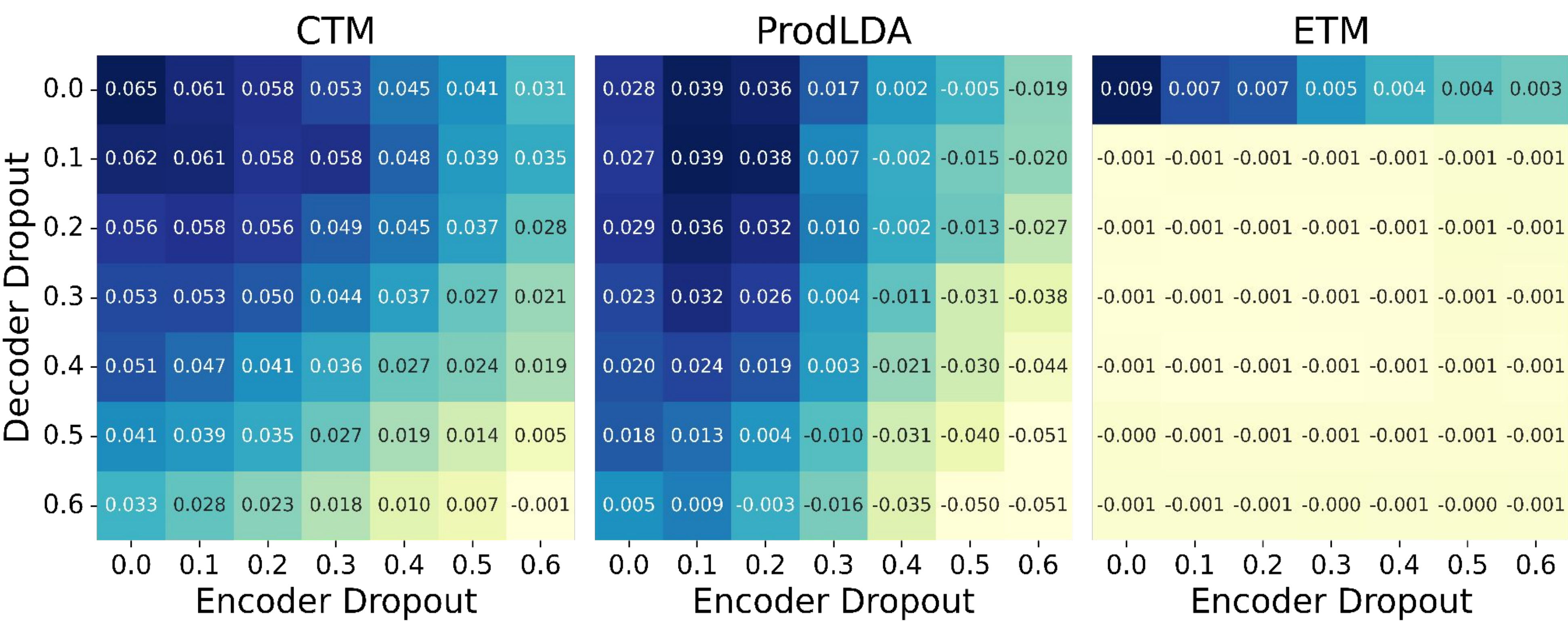


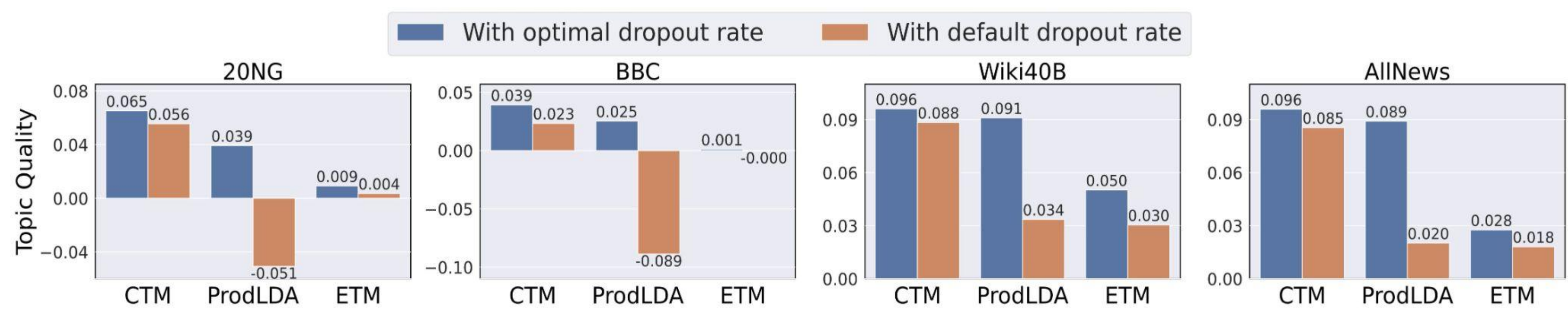
# Do Neural Topic Models Really Need Dropout?

## Analysis of the Effect of Dropout in Topic Modeling



Topic quality on 20NG dataset.

