Topic Modeling with LDA, LSA & Top2Vec using #MeeToo Twitter Data

Dileep kumar Kommineni | S V Arun Varma Vanaparthi | Sai Kiran Togiti | Nivedita Gadade | Sagarika Patial

Abstract:

In October 2017, many women accused Harvey Weinstein, a producer, for sexual harassment. Their experiences prompted more women to file sexual harassment charges against well-known figures such as politicians, actors, and producers. Because of the popular use of the hashtag, these instances have become known as the "#MeToo movement." Twitter has become a widespread platform for all the people who can share their thoughts on the internet, One of them though is the #MeToo social movement, people share their stories of sexual assault and harassment in an effort to end rape culture and other forms of sexual violence. Sexual assault survivor and activist Tarana Burke first introduced the phrase "Me Too" in this context on social media in 2006 on Myspace. By using #MeToo twitter data we are going to perform topic modeling. We are going to use LDA[Latent Dirichlet allocation], LSA[Latent Semantic Analysis] and Top2Vec. Generate a report of which is the best one or combine all the algorithms. Finally, we can conclude, we can go with LSA by looking at the performance with very large datasets, but LDA is standard for topic modeling it will be helpful for small datasets and Top2Vec is more accurate finding the topics and words regardless of size of the dataset.

Key Words: Latent Dirichlet Allocation (LDA), Latent Semantic Analysis(LSA),Top2Vec,Topic modeling, Twitter, #MeeToo

Introduction:

Tarana Burke established the #MeToo campaign in 2006 with the intention of instilling hope and unity among women who have faced sexual harassment or assault (Ohlheiser 2018). Following the wave of sexual harassment claims leveled against producer Harvey Weinstein in October 2017, actress Alyssa Milano used the hashtag #MeToo in a tweet and encouraged others to do the same. Her statement inspired a national movement, raising attention to the prevalence of sexual harassment and urging people to speak out about their experiences. According to Tarana Burke, the major purpose of the movement is "empowerment through empathy."

Use the hashtag "#metoo" on social media sites such as Twitter and Facebook. The movement is frequently referred to as feminism since it has given previously unheard women's voices greater significance than those of traditionally powerful males. The movement is frequently referred to as "empowering" because it has given previously unheard women's voices greater importance than those of traditionally prominent men. We examine how sentiment, power, and agency dynamics play out in online media coverage of these events in our study. We present a contextual emotional analysis of internet media pieces regarding the #MeToo movement.

Impact on #MeeToo in Social media:

According to Twitter, the hashtag was used nearly a million times in two days after Milan's tweet. The move spread to his Facebook, where around 4.7 million followers shared his 12 million posts within 24 hours. People are still using the hashtag #MeToo to express their stories on social media platforms years later. The reaction was especially important for those who deal daily with sexual assault survivors and harassment. Finally, the problem they had been fighting persistently to solve was getting traction and attracting international attention. Burke's local grassroots initiative had now grown to include a network of survivors from all walks of life.

As a result, the quiet surrounding sexual harassment and assault is being shattered. Many people are now willing and enthusiastic about discussing the concerns. The #MeToo campaign has resulted in significant improvements, including:

- It has been confirmed for victims that they're not alone.
- Developed a stronger society in which victims can speak out
- shown how pervasive the problem is
- Social conventions and attitudes toward the issue have shifted.
- Exposing thought systems that allow for violence
- Compassion towards survivors has grown.
- modified and adopted legislation and policies

Statistics of Sexual Harassment:

Sexual aggressiveness is a widespread problem. According to 2018 research performed by the University of California and the non-profit Stop Street Harassment, 81% of women and 43% of males reported experiencing some type of sexual harassment or assault.

Although the #Metoo campaign has done a lot in a short period of time, some supporters remain suspicious of its success. Although the topic remains on the public's awareness, sexual assault occurs. It's especially pernicious for transgender individuals, Native American women, college students, military personnel, and people of color. Women continue to be more vulnerable to sexual assault than males.

Military Sexual Trauma Leaving Permanent Marks

Violence and bullying can have terrible consequences, frequently leading to drug and alcohol abuse, homicide, and brain disorders.

Literature review:

The literature on social media analysis has previously been examined in a variety of applications that investigate the crucial topic #MeToo which has been a widely discussed subject on social media since 2017. During this period people organized pro-tests, raised awareness, shared tales, and, most crucially, documented violence and violations of people's rights to free speech and civic participation . The Me-Too movement (or #MeToo movement), which began in 2006, was elevated and picked up steam in 2017 when prominent female artists and actors chose to use the hashtag #MeToo to discuss their experiences with sexual harassment. This increased public awareness of sexual harassment issues and resulted in the conviction of several prominent individuals in Hollywood and the film business.

The authors of say that by enabling women to express their rage and stories which is difficult to do otherwise—the MeToo movement is bringing about change in the era of feminism. They add that a few men have also disclosed their stories as part of the #MeToo movement. The psychology of males and masculinity in the MeToo movement is a topic of study for a psychologists. They point out that some powerful men are reluctant to engage in mentor-ship relationships with women because of the MeToo movement. In, authors assert that being reluctant to mentor women is motivated by a desire to discredit the women who speak out against sexual assault and harassment rather than a simple fear of receiving false claims of sexual misconduct.

In the previous works, the author discussed about the occurrences of SH (Sexual Harassment) and SA (Sexual assault) can occur in respected and educated professions. Despite this, there are few SH cases that are reported, and this is primarily because people shy away from or are scared to bring up SA. Defense forces are also affected by the SA and SH problem. For instance, several studies have been done to help and protect victims as well as hold offenders accountable in the army.

The authors of a different book concentrate on addressing the key reasons why the experiences of color of women are disregarded in the MeToo movement. They contend that the existence of racial biases within the movement demonstrates the need for some logical modifications to the SH philosophy.

In another part of work, authors describe how users on various platforms (such as Twitter and Reddit) share their own experiences and react to those of others. According to their research, people prefer to follow the MeToo movement on Twitter instead of sharing their personal experiences in-depth on Reddit.

