TOPIC MODELLING ON CYBERSECURITY

BY

K NIKITHA REDDY

INTRODUCTION

Cybersecurity means protecting data, networks, programs and other information from unauthorized or unattended access, destruction or change. In today's world, cybersecurity is very important because of some security threats and cyber-attacks. For data protection, many companies develop software. This software protects the data. Cybersecurity is important because not only it helps to secure information but also our system from virus attack. After the U.S.A. and China, India has the highest number of internet users.

Figure 1: Cybersecurity

Cybersecurity is a growing concern in today's interconnected world. With the increasing dependence on technology and the internet in nearly every aspect of our lives, the need for strong cybersecurity measures is more important than ever. One of the primary concerns in cybersecurity is the protection of personal and sensitive information. With the proliferation of social media and online shopping, more and more personal information is being shared online. Hackers and cybercriminals can use this information for identity theft, financial fraud, and other crimes. To protect against this, it is important to use strong, unique passwords for all online accounts, to regularly update those passwords, and to be cautious about sharing personal information online.

Another major concern in cybersecurity is the protection of critical infrastructure. This includes power grids, transportation systems, and other vital infrastructure that is essential to the functioning of society. Hackers who gain access to these systems can cause significant damage and disruption. To protect against this, it is important for companies and organizations to invest in strong cybersecurity measures and to regularly update and test those measures to ensure their effectiveness.

In addition to personal and critical infrastructure protection, cybersecurity is also important for protecting businesses and organizations. Cyberattacks can result in the theft of sensitive information, the disruption of operations, and financial losses. To protect against this, businesses and organizations should have robust cybersecurity protocols in place, including the use of firewalls and other security measures, as well as employee education and training.

PROBLEM STATEMENT

Cybersecurity is a critical concern in today's world. It is important for individuals, businesses, and organizations to take steps to protect against cyber threats and to stay up to date on the latest cybersecurity best practices. By doing so, we can help ensure the safety and security of our personal information, critical infrastructure, and businesses. It becomes important to know what all researchers are doing to protect people from cyber threats. This mini project aims at identifying the most recent topics under Cybersecurity.

DATASET DESCRIPTION

The dataset is basically the collection of 45 articles overall in well-known research publications such as ScienceDirect, Research gate, IEEE, etc.. The data here collected is about the research articles related to Cyber Security to understand the different latest topics which comes under it. The articles used for data collection will be reflected under references.

METHODOLOGY

- **Text Preprocessing methods used**

  Text preprocessing is the process of preparing text data for natural language processing (NLP) tasks, such as text classification or sentiment analysis. There are several common text preprocessing methods that can be applied to raw text data:

  1. **Tokenization:** This involves dividing the text into smaller units called tokens, which can be individual words, phrases, or even characters. Tokenization helps to identify the structure of the text and make it easier to analyze.
  2. **Stopword removal:** Stopwords are common words that do not carry much meaning and are often removed from text data to reduce the size of the dataset and improve the efficiency of NLP algorithms. Examples of stopwords include articles (e.g., a, an, the) and prepositions (e.g., in, on, at). In stopword removal few extra words are added which doesn’t add any meaning to the analysis.

3. **Lemmatization:** This is similar to stemming, but instead of reducing words to their base form, lemmatization converts words to their base form, or lemma, taking into account the part of speech of the word. For example, the lemma of the word "jumping" would be "jump," while the lemma of the word "jumps" would be "jump."
4. **Lowercasing:** This involves converting all words in the text to lowercase in order to reduce the dimensionality of the dataset and simplify the analysis.
5. **Part-of-speech tagging:** This involves identifying the part of speech of each word in the text, such as noun, verb, adjective, etc. Part-of-speech tagging can be useful for identifying the structure and meaning of the text.

These are just a few of the many text preprocessing methods that can be applied to raw text data. The specific methods used will depend on the specific NLP task and the needs of the analysis.

