Social Media Analysis

Ashish Kumar Jha

Agenda

Advanced text mining

- Ngrams
- Correlations

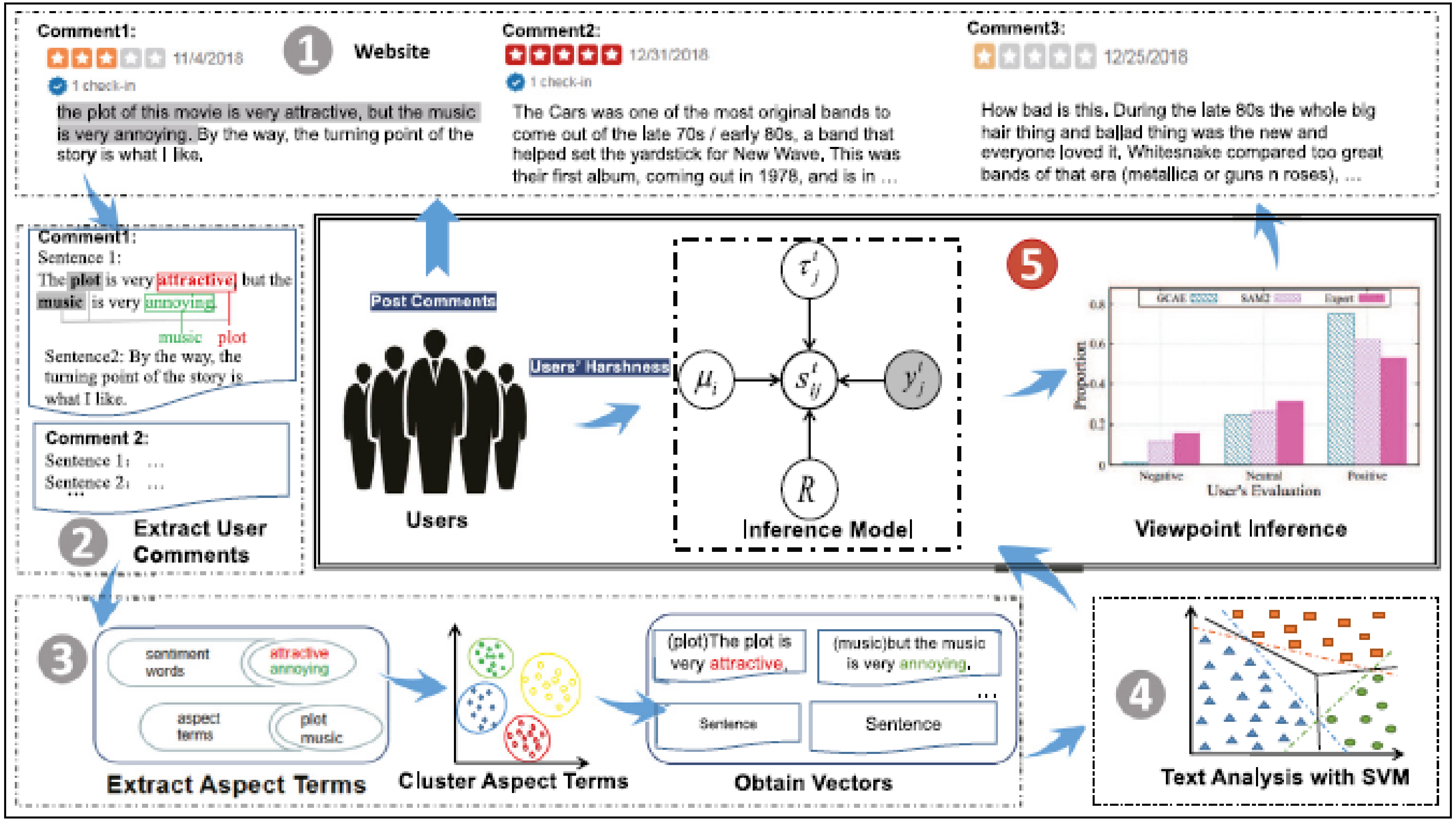
Text classification

Twitter Data

Topic modelling

- Corpus
- Text cleaning
- Topic modelling

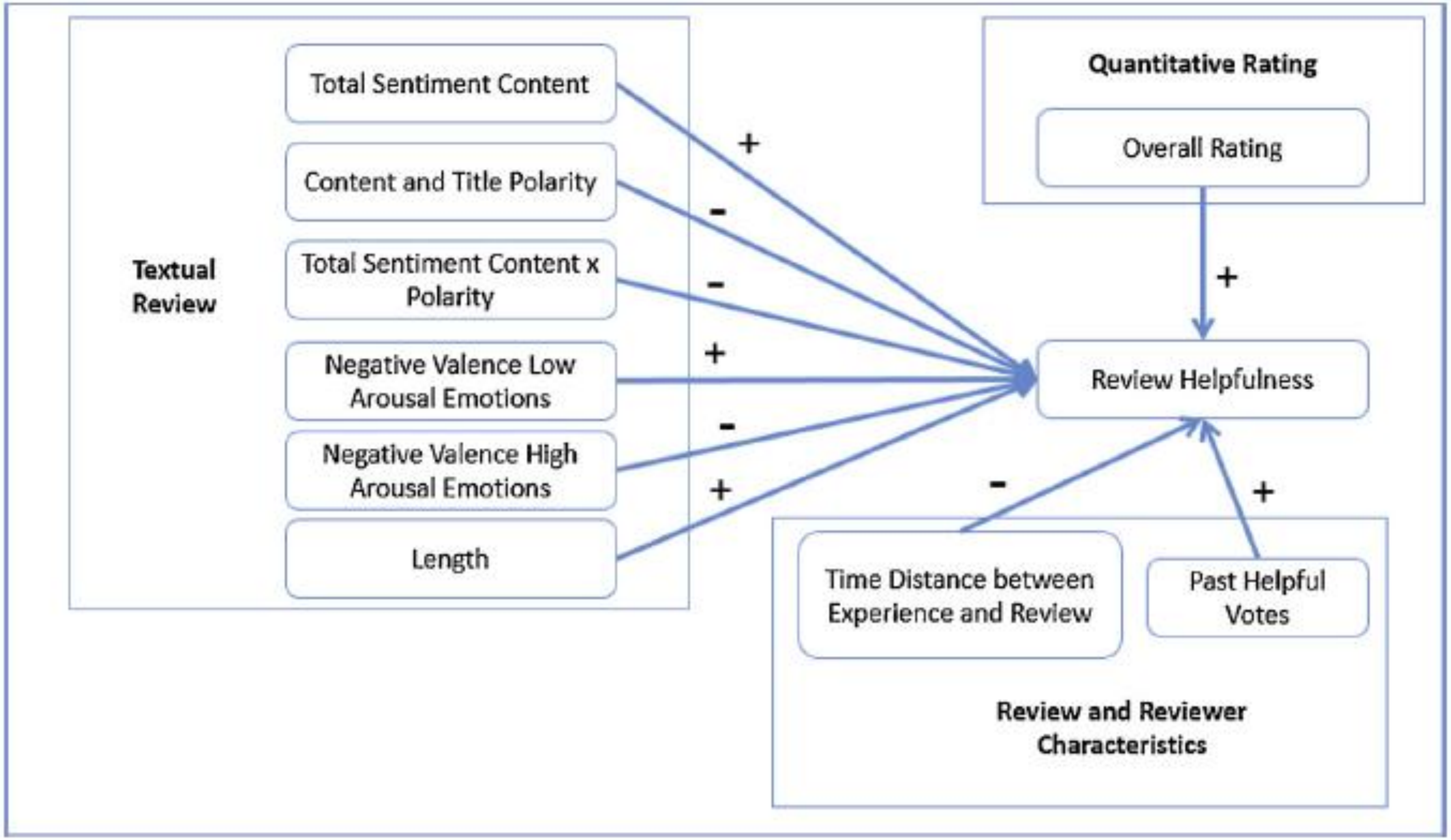
Harshness Aware



Challenges identified- Assignments

- One potential major issue that can be forecasted is reducing the influence of 'dissident' people that diverges from the majority opinion. Users who disagree with the majority are getting negative *i (harshness of user) (harshness parameter* μi *)* and in the algorithm assumes the probability of them telling the actuality is lowered, so their overall voice could be interpreted as neutral or sometimes positive.
- To obtain true product evaluation, this method counts the number of positive, negative and neutral comments. This method does not take into consideration the priority of the users for the product features. For example, a user comments about the bad quality of the mobile phone camera and its sound but liked the battery performance and he decided to give the product a 4 out of 5 stars because, for the user, battery performance matters more than the camera or sound quality. But under HBF method, the overall comment would have a higher negative weightage.
- The model works on word identification hence it wrongly classifies the negative words used positively. For example, "the painting is terribly beautiful". This makes it hard for the algorithm to classify this comment. It would ideally classify it as neutral but in reality, it is a positive comment.