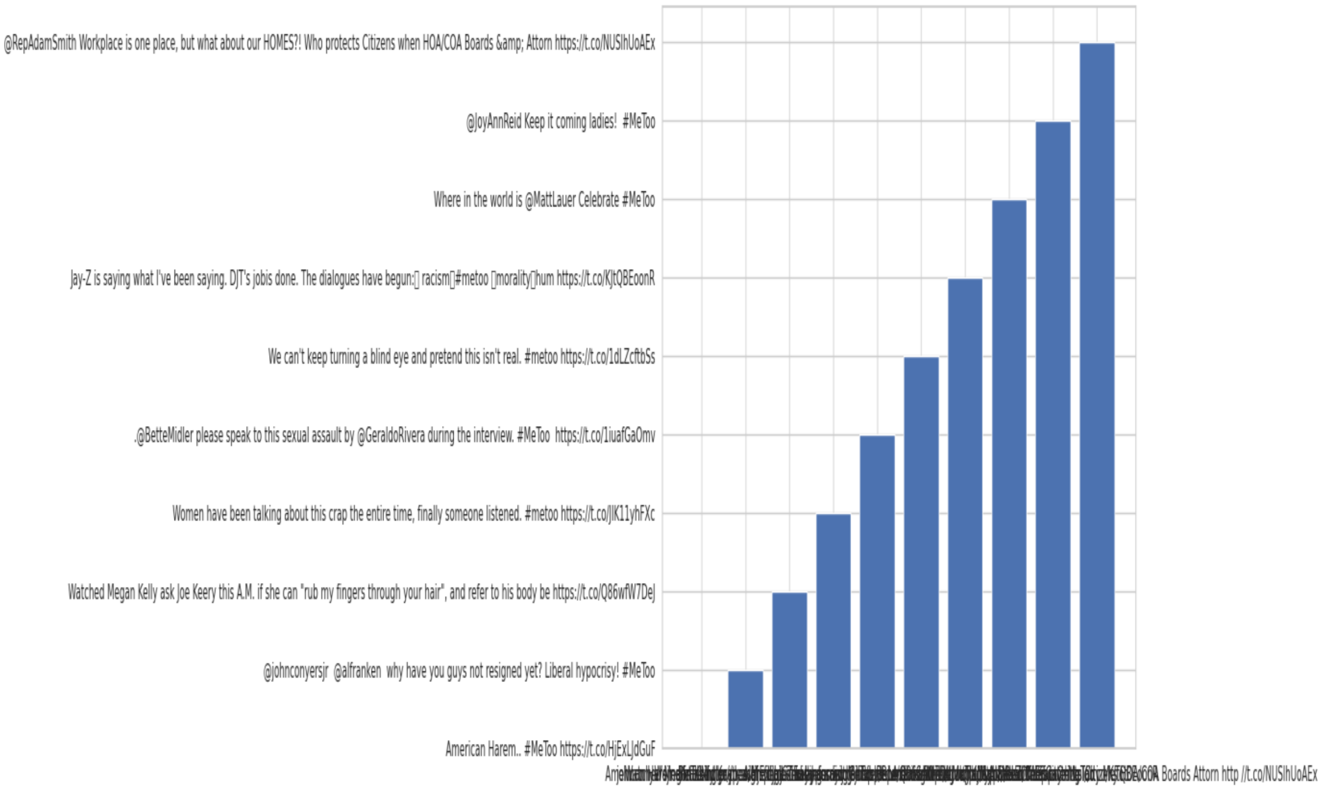
This study differs from others since it focuses on examining the MeToo movement rather than just one aspect while considering preconceived notions or information. But in this study, we uncover #MeToo's hidden aspects outside of SH and SA. Without making any assumptions ahead of time, these subtopics are immediately retrieved from the data of the gathered tweets.

Anjalie Field et al in their paper Contextual Affective Analysis: mentioned NLP's contributions (Bamman 2015). However, most approaches (Iyyer et al. 2016; Chambers and Jurafsky 2009; Bamman, O'Connor, and Smith 2013; Card et al. 2016) rely on unsupervised models, which can capture high-level patterns but are difficult to interpret and do not target specific dimensions.To tackle this problem they used Contextual Affective Analysis: which was effective for the binary format of the output and could not predict the hate speech correctly.

Methodology:

Our model's approach is based on natural language processing (NLP), and the technique utilized is known as sentiment analysis, which is the act of identifying positive or negative sentiments in text. Businesses frequently utilize it to identify sentiment in social data, assess brand reputation, and better understand customers. Classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the conveyed opinion in a document, phrase or entity feature/aspect is positive, negative, or neutral—is a fundamental job in sentiment analysis. Advanced sentiment categorization "beyond polarity" considers emotional states such as delight, anger, contempt, sadness, fear, and surprise. Existing sentiment analysis methodologies may be divided into three categories: knowledge-based techniques, statistical methods, and hybrid approaches.

Bar Graph On Raw Data for Top 10 Topics Word Frequency :



Data Collection and Preprocessing:

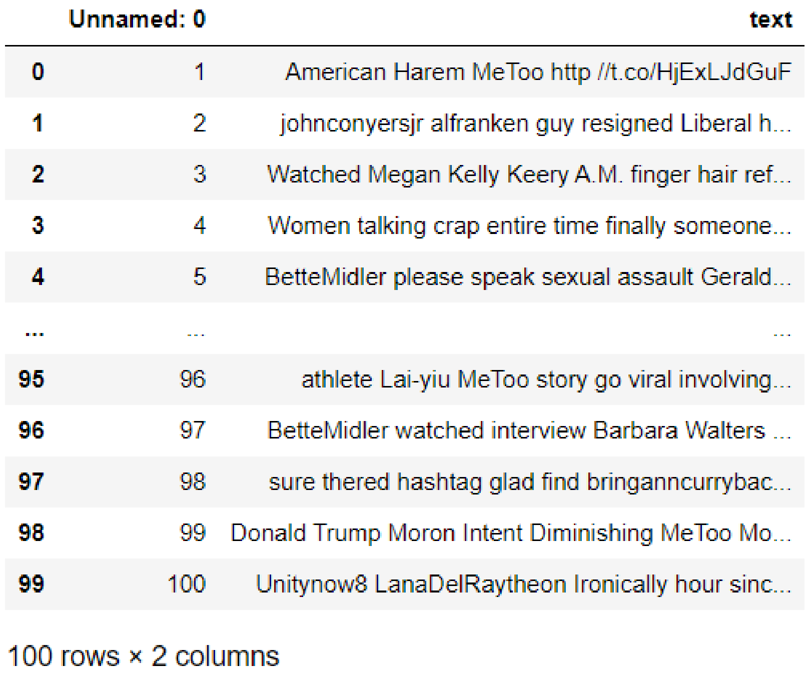
Text mining is a process of extracting information from unstructured text data. It involves prepossessing the text to convert it into a structured form that is more amenable to analysis. Preprocessing steps such as cleansing, case folding, normalization, stemming, stop-word removal, and tokenizing are applied to the text data to improve its quality and accuracy.

