Regulatory Analytics - Team 31

Jeremy Miller	Deval Shah	Helly Jain
Jyothsna Krishnamurthy	Duong Vo	Aaron Roell

Introduction - Motivation

The federal rulemaking process of the United States plays an incredibly important role in the lives of millions of people and can be quite complicated with many interactions taking place between several different groups. For a new rule, Congress passes a law and delegates authority to regulatory agencies to create the rules to enforce those laws. After drafting a proposed rule, the agency seeks public comment before writing the final rule and adding it to the Code of Federal Regulations (CFR) (Carey, 2013). Even from this simplified process, it is apparent that the creation of a rule creates a large amount of data with more generated from the details embedded with each step of the process and the legacy that the rule leaves behind (for example, court challenges).

To then, retroactively, analyze the relationship between rules and their source laws or the public comments from when the rule was being proposed becomes a daunting task. Currently, tracing these relationships is done in silos on an as-needed basis for the impact of an individual law or rule requiring significant effort. Although public comments are considered when drafting the final rule, no formal analysis is utilized when making these decisions. Public comments can be further complicated if a rule is published for comment more than once or if interested groups attempt to use mass comment campaigns to sway the rule in their favor.

Problem Definition

This project has two main objectives: to show the relationships between federal statutes and the resulting regulations and to analyze the public's opinion toward proposed regulations. Our approach to improving the current practice is to create an interactive visualization of the relationships between statutes and regulations and provide an overall depiction of public opinion.

The end result takes data that is already generated (and freely available) in order to automatically visualize and analyze the relationship between statutes and regulations and summarize overall public sentiment about proposed rules. Regulatory agencies and regulated industries are significantly impacted by the current rulemaking process. Any improvements to the process will be beneficial for these groups resulting in the ability to measure the impact of the rules across each industry, identify the most influential rules, and review new rules with overall public opinion in mind.

Survey

The importance of such a tool is made clear by Hall (2004) who performed an analysis of the laws passed by two different Congresses and noted significant interdependency between agencies. Grabosky (2012) also examined the increase in non-government actors in the regulatory process. The importance of understanding these regulations is emphasized in Goltz (2017) which examines the financial impact of regulations by their penalties for non-compliance.

An intuitive way to visualize the relationship between laws and regulations is with a graph. Hamou-Lhadj (2009) describes how the structure of these documents lends itself to a graph visualization. Cleland-Huang (2010) demonstrates a machine learning algorithm to trace the regulation's requirements to the end result.

Several groups have analyzed the content of the regulations. Lau (2005) calculated similarity based on a text analysis of a rule but notes shortcomings when analyzing general intent. J. Zhang (2016) proposes an information extraction algorithm for construction regulations using Natural Language Processing. Similarity metrics could be used as well, but it may be difficult to account for the nuances and interpretations of each law; therefore, a node-based approach seems the most straightforward to implement and seems appropriate for this task.

While graph theory lends useful measures to determine the influence of a node, other methods can be used to measure the influence of written documents. Ravenscroft (2017) describes the correlation of the Mean Normalized Citation Score to the scientific impact of research. Similarly, Fetscherin (2015) applies bibliometric citation meta-analysis to examine consumer brand relationships and the overall influence of several research teams. Al-Ubaydli (2017) performs an analysis on US federal regulations using their word count which is highly correlated with the number of restrictions imposed.

Another possible metric is the relevance of a particular law. Opijnen (2017) discusses a multi-dimensional classification system for determining relevance but does not specify a particular method for analysis. Lv (2016) has a similar goal and discusses the use of vector space and statistical language models to determine relevance. All of these papers helped us explore the options for measuring the influence of a law.

Sentiment analysis would be a useful tool for policy decision-makers to understand the public opinion on the issue. Many groups have applied sentiment analysis in a variety of contexts.

