu persatu. Kajian ini menggunakan kaedah Guo et al. (2017) dan Hao et al. (2017) di mana sambungan logik antara kata-kata teratas topik dan berat relatif (berat kedudukan untuk satu perkataan) dalam Rajah 4.10 yang sepadan dan juga *WordCloud* dalam Rajah 4.15 untuk menamakan topik. Di samping itu, kajian ini turut mengunakan tafsiran kata kunci berdasarkan hasil pada dokumen yang paling relevan untuk setiap topik untuk menamakan setiap topik yang dibincangkan dalam Subsekyen 4.4 dan 4.5.

3.6.3 Ujian Wilcoxon-Signed Rank

Terdapat beberapa ujian statistik yang digunakan untuk membandingkan prestasi model (Bifet et al. 2015). Dalam kajian ini, ujian Wilcoxon-Signed Rank digunakan untuk menilai dan membandingkan prestasi antara dua jenis model LDA iaitu *LdaModel* dan *LdaMulticore* seperti yang dibincangkan dalam Subseksyen 3.5.3. ujian *Wilcoxon Signed-Rank* adalah ujian bukan parametrik dan menggunakan hasilnya pada setiap lipatan sebagai percubaan. Ujian ini mengira perbezaan prestasi kedua-dua model tersebut untuk setiap percubaan (Bifet et al. 2015). Setelah menentukan nilai mutlak perbezaan ini, ia menghitung jumlah peringkat di mana perbezaannya positif dan jumlah peringkat di mana perbezaannya negatif. Nilai minimum kedua-dua jumlah ini kemudian dibandingkan dengan nilai kritikal V_{α}. Sekiranya nilai minimum ini lebih rendah, hipotesis nol bahawa prestasi kedua-dua model adalah sama dapat ditolak pada tahap α keyakinan (Bifet et al. 2015).

BAB IV

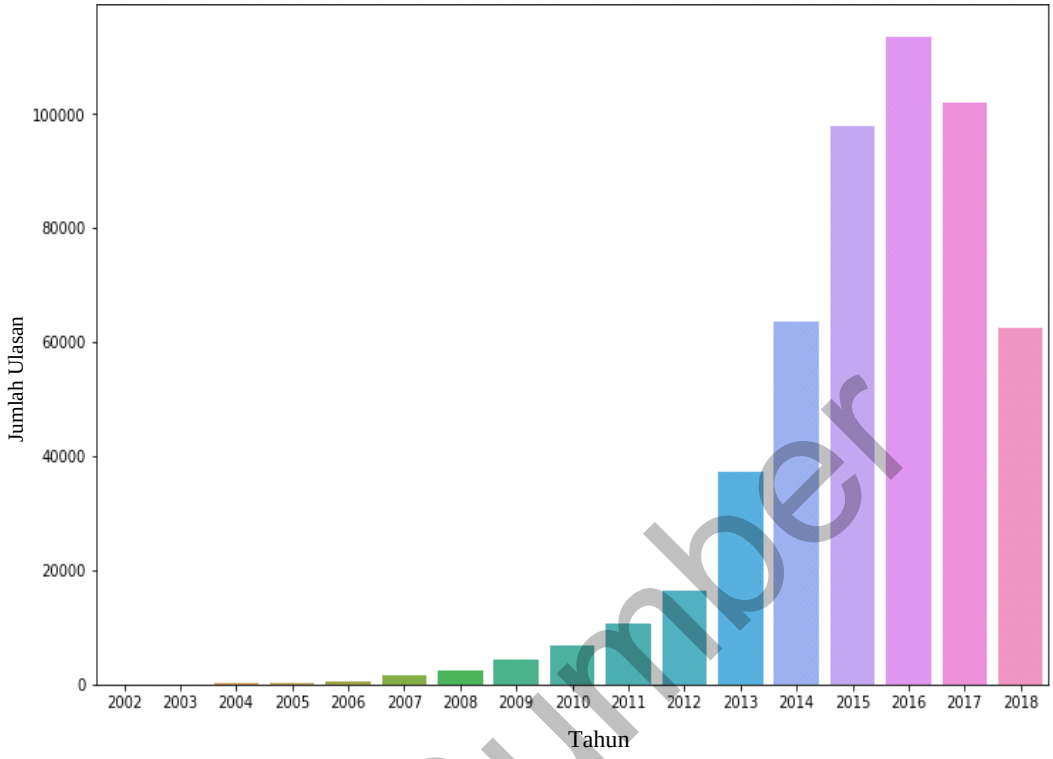
DAPATAN KAJIAN

4.1 ANALISIS DATA PENEROKAAN

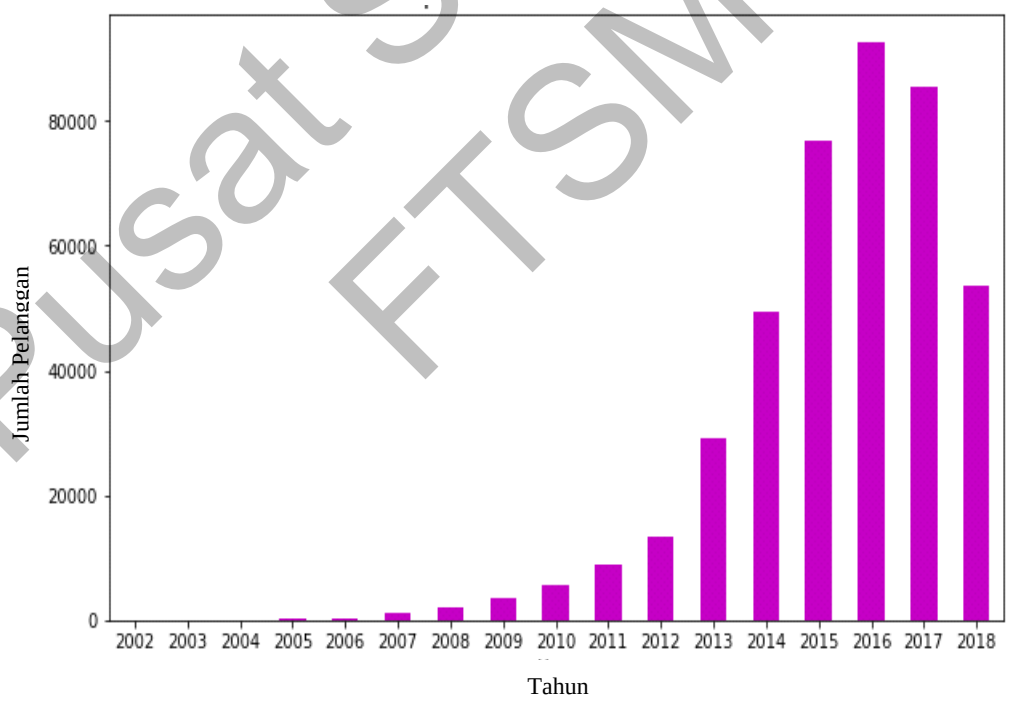
Statistik deskriptif set data ditunjukkan di Jadual 4.1. Daripada jadual tersebut, sebanyak 16,275 produk kopi berada di Amazon.com dari Jun 2002 sehingga September 2018 dengan lebih dari 2,484 jenama. Dalam set data, terdapat sejumlah 520, 038 ulasan pelanggan terhadap produk kopi dan seperti dalam Rajah 4.1, jumlah ulasan meningkat sepanjang tahun seiring dengan peningkatan jumlah pelanggan (Rajah 4.2) dan jumlah produk (Rajah 4.3) pada tahun-tahun berikutnya. Apabila kajian ini menyelami setiap penilaian, terdapat 334,862 produk diberi 5 bintang, 60,791 dengan 4 bintang, 34,456 dengan 3 bintang, 22,980 dengan 2 bintang dan 34,304 dengan 1 bintang.

Jadual 4.1 Statistik Deskriptif daripada Set Data

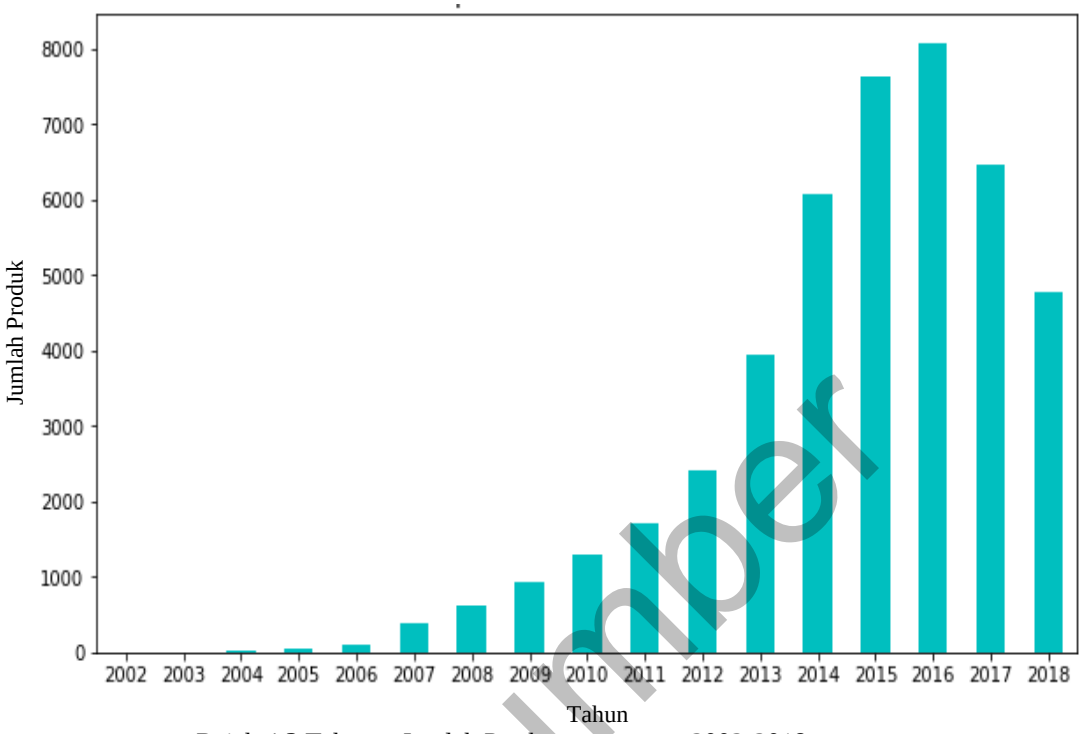
Penjelasan	Statistik
Jumlah produk	16,275
Jumlah jenama	2,484
Jumlah tinjauan	520,038
Ulasan keseluruhan dengan penilaian yang baik	458,259
Ulasan keseluruhan dengan penilaian yang kurang baik	61,779
Purata penilaian produk	4.307
Jumlah produk dengan satu bintang	34,304
Jumlah produk dengan dua bintang	22,980
Jumlah produk dengan tiga bintang	34,456
Jumlah produk dengan empat bintang	60,791
Jumlah produk dengan lima bintang	334,862



Rajah 4.1 Taburan Jumlah Ulasan dari Tahun 2002-2018

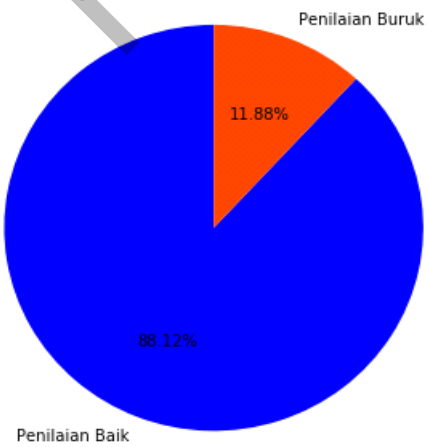


Rajah 4.2 Taburan jumlah pelanggan dari tahun 2002-2018



Rajah 4.3 Taburan Jumlah Produk dan Tahun 2002-2018

Semasa memilih satu produk di platform e-dagang terutamanya di Amazon iaitu sebelum mengklik dan melihat paparan produk, pelanggan kebanyakannya akan terlebih dahulu menapis harga dan penilaian produk. Berdasarkan carta pai dalam Rajah 4.4, 88.12 peratus produk kopi mempunyai penilaian yang baik sementara 11.88 peratus dengan penilaian buruk.



