INFO 5731 Section 020 Computational Methods

GROUP-1

FIGHTING MISINFORMATION WITH NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING

**Instructor Name: HAIHUA CHEN**

**Date: 11/29/2022**

**Team Members**

1. Penumatsa Krishnamraju
2. Gujjula Madhubala
3. Kotha Sindhu
4. Chalasani Mounisha
5. Adulla Manisha

# Abstract

In recent years, the spreading of false information has become a concerning phenomenon, giving rise to the use of terms like misinformation, disinformation, and fake news. Although information professionals have been asked to take the lead in the battle against false news, existing efforts have been fragmented and ineffective due to a lack of a thorough grasp of the situation. Most information behaviour models focus on information, not false information.

Studying the issue of misinformation psychology processes (manipulation of the formation process), especially how social psychology is evolving — psychology manipulation in public forums. Take note of psychological phenomena like enduring involvement, polarization (echo cameras), motivated thinking, which affirms pre-existing ideas, normative influence, and negative group feelings. Along with theoretical debates, there are also studies in Russian and foreign psychology, sociology, politics, public relations, and marketing. There is discussion of how to stop internet astroturfing. Additionally, there are still a lot of psychological issues that need to be clarified in relation to enhanced informal psychology protection identity, such as the relationship between values and how one views false information, as well as issues with trust, anxiety, social identities, and social capital.

Table of content

[**Abstract**](#_76q1wa4baacr) **1**

[**1. Introduction**](#_ekni68byiv94) **2**

[**2. Literature review**](#_8lyzcukxz178) **2**

[**3. Research questions**](#_k21ccssn3kjw) **3**

[**4. Research methods**](#_7kin2grh5hf0) **4**

[4.1. Data collection](#_b8f7d6nuxi34) 4

[4.2. Data analysis](#_ur9phb6avku6) 4

[4.3. Sample size](#_dtqyt0y8daa5) 4

[**5. Experiment and Data analysis plan**](#_hbifyz0qwly)6

[**6. Contributions of each team member**](#_5dwxdp1rtvot) **7**

[**7. Conclusion**](#_98oeq1scj395)8

[**8. Reference page**](#_3ookg7ai6pf4) **8**

# 

# 

# 

# 

# 

# **1. Introduction**

As the field's own research topics have evolved over time, so has social science research. The classic analysis methods need to be revised and adjusted for the current era due to the emergence of social networks, new media, technological developments, and various forms of media consumption, among other changes.

The investigation of fake news and disinformation in its broadest sense, from its creation to the consequences it has on citizens, is one of the subjects currently generating the most attention in the scientific community. Many authors investigate what makes someone more or less susceptible to misinformation.

The study of fake news, false information, and disinformation has been a major focus of the field of library and information science (LIS). Fake news is one of the top trends in academic libraries, according to the Association of College and Research Libraries (2016). However, existing efforts are frequently deemed insufficient (Sullivan, 2018).

The terms fake news, disinformation, misinformation, etc., have been used frequently over the past few years. Disinformation, often known as fake news, has been the subject of intense controversy, especially in the wake of the 2016 US elections, at which time it spread widely on social media.

**2. Literature review**

The vocabulary and phenomena of false information phenomenon known as fake news appears anytime a new technological advancement becomes popular. For instance, during the Second World War, there was a high demand for radios, newspapers, and magazines as a result of new international communication broadcast technologies that released an overwhelming volume of information (Soll, 2016). German radio and broadcast journalism were used by the Nazi Party to propagate its beliefs. The party at the time used the abundance of information to create rumours regarding the murder of Jews (Soll, 2016). As a result, bogus news circulated widely and was exploited for political gain.

propagation of false information through social media The tremendous information overload brought on by the widespread usage of social media platforms as a source of everyday information is a difficult component of the fake news problem. People's information behaviour will include accessing information from a variety of synchronous and asynchronous sources and channels by the year 2020 (Agarwal et al., 2011; Agarwal, 2011), such as face-to-face or phone-based friends and contacts, email, messaging apps, audio/video calls, online information, and social media sites like Facebook, Twitter, Instagram, and LinkedIn.