Drivers of Helpfulness



Drivers of Helpfulness

Table 4
Results of Regression Models Explaining Review Helpfulness.

	Poisson Regression 1	Poisson Regression 2	Negative Binomial Regression 1	Negative Binomial Regression 2	Hypothesis Supported
AIC Value	3161.7	3158.9	2106.9	2108.7	
(Intercept)	0.19^	0.1^	−0.10^	−0.12^	
RC	−0.00***	−0.00***	0.00^	0.00^	Not H7
OS	−0.00^	0.00^	−0.00^	0.00^	Not H1
PL	−0.04***	−0.05***	0.02^	0.02^	H2
TP	−0.11***	−0.11***	−0.10***	−0.11***	H4
OR	0.10**	0.09*	0.14˜	0.14˜	H8
TD	−0.00*	−0.00*	−0.00^	−0.00^	H9
DSG	0.14***	0.12***	0.14*	0.13*	H5
FR	−0.11**	−0.12***	−0.14*	−0.14*	H6
SAD	0.06*	0.06*	0.09˜	0.10˜	H5
THV	0.06***	0.06***	0.07**	0.07**	H10
OS x PL		0.00*		0.00^	H3

*** means $p < 0.001$; ** means $p < 0.01$; * means $p < 0.05$, ˜ means $p < 0.1$, ^ means NS.
RC = Review Count, OS = Overall Review Sentiment, PL = Review Polarity, TP = Title Polarity, OR = Overall Quantitative Rating, TD = Time Distance Between Experience and Review, THV = Total Helpful Votes by the reviewer, DSG = Disgust, FR = Fear, SAD = Sadness.

Regression types

- **Negative binomial regression** and **Poisson regression** are two types of regression models that are appropriate to use when the response variable is represented by discrete count outcomes.

- If the variance is roughly equal to the mean, then a Poisson regression model typically fits a dataset well.
- However, if the variance is significantly greater than the mean, then a negative binomial regression model is typically able to fit the data better.

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Trinity Business School

Text Mining
Advanced- n-grams

Dealing with n-grams

Necessary to deal with n-grams in many contexts

Most popularly used in case of bi-grams

- Sentiment mining
- Understanding contexts
- Filtering

Create Network Graphs

Necessary to understand the edges and nodes

– Details in session 4

•**from**: the node an edge is coming from

•**to**: the node an edge is going towards

•**weight**: A numeric value associated with each edge

A representation of Markov Chain

Useful for understanding the contexts in a visual fashion

Wordwise correlation

Basic correlation

- Useful function widyr

Phi Function

- The Phi Coefficient is a measure of association between two binary variables (i.e. living/dead, black/white, success/failure)
- It is also called the Yule phi or Mean Square Contingency Coefficient and is used for contingency tables
- The focus of the phi coefficient is how much more likely it is that either **both** word X and Y appear, or **neither** do, than that one appears without the other.

Wordwise correlation

Insert the counts into the formula and solve.
Φ = ad – bc / √((a + b)(c + d)(a + c)(b + d))
Φ = 14*13 – 10*6 / √((14 + 10)(6 + 13)(14 + 6)(10 + 13))
Φ = 182 – 60 / √((24)(19)(20)(23))
Φ = 122/ √((24)(19)(20)(23))
Φ = 122/ 458
Φ = 0.266.