Rajah 4.4 Carta Pai untuk Perbandingan antara Produk Baik dan Produk Buruk

4.2 MODEL ASAS DAN TETAPAN EKSPERIMEN

Projek ini dimulakan dengan model dasar LDA di mana jumlah topik, *K* adalah 10 dan semua parameter dalam LDA seperti yang disebutkan sebelumnya dalam Subseksyen 2.3.1 ditetapkan secara lalai. Ini kerana salah satu fokus projek ini adalah untuk mencari bilangan topik yang optimum, *K*. Seperti yang dinyatakan dalam Jadual 15 dan dibincangkan sebelumnya dalam Subseksyen 3.5.4, walaupun *Wilcoxon Signed Rank Test* menunjukkan bahawa tidak ada perbezaan kepentingan pada prestasi antara *LdaModel* dan *LdaMulticore*, *LdaMulticore* mempunyai skor koheren yang lebih tinggi berbanding *LdaModel*. Oleh kerana, *LdaMulticore* memproses beberapa kumpulan secara serentak, terdapat satu rangkaian agregasi tunggal yang menggabungkan hasilnya secara serentak (Vorontsov et al. 2015), *LdaMulticore* lebih pantas berbanding *LdaModel.*

Jadual 4.2 Perbezaan dua model LDA yang berbeza dengan bilangan topik, *K* = 10

No	Bahagian Penandaan Ucapan	Bilangan Token	*LdaModel*	*LdaMulticore*
1	Kata nama, kata kerja, kata keterangan dan kata adjektif	73,382,118	0.571	0.614
2	**Kata nama, kata kerja dan kata keterangan**	**55,403,763**	**0.667**	**0.714**
3	Kata nama, kata kerja dan kata adjektif	67,328,453	0.589	0.653
4	Kata nama, kata keterangan dan kata adjektif	58,859,046	0.612	0.626
5	Kata kerja, kata keterangan dan kata adjektif	39,622,992	0.669	0.672
6	Kata nama dan kata kerja	49,350,940	0.629	0.678
7	Kata nama dan kata keterangan	40,882,701	0.623	0.642
8	Kata nama dan kata adjektif	52,806,159	0.603	0.643
9	kata kerja dan kata keterangan	21,843,969	0.595	0.593
10	Kata kerja dan kata adjektif	33,576,069	0.637	0.629
11	Kata keterangan dan kata adjektif	25,159,278	0.547	0.556
12	Kata nama	34,832,172	0.595	0.600
13	Kata kerja	15,830,648	0.536	0.524
14	Kata keterangan	7,636,467	0.333	0.323
15	Kata adjektif	19,560,913	0.479	0.465

Objektif lain projek ini adalah untuk mencari gabungan Bahagian Ucapan (*Part of Speech Tagging*) yang terbaik dan berdasarkan Jadual 4.2, kata nama, kata kerja dan kata keterangan adalah gabungan terbaik kerana kedua *LdaModel* dan *LdaMulticore* menunjukkan skor koherensi tertinggi berbanding kombinasi lain. Rajah 4.5 dan Jadual 4.3 menunjukkan keluaran model LDA dengan setiap topik dengan set kata kunci dan pemberat kata kunci sendiri.

Jadual 4.3 Topik dengan kumpulan kata kunci sendiri

Topik	Kata kunci Per Topik
Topik 1	taste, try, buy, bag, say, review, think, get, smell, know, give, even, money, go, see, could, never, thing, may, look
Topik 2	order, product, box, time, receive, package, purchase, gift, arrive, come, delivery, selection, item, packaging, get, quality, send, quickly, pack, ship
Topik 3	price, buy, love, find, store, brand, get, quality, much, deal, time, purchase, year, wish, order, cost, husband, product, grocery, folger
Topik 4	star, love, taste, price, value, product, thank, stuff, always, service, seller, good, expect, deal, favorite, great, quality, repeat, shipping, awhile
Topik 5	flavor, variety, love, try, coffee, enjoy, pack, cup, like, chocolate, taste, brand, definitely, lot, favorite, far, one, purchase, choice, vanilla
Topik 6	best, bean, decaf, taste, find, try, make, grind, far, packet, always, instant, home, brew, hand, use, year, ground, travel, press
Topik 7	drink, love, morning, day, taste, recommend, caffeine, make, get, highly, start, give, feel, product, enjoy, friend, time, stuff, want, need
Topik 8	cup, pod, work, machine, make, ground, use, brew, capsule, maker, espresso, problem, get, water, well, seal, come, brand, kcup, time
Topik 9	roast, taste, blend, bean, try, prefer, brew, bit, smooth, coffee, medium, light, aroma, burn, bitterness, breakfast, find, sample, quality, quit
Topik 10	taste, add, creamer, make, water, sugar, tea, milk, drink, mix, cream, pretty, vanilla, need, product, ice, half, use, tasting, be

```
# Print the Keyword in the 10 topics
pprint(lda_model.print_topics())
doc_lda = lda_model[corpus]
```

```
[(0,
  '0.060*"taste" + 0.035*"try" + 0.030*"buy" + 0.028*"bag" + 0.026*"say" + '
  '0.024*"review" + 0.022*"think" + 0.021*"get" + 0.020*"smell" + '
  '0.017*"know"'),
 (1,
  '0.091*"order" + 0.072*"product" + 0.047*"box" + 0.038*"time" + '
  '0.033*"receive" + 0.027*"package" + 0.025*"purchase" + 0.020*"gift" + '
  '0.019*"arrive" + 0.017*"come"'),
 (2,
  '0.127*"price" + 0.112*"buy" + 0.057*"love" + 0.050*"find" + 0.032*"store" + '
  '0.026*"brand" + 0.022*"get" + 0.017*"quality" + 0.015*"much" + '
  '0.013*"deal"'),
 (3,
  '0.542*"star" + 0.146*"love" + 0.110*"taste" + 0.037*"price" + 0.026*"value" '
  '+ 0.021*"product" + 0.013*"thank" + 0.012*"stuff" + 0.007*"always" + '
  '0.006*"service"'),
 (4,
  '0.327*"flavor" + 0.062*"variety" + 0.048*"love" + 0.046*"try" + '
  '0.029*"coffee" + 0.029*"enjoy" + 0.025*"pack" + 0.024*"cup" + 0.017*"like" '
  '+ 0.017*"chocolate"'),
 (5,
  '0.073*"best" + 0.070*"bean" + 0.040*"decaf" + 0.035*"taste" + 0.032*"find" '
  '+ 0.022*"try" + 0.021*"make" + 0.020*"grind" + 0.020*"far" + '
  '0.017*"packet"'),
 (6,
  '0.060*"drink" + 0.048*"love" + 0.039*"morning" + 0.034*"day" + '
  '0.031*"taste" + 0.028*"recommend" + 0.023*"caffeine" + 0.018*"make" + '
  '0.018*"get" + 0.016*"highly"'),
 (7,
  '0.100*"cup" + 0.060*"pod" + 0.055*"work" + 0.043*"machine" + 0.038*"make" + '
  '0.029*"ground" + 0.022*"use" + 0.021*"brew" + 0.020*"capsule" + '
  '0.016*"maker"'),
 (8,
  '0.096*"roast" + 0.060*"taste" + 0.049*"blend" + 0.025*"bean" + 0.020*"try" '
  '+ 0.017*"prefer" + 0.017*"brew" + 0.014*"bit" + 0.014*"smooth" + '
  '0.013*"coffee"'),
 (9,
  '0.064*"taste" + 0.039*"add" + 0.037*"creamer" + 0.029*"make" + '
  '0.028*"water" + 0.026*"sugar" + 0.025*"tea" + 0.021*"milk" + 0.020*"drink" '
  '+ 0.020*"mix"')]
```

Rajah 4.5 Pengeluaran hasil gabungan terbaik *Part of Speech* model LDA

Rajah 4.6, *Wordcloud* berfungsi untuk memudahkan visualisasi kata kunci setiap topik di mana semakin banyak kata tertentu muncul dalam sumber data teks semakin besar dan lebih tebal muncul dalam *Wordcloud.* Setiap topik dalam *Wordcloud* ada kata kunci paling besar dan tebal di mana topik 1: '*taste',* topik 2: '*order',* topik 3: '*price*', topik 4: 'star', topik 5: 'flavour', topik 6: 'best', topik 7: 'drink', topik 8: '*cup*', topik 9: 'roast' dan topik 10: 'taste'.

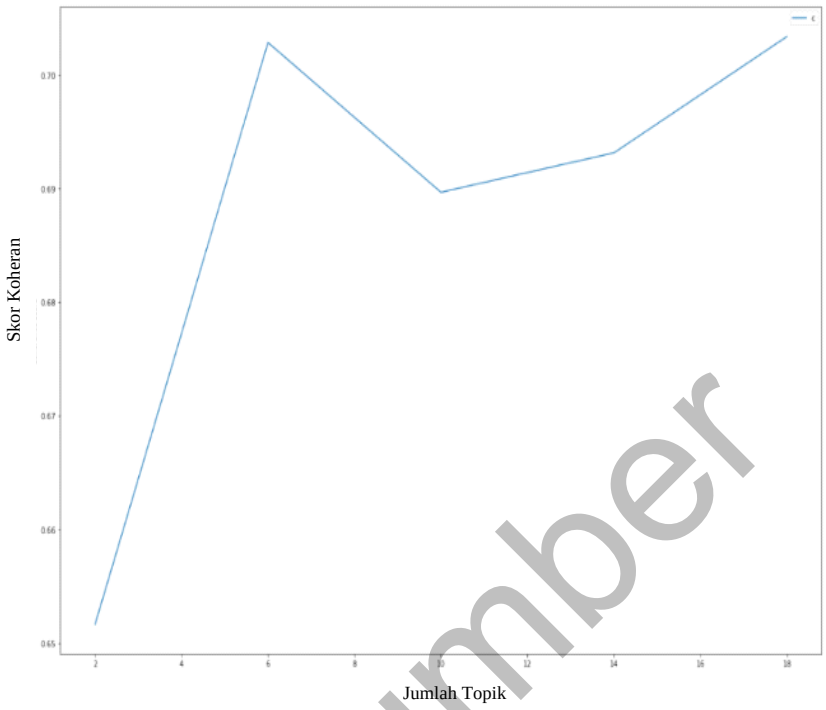
Rajah 4.6 WordCloud model LDA POSTAGs terbaik, K = 10

4.3 MODEL PERUNTUKAN *LATENT DIRICHLET* AKHIR DAN PENGAMBILAN TOPIK

Bilangan laten topik adalah keputusan penting dalam pemodelan topik. Bilangan topik optimum yang dioptimumkan menggunakan fungsi *compute_coherence_values()*. Berdasarkan Jadual 4.4 dan Rajah 4.7, bilangan topik optimum, K adalah 6 kerana ia mempunyai skor koherensi tertinggi setelah pengoptimuman. Oleh itu, model LDA akhir untuk projek ini adalah dengan gabungan POSTAG terbaik (kata nama, kata kerja dan kata keterangan) dan bilangan topik, K = 6.