- **Topic Modelling Algorithm used**

Latent Dirichlet Allocation (LDA)

LDA is a statistical model used for discovering the underlying topics in a collection of documents. It is a generative model that assumes that each document is a mixture of a fixed number of topics and that each word in the document is associated with one of the topics. LDA is based on the idea of a "latent" or hidden topic that is not directly observed, but can be inferred from the words in the document. The model estimates the probability of each word in the document being generated by each topic and uses this information to identify the most likely topic for each word.

One of the main advantages of LDA is its ability to handle large datasets and identify multiple topics within a single document. It is often used in text mining and information retrieval applications to automatically extract meaningful themes from large collections of documents. To use LDA, the number of topics must be specified in advance. Here the number of topics we specified is 4. The model then assigns each document to a set of topics and estimates the probability of each word in the document being generated by each topic. The resulting topics can be interpreted based on the most probable words for each topic, allowing for the automatic discovery of themes within the text data. It is a generative model that assumes that each document is a mixture of a fixed number of topics and that each word in the document is associated with one of the topics.

Here is a brief overview of how LDA works:

- The number of topics is specified in advance.
- For each article/document, the model estimates the probability of each word being generated by each topic.
- The model assigns each document to a set of topics based on the probabilities estimated in step 2.
- The model estimates the probability of each word being generated by each topic, taking into account the assignments of documents to topics in step 3.

The resulting topics can be interpreted based on the most probable words for each topic, allowing for the automatic discovery of themes within the text data. LDA is typically implemented using a variation of the Expectation-Maximization (EM) algorithm, which iteratively estimates the parameters of the model and refines the estimates until convergence. The EM algorithm is used to estimate the probability of each word being generated by each topic and the probability of each document being generated by each topic.

RESULTS AND DISCUSSION

Discovering Topics

The word cloud of the topics obtained are as follows:

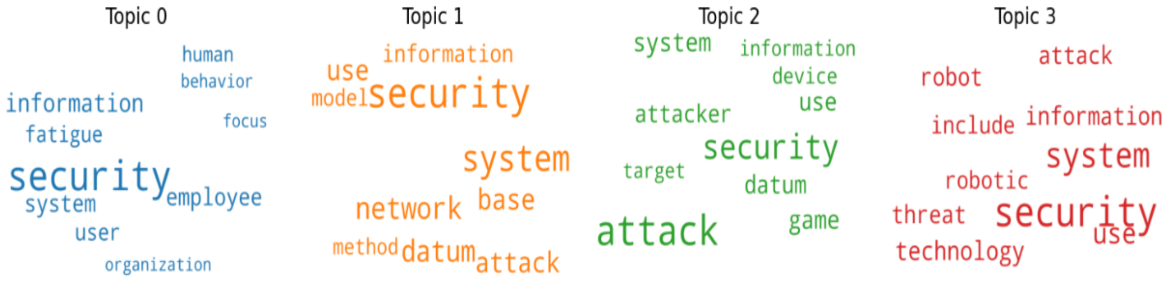
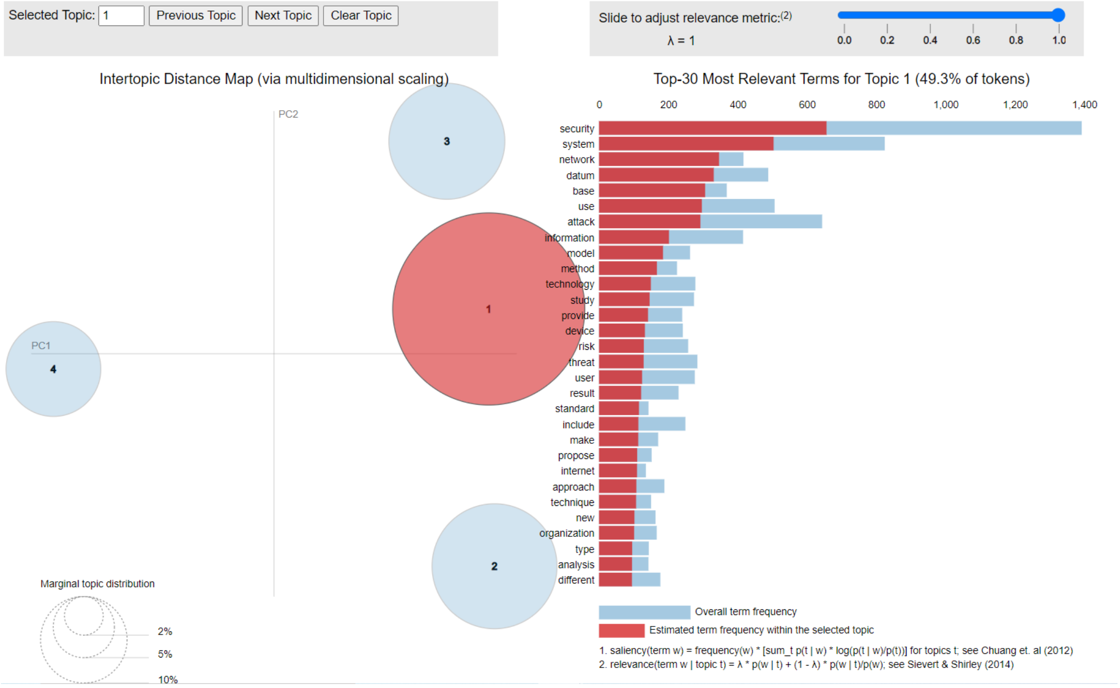