Lv (2018) highlights the importance of analyzing sentiment and concern with methods for reviewing large scale projects with environmental impact for these aspects. Lemieux (2016) discusses the application of Mixed-Initiative Social Media Analytics for current social and political issues discussed on social media and the use of this data in the regulatory assessment. Agarwal (2011) analyzed microblog text using a prior polarity technique to establish the positivity or negativity of the text. Lim (2020) compares three approaches of text representation for sentiment analysis: Latent Semantic Analysis, Word2Vec, and Embedding from Language Models. Bakshi (2016) details the steps for sentiment analysis using Natural Language Processing to assign sentiment scores to the text. L. Zhang (2018) and Cambria (2017) survey the considerations that must be made when implementing a deep learning algorithm for sentiment analysis such as detecting emotion, but also nuances of language such as sarcasm and metaphor. We utilized multiple approaches to sentiment analysis, given the formal language used to write laws, and hand-tune the understanding. This introduces bias, but may lead to an increased readability and usefulness of the resulting sentiment.

Spam comments from particularly interested groups must also be considered. Balla (2019) finds evidence of such behavior in US rulemaking but does not explore its influence. This possibility must be considered during data cleaning to filter out low or no content comments.

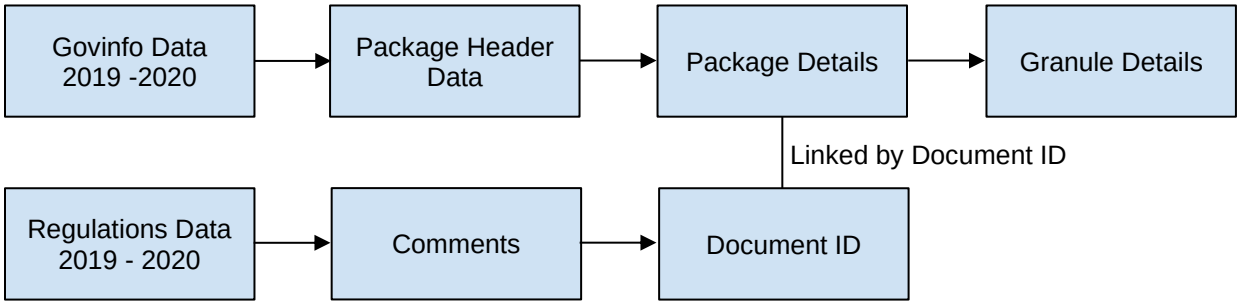
Proposed Method

There are three primary data sources in order to achieve the objectives of this project, and all are publicly available on the internet.

Govinfo.gov contains official publications from the federal government. The main publications of interest for this project are the Federal Register, CFR, and statutory citations. Regulations.gov

contains information related to the development of rules and houses the public comments for proposed rules. Reginfo.gov provides other information about regulations such as the Unified Agenda, regulatory reviews, and information collections. Govinfo.gov and Regulations.gov have APIs available in order to obtain data directly from the website. Data from Reginfo.gov must be downloaded as an XML file.

Acquiring the data from govinfo.gov and regulations.gov posed some challenges in data retrieval due to the limitations imposed by the API. To acquire enough data, the team acquired several API keys and ran the ETL process in iterations. The linkages between Document ID from reginfo.gov and comments in regulations.gov were done by mapping Document ID to Docket ID from the regulations.gov and iterating each Docket ID 10000 times to gather all the comments for the document. The merged dataset was around 6 million rows. The data acquisition was done using Python. One module was primarily used to get the comments data from Regulations.gov. The other module retrieved data from govinfo.gov. First, the header data for the years 2019 - 2020 was obtained for both sources. Then, the rules and associated granules were retrieved by making API calls in iterations and simultaneously accounting for the data limitations from the API by constructing dynamic API calls by including the Package ID and offsets. A summary of the process is shown below. After extracting all of the required data, it was stored in a SQLite database for ease of retrieval and analysis.



The document information, comments, CFRs, and specific parts of a document (called granules) from these three data sources were merged into one file using SQL and Python to create the final dataset, using string normalization and matching. This final dataset was used to determine the dataframes for nodes and edges used to create the graph visualization with D3. The visualization is a force directed graph with colors representing different types of nodes. Forces include a many body force to repel nodes from each other and both center and link forces to pull attract nodes. A collision force keeps nodes from overlapping one another.