Jadual 4.4 Bilangan Topik, K dan Skor Koheren

Bilangan Topik, *K*	**Skor Koheren**
2	0.6516
6	**0.7029**
10	0.6897
14	0.6931
18	0.7034



Rajah 4.7 Bilangan plot topik, *K* dan skor koheran

Jadual 4.5 Topik dengan Set Kata Kunci Tersendiri

Topik	**Terma Setiap Topik**
Topik 1	taste, roast, flavor, try, blend, smell, buy, coffee, pretty, think, expect, better, say, brand, even, bit, look, aroma, prefer, give
Topik 2	cup, pod, work, product, order, box, machine, receive, package, time, come, get, ground, make, buy, money, gift, purchase, problem, packaging
Topik 3	price, flavor, love, buy, find, variety, try, brand, order, purchase, pack, cup, store, get, decaf, capsule, time, far, coffee, enjoy
Topik 4	star, love, taste, product, price, best, thank, delivery, buy, stuff, value, order, discount, time, service, coconut, always, arrive, fast, quality
Topik 5	flavor, love, taste, drink, morning, creamer, add, recommend, make, caffeine, day, sugar, enjoy, milk, tea, need, vanilla, cup, highly, mix
Topik 6	bean, make, brew, use, bag, get, try, product, drink, water, grind, buy, find, go, review, time, espresso, know, year, take

```
#the keywords for each topic and the weightage(importance) of each keyword using lda_model.print_topics()

from pprint import pprint

# Print the Keyword in the 10 topics
pprint(lda_model.print_topics())
doc_lda = lda_model[corpus]
```

```
[(0,
  '0.127*"taste" + 0.054*"roast" + 0.052*"flavor" + 0.030*"try" + '
  '0.019*"blend" + 0.018*"smell" + 0.015*"buy" + 0.014*"coffee" + '
  '0.014*"pretty" + 0.013*"think"'),
 (1,
  '0.046*"cup" + 0.038*"pod" + 0.035*"work" + 0.029*"product" + 0.025*"order" '
  '+ 0.024*"box" + 0.023*"machine" + 0.019*"receive" + 0.017*"package" + '
  '0.015*"time"'),
 (2,
  '0.060*"price" + 0.051*"flavor" + 0.041*"love" + 0.038*"buy" + 0.036*"find" '
  '+ 0.033*"variety" + 0.032*"try" + 0.028*"brand" + 0.026*"order" + '
  '0.018*"purchase"'),
 (3,
  '0.397*"star" + 0.125*"love" + 0.071*"taste" + 0.050*"product" + '
  '0.039*"price" + 0.024*"best" + 0.016*"thank" + 0.015*"delivery" + '
  '0.012*"buy" + 0.012*"stuff"'),
 (4,
  '0.077*"flavor" + 0.043*"love" + 0.042*"taste" + 0.037*"drink" + '
  '0.029*"morning" + 0.026*"creamer" + 0.022*"add" + 0.021*"recommend" + '
  '0.020*"make" + 0.017*"caffeine"'),
 (5,
  '0.032*"bean" + 0.025*"make" + 0.019*"brew" + 0.017*"use" + 0.017*"bag" + '
  '0.016*"get" + 0.014*"try" + 0.013*"product" + 0.012*"drink" + '
  '0.011*"water"')]
```

Rajah 4.8 Hasil keluaran model terakhir LDA

Rajah 4.9 *WordCloud* Model LDA Akhir

4.4 TOPIK DOMINAN DAN SUMBANGAN PERATUSNYA DALAM SETIAP DOKUMEN

Dalam model LDA, setiap dokumen terdiri daripada pelbagai jenis topik. Namun begitu, selalunya hanya salah satu topik sahaja yang dominan. Kod di bawah ini mengekstrak topik dominan ini untuk setiap ayat dan menunjukkan berat topik dan kata kunci dalam output yang diformat dengan baik. Dengan cara ini, pengkaji dapat mengenal pasti dokumen yang akan menjadi topik utama.

```
def format_topics_sentences(ldamodel=lda_model, corpus=corpus, texts=coffee):
    # Init output
    sent_topics_df = pd.DataFrame()

    # Get main topic in each document
    for i, row in enumerate(ldamodel[corpus]):
        row=row[0]
        row = sorted(row, key=lambda x: (x[1]), reverse=True)
        # Get the Dominant highest weighted topic, Perc Contribution and Keywords for each document
        for j, (topic_num, prop_topic) in enumerate(row):
            if j == 0:  # => dominant topic
                wp = ldamodel.show_topic(topic_num)
                topic_keywords = ", ".join([word for word, prop in wp])
                sent_topics_df = sent_topics_df.append(pd.Series([int(topic_num), round(prop_topic,4), topic_ke
            else:
                break
    sent_topics_df.columns = ['Dominant_Topic', 'Perc_Contribution', 'Topic_Keywords']

    # Add original text to the end of the output
    contents = pd.Series(texts)
    sent_topics_df = pd.concat([sent_topics_df, contents], axis=1)
    return(sent_topics_df)
```

Rajah 4.10 Kod untuk mengeluarkan topik dominan untuk setiap ayat

	DocumentNo	Dominant_Topic	Perc_Contribution	Topic_Keywords	texts
176258	176258	9.0	0.6199	taste, make, creamer, add, brew, water, sugar,...	[star, chai, well, spice]
174152	174152	9.0	0.2604	taste, make, creamer, add, brew, water, sugar,...	[highly, gluten, also, highly, gluten, even, p...
423686	423686	4.0	0.3113	flavor, love, enjoy, taste, try, coffee, choco...	[husband, drink, terribly, vacation, wake, hot...
313902	313902	4.0	0.4300	flavor, love, enjoy, taste, try, coffee, choco...	[star, robust, definitely, hit, spot]
433202	433202	7.0	0.4558	cup, pod, work, machine, make, ground, use, ca...	[product, product, machine, worse, yet, work, ...
299732	299732	6.0	0.2801	drink, love, morning, day, recommend, taste, g...	[love, yourseldf, favor, first, also, soo, see...
426896	426896	6.0	0.7000	drink, love, morning, day, recommend, taste, g...	[highly, recommend]
273472	273472	0.0	0.3644	taste, bag, try, say, buy, review, think, get,...	[chai, tea, blend, try, savory]
108597	108597	3.0	0.4827	star, love, taste, price, product, thank, stuf...	[star, value, product, extremely]
128503	128503	2.0	0.5320	price, buy, find, love, taste, store, brand, q...	[taste, office, like, price, get, order, super...

Rajah 4.11 Contoh hasil dikeluarkan daripada topik dominan

4.5 TAFSIRAN TOPIK

Membaca hanya kata kunci topik mungkin tidak mencukupi untuk memahami topik itu. Untuk membantu memahami topik, projek ini mencari dokumen dengan topik yang paling banyak menyumbang dan menyimpulkan topik yang paling bermakna dengan membaca dokumen itu seperti yang ditunjukkan pada Rajah 4.12 dan Rajah 4.13.

```
# showing best relevant document under each topic
topic_sentences_df =pd.DataFrame()
df_topic_sents_grped=df_dominant_topic.groupby('Dominant_Topic')

for i,grp in df_topic_sents_grped:
    topic_sentences_df=pd.concat([topic_sentences_df,grp.sort_values(['Perc_Contribution'], ascending=[0]).head(1)], axis=0)


#reset index
topic_sentences_df.reset_index(drop=True,inplace=True)

#Format
topic_sentences_df.columns=['Document No','Topic_Num', "Topic_Perc_Contrib", "Keywords", "Text"]

topic_sentences_df.head()
```

Rajah 4.12 Kod untuk mencari dokumen paling berkaitan untuk setiap topik

	Document No	Topic_Num	Topic_Perc_Contrib	Keywords	Text
0	322461	0.0	0.9640	taste, try, buy, bag, say, review, think, get,...	[taste, crap, may, like, taste, crap, may, lik...
1	43044	1.0	0.9500	order, product, box, time, receive, package, p...	[probably, ship, error, box, arrive, today, da...
2	162417	2.0	0.9437	price, buy, love, find, store, brand, get, qua...	[love, finally, buy, couple, month, ago, can, ...
3	133236	3.0	0.9877	star, love, taste, price, value, product, than...	[love, love, love, love, love, love, love, lov...
4	425174	4.0	0.9308	flavor, variety, love, try, coffee, enjoy, pac...	[flavor, get, variety, pack, brand, flavor, ge...

Rajah 4.13 Contoh hasil untuk mencari dokumen paling berkaitan untuk setiap topik

6 topik yang diekstrak melalui model *Latent Dirichlet Allocation* diringkaskan dalam Jadual 4.6 di mana projek ini menggunakan kaedah oleh Guo et al. (2017) dan Hao et al. (2017) untuk mengestrak topik dan menamakan setiap topik. Sambungan logik antara kata-kata teratas mereka dan berat relatif yang sepadan (Rajah 4.8), *WordCloud* dalam Rajah 4.9 dan juga tafsiran kata kunci berdasarkan hasil pada dokumen yang paling relevan untuk setiap topik untuk menamakan setiap topik.

