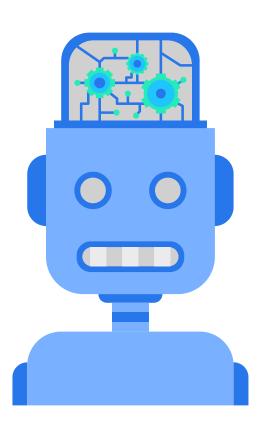


AN EXPLORATORY ANALYSIS OF GSDMM & BERTOPIC ON **SHORT TEXT TOPIC MODELLING**



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.

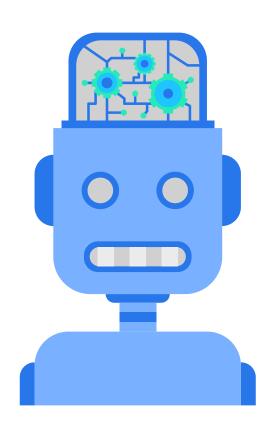
LEARNING OUTCOMES

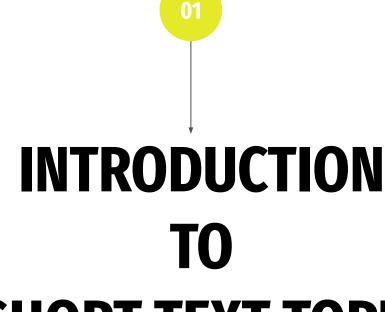


Learning Outcomes

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

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





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

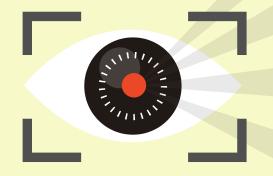
Topic modeling is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents.

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

Topic Modelling Applications



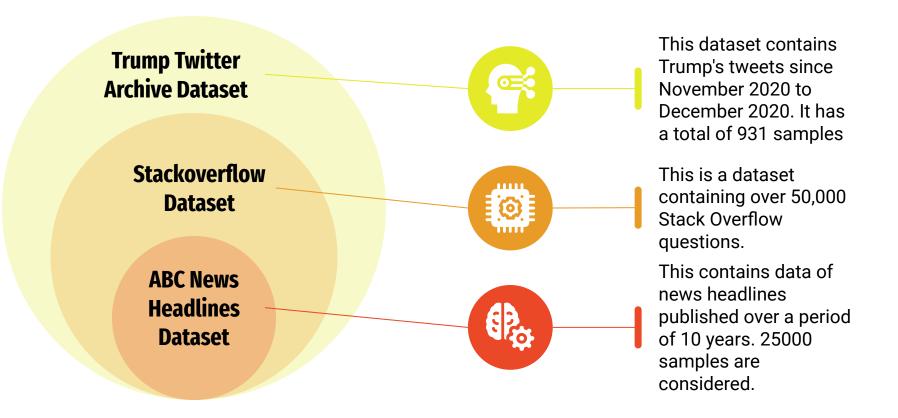
Text summarisation

Sentiment analysis

Recommender systems

Understanding scientific publications

DATASETS



DATASETS - Trump Twitter Archive Dataset

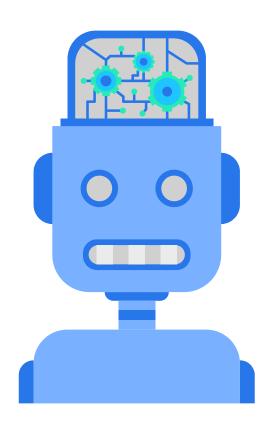
```
* 8 : {
   "date": 1688388872888
   "favorites": "212444"
   "id": "1339937891787351846"
   "isRetweet" : false
   "retweets": "54898"
   "text" :
   ".@senatemajldr and Republican Senators have to get tougher, or you
   won't have a Republican Party anymore. We won the Presidential
   Election, by a lot. FIGHT FOR IT. Don't let them take it away!"
*1:{
   "date": 1688299181888
   "favorites": "77814"
   "1d": "1339929663884747778"
   "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/tT4gdn8f9V"
```

