Text Analytics & Business Application

Text Analytics in E-commerce

Qinglai He

Department of Operations and Information Management

Wisconsin School of Business

WISCONSIN SCHOOL OF BUSINESS

IC5 Discussions

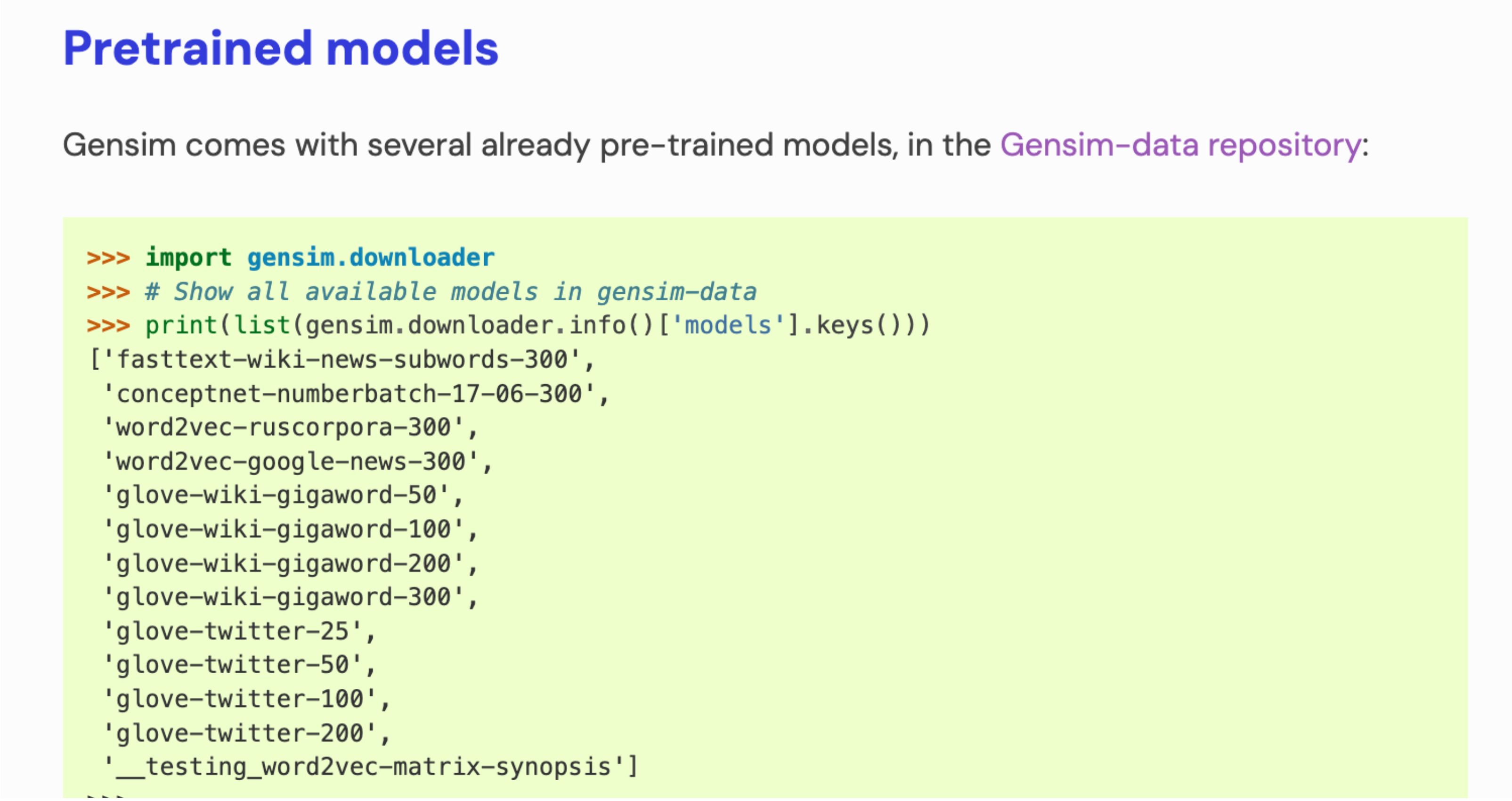
- Why using pre-trained word embedding as text representation could not give us better classification results than BoW?
    - The effectiveness of word embeddings and BoW depends on various factors, such as the data and complexity of the task. So, it is always beneficial to experiment with different approaches for a particular task.
    - The performance of pre-trained word embeddings also depends on other factors, such as overlap of words between data the pre-trained model, corpus used for training, training architectural, and its dimensionality.
    - Pre-trained model aims to capture more general meaning. In contrast, BoW is generated based on data in a specific domain, therefore it is more flexible for different tasks.