Jadual 4.6 Nama Topik dan Kata Kunci per Topik

Topik	Tafsiran	Kata Kunci per Topik
Topik 1	Kualiti kopi	taste, roast, flavor, try, blend, smell, buy, coffee, pretty, think, expect, better, say, brand, even, bit, look, aroma, prefer, give
Topik 2	Pembungkusan	cup, pod, work, product, order, box, machine, receive, package, time, come, get, ground, make, buy, money, gift, purchase, problem, packaging
Topik 3	Harga	price, flavor, love, buy, find, variety, try, brand, order, purchase, pack, cup, store, get, decaf, capsule, time, far, coffee, enjoy
Topik 4	Khidmat Pelanggan	star, love, taste, product, price, best, thank, delivery, buy, stuff, value, order, discount, time, service, coconut, always, arrive, fast, quality
Topik 5	Rasa kopi	flavor, love, taste, drink, morning, creamer, add, recommend, make, caffeine, day, sugar, enjoy, milk, tea, need, vanilla, cup, highly, mix
Topik 6	Biji Kopi	bean, make, brew, use, bag, get, try, product, drink, water, grind, buy, find, go, review, time, espresso, know, year, take

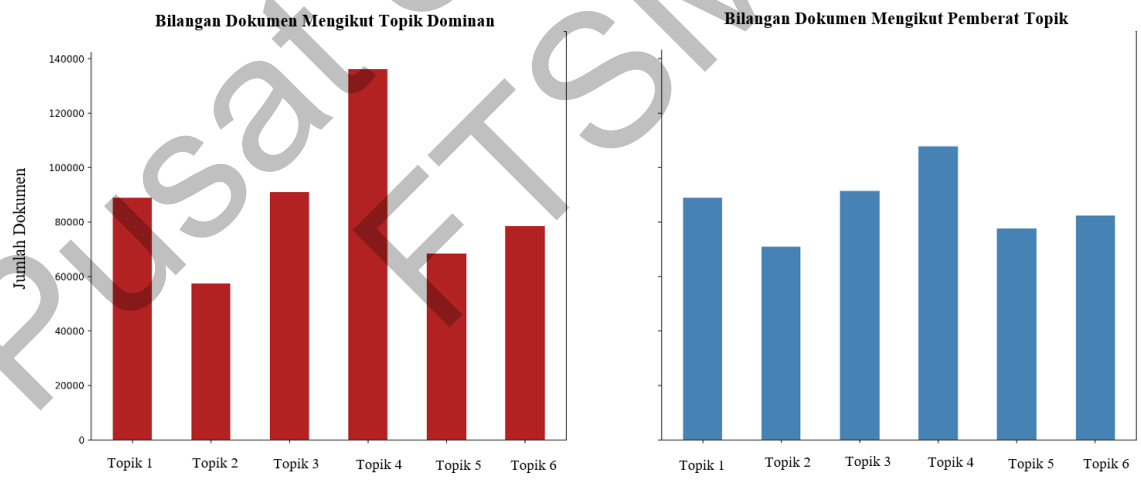
4.6 PEMBAHAGIAN TOPIK DI SELURUH DOKUMEN

Seterusnya, projek ini ingin memahami jumlah dan taburan setiap topik untuk menilai berapa banyak perbincangan masing-masing. Projek ini memaparkan jumlah dokumen untuk setiap topik dengan memberikan dokumen tersebut kepada topik yang mempunyai berat paling banyak dalam dokumen tersebut dan yang lain adalah jumlah dokumen untuk setiap topik dengan menjumlahkan sumbangan berat sebenar setiap topik ke dokumen masing-masing.

Berdasarkan kedua plot di Rajah 4.15, ulasan terbesar adalah topik 4 iaitu ‘Khidmat Pelanggan’. Kajian ini turut memahami bahawa banyak pelanggan akan memberikan banyak maklum balas kepada penjual atau pemilik perniagaan mengenai produk dan perkhidmatan mereka. Seterusnya, ‘Kualiti kopi’ dan ‘Harga’ adalah topik kedua dan ketiga terbanyak di Amazon.com diikuti oleh ‘Biji Kopi’, ‘Rasa kopi’ dan ‘Pembungkusan’.

```
def topics_per_document(model, corpus, start=0, end=1):
    corpus_sel = corpus[start:end]
    dominant_topics = []
    topic_percentages = []
    for i, corp in enumerate(corpus_sel):
        topic_percs, wordid_topics, wordid_phivalues = model[corp]
        dominant_topic = sorted(topic_percs, key = lambda x: x[1], reverse=True)[0][0]
        dominant_topics.append((i, dominant_topic))
        topic_percentages.append(topic_percs)
    return(dominant_topics, topic_percentages)

dominant_topics, topic_percentages = topics_per_document(model=lda_model, corpus=corpus, end=-1)

# Distribution of Dominant Topics in Each Document
df = pd.DataFrame(dominant_topics, columns=['Document_Id', 'Dominant_Topic'])
dominant_topic_in_each_doc = df.groupby('Dominant_Topic').size()
df_dominant_topic_in_each_doc = dominant_topic_in_each_doc.to_frame(name='count').reset_index()

# Total Topic Distribution by actual weight
topic_weightage_by_doc = pd.DataFrame([dict(t) for t in topic_percentages])
df_topic_weightage_by_doc = topic_weightage_by_doc.sum().to_frame(name='count').reset_index()

# Top 3 Keywords for each Topic
topic_top3words = [(i, topic) for i, topics in lda_model.show_topics(formatted=False)
                                 for j, (topic, wt) in enumerate(topics) if j < 3]

df_top3words_stacked = pd.DataFrame(topic_top3words, columns=['topic_id', 'words'])
df_top3words = df_top3words_stacked.groupby('topic_id').agg(', \n'.join)
df_top3words.reset_index(level=0,inplace=True)
```

Rajah 4.14 Kod Taburan Topik Merentasi Dokumen



Rajah 4.15 Pembahagian plot berdasarkan topik dominan dan pemberat topik

4.7 ANALISIS SENTIMEN BERASASKAN CIRI

Projek ini bertujuan untuk mengekstrak ciri kopi menggunakan analisis sentimen berasaskan ciri untuk mengetahui pilihan dan tingkah laku pengguna kopi terhadap produk kopi di Amazon.com. Setelah mengeluarkan topik dari *Latent Dirichlet Allocation*, topik tersebut dianggap sebagai ciri kopi. Seterusnya, untuk mengetahui pilihan dan tingkah laku pelanggan terhadap ciri kopi, projek ini menggunakan analisis sentimen VADER untuk melabelkan ciri kopi sebagai ciri positif, neutral dan negatif. Berdasarkan Jadual 4.7, semua ciri kopi adalah ciri positif kerana sentimen positifnya mempunyai jumlah tinjauan tertinggi berbanding dengan sentimen negatif dan neutral.

Jadual 4.7 Ringkasan Analisis Sentimen Berasaskan Ciri

Ciri	**Negatif**		**Neutral**		**Positif**		**Sentimen Ciri**
	Bilangan Ulasan	Skor Purata Sentimen	Bilangan Ulasan	Skor Purata Sentimen	Bilangan Ulasan	Skor Purata Sentimen	
Kualiti kopi	8,030	-0.5101	4,294	0.0002	71,431	0.7512	Positif
Pembungkusan	5,164	-0.5146	2,769	0.0003	45,619	0.7501	Positif
Harga	8,125	-0.5133	4,373	0.0002	72,988	0.7523	Positif
Khidmat Pelanggan	12,152	-0.5093	6,624	-0.0002	108,333	0.7493	Positif
Rasa kopi	6,037	-0.5134	3,388	0.0002	54,296	0.7520	Positif
Biji Kopi	6,946	-0.5134	3,595	0.0001	63,229	0.7561	Positif

BAB V

RUMUSAN DAN CADANGAN MASA DEPAN

5.1 RUMUSAN

Lebih dari 16,275 produk kopi, projek ini mengkaji pilihan dan tingkah laku pengguna di pasaran kopi Amazon. Ketika coronavirus menjadi pandemik global pada bulan Mac 2020 dan kemajuan teknologi, Amazon menyaksikan lonjakan permintaan awal dalam industri seperti runcit, di mana orang lebih bergantung pada pembelian dalam talian daripada sebelumnya menyebabkan keperluan mendesak untuk mengumpulkan lebih banyak maklumat mengenai pilihan pengguna dan mengembangkan produk yang memenuhi keperluan mereka.

Projek ini menggunakan pendekatan pemodelan sentimen berasaskan ciri menggunakan *Latent Dirichlet Allocation* (LDA) dan analisis sentimen *VADER* untuk mengekstrak ciri produk dan mengetahui pilihan dan tingkah laku pengguna terhadap produk kopi di Amazon. Di samping itu, projek ini juga meneliti peningkatan model LDA melalui kombinasi yang berbeza dari *Part of Speech Tagging* (POSTAG) dan mengoptimumkan jumlah topik, *K*. Hasil kajian menunjukkan bahawa Kata nama, Kata kerja dan Kata keterangan adalah gabungan terbaik POSTAG, dan bilangan topik yang optimum *K* adalah 6. Ciri kopi yang diekstrak dari model tersebut adalah ‘Kualiti Kopi’, ‘Pembungkusan’, ‘Harga’, ‘Khidmate Pelanggan’, ‘Rasa Kopi’ dan ‘Biji Kopi’ dan semua ciri ini adalah ciri positif kerana sentimen positif mereka memiliki jumlah tinjauan yang tertinggi berbanding dengan sentimen negatif dan neutral. Di samping itu, projek ini juga membincangkan mengenai keperluan kekuatan komputasi ketika berurusan dengan korpus besar dan menggunakan model LDA untuk meningkatkan kelajuan masa pemprosesan.

5.2 CADANGAN UNTUK KAJIAN MASA HADAPAN

Untuk cadangan masa depan, projek ini akan mencadangkan untuk membina Pelabelan Topik Automatik untuk menamakan setiap topik tanpa memerlukan pertimbangan manusia. Selain itu, terdapat banyak jenis *Part of Speech Tagging* dan kajian ini telah melakukan kombinasi pada kata nama, kata kerja, kata keterangan dan kata adjektif untuk membina model LDA, kajian yang akan datang boleh mencuba kombinasi lain untuk mengekstrak kata kunci dan menilai prestasi model. Akhir sekali, selain jumlah topik *K*, terdapat hiperparameter lain untuk dikaji bagi melihat dan melakukan pengoptimuman untuk meningkatkan lagi model LDA.

RUJUKAN

Bhowmik, T. *et al.* (2015) 'Leveraging topic modeling and part-of-speech tagging to support combinational creativity in requirements engineering', *Requirements Engineering*, 20(3), pp. 253–280. doi: 10.1007/s00766-015-0226-2.

Bifet, A. *et al.* (2015) *Efficient Online Evaluation of Big Data Stream Classifiers*. doi: 10.1145/2783258.2783372.

Bonta, V. and Janardhan, N. K. N. (2019) 'A comprehensive study on lexicon based approaches for sentimen analysis', *Asian Journal of Computer Science and Technology*, 8(S2), pp. 1–6.

Bouma, G. (2009) 'Normalized (pointwise) mutual information in collocation extraction', *Proceedings of GSCL*, pp. 31–40.

Browne, M. (2020) *Amazon online grocery sales triple in second quarter, Supermarket News*. Available at: https://www.supermarketnews.com/online-retail/amazon-online-grocery-sales-triple-second-quarter.

Chakraborty, G. (2014) *Analysis of Unstructured Data: Applications of Text Analytics and Sentimen Mining*.

Chen, S. (2021) 'Keyword Extraction for Privacy Policy Analysis Using Topic Modelling Approaches'.

Chen, Y., Chen, L. and Takama, Y. (2015) 'Proposal of LDA-Based Sentimen Visualization of Hotel Reviews', in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 687–693. doi: 10.1109/ICDMW.2015.72.

Darling, W. M., Paul, M. and Song, F. (2012) 'Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm', in *Proceedings of the Workshop on Semantic Analysis in Social Media*, pp. 1–9.

Darling, W. M. and Song, F. (2013) 'Probabilistic topic and syntax modeling with part-of-speech LDA', *arXiv preprint arXiv:1303.2826*.

