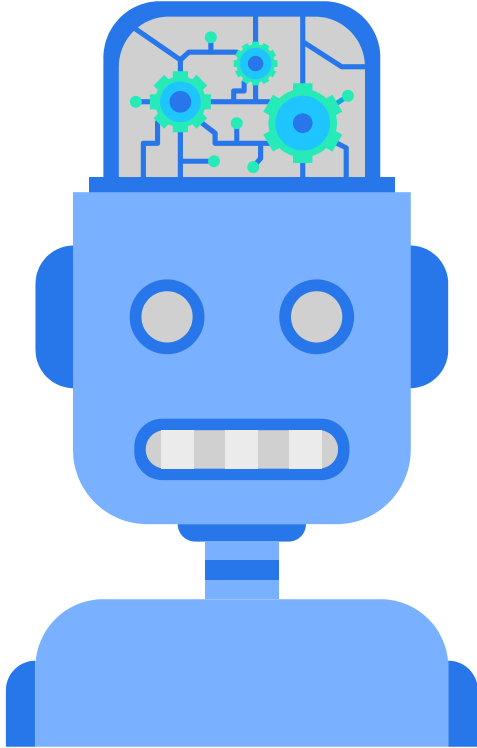




# **AN EXPLORATORY ANALYSIS OF GSDMM & BERTOPIC ON SHORT TEXT TOPIC MODELLING**



# PROBLEM STATEMENT

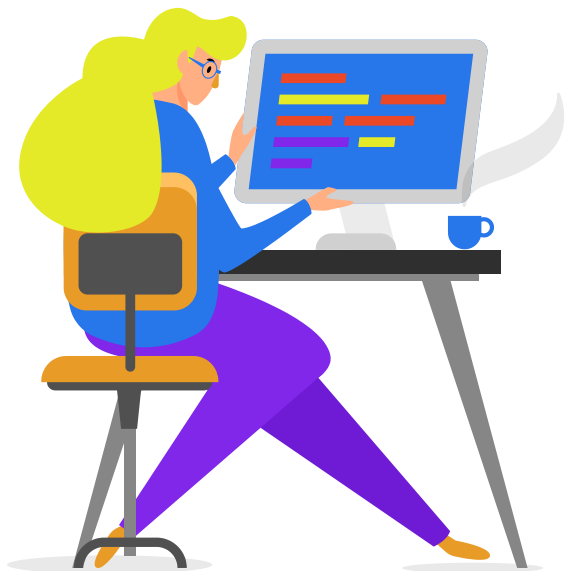
**To gain a thorough understanding of Short Text Topic Modelling and the application of BertTopic and GSDMM on three selected datasets. The two modelling methods used are compared in terms of Coherence.**

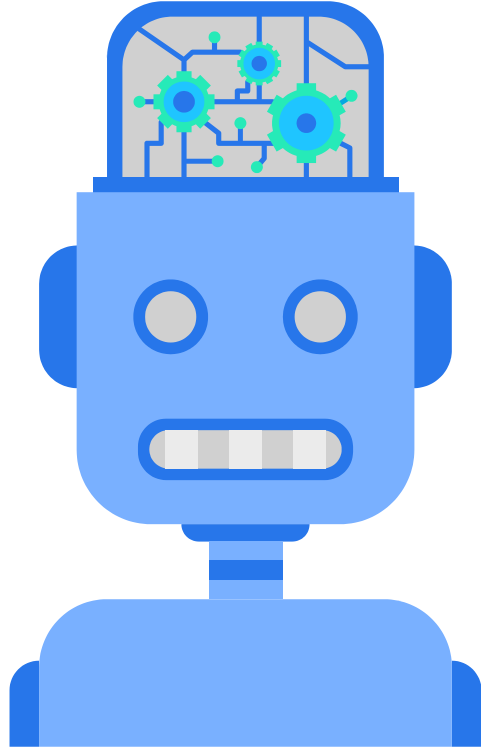
# LEARNING OUTCOMES

## Learning Outcomes

- Short Text Topic Modelling (STTM)
- GSDMM
- BERTopic

Applying these algorithms on tweets, headlines and stackoverflow queries, we have inferred that BERTopic performs better.





01



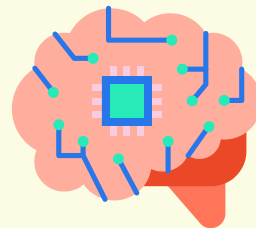
# **INTRODUCTION TO SHORT TEXT TOPIC MODELLING (STTM)**

# SHORT TEXTS

Short texts have become an important information source including news headlines, status updates, web page snippets, tweets, question/answer pairs, feedback, etc.

**For example:** There are huge floods in Karnataka

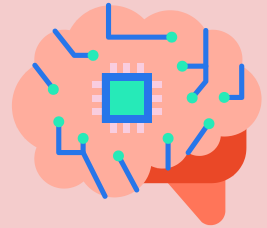
Short text analysis has been attracting increasing attention in recent years due to the ubiquity of short text in the real-world. Unlike paragraphs or documents, short texts are more ambiguous since they have not enough contextual information, which poses a great challenge for classification.



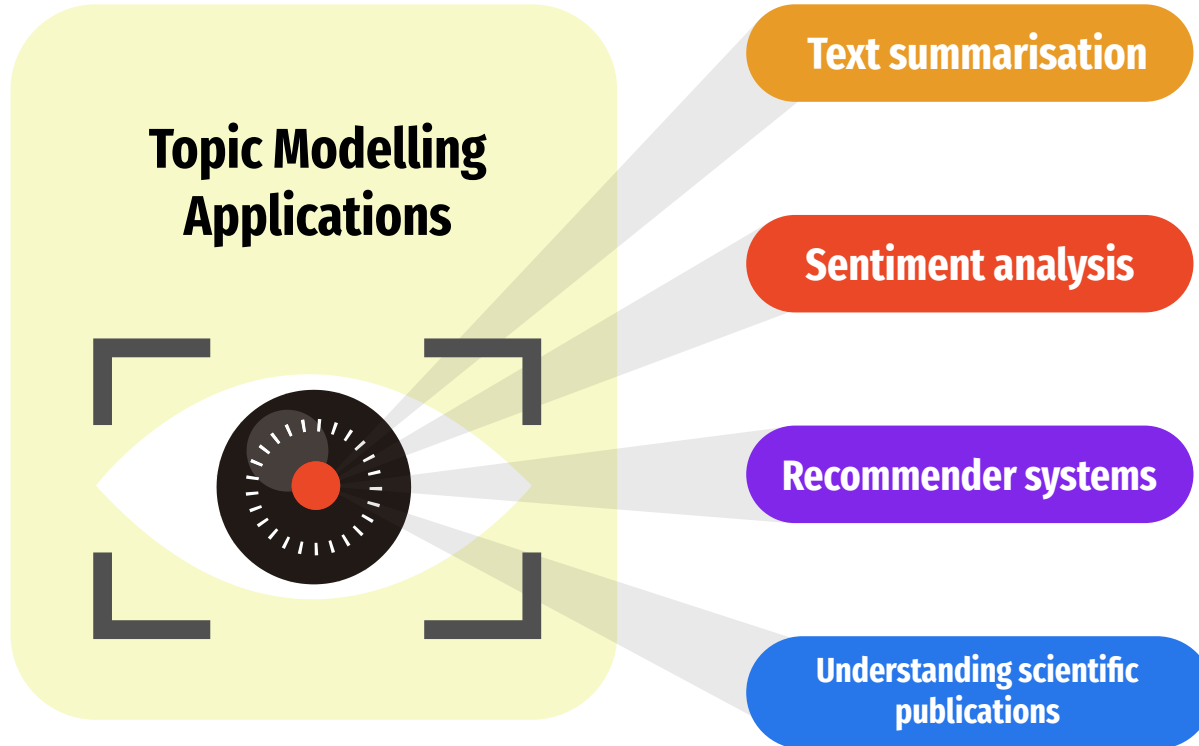
# TOPIC MODELLING

Topic modeling is an **unsupervised** machine learning technique that's capable of scanning a set of documents, **detecting word and phrase patterns** within them, and **automatically clustering word groups** and similar expressions that best characterize a set of documents.