Cleansing is the process of removing unwanted words or characters from the text data, such as HTML tags, emoticons, hashtags, usernames, and URLs. Case folding, on the other hand, converts all characters to lowercase letters to make the text data more uniform. Normalization is the process of standardizing the text data, such as converting numbers to their word form or converting abbreviations to their full form.

Stemming is the process of reducing words to their root form by removing affixes such as prefixes, suffixes, and inflections. This step is useful for reducing the size of the text data and for improving the performance of text analysis algorithms. Stop-word removal is the process of removing words that are considered to be uninformative or irrelevant, such as prepositions, conjunctions, and pronouns. These words are often common in text data but do not contribute to the meaning of the text.

Finally, tokenization is the process of breaking the text data into smaller pieces, such as words or phrases. This step is useful for preparing the text data for further analysis, such as identifying patterns or relationships between words. Overall, text mining is a valuable technique for extracting useful information from unstructured text data.

First 100 Tweets after Preprocessing of Data:



	Unnamed: 0	text
0	1	American Harem MeToo http //t.co/HjExLJdGuF
1	2	johnconyersjr alfranken guy resigned Liberal h...
2	3	Watched Megan Kelly Keery A.M. finger hair ref...
3	4	Women talking crap entire time finally someone...
4	5	BetteMidler please speak sexual assault Gerald...
...	...	...
95	96	athlete Lai-yiu MeToo story go viral involving...
96	97	BetteMidler watched interview Barbara Walters ...
97	98	sure thered hashtag glad find bringanncurrybac...
98	99	Donald Trump Moron Intent Diminishing MeToo Mo...
99	100	Unitynow8 LanaDelRaytheon Ironically hour sinc...

100 rows × 2 columns

Latent Dirichlet Allocation :

The Latent Dirichlet Allocation (LDA) method is a popular technique for topic modeling, which is the process of automatically identifying the topics present in a collection of documents. LDA treats each document as a mixture of topics, and each topic as a mixture of words. This allows for documents to overlap with each other in terms of content, rather than being separated into discrete groups.

LDA is a probabilistic model, which means that it uses probabilities to predict the likelihood that a given topic or word will appear in a document. This allows LDA to produce more accurate results than other methods that rely on heuristics or rules of thumb.

LDA is widely used in many machines learning, natural language processing (NLP), and information retrieval applications. For example, researchers have used LDA to identify scientific topics in a collection of documents, or to classify documents based on their content.

The LDA model works in three steps (Blei, Ng, & Jordan, 2003):

1. Each document in the collection is represented as a distribution over topics, where the distribution is sampled for that document based on a Dirichlet distribution.
2. Each word in the document is associated with a single topic, based on the chosen Dirichlet distribution.
3. Each topic is represented as a multinomial distribution over words that are assigned to the topic.

The following notations will be used to describe the LDA process:

- M: the number of documents in the corpus
- N: the number of words in each document
- w: a vector representing a document, where wn is the nth word in the sequence and the vector has a single component equal to one and all other components equal to zero
- V: the size of the vocabulary, where the vth word in the vocabulary is represented by a vector such that wv = 1 and wu = 0 for u ≠ v
- D: the corpus, represented as a collection of M documents
- k: the number of topics a document belongs to
- z: a topic from a set of k topics.

LDA uses these notations to estimate the topic distributions in the corpus and the word distributions for each topic. This allows it to identify the topics that are most strongly associated with specific words, and to determine how each document in the corpus relates to these topics.

The following formula is used to compute and derive the probability of the observed dataset from the corpus D:

$$P\,D|\alpha,\beta) =$$

$$= \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{d}} p(z_{dn}|\theta_d)p(w_{\,n}|z\;\;,\beta)\right)d\theta_d$$

In LDA, the document-level variable θ represents the distribution of topics in the dth document and is sampled once per document. The word-level variables z{d,n} and w{d,n} represent the topic and word associated with the nth word in the dth document, respectively, and are sampled once for each word in each document.

In other words, θ is a matrix of probabilities that indicate the likelihood of a given document containing words from each of the K topics. The Dirichlet distribution is a distribution over multinomial distributions, which is suitable for modeling the distribution of topics in a document. The parameter α controls the concentration of the Dirichlet distribution, which in turn controls the concentration of the topic distribution in each document.

