Public Opinion on Student Loan Forgiveness

Arunaggiri Pandian Karunanidhi and Amaan Mujtaba Jaweed

Guided by Professor Jiebo Luo

CSC 440 DATA MINING FINAL PROJECT

December 11, 2022

Introduction

Background and Motivation

A Student loan is the money a student borrows from the federal government or private lender that would help them pay for their university. It is estimated that around 42.8 million people have student debt, and the total would amount to 1.77 trillion dollars (in 2022 December). An average student borrows around 35000 Dollars to attain a bachelor's degree. The benefits of student loans are that they are easily accessible by students, and it can help them finance their college education without paying much interest. On the other hand, student loans can become financially crippling as the borrower would have to typically make a payment of between 200$ and 300$ (the payment is higher if the loan is higher than the average amount!). As one can imagine, this would mean that the borrower would have to put his other financial goals aside, such as buying a house, or saving for retirement. In order to make the lives of these borrowers easier, Joe Biden has announced student loan forgiveness in which, the Biden administration is forgiving up to $10,000 in federal student loans for those making less than $125,000 a year for individuals or $250,000 for married couples or heads of households and up to $20,000 for Pell Grant recipients who meet the income threshold.

Since this announcement has been made, there has been a huge controversy going on regarding this. People have conflicting ideas regarding this announcement. The Economists are concerned about how the plan would affect the country's economy. Moreover, students who have recently paid off their loans are also against this as they have no perk from this plan while the students who are currently on debt and have taken loans for their education are welcoming the plan with open arms. On the other hand, the current students and students who have recently graduated that are in huge amounts of student debts are optimistic about the announcements and fully support it.

In the current study, we adopted a machine learning framework based on state-of-the-art transformer language models to capture individual opinions on student loan forgiveness, and categorize these opinions into three groups: positive, Neutral, and negative. We use more than 20000 rigorously selected tweets distinct Twitter users ranging from August to December of 2022. We extract and infer individual-level features, i.e., age and gender to characterize the opinion groups. From this study, we intend to answer the following questions after obtaining the results:

1) What is the public opinion on student loan forgiveness? And determing the percentage of people supporting it.
2) Determine what age group is likely to support the plan, and what age group is inclined to have negative opinions.
3) Determine the difference between opinions held by men and women and see if they are significantly different

Related Works

Hanjia Lyu et al., [1] attempted to evaluate public opinion on Twitter and analyze how it influences racially motivated hate crimes in the United States. Using hashtags associated with the Asian Hate Movement, tweets were gathered for this study, and statistics on crime were collected. Each tweets feature inference and topic encoding were completed. They discovered the differences in opinion across user attributes and its connection to hate crimes.

Hanjia Lyu et al., [2] adopted a human-guided machine learning framework using more than six million tweets from almost two million unique Twitter users to capture public opinions on the vaccines for SARS-CoV-2, classifying them into three groups: pro-vaccine, vaccine-hesitant, and anti-vaccine. They also determined how the opinions change over time, and how different states differ in their opinions regarding the vaccine. They also figured out a relationship between gender and the opinions they hold, and how a person's age would affect their opinion. They also performed multinomial logistic regression to predict the opinion on potential COVID-19 vaccines against demographics and other variables of interest.

Using information from Twitter Hanjia Lyu et al., [5], tried to understand how the general public felt about the Israel-Palestine Conflict. To measure the population biases across user types and demographics in describing the demographics behind the Black Lives Matter movement, Olteanu et al., [15] looked at the Black Lives Matter movement and hashtag on Twitter. In terms of Twitter users' geolocation and political ideologies, Badawy et al., [16] made analysis of the digital footprints of political manipulation connected to Russian meddling in 2016 took place.

Different online signals can be used to model stance on social media data. The two primary categories of these signals are content signals (such as the text of tweets) and network signals (such as connections and interactions between members on social networks).This was proposed in AL Dayel et al., [17]. Li et al., [18] showed that to identify position, sentiment might be utilized as a stand-in. Because of how recent this incident is, there has not been a systematic analysis of it in the past. Newspaper articles on the subject have taken positions over the course of the time. But because that is just the raw data, no analysis in the traditional sense has been done on the subject.

Materials and Methods

The Methods section is structured as follows. We describe the datasets we use in Methods M1 and how we extract age and gender features in Methods M2. We describe our strategy for sentiment analysis in Methods M3, and topic modelling using LDA in M4.

1) M1 Dataset:

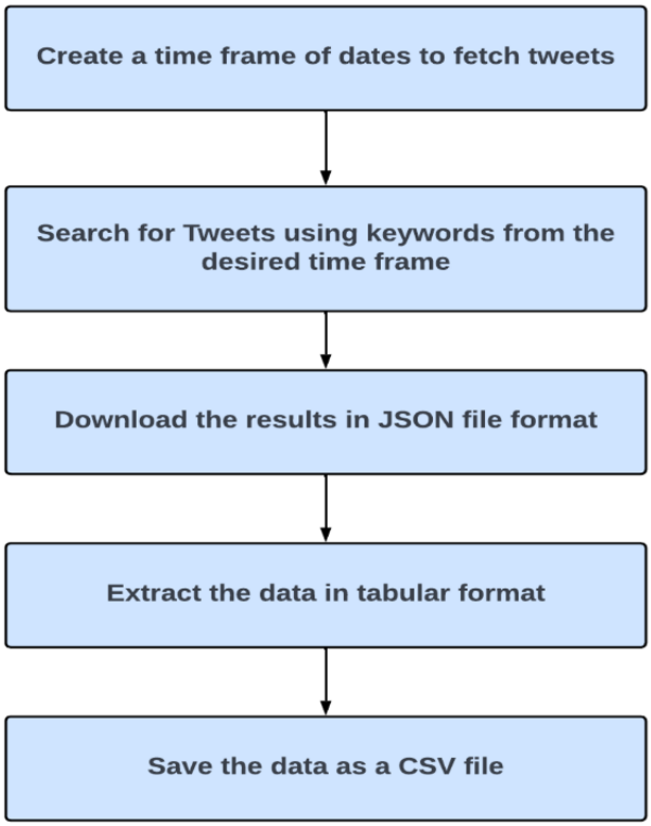

Figure 1: Data collection

We use the snscrape scrapper to collect the related tweets which are publicly available. The search keywords and hashtags are student loan forgiveness-related including "studentloan", "studentloanforgiveness", "studentloandebt", "studentloanbomb", "studentdebt", "loanforgiveness" It is noteworthy that the capitalization of non-hastag keywords does not matter in the query. In the end, 20267 tweets from August 14 to December 8th, 2022, are collected.

Along with the tweet, we also extract twitter profile information, such as twitter user id and twitter profile picture url which would help us determine the age and gender of the twitter user.

2) M2: Feature Inference

Following the methods of Lyu et al. [29], we used the pretrained deep learning model for age and gender prediction trained using Caffe model and OpenCV [2] to infer the gender and age information of the users using their profile images.

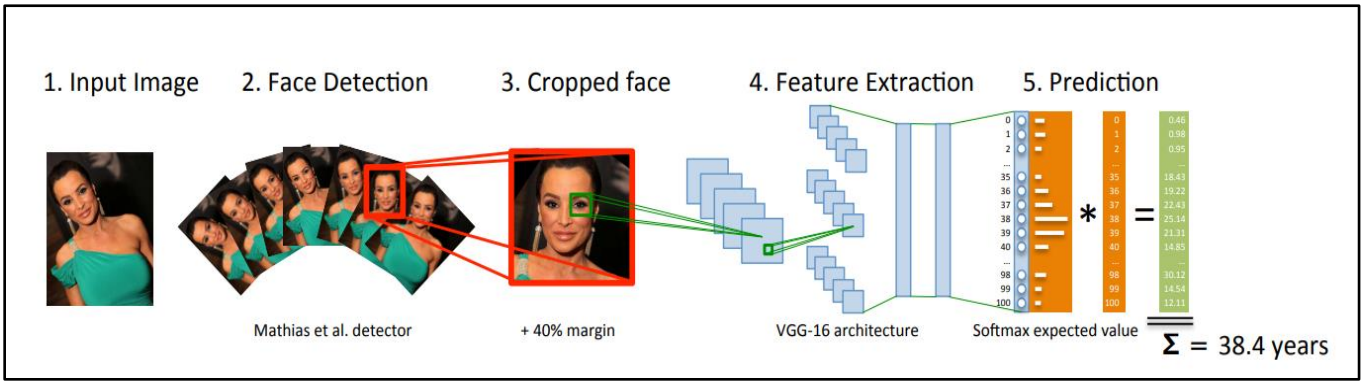

Figure 2: Age and Gender predictor

Figure 3: Age and gender predictor examples

The predictor worked well with 97.5% and thus, we use it in our study to predict the age and gender of our twitter users. The invalid image urls and images with multiple or zero faces are excluded. The gender and age information of the remaining users (i.e., there is only one intelligible face in the profile image) is inferred. Since our study focuses on the opinions of United States adults, the users who are younger than 18 are removed.