IC5 Discussions

- What is the pros and cons of using word embeddings?
  - Pros:
    - It captures more semantic relationships between words. Therefore, it is good for tasks where the meaning of words is crucial.
    - It can capture more contextual information.
    - It provides dense and low-dimensional representations.
  - Cons:
    - Fixed vocabulary: The vocabulary of the pre-trained model are from training corpus. It may not cover all words in a specific domain or dataset.
    - May not capture the domain-specific nuances, limiting the effectiveness for certain tasks.
    - Fine-tuning the pre-trained model is usually required for a specific domain.
    - We are unable to interpret the embeddings.
    - Still couldn’t capture the word order in your given sentence.

IC5 Discussions

- Available pre-trained model offered by Gensim
  - Read more here https://radimrehurek.com/gensim/models/word2vec.html



Pretrained models

Gensim comes with several already pre-trained models, in the Gensim-data repository:

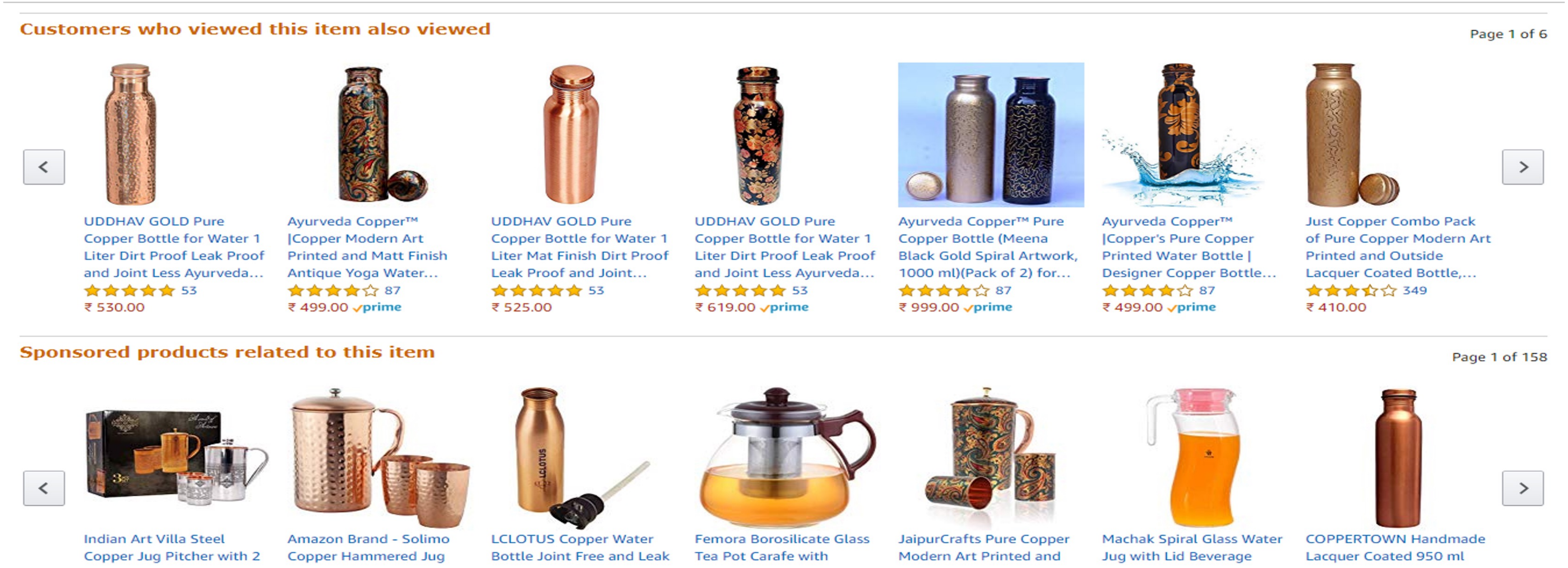
```
>>> import gensim.downloader
>>> # Show all available models in gensim-data
>>> print(list(gensim.downloader.info()['models'].keys()))
['fasttext-wiki-news-subwords-300',
 'conceptnet-numberbatch-17-06-300',
 'word2vec-ruscorpora-300',
 'word2vec-google-news-300',
 'glove-wiki-gigaword-50',
 'glove-wiki-gigaword-100',
 'glove-wiki-gigaword-200',
 'glove-wiki-gigaword-300',
 'glove-twitter-25',
 'glove-twitter-50',
 'glove-twitter-100',
 'glove-twitter-200',
 '__testing_word2vec-matrix-synopsis']
>>>
```

Outline of Today’s Class

- E-Commerce applications
- E-Commerce catalog
- Sentiment analysis
  - Review analysis
  - Aspect-Level sentiment analysis
- Case study