In order to perform the sentiment analysis, we started with rule-based models like VADER but due to limited vocabulary the sentiment score for this model was not appropriate. This led us to use a convolution neural network model trained using Word2Vec encoding and IMDB Large Movie Review Dataset (Maas, 2011). We then extracted the comments attached to the regulations and predicted the sentiment using this model. We appended the score to the json as another field, for ease of visualization.

Experiments/Evaluation

The number of citations to a rule can indicate the importance and relevance. The linkage was seen in our visualizations between citations and rules. Using our data driven insights we validated these by manually tracing back to the rules from govinfo to understand if the rules that had the most mapping were indeed important and relevant. Establishing this ground truth helps measure

the accuracy of our mapping and visualizations. We also validated various independent comments as a spot-check for our algorithm.

An initial visualization was created using approximately 10,000 rows from the dataset (consisting of about 25,000 nodes and edges) which led to new ideas and improvements for the visualization. The resulting network was too sparse (multiple nodes with few to no edges) to glean meaningful insight. As a result, the next iteration was done on a filtered dataset by removing nodes that do not have any edges to help fine-tune the visualization before incorporating the rest of the data. The initial visualization also highlighted that a single visualization with all of the data may be impractical in terms of both interpreting the visualization as well as computing resources needed (the initial visualization required a significant amount of RAM). Therefore, a way for the user to change the scope of the visualization would be useful functionality to control the visualization to suit their needs. Finally, the rules with comments appeared in the initial visualization as starbursts where many nodes all emanated from a single point. This led to the idea to aggregate the sentiment from all the comments for a rule into a single node. The node could then convey sentiment using color and the size of the node could represent the number of comments on a particular rule. This can demonstrate how relevant a rule is to various actors, and the overall sentiment toward it. This could then be further analyzed for trends in the data, such as how the sentiment of the public comments is correlated with the overall volume of comments for that rule, or the sentiment of comments made by different types of groups, such as religious groups, individuals, cultural organizations, and other interest groups.

Our goal is to augment the analysis and visualizations by further providing insights into the most impactful and relevant rules and their sentiment associations. Other forms of analysis will include ranking, and status progress of the rules where available.

Our project traces regulatory codes to their relevant citations and comments in a unified form. Users are often forced to manually pore over documentation manuals to identify relevant sections of a rule or code, and then painstakingly trace them to the comments and their impacts. Our method is the first of this kind in this space which links codes to comments and their sentiments. The traceability problems have been partially addressed by enabling the linkages in our data wrangling process. The visualizations and sentiment analysis demonstrate the concentration of citations and sentiment linked to rules. This equips a user with single place to view the data.

A Keras model was trained on Large movie review dataset available on Stanford.edu. Google Word2Vec with 200,000 vocabulary was used to vectorize the text. Sentences were limited to 400 words to make it consistent with the neural network input layer. The Keras model used an epoch size of 5 with batch size of 32. The convolution filter size was 3. Relu was used as the activation for the Dense layer and Sigmoid for the output layer. A 1 layer convolutional layer was used to train the model with sigmoid used in the fully connected layer for the results. A dropout of 20 % was used in the hidden layer to introduce some bias in the model and hence avoid overfitting. The network was then trained using ‘binary_crossentropy’ loss, ‘adam’ optimizer on ‘accuracy’ matrix and 80-20 train-validation split. The results we got aligned well with the sentiments we were able to validate manually.

To perform experiments and to evaluate the sentiment analysis model, we prepared a small subset of data by manually labeling comments. Through the progress, we discovered some interesting challenges and insights. At the start, the base model only gave less than 50% accuracy. The first identified problem related to the nature of the dataset. Classifying reviews of movies is often straightforward since a reviewer often chooses either the positive side or the negative side. In contrast, the regulatory public comment dataset contains more negative and

neutral comments than positive ones. From our labeling activity, we observed over 50% of the negative comments while neutral and positive comments shared equal portions. The line between neutral comments and the other two types were quite blurry. Many long comments included both pros and cons. Some commenters gave observations and used scientific languages. Others showed support for a regulatory document by presenting the downside of the current situation. Their use of strong words can mislead the algorithm, resulting in the wrong classification. The second problem is data quality. Even though we applied a data cleaning step to check for duplicates, some long comments still bypassed the check. In addition, approximately 7.8% of the comment data are attachments and many of them were misclassified.

