Analisis Trend Pasar Berdasarkan Dataset Judul Produk dari 20 Kategori Berbeda di Amazon

1st *Wahyudiyanto*
121450040
wahyudiyanto.121450040@stundent.itera.ac.id

2nd Annisa Novantika
121450005
annisa.121450005@student.itera.ac.id

3rd Putri Intan Kirani
121450055
putri.121450055@student.itera.ac.id

4th Berliyana Kesuma Hati
121450086
berliyana.121450086@student.itera.ac.id

5th Nazwa Nabilla
121450122
nazwa.121450122@student.itera.ac.id

Abstract:

This study analyzes market trends by collecting and analyzing product data from the Amazon website. Product title data is stored in the Hadoop HDFS system and analyzed using PySpark to compute word frequency. The research reveals dominant keywords across various product categories, offering insights into consumer preferences and market trends. Data collection was conducted using the Google Chrome extensions "simple scraper" and "infy scroll," resulting in approximately one million rows of data. The analysis indicates variations in consumer interest, such as in the Fashion Accessories category where "watch" emerges as the most frequent keyword. These findings are valuable for researchers and marketers to understand consumer behavior and develop more effective marketing strategies. The utilization of PySpark and Hadoop demonstrates the potential of big data tools in uncovering market trends and consumer behavior comprehensively.

Keywords: *Amazon, Word Count, PySpark, Hadoop, Market Trends, Consumer Behavior, E-commerce.*

Abstrak:

Penelitian ini menganalisis tren pasar dengan mengumpulkan dan menganalisis data produk dari situs web Amazon. Data judul produk disimpan dalam sistem Hadoop HDFS dan dianalisis menggunakan PySpark untuk menghitung frekuensi kata. Penelitian ini mengungkapkan kata kunci dominan di berbagai kategori produk, memberikan wawasan tentang preferensi konsumen dan tren pasar. Data dikumpulkan dengan ekstensi Google Chrome "simple scraper" dan "infy scroll", menghasilkan sekitar satu juta baris data. Analisis menunjukkan variasi minat konsumen, seperti dalam kategori Aksesori Fashion di mana "watch" adalah kata kunci paling sering muncul. Temuan ini berguna bagi peneliti dan pemasar untuk memahami perilaku konsumen dan mengembangkan strategi pemasaran yang lebih efektif. Penggunaan PySpark dan Hadoop menunjukkan potensi alat big data dalam mengungkap tren pasar dan perilaku konsumen secara mendalam.

Kata Kunci : Amazon, *Word Count, PySpark, Hadoop*, Tren Pasar, Perilaku Konsumen, *E-commerce.*

Pendahuluan

Dalam era digital yang kian semakin pesat, platform *e-commerce* telah menjadi salah satu kanal utama bagi perusahaan untuk menjual produk mereka dan bagi konsumen untuk membeli barang secara online. *E-commerce* telah merevolusi dunia bisnis, membawa pengaruh signifikan pada strategi marketing perusahaan dalam meraih tujuannya[1]. Platform ini menawarkan berbagai fitur dan tools yang canggih untuk membantu perusahaan dalam mempromosikan produk mereka, menjangkau target audience yang tepat, dan membangun hubungan yang kuat dengan konsumen. Menurut laporan

We Are Social, sekitar 56,1% pengguna internet global belanja online setiap pekan pada Januari 2024.

Tingginya jumlah konsumen yang beralih ke berbelanja secara online menandakan bahwa *e-commerce* telah menjadi pasar yang sangat relevan dan menjanjikan. Untuk membangun bisnis yang sukses di era digital saat ini, analisis tren pasar menjadi hal yang sangat penting. Amazon, sebagai salah satu perusahaan teknologi terdepan secara global menyediakan berbagai macam produk dari berbagai kategori kepada jutaan pengguna di seluruh dunia[2]. Oleh karena itu, analisis tren pasar berdasarkan dataset penjualan produk dari 20 kategori berbeda di Amazon menjadi kunci untuk memahami perilaku pembeli, pola pembelian, dan kecenderungan konsumsi di pasar daring. Dengan memanfaatkan data penjualan yang tersedia, kita dapat mengidentifikasi tren yang sedang naik, memahami preferensi konsumen, dan mengantisipasi permintaan pasar di masa mendatang.