David Court, Dave Elzinga, Susan Mulder, and O. J. V. (2009) *The consumer decision journey, Mckinsey*. Available at: https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-consumer-decision-journey.

Ding, K. *et al.* (2020) 'Employing structural topic modelling to explore perceived service quality attributes in Airbnb accommodation', *International Journal of Hospitality Management*, 91, p. 102676. doi: https://doi.org/10.1016/j.ijhm.2020.102676.

Forrester (2016) *Forrester Data Web-Influenced Retail Sales Forecast, 2016 To 2021 (EU-7)*. Available at: Forrester Data Web-Influenced Retail Sales Forecast, 2016 To 2021 (EU-7).

García-Pablos, A., Cuadros, M. and Rigau, G. (2018) 'W2VLDA: almost unsupervised system for aspect based sentimen analysis', *Expert Systems with Applications*. Elsevier, 91, pp. 127–137.

Hatami, A., Akbari, A. and Nasersharif, B. (2013) 'N-gram adaptation using Dirichlet class language model based on part-of-speech for speech recognition', in *2013 21st Iranian Conference on Electrical Engineering (ICEE)*. IEEE, pp. 1–5.

Heng, Y. *et al.* (2018) 'Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach', *Journal of Retailing and Consumer Services*. Elsevier, 42, pp. 161–168.

Heng, Y., Chen, J. and Gao, Z. (2019) 'A Topic Mining Approach to Understand What Matters to Online Grocery Consumers: the Case of Coconut Oil'.

Irawan, H., Akmalia, G. and Masrury, R. (2019) *Mining Tourist's Perception toward Indonesia Tourism Destination Using Sentimen Analysis and Topic Modelling, CCIOT 2019: Proceedings of the 2019 4th International Conference on Cloud Computing and Internet of Things*. doi: 10.1145/3361821.3361829.

Jacobi, C., van Atteveldt, W. and Welbers, K. (2016) 'Quantitative analysis of large amounts of journalistic texts using topic modelling', *Digital Journalism*. Routledge, 4(1), pp. 89–106. doi: 10.1080/21670811.2015.1093271.

Jauhari, T. M. *et al.* (2020) 'Assessing Customer Needs Based On Online Reviews: A Topic Modeling Approach', in *CEUR Workshop Proceedings*. CEUR-WS, pp. 57–62.

Jo, Y. and Oh, A. H. (2011) 'Aspect and sentimen unification model for online review analysis', in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 815–824.

Kwon, H.-J. *et al.* (2021) 'Topic Modeling and Sentimen Analysis of Online Review for Airlines', *Information* . doi: 10.3390/info12020078.

Li, R. *et al.* (2021) 'Meal Kit Preferences during COVID-19 Pandemic: Exploring User-Generated Content with Natural Language Processing Techniques'.

Lin, C. and He, Y. (2009) 'Joint sentimen/topic model for sentimen analysis', in *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 375–384.

Liu, Y., Bi, J.-W. and Fan, Z.-P. (2017) 'Ranking products through online reviews: A method based on sentimen analysis technique and intuitionistic fuzzy set theory', *Information Fusion*. Elsevier, 36, pp. 149–161.

Lucini, F. R. *et al.* (2020) 'Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews', *Journal of Air Transport Management*. Elsevier, 83, p. 101760.

Luo, J. M. *et al.* (2020) 'Topic modelling for theme park online reviews: analysis of Disneyland', *Journal of Travel & Tourism Marketing*. Routledge, 37(2), pp. 272–285. doi: 10.1080/10548408.2020.1740138.

Martin, F. and Johnson, M. (2015) 'More efficient topic modelling through a noun only approach', in *Proceedings of the Australasian Language Technology Association Workshop 2015*, pp. 111–115.

McAuley, J. and Leskovec, J. (2013) 'Hidden factors and hidden topics: understanding rating dimensions with review text', in *Proceedings of the 7th ACM conference on Recommender systems*, pp. 165–172.

Mou, J. *et al.* (2019) 'Understanding the topics of export cross-border e-commerce consumers feedback: an LDA approach', *Electronic Commerce Research*. Springer, 19(4), pp. 749–777.

Nguyen, H. *et al.* (2018) 'Comparative study of sentimen analysis with product reviews using machine learning and lexicon-based approaches', *SMU Data Science Review*, 1(4), p. 7.

Ni, J., Li, J. and McAuley, J. (2019) *Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects*. doi: 10.18653/v1/D19-1018.

Nielsen (2018) *Total Consumer Report*. United States. Available at: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/viewer.html?pdfurl=https%3A%2F%2Fwww.nielsen.com%2Fwp-content%2Fuploads%2Fsites%2F3%2F2019%2F04%2Ftotal-consumer-report-june-2018.pdf&clen=2825634&chunk=true.

Research, C. (2020) *US Online Grocery Survey 2020: Many More Shoppers Buying More Categories from More Retailers*. Available at: http://coresight.com/research/us-online-grocery-survey-2020-many-more-shoppers-buying-more-categories-from-more-retailers/.

Röder, M., Both, A. and Hinneburg, A. (2015) 'Exploring the space of topic coherence measures', in *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408.

Sahin, E. K. (2020) 'Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest', *SN Applied Sciences*, 2(7), p. 1308. doi: 10.1007/s42452-020-3060-1.

Stankevich, A. (2017) 'Explaining the consumer decision-making process: Critical literature review', *Journal of International Business Research and Marketing*, 2(6).

Stevens, K. *et al.* (2012) 'Exploring topic coherence over many models and many topics', in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 952–961.

Sutherland, I. *et al.* (2020) 'Topic modeling of online accommodation reviews via latent dirichlet allocation', *Sustainability*. Multidisciplinary Digital Publishing Institute, 12(5), p. 1821.

Tang, F. *et al.* (2019) 'Aspect based fine-grained sentimen analysis for online reviews', *Information Sciences*. Elsevier, 488, pp. 190–204.

US Census Bureau (2020) *Quarterly Retail E-Commerce Sales 2nd Quarter 2020*. Available at: https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf.

Usop, E. S., Isnanto, R. R. and Kusumaningrum, R. (2017) 'Part of speech features for sentimen classification based on Latent Dirichlet Allocation', in *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pp. 31–34. doi: 10.1109/ICITACEE.2017.8257670.

Vorontsov, K. *et al.* (2015) 'BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections BT - Analysis of Images, Social Networks and Texts', in Khachay, M. Y. et al. (eds). Cham: Springer International Publishing, pp. 370–381.

Wang, B. *et al.* (2014) 'Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology', *Scientometrics*. Springer, 101(1), pp. 685–704.

Wang, W., Feng, Y. and Dai, W. (2018) 'Topic analysis of online reviews for two competitive products using latent Dirichlet allocation', *Electronic Commerce Research and Applications*, 29, pp. 142–156. doi: https://doi.org/10.1016/j.elerap.2018.04.003.

Westerlund, M. *et al.* (2019) 'Topic modelling analysis of online reviews: Indian restaurants at Amazon. Com', in *ISPIM Conference Proceedings*. The International Society for Professional Innovation Management (ISPIM), pp. 1–14.

Xianghua, F. *et al.* (2013) 'Multi-aspect sentimen analysis for Chinese online social reviews based on topic modeling and HowNet lexicon', *Knowledge-Based Systems*. Elsevier, 37, pp. 186–195.

Xiong, S. *et al.* (2018) 'A short text sentimen-topic model for product reviews', *Neurocomputing*. Elsevier, 297, pp. 94–102.

LAMPIRAN A

KOD SUMBER UNTUK KESELURUHAN PROJEK

PREPARING DATA

In [1]:

```
import os
import json
import gzip
import pandas as pd
from urllib.request import urlopen

import random
import numpy as np
from tqdm import tqdm_notebook as tqdm
from collections import defaultdict
```

In [2]:

```
!wget http://deepyeti.ucsd.edu/jianmo/amazon/metaFiles2/meta_Grocery_and_Gourmet_Food.json.gz
```

```
'wget' is not recognized as an internal or external command,
operable program or batch file.
```

In [3]:

```
#Load the meta data

data = []
with gzip.open('meta_Grocery_and_Gourmet_Food.json.gz', 'r') as f:
    for l in tqdm(f):
        data.append(json.loads(l))
```

```
C:\Users\Khadijah Jasni\Anaconda3\lib\site-packages\ipykernel_launcher.py:5: TqdmDeprecationWarning: This function will be removed in tqdm==5.0.0
Please use `tqdm.notebook.tqdm` instead of `tqdm.tqdm_notebook`
  """
```

In [4]:

```
# Convert metadata list into pandas dataframe

df = pd.DataFrame.from_dict(data)
df.head()
```

Out[4]:

	category	tech1	description	fit	title	also_buy	tech2	brand	feature	rank	also_view	main_
0	[Grocery & Gourmet Food, Dairy, Cheese & Eggs,...		[BEEMSTER GOUDA CHEESE AGED 18/24 MONTHS, Stat...		Beemster Gouda - Aged 18/24 Months - App. 1.5 Lbs	[]		Ariola Imports	[]	165,181 in Grocery & Gourmet Food (	[B0000D9MYM, B0000D9MYL, B00ADHIGBA, B00H9OX59...	Groc
1	[Grocery & Gourmet Food, Cooking & Baking, Sug...		[Shipped from UK, please allow 10 to 21 busine...		Trim Healthy Mama Xylitol	[B01898YHXK, B01BCM6LAC, B00Q4OL47O, B00Q4OL5Q...			[]	315,867 in Grocery & Gourmet Food (	[]	Groc
	[Grocery &		[Jazz up		Letter C -					[>#669,941		

	category	tech1	description	fit	title	also_buy	tech2	brand	feature	rank	also_view	main_cat
2	[Grocery & Gourmet Food, Cooking & Baking, Fro...		your cakes with a sparkling monogram ...		Swarovski Crystal Monogram Wedding ...	[]		Unik Occasions	[]	[>#832,581 in Kitchen & Dining (See Top 100 in...	[B07DXN65TF]	Ama: Ho
3	[Grocery & Gourmet Food, Cooking & Baking, Fro...		[Large Letter - Height 4.75"]		Letter H - Swarovski Crystal Monogram Wedding ...	[]		Other	[Large Letter - Height 4.75"]	[>#832,581 in Kitchen & Dining (See Top 100 in...	[]	Ama: Ho
4	[Grocery & Gourmet Food, Cooking & Baking, Fro...		[4.75"]		Letter S - Swarovski Crystal Monogram Wedding ...	[]		Unik Occasions	[4.75" height]	[>#590,999 in Kitchen & Dining (See Top 100 in...	[]	Ama: Ho

In [5]:

```
#Extracting coffee from TITLE column

metadata = df[df["title"].str.contains("coffee|Coffee|Coffees|coffees|COFFEE|COFFEES|kopi
|KOPI|Kopi|Café|Kaffee|Kofe|Кофе|Koffie|Kahvi|Kafés|Καφές|Kofee|Kaffi|Cafea|Kaffe")]

metadata.shape
```

Out[5]:

```
(18411, 19)
```

Merge Metadata with Review File

In [6]:

```
!wget http://deepyeti.ucsd.edu/jianmo/amazon/categoryFiles/Grocery_and_Gourmet_Food.json
.gz
```

```
'wget' is not recognized as an internal or external command,
operable program or batch file.
```