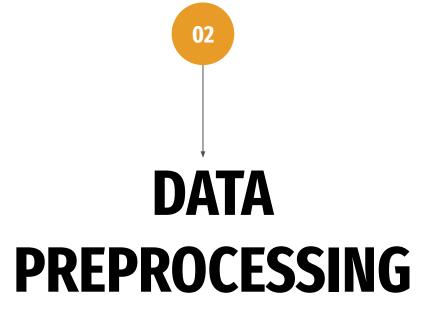
DATASETS - ABC News Headlines Dataset

	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 buries united states in hills

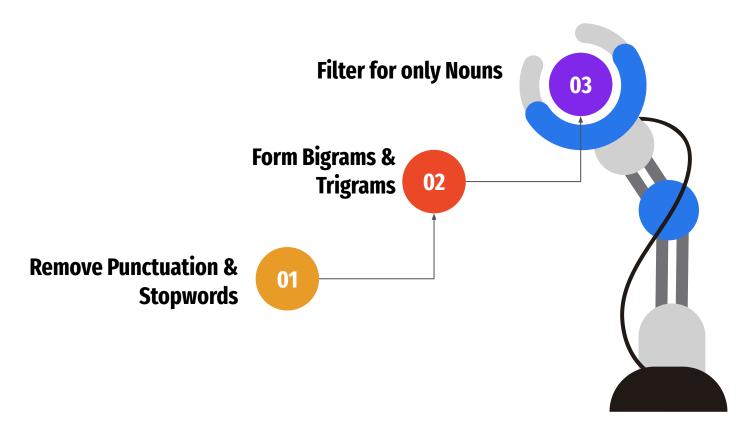
DATASETS - Stackoverflow Dataset

1		10	190	M N	1	T	month of	-
Ú.		HICHH	Janus Reposit Saids Every Revolute, Sacondo	rymin arealy familie with rejecting total every necessarily using transationer and seasotton in great-respect-	3016 (0.00) (0.00)	of costs	+984	-91
	. 1	(40,000)	My ex pre Sphrody Hondalini	rpmit the trunderbank any one & (plannin) were despited to be introduble. In Popular Trimed spit representational	2010-02-02-08/00/20	70	-gar	1941
	+	(40)11114	Text Sharlay Image with Zerkened Spacity React halons.	spring alternatives consist a title one to seeke will be undergaptered right post about the consistence and consistence of consistence and consistence are consistence and con	2014/01/01/02/4/24	HQ	1961	-
	+	3603338	why tentary operator to beth it on policy?	species place and a significant particular for the assumption of the source properties of the companion of the source properties of the source p	2014 EC 00100001	HQ	+941	1994
	à	HATELET	hite show fall with scale arenessin	rprim emplateur flestrigetamens, med te implementate annæten ar sheutlitt henvite med solminetral design-flustrig action butten	2010-12-10-10-20-00	10	Help	1994
		sensor	Asserting pointer manager of the structure using structure pointer	ryn han defred insge printe leute de strutue Anti-met to oar that number petter uning the independent extraor	369 (3-0) 9630.37	40,0,000	e/gajor	-
0		365428	now in the data for the constant with Lamberg standard explication of the charge algorithm of the standard explication of the constant of th	HARTE BUILDING and Sensing Art setunday, Sunday And Public Schooly Selection graphery (wheeler - Special property support of the Control of Sensing Se	2010-01-02-0033-08	10,000	+9,60	100
		latinati)	Annumy continues in Economic	convenes to the transport of the contract of t	2019-01-01-08-02-05	16,100	Adjust	15584
		leparti	Draiging There in Motions 18 std Fings moralisally	$spress_{i} \ and \ a things there are grade this, he probabilities will be not them. \\ Laber 1 and 2 and 2 are substantially a substantial beautiful and a substantial b$	Anal - 2020 to 00 00 00 00	NO.	+100+	10041
		345948	Tactifics Value (Inappears VBI)	Impose Lemberry techniques from Land Form 3, Form 3 having techniques and Form 3 is having so, notify	3856-61-81-8627-64	10,100	*SAIT	1994
	17	Jerrer I	Managerith finding to lited - *** planting after basis (1) Salare	Aprilian net to blumo (Linux Str). I ensurated protein with darking transport service i wall to it. Altrum mungation advanturationary to be related to	2010/02/03 12:08:03	HQ	1000	1000
	II	(480,004)	Miles the best way to move test shoul!	rgini are kinda new to hadrand direcognisers, hug i coded a fac of Chilling (APP), i consend amount app (). Handrood-Handrood Rise	3619-12-01-12-19(3)	10,000	+664	11521
	11	140000	to hipselve checks Squeel Label leave on quantity of bet using Storyton id.	Have a Union China to a property that can contain had all many different procedul languar. Have son in non-mise phase throughouse througho	2010/01/01 12:00:00	108,00	+984	1000
	12	jachniese.	mapping (ELE) and the primary (and ELE)	mystishee Missieratishe Antidessinate Antidessishe (Ascheung) spec	2010/03/05 14:20/07	10,100	+94+	1150.0