$$\Phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

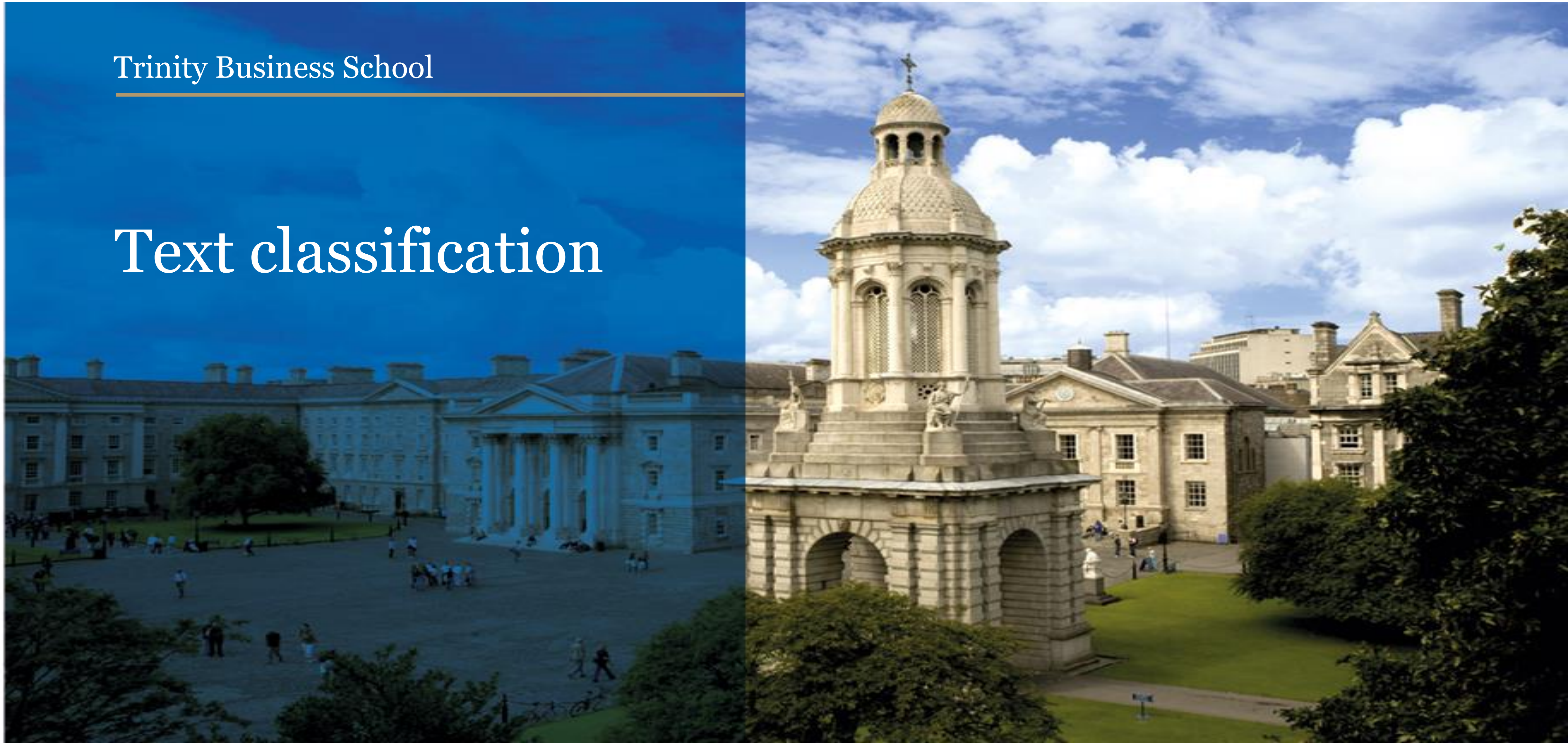
Example: Find phi for the following contingency table:

		Politicans	
		Truthful	Not Truthful
Scientists	Truthful	14	10
	Not Truthful	6	13

	Has word Y	**No word Y**	**Total**
Has word X	n_{11}	n_{10}	$n_{1.}$
No word X	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	n

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Trinity Business School



Text classification

Naïve Bayesian classifiers

- P(A|B): Conditional probability of event A occurri

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- P(A): Probability of event A occurring
- P(B): Probability of event B occurring
- P(B|A): Conditional probability of event B occurring, given the event A

Given a Hypothesis H and evidence E, Bayes Theorem states that the relationship between the probability of Hypothesis before getting the evidence P(H) and the probability of the hypothesis after getting the evidence P(H|E)

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Naïve Bayesian classifiers

$$P(C_i \mid x_1, x_2 \dots, x_n) = \frac{P(x_1, x_2 \dots, x_n \mid C_i).P(C_i)}{P(x_1, x_2 \dots, x_n)} \; for\, 1 < i < k$$

Naïve Bayesian classifiers- Example

Type	Swim	Wings	Green	Sharp Teeth
Cat	450/500	0	0	500/500
Parrot	50/500	500/500	400/500	0
Turtle	500/500	0	100/500	50/500