3) M3 Sentiment Analysis

A) Data cleaning and data preprocessing

Firstly, we clean the tweets, and remove the tweets that we cannot use in our analysis. Tweets to be removed:

i. Rows in which any of the values are missing
ii. Rows for which there are more than 1 face in the profile picture
iii. Rows in which there are no faces in the profile picture.
iv. Rows in which tweet becomes an empty string after data preprocessing

Secondly, in the data preprocessing, we perform the steps shown below. We used regular expression and NLTK package in python to do so. We remove the hashtags, urls,

punctuations etc. using regular expressions from the tweets as it was interfering with sentiment analysis. We remove stop words and take care of lower-casing using NLTK.

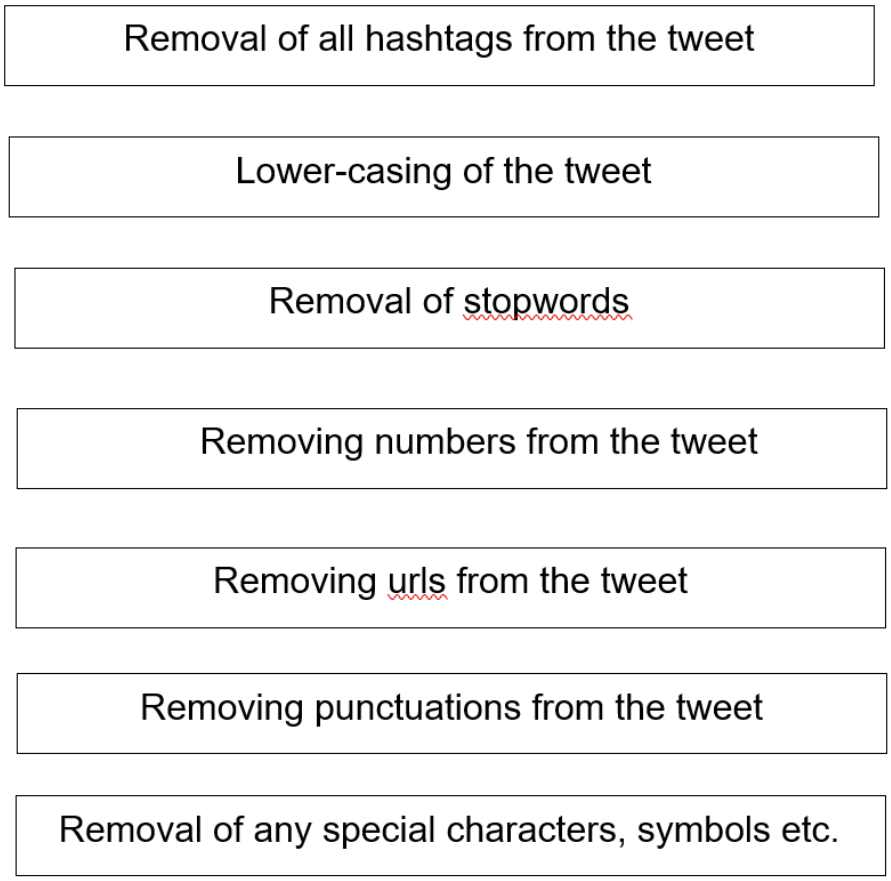

Figure 4: Data Preprocessing Tasks

B) Sentiment analysis using VADER

For sentiment analysis, we use VADER sentiment analyzer from nltk package. VADER is a lexicon and rule-based sentiment analysis tool built on top of LIWC and ANEW, that is specifically attuned to sentiments expressed in social media. VADER not only tells us if the sentiment is positive or negative, but it also tells us how positive or negative a sentiment is.

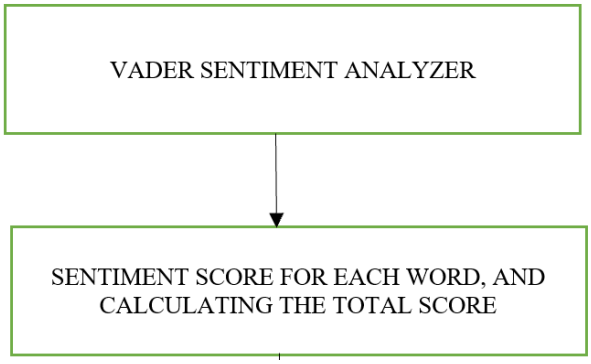

Figure 5: VADER Sentiment Analyzer

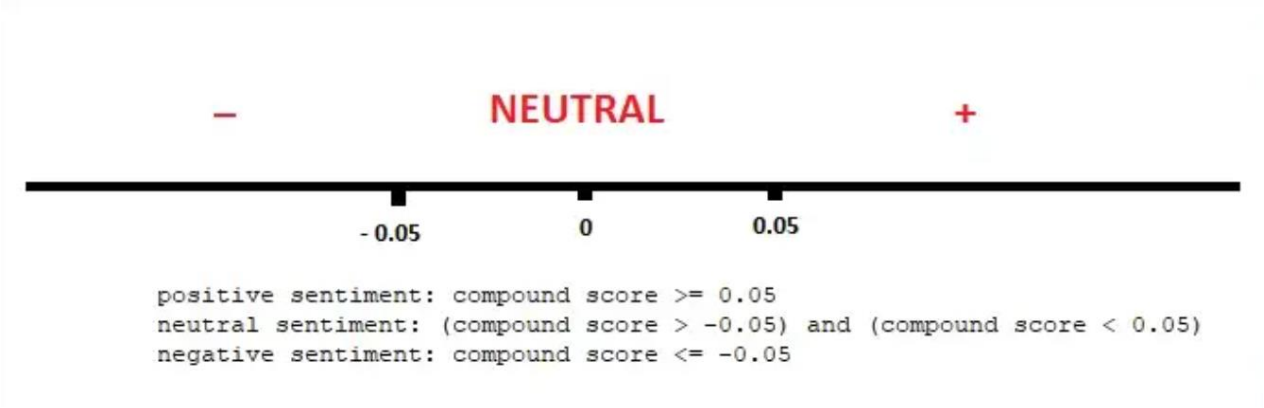

Figure 6: VADER CLASSIFIER-Scoring

VADER takes in a tweet as its input and gives a compound score to it, if this score is greater than 0.05, we label the tweet as positive. If the score is between -0.05 and 0.05, we label the tweet as neutral, and if the score is less than -0.05, we label the tweet as negative.

