# gregoriades_2021_supporting_digital_content_marketing_and_messaging_through_topic_modelling_and_decision_trees

## Year

2021

## Author(s)

Andreas Gregoriades and Maria Pampaka and Herodotos Herodotou and Evripides Christodoulou

## Title

Supporting digital content marketing and messaging through topic modelling and decision trees

## Venue

Expert Systems With Applications

---

## Topic labeling

Partially automated

## Focus

Secondary

## Type of contribution

Established approach

## Underlying technique

Manual labeling assisted by topic polarity & associated documents

## Topic labeling parameters

\

## Label generation

In each STM model, each review (i.e. document) is associated with a distribution of a finite set of topics; topics are distributions of words in the corpus of all reviews. The probability distribution of topics per review denotes the probability of each topic discussed in a review and the sum of all topics' probabilities in each review totalled to 1.

–

A popular approach for labelling topics, also used herein, is to

1. Consider words highly associated with each topic and
2. Inspect the most prevalent reviews related to that topic.

For the first task, extracted topics were inspected to make the necessary connections between the main words associated with the topic that emerged from different metrics such as highest probability, Lift, and FREX (Roberts et al., 2019).
Lift weights words by giving higher weight to words that appear less frequently in other topics and FREX weights words by their overall frequency and how exclusive they are to the topic.

To assist the interpretation of topics, it was essential to identify the polarity of each topic based on its association to satisfaction (sentiment).
To that end, the STM's metadata capability was used with sentiment set as prevalence variable (metadata) to assess the association of each topic with the outcome (sentiment) using LASSO based estimates. For topics that were not included in the LASSO output a logistic regression model was generated per topic with the reviews' sentiment as the dependent variable and the topics' theta values from STM models as independent variable.
The evaluated topics' polarity helped to refine the naming of topics to have either positive or negative connotations, to later enable the interpretation of the DTs. The main topics that emerged from the analysis are depicted in Table 1.

**Table 1**
Naming of topics derived from the two STM models for four/five (right) and two/three star hotels (left).

| Defined Topics for 2/3-star hotels based on STM model | Evaluated Topic Polarity | Defined Topics for 4/5-star hotels based on STM model | Evaluated Topic Polarity |
|---|---|---|---|
| V1_clubbing_holidays (Noise) | Negative | V1_revisited_refurnished | Positive |
| V2_renovation | Negative | V2_bar_food_cats | Negative |
| V3_good_rooms | Positive | V3_nice_pool_drinks | Positive |
| V4_dirty_bathroom | Negative | V4_spa_gym_massage | Positive |
| V5_dirty-room-unprofessional_staff | Negative | V5_staff_always_willing_to_help | Positive |
| V6_helpfull_staff | Positive | V6_rude_staff | Negative |
| V7_other_guests | Negative | V7_excellent_service | Positive |
| V8_limited_options_all_inclusive | Negative | V8_best_stay_ever | Positive |
| V9_good_location | Positive | V9_nice_room | Positive |
| V10_amazing_staff | Positive | V10_extra_costs_for_amenities (WIFI, coffee machines) | Negative |
| V11_not_well_equipped_room(e.g. fridge) | Negative | V11_extra_charges | Negative |
| V12_good_mobility_options(bus, walk) | Positive | V12_amazing_dinners (buffet, a la carte) | Positive |
| V13_comfortable_room | Positive | V13_pool_area_issues (kids, noise) | Negative |
| V14_poor_entertaiment_and_food | Negative | V14_good_transportation_options | Positive |
| V15_limited_breakfast_options | Negative | V15_located_close_to_beach | Positive |
| V16_pool_area | Positive | V16_great_wedding_venue | Positive |
| V17_value_for_money | Positive | V17_birthday/anniversaries | Positive |
| V18_close_to_beach_with_pool | Positive | V18_perfect_location(walk around) | Positive |
| V19_basic_accomotation | Negative | V19_disagree_with_hotel_reviews | Positive |
| V20_accept_late_arrivals | Positive | V20_luxurius/quality -amenities | Positive |
| V21_good_dinner | Positive | V21_great_team_of_staff | Positive |
| V22_low_quality_food_and_drinks | Negative | V22_room_with_seaview | Positive |
| V23_smelly_room | Negative | V23_lovely_staff | Positive |
| V24_good_bar_service | Positive | V24_dirty_room_beddings_ shower | Negative |
| V25_ideal_for_summer_holidays (pool_area) | Positive | V25_missing_services | Negative |
| V26_bad_customer_service | Negative | V26_great_entertaiment (bingo, darts) | Positive |
| V27_old_room | Negative | V27_great_for_families | Positive |
| V28_good_entertaiment | Positive | V28_checkin/front_desk_issues | Negative |
| V29_friendly_staff | Positive | V29_beach_hotel_with_convenient_location | Positive |
| V30_will_revisit | Positive | V30_exceptional_staff-food-pool | Positive |
| V31_basic_apartment_no_luxuries | Negative | | |
| V32_the_best_experiance | Positive | | |
| V33_close_to_beach | Positive | | |
| V34_bad_facilities_maintenance | Negative | | |
| V35_clean_good_service | Positive | | |

## Motivation

1. Topics naming was motivated by prevalent hotel service quality factors in hospitality literature (Banerjee & Chua, 2016) such as convenient location, service quality, reputation, and friendliness of staff and factor groupings such as tangibles (e.g., equipment), reliability (e.g., punctuality), responsiveness (e.g., prompt service), assurance (e.g., politeness), and empathy (e.g., personal attention).
2. Enable for the interpretation of the generated DTs.