	Swim	Wings	Green	Sharp Teeth
Observation	True	False	True	False

Example source- https://www.edureka.co/blog/naive-bayes-in-r/

Naïve Bayesian classifiers- Example

*P(Cat | Swim, Green) = P(Swim|Cat) * P(Green|Cat) * P(Cat) / P(Swim, Green)*
*= 0.9 * 0 * 0.333 / P(Swim, Green)*
= 0

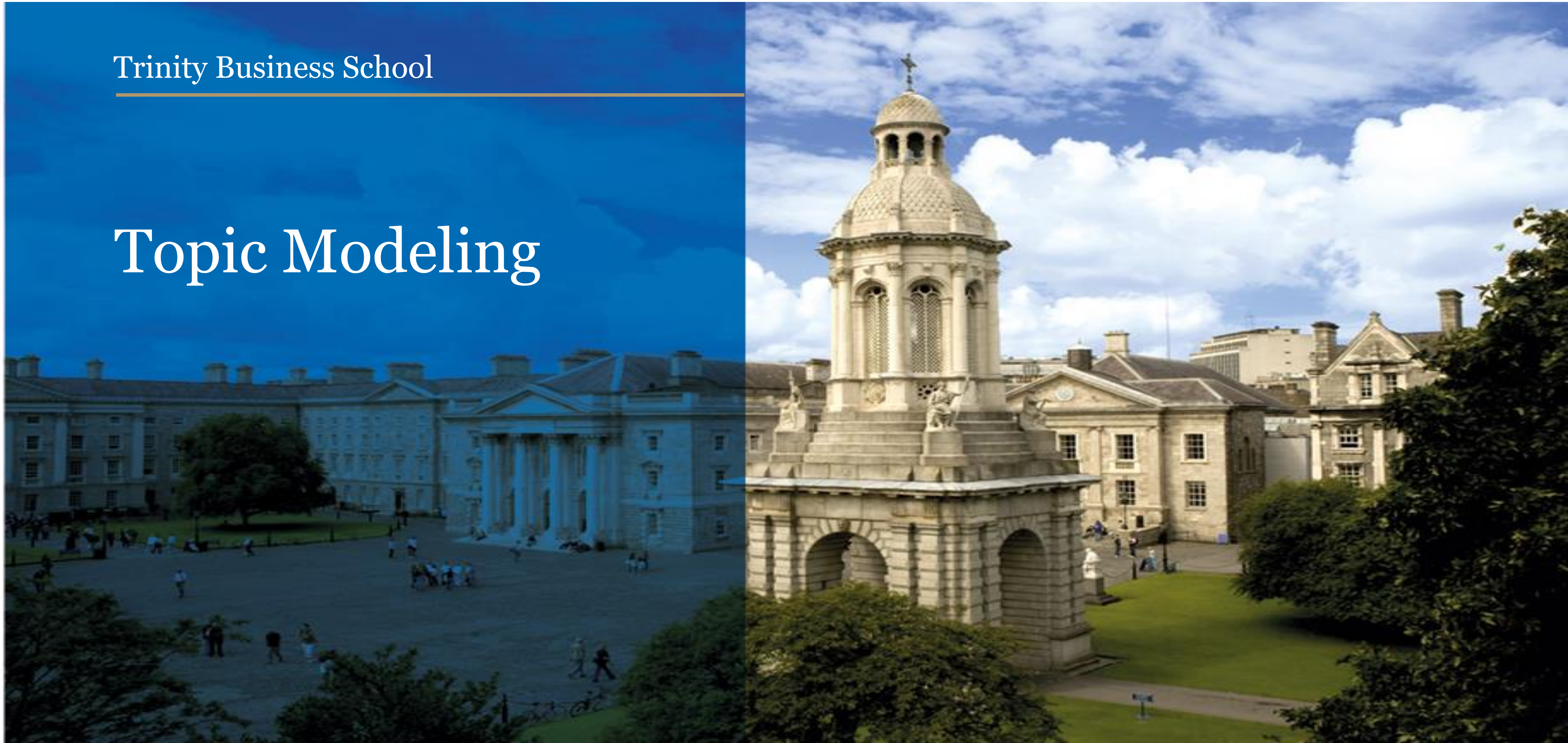
To check if the animal is a Parrot:
*P(Parrot| Swim, Green) = P(Swim|Parrot) * P(Green|Parrot) * P(Parrot) / P(Swim, Green)*
*= 0.1 * 0.80 * 0.333 / P(Swim, Green)*
= 0.0264/ P(Swim, Green)