Low dropout rate leads to a significant improvement in the performance of VAE-NTMs



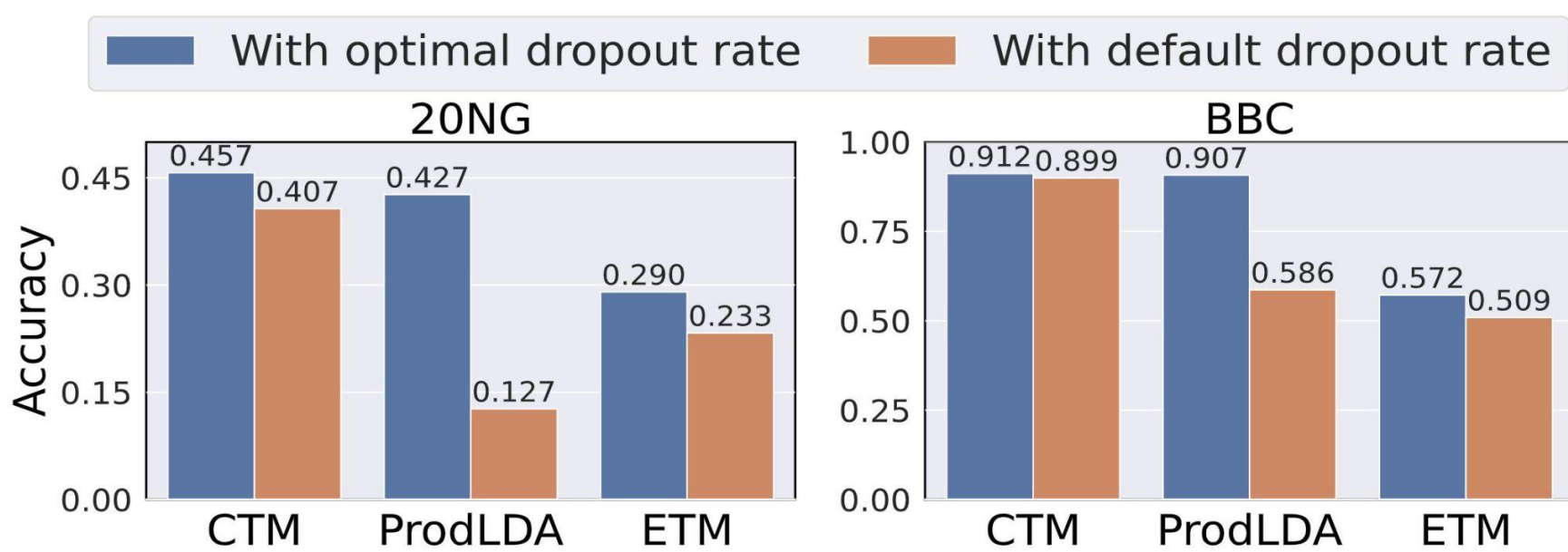
Topic quality for models with optimal and default dropout rates.