Text Analytics Applications in Ecommerce

Recommendation System



Customers who viewed this item also viewed

Page 1 of 6

- UDDHAV GOLD Pure Copper Bottle for Water 1 Liter Dirt Proof Leak Proof and Joint Less Ayurveda… ★★★★★ 53 ₹ 530.00
- Ayurveda Copper™ |Copper Modern Art Printed and Matt Finish Antique Yoga Water… ★★★★☆ 87 ₹ 499.00 ✓prime
- UDDHAV GOLD Pure Copper Bottle for Water 1 Liter Mat Finish Dirt Proof Leak Proof and Joint… ★★★★★ 53 ₹ 525.00
- UDDHAV GOLD Pure Copper Bottle for Water 1 Liter Dirt Proof Leak Proof and Joint Less Ayurveda… ★★★★★ 53 ₹ 619.00 ✓prime
- Ayurveda Copper™ Pure Copper Bottle (Meena Black Gold Spiral Artwork, 1000 ml)(Pack of 2) for… ★★★★☆ 87 ₹ 999.00 ✓prime
- Ayurveda Copper™ |Copper's Pure Copper Printed Water Bottle | Designer Copper Bottle… ★★★★☆ 87 ₹ 499.00 ✓prime
- Just Copper Combo Pack of Pure Copper Modern Art Printed and Outside Lacquer Coated Bottle,… ★★★½☆ 349 ₹ 410.00

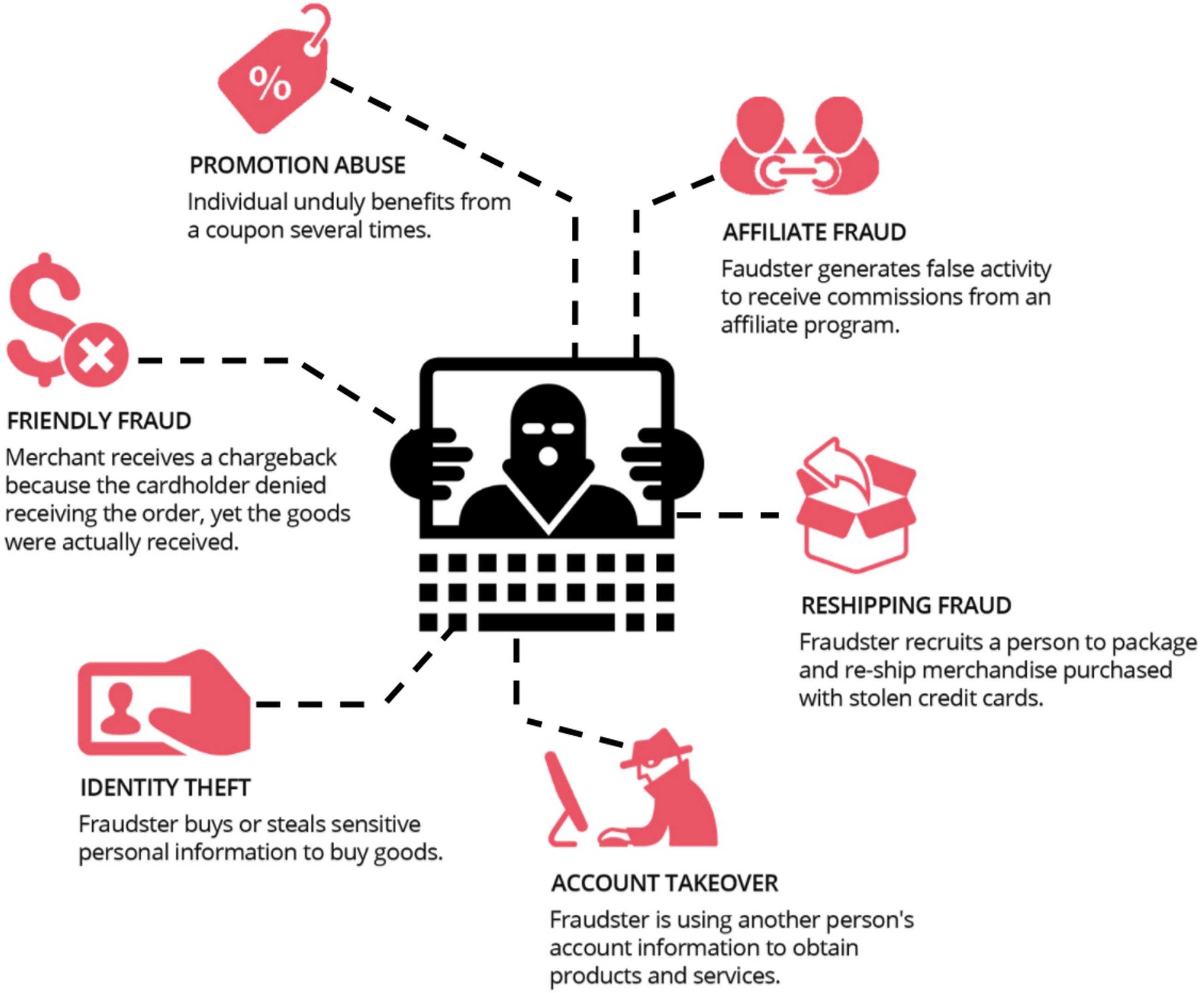
Sponsored products related to this item

Page 1 of 158

- Indian Art Villa Steel Copper Jug Pitcher with 2
- Amazon Brand - Solimo Copper Hammered Jug
- LCLOTUS Copper Water Bottle Joint Free and Leak
- Femora Borosilicate Glass Tea Pot Carafe with
- JaipurCrafts Pure Copper Modern Art Printed and
- Machak Spiral Glass Water Jug with Lid Beverage
- COPPERTOWN Handmade Lacquer Coated 950 ml

Delighting Customers with Personalized Discounts

Fighting Cyberfraud



PROMOTION ABUSE

Individual unduly benefits from a coupon several times.