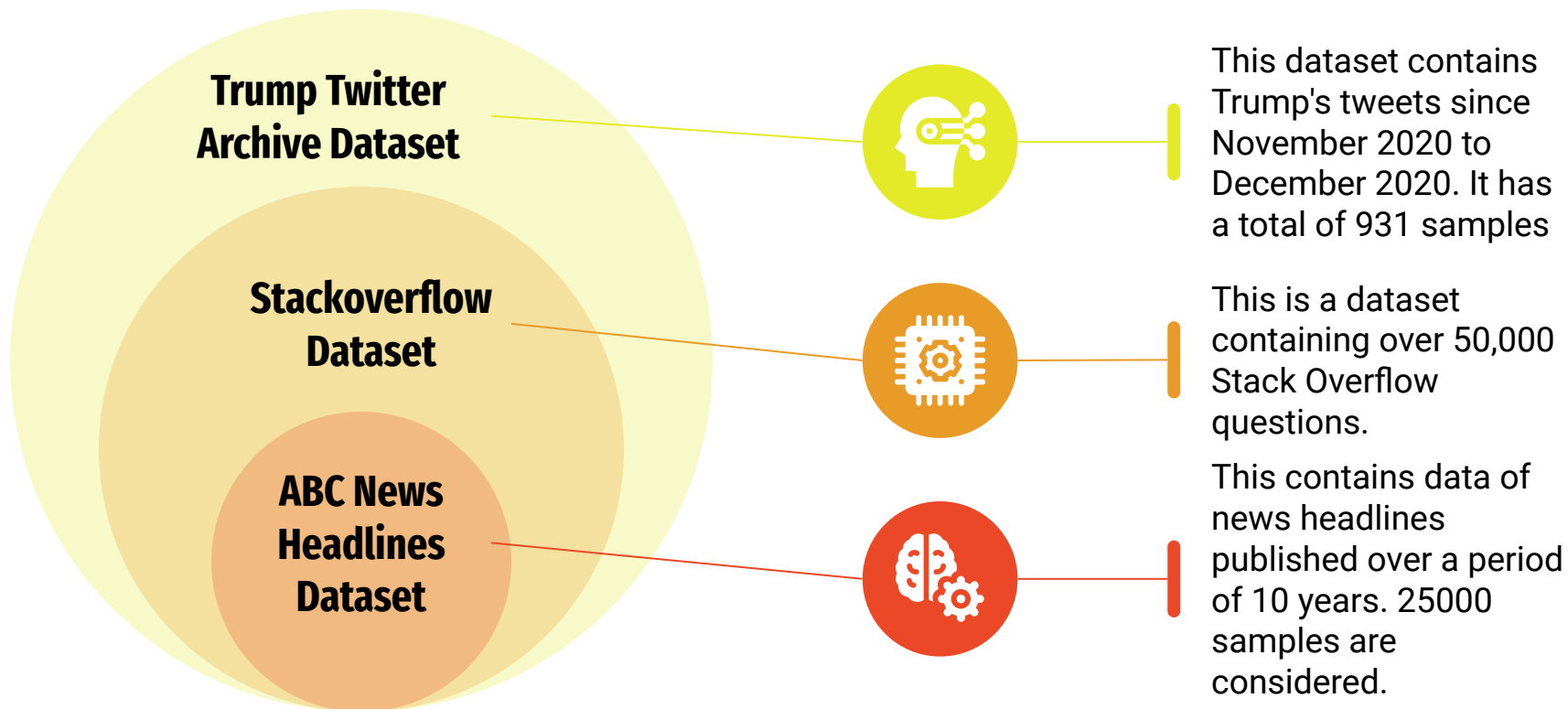
Topic Modeling basically aims to find the topics (or clusters) inside a corpus of texts (like mails or news articles), without knowing those topics at first. Here lies the real power of Topic Modeling, you don't need any labeled or annotated data, only raw texts.



# Topic Modelling Applications



# DATASETS





# DATASETS - Trump Twitter Archive Dataset

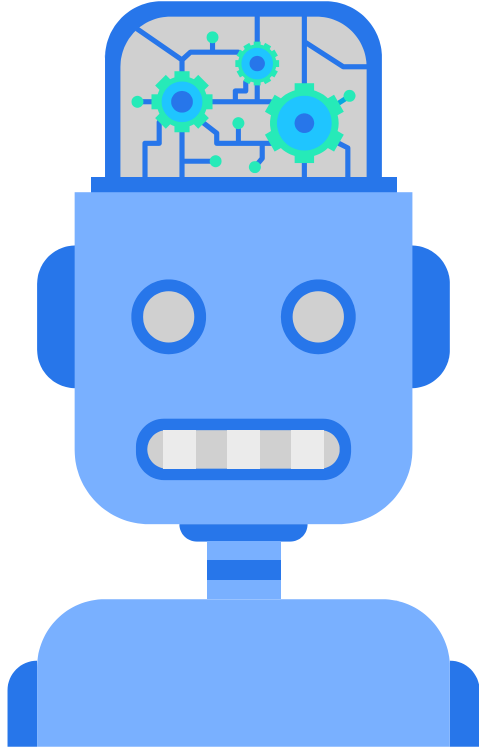
```
{
  0: {
    "date": 1608300872000
    "favorites": "212444"
    "id": "1339937091707351846"
    "isRetweet": false
    "retweets": "54898"
    "text":
      "@senatenajldr and Republican Senators have to get tougher, or you
      won't have a Republican Party anymore. We won the Presidential
      Election, by a lot. FIGHT FOR IT. Don't let them take it away!"
  }
  1: {
    "date": 1608299101000
    "favorites": "77814"
    "id": "1339929663804747778"
    "isRetweet": false
    "retweets": "16670"
    "text":
      "Well, at least she was happy when I pardoned Scooter Libby. We got
      a GREAT new Senator from her state, not Liz!
      https://t.co/tI4gdn8f9y"
  }
}
```

# DATASETS - ABC News Headlines Dataset

	headline_text
0	aba decides against community broadcasting licence
1	act fire witnesses must be aware of defamation
2	a g calls for infrastructure protection summit
3	air nz staff in aust strike for pay rise
4	air nz strike to affect australian travellers
5	ambitious olsson wins triple jump
6	antic delighted with record breaking barca
7	aussie qualifier stosur wastes four memphis match
8	aust addresses un security council over iraq
9	australia is locked into war timetable opp
10	australia to contribute 10 million in aid to iraq
11	barca take record as robson celebrates birthday in
12	bathhouse plans move ahead
13	big hopes for launceston cycling championship
14	big plan to boost paroo water supplies
15	blizzard huries united states in hills

## DATASETS - Stackoverflow Dataset

[illegible]

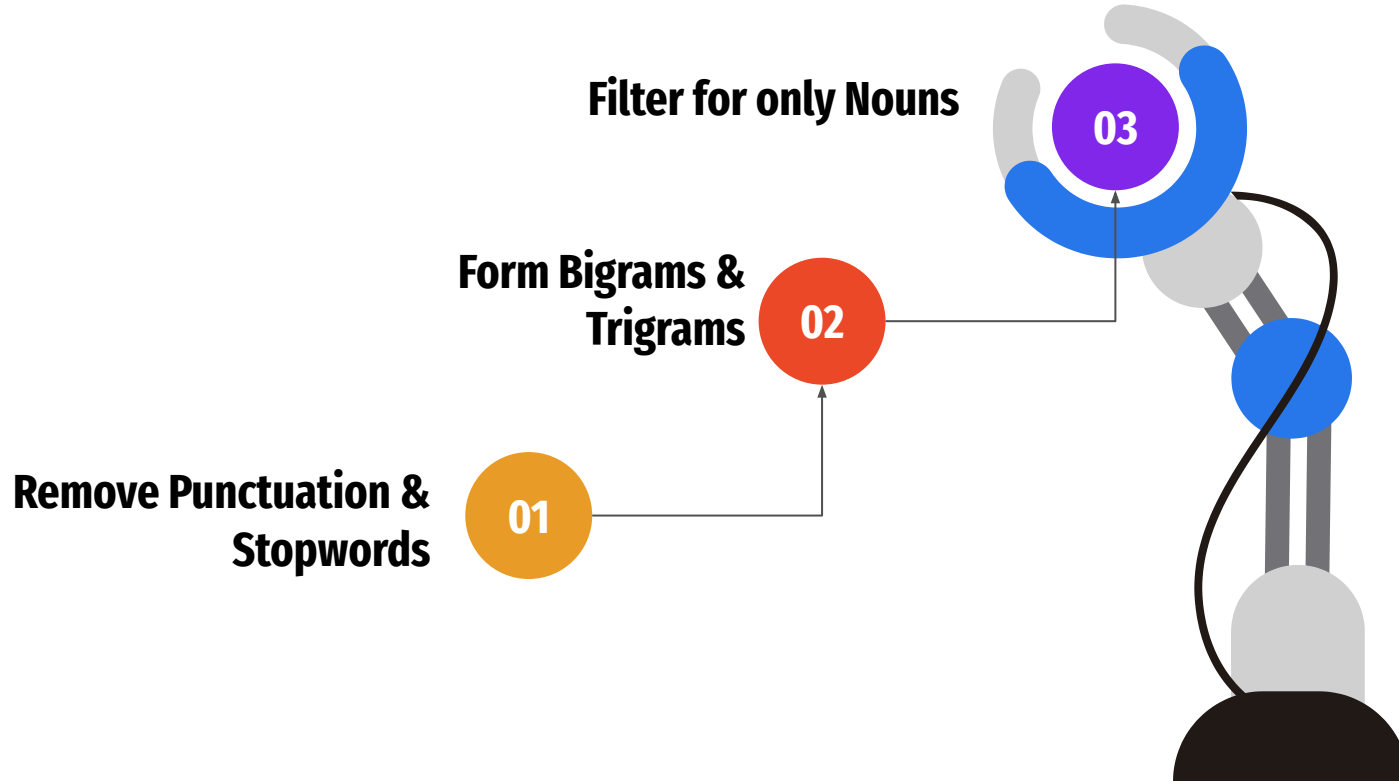


02



# **DATA PREPROCESSING**

# Steps Involved in Data Preprocessing



# Example for Data Preprocessing

How do we run "Streamlit App" on our computer?