For the data quality problem, the team removed all comments that contain attachment since it would not be possible to scan pdf documents within the timeline of the project. To handle the problem of the unbalanced data, we put thresholds on the probability score to categorize comments into 3 groups - positive, neutral and negative instead of labelling only negative and positive. We performed iterations over distinct combinations of thresholds to pick the best pair for different models. Even though training and testing different models took a long time, the team was able to test multiple models with various hyperparameters (epoch, vector size, dropout) to pick out one with the best accuracy of approximately 75.96% (refer to Appendix 1). As we experimented with different hyperparameters, the team was aware of their impacts on the model and intentionally picked out ones that would complement each other. For example, we increased the number of epochs and the drop-out rate at the same time to avoid overfitting. The best model is one with 5 epochs, 0.2 drop-out rate, and size of word vectors of up to two hundred thousand.

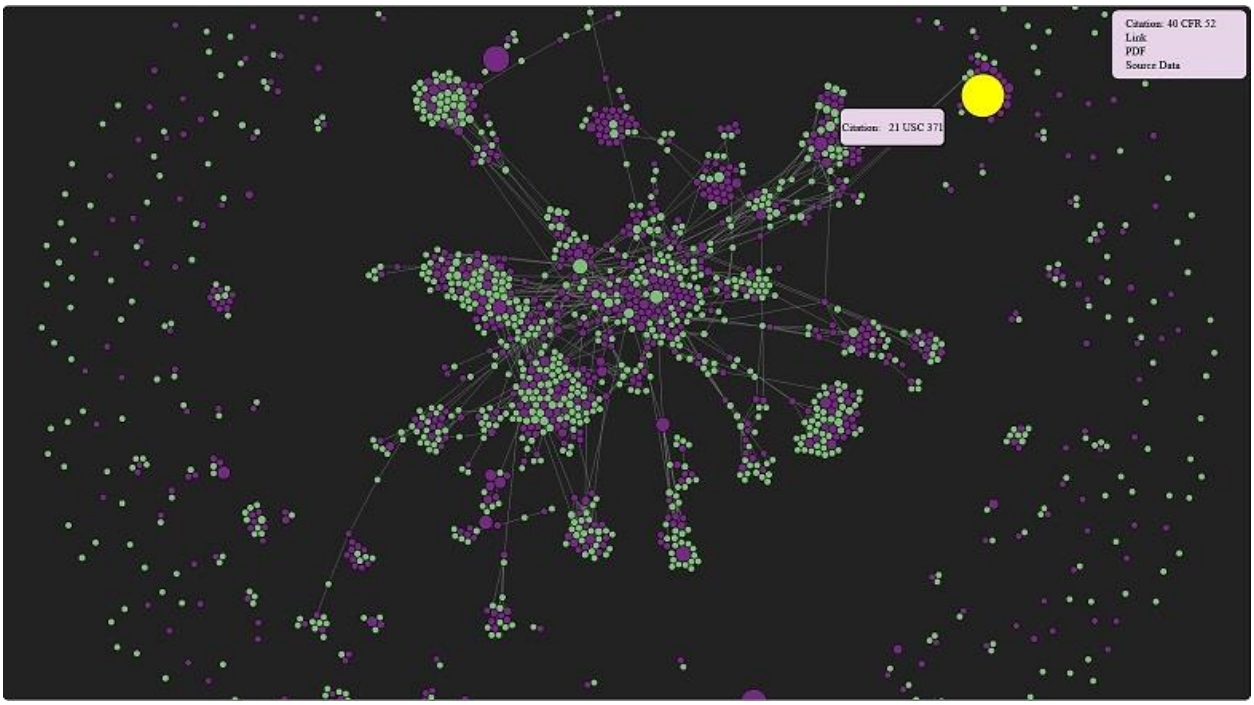
Conclusions and Discussion

During the course of this project, we made significant progress in scraping to get, merge, and visualize the data, build a sentiment analysis model to better understand the comments attached to the regulations and iterative toward a robust and easy-to-understand visualization while making it easy to discover comments and regulations.