AFFILIATE FRAUD

Faudster generates false activity to receive commissions from an affiliate program.

FRIENDLY FRAUD

Merchant receives a chargeback because the cardholder denied receiving the order, yet the goods were actually received.

RESHIPPING FRAUD

Fraudster recruits a person to package and re-ship merchandise purchased with stolen credit cards.

IDENTITY THEFT

Fraudster buys or steals sensitive personal information to buy goods.

ACCOUNT TAKEOVER

Fraudster is using another person's account information to obtain products and services.

E-commerce Catalog

E-commerce Catalog

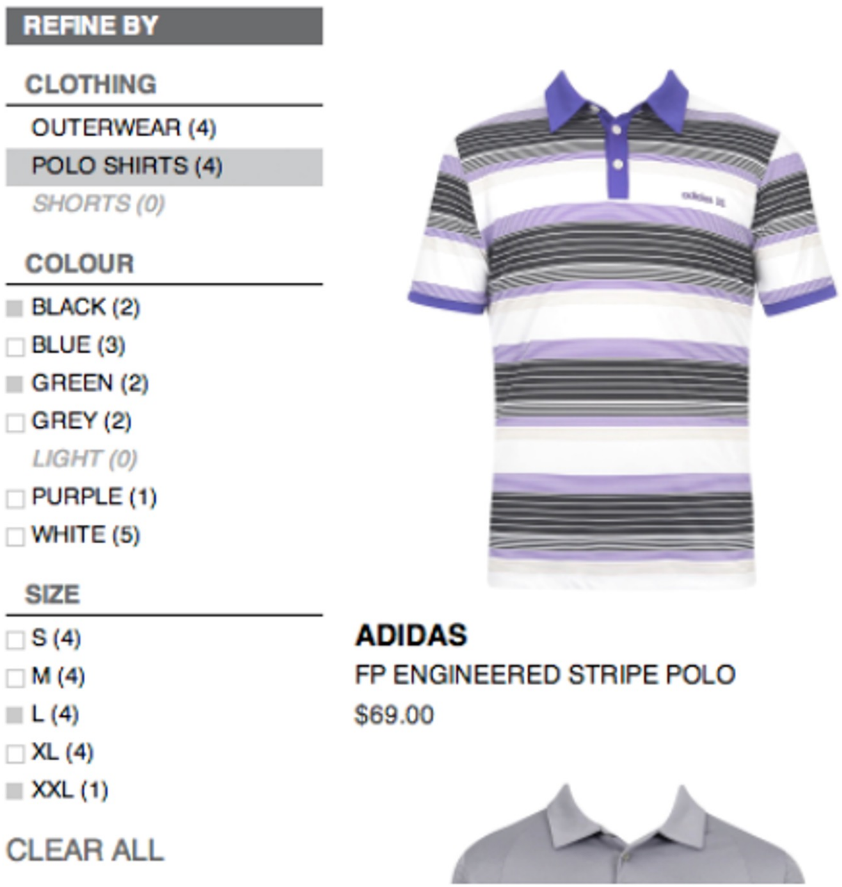
- A product catalog is a database of the products the enterprise deals with, containing product descriptions and images.
- Better product descriptions with relevant information help customers choose the right product through the catalog.
- Product information can also help with search and recommendations.

Building an E-Commerce Catalog

- Building an informative catalog is one of the primary problems in e-commerce. It can be split into several subproblems:
  - Attribute extraction
  - Product categorization and taxonomy creation
  - Product enrichment
  - Product deduplication and matching

1. Attribute Extraction

- Attributes define product properties in e-commerce and provide a complete overview of the product for the customer.
- A rich set of attributes improves clicks and click-through rates, which influences product sales.
- Obtaining attributes directly from sellers for all products is difficult, and the quality of attributes must be **consistent** for the customer to have the correct and relevant information.



Obtaining Product Attributes

- Traditionally, e-commerce websites employed manual labeling or crowdsourcing techniques to obtain product attributes, which is expensive and not scalable with an increase in the volume of products.
- Machine learning techniques are used to obtain product attributes.
- There are two types of attribute extraction algorithms:
  - Direct (will talk about information extraction in Week 9)
  - Derived (Indirect, e.g., classification task)

2. Product Categorization and Taxonomy

- Product categorization is a process of dividing products into groups.
- These groups can be defined based on similarity
  - e.g., products of the same brand or products of the same type can be grouped together. Generally, e-commerce has pre-defined broad categories of products, such as electronics, personal care products, and foods.