Figure 2: Wordcloud of Topics specified

Therefore, just obtaining the word cloud is not enough. We need to discover the topics that are there and we need to give a proper nomenclature by our human understanding. In order to do that we look at 30 high frequent occurring words by each topic through visualization and then justify why the topic is given that name.

Topic 1:

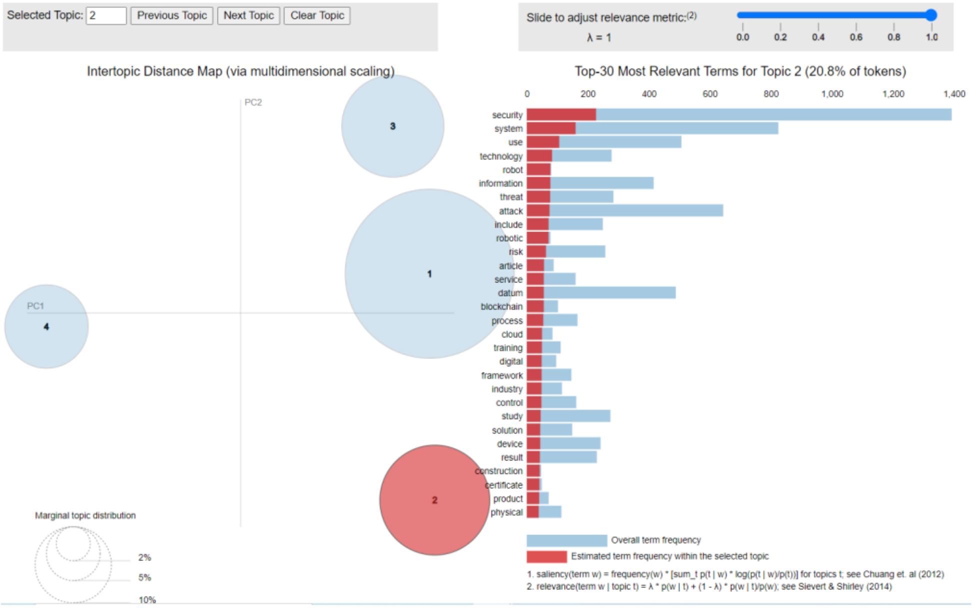


1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3: Topic 1

From the most relevant terms highlighted when we select on Topic 1 by using LDA, we see the common terms here are about system, security, network, technology, etc. So, the topic assigned to it under Cybersecurity would be **Network Security** because it involves implementing the hardware and software to secure a computer network from unauthorized access, intruders, attacks, disruption, and misuse. This security helps an organization to protect its assets against external and internal threats. Since this topic model also has words like user, threat, risk and network. This can be the topic assigned to Topic 1.