Dalam melakukan analisis tren pasar pada e-commerce terbesar juga berarti mengelola volume data yang besar. Oleh karena itu, dalam analisis ini akan menggunakan alat-alat seperti PySpark dan Hadoop. Beberapa peneliti terdahulu telah melakukan analisis menggunakan spark dan hadoop. Pada tahun 2020, Irfan Rizqi Prabaswara dan Ragil Saputra melakukan penelitian dengan judul "Implementasi Hadoop Dan Spark Untuk Analisis Penyebaran Demam Berdarah Dengue Berdasarkan Data Twitter". Peneliti ini melihat performa spark dan hadoop dalam menganalisis Penyebaran Demam Berdarah Dengue Berdasarkan Data Twitter. Hasil penelitian menunjukan bahwa performa terbaik dalam menggunakan Hadoop dan spark adalah waktu eksekusi 5,3 menit[3]. Pada tahun 2021, Reza Apriliana Fauzi, Imam Cholissodin, dan Bayu Rahayudi melakukan penelitian dengan judul "Pemanfaatan Spark untuk Analisis Sentimen Mengenai Netralitas Berita dalam Membahas Pemilu Presiden 2019 Menggunakan Metode Naïve Bayes Classifier". Peneliti ini menggunakan naive bayes dalam mengklasifikasi dan menerapkan spark dalam pengolahan datanya. Hasil penelitian menunjukan nilai akurasi tertinggi didapatkan pada skenario 10-fold cross validation ke-8 sebesar 100%[4].

Peneliti ini akan memperluas penelitian sebelumnya dengan melakukan analisis tren pasar pada e-commerce terbesar, yaitu Amazon, menggunakan PySpark dan Hadoop. Peneliti akan melakukan analisis mendalam mengenai minat konsumen dalam setiap 20 kategori produk yang tersedia di platform tersebut. Dengan melakukan penelitian ini, diharapkan peneliti dapat mengidentifikasi dan memahami tren pasar berdasarkan penjualan produk di Amazon, serta mengeksplorasi pola-pola yang muncul dari data yang dianalisis.

Metode

Pengumpulan Data

Data dalam analisis ini diperoleh melalui proses scraping dari situs web Amazon menggunakan ekstensi *Google Chrome* "*Simple Scraper*" dan "*Infy Scroll*". Metode ini memungkinkan pengambilan data dalam jumlah besar secara otomatis dan menyimpannya dalam format yang mudah diolah, seperti CSV atau JSON. Informasi yang dikumpulkan mencakup judul produk dari 20 kategori berbeda, menghasilkan sekitar 1 juta baris data [5].

Berikut adalah 20 kategori produk yang dikumpulkan:

1. *Fashion Accessories*
2. *Data Storage*
3. *Perfume & Cologne*
4. *Automotive Tools*
5. *Beauty & Personal Care*
6. *Bath & Body*
7. *Shaving & Hair Removal Products*
8. *Handmade Jewellery*
9. *Kids & Babies*
10. *Luggage & Travel Gear*
11. *Home Decor*
12. *Pets*
13. *Handmade Kitchen & Dining*
14. *Outdoor Cooking*
15. *Men*
16. *Women*
17. *Grocery*
18. *Work & Safety*
19. *Hobbies & Crafts*
20. *Toys & Games*

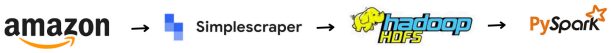
Langkah-langkah pengumpulan data adalah sebagai berikut:

- Pemilihan Kategori Produk: Menentukan 20 kategori produk yang akan diambil datanya dari situs Amazon.
- Konfigurasi Ekstensi *Simple Scraper* : Mengkonfigurasi ekstensi untuk mengambil elemen seperti judul produk.
- Aktivasi *Infy Scrol*l: Mengaktifkan ekstensi "*Infy Scroll"* untuk menggulirkan halaman secara otomatis dan memuat lebih banyak item.
- Proses *Scraping*: Data diambil dengan cara melakukan pencarian berdasarkan kategori, kemudian discraping sampai halaman web habis.

- Penyimpanan Data: Menyimpan data yang diambil dalam format CSV.