### Qualitative Evaluation

| Model              | Topics  |
|--------------------|---|
| ProdLDA (0.1, 0.1) | window, driver, <i>mode</i> , run, mouse, session, server, program, manager, install              |
|                    | car, engine, buy, company, vehicle, <i>make</i> , brake, tire, dealer, road                       |
|                    | signal, voltage, output, circuit, noise, power, switch, wire, connector, <i>degree</i>            |
| ProdLDA (0.6, 0.6) | <i>line</i> , window, <i>gun</i> , read, space, run, statement, datum, drive, <i>make</i>         |
|                    | <i>make</i> , battery, engine, <i>homosexual</i> , assault, reason, place, single, large, attempt |
|                    | voltage, <i>damn</i> , signal, usual, label, hour, bio, leg, bullet, hundred                      |

Some selected topics from 20NG.

### Extrinsic Evaluation

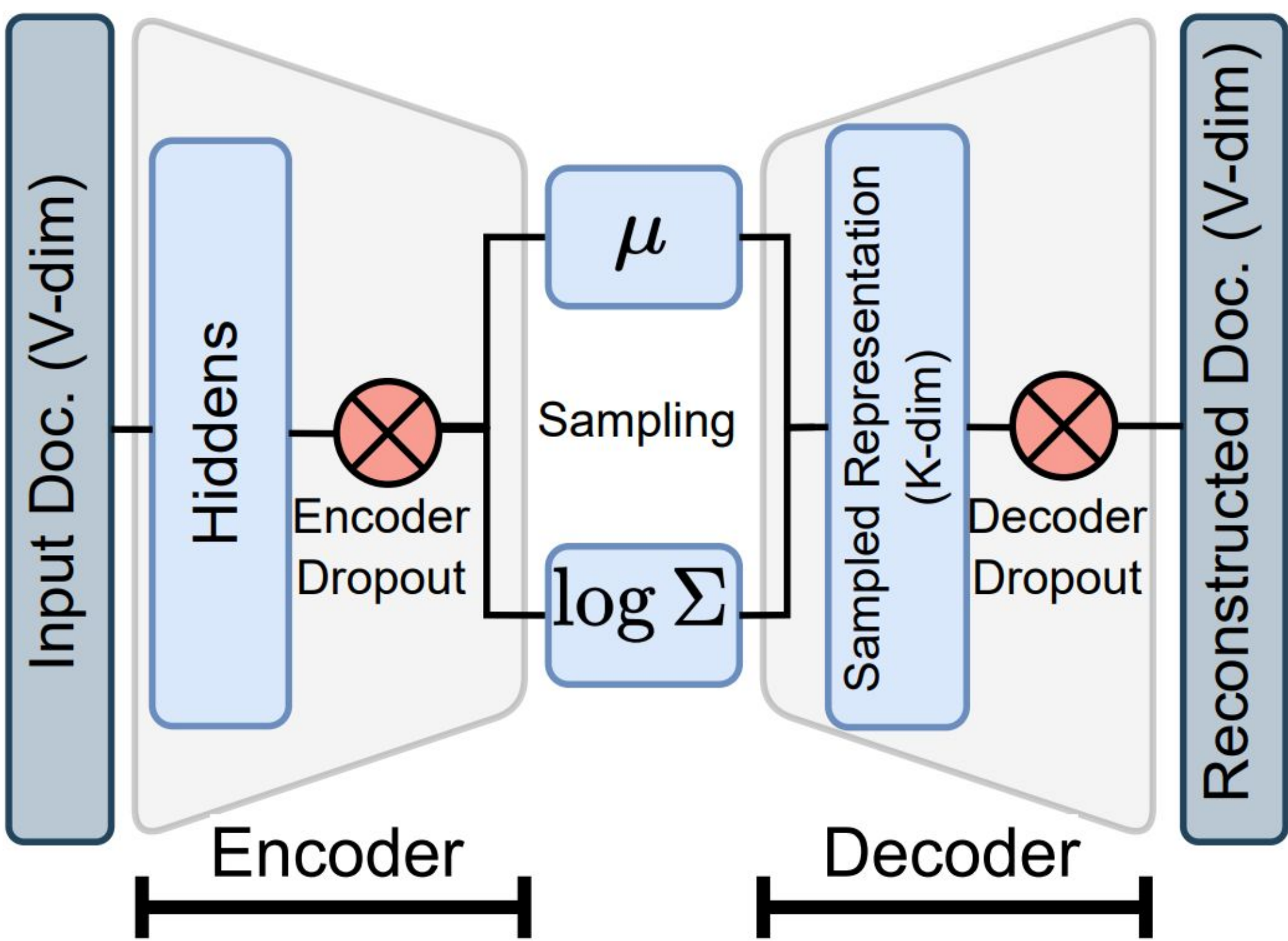


Accuracy scores for topic models with **optimal** and **default** dropout rates in document classification task.

Suman Adhya  
Avishek Lahiri  
Debarshi Kumar Sanyal



### VAE Framework in Neural Topic Models (NTMs)



#### Encoder:

Input: Document representation

Dropout on: Output of the hidden layer(s)