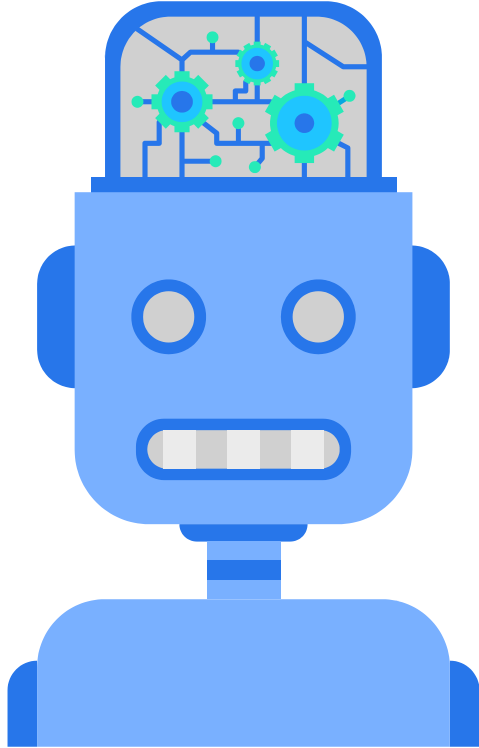
how do we run streamlit app on our computer



how do we run streamlit\_app on our\_computer



streamlit\_app, our\_computer

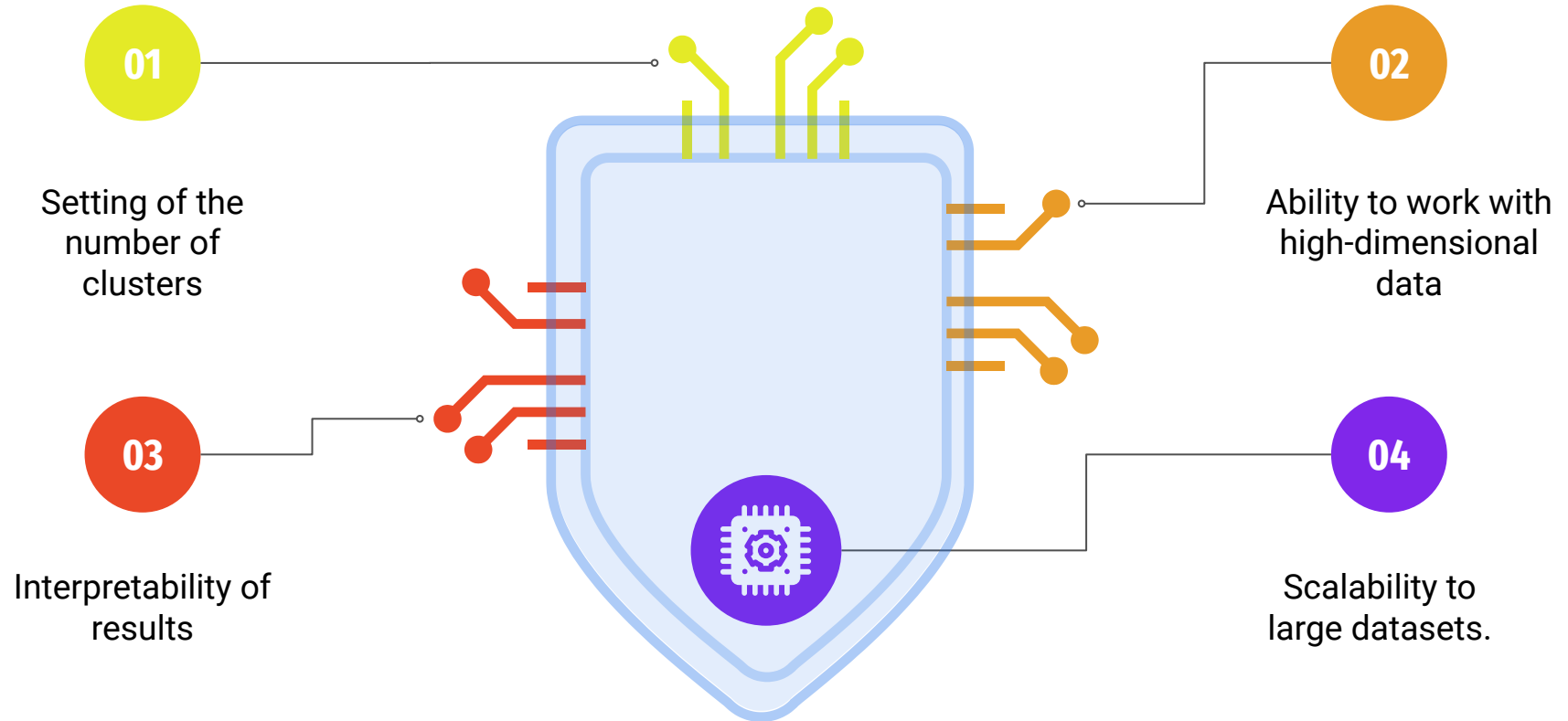


03



**GSDMM**

# Challenges faced in Short Text Topic modelling





# Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM)



- GSDMM is a short text clustering model that leverages Gibbs Sampling to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions.
- GSDMM can infer the number of clusters automatically with a good balance between the completeness and homogeneity of the clustering results, and is fast to converge.
- The basic principle of GSDMM is described using an analogy called “Movie Group Approach”.

# Algorithm Learning

**Data:** Documents in the input,  $\vec{d}$ .

**Result:** Cluster labels of each document,  $\vec{z}$ .

**begin**

    initialize  $m_z, n_z$ , and  $n_z^w$  as zero for each cluster  $z$

**for** each document  $d \in [1, D]$  **do**

        sample a cluster for  $d$ :

$z_d \leftarrow z \sim \text{Multinomial}(1/K)$

$m_z \leftarrow m_z + 1$  and  $n_z \leftarrow n_z + N_d$

**for** each word  $w \in d$  **do**

$n_z^w \leftarrow n_z^w + N_d^w$

**for**  $i \in [1, I]$  **do**

**for** each document  $d \in [1, D]$  **do**

            record the current cluster of  $d$ :  $z = z_d$

$m_z \leftarrow m_z - 1$  and  $n_z \leftarrow n_z - N_d$

**for** each word  $w \in d$  **do**

$n_z^w \leftarrow n_z^w - N_d^w$

            sample a cluster for  $d$ :

$z_d \leftarrow z \sim p(z_d = z | \vec{z}_{-d}, \vec{d})$  (Equation 4)

$m_z \leftarrow m_z + 1$  and  $n_z \leftarrow n_z + N_d$

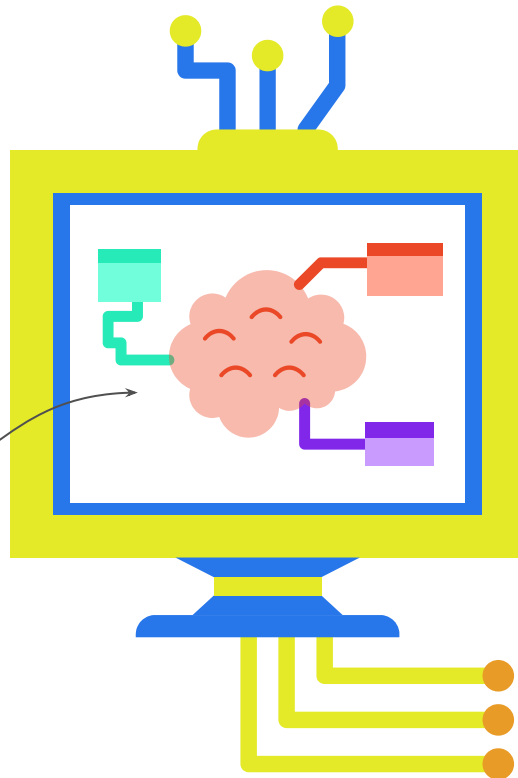
**for** each word  $w \in d$  **do**

$n_z^w \leftarrow n_z^w + N_d^w$

# Influence of Parameters

## Alpha

GSDMM will get larger slightly with the increase of  $\alpha$ , and GSDMM will result in more clusters with only one document.