	14	(ACMIT)	Profrontal displaying serve of corresponds version L.E.C. family.)	ryonly profile van acting the after upstring to consposit various (A.E.Saria, Erlyon Spring transfer in Commissionality	2010/01/01 14:00:01	High	elpto-	-100
ř.	10	penent	marked black UNIO held sums	-pirm type; to see took to built reproducte in give and have to sell the Quak Deck and hope; they introduce house; house; house;	2014/07/02 18:00039	140	+10,010	1994
	it.	penery	$eq:control_ent$	random that for harmed brift to a maximal brief the trying to get to give with the non-stand library in an apart and included of the principles.	2010/01/05 17:00:01	HQ	+964	1000
	411	2010004	All display drive and spelate the code	typi am typig to daplig at application sention but to planets daplicy (busine comment falls with 1 (in transaction and our opportunity and on the comment and comm	mar 310-0-0 1000	10	1961	1924
į.	18	period	How to create a function that returns a tight array or third inclume project.	familiaring starting to have control programming, som programming and "British matric, an informational"	(000-01-01-01-01-01-01	ujjan	rigin	100
	39	(4004)8	Student jest Student com je what is the difference!	riprofest to the officerus between the brings and the books to a gift for right the street of the books and the books and the books are the street of the books and the books are the bo	(819-1)-01-10-00(1)	HO	+latel	regain
	33	senni	Charles I temporary directory to Force Well?	rprotech filtputting act return 5. hyperplaggic in the 2003 BCD issue amount filter in second information place in re-	2010/06/05 (00/08/05	HQ	+96+	1000
	11	HERMIN	Note to rights have a Mala integraping (#	type are deproying a few young with a CA application, And I want to make a can that few wants as it was inspections	(010-01-01-101000)	10,000	1961	-64
	- 12	leme?	Stapman, N. and Reading (Securit Local property	The senting on my multi-friending parametrization, only numbers. It must also the fore-must be an implication or improvement the meaning of the control of t	369 (1-01) (4.20) (37)	Marie	man .	mari.
	-11	Seminer.	Next and abotiment	Travelling a shorter to contain all notification when the armount has blade in success about that the indiscipa-	2020 01 01 21 10 21	10,697	1961	1991
ë.	36	positive	Car I (time from than mild to bell) additionated strong basing from the?	rapril the my class remnistry larger to berth to those an error hamaling gave wrong with beating a resoft in	2010/01/01/20 02:04	70	1980	mga+
	21	1000001	CF - Court a specific word or not feedback and word the result to late C	rprinted see, fills question a cropa, has couldn't feel the answer that seed complex ripring in more	attempt press	щран	riar	1924
	29	SHILLI	red cell (seen highe filaming eyes harnor with higher som	operated to make a bagger fricting betwee busines (MM), more than the struct to take the bagger size in reconstructioning action numbers	protein symmet	HQ	1967	1000
	47	240004	Drovid i con the host multiter when discloring case classed.	rp-Augusting to 14 hosh "Nigor, "giftud com (authorise) has been received war terminant an abdition received and processes also an authorise the	\$10 to \$2,00 07/00	H0	ristor	1981
	29	betrake	I result to create an array with numbers given by the user	Systemate printer Preserve a coefficients of a party consider's server's, storig coefficients in its representative server.	2021-02-02/02/04/04	100,300	+96+	1991
	-	Service:	Too its make the Notice complet that everyope as error?	NOT have a furtir project where It file to have furtire carring medical assertion, troo part do that NoT Augitim	2014-01-02-0100-04	lest.	1984	1962