4) M4: Topic Modeling

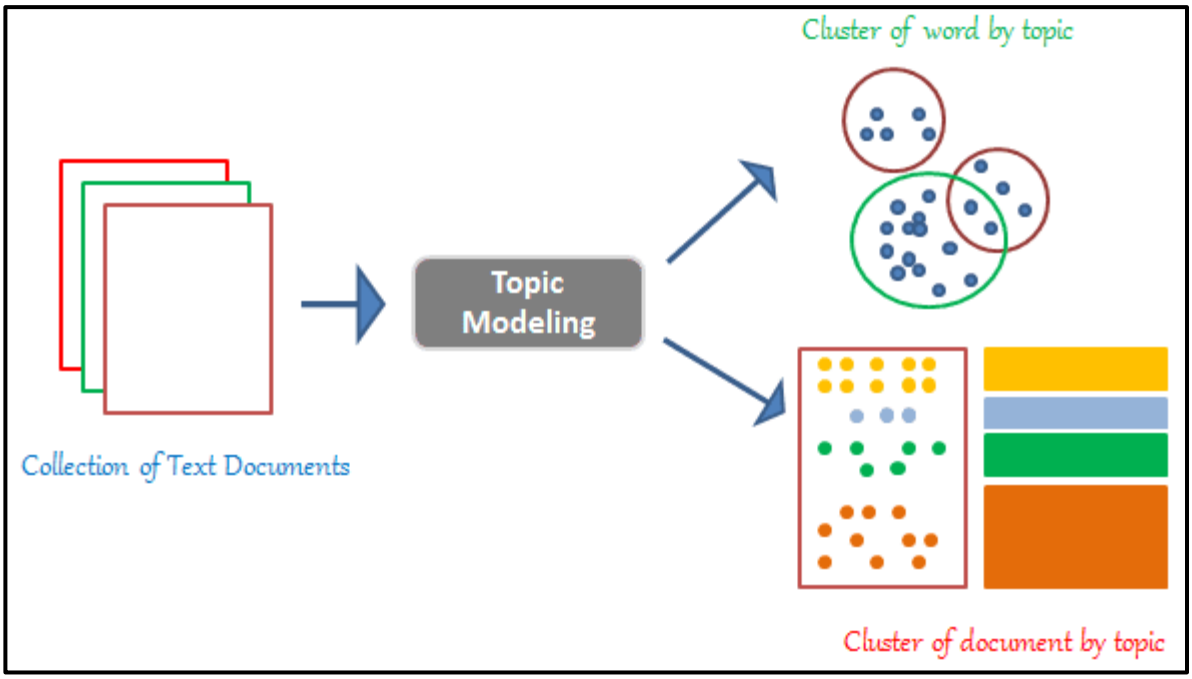

Figure 7: Topic Modeling

Latent Dirichlet Allocation (LDA) is a generative probabilistic model in which a varying proportion of themes are considered to make up each text in the model.

Steps involved in LDA:

1. For each cleaned tweet, randomly assign each word to a topic where we have K number of pre-defined topics.
2. For each tweet tw:
   For each word w in tweet tw, compute:
   1. P(topic T | tweet tw): Proportion of words in tweet assigned to topic
   2. P(word w| topic T): Proportion of assignments to topic across all tweets from words that come from w
3. Reassign topic T' to word w with probability p(t' | tw) * p(w | t') considering all other words and their assignments.
4. Repeat step 3 multiple times until the topic assignments do not change further. The proportion of topic for each tweet is then determined from these topic assignments.

Consider four tweets given below as the corpus

Tweet 1: I watch YouTube videos while eating.

Tweet 2: YouTube videos helps a lot in visual learning.

Tweet 3: I can understand complex concepts easily by reading some online technical blog

Tweet 4: I am a visual learner, so I prefer YouTube videos to blogs.

We can assign topic mixes to each of the documents and find subjects in the corpus mentioned above with the aid of LDA modeling. The model might provide anything like what is shown below as an example.

Topic 1: 40% videos, 60% YouTube

Topic 2: 95% blogs, 5% YouTube

Along with the Mallet's implementation, we utilized the Latent Dirichlet Allocation (LDA) from the Gensim package (via Gensim). Mallet has a successful LDA implementation. It is reputed to operate more quickly and provide greater topic separation.

To determine how important a topic is, we will also extract the volume and percentage contribution of each topic.

i) Positive Tweets

Topics identified in Positive Tweets could be classified as:

Topic – 1: Benefits. **Topic – 2**: Relief **Topic – 3**: Money

ii) Negative Tweets

Topics identified in Negative Tweets could be classified as:

Topic – 1: Borrower. **Topic – 2**: Plan. **Topic – 3**: Debt

iii) Neutral Tweets

Topics identified in Neutral Tweets could be classified as:

Topic – 1: People**. Topic – 2**: Application**. Topic – 3**: Education

Word Cloud of Top 10 Keywords in each Topic

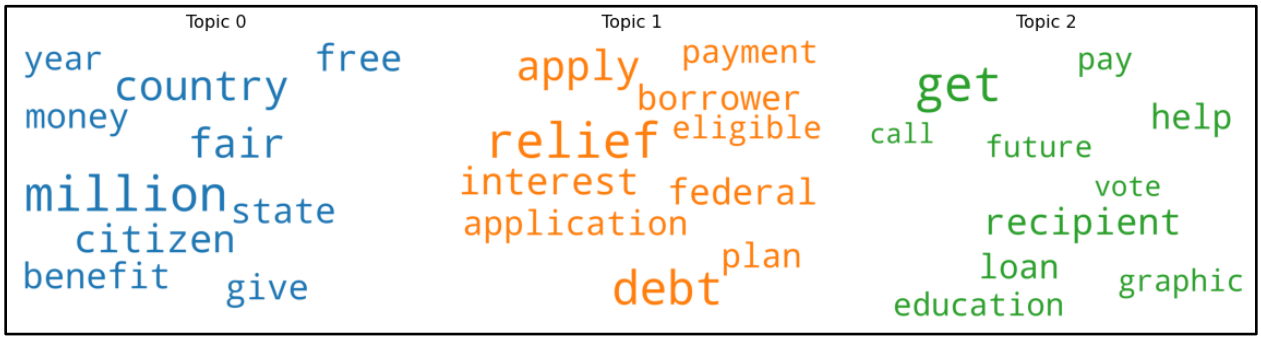

Figure 8: Word clouds for Positive Tweets

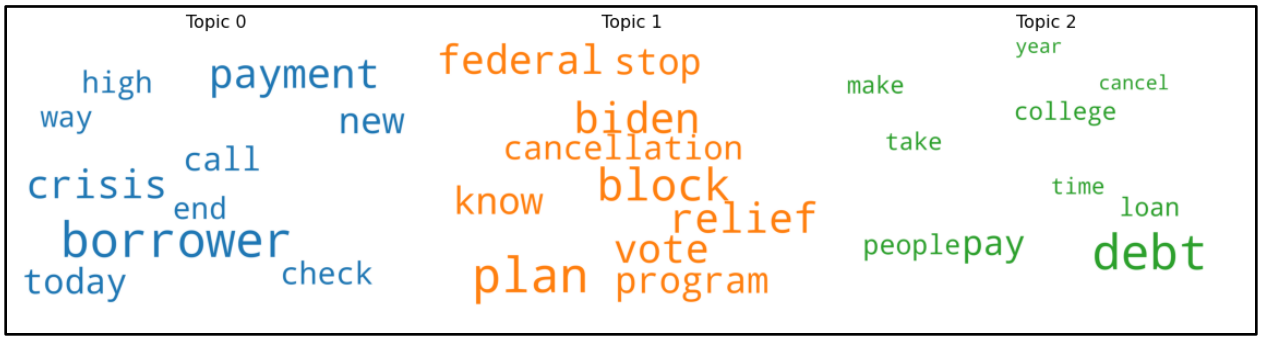

Figure 9: Word Clouds for Negative Tweets

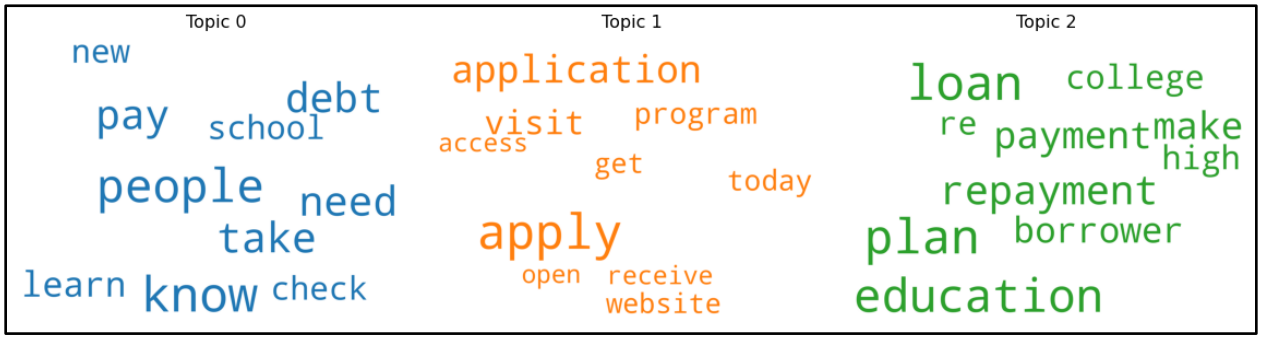

Figure 10: Word Clouds for Neutral Tweets

Word Counts of Topic Keywords

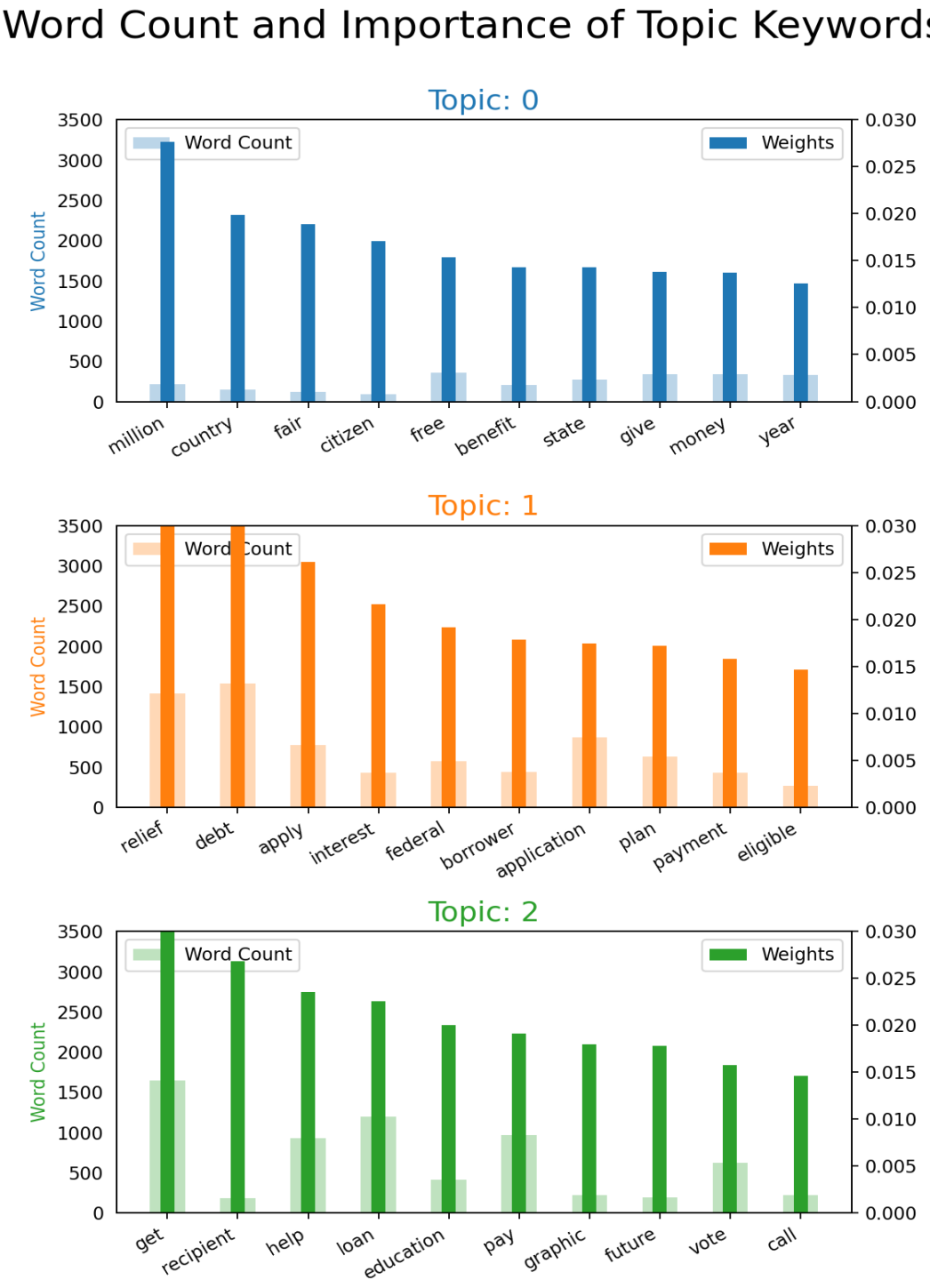

Figure 11: Importance of topic keywords (Positive)

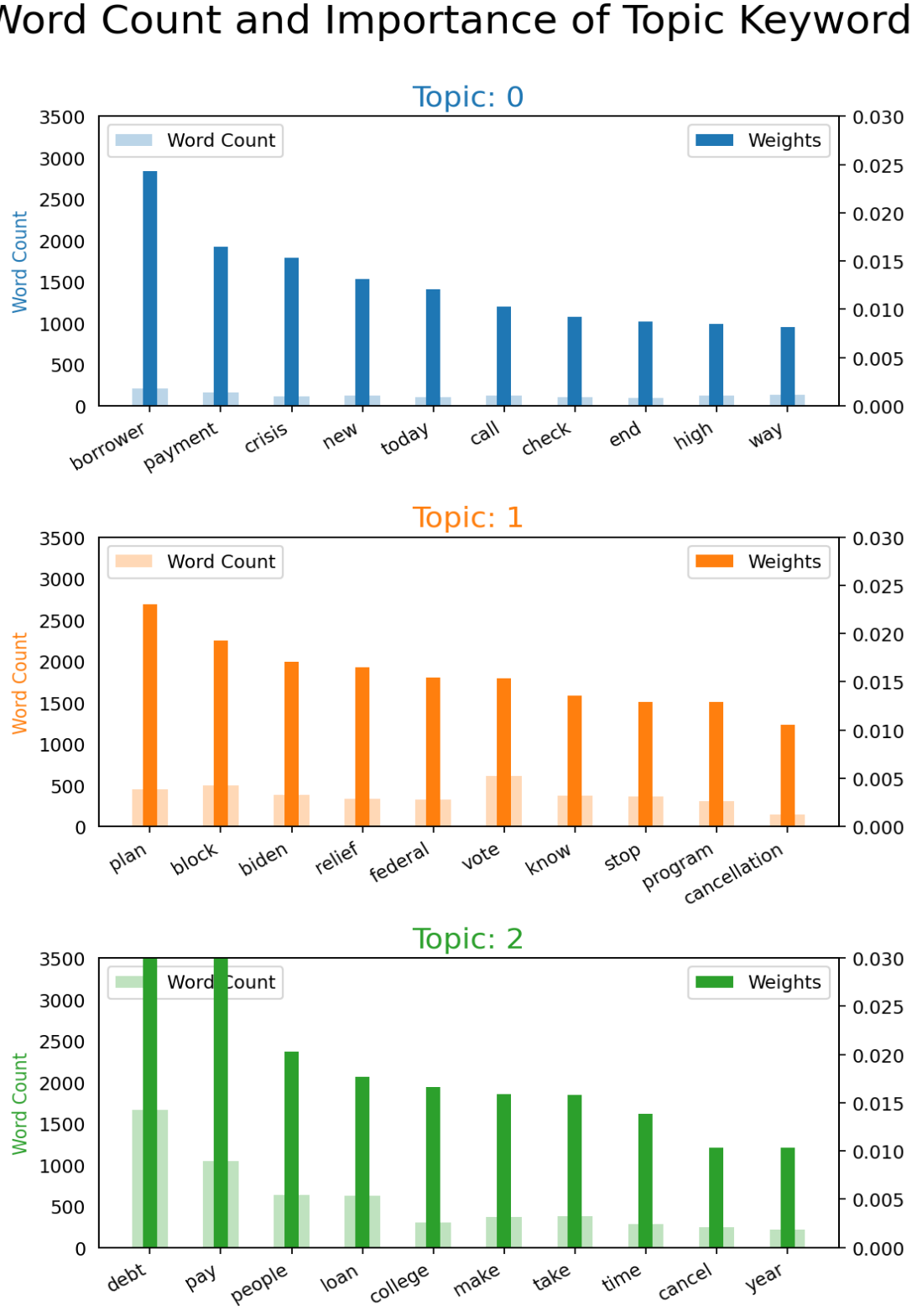

Figure 12: Importance of topic keywords (Negative)