During the course of this project we used innovative techniques to gather and analyze the data. Acquiring the data from multiple sources required coordination amongst the team to create several API keys to loop through to overcome limitations in downloading the data. Performing sentiment analysis on federal regulations has not been performed to the same extent we are pursuing in this project. We also had to design the visualization in an innovative manner - to make sense of federal regulations is not trivial, and making the data discoverable can lead to more participation in the democratic process.

The primary graph visualization is a force directed graph representing the interconnections between US Code (USC), Code of Federal Regulations (CFR) and Federal Register (FR) citations related to the public comments analyzed. The full graph is approximately 150,000 nodes across 4 node types, but has been reduced to show only the top 2000 by page rank. Size represents the node's Page Rank; color represents the node type. Green nodes are statutes from the US Code. Purple nodes are regulations from the CFR. No FR or public comment nodes had a high enough page rank to be displayed. The Yellow node is the selected node and its details are displayed in the box on the top right which includes links to the web, PDF and XML versions of the node. The is also a tool tip for the hovered node

Forces include a many body force to repel nodes from each other and both center and link forces to pull attract nodes. A collision force keeps nodes from overlapping one another.



Additionally, we have identified further recommendations for this project to expand the functionality of the project. Entity extraction could be used to further analyze comments and free-text fields for mentions of prominent politicians or policies. Additional filtering capabilities by year or industry affected could be valuable for policy-makers and leaders in industry to analyze the regulatory landscape from a big picture perspective. Furthermore, automating or streamlining the data scraping process could be useful to have a consistent refresh of the data needed, although we recognize that the structure and the rate limiting aspect of the various API's used make it so this may be hard to achieve.

All team members have contributed a similar amount of effort to the deliverables of the project. The efforts were distributed across project ideation, data scraping and linking, analytics , visualizations, project management and documentation. Every team member put their best effort to construct the deliverables while keeping an eye on quality.

References:

Agarwal, Apporv, Rebecca Passonneau, Owen Rambow, and Boyi Xie. (2011). “Sentiment Analysis with Twitter Data.” Department of Computer Science at Columbia University, 2018. https://doi.org/10.4135/9781526468857.

Al-Ubaydli, O., & Mclaughlin, P. A. (2017). RegData: A numerical database on industry-specific regulations for all United States industries and federal regulations, 1997–2012. *Regulation & Governance*, *11*(1), 109–123. https://doi.org/10.1111/rego.12107

Bakshi, R. K., Kaur, N., Kaur, R. and Kaur, G., "Opinion mining and sentiment analysis," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 452-455.

Balla, S. J., Beck, A. R., Cubbison, W. C., & Prasad, A. (2019). Where's the Spam? Interest Groups and Mass Comment Campaigns in Agency Rulemaking. *Policy & Internet*, *11*(4), 460–479. https://doi.org/10.1002/poi3.224

Cambria, Erik, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. “Sentiment Analysis Is a Big Suitcase.” IEEE Intelligent Systems 32, no. 6 (2017): 74–80. https://doi.org/10.1109/mis.2017.4531228.

Carey, Maeve P. “The Federal Rulemaking Process: An Overview.” Congressional Research Service, 2013. https://fas.org/sgp/crs/misc/RL32240.pdf.

Cleland-Huang, J., Czauderna, A., Gibiec, M., & Emenecker, J. (2010). A machine learning approach for tracing regulatory codes to product specific requirements. *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - ICSE '10*. https://doi.org/10.1145/1806799.1806825

Fetscherin, M., & Heinrich, D. (2015). Consumer brand relationships research: A bibliometric citation meta-analysis. *Journal of Business Research*, *68*(2), 380–390. https://doi.org/10.1016/j.jbusres.2014.06.010

Grabosky, Peter. “BeyondResponsive Regulation: The Expanding Role of Non-State Actors in the Regulatory Process.” Regulation & Governance 7, no. 1 (2012): 114–23. https://doi.org/10.1111/j.1748-5991.2012.01147.x.

Goltz, N., & Mayo, M. (2017, June 4). *Enhancing Regulatory Compliance by Using Artificial Intelligence Text Mining to Identify Penalty Clauses in Legislation*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2977570.