Product Categorization and Taxonomy

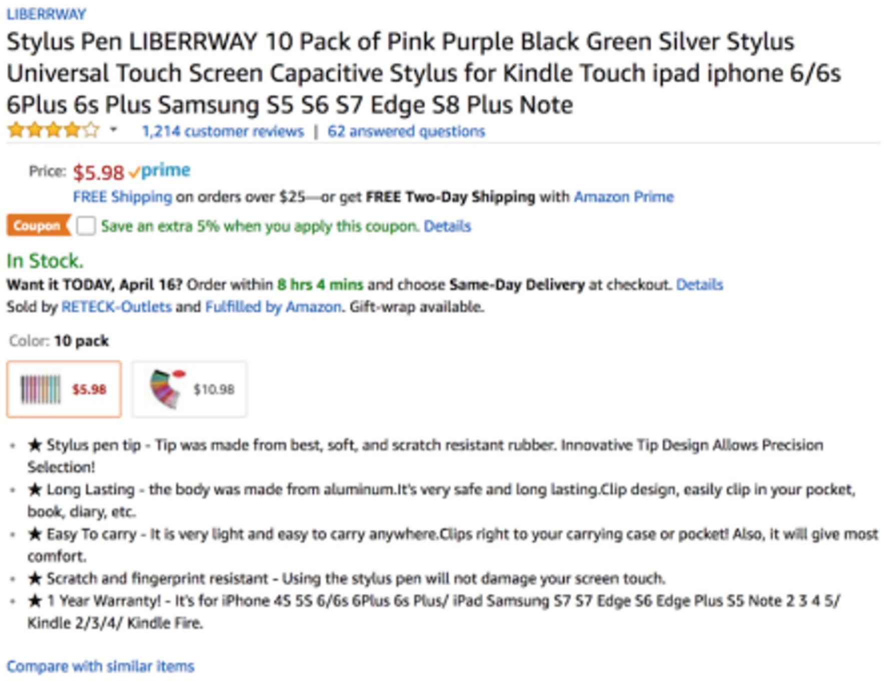
- A good taxonomy and properly linked products can be critical for business operations.
- This categorization process is typically manual to start at small scale.
- At scale, this categorization is typically posed as a classification task where the algorithm takes information from a variety of sources and applies the classification technique to solve it.

3. Product Enrichment

- To improve search and recommendations in e-commerce platforms, it's important to gather richer product information.
- Short and long titles, product images, and product descriptions are potential sources of this information.
- However, this information is often incorrect or incomplete, which can hamper the faceted search in e-commerce platforms.

Product Enrichment: Misleading Titles

- Misleading titles can hamper faceted search in e-commerce platforms and negatively affect click-through and conversion rates.



LIBERRWAY

Stylus Pen LIBERRWAY 10 Pack of Pink Purple Black Green Silver Stylus Universal Touch Screen Capacitive Stylus for Kindle Touch ipad iphone 6/6s 6Plus 6s Plus Samsung S5 S6 S7 Edge S8 Plus Note

★★★★☆ 1,214 customer reviews | 62 answered questions

Price: $5.98 ✓prime

FREE Shipping on orders over $25—or get **FREE Two-Day Shipping** with Amazon Prime

Coupon □ Save an extra 5% when you apply this coupon. Details

In Stock.

Want it TODAY, April 16? Order within 8 hrs 4 mins and choose **Same-Day Delivery** at checkout. Details

Sold by RETECK-Outlets and Fulfilled by Amazon. Gift-wrap available.

Color: **10 pack**

$5.98 $10.98

- ★ Stylus pen tip - Tip was made from best, soft, and scratch resistant rubber. Innovative Tip Design Allows Precision Selection!
- ★ Long Lasting - the body was made from aluminum.It's very safe and long lasting.Clip design, easily clip in your pocket, book, diary, etc.
- ★ Easy To carry - It is very light and easy to carry anywhere.Clips right to your carrying case or pocket! Also, it will give most comfort.
- ★ Scratch and fingerprint resistant - Using the stylus pen will not damage your screen touch.
- ★ 1 Year Warranty! - It's for iPhone 4S 5S 6/6s 6Plus 6s Plus/ iPad Samsung S7 S7 Edge S6 Edge Plus S5 Note 2 3 4 5/ Kindle 2/3/4/ Kindle Fire.