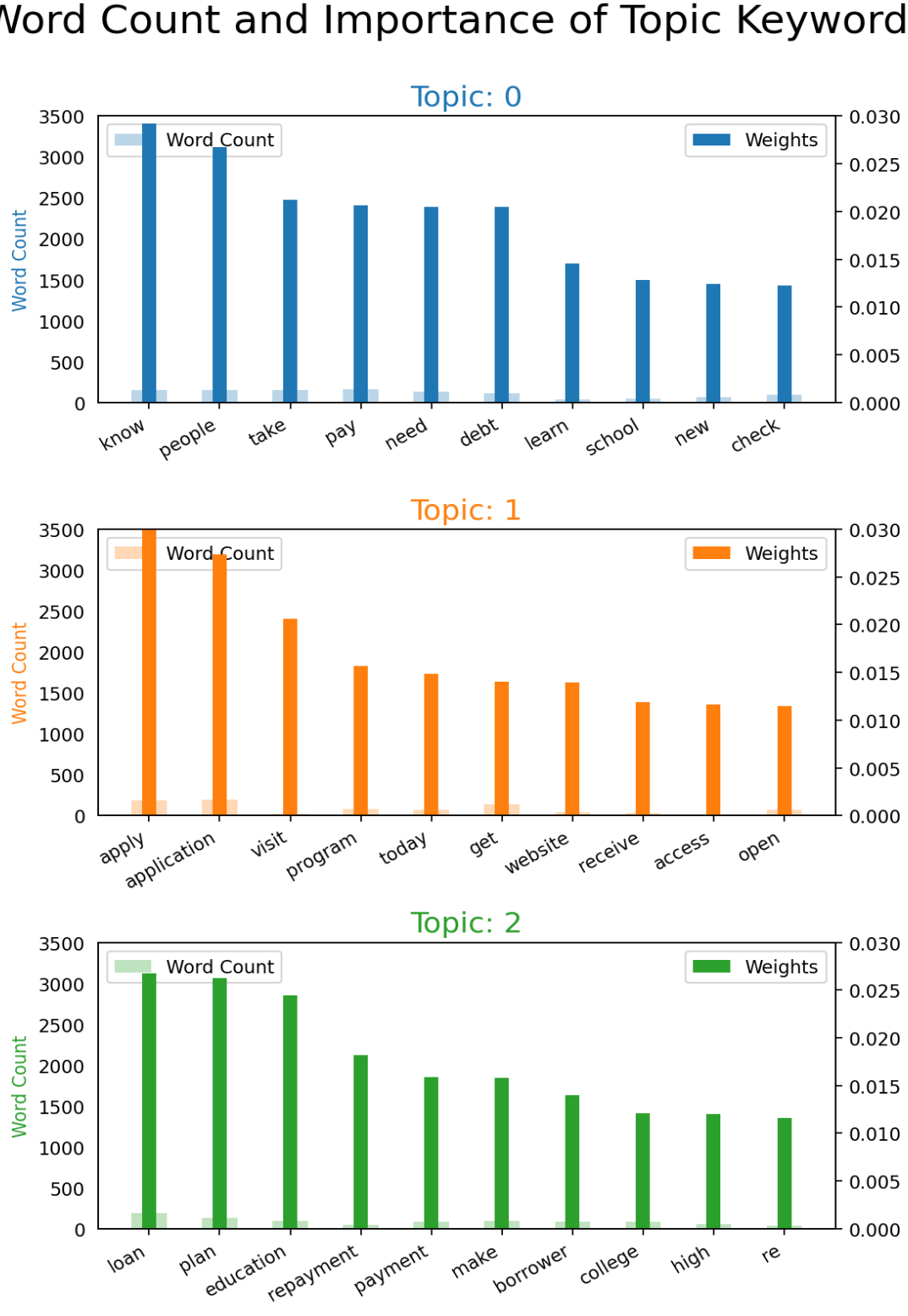

Figure 13: Importance of topic keywords (Neutral)

Sentence Chart Colored by Topic

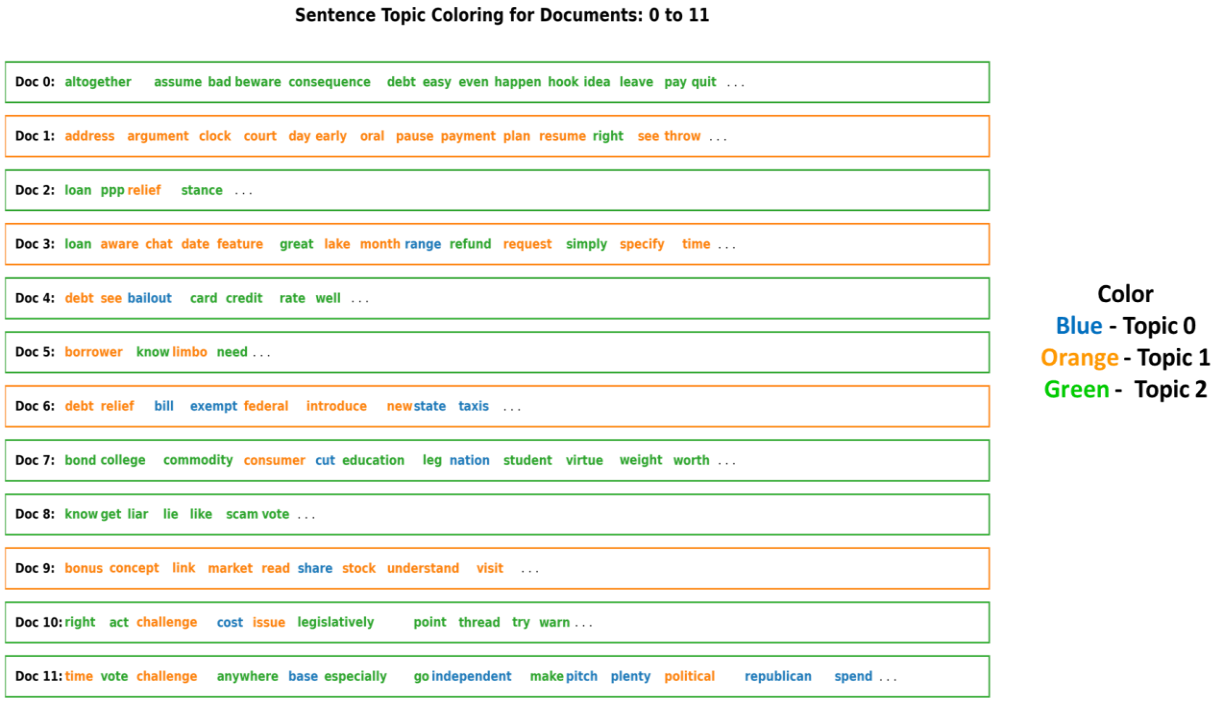

Figure 14: Sentence Chart Colored by Topic (Positive)

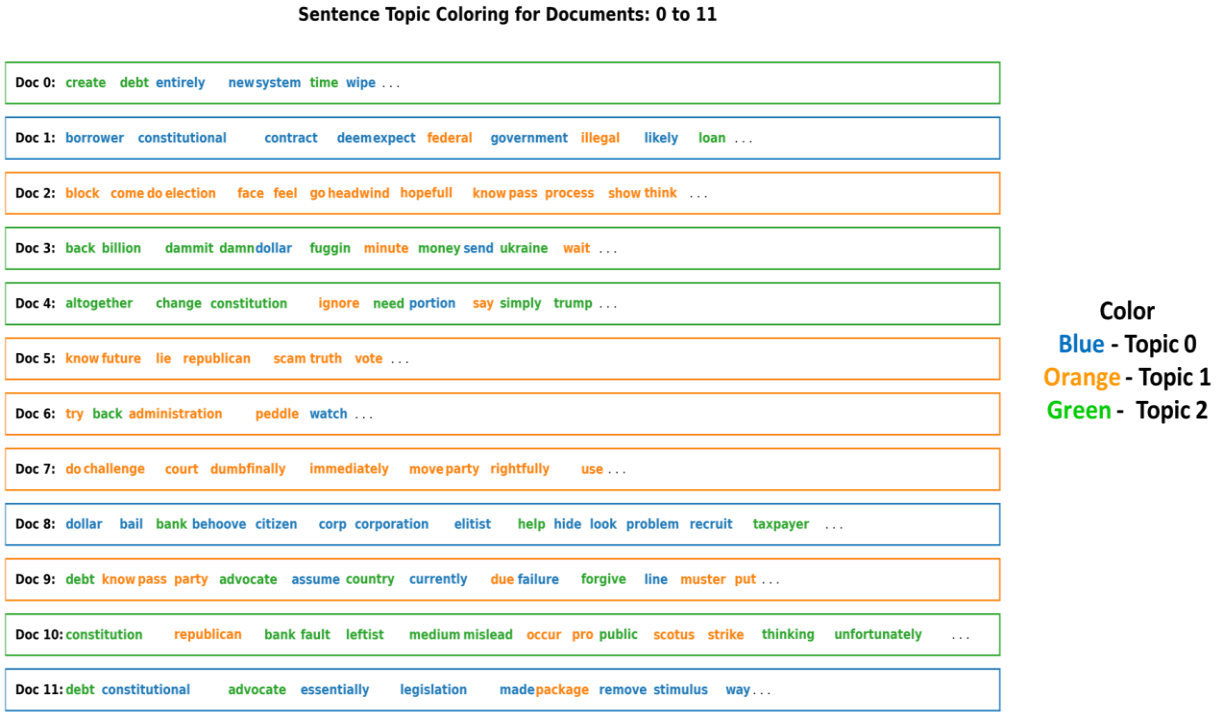

Figure 15: Sentence Chart Colored by Topic (Negative)