The connection between fake news and LIS Since it can be challenging to determine the authority and credibility of information resources, LIS organizations and professionals can assist in educating users by facilitating discussions about trustworthy sources and critical thinking, as well as by updating information literacy instruction in light of the fake news problem (Jacobson, 2017; Creation, dissemination and mitigation 643 Kenney, 2017). Through information literacy (IL) education and research, LIS researchers and information professionals address the issue of recognizing and evaluating the dependability of information sources. The Association for College and Research Libraries is renowned for its work in developing IL frameworks and standards in order to assist LIS professionals in creating educational initiatives that will enhance people's capacity to find, assess, and use information successfully. Access to information becomes crucial in addition to information literacy.

Impact of fake news According to Bessi and Ferrara (2016), "the presence of social bots in online political discussion can create three tangible issues: first, influence can be redistributed across suspicious accounts that may be operated with malicious intentions; second, the political discussion can become even more polarized; and third, the spreading of misinformation and unverified information can be enhanced" (para.1). The phenomenon of filter bubbles is a significant contributor to the polarization of political discourse (Pariser, 2011). People have a list of in-person and online connections who "agree" with their worldview, political beliefs, mental models, and prejudices as well as particular news websites and social media accounts that they follow. Through the use of machine-learning models, search engines, news aggregators, and social networks are increasingly responsible for this "personalization" of content.

# **3. Research questions**

The following research questions will be addressed in this study:

1. How can we comprehend the phenomenon of false information?
2. What connection does it have to LIS?
3. How can information specialists combat it?
4. Gain knowledge of the conditioning elements that affect the potency of misinformation.
5. Examine the demographics of the most affected audiences to determine their weak points.
6. Make a list of potential steps that are carefully targeted and tailored to the various audiences found.
7. Identify crucial elements when developing solutions to lessen the effects of misinformation.
8. Determine whether education level is related to a higher propensity for misinformation.
9. Identify whether age plays a role in being more susceptible to misinformation.
10. Determine the frequency of the following other dependent variables of the subjects: sociodemographic factors, such as income level, social standing, and place of residence.

**4. Research methods**

4.1. Data collection

The data for this study will be collected from Google scholar based on author and title. Extract and identify the best methodology based off of given Abstract, Keywords and methodologies among the collected articles.

4.2. Data analysis

The data will be analysed to identify the different methodologies that are being expressed in relation to field of study.

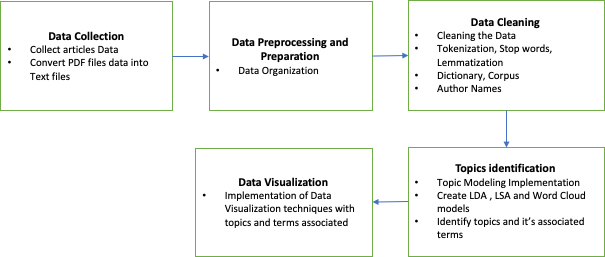
4.3. Sample size

A total of 1000 articles will be collected from Google scholar and extract Methodologies implemented, Abstract and Keywords.

# **5. Experiment and Data analysis plan**

Extract and identify the methodology addressed based off of various pdf articles for different field of study by using Python language Topic Modelling concepts (e.g., LDA and LSA). Data were pre-processed after data extraction in order to extract the terms and topics from the dataset before topic models were built. Several techniques were used for data pre-processing.

In order to comprehend the insights of the dataset, extract key variables, and find outliers and anomalies, exploratory data analysis (EDA) applies a variety of methodologies. EDA was used in this study to comprehend the dataset, find any null values in the columns, and clean the data by removing extraneous material. EDA was also used to examine the dataset and use visual approaches to highlight its key features.