To check if the animal is a Turtle:
*P(Turtle| Swim, Green) = P(Swim|Turtle) * P(Green|Turtle) * P(Turtle) / P(Swim, Green)*
*= 1 * 0.2 * 0.333 / P(Swim, Green)*
= 0.0666/ P(Swim, Green)

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Trinity Business School

Topic Modeling



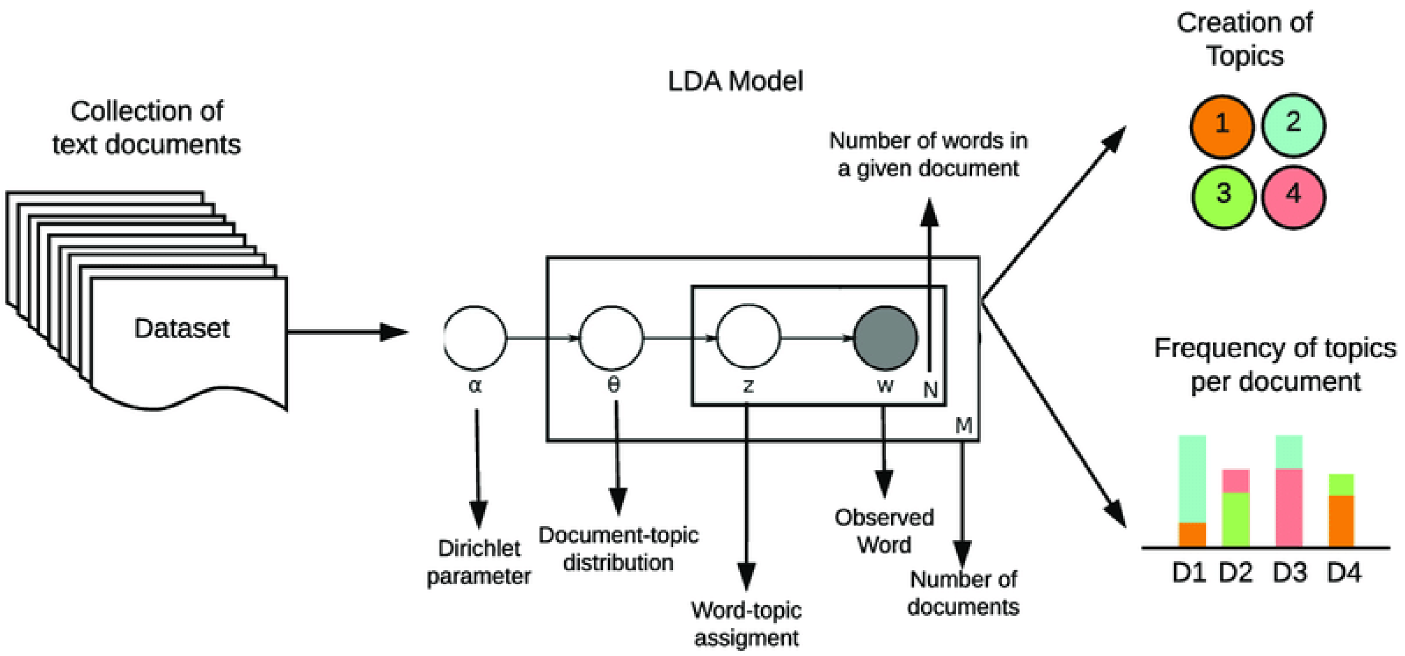
LDA Theory

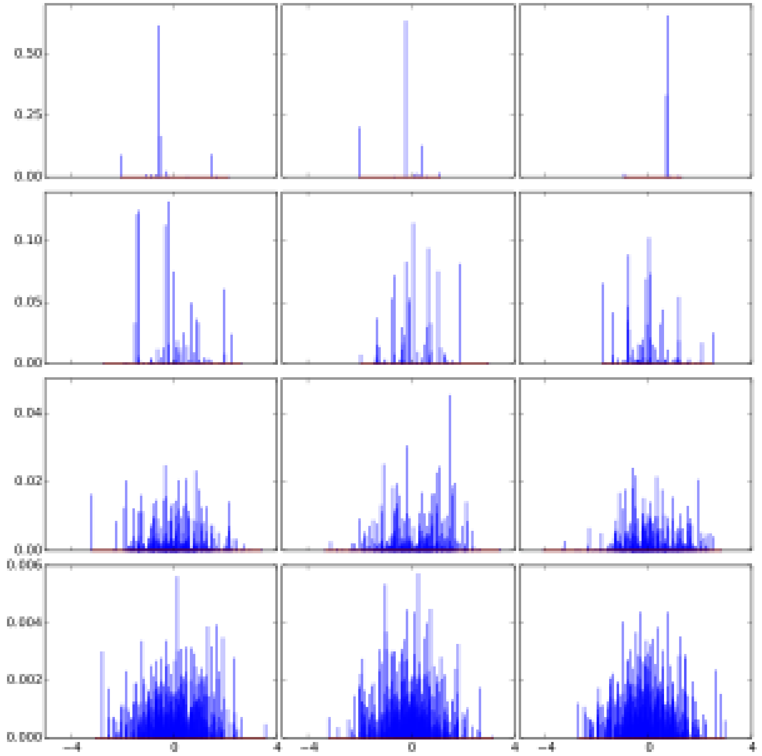
LDA assumes that each document in a corpus contains a mix of topics that are found throughout the entire corpus.