Compare with similar items

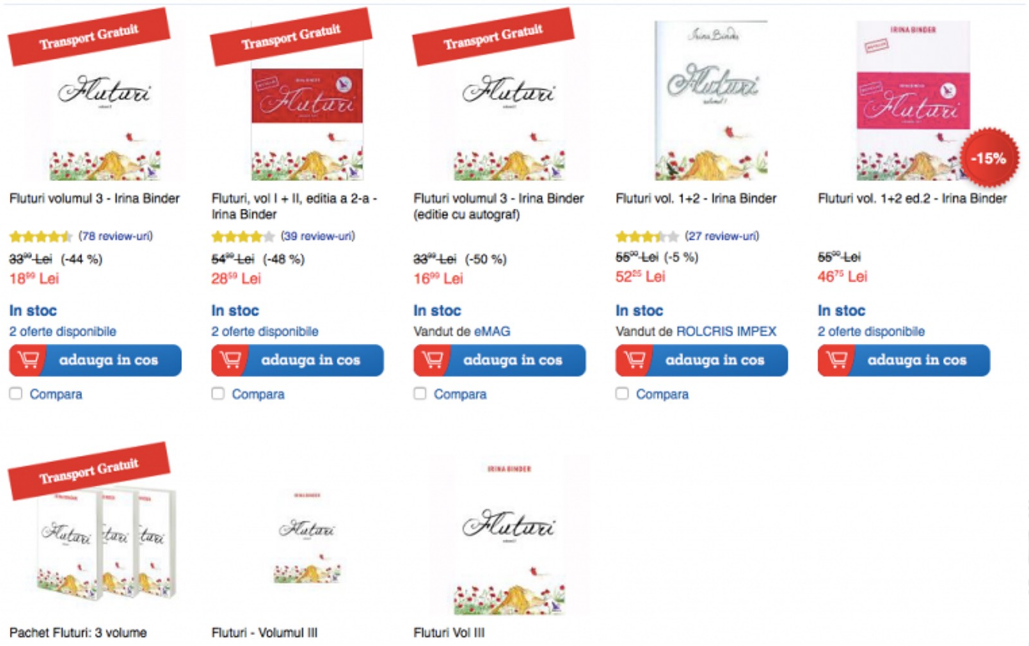
This text is too complicated even for a human to parse and make sense of, let alone a machine.

Improving Product Titles

- Filtering out misleading tokens
  - Direct string matching can be used to improve product titles.
  - Tokens that are not part of the product's attribute values should be filtered out to avoid misleading search results.
- Using pre-defined templates
  - Pre-defined templates can help maintain consistency across product titles.
  - A template can be built using attributes from the taxonomy tree.

4. Product Deduplication and Matching

- Product duplication in e-commerce
  - Third-party sellers often add products to e-commerce platforms with different names and terminology.
  - This can result in the same product being listed with multiple titles and product images, leading to confusion for customers.
- Ways to handle product duplication
  - Attribute match
  - Title match
  - Image match



Transport Gratuit
Fluturi volumul 3 - Irina Binder
(78 review-uri)
~~33^{99} Lei~~ (-44 %)
18^{99} Lei
In stoc
2 oferte disponibile
adauga in cos
□ Compara

Transport Gratuit
Fluturi, vol I + II, editia a 2-a - Irina Binder
(39 review-uri)
~~54^{99} Lei~~ (-48 %)
28^{59} Lei
In stoc
2 oferte disponibile
adauga in cos
□ Compara

Transport Gratuit
Fluturi volumul 3 - Irina Binder (editie cu autograf)
~~33^{99} Lei~~ (-50 %)
16^{99} Lei
In stoc
Vandut de eMAG
adauga in cos
□ Compara

Fluturi vol. 1+2 - Irina Binder
(27 review-uri)
~~55^{00} Lei~~ (-5 %)
52^{25} Lei
In stoc
Vandut de ROLCRIS IMPEX
adauga in cos
□ Compara

-15%
Fluturi vol. 1+2 ed.2 - Irina Binder
~~55^{00} Lei~~
46^{75} Lei
In stoc
2 oferte disponibile
adauga in cos