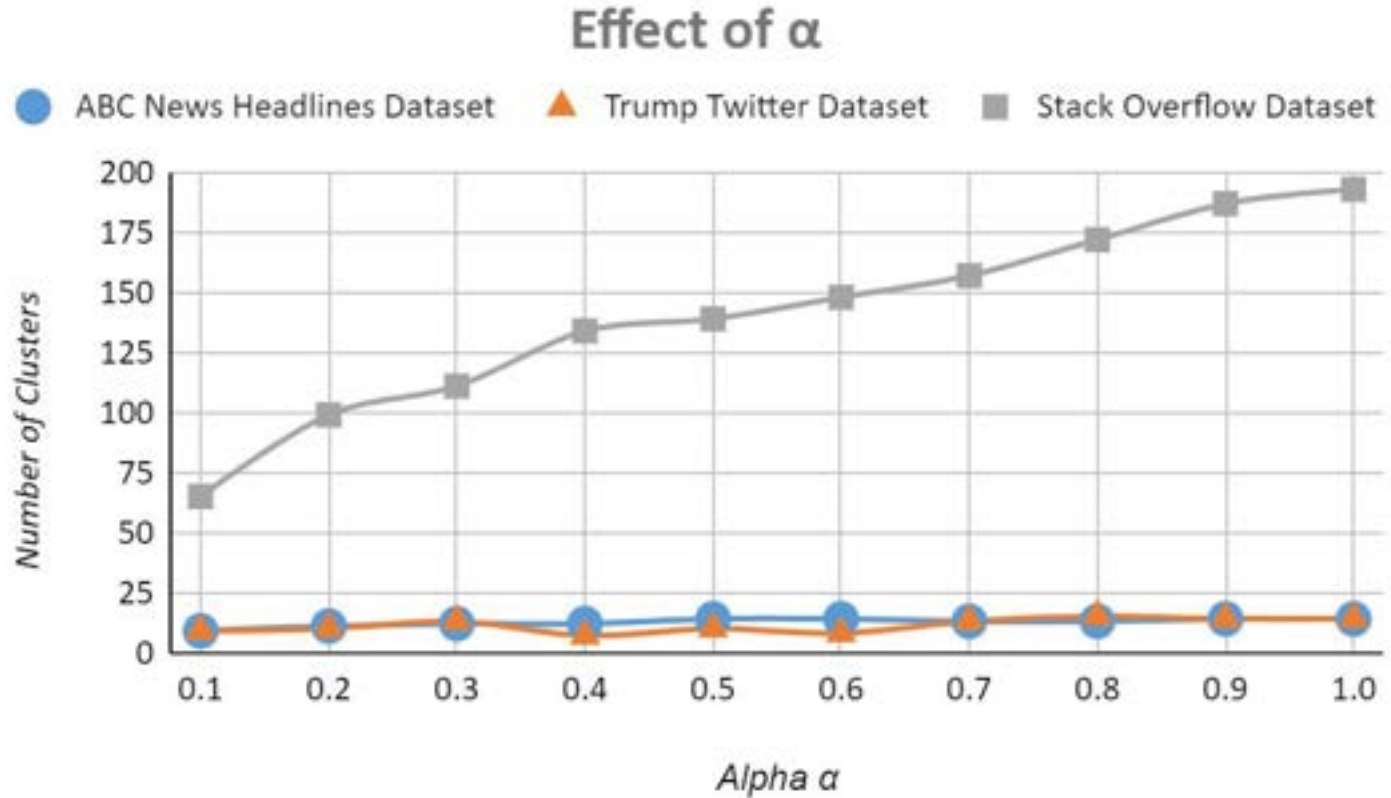


## Beta

GSDMM gives more emphasis on the similarity of words when  $\beta$  is small, and the words will have a larger probability to get into a particular cluster

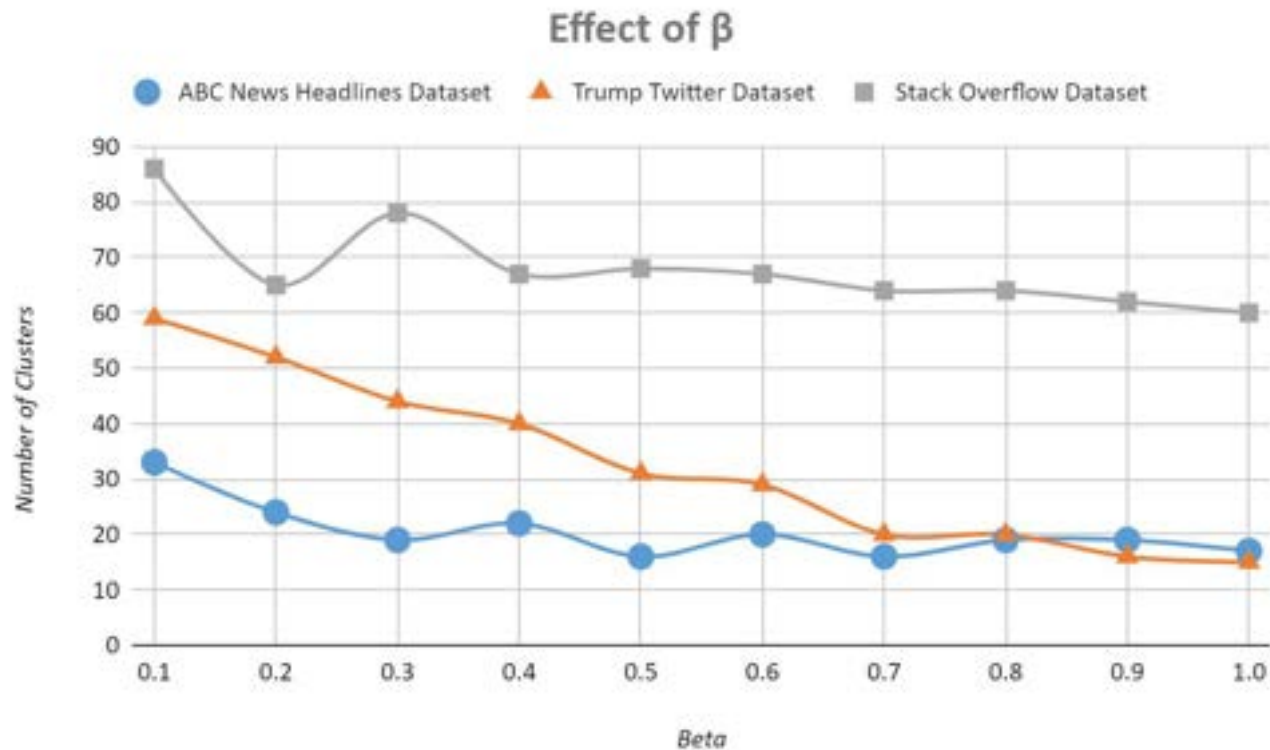
# Influence of Parameters

## Effect of $\alpha$



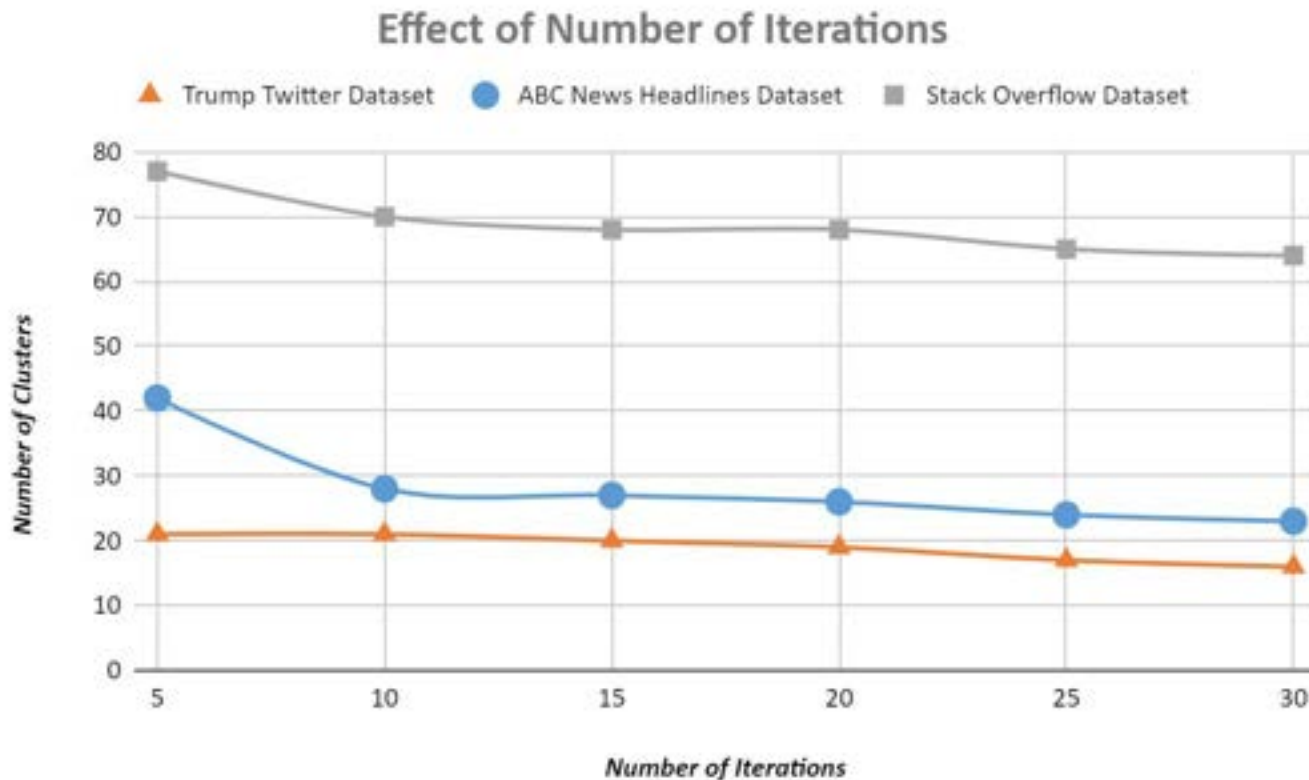
# Influence of Parameters

## Effect of $\beta$



# Influence of Parameters

## Effect of Number of Iterations



# FRONTEND INTERFACE FOR GSDMM

Check out GSDMM Yourself! 

Which dataset would you like to use?

Trump Tweets

Set alpha  $\alpha$

0.10 1.00

0.10 1.00

Set beta  $\beta$

0.10 1.00

1.00 1.00

Number of Iterations: 30 Adjust k: 15

Run GSDMM!

