Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

SBD3 | Text Mining Project Briefing (Group Work 1)

Prof. Dr. Branka Hadji Misheva &
Prof. Dr. Patrick Cichy

Berner Fachhochschule | Haute école spécialisée bernoise | Bern University of Applied Sciences

Text mining project – Overview

Work on an empirical case study and derive insights from text data that help answering your “clients” questions.

Online Store Management (Fashion)

Technology Hype: ChatGPT

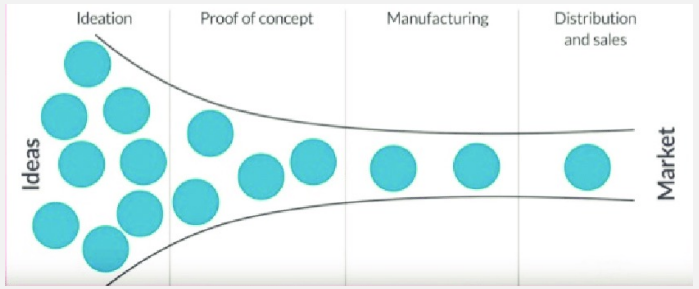
BFH Social Media communication

Disneyland Park Management

Text Mining Project – Briefing

Group Work

- Try to get a proper understanding of the context/field your project is located in
- Invest time to really understand the dataset
- Involve all (!) team members in the explorative process to unleash the creative potential of the group
- Fail early (agile approach, prioritize ideas)



- Apply text mining methods that we covered in class

Text Mining Project – Briefing

Presentation

- Have your audience in mind
  - Client + listeners without specific knowledge of your study context/ products and technologies that your case relates to ▸ *Please briefly introduce what you think is important.*
- Zoom-In / Zoom-Out
- Don‘t forget the time limit for your presentation (20 min)
- Use storytelling and data visualizations to communicate your message effectively

We are a successful online fashion retailer however struggle with the amount of items that our customers return. Clearly, we are not the only one with such a problem. Studies repeatedly show that this phenomenon is observed across the entire online vendor industry. However, fashion items are being returned most often: Almost 40% of orders in this category are being sent back to the vendor (EHI Retail Institute, 2019). This generates enormous costs and conflicts with our sustainability goals, as you can imagine. Can you help us to understand our customer better and to derive measures to decrease the rate of returns? In particular, we are interested in the following:

1. What can you tell us about the customers that write reviews?
2. What problems/dislikes – as potential reasons for a product return – are customers describing in the reviews and how do these relate to sentiment/ratings?
3. What differences can you detect in the various product categories (classes)?
4. What specific advice can you give to our assortment management team based on your analysis? How can we integrate the analysis of reviews in our internal processes, can you think of any data products that would be of value for us?

Dataset	Customer reviews, ratings and some meta data for items bought in our online store.
Quantity	23.486 Reviews
File	E_commerce.rda
Variables	**Clothing ID** (refers to the specific piece being reviewed), **Age** (of the reviewer), **Title** (of review), **Review Text**, Rating (from 1 Worst, to 5 Best), Recommended IND (1 if the reviewer recommends the product to others, 0 if not), Positive Feedback Count (No of other customers who found this review helpful), Division Name, Department Name, Class Name. *Note. This is real commercial data, it has been anonymized, and references to the company in the review text and body have been replaced with "retailer"*
Further hints.	How to deal with fake or sponsored reviews, different languages and spelling mistakes in reviews?

We are Disney Parks and operate several large theme parks worldwide. In our famous Disneyland we host several millions guests each year and are eager to continuously improve our visitors‘ experience. As Walt Disney said "Disneyland will never be completed. It will continue to grow as long as there is imagination left in the world.“ Can you help us with understanding our visitors better by examining their post-experience reviews?

1. What can you tell us about the customers that write reviews?
2. What do the visitors talk about in their reviews and how does it relate to sentiment/ratings?
3. What differences can you detect for the three different locations and are there any interesting trends over time?
4. What specific advice can you give to our park management based on your analysis? How can we integrate the analysis of reviews in our internal processes, can you think of any data products that would be of value for us?

Dataset	Customer reviews, ratings and some additional information about the visitor and the park he/she visited
Quantity	42.656 Reviews
File	Disneyland.rda
Variables	Review_ID, Rating (from 1 =unsatisfied to 5 =satisfied), Year_Month (of visit), Year (of visit), Reviewer_Location (country of origin of visitor), Review_Text, Disneyland_Branch (location of Disneyland Park)
Further hints.	How to deal with fake or sponsored reviews, different languages and spelling mistakes in reviews?

We are the communication department of Berner Fachhochschule and manage various channels of which Twitter is one. Next to our own activities on this social media platform, we are interested in finding out how other Universities of Applied Sciences use Twitter. Any insights are of value, because such can help us to further improve our communication strategy on Twitter and beyond.

1. How many tweets are being posted by the various Universities when? Are there any “release“ strategies visible?
2. What are the tweets about and how do other Twitter users react to them (likes, etc.)?
3. How do the university tweets differ in terms of content, style, emotions, etc?
4. What specific advice can you give us as communication department of BFH based on your analysis? How can we integrate the analysis of tweets in our internal processes, can you think of any data products that would be of value for us?

Dataset	Tweets created by one of the swiss universities of applied sciences (until Jan/23)
Quantity	19.575 tweets
File	Tweets_all.rda
Variables	ID & Id_str(of the tweet as integer and as string), full_text, in_reply_to_screen_name, retweet_count (how often a tweet was re-posted by Jan/23), favorite_count (how often a tweet was liked by Jan/23), lang (language of the tweet), university (which univ. of applied science created the tweet), Various forms the tweet‘s timestamp (e.g. created_at, tweet_date, tweet_month, tweet_minute).
Further hints.	How to deal with re-tweets, automated tweets & chatbots and emojis?