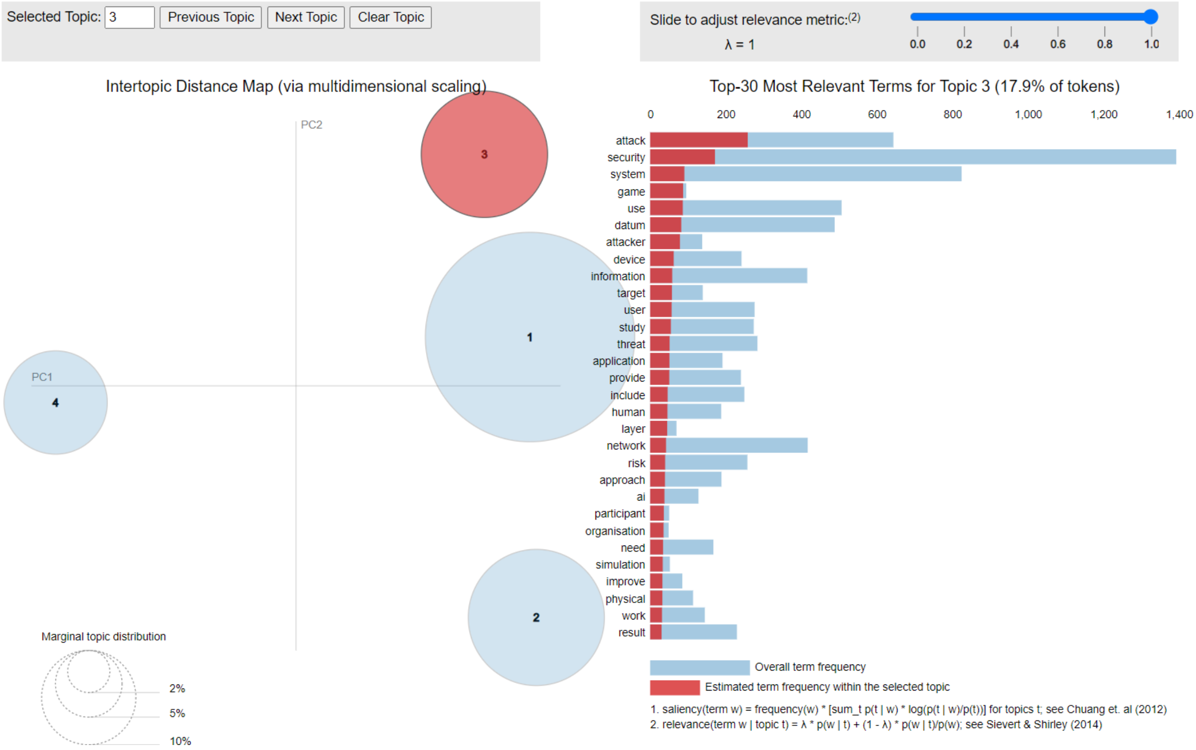
Topic 2:



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

From the most relevant terms highlighted when we select on Topic 2 by using LDA, we see the common terms here are about system, security, technology, robot, information etc. So, the topic assigned to it under Cybersecurity would be **Robotics Cyber Security** because it refers to the measures and practices that are implemented to protect robotic systems and devices from cyber threats and vulnerabilities. With the increasing use of robots in various industries, it is important to ensure that these systems are secure from cyber-attacks that could compromise their functionality or the safety of their users. Since this topic model also has words like attack, threat, risk, service and training. This can be the topic assigned to Topic 2.