Transport Gratuit
Pachet Fluturi: 3 volume

Fluturi - Volumul III

Fluturi Vol III

Attribute Match

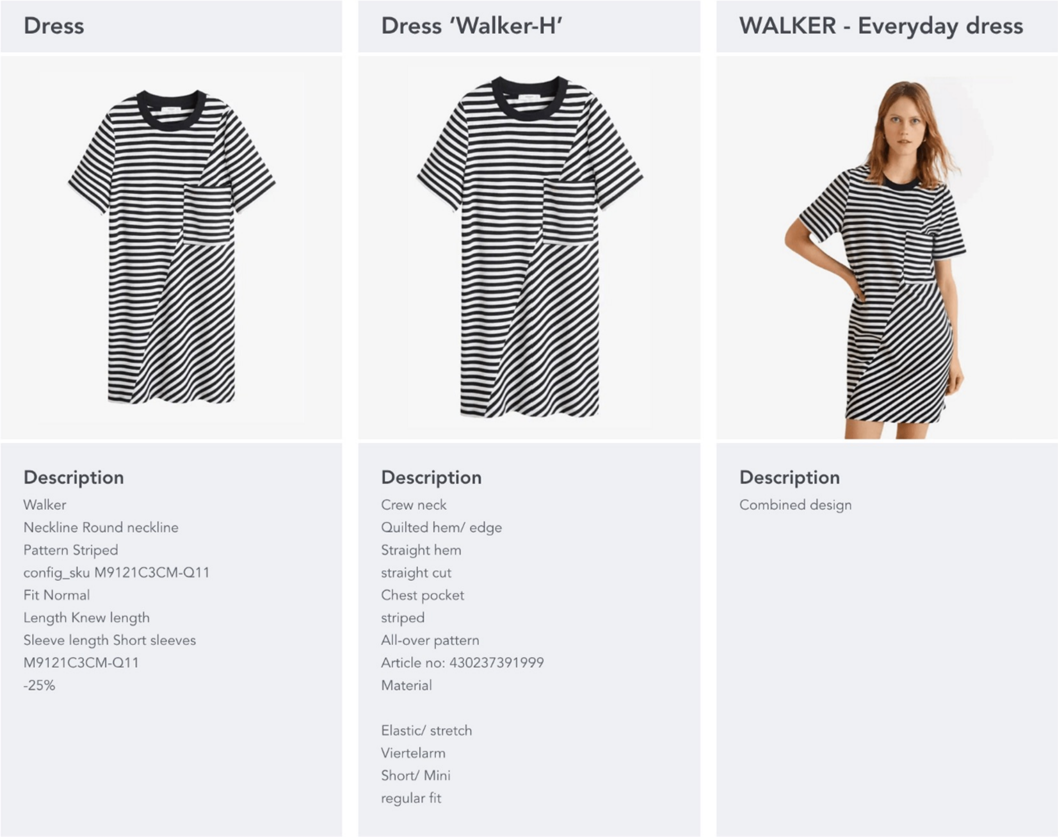
- Matching attribute values
  - Maximum overlap of attributes indicates strong product matching.
- Using string matching
  - String matching can be used to match attribute values.
  - Two strings can be matched using exact character match or string similarity metrics that take into account slight spelling mistakes, abbreviations, etc.

Title Match

- One product can often have multiple title variants.
- A simple method could be to compare **bigrams** and **trigrams** among these.
- It’s also possible to generate title-level features (such as counts of common bigrams and trigrams) and then calculate the Euclidean distance between them

Image Match

- Different techniques like pixel-to-pixel match, feature map matching, and advanced image-matching techniques can be used for image matching.



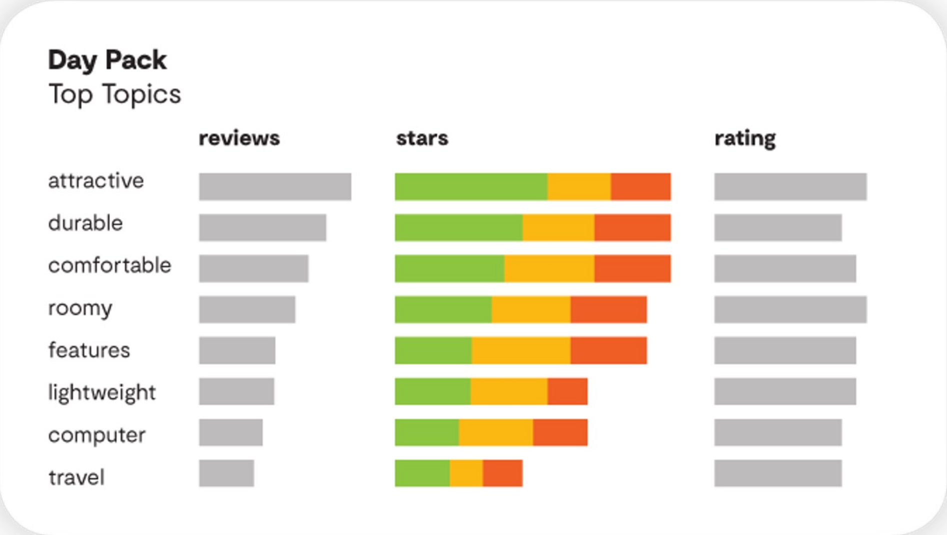
Dress	Dress 'Walker-H'	WALKER - Everyday dress
Description Walker Neckline Round neckline Pattern Striped config_sku M9121C3CM-Q11 Fit Normal Length Knew length Sleeve length Short sleeves M9121C3CM-Q11 -25%	**Description** Crew neck Quilted hem/ edge Straight hem straight cut Chest pocket striped All-over pattern Article no: 430237391999 Material Elastic/ stretch Viertelarm Short/ Mini regular fit	**Description** Combined design

Sentiment Analysis

Why Sentiment Analysis?