Returns: Posterior distribution

#### Decoder:

Input: Document-topic distribution vector

Dropout on: Document-topic distribution

Returns: Reconstructed document

### Experimental Setup

| Dataset | Type                   | #Docs |
|---------|------------------------|-------|
| 20NG    | Newsgroups posts       | 16309 |
| BBC     | News articles from BBC | 2225  |
| Wiki40B | Wikipedia text         | 24774 |
| AllNews | News articles          | 49754 |

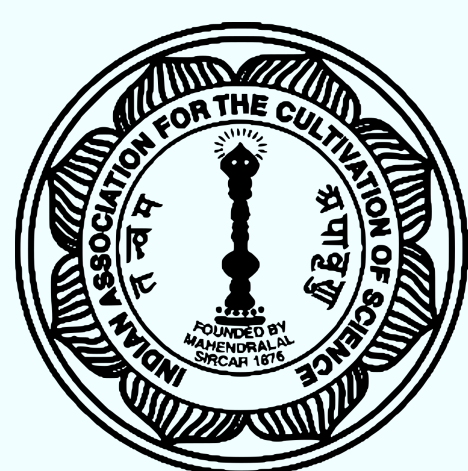
#### Evaluation metrics:

- NPMI: Topic-word relevancy
- TD: Topic diversity
- Topic Quality: NPMI  $\times$  TD

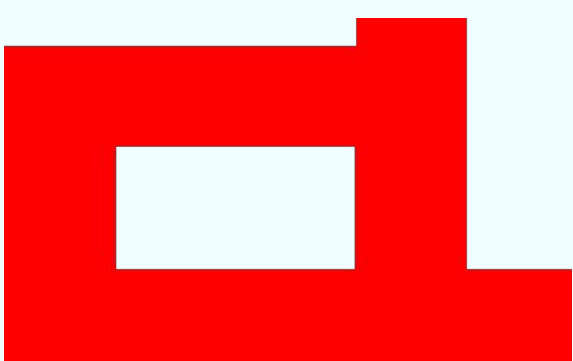
### Optimal Dropout Rates

| Model              | 20NG       | BBC        | Wiki40B    | AllNews    |
|--------------------|------------|------------|------------|------------|
| CTM (0.2, 0.2)     | (0.0, 0.0) | (0.0, 0.0) | (0.2, 0.1) | (0.0, 0.1) |
| ProdLDA (0.6, 0.6) | (0.1, 0.1) | (0.0, 0.0) | (0.1, 0.1) | (0.1, 0.1) |
| ETM (0.5, 0.0)     | (0.0, 0.0) | (0.1, 0.0) | (0.0, 0.0) | (0.1, 0.0) |

Default dropout rates and dataset-dependent optimal dropout rates for (encoder, decoder) of each model.



Indian Association for the Cultivation of Science  
Jadavpur, Kolkata 700032, INDIA



EACL  
2023