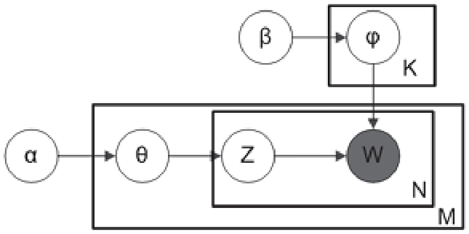


Fig: Probabilistic Graph Model of LDA

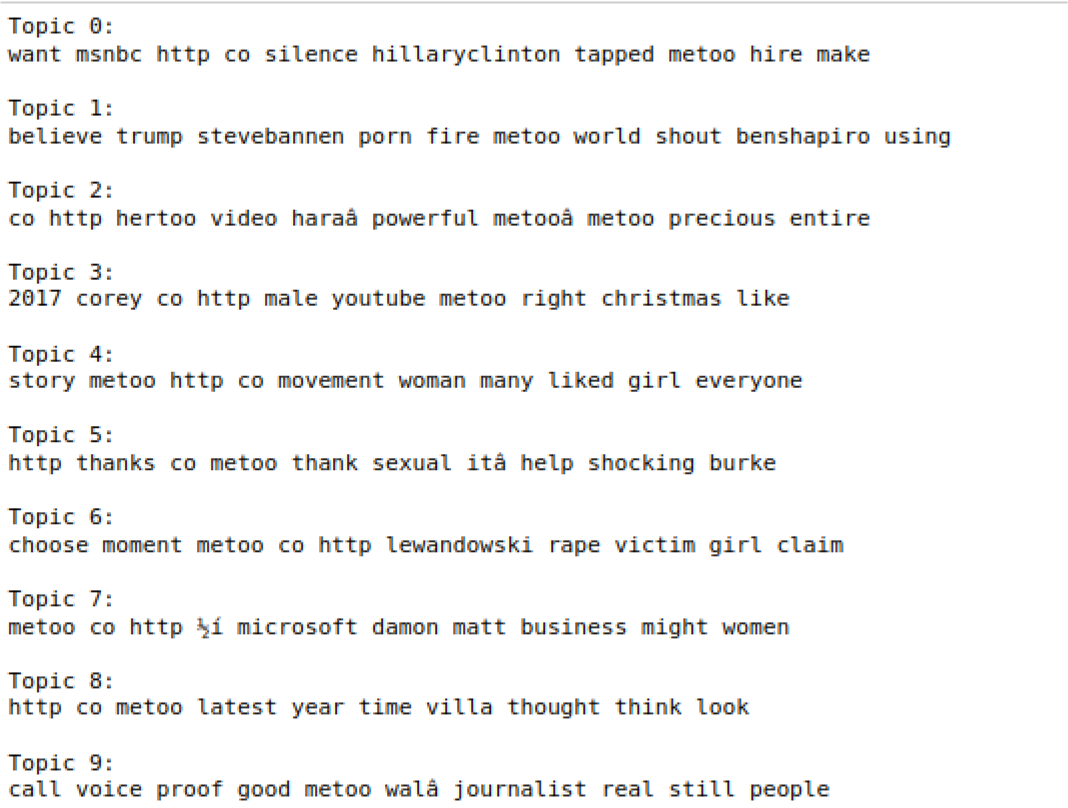
The graphical model shown in Figure 1 is a directed acyclic graph (DAG) that illustrates the relationships between the variables in the LDA model. In this model, the corpus-level parameters alpha and beta are connected to the document-level variable theta, which represents the distribution of topics over the document. Theta is then connected to the word-level variables z and w, which represent the topic and the word, respectively.

The joint distribution of the DAG in Figure 1 can be written as follows:

$$p(w,z,\theta,\varphi|\alpha,\beta) = p(\theta|\alpha)p(z|\theta)p(\varphi|\beta)p(w|z,\varphi).$$

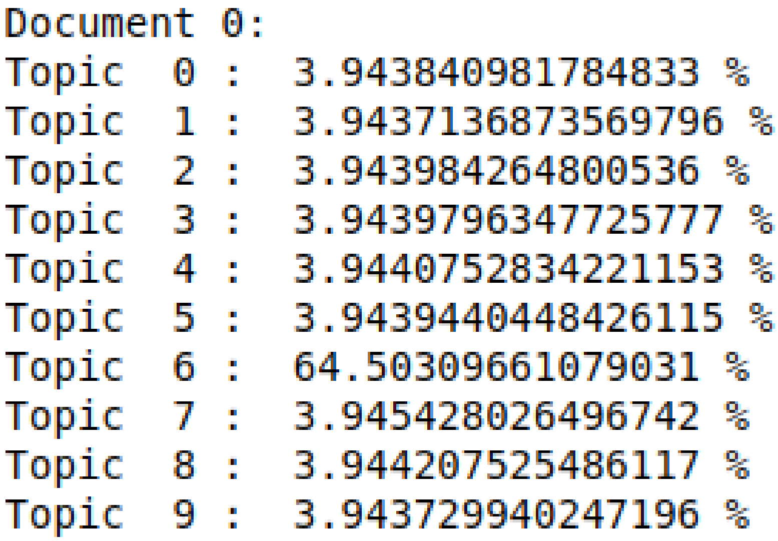
This equation describes the probability of the observed dataset, as well as the probability of the latent variables (alpha, beta, theta, z, and w) given the observed data. The joint distribution of these variables can be used to infer the latent topics in the document collection and to generate new samples that resemble the input data.

LDA top 10 Words Results:



```
Topic 0:
want msnbc http co silence hillaryclinton tapped metoo hire make

Topic 1:
believe trump stevebannen porn fire metoo world shout benshapiro using

Topic 2:
co http hertoo video haraâ powerful metooâ metoo precious entire

Topic 3:
2017 corey co http male youtube metoo right christmas like

Topic 4:
story metoo http co movement woman many liked girl everyone

Topic 5:
http thanks co metoo thank sexual itâ help shocking burke

Topic 6:
choose moment metoo co http lewandowski rape victim girl claim

Topic 7:
metoo co http ½í microsoft damon matt business might women

Topic 8:
http co metoo latest year time villa thought think look

Topic 9:
call voice proof good metoo walâ journalist real still people
```

LDA % of Contribution:



```
Document 0:
Topic  0 :  3.943840981784833 %
Topic  1 :  3.9437136873569796 %
Topic  2 :  3.943984264800536 %
Topic  3 :  3.9439796347725777 %
Topic  4 :  3.9440752834221153 %
Topic  5 :  3.9439440448426115 %
Topic  6 :  64.50309661079031 %
Topic  7 :  3.945428026496742 %
Topic  8 :  3.944207525486117 %
Topic  9 :  3.943729940247196 %
```

Latent Semantic Analysis (LSA):

Latent Semantic Analysis (LSA) is a method used to extract and represent the meaning of words in a text document. It uses statistical calculations and natural language processing techniques to analyze a large corpus of text documents and identify common themes or topics.

