**Topic Modeling for Smart Watches**

**Scope:** Our goal is to conduct topic modeling on customer reviews for different smartwatch brands in the world, to uncover the underlying themes or topics in a collection of documents. The final objective of this project would be to determine which smartwatch brand has the highest percentage of positive reviews and the least percentage of negative reviews and identify which features and aspects of the smartwatches are most important to customers and which brands excel in those areas.

We’ve collected reddit comments corresponding to 5 different smartwatch brands –

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | **Brand Name** | **No. of Comments** |  |
|  | Apple Watch | 1,011 |  |
|  | Fitbit | 1,025 |  |
|  | Galaxy Watch | 1,000 |  |
|  | Garmin Watch | 1,002 |  |
|  | Huawei Watch GT | 1,006 |  |
|  | **Combined** | **5,044** |  |
|  |  |  |  |

1. **Top 15 words from the dataset:**

The words which are repeated the most in the combined dataset is given below. The frequency of the words can give us an idea about the trend the users are following and the choice of words they are using to express their sentiment towards a particular smart watch brand.

|  |  |
| --- | --- |
| **Words** | **Count** |
| 'watch' | 1788 |
| 'like' | 663 |
| 'sleep' | 550 |
| 'time' | 435 |
| 'use' | 431 |
| battery' | 415 |
| 'phone' | 407 |
| 'app' | 394 |
| 'garmin' | 360 |
| 'apple' | 331 |
| huawei' | 320 |
| 'good' | 319 |
| 'think' | 317 |
| 'day' | 315 |
| 'got' | 264 |

1. **The parameters for the best-performing topic model:**

LatentDirichletAllocation is used to find the important topics based on the frequency of the words used in the dataset. We have identified the parameters below as the important parameters in order to obtain optimal results. This is done by running a grid search through the various available parameters and their corresponding values.

|  |  |
| --- | --- |
| **Parameters** | **Values** |
| 'doc\_topic\_prior' | 0.1 |
| 'learning\_method' | 'online' |
| 'max\_iter' | 10 |
| 'n\_components' | 3 |
| 'topic\_word\_prior' | 1 |

'doc\_topic\_prior' - A hyperparameter that determines the prior probability distribution of topics in each document.

'learning\_method' – Determines the method used to estimate the model parameters from the input corpus.

'max\_iter' – A hyperparameter that determines the maximum number of iterations to run during model training

'n\_components' – This determines the number of topics the model should extract from the available data.

'topic\_word\_prior' - A hyperparameter that determines the prior probability of each word given a topic.

1. **Identify 2-3 topics that are meaningful to the project objectives, report the frequent words associated with these topics, and explain how these topics are relevant to the project:**

**Based on the top words for each topic, LatentDirichletAllocation model has identified three topics related to smartwatches:**

**Topic 0:** This topic seems to be related to Comparison of Smartwatches from Huawei and Garmin and features associated with them such as their dimensions, face, and features like "mm" and "celest".

**Topic 1:** This topic seems to be related to Review and Comparison of Smartwatches for Daily Use. The presence of "sleep," "time," "day," and "battery" indicate that the topic could be related to features and functionality that are useful for daily wear.

**Topic 2:** This topic seems to be related to Cleaning and Maintenance of Devices. The words "water," "clean," "dry," "wash," "swim," "soap," "shower," and "washing" suggest that this topic could involve the cleaning and maintenance.

Using these topics, we can obtain a clear understanding of the theme of a particular brand and product. This can help us to perform further analysis in order to obtain the best and the most sort after smart watches.

**The most frequently used words in these topics are:**

**Topic 0:** ['https', 'com', 'watch', 'huawei', 'garmin', 'mm', 'www', 'celest', 'gt', 'en', 'watches', 'face', '46', '3', 'r']

**Topic 1:** ['watch', 'like', 'sleep', 'time', 'work', 'day', 'app', 'use', 'phone', 'good', 'battery', 'think', 'apple', '2', 'look']

**Topic 2:** ['thank', 'water', 'clean', 'afib', 'dry', 'wash', 'swim', 'soap', 'white', 'blue', 'shower', 'ok', 'hand', 'machine', 'washing']

**Visualizing the topics and words:**

Graphical user interface

Description automatically generated

The visualization displayed on the right-hand side depicts the 30 most pertinent words per topic. The blue shaded bar reflects the frequency of the word across all reviews, while the red bar indicates the occurrence of the word within the selected topic. Additionally, a slide is available to adjust the relevance metric λ, where 0 ≤ λ ≤ 1. When λ = 1, the visualization prioritizes words that are most likely to occur within each topic, and when λ = 0, the visualization emphasizes words that are unique to the selected topic.