In [7]:

```
#Load review data

review = []
with gzip.open('Grocery_and_Gourmet_Food.json.gz', 'r') as f:
    for l in tqdm(f):
        review.append(json.loads(l))
```

```
C:\Users\Khadijah Jasni\Anaconda3\lib\site-packages\ipykernel_launcher.py:5: TqdmDeprecat
ionWarning: This function will be removed in tqdm==5.0.0
Please use `tqdm.notebook.tqdm` instead of `tqdm.tqdm_notebook`
  """
```

In [8]:

```
# convert list into pandas dataframe

df_review= pd.DataFrame.from_dict(review)
df_review.head()
```

Out[8]:

	overall	verified	reviewTime	reviewerID	asin	reviewerName	reviewText	summary	unixReviewTime	v
0	5.0	True	06 4, 2013	ALP49FBWT4I7V	1888861614	Lori	Very pleased with my purchase. Looks exactly l...	Love it	1370304000	I
1	4.0	True	05 23, 2014	A1KPIZOCLB9FZ8	1888861614	BK Shopper	Very nicely crafted but too small. Am going to...	Nice but small	1400803200	I
2	4.0	True	05 9, 2014	A2W0FA06IYAYQE	1888861614	daninethequeen	still very pretty and well made...i am super p...	the "s" looks like a 5, kina	1399593600	I
3	5.0	True	04 20, 2014	A2PTZTCH2QUYBC	1888861614	Tammara	I got this for our wedding cake, and it was ev...	Would recommend this to a friend!	1397952000	I
4	4.0	True	04 16, 2014	A2VNHGJ59N4Z90	1888861614	LaQuinta Alexander	It was just what I want to put at the top of m...	Topper	1397606400	I

In [9]:

```
#Merge both files

coffee=pd.merge(metadata, df_review,on='asin',how='left')

coffee.head()
```

Out[9]:

	category	tech1	description	fit	title	also_buy	tech2	brand	feature	rank	...	verified	re
0	[Grocery & Gourmet Food, Beverages, Coffee, Te...		[]		IKEA - BRYGGKAFFE MELLANROST Decaffeinated Cof...	[]			[]	1,229,560 in Grocery & Gourmet Food (	...	False	0
1	[Grocery & Gourmet Food, Beverages, Coffee, Te...		[Coffee Whole Beans Dark Roast, Statements reg...		IKEA - KAFFE HELA BÖNOR MÖRKROST Cof...	[]		IKEA	[]	676,857 in Grocery & Gourmet Food (	...	True	1
2	[Grocery & Gourmet Food, Beverages, Coffee, Te...		[This replacement container for Vacu Vin's Cof...		Vacu Vin Coffee Saver Refill Container	[]		Vacu Vin	[Replacement container for Vacu Vin Coffee Sav...	[>#1,646,318 in Kitchen & Dining (See Top 100 ...	...	False	
3	[Grocery & Gourmet Food, Beverages, Coffee, Te...		[This replacement container for Vacu Vin's Cof...		Vacu Vin Coffee Saver Refill Container	[]		Vacu Vin	[Replacement container for Vacu Vin Coffee Sav...	[>#1,646,318 in Kitchen & Dining (See Top 100 ...	...	True	1

	category	tech1	description	fit	title	also_buy	tech2	brand	feature	rank	...	verified	re
4	[Grocery & Gourmet Food, Beverages, Coffee, Te...		[This replacement container for Vacu Vin's Cof...		Vacu Vin Coffee Saver Refill Container	[]		Vacu Vin	[Replacement container for Vacu Vin Coffee Sav...	[>#1,646,318 in Kitchen & Dining (See Top 100 ...	...	True	

5 rows × 30 columns

In [10]:

```
coffee.shape
```

Out[10]:

```
(555639, 30)
```

In [11]:

```
coffee['category'] = coffee['category'].astype(str).str.replace(r'\[|\]|,', '')
coffee['description'] = coffee['description'].astype(str).str.replace(r'\[|\]|,', '')
coffee['feature'] = coffee['feature'].astype(str).str.replace(r'\[|\]|,', '')
coffee['rank'] = coffee['rank'].astype(str).str.replace(r'\[|\]|,', '')

coffee.head()
```

```
C:\Users\Khadijah Jasni\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: FutureWarnin
g: The default value of regex will change from True to False in a future version.
  """Entry point for launching an IPython kernel.
C:\Users\Khadijah Jasni\Anaconda3\lib\site-packages\ipykernel_launcher.py:2: FutureWarnin
g: The default value of regex will change from True to False in a future version.

C:\Users\Khadijah Jasni\Anaconda3\lib\site-packages\ipykernel_launcher.py:3: FutureWarnin
g: The default value of regex will change from True to False in a future version.
  This is separate from the ipykernel package so we can avoid doing imports until
C:\Users\Khadijah Jasni\Anaconda3\lib\site-packages\ipykernel_launcher.py:4: FutureWarnin
g: The default value of regex will change from True to False in a future version.
  after removing the cwd from sys.path.
```

Out[11]:

	category	tech1	description	fit	title	also_buy	tech2	brand	feature	rank	...	verified	rev
0	'Grocery & Gourmet Food' 'Beverages' 'Coffee T...				IKEA - BRYGGKAFFE MELLANROST Decaffeinated Cof...	[]				1229560 in Grocery & Gourmet Food (	...	False	01
1	'Grocery & Gourmet Food' 'Beverages' 'Coffee T...		'Coffee Whole Beans Dark Roast' 'Statements re...		IKEA - KAFFE HELA BÖNOR MÖRKROST Cof...	[]		IKEA		676857 in Grocery & Gourmet Food (	...	True	11
2	'Grocery & Gourmet Food' 'Beverages' 'Coffee T...		"This replacement container for Vacu Vin's Cof...		Vacu Vin Coffee Saver Refill Container	[]		Vacu Vin	'Replacement container for Vacu Vin Coffee Sav...	'>#1646318 in Kitchen & Dining (See Top 100 in...	...	False	0
3	'Grocery & Gourmet Food' 'Beverages' 'Coffee T...		"This replacement container for Vacu Vin's Cof...		Vacu Vin Coffee Saver Refill Container	[]		Vacu Vin	'Replacement container for Vacu Vin Coffee Sav...	'>#1646318 in Kitchen & Dining (See Top 100 in...	...	True	12

	category	tech1	description	fit	title	also_buy	tech2	brand	feature	rank	...	verified	rev
4	'Grocery & Gourmet Food' 'Beverages' 'Coffee T...		"This replacement container for Vacu Vin's Cof...		Vacu Vin Coffee Saver Refill Container	[]		Vacu Vin	'Replacement container for Vacu Vin Coffee Sav...	'>#1646318 in Kitchen & Dining (See Top 100 in...	...	True	0

5 rows × 30 columns

In [12]:

```
#Select GROCERY from main_Category

coffee1 = coffee[coffee["main_cat"].str.contains("Grocery")]

coffee1.shape
```

Out[12]:

```
(538460, 30)
```

In [13]:

```
#Extracting BEVERAGE from category

coffee2 = coffee1[coffee1["category"].str.contains("Beverages|beverage|beverages|Beverage
")]

coffee2.shape
```

Out[13]:

```
(520384, 30)
```

In [14]:

```
#Check missing values

coffee2.isnull().sum()
```

Out[14]:

```
category                0
tech1                   0
description             0
fit                     0
title                   0
also_buy                0
tech2                   0
brand                   0
feature                 0
rank                    0
also_view               0
main_cat                0
similar_item            0
date                    0
price                   0
asin                    0
imageURL                0
imageURLHighRes         0
details                 0
overall                 1
verified                1
reviewTime              1
reviewerID              1
reviewerName           22
reviewText            245
summary               115
unixReviewTime          1
vote               475429
image              515621
style              218914
dtype: int64
```

```
dtype: int64
```

In [15]:

```
#Dropping variables: tech1, tech2, also_buy, also_view, similar_item, imageURL, imageURLH
ighRes, details, rank,price

coffee3=coffee2.drop(['category','tech1','fit', 'tech2', 'also_buy', 'also_view','main_c
at', 'similar_item', 'imageURL', 'imageURLHighRes','details','verified','reviewerName','u
nixReviewTime','vote','image','style','rank','price'], axis=1)

coffee3.shape
```

Out[15]:

```
(520384, 11)
```

In [16]:

```
# Drop missing values from row

coffee4=coffee3.dropna(subset=['overall','reviewTime','reviewerID','reviewText','summary
'])

coffee4.shape
```

Out[16]:

```
(520038, 11)
```

In [17]:

```
#Check missing values

coffee4.isnull().sum()
```

Out[17]:

```
description     0
title           0
brand           0
feature         0
date            0
asin            0
overall         0
reviewTime      0
reviewerID      0
reviewText      0
summary         0
dtype: int64
```

In [18]:

```
# Conconcateneate reviewtext and summary

coffee4['review_text'] = coffee4[['summary', 'reviewText']].apply(lambda x: " ".join(str
(y) for y in x if str(y) != 'nan'), axis = 1)
coffee5 = coffee4.drop(['reviewText', 'summary'], axis = 1)

coffee5.head()
```

```
C:\Users\Khadijah Jasni\Anaconda3\lib\site-packages\ipykernel_launcher.py:3: SettingWithC
opyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_g
uide/indexing.html#returning-a-view-versus-a-copy
  This is separate from the ipykernel package so we can avoid doing imports until
```

Out[18]:

	description	title	brand	feature	date	asin	overall	reviewTime	reviewerID	re

	description	title	brand	feature	date	asin	overall	reviewTime	reviewerID	r
0		IKEA - BRYGGKAFFE MELLANROST Decaffeinated Cof...				9178894018	5.0	01 27, 2013	AGPPE7SALG8D	
1	'Coffee Whole Beans Dark Roast' 'Statements re...	IKEA - KAFFE HELA BÖNOR MÖRKROST Cof...	IKEA			9178902568	4.0	11 23, 2014	AY8J9E7FHV400	
23	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	1.0	06 4, 2002	A1EWE4NF4LNKS2	C ex
24	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	5.0	08 9, 2013	A1LDTUKPKOHHLG	p
25	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	4.0	07 14, 2013	A2CN1RIGQ8ECPZ	e

In [19]:

```
# change column name
coffee5 = coffee5.rename(columns={'overall': 'Rating'})

print ("Total data:", str(coffee5.shape))
coffee5.head()
```

```
Total data: (520038, 10)
```

Out[19]:

	description	title	brand	feature	date	asin	Rating	reviewTime	reviewerID	re
0		IKEA - BRYGGKAFFE MELLANROST Decaffeinated Cof...				9178894018	5.0	01 27, 2013	AGPPE7SALG8D	
1	'Coffee Whole Beans Dark Roast' 'Statements re...	IKEA - KAFFE HELA BÖNOR MÖRKROST Cof...	IKEA			9178902568	4.0	11 23, 2014	AY8J9E7FHV400	
23	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	1.0	06 4, 2002	A1EWE4NF4LNKS2	C ex
24	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	5.0	08 9, 2013	A1LDTUKPKOHHLG	p
	'Espressione			'Pods						e

	description	title	brand	feature	date	asin	Rating	reviewTime	reviewerID	re
25	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	4.0	07 14, 2013	A2CN1RIGQ8ECPZ	f