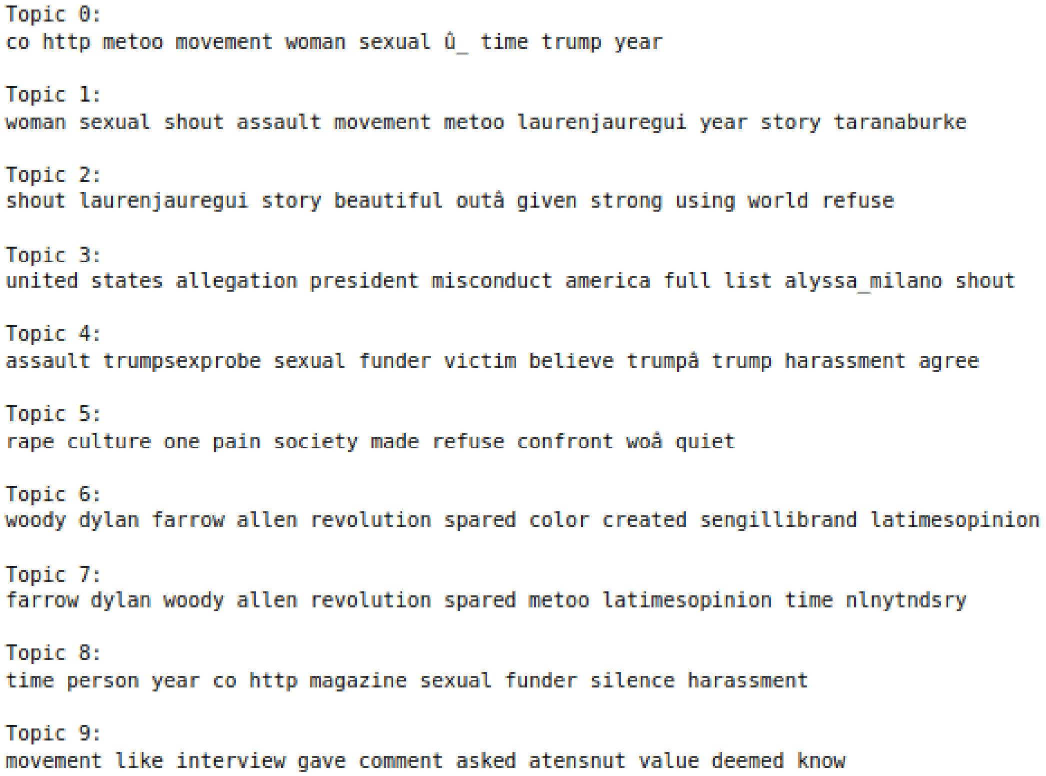
LSA is a form of semantic analysis that uses Singular Value Decomposition (SVD) to identify generalizations in the text. The first step in LSA is to represent the text as a matrix, where each row represents a unique word and each column represents a part of the text or another context. The cells in the matrix contain the frequency of word occurrence, weighted by a function that represents the importance of the word in the context and its overall information content.

Next, LSA applies the SVD to the matrix. SVD is based on a theorem that states that a square matrix can be decomposed into three matrix multiplications. These three matrices are an orthogonal matrix U, a diagonal matrix S, and the transpose of the orthogonal matrix V. The theorem can be written as follows:

$$\mathbf{A}_{m\times n} = \mathbf{U}_{m\times m} \times \mathbf{S}_{m\times n} \times \mathbf{V}^{T}_{n\times n} \,.$$

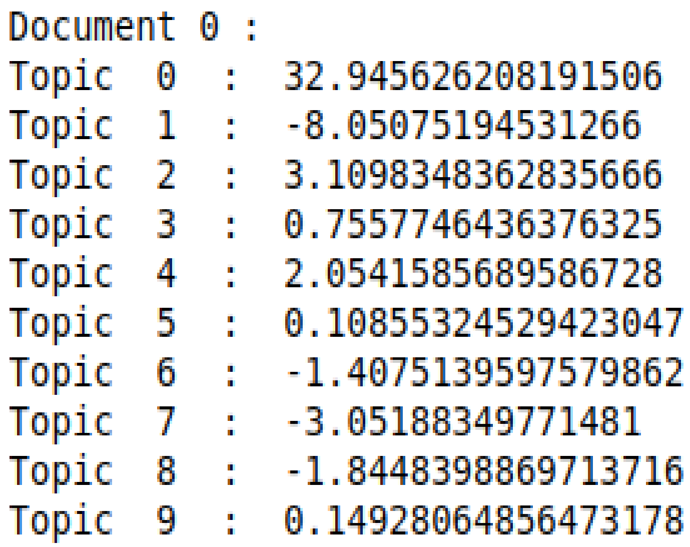
The resulting matrices can be used to identify common themes or topics in the text and to generate new samples that resemble the input data. LSA is commonly used in natural language processing and information retrieval applications.

LSA top 10 Words Results:



```
Topic 0:
co http metoo movement woman sexual û_ time trump year

Topic 1:
woman sexual shout assault movement metoo laurenjauregui year story taranaburke

Topic 2:
shout laurenjauregui story beautiful outâ given strong using world refuse

Topic 3:
united states allegation president misconduct america full list alyssa_milano shout

Topic 4:
assault trumpsexprobe sexual funder victim believe trumpâ trump harassment agree

Topic 5:
rape culture one pain society made refuse confront woâ quiet

Topic 6:
woody dylan farrow allen revolution spared color created sengillibrand latimesopinion

Topic 7:
farrow dylan woody allen revolution spared metoo latimesopinion time nlnytndsry

Topic 8:
time person year co http magazine sexual funder silence harassment

Topic 9:
movement like interview gave comment asked atensnut value deemed know
```

LSA % of Contribution:

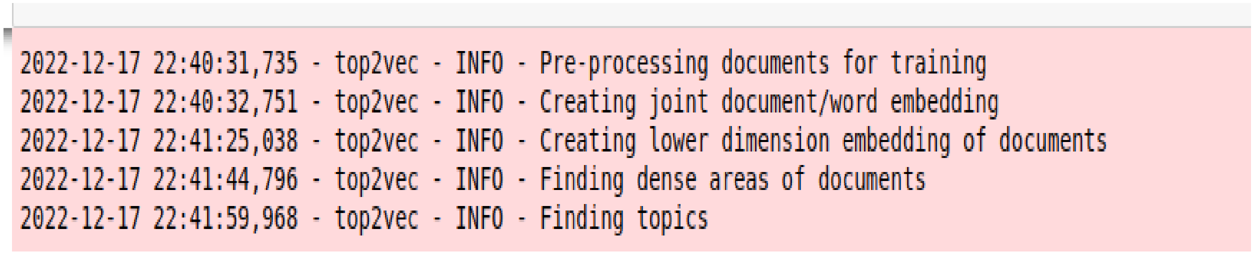


```
Document 0 :
Topic  0  :  32.945626208191506
Topic  1  :  -8.05075194531266
Topic  2  :  3.1098348362835666
Topic  3  :  0.7557746436376325
Topic  4  :  2.0541585689586728
Topic  5  :  0.10855324529423047
Topic  6  :  -1.4075139597579862
Topic  7  :  -3.05188349771481
Topic  8  :  -1.8448398869713716
Topic  9  :  0.14928064856473178
```

Top2Vec Model Application:

Top2Vec is used to perform unsupervised topic modelling using embedding vectors and clustering techniques. It detects topics present in the text and generates combined embedded topics, document, and word vectors. It works in 5 steps:

The following screenshot is the actual copy of commands/processes that were displayed as output on running the algorithm.



```
2022-12-17 22:40:31,735 - top2vec - INFO - Pre-processing documents for training
2022-12-17 22:40:32,751 - top2vec - INFO - Creating joint document/word embedding
2022-12-17 22:41:25,038 - top2vec - INFO - Creating lower dimension embedding of documents
2022-12-17 22:41:44,796 - top2vec - INFO - Finding dense areas of documents
2022-12-17 22:41:59,968 - top2vec - INFO - Finding topics
```

STEP1 : To generate embedding vectors for documents and words.