## RESULTS - Topic Word Clouds



Topic #0



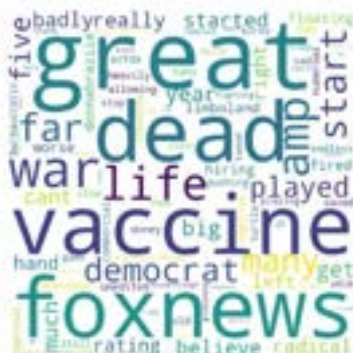
### Topic #1



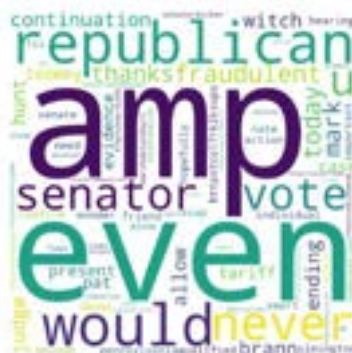
## Topic #2



### Topic #3



#### Topic #4

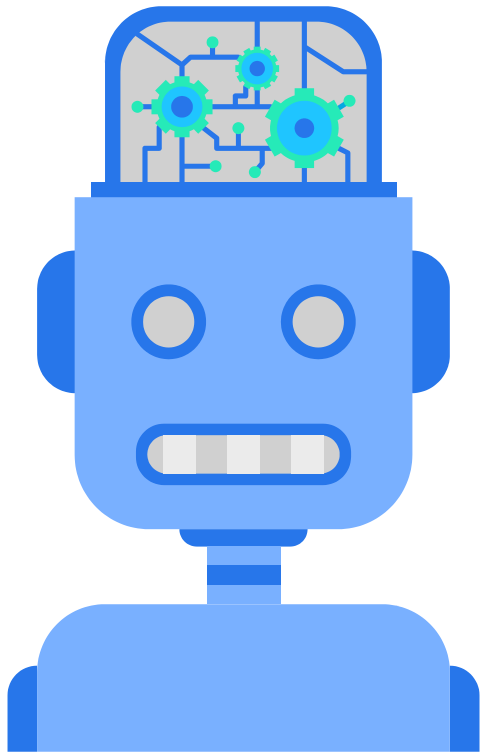


Topic #5



# RESULTS - Topic Assigned to each Data Sample

	text	topic	stems
0	.@senatmajdr and Republican Senators have to get tougher, or you won't have a Republican Party a	Topic #0	senatmajdr republican senator get tougher wont republican party anymore
1	Well, at least she was happy when I pardoned Scooter Libby. We got a GREAT new Senator from her sta	Topic #0	well least happy pardoned scooter libby got great new senator state
2	The Russia Hoax becomes an even bigger lie! <a href="https://t.co/nbtflGg2Ew">https://t.co/nbtflGg2Ew</a>	Topic #1	russia hoax becomes even bigger lie httpsconbtfllgg2ew
3	Europe and other parts of the World being hit hard by the China Virus - Germany, France, Spain and Ita	Topic #2	europe part world hit hard china virus germany france spain italy pa
4	Moderna vaccine overwhelmingly approved. Distribution to start immediately.	Topic #1	moderna vaccine overwhelmingly approved distribution start immediately
5	I am very disappointed in the United States Supreme Court, and so is our great country!	Topic #0	disappointed united state supreme court great country
6	New Peter Strzok Texts Undermine Official Narrative on Start of 'Russia Collusion' Investigation <a href="https://t.co/TrBRmBimTv">https://t.co/TrBRmBimTv</a>	Topic #1	new peter strzok text undermine official narrative start russia collusion
7	We won Wisconsin big. They rigged the vote! <a href="https://t.co/TrBRmBimTv">https://t.co/TrBRmBimTv</a>	Topic #0	wisconsin big rigged vote httpstcotbrmbimtv
8	Thank you! <a href="https://t.co/Du7bro3qls">https://t.co/Du7bro3qls</a>	Topic #1	thank httpstcodu7bro3qls
9	Tommy will be more popular than ever before - a hero! <a href="https://t.co/dTAXJyENlr">https://t.co/dTAXJyENlr</a>	Topic #1	tommy popular ever hero httpstcodtaxjyenlr
10	That's because he is a great champion and man of courage. More Republican Senators should follow h	Topic #0	thats great champion man courage republican senator follow lead landslit
11	Democrats would never put up with a Presidential Election stolen by the Republicans!	Topic #0	democrat would never put presidential election stolen republican
12	Just released data shows many thousands of noncitizens voted in Nevada. They are totally ineligible to	Topic #0	released data show many thousand noncitizen voted nevada totally inelig
13	I will Veto the Defense Bill, which will make China very unhappy. They love it. Must have Section 230 te	Topic #2	veto defense bill make china unhappy love must section 230 terminatio
14	Crazy! <a href="https://t.co/rO3EkzQvN8">https://t.co/rO3EkzQvN8</a>	Topic #0	crazy httpstcoro3ekzqvN8
15	I have NOTHING to do with the potential prosecution of Hunter Biden, or the Biden family. It is just ma	Topic #0	nothing potential prosecution hunter biden biden family fake news actual



04



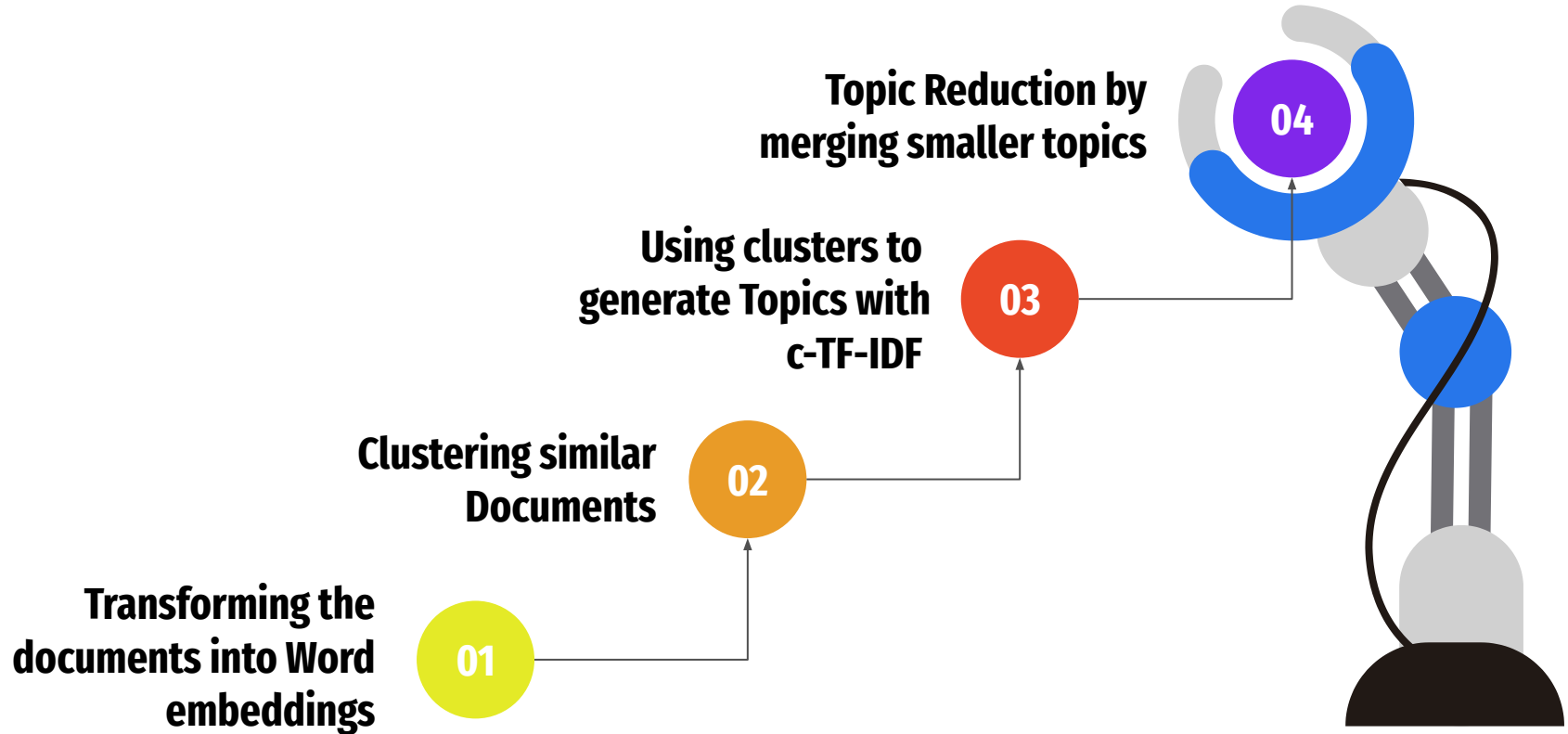
**BERTopic**

# BERTopic



- BERTopic is a topic modeling technique that leverages BERT embeddings and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions.
- The two greatest advantages to BERTopic are arguably its straight forward out-of-the-box usability and its novel interactive visualization methods.
- Having an overall picture of the topics that have been learned by the model allows us to generate an internal perception of the model's quality and the most notable themes encapsulated in the corpus.