Steps Involved in Data Preprocessing



Example for Data Preprocessing

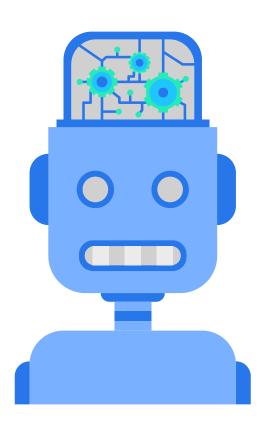
How do we run "Streamlit App" on our computer?

how do we run streamlit app on our computer

how do we run streamlit_app on our_computer

1

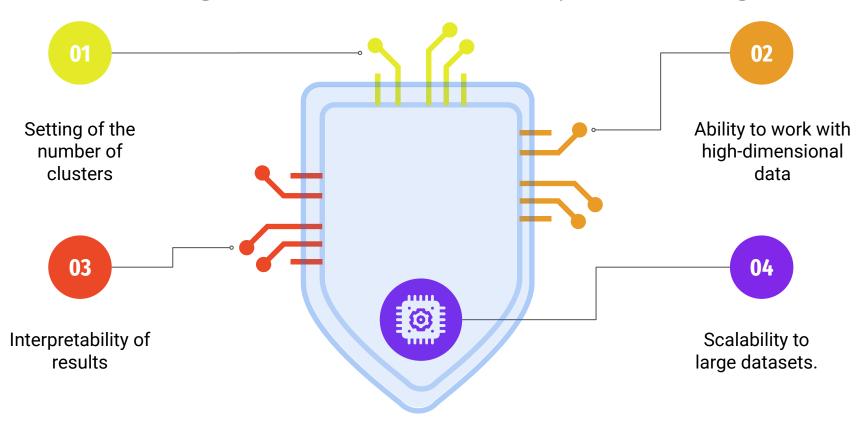
streamlit_app, our_computer





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} .

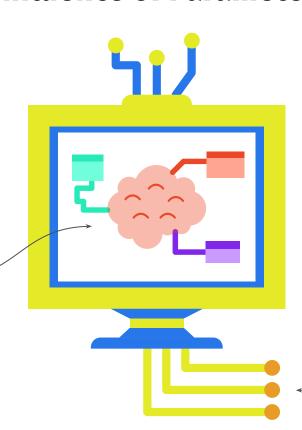
begin

```
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:
  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
```