Embedding vector allows developer to represent word or text document in multi-dimensional space. The purpose of embedding vectors is to wrap up similar words and text documents to have similar vectors. Creating embedding vectors for every document allows developer to treat each document as a single point in multi-dimensional space. Word Vectors helps to determine topic keywords later at the time of implementation.

STEP2 : Performing dimensionality reduction on the vectors using an algorithm UMAP.

Once we have unique vectors for each document, the next step would be to divide them into clusters. We can have enormous numbers of components, even more than 500 depending upon the embedded models we have used. Due to such large components we need to perform dimentionality reduction process to reduce the number of dimensions in the data. Top2Vec uses an algorithm called UMAP (Uniform Manifold Approximation and Projection).

STEP3 : Cluster the vectors using clustering algorithms such as HDBSCAN.

Top2Vec uses an algorithm HDBSCAN (hierarchical density-based clustering algorithm) it is used dense areas of documents. HDBSCAN is very efficient for very large topics of diverse subtopics.

STEP4 : Assign topics to each cluster.

After obtaining clusters for each document, we can use each cluster as a separate topics for the topic modelling. Every unique topic is represented as a topic vector which can used as a centriod (Average of all vector points). We can computer n-closed words to a topic centroid vector. Once we obtain all the keywords for each topic and its sub-topics used accordingly.

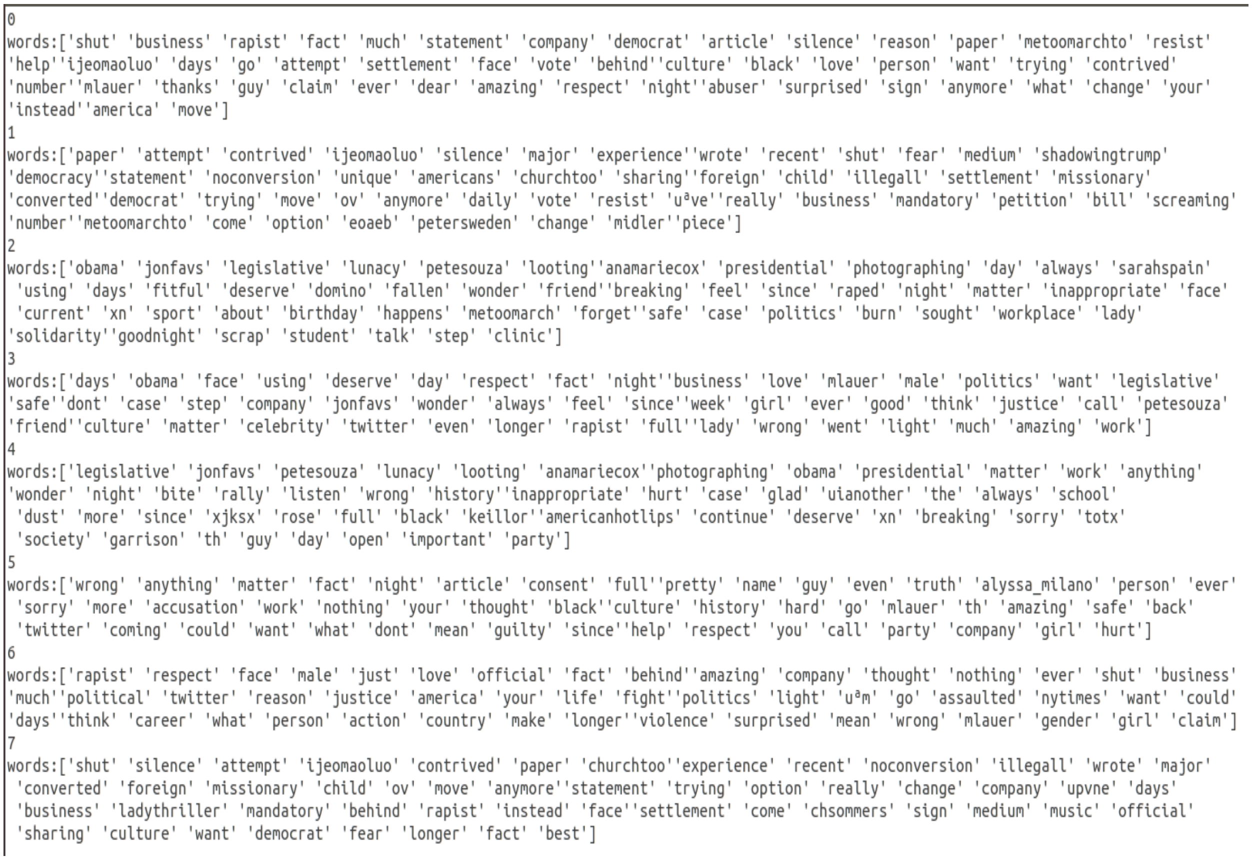
STEP5 : Topic assignment to the words.

After having topics and sub-topics we need to identify words as per each cluster assigned. It should be readable and helpful for us to identifying or predicting trends. We can further use these words for to for frequency of words or defining the negative, positive or neutral trends or response from users.

Why did we use Top2Vec?

Top2Vec is does not require us to remove stop-words. No, stop-words will appear in almost the whole document in a corpse. All the topics will be equidistant and will not appear as nearest word to any topic. We have used it for topic modelling in our project as it's a newly implemented algorithms as compared to other traditional algorithms which gives a room of experimentation. We have used approx. 40% of data-set to implement Top2Vec. As the data-set was huge working on the whole data-set was too time consuming. It is approx. Taking (8-12) minutes for 40% of the data-set.