- **Movie**: is this review positive or negative?
- **Products**: what do people think about the new iPhone?
- **Public** sentiment: how is consumer confidence? Is despair increasing?
- **Politics**: what do people think about this candidate or issue?
- **Prediction**: predict election outcomes or market trends from sentiment

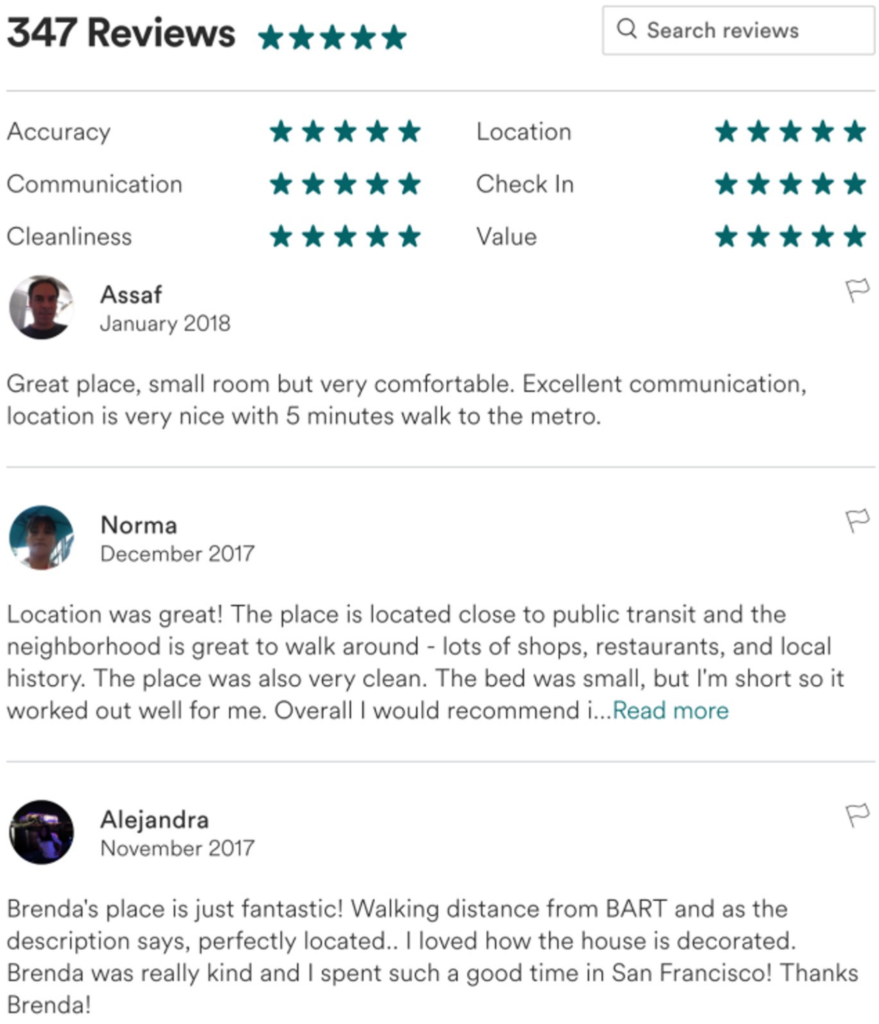
1. Review Analysis



- They capture direct feedback from customers about products, and they can directly affect the sales of the products.
- **Brand monitoring**: reveal key insights about how your brand, product, or company is viewed by your customers and stakeholders
- **Product improvement**: companies can use them to further improve the customer experience.

2. Aspect-Level Sentiment Analysis

- An aspect is a collection of words that indicates certain properties or characteristics of the product.
- Aspects can also include anything related to the supply, presentation, delivery, return, quality, etc.
- Supervised and unsupervised techniques:
  - Supervised techniques involve using labeled data and predefined features.
  - Unsupervised techniques involve clustering and topic modeling.



347 Reviews ★★★★★

Accuracy	★★★★★	Location	★★★★★
Communication	★★★★★	Check In	★★★★★
Cleanliness	★★★★★	Value	★★★★★

Assaf
January 2018

Great place, small room but very comfortable. Excellent communication, location is very nice with 5 minutes walk to the metro.

Norma
December 2017

Location was great! The place is located close to public transit and the neighborhood is great to walk around - lots of shops, restaurants, and local history. The place was also very clean. The bed was small, but I'm short so it worked out well for me. Overall I would recommend i...Read more

Alejandra
November 2017

Brenda's place is just fantastic! Walking distance from BART and as the description says, perfectly located.. I loved how the house is decorated. Brenda was really kind and I spent such a good time in San Francisco! Thanks Brenda!

Challenges of Sentiment Analyses

- Tone
- Polarity
- Sarcasm
- …

Case Study & Tools

Existing Tools for Sentiment Analyses

- TextBlob - an easy way to calculate sentiment polarity. TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.
- Vader Sentiment Analysis
- Entity Sentiment Analysis

Examples - TextBlob

```
text_1 = "The material feels amazing and it seems durable"
TextBlob(text_1).sentiment.polarity
```

=> 0.6

```
text_2 = "I have purchased two of these dresses in the past and was
 excited to order one from Amazon. Unfortunately quality is very lo
w. Arrived with stitching already unraveling."
TextBlob(text_2).sentiment.polarity
```

=> -0.09375

Examples with Google Colab

Take 10 minutes break…