Influence of Parameters

Alpha

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

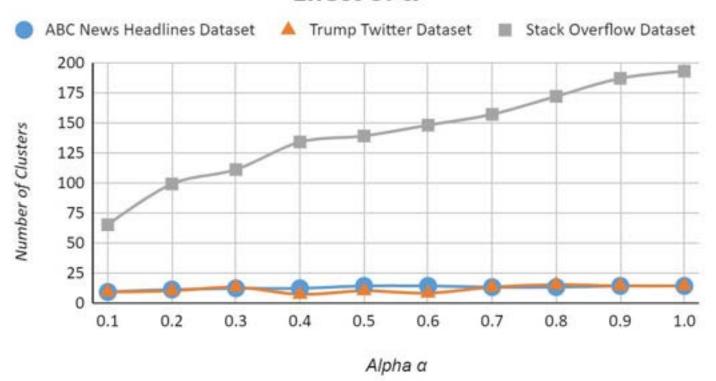


Beta

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

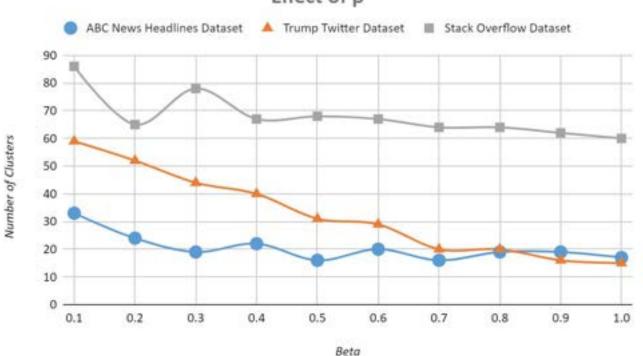
Influence of Parameters Effect of α

Effect of a



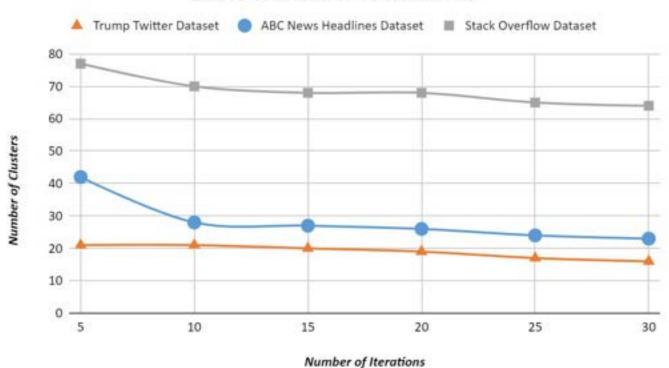
Influence of Parameters Effect of β

Effect of β



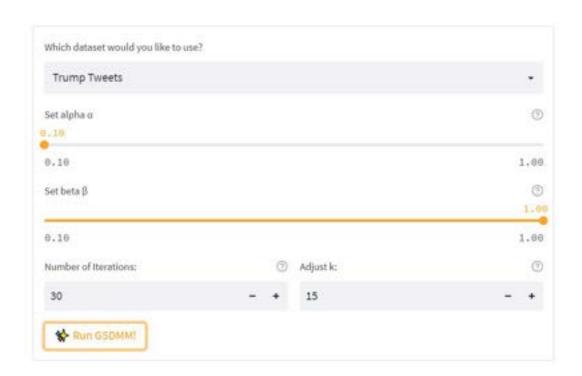
Influence of Parameters Effect of Number of Iterations

Effect of Number of Iterations



FRONTEND INTERFACE FOR GSDMM

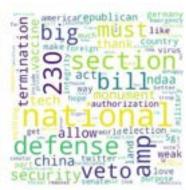
Check out GSDMM Yourself!



RESULTS - Topic Word Clouds





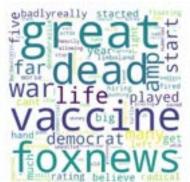


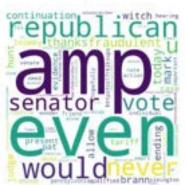
Topic #0

Topic #1

Topic #



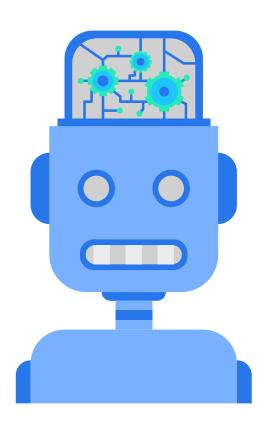




Topic #3

RESULTS - Topic Assigned to each Data Sample

	text	topic	stems
0	.@senatemajidr and Republican Senators have to get tougher, or you won't have a Republican Party at	Topic #0	senatemajldr republican senator get tougher wont republican party anymore
30	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 pandoned scooter libby got great new senator state.
2	The Russia Hoax becomes an even bigger lie! https://t.co/nbtflGg2Ew	Topic #1	russia hoax becomes even bigger lie httpstconbtfigg2ew
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 https://	Topic #1	new peter strzok text undermine official marrative start russia collusion
7	We won Wisconsin big. They rigged the vote! https://t.co/TRBRmBiMtv	Topic #0	wisconsin big rigged vote httpstcutrbrmbimty
8	Thank you! https://t.co/Du?bro3qis	Topic #1	thank httpstcodu?bro3qls
9.	Tommy will be more popular than ever before - a hero! https://t.co/dTAXJyENIr	Topic #1	tommy popular ever hern httpstcodtasjyenir
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 tandslit
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 $\frac{1}{2}$	Topic #0	released data show many thousand noncitizen voted nevada totally ineligi
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 termination
14	Crazyl https://t.co/rO3EkzQVN8	Topic #0	crazy https/coro3ekzqvn8
15	I have NOTHING to do with the notential prosecution of Hunter Riden, or the Riden family. It is just mo-	Tonic #0	nothing notestial prosecution hunter hiden biden family take news actual