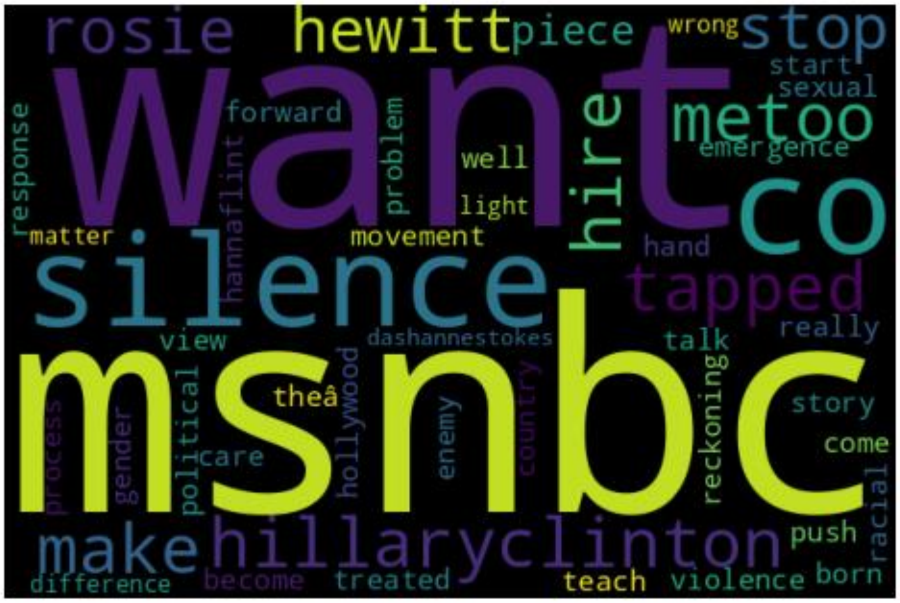
Output From Top2Vec After Preprocessing Data :



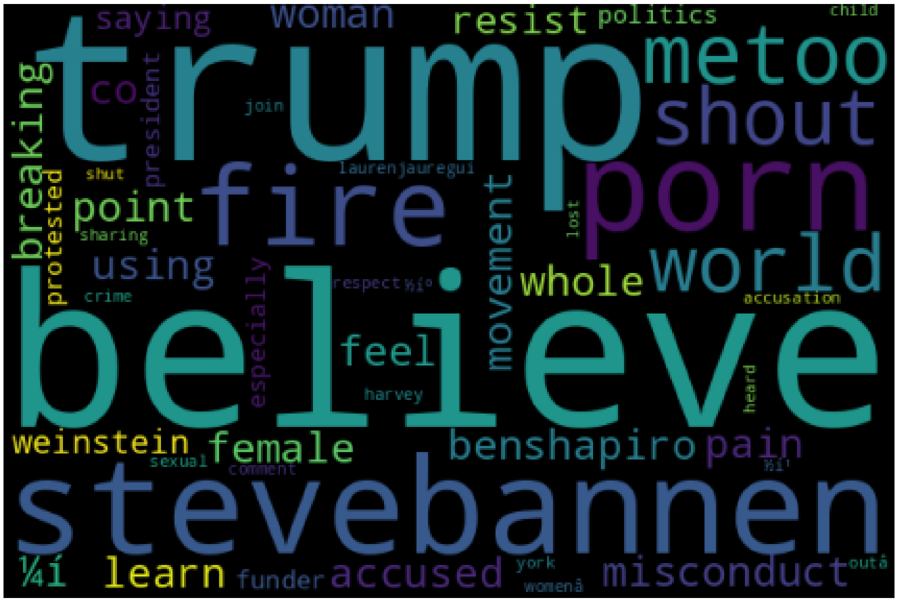
```
0
words:['shut' 'business' 'rapist' 'fact' 'much' 'statement' 'company' 'democrat' 'article' 'silence' 'reason' 'paper' 'metoomarchto' 'resist'
'help''ijeomaoluo' 'days' 'go' 'attempt' 'settlement' 'face' 'vote' 'behind''culture' 'black' 'love' 'person' 'want' 'trying' 'contrived'
'number''mlauer' 'thanks' 'guy' 'claim' 'ever' 'dear' 'amazing' 'respect' 'night''abuser' 'surprised' 'sign' 'anymore' 'what' 'change' 'your'
'instead''america' 'move']
1
words:['paper' 'attempt' 'contrived' 'ijeomaoluo' 'silence' 'major' 'experience''wrote' 'recent' 'shut' 'fear' 'medium' 'shadowingtrump'
'democracy''statement' 'noconversion' 'unique' 'americans' 'churchtoo' 'sharing''foreign' 'child' 'illegall' 'settlement' 'missionary'
'converted''democrat' 'trying' 'move' 'ov' 'anymore' 'daily' 'vote' 'resist' 'uªve''really' 'business' 'mandatory' 'petition' 'bill' 'screaming'
'number''metoomarchto' 'come' 'option' 'eoaeb' 'petersweden' 'change' 'midler''piece']
2
words:['obama' 'jonfavs' 'legislative' 'lunacy' 'petesouza' 'looting''anamariecox' 'presidential' 'photographing' 'day' 'always' 'sarahspain'
 'using' 'days' 'fitful' 'deserve' 'domino' 'fallen' 'wonder' 'friend''breaking' 'feel' 'since' 'raped' 'night' 'matter' 'inappropriate' 'face'
 'current' 'xn' 'sport' 'about' 'birthday' 'happens' 'metoomarch' 'forget''safe' 'case' 'politics' 'burn' 'sought' 'workplace' 'lady'
'solidarity''goodnight' 'scrap' 'student' 'talk' 'step' 'clinic']
3
words:['days' 'obama' 'face' 'using' 'deserve' 'day' 'respect' 'fact' 'night''business' 'love' 'mlauer' 'male' 'politics' 'want' 'legislative'
'safe''dont' 'case' 'step' 'company' 'jonfavs' 'wonder' 'always' 'feel' 'since''week' 'girl' 'ever' 'good' 'think' 'justice' 'call' 'petesouza'
'friend''culture' 'matter' 'celebrity' 'twitter' 'even' 'longer' 'rapist' 'full''lady' 'wrong' 'went' 'light' 'much' 'amazing' 'work']
4
words:['legislative' 'jonfavs' 'petesouza' 'lunacy' 'looting' 'anamariecox''photographing' 'obama' 'presidential' 'matter' 'work' 'anything'
'wonder' 'night' 'bite' 'rally' 'listen' 'wrong' 'history''inappropriate' 'hurt' 'case' 'glad' 'uianother' 'the' 'always' 'school'
 'dust' 'more' 'since' 'xjksx' 'rose' 'full' 'black' 'keillor''americanhotlips' 'continue' 'deserve' 'xn' 'breaking' 'sorry' 'totx'
 'society' 'garrison' 'th' 'guy' 'day' 'open' 'important' 'party']
5
words:['wrong' 'anything' 'matter' 'fact' 'night' 'article' 'consent' 'full''pretty' 'name' 'guy' 'even' 'truth' 'alyssa_milano' 'person' 'ever'
 'sorry' 'more' 'accusation' 'work' 'nothing' 'your' 'thought' 'black''culture' 'history' 'hard' 'go' 'mlauer' 'th' 'amazing' 'safe' 'back'
 'twitter' 'coming' 'could' 'want' 'what' 'dont' 'mean' 'guilty' 'since''help' 'respect' 'you' 'call' 'party' 'company' 'girl' 'hurt']
6
words:['rapist' 'respect' 'face' 'male' 'just' 'love' 'official' 'fact' 'behind''amazing' 'company' 'thought' 'nothing' 'ever' 'shut' 'business'
'much''political' 'twitter' 'reason' 'justice' 'america' 'your' 'life' 'fight''politics' 'light' 'uªm' 'go' 'assaulted' 'nytimes' 'want' 'could'
'days''think' 'career' 'what' 'person' 'action' 'country' 'make' 'longer''violence' 'surprised' 'mean' 'wrong' 'mlauer' 'gender' 'girl' 'claim']
7
words:['shut' 'silence' 'attempt' 'ijeomaoluo' 'contrived' 'paper' 'churchtoo''experience' 'recent' 'noconversion' 'illegall' 'wrote' 'major'
 'converted' 'foreign' 'missionary' 'child' 'ov' 'move' 'anymore''statement' 'trying' 'option' 'really' 'change' 'company' 'upvne' 'days'
 'business' 'ladythriller' 'mandatory' 'behind' 'rapist' 'instead' 'face''settlement' 'come' 'chsommers' 'sign' 'medium' 'music' 'official'
 'sharing' 'culture' 'want' 'democrat' 'fear' 'longer' 'fact' 'best']
```