We are Bern DigiLab, a medium-sized consulting company offering services in the field of digital transformation and analytics. Since the release of ChatGPT, our clients are increasingly asking us whether it would make sense to use this technology in their business. We are equally excited about this novel technology, yet cannot fully grasp of what it actually can do for us. Can you help us to tap the „wisdom of the crowd“ to find out?

1. What can you tell us about the users that tweet about ChatGPT?
2. What are the tweets about, what do users associated the new technology with (e.g. industries, specific applications, and also emotions)?
3. How did the excitement and topics developed over time?
4. What specific advice can you give us as consulting company, how could we use ChatGPT in clients projects, what use cases are being discussed? Is ChatGPT just a hype or here to stay? How can we integrate the analysis of tweets related to new trends/technologies in our internal processes, can you think of any data products that would be of value for us (technology scouting)?

Dataset	A collection of tweets with the hashtag #chatgpt : discussions about the chatgpt language model, sharing experiences with using chatgpt, or asking for help with chatgpt-related issues. Overall, the collection of tweets provide a glimpse into the online conversation surrounding the new technology. (Nov/22-Jan/23)
Quantity	179.252 tweets
File	Tweets_all.rda
Variables	Tweet (the actual tweet text), additional information of the tweet (e.g. Retweets, Likes) and the user who created it (e.g. UserFollowers, UserFriends, Location),various forms the tweet‘s timestamp (e.g. created_at, tweet_date, tweet_month, tweet_minute).
Further hints.	How to deal with re-tweets, automated tweets & chatbots and emojis?

Text mining project – Case Study

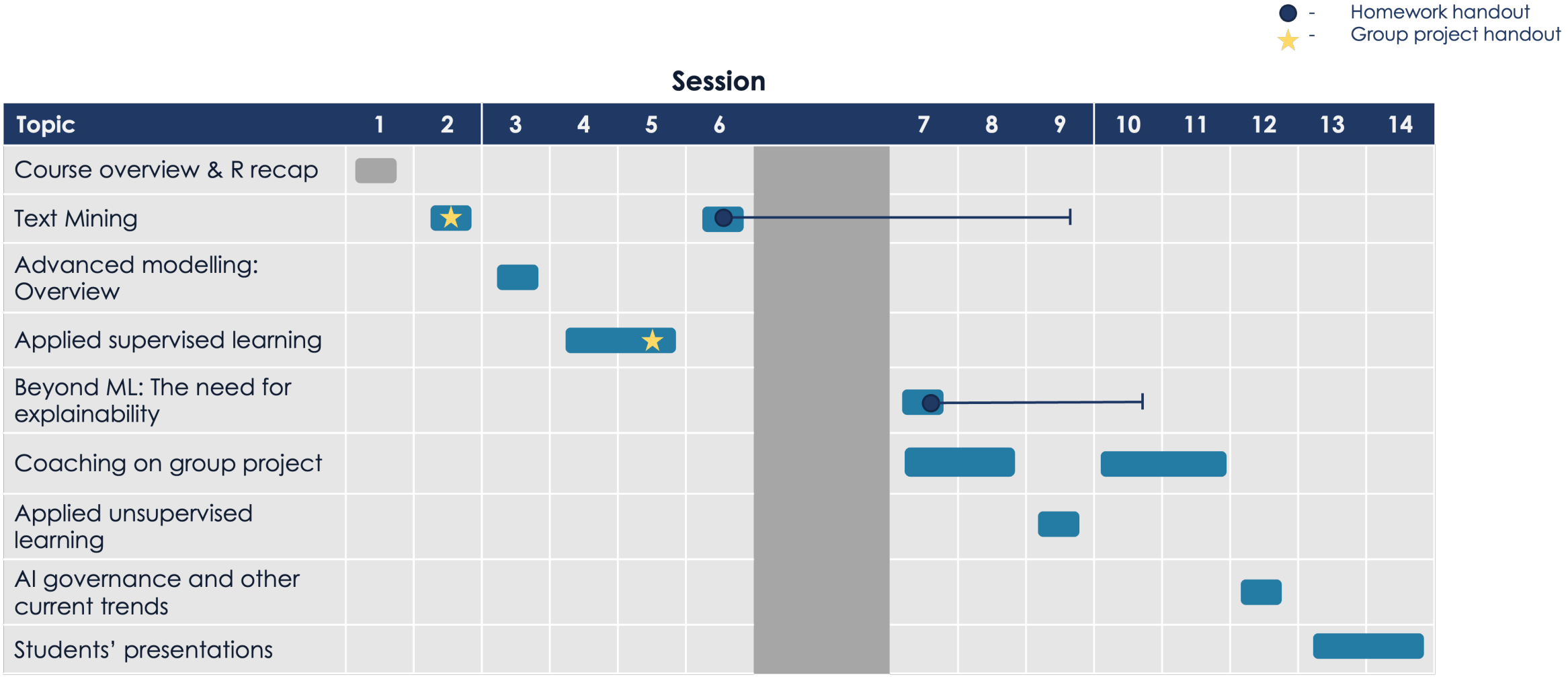
- Discuss your preferences regarding the case studies in your group and rank case studies from 1 (=like most) to 4 (=like least). *If you have further questions, don't hesitate to contact me.*

- Send us your preferences, i.e. your ranking, per e-mail: patrick.cichy@bfh.ch until **2nd of March 2023** at the very latest

- We will try our best to consider your preferences, but cannot promise that you get your first choice

- Each case study will be assigned two times (at maximum) → Allocation of groups to case study will be announced on Moodle

Course overview



Notes:

- Easter break (no classes): 4th and 11th of April

Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Thank you and see you next week!