Pengolahan Data

Data yang dikumpulkan diunggah ke lingkungan Hadoop untuk diolah lebih lanjut. Setelah itu, dilakukan eksplorasi data untuk setiap kategori guna memahami jumlah baris dan melihat contoh data. Kemudian, dilakukan analisis Word Count untuk menghitung seberapa sering setiap kata muncul dalam setiap kategori produk. Setiap kategori memiliki hasil *Wordcount* tersendiri yang disimpan dalam format data terpisah. Alur pengolahan data dapat dilihat seperti pada **Gambar 1.**



Gambar 1. Alur pengolahan data

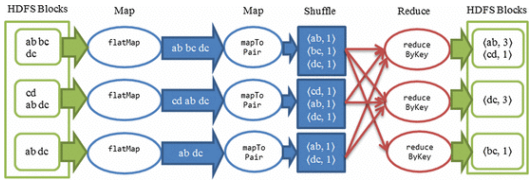
Langkah-langkah pengolahan data meliputi:

- Penyiapan Lingkungan Hadoop: Instalasi dan konfigurasi Hadoop serta memastikan ketersediaan sumber daya.
- Penyiapan Direktori Hadoop: Membuat direktori untuk setiap kategori produk di Hadoop.
- Pengunggahan Dataset: Mengunggah dataset judul produk ke direktori yang sesuai di Hadoop.
- Penggunaan PySpark: Menggunakan PySpark untuk memproses data secara distribusi.
- Eksplorasi Data: Memeriksa jumlah baris dan menampilkan sampel data untuk memahami karakteristik dataset.
- *Word Count*: Menghitung frekuensi kemunculan setiap kata dalam judul produk dari setiap kategori.
- Penyimpanan Hasil *Word Count*: Menyimpan hasil word count dalam sistem penyimpanan terdistribusi Hadoop HDFS, dengan setiap kategori produk disimpan dalam file teks terpisah.

Analisis Statistik

Analisis statistik dilakukan untuk mendapatkan wawasan tentang tren pasar berdasarkan judul produk dari berbagai kategori di Amazon:

- *Word Count*: Menghitung frekuensi kemunculan setiap kata dalam judul produk untuk mengidentifikasi kata kunci yang umum. *Word count* dilakukan dengan menerapkan *ignored words* yang mengabaikan simbol, kata-kata sensitif dan konjugasi. Cara kerja *word count* pada Pyspark dapat dilihat seperti pada **Gambar 2**.



Gambar 2. *word count in pyspark*

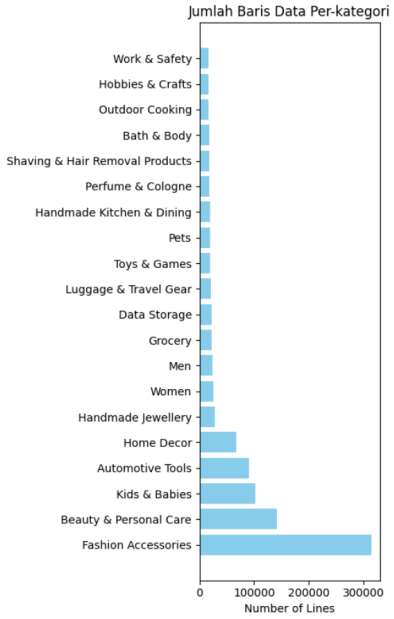
- Visualisasi Data: Menyajikan hasil word count dalam bentuk *bar chart* dan word cloud untuk memvisualisasikan frekuensi kemunculan kata secara grafis.

Melalui analisis ini, diperoleh pemahaman lebih dalam tentang tren pasar dan preferensi pelanggan untuk setiap kategori produk di Amazon, yang berguna bagi peneliti, pemasar, dan pengambil keputusan dalam mengembangkan strategi pemasaran yang efektif.

Hasil dan Diskusi

Eksplorasi Data

Tiap kategori data yang telah dikumpulkan memiliki jumlah data yang berbeda-beda. Dalam penelitian ini, data Fashion Accessories dan Beauty & Personal Care adalah dua dataset dengan jumlah baris data terbanyak. Jumlah baris ini sesuai dengan jumlah ketersediaan produk per-kategori di Amazon. Jumlah baris data dari 20 kategori dapat dilihat pada **Gambar 3**.