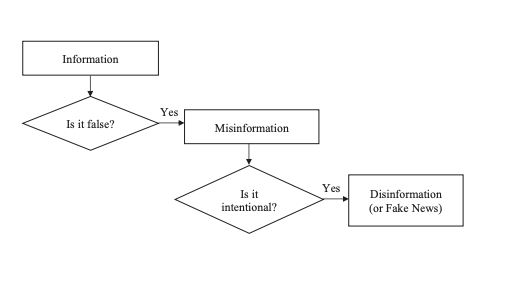


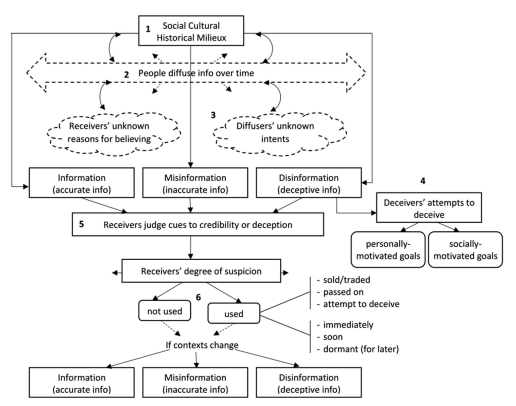
Research Design

1. Establishing and using stop words: Stop words are often used English words that don't add anything to a sentence's meaning. Therefore, these words were defined first, then Python was used to implement them. English terms like me, us, and will that be often used in texts and data were eliminated once the stop words were implemented. The stop words were defined and implemented using the Python library of the Natural Language Toolkit (NLTK).
2. Tokenization: Tokenization is a technique used to break up sentences into their component words. It is said of each word as a "token." Letter accents and punctuation are not used in this system, and the tokens are lowercase. Additionally, in this method, shorter tokens are ignored. Word tokenization, character tokenization, and sub word (n-grams) tokenization are the three types of tokenization. Word and sub word tokenization were done in this study.
3. N-gram implementation: This technique is used to extract recurring "n" word sequences from the corpus. Only single words are extracted during tokenization. Bigrams and trigrams, which are two words and three words in order, respectively, were retrieved and used in this study. (For instance, the bigrams network theory and complex network theory are trigrams.) The bigrams and trigrams used in this study were constructed using Gensim's phrases model.
4. Lemmatization: Lemmatization is the process of stripping tokens of their inflectional endings and reverting to the word's base or dictionary form. For instance, lemmatization on the word "using" causes the word to change from "using" to "use," where use is the word's base form.
5. Making a corpus and dictionary: A word dictionary and corpus were generated using the pre-processed data after the data had undergone pre-processing. The pre-processed data's unique words were all included in the word dictionary. A corpus that provided details on word frequencies was created.

Different topic modelling techniques were utilized to generate various topic models after pre-processing the text to find the keywords. To find abstract "themes" that appear in a group of documents or a text, topic modelling is a form of statistical model. One text-mining approach used to unearth latent semantic structures in a text body is topic modelling. It is expected that specific terms will appear in a text because it is about a particular topic. Topic modeling techniques generate "themes" that are collections of related words. This intuition is captured in a mathematical framework by a topic model, enabling the exploration and analysis of documents. Relevant subjects are found and their frequency is calculated using word data from each text. One or more subjects may be present in a text or document, and each topic may include one or more terms associated with it. The following list of topic modeling strategies was employed in this study:

1. Latent Dirichlet Allocation (LDA): LDA expects that a variety of themes be used to create documents. After then, words are generated from those themes depending on their probability distribution. Given a dataset of documents, LDA goes back and tries to determine what subjects would have initially generated those documents.
2. Latent Semantic Analysis (LSA): LSA aims to minimize classification-related dimensions. LSA believes that words with similar meanings will appear in texts with a similar structure (the distributional hypothesis). Singular value decomposition (SVD), a mathematical method, is used to condense the number of rows in a matrix storing word counts per document while maintaining the similarity structure between the columns. This model was created using the "Gensim" and "LsiModel" inbuilt packages, libraries, and functions. This LSA model was created following the standard method, which involved creating a word matrix, reducing the matrix, and identifying the subjects and their terms.
3. K-Means: An unsupervised learning algorithm is K-means. It starts out with a specific number of clusters. To reduce the sum of squares within a cluster, each observation is given a cluster assignment. The new cluster centroid is then determined by taking the mean of the clustered observations. Then, in an iterative process, data are redistributed to clusters and centroids are recalculated until the algorithm reaches convergence.





# 

# **6. Contributions of each team member**

The team members were also involved in the development of the individual sections of the proposal. Kirshamraju was involved in the development of the introduction, literature review, and research methods. Sindhu was involved in the development of literature review and research methods. Mounisha was involved in the development of the literature review. Madhubala was involved in the development of the research methods. Manisha was involved in the development of the research methods.