Topic 3:



Selected Topic: 3

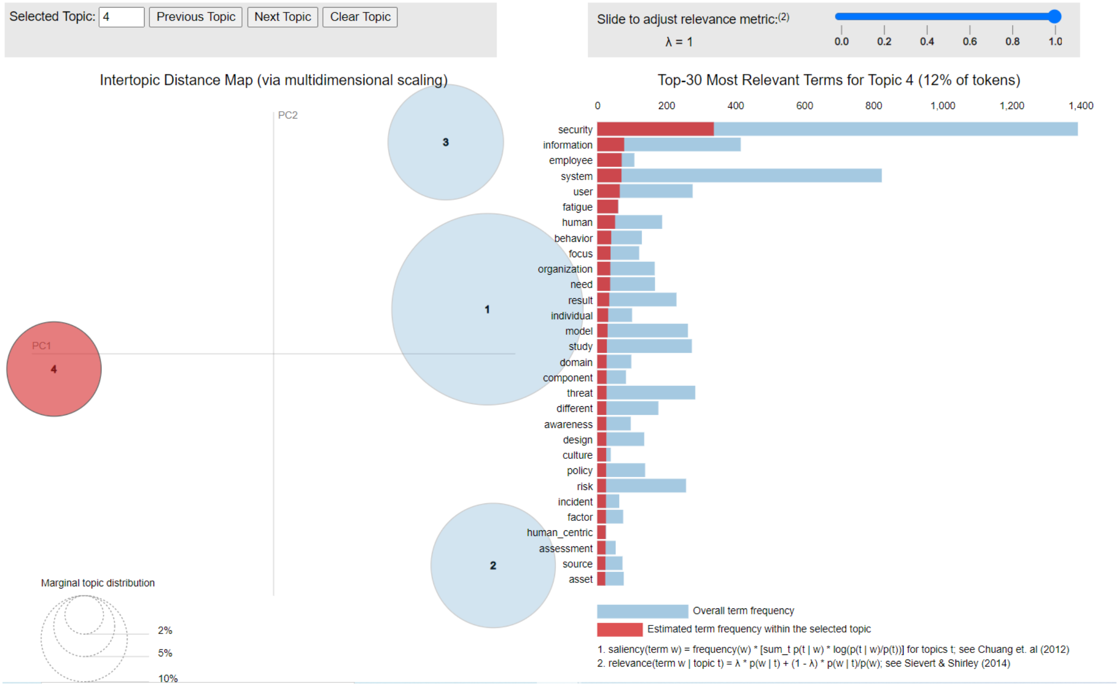
Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5: Topic 3

From the most relevant terms highlighted when we select on Topic 3 by using LDA, we see the common terms here are about attack, system, security, network, information, etc. So, the topic assigned to it under Cybersecurity would be **Application Security** because it involves protecting the software and devices from unwanted threats. This protection can be done by constantly updating the apps to ensure they are secure from attacks. Successful security begins in the design stage, writing source code, validation, threat modeling, etc., before a program or device is deployed. Since this topic model also has words like attacker, threat, approach, simulation and application. This can be the topic assigned to Topic 3.

Topic 4:



Selected Topic: 4

Slide to adjust relevance metric:(2)

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 6: Topic 4

From the most relevant terms highlighted when we select on Topic 4 by using LDA, we see the common terms here are about security, information, employee, user, etc. So, the topic assigned to it under Cybersecurity would be **Disaster Recovery and Business continuity Planning** because it deals with the processes, monitoring, alerts, and plans to how an organization responds when any malicious activity is causing the loss of operations or data. Its policies dictate resuming the lost operations after any disaster happens to the same operating capacity as before the event. Since this topic model also has words like user, human, threat, policy, and assessment. This can be the topic assigned to Topic 4.