Gambar 3. Jumlah baris data

Word Count

Proses ini bertujuan untuk melakukan penghitungan jumlah kemunculan kata dalam setiap kategori produk yang telah dijelaskan sebelumnya. Dalam implementasinya, dilakukan pemrosesan teks untuk setiap kategori produk dengan mengabaikan kata-kata yang tidak relevan atau diabaikan seperti '*and*', *'for'*, dan kata-kata umum lainnya. Setelah itu, dilakukan penghitungan jumlah kemunculan kata-kata yang tersisa dalam setiap kategori produk. Hasil akhirnya adalah jumlah kemunculan kata-kata yang relevan dalam tiap kategori produk, yang nantinya dapat digunakan untuk analisis lebih lanjut terhadap tren pasar dan preferensi pelanggan dalam masing-masing kategori tersebut.

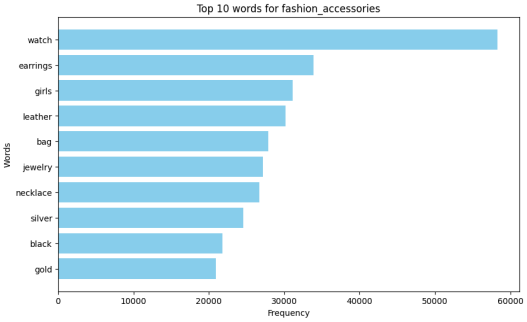
Berikut adalah hasil top-5 wordcount untuk tiap kategori yang ditunjukan pada **Gambar 4** :

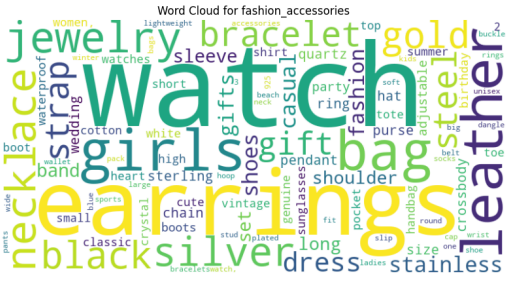
Gambar 4. top-5 kata per kategori

Analisis

Pada analisis ini, kami mengeksplorasi tren dan preferensi konsumen dalam tiga kategori produk utama: *fashion accessories*, *beauty & personal care*, dan *home decor*. Dengan menggunakan data frekuensi kata dari deskripsi produk, kami mengidentifikasi sepuluh kata teratas dalam setiap kategori. Analisis ini bertujuan untuk memahami produk apa yang paling diminati oleh konsumen, bahan dan fitur apa yang mereka cari, serta pola perilaku yang mungkin ada dalam setiap kategori.

- **Kategori Fashion Accessories**





Gambar 5. fashion accessories

Watch: Kata ini mendominasi, menunjukkan bahwa jam tangan adalah salah satu produk yang paling diminati dalam kategori ini. Jam tangan mungkin berfungsi ganda sebagai alat penunjuk waktu dan aksesori mode.

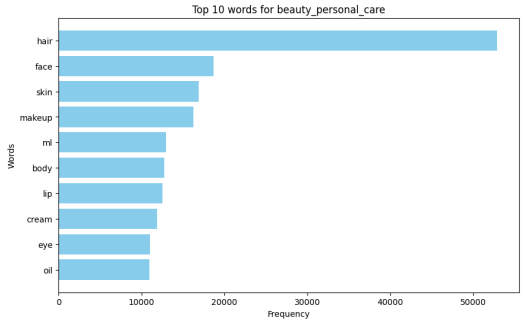
Earrings, Necklace, Jewelry: Banyaknya sebutan ini menunjukkan tingginya minat terhadap perhiasan, khususnya anting-anting dan kalung.

Girls: Mengindikasikan bahwa banyak produk aksesori yang ditargetkan untuk perempuan muda.

Leather, Silver, Gold: Referensi terhadap bahan-bahan ini menunjukkan preferensi konsumen terhadap material berkualitas tinggi.

Bag, Black: Menunjukkan popularitas tas dan preferensi untuk warna hitam yang serbaguna.