In [20]:

```
# Convert time object to datetime and create a new column named 'time'#

coffee5['time'] = coffee5.reviewTime.str.replace(',', "")
coffee5['time'] = pd.to_datetime(coffee5['time'], format = '%m %d %Y')

# Drop redundant 'reviewTime' column

coffee6 = coffee5.drop('reviewTime', axis = 1)

coffee6.head()
```

Out[20]:

	description	title	brand	feature	date	asin	Rating	reviewerID	review_text	t
0		IKEA - BRYGGKAFFE MELLANROST Decaffeinated Cof...				9178894018	5.0	AGPPE7SALG8D	Strong but not bitter I had a cup while on bre...	20 07
1	'Coffee Whole Beans Dark Roast' 'Statements re...	IKEA - KAFFE HELA BÖNOR MÖRKROST Cof...	IKEA			9178902568	4.0	AY8J9E7FHV400	Four Stars Excellent for espresso.	20 11
23	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	1.0	A1EWE4NF4LNKS2	Cappuccino without flavor I am an experienced ...	20 06
24	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	5.0	A1LDTUKPKOHHLG	Was skeptical about minimum amout to purchase ...	20 08
25	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	4.0	A2CN1RIGQ8ECPZ	easy to use No mess and good flavor. The pods ...	20 07

In [21]:

```
# Create new columns

coffee6['day'] = coffee6['time'].dt.day
coffee6['month'] = coffee6['time'].dt.month
coffee6['year'] = coffee6['time'].dt.year

coffee6.head()
```

Out[21]:

	description	title	brand	feature	date	asin	Rating	reviewerID	review_text	t
		IKEA - BRYGGKAFFE							Strong but not bitter I	

	description	title	brand	feature	date	asin	Rating	reviewerID	review_text	t
0		IKEA - BRYGGKAFFE MELLANROST Decaffeinated Cof...				9178894018	5.0	AGPPE7SALG8D	Strong but not bitter I had a cup while on bre...	2(0
1	'Coffee Whole Beans Dark Roast' 'Statements re...	IKEA - KAFFE HELA BÖNOR MÖRKROST Cof...	IKEA			9178902568	4.0	AY8J9E7FHV400	Four Stars Excellent for espresso.	2(1
23	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	1.0	A1EWE4NF4LNKS2	Cappuccino without flavor I am an experienced ...	2(0
24	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	5.0	A1LDTUKPKOHHLG	Was skeptical about minimum amout to purchase ...	2(0
25	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	4.0	A2CN1RIGQ8ECPZ	easy to use No mess and good flavor. The pods ...	2(0

In [22]:

```
# Classify ratings as good

good_rate = len(coffee6[coffee6['Rating'] >= 3])
bad_rate = len(coffee6[coffee6['Rating'] < 3])

# Printing rates and their total numbers
print ('Good ratings : {} reviews for coffee products'.format(good_rate))
print ('Bad ratings : {} reviews for coffee products'.format(bad_rate))
```

```
Good ratings : 458259 reviews for coffee products
Bad ratings : 61779 reviews for coffee products
```

In [23]:

```
# Apply the new classification to the ratings column##

coffee6['rating_class'] = coffee6['Rating'].apply(lambda x: 'bad' if x < 3 else'good')
coffee6.head()
```

Out[23]:

	description	title	brand	feature	date	asin	Rating	reviewerID	review_text	t
0		IKEA - BRYGGKAFFE MELLANROST Decaffeinated Cof...				9178894018	5.0	AGPPE7SALG8D	Strong but not bitter I had a cup while on bre...	2(0
1	'Coffee Whole Beans Dark Roast' 'Statements re...	IKEA - KAFFE HELA BÖNOR MÖRKROST Cof...	IKEA			9178902568	4.0	AY8J9E7FHV400	Four Stars Excellent for espresso.	2(1
23	'Espressione Classic Espresso is a rich blend	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the	June 20, 2004	B00005IX97	1.0	A1EWE4NF4LNKS2	Cappuccino without flavor I am an experienced	2(0

	description	title	brand	feature	date	asin	Rating	reviewerID	review_text	t
24	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	5.0	A1LDTUKPKOHHLG	Was skeptical about minimum amout to purchase ...	2(0(
25	'Espressione Classic Espresso is a rich blend ...	Espressione 100% Arabica Coffee, 150-Count Pods	Espressione	'Pods compatible with all machines using the E...	June 20, 2004	B00005IX97	4.0	A2CN1RIGQ8ECPZ	easy to use No mess and good flavor. The pods ...	2(0(

In [24]:

```
coffee6.shape
```

Out[24]:

```
(520038, 14)
```

In [27]:

```
#DESCRIPTIVE STATISTICS

print ("================================================")

# Total reviews
total = len(coffee6)
print ("Number of reviews: ",total)
print ()

# How many unique products?
print ("Number of unique products: ", len(coffee6.asin.unique()))
product_prop = float(len(coffee6.asin.unique())/total)
print ("Prop of unique products: ",round(product_prop,3))
print ()

# How many unique brands?
print ("Number of unique brands: ", len(coffee6.brand.unique()))
product_prop = float(len(coffee6.brand.unique())/total)
print ("Prop of unique brands: ",round(product_prop,3))
print ()

# Average star score
print ("Average rating score: ",round(coffee6.Rating.mean(),3))

print ("================================================")
```

```
================================================
Number of reviews:  520038

Number of unique products:  16275
Prop of unique products:  0.031

Number of unique brands:  2484
Prop of unique brands:  0.005

Average rating score:  4.307
================================================
```

In [28]:

```
coffee6.to_csv(r'C:\Users\Khadijah Jasni\Desktop\rawcoffee_review.csv')
```

In [29]:

```
!pip install plotly
```

```
Requirement already satisfied: plotly in c:\users\khadijah jasni\anaconda3\lib\site-packa
ges (5.2.1)
Requirement already satisfied: six in c:\users\khadijah jasni\anaconda3\lib\site-packages
(from plotly) (1.16.0)
Requirement already satisfied: tenacity>=6.2.0 in c:\users\khadijah jasni\anaconda3\lib\s
ite-packages (from plotly) (8.0.1)
```

```
WARNING: Ignoring invalid distribution -umpy (c:\users\khadijah jasni\anaconda3\lib\site-
packages)
WARNING: Ignoring invalid distribution -cipy (c:\users\khadijah jasni\anaconda3\lib\site-
packages)
WARNING: Ignoring invalid distribution -umpy (c:\users\khadijah jasni\anaconda3\lib\site-
packages)
WARNING: Ignoring invalid distribution -cipy (c:\users\khadijah jasni\anaconda3\lib\site-
packages)
WARNING: Ignoring invalid distribution -umpy (c:\users\khadijah jasni\anaconda3\lib\site-
packages)
WARNING: Ignoring invalid distribution -cipy (c:\users\khadijah jasni\anaconda3\lib\site-
packages)
WARNING: Ignoring invalid distribution -umpy (c:\users\khadijah jasni\anaconda3\lib\site-
packages)
WARNING: Ignoring invalid distribution -cipy (c:\users\khadijah jasni\anaconda3\lib\site-
packages)
WARNING: Ignoring invalid distribution -umpy (c:\users\khadijah jasni\anaconda3\lib\site-
packages)
WARNING: Ignoring invalid distribution -cipy (c:\users\khadijah jasni\anaconda3\lib\site-
packages)
```

In [30]:

```
!pip install cufflinks
```

```
Requirement already satisfied: cufflinks in c:\users\khadijah jasni\anaconda3\lib\site-pa
ckages (0.17.3)
Requirement already satisfied: plotly>=4.1.1 in c:\users\khadijah jasni\anaconda3\lib\sit
e-packages (from cufflinks) (5.2.1)
Requirement already satisfied: setuptools>=34.4.1 in c:\users\khadijah jasni\anaconda3\li
b\site-packages (from cufflinks) (41.4.0)
Requirement already satisfied: ipython>=5.3.0 in c:\users\khadijah jasni\anaconda3\lib\si
te-packages (from cufflinks) (7.8.0)
Requirement already satisfied: colorlover>=0.2.1 in c:\users\khadijah jasni\anaconda3\lib
\site-packages (from cufflinks) (0.3.0)
Requirement already satisfied: ipywidgets>=7.0.0 in c:\users\khadijah jasni\anaconda3\lib
\site-packages (from cufflinks) (7.5.1)
Requirement already satisfied: pandas>=0.19.2 in c:\users\khadijah jasni\anaconda3\lib\si
te-packages (from cufflinks) (1.3.2)
Requirement already satisfied: numpy>=1.9.2 in c:\users\khadijah jasni\anaconda3\lib\site
-packages (from cufflinks) (1.21.2)
Requirement already satisfied: six>=1.9.0 in c:\users\khadijah jasni\anaconda3\lib\site-p
ackages (from cufflinks) (1.16.0)
Requirement already satisfied: pickleshare in c:\users\khadijah jasni\anaconda3\lib\site-
packages (from ipython>=5.3.0->cufflinks) (0.7.5)
Requirement already satisfied: colorama in c:\users\khadijah jasni\anaconda3\lib\site-pac
kages (from ipython>=5.3.0->cufflinks) (0.4.1)
Requirement already satisfied: decorator in c:\users\khadijah jasni\anaconda3\lib\site-pa
ckages (from ipython>=5.3.0->cufflinks) (4.4.0)
Requirement already satisfied: prompt-toolkit<2.1.0,>=2.0.0 in c:\users\khadijah jasni\an
aconda3\lib\site-packages (from ipython>=5.3.0->cufflinks) (2.0.10)
Requirement already satisfied: jedi>=0.10 in c:\users\khadijah jasni\anaconda3\lib\site-p
ackages (from ipython>=5.3.0->cufflinks) (0.15.1)
Requirement already satisfied: pygments in c:\users\khadijah jasni\anaconda3\lib\site-pac
kages (from ipython>=5.3.0->cufflinks) (2.4.2)
Requirement already satisfied: backcall in c:\users\khadijah jasni\anaconda3\lib\site-pac
kages (from ipython>=5.3.0->cufflinks) (0.1.0)
Requirement already satisfied: traitlets>=4.2 in c:\users\khadijah jasni\anaconda3\lib\si
te-packages (from ipython>=5.3.0->cufflinks) (4.3.3)
Requirement already satisfied: nbformat>=4.2.0 in c:\users\khadijah jasni\anaconda3\lib\s
ite-packages (from ipywidgets>=7.0.0->cufflinks) (4.4.0)
Requirement already satisfied: widgetsnbextension~=3.5.0 in c:\users\khadijah jasni\anaco
nda3\lib\site-packages (from ipywidgets>=7.0.0->cufflinks) (3.5.1)
Requirement already satisfied: ipykernel>=4.5.1 in c:\users\khadijah jasni\anaconda3\lib\
site-packages (from ipywidgets>=7.0.0->cufflinks) (5.1.2)
```

```
>=7.0.0->cufflinks) (0.5.1)
```

```
WARNING: Ignoring invalid distribution -umpy (c:\users\khadijah jasni\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution -cipy (c:\users\khadijah jasni\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution -umpy (c:\users\khadijah jasni\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution -cipy (c:\users\khadijah jasni\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution -umpy (c:\users\khadijah jasni\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution -cipy (c:\users\khadijah jasni\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution -umpy (c:\users\khadijah jasni\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution -cipy (c:\users\khadijah jasni\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution -umpy (c:\users\khadijah jasni\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution -cipy (c:\users\khadijah jasni\anaconda3\lib\site-packages)
```