# Stages in Topic Modelling with BERTopic



# 01

## Transforming the data into Word embeddings

---

- In the very first step we convert the documents to numerical data. We use [BERT](#) for this purpose as it extracts different embeddings based on the context of the word.
- The sentence-transformers package will be used here as the resulting embeddings have shown to be of high quality and typically work quite well for document-level embeddings.
- Sentence-transformers transform the data ie. your documents into 378-dimensional vectors.

## 02.A

# Clustering similar Documents - Reducing the Dimensionality

---

- Here we cluster together documents which have similar topics in a way that enables us to find similar cluster in each of the clusters.
- It will be good to reduce the dimensionality of the embedding as most clustering algorithms perform poorly with data of high dimensionality.
- UMAP(Uniform Manifold Approximation and Projection) is used here as it is good in preserving the original structure of the document after reducing the dimensionality.
- 15 would be a good size for each cluster (local neighbourhood) as it is a good balance between performance and the quality of clusters generated.
- The dimensionality of the documents will be reduced to 5 using UAMP with the parameter described above.

## 02.B Clustering similar Documents - Clustering the documents

---

- HDBSCAN is used to cluster similar documents
- HDBSCAN is a density-based algorithm that works quite well with UMAP since UMAP maintains a lot of local structure even in lower-dimensional space.
- Now the similar documents will be clustered together and these clusters represent the topics that they consist of.
- HDBSCAN does not force data points to clusters as it considers them outliers.

## 03 Using clusters to generate Topics with c-TF-IDF

---

- We want to know from the clusters generated what makes one cluster based on their content different from another.
- All documents in a single category (i.e a cluster) are treated as a single document and then TF-IDF is applied resulting in a very long document per category and the TF-IDF score here would demonstrate the important words in a topic.
- Then class-based TF-IDF is applied:

$$c - TF - IDF_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_j^n t_j}$$



## 03

# Using clusters to generate Topics with c-TF-IDF

---

- We get a single importance value for each word in a cluster which can be used to create the topic.
- We can take (for example top 10) most important words in each cluster and hence get a good representation of a cluster and thereby a topic.
- The great thing about HDBSCAN is that not all documents are forced towards a certain cluster. If no cluster could be found, then it is simply an outlier.

## 03

# Using clusters to generate Topics with c-TF-IDF

- For example we can see the largest clusters here as well as the words belonging to those topics which nicely seem to represent the interpretable topics.

```
top_n_words[7][:10]
```

```
[('game', 0.010457064574205876),  
 ('team', 0.009330623698817741),  
 ('hockey', 0.008341477022610098),  
 ('games', 0.006831457696895118),  
 ('players', 0.006753830927421891),  
 ('play', 0.006293209317999615),  
 ('season', 0.006227752030983029),  
 ('baseball', 0.0060984850344868195),  
 ('year', 0.005789161738305711),  
 ('nhl', 0.005736180378958607)]
```

```
top_n_words[12][:10]
```

```
[('nasa', 0.019843378603487997),  
 ('space', 0.019422587182152878),  
 ('gov', 0.01211931116301047),  
 ('henry', 0.00885814675356012),  
 ('launch', 0.008563915961385683),  
 ('orbit', 0.00826107575349062),  
 ('moon', 0.007999346663885938),  
 ('earth', 0.007568500945765267),  
 ('shuttle', 0.0075630804907155895),  
 ('jpl', 0.007471242802444713)]
```

```
top_n_words[43][:10]
```

```
[('dos', 0.014029062524679042),  
 ('windows', 0.010633883955998472),  
 ('problem', 0.007022143162108724),  
 ('help', 0.0052383622429981215),  
 ('disk', 0.005146911575725927),  
 ('thanks', 0.005086509908619605),  
 ('file', 0.0049983442954192),  
 ('program', 0.004945085186682861),  
 ('pc', 0.004936689541825903),  
 ('files', 0.004886658254378234)]
```

```
top_n_words[41][:10]
```

```
[('jesus', 0.017941948810926055),  
 ('god', 0.017222920386950193),  
 ('church', 0.011782741914097233),  
 ('christian', 0.011198341988130087),  
 ('christians', 0.010921245789061267),  
 ('christ', 0.010734557826855092),  
 ('bible', 0.010671425554580173),  
 ('faith', 0.010616988475347046),  
 ('sin', 0.007795019887970474),  
 ('christianity', 0.007527424973714497)]
```

## 04

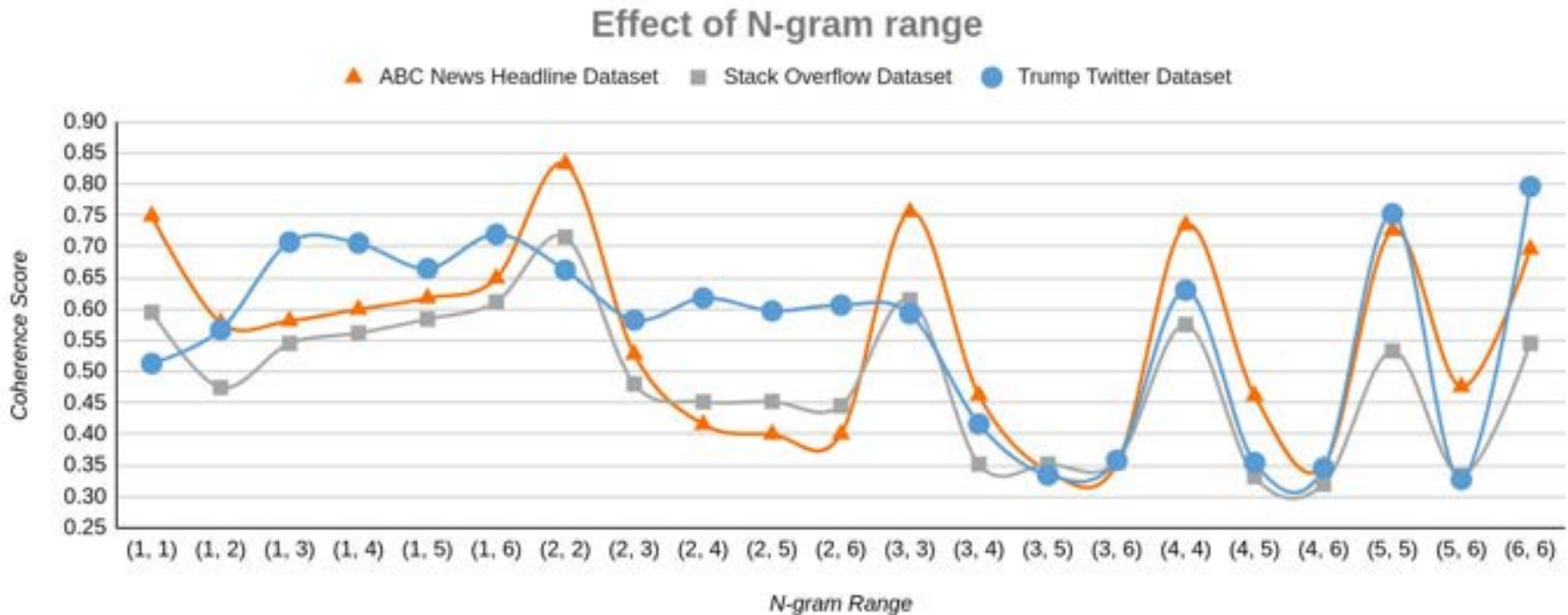
# Topic Reduction by merging smaller topics

---

- The parameters of HDBSCAN can be tweaked such that fewer topics are obtained but it does not allow you to specify the exact number of clusters.
- For this we have compared the c-TF-IDF vectors among topics, merge the most similar ones and finally re-calculate the c-TF-IDF vectors to update the representation of our topics(this can be re-iterated several times as needed).