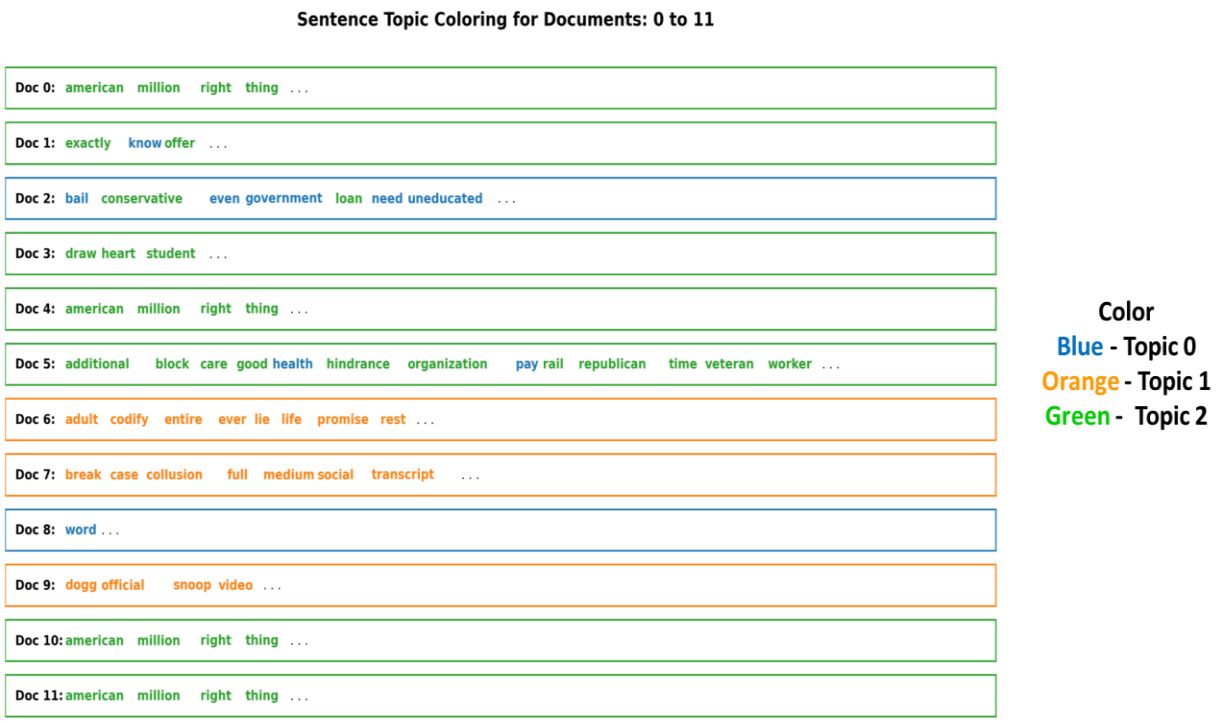

Figure 16: Sentence Chart Colored by Topic (Neutral)

Most Discussed Topics in the Document

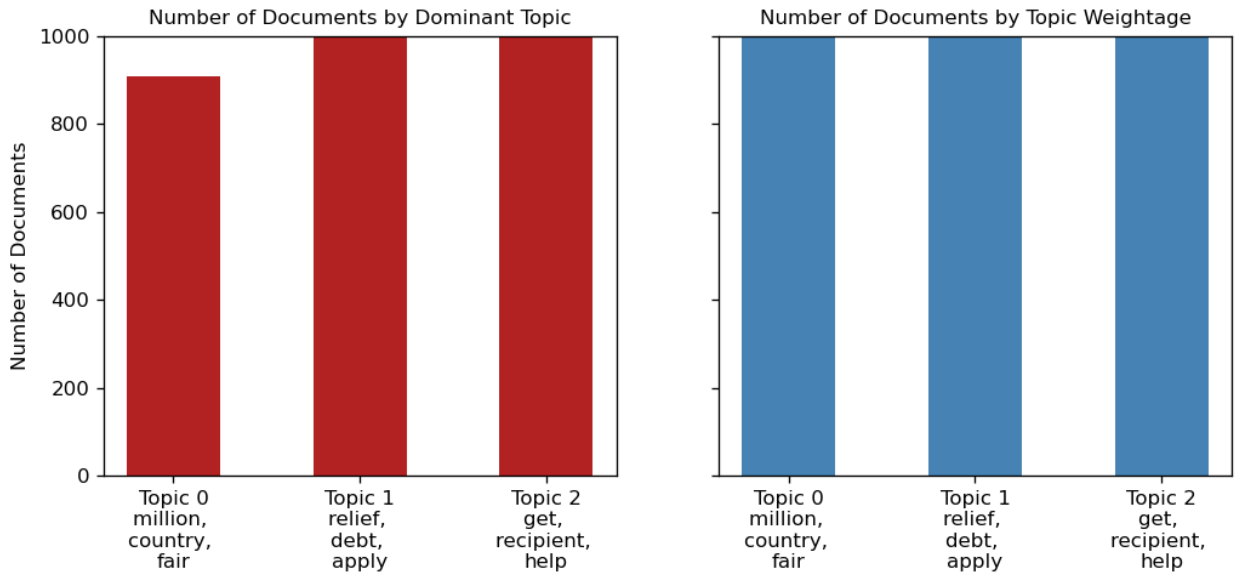

Figure 17: Most Discussed Topics in the Document (Positive)

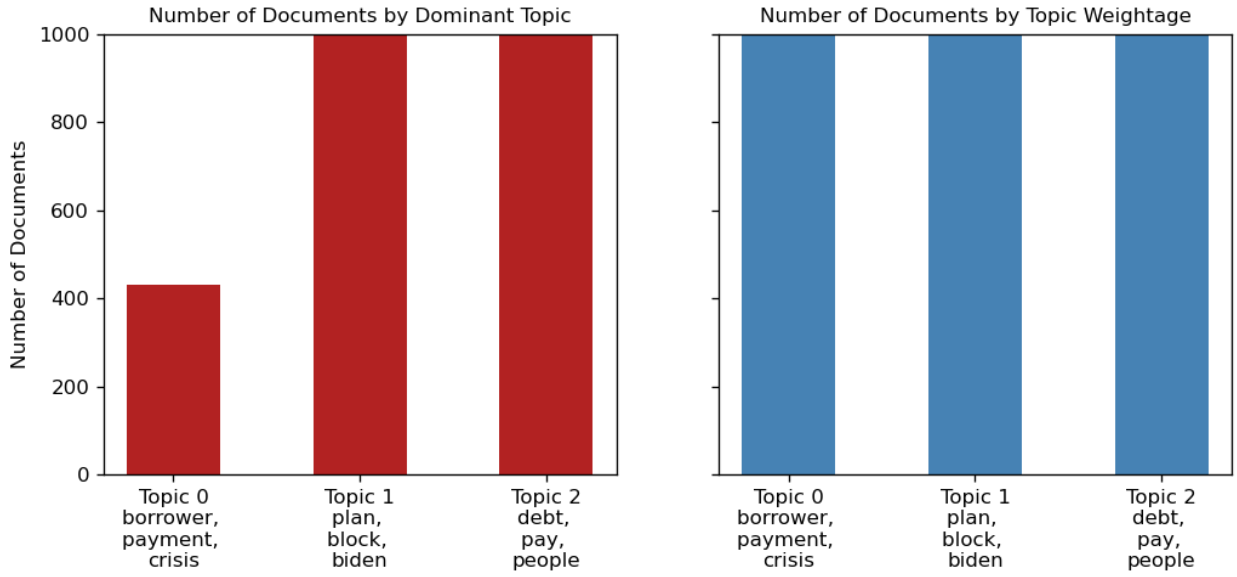

Figure 18: Most Discussed Topics in the Document (Negative)