---

## Topic modeling

STM

## Topic modeling parameters

Nr of topics (k): 10 to 100

## Nr. of topics

One model per review type and corresponding data (2/3-star and 4/5-star hotels)
30 topics for the 4/ 5 hotels and 35 topics for the 2/3 star hotels

## Label

Single or multi-word labels linked to hotel service quality factors in hospitality literature (Banerjee & Chua, 2016)

## Label selection

\

## Label quality evaluation

\

## Assessors

\

## Domain

Paper:
Dataset: Tourism and Hospitality

## Problem statement

This paper presents a machine learning approach involving tourists' electronic word of mouth (eWOM) to support destination marketing campaigns.
This approach enhances the communication of the right content to the right consumers. The proposed method further considers aggregate cultural and economic-related information of the tourists' country of origin with topic modelling and Decision Tree (DT) models. Each DT addresses different dimensions of culture and purchasing power and the way these dimensions are associated with the topics discussed in eWOM, thus revealing patterns relating tourists' experiences with potential explanations for their dissatisfaction/ satisfaction.
Patterns emerged from the extraction of rules from DTs illuminate combinations of variables associated with tourist experience (negative or positive) for each of the two hotel categories and verify the asymmetric relationship between service performance and satisfaction.

## Corpus

Origin: TripAdvisor
Nr. of documents: 42.000
Details:

- eWOM data, written in English, by tourists in Cyprus who stayed in (reviewed) hotels between 2014 and 2019
- The data were retrieved using location filtering criteria in TripAdvisor
- Two main categories of hotels, namely 2/3-star and 4/5-star hotels, the reviews were split into two datasets, corresponding to each hotel category and processed separately. Each dataset was used to develop a topic model.

## Document

Information from eWOM related to the generator, the hotel, and timing of visit and review, and included the reviewer's nationality based on their self-declared location, their past eWOM contributions, helpfulness vote, hotel rating, hotels star rating, and actual review text.
+
Level of economic development (as measured by nominal GDP per capita in US$ at the time of the review) and cultural characteristics (as measured by scores on the six cultural dimensions) for both Cyprus (i.e., the destination) and the reviewer's home country.

## Pre-processing

- stop-word removal
- stemming
- tokenization
- punctuation removal
- custom words removal (names of hotels, cities, and resorts)
- numbers removal
- converting all text to lowercase
- Reviews from local tourists were also eliminated

---

```
@article{gregoriades_2021_supporting_digital_content_marketing_and_messaging_th
rough_topic_modelling_and_decision_trees,
```

```
  abstract = {This paper presents a machine learning approach involving tourists' electronic word of mouth (eWOM) to support destination marketing campaigns. This approach enhances optimisation of a critical aspect of marketing campaigns, that is, the communication of the right content to the right consumers. The proposed method further considers aggregate cultural and economic-related information of the tourists' country of origin with topic modelling and Decision Tree (DT) models. Each DT addresses different dimensions of culture and purchasing power and the way these dimensions are associated with the topics discussed in eWOM, thus revealing patterns relating tourists' experiences with potential explanations for their dissatisfaction/satisfaction. The method is implemented in a case study in the context of tourism in Cyprus focusing on two hotel groups (2/3 and 4/5 stars) to account for their differences. Patterns emerged from the extraction of rules from DTs illuminate combinations of variables associated with tourist experience (negative or positive) for each of the two hotel categories and verify the asymmetric relationship between service performance and satisfaction. The approach can be used by management during marketing campaigns to design messages to better address the desires and needs of tourists from different cultural and economic backgrounds, as these emerge from the data analysis.},
  author = {Andreas Gregoriades and Maria Pampaka and Herodotos Herodotou and Evripides Christodoulou},
  date-added = {2023-03-19 17:37:44 +0100},
  date-modified = {2023-03-19 17:37:44 +0100},
  doi = {https://doi.org/10.1016/j.eswa.2021.115546},
  issn = {0957-4174},
  journal = {Expert Systems with Applications},
  keywords = {Topic modelling, Cultural and economic distance, Decision trees, Shapley additive explanation, Tourists' reviews},
  pages = {115546},
  title = {Supporting digital content marketing and messaging through topic modelling and decision trees},
  url = {https://www.sciencedirect.com/science/article/pii/S0957417421009532},
  volume = {184},
  year = {2021}}
```

#Thesis/Papers/Initial