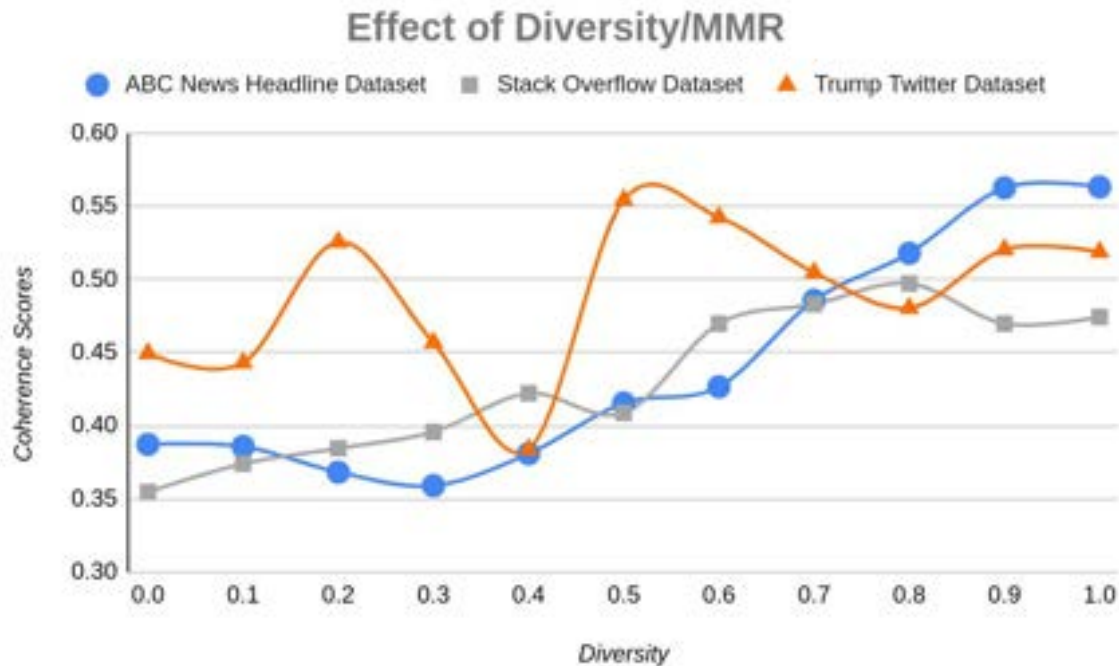
# Influence of Parameters

## Effect of the N-gram range



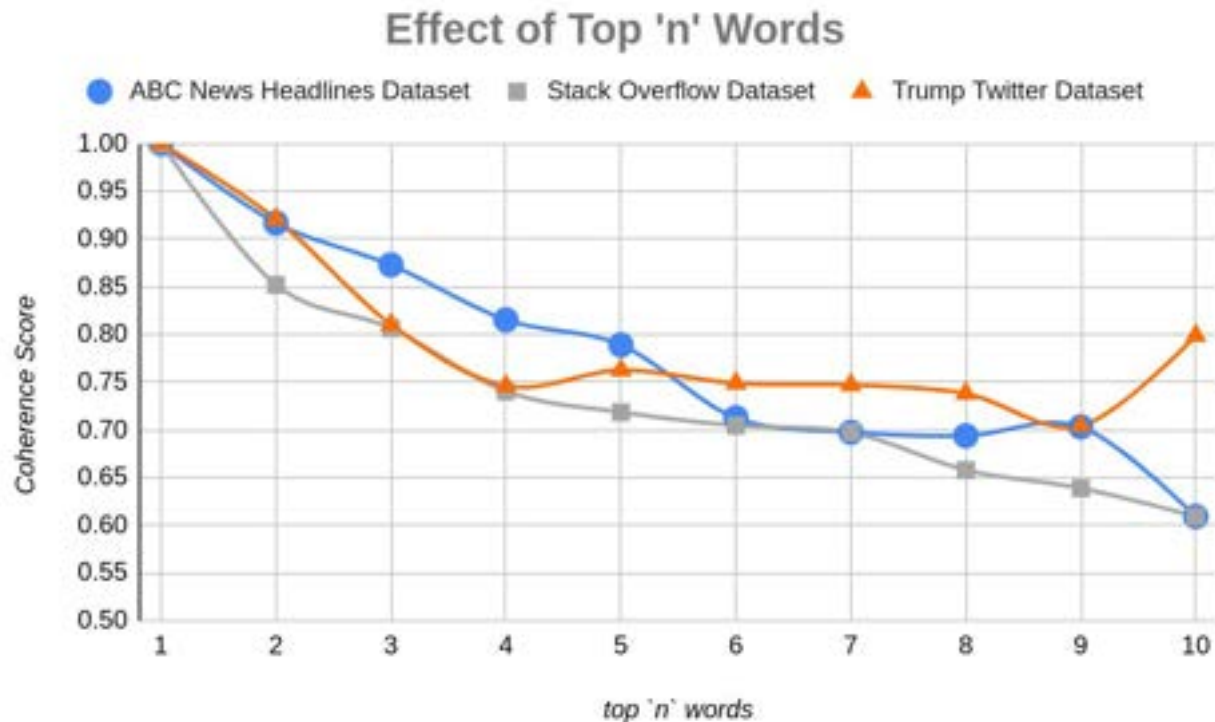
# Influence of Parameters

## Effect of Diversity/MMR



# Influence of Parameters

## Effect of the top 'N' words



# FRONTEND INTERFACE FOR BERTopic

Check out BERTopic Yourself! 

Which dataset would you like to use?

Trump Tweets

# of topics:

20

20 100

Minimum size of each topic cluster:

10

1 40

How diverse should the topics be? (0 for no diversity and 1 for maximum diversity):

0.50

0.00 1.00


Minimum Ngram

2

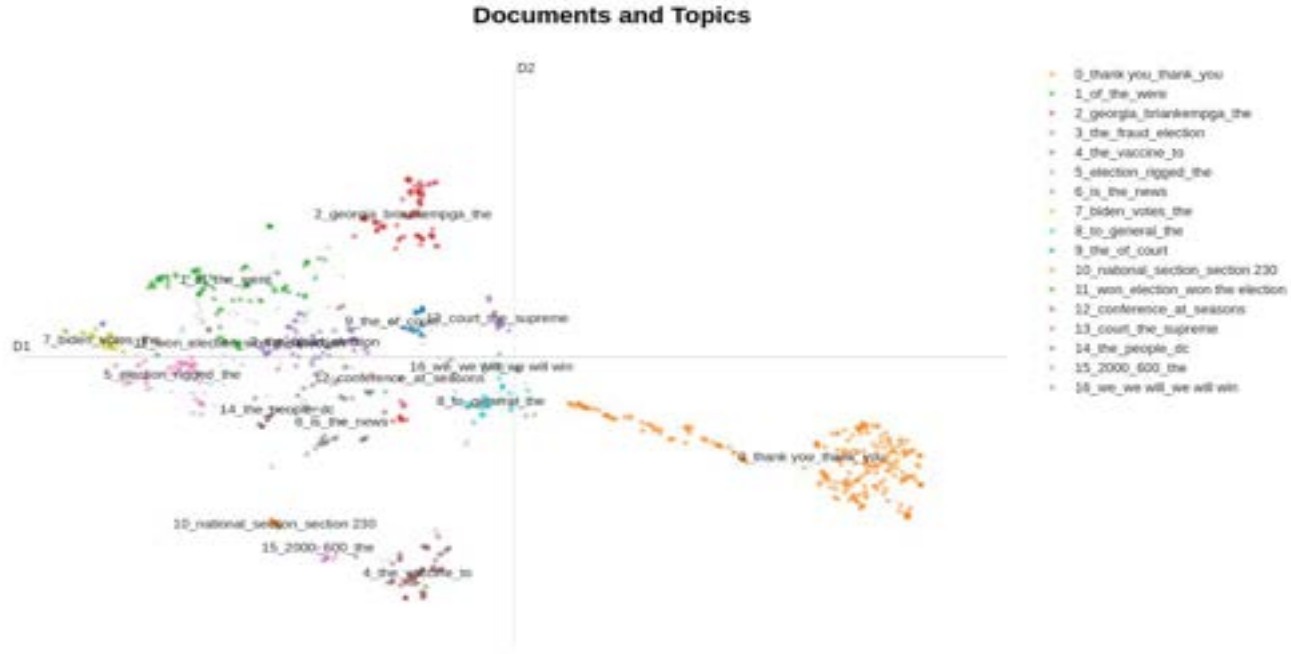
Maximum Ngram

3

BERTopic

 Run BERTopic!