The topic structure is hidden - we can only observe the documents and words, not the topics themselves.

Because the structure is hidden (also known as **latent**), this method seeks to infer the topic structure given the known words and documents.

Dirichlet process





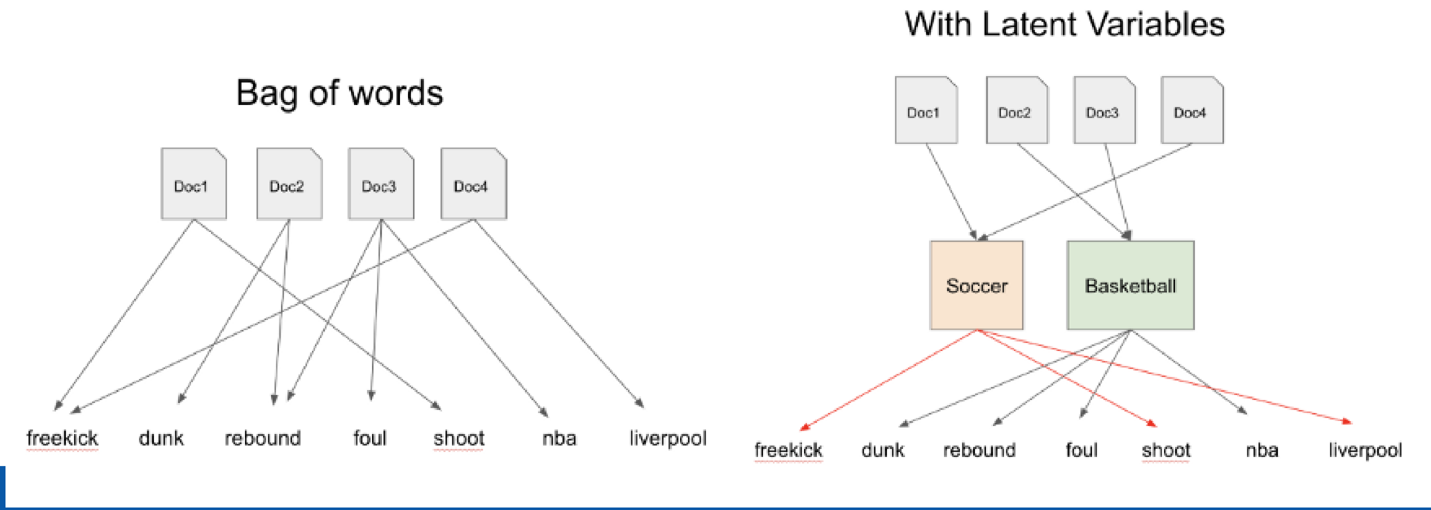
LDA Theory

First Iteration:

In the first iteration, it randomly assigns the topics to each word in the document.

