As is known, STM was originally developed with the R package to examine the relationship between the topics and the data dimensions, which are referred to as covariates and metadata.

Understanding and focusing on the terms **Metadata** and **Covariate** related to STM is important. For this purpose, you can take a look at the quotes I have made from the links below.

**1.** <https://cran.r-project.org/web/packages/stm/stm.pdf>

"The Structural Topic Model (STM) allows researchers to estimate topic models with document-level covariates."

**2.** <https://towardsdatascience.com/introduction-to-the-structural-topic-model-stm-34ec4bd5383>

"STM allow us to incorporate metadata into our model and uncover how different documents might talk about the same underlying topic using different word choices.

We might imagine that both the topic prevalence and topic content for a specific document is correlated with “metadata” about the document. For instance, certain sources may be more likely to write about politics or write about politics in a particular way. Metadata can include date published, author, publication, likes on social media, or any number of categorical or numerical variables about a document."

**3.** <https://github.com/mkrcke/strutopy>

"Structural Topic Model (Roberts et al. 2016) can be used to extend the former (LDA) topic modelling approaches by including text metadata on a document level. The meta information can be introduced to the estimation procedure two-fold, via:

1. topical content covariates that shape the word usage within topics
2. topical prevalence covariates that shape the frequency of topic occurrences."

**4.** Finally, I will provide feedback to complement the above, as follows.

We plan to compare LDA, SMT, and BERTopic topic modeling techniques. SMT and BERTopic are updated and improved methods compared to LDA.

The Twitter dataset we are working on contains a time column/dimension in addition to a content column (Twitter messages) that we will use to predict topics. So, in our data, the topic is a dimension/component and time is another dimension/component.  Each topic we will examine evolves over time. While modeling the topics, we should consider the time dimension, as well. In other words, wee will model how topics (in a corpus of documents) evolve over time.

If a time-based analysis is performed, one should choose one of the temporal algorithms. Dynamic/temporal topic modeling refers to the inclusion of a temporal (time) dimension in the topic modeling analysis. In other words, it means to examine the change over time of specific topics.

Here, I would like to explain the concept of "**covariate**", which we often come across while reading papers. In statistic and machine learning fields, it is called “covariate” any variable that you consider during your analysis in addition to your variable of interest (targeted variable) [1,2]. In your data/project, topics to be analyzed or modeled are the main or targeted variables in the data. Then, the time dimension is covariate.

LDA is well-know and classical/conventional/traditional topic modeling methods. The traditional LDA model is not intended for modeling temporal document collections; however, Griffiths et al demonstrated how simple time-stratified estimators can be used to illustrate the evolution of latent topical vectors over time [3]. If the complete dataset is split into corpora covering a specific time interval, the LDA can be applied to each individual corpus and then it is possible to analyze how each topic evolves over time [4]. The structural topic model (STM), extends the classical LDA model, allowing either (1) the matrix of per-document topical prevalence weights or (2) the matrix of per-topic word probabilities to deterministically vary according to covariate information parameterized using a generalized linear model [5]. In other words, STM is an extension of the LDA process which allows "covariates" of interest (such as the temporal origin or location of the topic) to be included in the prior distributions for topic proportions and topic-word distributions [6]. So, the most important feature that distinguishes the STM algorithm from the others is that covariates can be included in the model for each document.

References:

1. <https://www.statology.org/covariate/>

2. <https://www.foldercase.com/blog-covariates-and-confounders-an-introduction.php>

3. Comparison of Methods for Estimating Temporal Topic Models From Primary Care Clinical Text Data: Retrospective Closed Cohort Study

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9808604/>  **Note:** Please read the section of "Topic Models" in this article for comparision aspects of the models.

4. <https://www.statcan.gc.ca/en/data-science/network/topic-modelling> **Note:** Please read this carefully.

5. Roberts ME, Stewart BM, Airoldi EM. A model of text for experimentation in the social sciences. J Am Stat Assoc. 2016 Oct 18;111(515):988–1003.

6. A structural topic model approach to scientific reorientation of economics and chemistry after German reunification

<https://link.springer.com/article/10.1007/s11192-020-03640-0>