**7. Conclusion**

In conclusion, the false news pandemic is a multifaceted issue with many root causes. This issue is significant, and it cannot be solved by solitary efforts. A team effort that can address the political, social, technological, and human dimensions of the fake news problem is necessary to solve this challenge. LIS specialists can aid in the fight against false information by enlightening the public, creating awareness, and researching the information behaviour aspects of false information. To combat false news, all LIS experts in the US and around the world must band together and teach critical thinking to professionals and students. Fighting false information must be a key learning objective for academic programs.

**Reference**

Babbie, E. R. (2012). The Practice of Social Research(13th ed.). Wadsworth Publishing.Bensley, D. A., & Haynes, C. (1995). The Acquisition of General Purpose Strategic Knowledge for Argumentation. Teaching of Psychology, 22(1), 41–45. <https://doi.org/10.1207/s15328023top2201_13>

Collis, J., & Hussey, R. (2013). Business Research: A Practical Guide for Undergraduate and Postgraduate Students(4th ed.). Red Globe Press.

Cooper, D. R., & Schindler, P. S. (2013). Business Research Methods(12th ed.). McGraw-Hill Education.

Council of Europe. (2018). Dealing with propaganda, misinformation and fake news. Democratic Schools for All.

https://www.coe.int/en/web/campaign-free-to-speak-safe-to-learn/dealing-with-propaganda-misinformation-and-fake-news

Allied Joint Doctrine for Psychological Operations. AJP-3.10.1 [Elektronnyi resurs]. September, 2014. URL: <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/450521/20150223-AJP_3_10_1_PSYOPS_with_UK_Green_pages.pdf>. (Accessed 14.01.2018).

Bessi A. Personality traits and echo chambers on Facebook *Computers in Human Behavior.* 2016. Vol. 65, pp. 319—324. doi: 10.1016/j.chb.2016.08.016. URL: https:// dl.acm.org/citation.cfm?id=3043234 (Accessed 14.01.2018).

Bessi A., Petroni F., Del Vicario M. et al. Viral misinformation: The role of homophily and polarization *Proceedings of the 24th International Conference on World Wide Web Companion.* 18 May, 2015. pp. 355—356.

Cialdini R.B. Influence. Science and Practice [Elektronnyi resurs]. 4th edition, 2001. pp. 164—165 URL: <https://www.influenceatwork.com/wp-content/uploads/2012/02/Influence_SP.pdf> (Accessed 08.02.2018).

Del Vicario M., Bessi A., Zollo F., et al. The spreading of misinformation online [Elektronnyi resurs] *Proceedings of the National Academy of  Sciences.* 2016. Vol. 113,  no. 3, pp. 554—559. doi: 10.1073/pnas.1517441113. URL: <http://www.pnas.org/content/113/3/554> (Accessed 16.04.2018).

Ecker U.K.H., Lewandowsky S., Fenton O., Kelsey K. Do people keep believing because they want to? Preexisting attitudes and the continued influence of misinformation. *Memory & Cognition,* 2014. Vol. 42, pp. 292—304. doi: 10.3758/s13421-013-0358-x.

Edelman D.J. *Edelman Trust Barometer Executive Summary* [Elektronnyi resurs], 2015, pp. 5—6. URL: <https://ru.scribd.com/doc/252750985/2015-Edelman-Trust-Barometer-Executive-Summary> (Accessed 08.02.2018).

Frenda S.J., Nichols R.M., Loftus E.F. Current Issues and Advances in Misinformation Research. *Current Directions in Psychological Science.* 2011. Vol. 20, no. 1, pp. 20—23.

Harkins G.S., Szymanski K. Social Loafing and Self-Evaluation with a social standard. Journal of Personality and social psychology [Elektronnyi resurs] *Journal of Personality and social psychology.* 1987. Vol. 24, pp. 354—365. doi: <https://www.sciencedirect.com/science/article/pii/002210318890025X>. URL: <https://pdfs.semanticscholar.org/65c6/162ff35fc61ea7b59945858f1235e80cdbc3.pdf>  (Accessed 14.01.2018).

How does misinformation spread online? [Elektronnyi resurs] *World Economic Forum.* 2016. URL: <https://www.weforum.org/agenda/2016/01/q-a-walter-quattrociocchi-digital-wildfires/> (Accessed 14.01.2018).