Subsequently:

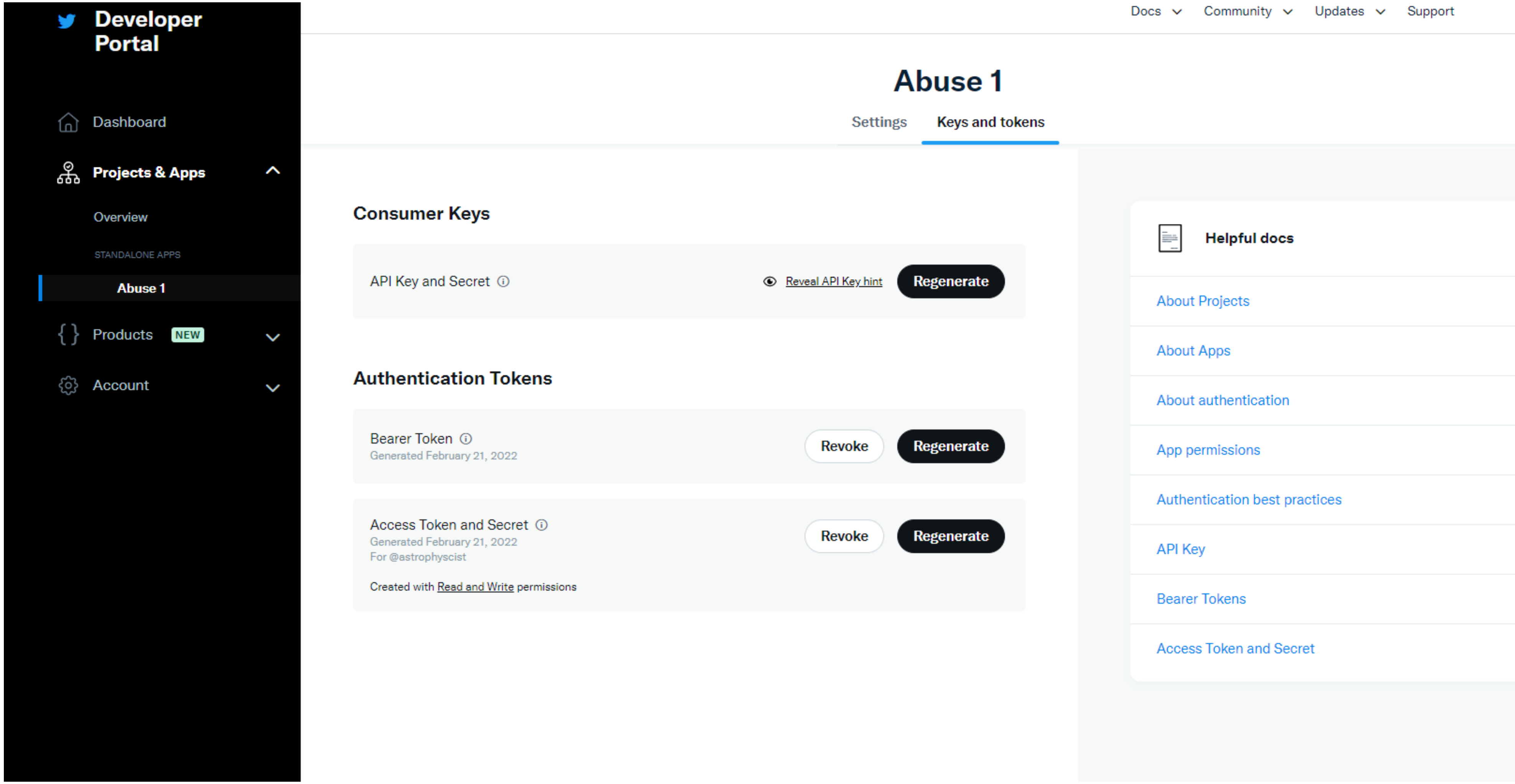
LDA makes another assumption that all the topics that have been assigned are correct except the current word. So, based on those already-correct topic-word assignments, LDA tries to correct and adjust the topic assignment of the current word with a new assignment



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Trinity Business School

Twitter Data collection


Developer Portal
Dashboard
Projects & Apps
Overview
STANDALONE APPS
Abuse 1
Products NEW
Account
Docs
Community
Updates
Support
Abuse 1
Settings
Keys and tokens
Consumer Keys
API Key and Secret
Reveal API Key hint
Regenerate
Authentication Tokens
Bearer Token
Generated February 21, 2022
Revoke
Regenerate
Access Token and Secret
Generated February 21, 2022
For @astrophyscist
Created with Read and Write permissions
Revoke
Regenerate
Helpful docs
About Projects
About Apps
About authentication
App permissions
Authentication best practices
API Key
Bearer Tokens
Access Token and Secret