So, hence the major four topics from the 45 articles collected are:

1. Network Security
2. Robotic Cybersecurity
3. Application security
4. Disaster Recovery and Business continuity planning

We can observe from the above diagrams that each topics are different from each other slightly in Cybersecurity since the bubbles are bit far and not too far.

CONCLUSION

Today due to high internet penetration, cybersecurity is one of the biggest needs of the world as cybersecurity threats are very dangerous to the country's security. Not only the government but also the citizens should spread awareness among the people to always update their system and network security settings and to the use proper anti-virus so that your system and network security settings stay virus and malware-free.

BIBLIOGRAPHY

https://www.sciencedirect.com/science/article/pii/S0142061517328946
https://link.springer.com/article/10.1007/s10462-021-09976-0
https://d197for5662m48.cloudfront.net/documents/publicationstatus/90321/preprint_pdf/bcff668d616b9c43ffde5be665cea136.pdf
https://link.springer.com/article/10.1007/s10207-021-00545-8
https://d197for5662m48.cloudfront.net/documents/publicationstatus/90291/preprint_pdf/c12f4b6dfcb0ece3a42a357ad2203fac.pdf
https://www.hindawi.com/journals/scn/2022/6200121/
https://d197for5662m48.cloudfront.net/documents/publicationstatus/90319/preprint_pdf/87c34d475885f4f2b553401959b483cb.pdf
https://www.sciencedirect.com/science/article/pii/S2352484721007289
https://www.mdpi.com/1207146
https://www.sciencedirect.com/science/article/pii/S2214785321016722
https://link.springer.com/article/10.1007/s42979-021-00557-0
https://www.sciencedirect.com/science/article/pii/S1574013721000010
https://www.mdpi.com/1099-4300/23/9/1112/htm
https://www.sciencedirect.com/science/article/pii/S0926580521004398
https://www.sciencedirect.com/science/article/pii/S0267364921000017
https://journals.sagepub.com/doi/pdf/10.1177/21582440211000049
https://link.springer.com/article/10.1007/s10639-022-11261-8
https://www.sciencedirect.com/science/article/pii/S2352484721007265
https://link.springer.com/article/10.1007/s43926-020-00001-4
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9562531
https://ieeexplore.ieee.org/abstract/document/9491638/
https://ibn.idsi.md/sites/default/files/imag_file/Cyber-Security-Threat-Analysis-In-Higher-Education-Institutions-As-A-Result-Of-Distance-Learning.pdf
https://www.mdpi.com/983864
https://www.sciencedirect.com/science/article/pii/S1319157821000203
https://www.mdpi.com/1150348
https://dl.acm.org/doi/abs/10.1145/3510410?casa_token=zP6G5y9cMs8AAAAA:yvx4xlCkWmq_wujTiWMkgy7Ae-qu5Uvrw_09qSZlraEsQ7I1wEsZegefTrwnFrhdi0eRgG_10fox_70
https://ieeexplore.ieee.org/abstract/document/9491691/
https://link.springer.com/article/10.1007/s42979-021-00535-6
https://www.mdpi.com/2079-9292/11/14/2181
https://link.springer.com/article/10.1007/s10796-021-10134-8
https://link.springer.com/article/10.1007/s10111-021-00683-y
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9353530
https://www.frontiersin.org/articles/10.3389/fdata.2021.583723/full#B16
https://scholarlypublications.universiteitleiden.nl/access/item%3A3247541/view
https://link.springer.com/chapter/10.1007/978-981-16-3246-4_55
https://academic.oup.com/cybersecurity/article-pdf/doi/10.1093/cybsec/tyab005/36597586/tyab005.pdf
https://www.sciencedirect.com/science/article/pii/S0167923621001093
https://www.mdpi.com/1060360
https://www.academia.edu/download/82132933/v1_n1_10.pdf