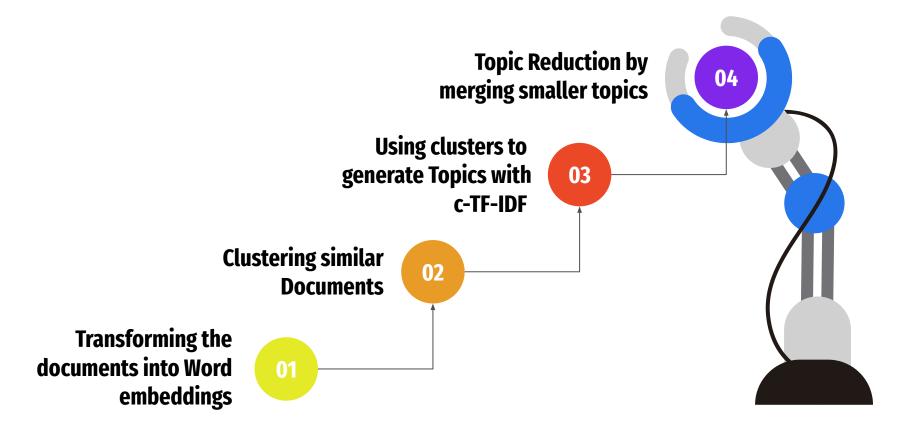


BERTopic



- → 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



Of the control of

- 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.

O2.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.

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:

$$c - TF - IDF_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_{j=1}^{n} t_j}$$

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.

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.