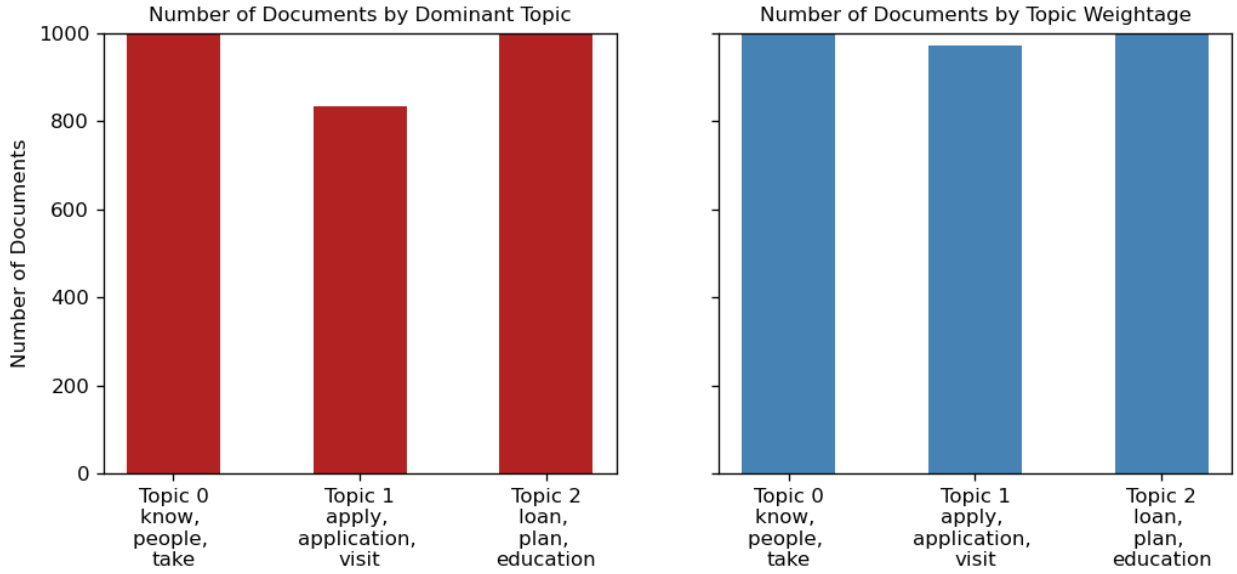

Figure 19: Most Discussed Topics in the Document (Neutral)

Clusters of Documents using t-SNE Algorithm

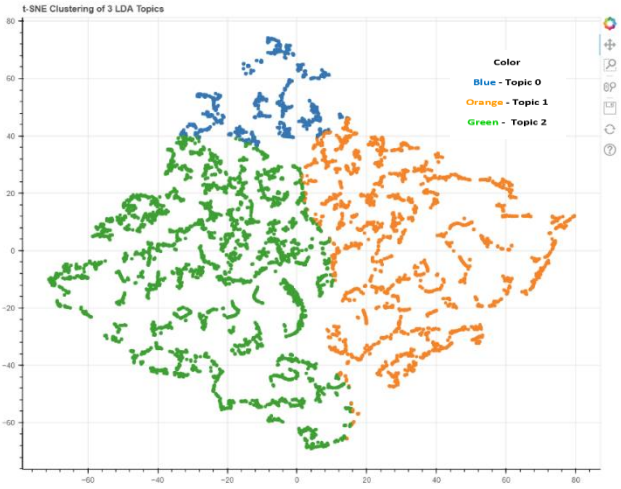

Figure 20:Clusters of Documents (Positive)

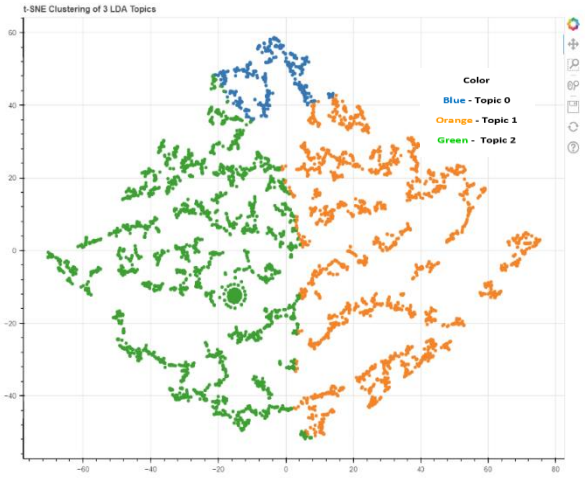

Figure 21:Clusters of Documents (Negative)

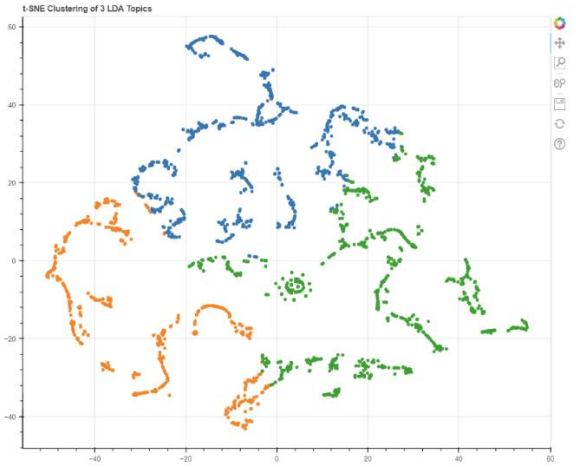

Figure 22: Clusters of Documents (Neutral)

Results

After sentiment analysis and topic modelling, we get word clouds for positive tweets, neutral tweets, and negative tweets.

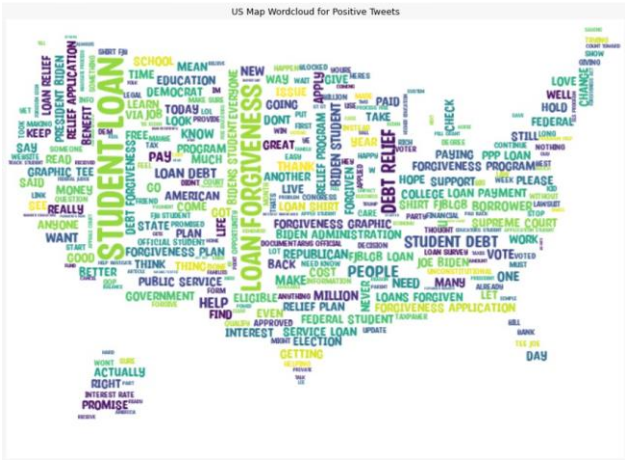

Figure 23: Word cloud for positive tweets

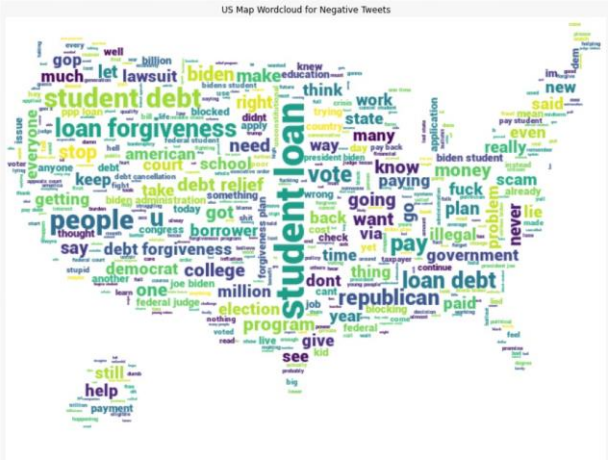

Figure 24: Word clouds for negative tweets

Figure 25: Word clouds for neutral tweets

We can also visualize the distribution of the tweets across each sentiment as a bar chart