Data Visualization

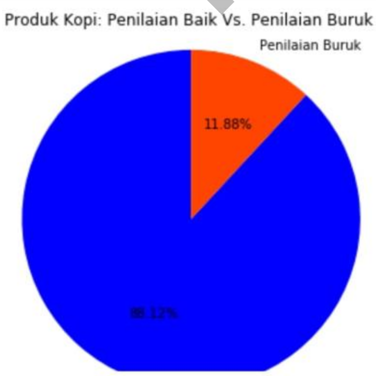
In [31]:

```
import seaborn as sns
import plotly.express as px
import matplotlib.pyplot as plt
%matplotlib inline
```

```
C:\Users\Khadijah Jasni\Anaconda3\lib\site-packages\statsmodels\tools\_testing.py:19: FutureWarning: pandas.util.testing is deprecated. Use the functions in the public API at pandas.testing instead.
  import pandas.util.testing as tm
```

In [32]:

```
fig = plt.figure()
ax = fig.add_axes([0,0,1,1])
values = [(good_rate),(bad_rate)]

ax.pie(values,
       labels = ['Penilaian Baik', 'Penilaian Buruk'],
       colors=['blue', 'orangered'],
       shadow=False,
       startangle=90,
       autopct='%1.2f%%')
ax.axis('equal')
plt.title('Produk Kopi: Penilaian Baik Vs. Penilaian Buruk ');

# save the figure
plt.savefig('plot1.png', dpi=300, bbox_inches='tight')
```



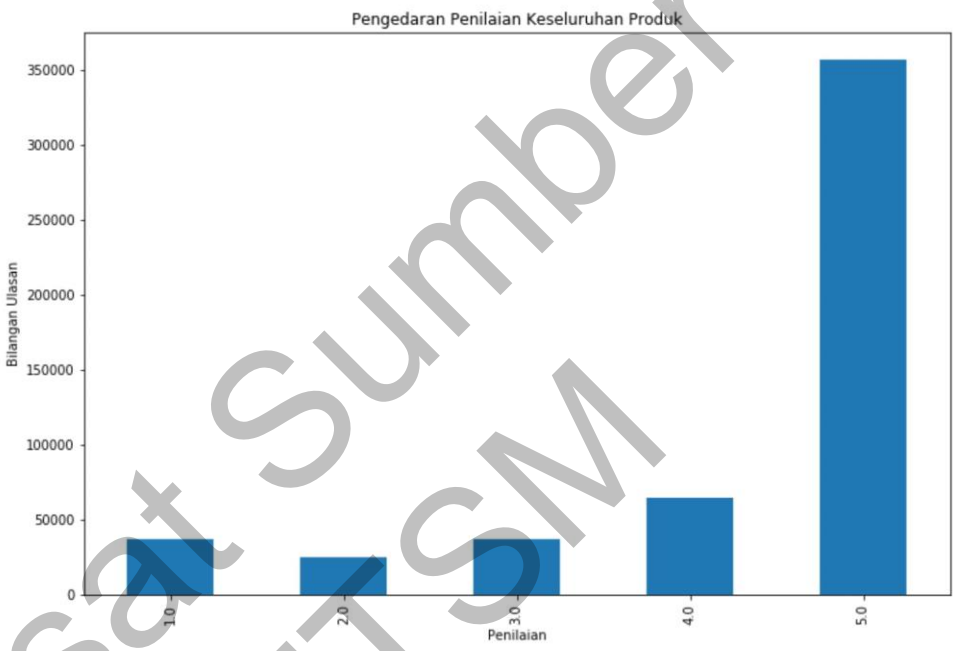
Penilaian Baik

In [33]:

```
## PLOT DISTRIBUTION OF RATING
########################################

plt.figure(figsize=(12,8))
# sns.countplot(df['Rating'])
coffee6['Rating'].value_counts().sort_index().plot(kind='bar')
plt.title('Pengedaran Penilaian Keseluruhan Produk')
plt.xlabel('Penilaian')
plt.ylabel('Bilangan Ulasan')
```

Out[33]:

```
Text(0, 0.5, 'Bilangan Ulasan')
```



In [34]:

```
#DISTRIBUTION OF RATING SCORE
class_counts = coffee6.groupby('Rating').size()
class_counts
```

Out[34]:

```
Rating
1.0     37125
2.0     24654
3.0     36701
4.0     64521
5.0    357037
dtype: int64
```

In [35]:

```
# Customer totals for each rating class
coffee6['rating_class'].value_counts()
```

Out[35]:

```
good    458259
bad      61779
Name: rating_class, dtype: int64
```

In [36]:

```
# Statistics of non-numeric variables

# Number of unique customers
print('\nNumber of unique customers : {}'.format(len(coffee6['reviewerID'].unique())))

# Number of unique products
print('\nNumber of unique products : {}'.format(len(coffee6['asin'].unique())))

# Review number per unique customer
print('\nReview per customer: {}'.format((len(coffee6)/len(coffee6['reviewerID'].unique(
)))))

# Review number per unique product
print('\nReview per product: {}'.format((len(coffee6)/len(coffee6['asin'].unique()))))
```

```
Number of unique customers : 379046

Number of unique products : 16275

Review per customer: 1.3719654078924457

Review per product: 31.953179723502306
```

In [37]:

```
# Read statistic summary of numeric variables
coffee6[['Rating']].describe()
```

Out[37]:

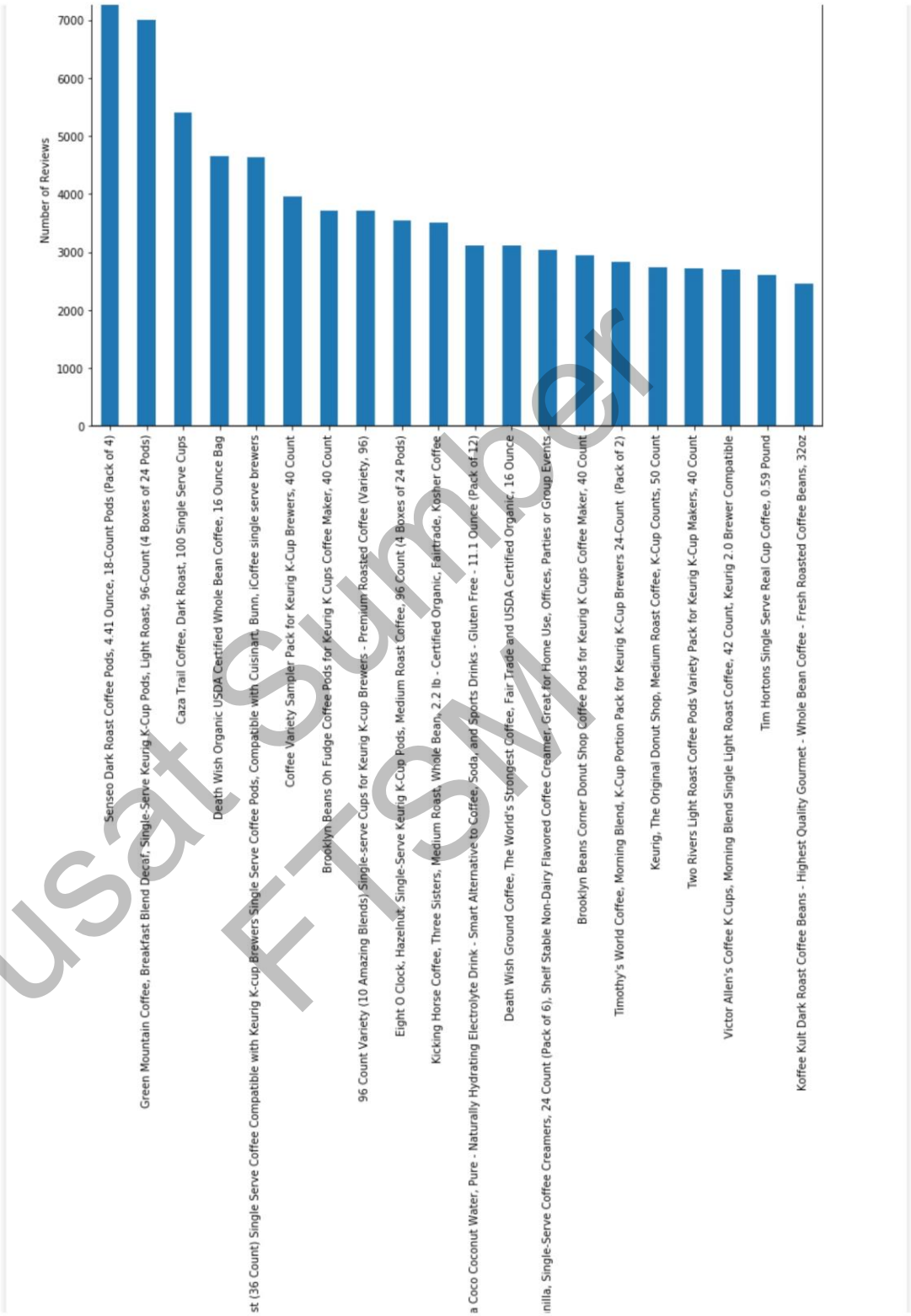
	Rating
count	520038.000000
mean	4.307003
std	1.222710
min	1.000000
25%	4.000000
50%	5.000000
75%	5.000000
max	5.000000

In [40]:

```
########################################
## PLOT NUMBER OF REVIEWS FOR TOP 20 PRODUCTS
########################################

products = coffee6["title"].value_counts()
plt.figure(figsize=(12,8))
products[:20].plot(kind='bar')
plt.title("Number of Reviews for Top 50 Products")
plt.xlabel('Product Name')
plt.ylabel('Number of Reviews')
```

Out[40]:

```
Text(0, 0.5, 'Number of Reviews')
```



In [41]:

```
coffee_fplot = coffee6.groupby(['month'])['Rating'].mean()
coffee_fplot
```

Out[41]:

```
month
1     4.283104
2     4.291622
3     4.307222
4     4.305104
5     4.313178
6     4.329493
7     4.336259
8     4.349684
9     4.321796
10    4.292552
11    4.270809
12    4.289164
Name: Rating, dtype: float64
```

In [43]:

```
# Total numbers of ratings in the home and kitchen product reviews
plt.figure(figsize = (10,6))
sns.countplot(coffee6['Rating'])
plt.title('Total Review Numbers for Each Rating', color='r')
plt.xlabel('Rating')
plt.ylabel('Number of Reviews')
plt.show()

# Customer totals for each rating class
coffee6['Rating'].value_counts()
```