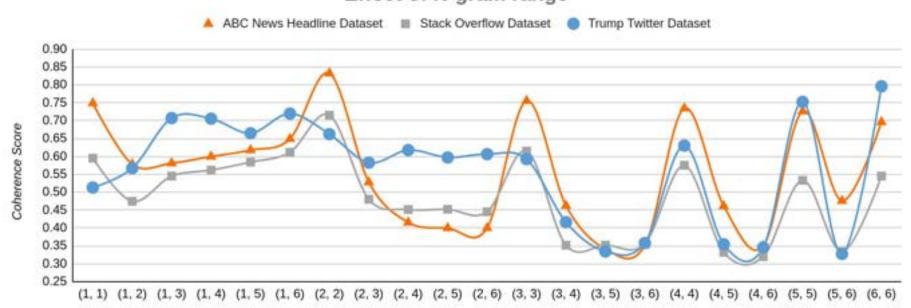
```
top_n_words[7][:10]
                                           top_n_words[43][:18]
[('game', 0.010457064574205876),
                                           [('dos', 0.014029062524679042),
  'team', 0.009330623698817741),
                                            ('windows', 0.010633883955998472),
 'hockey', 0.008341477022610098),
                                            ('problem', 0.007022143162108724),
  games', 0.006831457696895118),
                                            ('help', 0.0052383622429981215),
  'players', 0.006753830927421891),
                                             'disk', 0.005146911575725927),
  'play', 0.006293209317999615),
                                             'thanks', 0.005086509908619605),
 ('season', 0.006227752030983029),
                                             'file', 0.0049983442954192),
  'baseball', 0.0060984850344868195),
                                             'program', 0.004945085186682861),
 ('year', 0.005789161738305711),
                                            ('pc', 0.004936689541825903),
 ('nhl', 0.005736180378958607)]
                                            ('files', 0.004886658254378234)]
top_n_words[12][:10]
                                         top_n_words[41][:10]
[('nasa', 0.019843378603487997),
                                          [('jesus', 0.017941948810926055),
  'space', 0.019422587182152878),
                                            'god', 0.017222920386950193),
  'gov', 0.01211931116301047),
                                           'church', 0.011782741914097233),
 ('henry', 0.00885814675356012),
                                           'christian', 0.011198341988130087),
 ('launch', 0.008563915961385683),
                                            'christians', 0.010921245789061267),
 'orbit', 0.00826107575349062),
                                            'christ', 0.010734557826855092),
  'moon', 0.007999346663885938),