Visualizations For Comparisons:

Word Cloud For LDA:



Word Cloud For LSA:



Conclusion:

All three algorithms have had various levels of success. While LSA produced a more cohesive topic collection and a more acceptable topic organization, it was unable to achieve significant distinction between these groups. Instead, LDA generated nearly the dissimilar: extremely strong topic separation, but the subjects obtained were not particularly understandable, and their spread appeared odd. As usual, the No Free Meal concept remains true. LDA and LSA are more of probabilistic algorithms rather Top2Vec is helps to find topics which are significantly more informative and representative of a corpus trained data-set. Top2Vec provides joint document and words sentiment to find vectors which gives more clarity rather than estimations. Topic modeling is a really difficult subject.

Nonetheless, our findings are generally favorable. The potential for subject demarcation and coherence in highlights data has been demonstrated, as have numerous basic features of the collection. Of course, there is room for enhancement model parameters could be tweaked (the number of topic areas N might be differed to acquire a more easy-to-interpret topic set, LDA's Dirichlet priors could be personalized to best suit the data), and even more time could be managed to spend on feature construction to possibly further minimize the size of the vocabulary set.

However, the preceding experiments lay a solid framework for any future analysis of the headline’s dataset, and they also clearly illustrate the validity of topic modeling on an obviously unusual data format.

References:

1. Rathore, A.K.; Kar, A.K.; Ilavarasan, P.V. Social Media Analytics: Literature Review and Directions for Future Research. Decis. Anal. 2017, 14, 229–249. [Google Scholar] [CrossRef]

2. Mao, J.J.; Chung, A.; Benton, A.; Hill, S.; Ungar, L.; Leonard, C.E.; Hennessy, S.; Holmes, J.H. Online discussion of drug side effects and discontinuation among breast cancer survivors. Pharmacoepidemiol. Drug Saf. 2013, 22, 256–262. [Google Scholar] [CrossRef] [PubMed][Green Version]

3. Jaffe, S.: The collective power of# metoo. Dissent 65(2), 80–87 (2018)

4. PettyJohn, M.E., Muzzey, F.K., Maas, M.K., McCauley, H.L. # howiwillchange: Engaging men and boys in the# metoo movement. Psychology of Men & Masculinity (2018)

5. Soklaridis, S., Zahn, C., Kuper, A., Gillis, D., Taylor, V.H., Whitehead, C.: Men’s fear of mentoring in the# metoo era what’s at stake for academic medicine? (2018)

6. Kelly, L.: The (in) credible words of women: false allegations in European rape research. Violence Against Women 16(12), 1345–1355 (2010)

7. Ram, Y., Tribe, J., Biran, A.: Sexual harassment: overlooked and under-researched. Int. J. Contemp. Hosp. Manag. 28(10), 2110–2131 (2016)

8. Onwuachi-Willig, A.: What about# ustoo: the invisibility of race in the# metoo movement. Yale LJF 128, 105 (2018)

9. Manikonda, L., Beigi, G., Liu, H., Kambhampati, S.: Twitter for sparking a movement, Reddit for sharing the moment:# metoo through the lens of social media. arXiv preprint arXiv:1803.08022 (2018)

10. Prabhakaran, V., and Rambow, O. 2017. Dialog structure through the lens of gender, gender environment, and power. Dialogue & Discourse

11. Rho, E. H. R.; Mark, G.; and Mazmanian, M. 2018. Fostering civil discourse online: Linguistic behavior in comments of #MeToo articles across political perspectives. In CSCW.

12. Rashkin, H.; Singh, S.; and Choi, Y. 2016. Connotation frames: A data-driven investigation. In ACL.

13. Weiss, B. 2018. Aziz Ansari is guilty. Of not being a mind reader. The New York Times.

14. Way, K. 2018. I went on a date with Aziz Ansari. It turned into the worst night of my life. Babe.net.

15. Blei, D.M., A.Y. Ng, and M.I. Jordan, Latentdirichlet allocation. the Journal of machine Learningresearch, 2003. 3: p. 993-1022.

16. Feldman, R., & Sanger, J., The Text Mining Handbook, New York: Cambridge University Press, (2007).

17. Rifqi, N., Maharani, W., & Shaufiah., Analisis dan Implementasi Data Mining Menggunakan Jaringan Syaraf Tiruan dan Evolution Strategis. Konferensi Nasional Sistem dan Informatika, (2011).

18. Siti Qomariyah, Nur Iriawan and Kartika Fithriasari - Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis (2019)