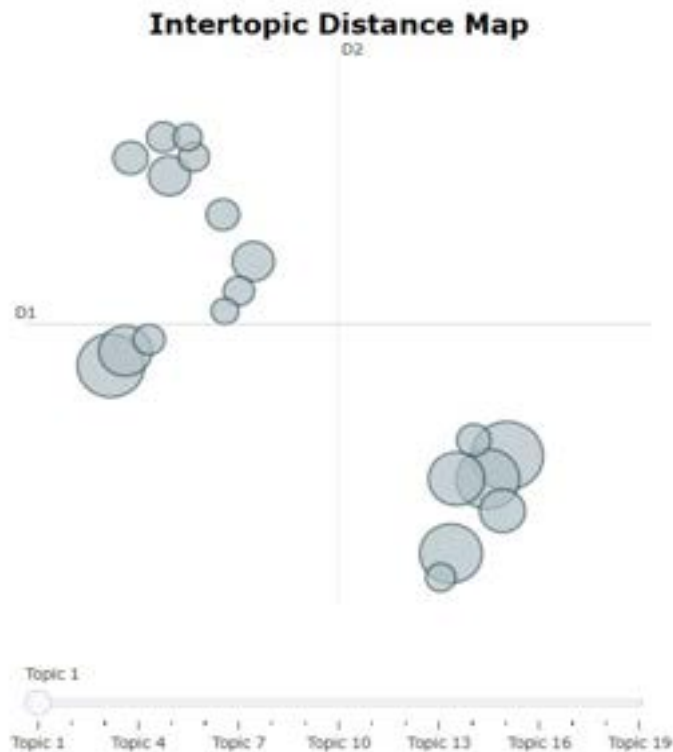
# RESULTS - Document Visualization





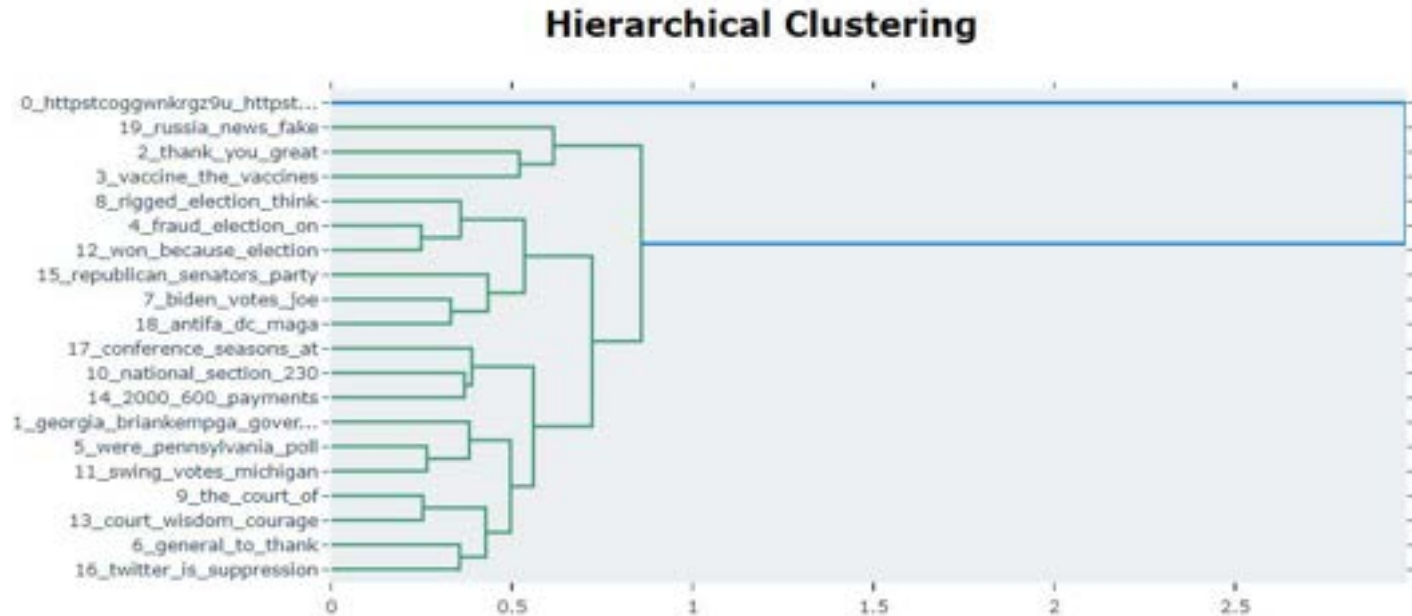
# RESULTS - Coherence Graph

---



# RESULTS - Hierarchical Clustering

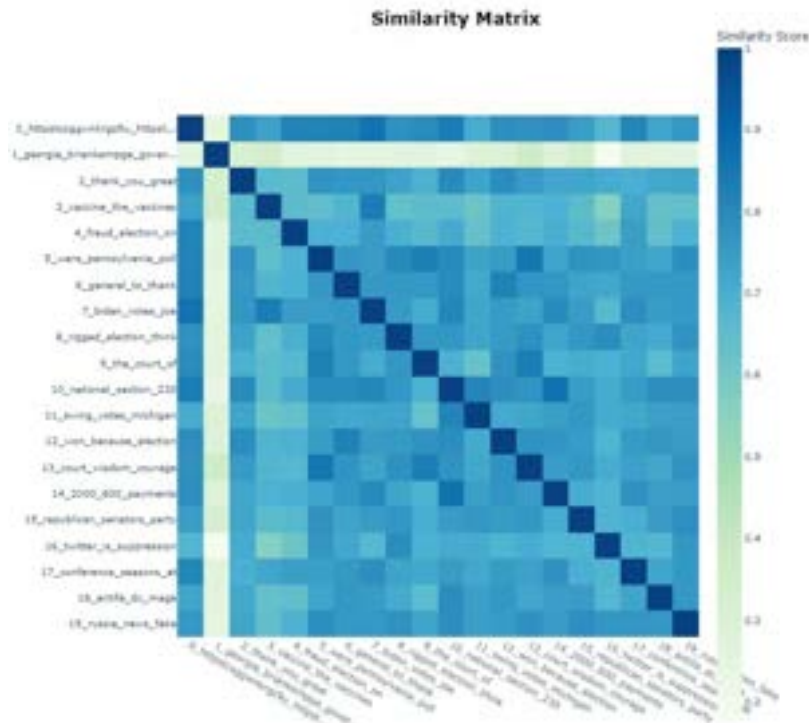
---

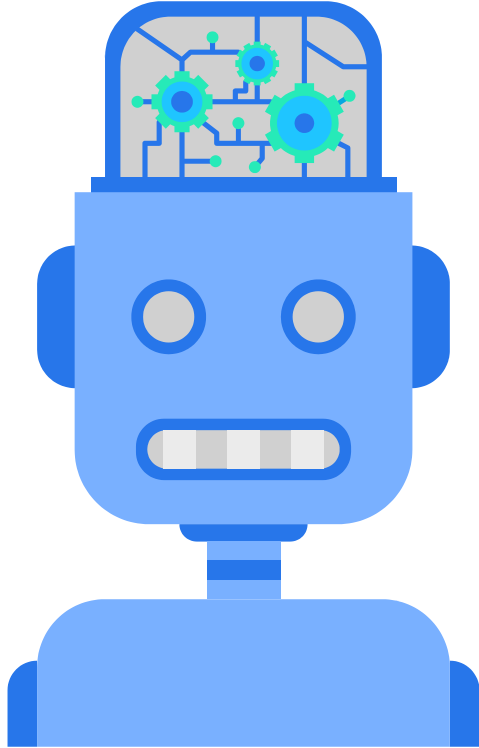


# RESULTS - Topic Word Scores



# RESULTS - Similarity Matrix





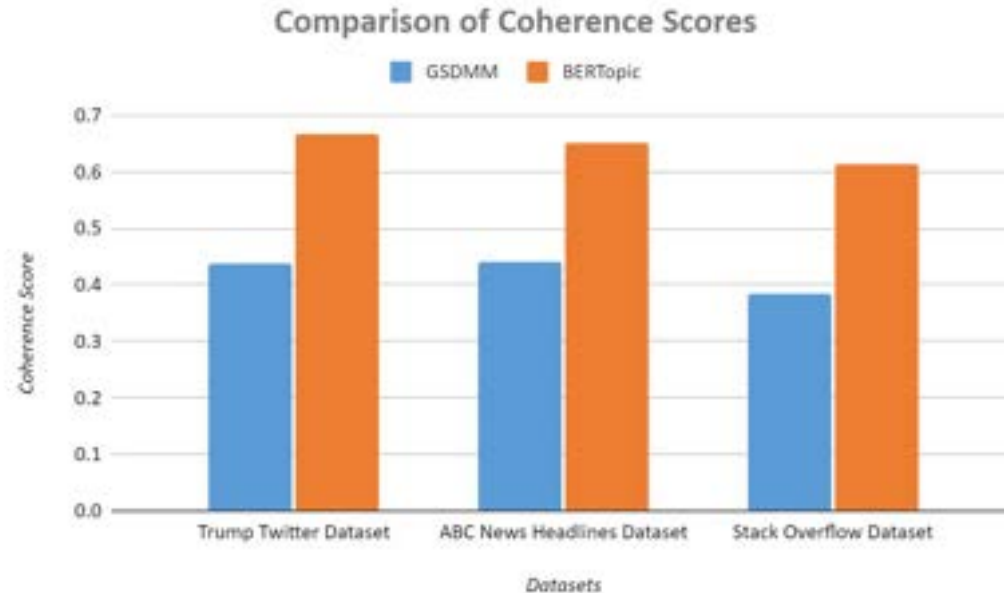
05



# **Results & Conclusion**

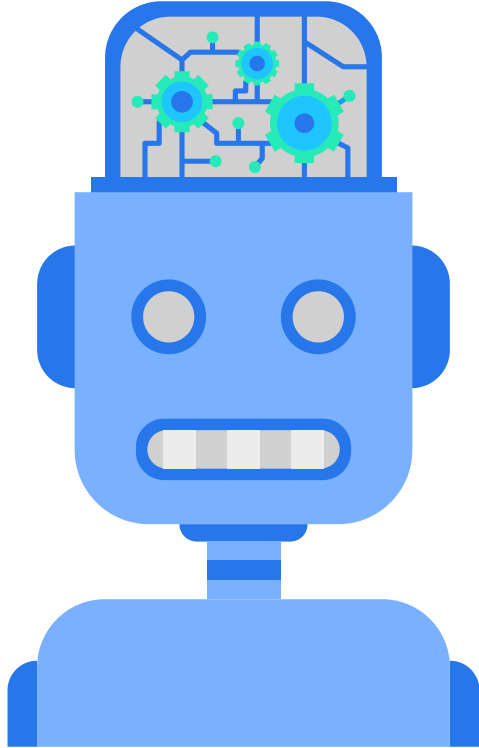
# COMPARISON - Coherence Score ( c\_v )

Algorithms	Trump Twitter Dataset	ABC News Headlines Dataset	Stack Overflow Dataset
GSDMM	0.4374	0.4404	0.3843
BERTopic	0.6649	0.6493	0.6119



# CONCLUSION

- From the results of testing the two algorithms on the three chosen datasets we can conclude that that BERTopic, which is a neural topic model, performs better than GSDMM which is statistical model.
- BERTopic uses a pre-trained BERT based model which captures the semantics and the context of the text.
- With this information BERTopic is able cluster similar topics effective giving more coherent topics with better contextual semantics.
- GSDMM is not able to use this contextual semantics because of which the coherence scores are not as good as those for BERTopic



# **THANK YOU**

**Hope you liked our presentation!**