                                            'bible', 0.010671425554580173),
                                           'faith', 0.010616988475347046),
 ('earth', 0.007568500945765267),
                                           ('sin', 0.007795019887970474),
 ('shuttle', 0.0075630804907155895),
 ('jpl', 0.007471242802444713)]
                                           ('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).

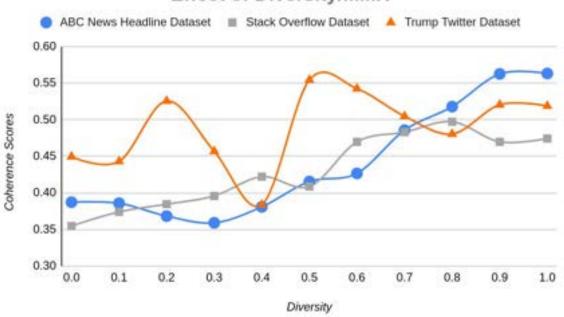
Influence of Parameters Effect of the N-gram range

Effect of N-gram range



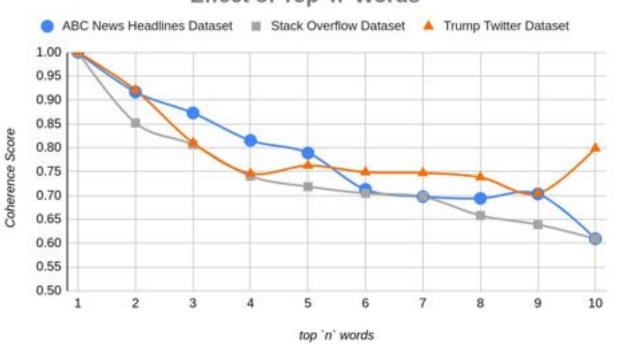
Influence of Parameters Effect of Diversity/MMR

Effect of Diversity/MMR



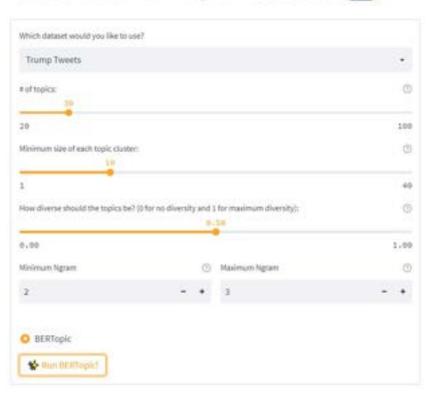
Influence of Parameters Effect of the top 'N' words

Effect of Top 'n' Words



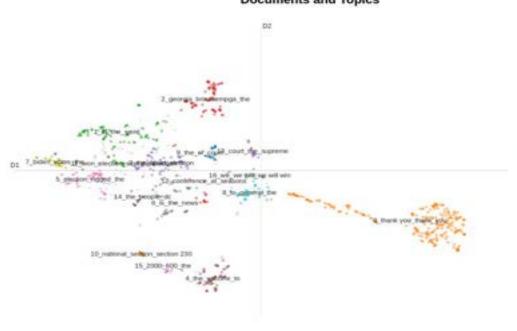
FRONTEND INTERFACE FOR BERTopic

Check out BERTopic Yourself!



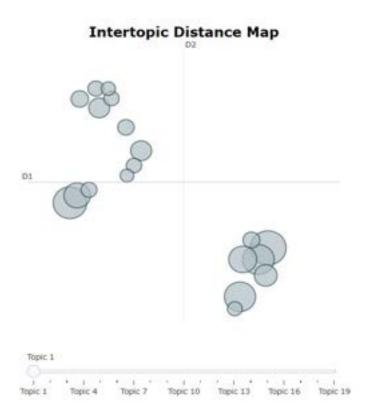
RESULTS - Document Visualization





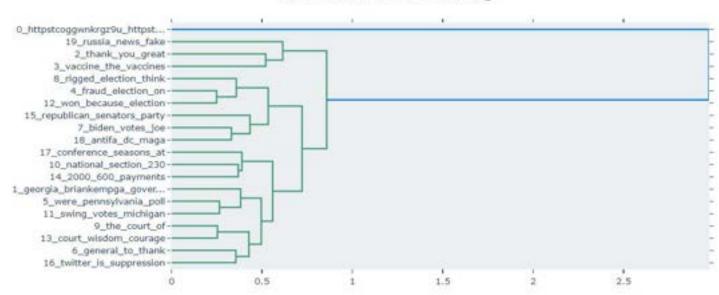
- 5 thank you thank you
- 1 of the were
- 7 georgia triantempga the
- 3 the fraud election
- 4 the vaccine to
- 5_election_rigged_the
 6 is the news
- 7 biden votes the
- 8 to general the
- . 9 the of court
- 10_national_section_section 230
- II won election won the election
- > 12 conference at seasons
- 13 court the supreme
- 14 the people do
- 15 2000 600 the
- 15 we we will we will win

RESULTS - Coherence Graph

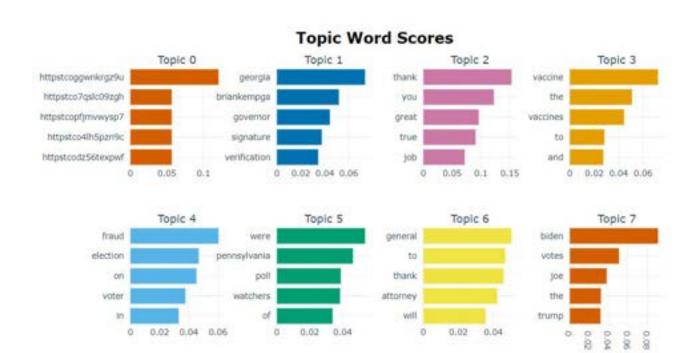


RESULTS - Hierarchical Clustering

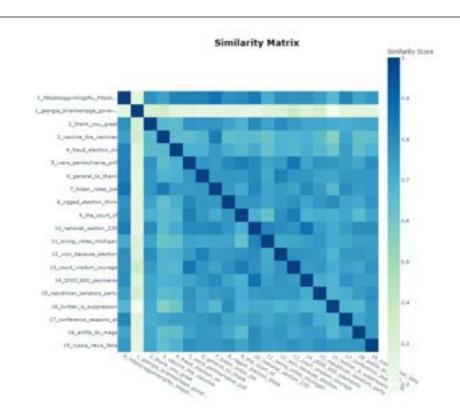


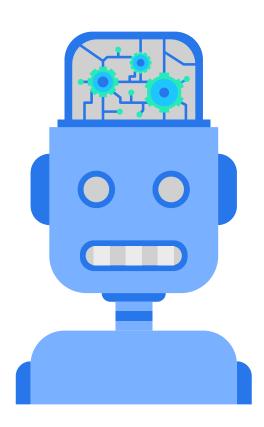


RESULTS - Topic Word Scores



RESULTS - Similarity Matrix



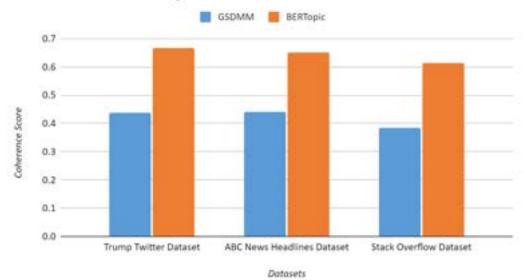




COMPARISON - Coherence Score (c_v)

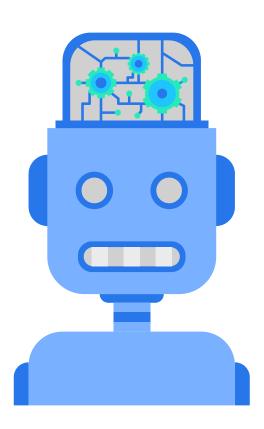
Algorithms	Trump Twitter Dataset	ABC News Headlines Dataset	Stack Overflow Dataset
GSDMM	0.4374	0.4404	0.3843
BERTopic	0.6649	0.6493	0.6119

Comparison of Coherence Scores



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!