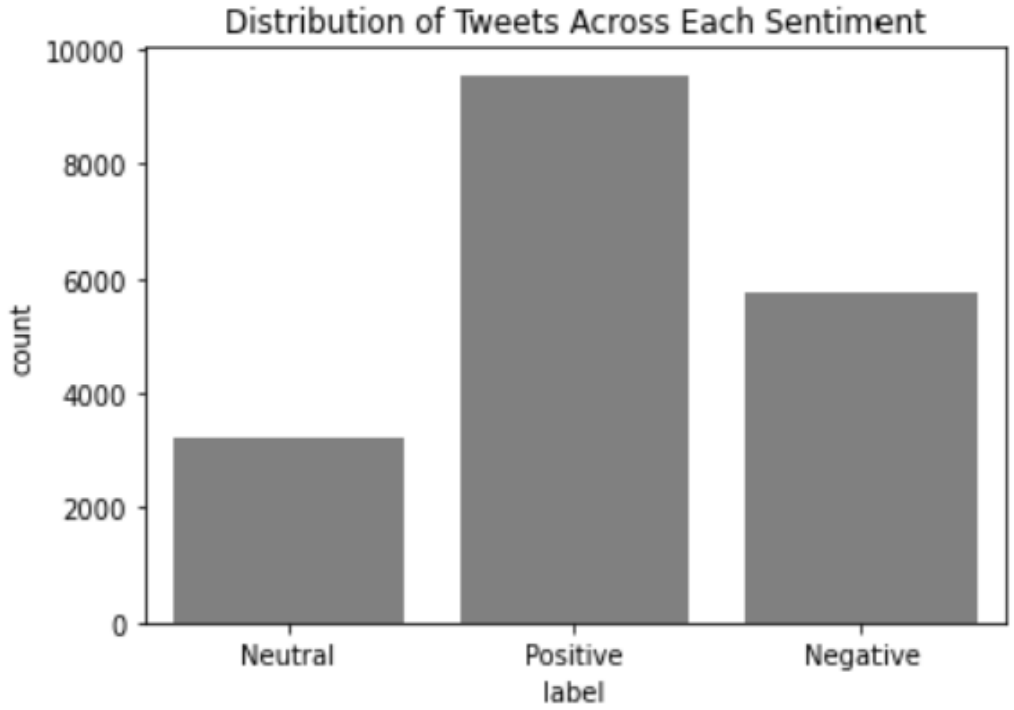

Figure 26: Distribution of tweets across sentiments

We can also determine the percentage of tweets with each sentiment as a pie chart

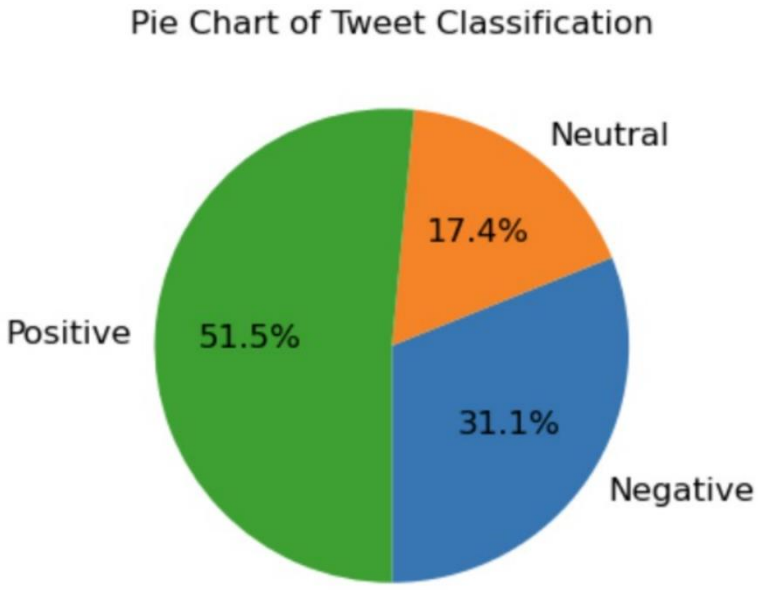

Figure 27: Tweet Classification

As we can see from the chart, the percentage of positive tweets is the maximum, with 51.5%, which means that majority of the tweets were in support of the student loan forgiveness. This means that most people have positive opinions towards the announcement and want it to go forward. This is predicted as most of the users in twitter are students/recently graduated students and they would be in support of it.

In contrast, 31.1% of total tweets taken were labelled as negative, which means that they are against the announcement. This would mostly be people who have already paid off the student debt or members who oppose Biden. And a small number of tweets (17.4% of total tweets) are labelled as neutral.

Further Research

In this study, we were successfully able to determine the opinion of the general public on student loan forgiveness announcement by Biden Administration by doing sentiment analysis using VADER and Topic modelling using LDA. We have a rough idea as to how public opinion changes with age, however, at this point of our study, we do not have enough data to draw correlations between age and gender with their opinion. This can be done when enough time has passed, and we can obtain sufficient amount of data to draw correct conclusions.

Additionally, another idea for future research would be to determine how the opinion varies according to the state. This could help us determine which state would have more people with positive opinion and which states have people with more negative opinion. We could also extend this particular research to find relations between opinions of the people and their political affiliations.

Finally, further research could be done to figure out the reasons for the opinion of people and what are the factors responsible for swaying their opinions.

References

[1] Hanjia Lyu, Yangxin Fan, Ziyu Xiong, Mayya Komisarchik, and Jiebo Luo, "Understanding Public Opinion toward the #StopAsianHate Movement and the Relation with Racially Motivated Hate Crimes in the US," IEEE Transactions on Computational Social Systems.

[2] Hanjia Lyu, Junda Wang, Wei Wu, Viet Duong, Xiyang Zhang, Timothy D. Dye, and Jiebo Luo, "Social Media Study of Public Opinions on Potential COVID-19 Vaccines: Informing Dissent, Disparities, and Dissemination," Intelligent Medicine, 2021.

[3] Nazan Öztürk, Serkan Ayvaz,Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis, Telematics and Informatics, Volume 35, Issue 1,2018, Pages 136-147, ISSN 0736-5853, https://doi.org/10.1016/j.tele.2017.10.006

[4] Sharif, W.; Mumtaz, S.; Shafiq, Z.; Riaz, O.; Ali, T.; Husnain, M.; Choi, G.S. An Empirical Approach for Extreme Behavior Identification through Tweets Using Machine Learning. Appl. Sci. 2019, 9, 3723.https://doi.org/10.3390/app9183723

[5] S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data," 2019 International Conference on Computer and Information Sciences (ICCIS), 2019, pp. 1-4, doi: 10.1109/ICCISci.2019.8716464.

[6] V. Prakruthi, D. Sindhu and D. S. Anupama Kumar, "Real Time Sentiment Analysis Of Twitter Posts," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), 2018, pp. 29-34, doi: 10.1109/CSITSS.2018.8768774.

[7] P. Garg, H. Garg and V. Ranga, "Sentiment analysis of the Uri terror attack using Twitter," 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 17-20, doi: 10.1109/CCAA.2017.8229812.

[8] N. Kumar, "Sentiment Analysis of Twitter Messages: Demonetization a Use Case," 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), 2017, pp. 1-5, doi: 10.1109/CSITSS.2017.8447796.

[9] M. R. Hasan, M. Maliha and M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data," 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 2019, pp. 1-4, doi: 10.1109/IC4ME247184.2019.9036670.

[10] V. Ikoro, M. Sharmina, K. Malik and R. Batista-Navarro, "Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers," 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2018, pp. 95-98, doi: 10.1109/SNAMS.2018.8554619.

[11] Kusrini and M. Mashuri, "Sentiment Analysis In Twitter Using Lexicon Based and Polarity Multiplication," 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), 2019, pp. 365-368, doi: 10.1109/ICAIIT.2019.8834477.

[12] A. R. Pai, M. Prince and C. V. Prasannakumar, "Real-Time Twitter Sentiment Analytics and Visualization Using Vader," 2022 2nd International Conference on Intelligent Technologies (CONIT), 2022, pp. 1-4, doi: 10.1109/CONIT55038.2022.9848043.

[13] A. J. Nair, V. G and A. Vinayak, "Comparative study of Twitter Sentiment On COVID - 19 Tweets," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1773-1778, doi: 10.1109/ICCMC51019.2021.9418320.

[14] H. Talpada, M. N. Halgamuge and N. Tran Quoc Vinh, "An Analysis on Use of Deep Learning and Lexical-Semantic Based Sentiment Analysis Method on Twitter Data to Understand the Demographic Trend of Telemedicine," 2019 11th International Conference on Knowledge and Systems Engineering (KSE), 2019, pp. 1-9, doi: 10.1109/KSE.2019.8919363.

[15] A. Olteanu, I. Weber, and D. Gatica-Perez, “Characterizing the demographics behind the#blacklivesmatter movement,” in Proc. AAAI Spring Symp. Series, 2016, pp. 310–313

[16] A. Badawy, E. Ferrara, and K. Lerman, “Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign,” in Proc. IEEE/ACM Int. Conf. Advances Soc. Netw. Anal. Mining, 2018, pp. 258–265

[17] ALDayel, A.; and Magdy, W. 2021. Stance detection on social media: State of the art and trends. Information Processing & Management 58(4): 102597. ISSN 0306-4573.

[18] Li, Y.; and Caragea, C. 2019. Multi-Task Stance Detection with Sentiment and Stance Lexicons. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)