- **Kategori Beauty & Personal Care**



Gambar 6. Beauty personal care

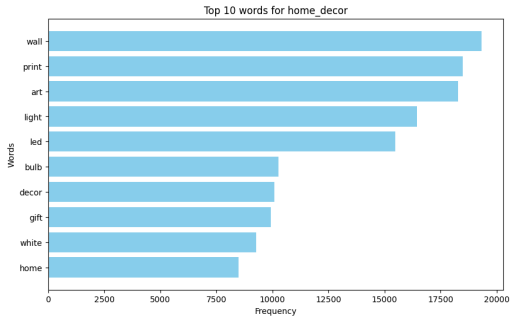
Hair, Face, Skin : Fokus utama pada perawatan rambut, wajah, dan kulit. Hal ini menandakan bahwa konsumen sangat memperhatikan kesehatan dan penampilan bagian-bagian tubuh ini.

Makeup: Menunjukkan pentingnya produk kosmetik dalam rutinitas kecantikan konsumen**.**

Ml : Mungkin terkait dengan ukuran produk, menunjukkan perhatian konsumen pada kuantitas dan volume produk yang mereka beli.

Body, Lip, Cream, Eye, Oil : Menunjukkan keragaman kebutuhan perawatan pribadi yang difokuskan pada berbagai bagian tubuh, mulai dari krim wajah hingga minyak perawatan rambut.

- **Kategori Home Decor**



Gambar 7. Bar chart top 10 word count home_decor

Wall, Print, Art: Menunjukkan pentingnya dekorasi dinding seperti cetakan dan karya seni dalam kategori ini.

Konsumen tampaknya sangat tertarik pada estetika dinding rumah mereka.

Light, LED, Bulb: Fokus pada pencahayaan rumah, mengindikasikan bahwa aspek ini adalah elemen penting dalam dekorasi rumah.

Decor, Gift: Menegaskan pentingnya elemen dekoratif dan produk yang sering dibeli sebagai hadiah, menunjukkan bahwa dekorasi rumah sering kali juga berfungsi sebagai pilihan hadiah yang populer.

White: Menunjukkan preferensi terhadap warna netral dalam dekorasi rumah, yang mungkin mencerminkan tren desain interior yang bersih dan minimalis.

Home: Menggarisbawahi bahwa produk-produk ini berhubungan dengan peningkatan dan estetika rumah secara keseluruhan, menekankan pentingnya kenyamanan dan penampilan ruang hunian.

Kesimpulan

Analisis ini mengidentifikasi *trend* dan preferensi konsumen di Amazon dalam 20 kategori dan 3 kategori pilihan yaitu kategori fashion accessories, beauty & personal care, dan home decor. Data scraping dan analisis word count menunjukkan jam tangan paling diminati dalam fashion accessories. Fokus utama dalam beauty & personal care adalah perawatan rambut, wajah, dan kulit. Dalam home decor, dekorasi dinding dan pencahayaan menjadi prioritas utama. Hasil analisis ini memberikan wawasan penting bagi peneliti dan pemasar dalam mengembangkan strategi pemasaran yang lebih efektif, dengan pemahaman mendalam mengenai preferensi bahan, fitur, dan produk yang dicari konsumen di setiap kategori.

References

[1]Aco, Ambo. n.d. “Analisis Bisnis E-Commerce pada Mahasiswa Universitas Islam Negeri Alauddin Makassar.” *Jurnal UIN (Universitas Islam Negeri) Alauddin Makassar*. Accessed Mei 13, 2024.

[2]Alwendi. 2020. “Penerapan E-Commerce Dalam Meningkatkan Daya Saing Usaha.” *Jurnal Manajemen Bisnis* 17 (3). https://journal.undiknas.ac.id/index.php/magister-manajemen/article/view/2486/732).

[3]i Prabaswara, Irfan R., and Ragil Saputra. 2020. “Implementasi Hadoop Dan Spark Untuk Analisis Penyebaran Demam Berdarah Dengue Berdasarkan Data Twitter.” *IT Journal Research and Development (ITJRD)* 4 (2): 164-171. https://journal.uir.ac.id/index.php/ITJRD/article/download/4099/2384.

[4]Fauzi, Reza A., Imam m Cholissodin, and Bayu Rahayudi. 2021. “Pemanfaatan Spark untuk Analisis Sentimen Mengenai Netralitas Berita dalam Membahas Pemilu Presiden 2019 Menggunakan Metode Naïve Bayes Classifier.” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* 5 (3): 1070-1077. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/8741.

[5]Ngalup. 2021. “Web Scraping Pengertian, Teknik, Manfaat dan Kendala adalah.” Ngalup. https://ngalup.co/artikel/pengertian-teknik-manfaat-kendala-web-scraping/.