Hall, T. E., & O'toole, L. J. (2004). Shaping Formal Networks through the Regulatory Process. *Administration & Society*, *36*(2), 186–207. https://doi.org/10.1177/0095399704263476

Hamou-Lhadj, A., & Hamdaqa, M. (2009). Citation Analysis: An Approach for Facilitating the Understanding and the Analysis of Regulatory Compliance Documents. *2009 Sixth International Conference on Information Technology: New Generations*. https://doi.org/10.1109/itng.2009.161

Hobson Lane, Cole Howard, Hannes Max Hapke “Natural Language Processing in Action”

Understanding, analyzing and generating text with Python. https://www.manning.com/books/natural-language-processing-in-action

Lau, G., Wang, H., Law, K., & Wiederhold, G. (2005, May 1). *A relatedness analysis approach for regulation comparison and e-rulemaking applications*. A relatedness analysis approach for regulation comparison and e-rulemaking applications | Proceedings of the 2005 national conference on Digital government research. https://dl.acm.org/doi/pdf/10.5555/1065226.1065246.

Lemieux, V. L. (2016). Innovating Good Regulatory Practice Using Mixed-Initiative Social Media Analytics and Visualization. *2016 Conference for E-Democracy and Open Government (CeDEM)*. https://doi.org/10.1109/cedem.2016.38

Lim, Yu Qing, Chun Ming Lim, Keng Hoon Gan, and Nur Hana Samsudin. "Text Sentiment Analysis on Twitter to Identify Positive or Negative Context in Addressing Inept Regulations on Social Media Platform." 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE), 2020. https://doi.org/10.1109/iscaie47305.2020.9108706.

Lv, Xuan. "Semantic Process Analysis, Context-Aware Information Retrieval, and Sentiment Analysis for Supporting Transportation Project Environmental Review." https://www.ideals.illinois.edu/handle/2142/101202, April 20, 2018. https://cm.fiu.edu/wp-content/uploads/2019/03/CV-Xuan-Lv.pdf.

Lv, X., & El-Gohary, N. M. (2016). Enhanced context-based document relevance assessment and ranking for improved information retrieval to support environmental decision making. *Advanced Engineering Informatics*, *30*(4), 737–750. https://doi.org/10.1016/j.aei.2016.08.004

Maas, A. L., Daly, R. E., Pham P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*

Opijnen, M. V., & Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, *25*(1), 65–87. https://doi.org/10.1007/s10506-017-9195-8

Ravenscroft, J., Liakata, M., Clare, A., & Duma, D. (2017). Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. Plos One, 12(3). https://doi.org/10.1371/journal.pone.0173152

Zhang, Jiansong, and Nora M. El-Gohary. "Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking." Journal of Computing in Civil Engineering 30, no. 2 (2016): 04015014. https://doi.org/10.1061/(asce)cp.1943-5487.0000346.

Zhang, Lei, Shuai Wang, and Bing Liu. "Deep Learning for Sentiment Analysis: A Survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8, no. 4 (2018). https://doi.org/10.1002/widm.1253.

Appendix 1: Accuracy score for Sentiment Analysis

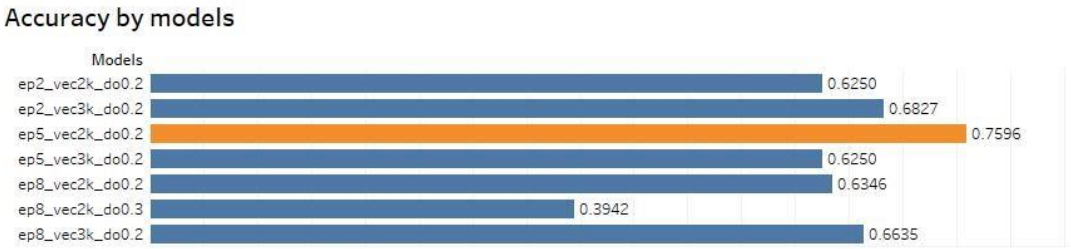


Figure SA1: Chart shows accuracy by different models.
(ep2_vec2k_do0.2 means 2 epochs, 200k vector and 0.2 drop-out rate)

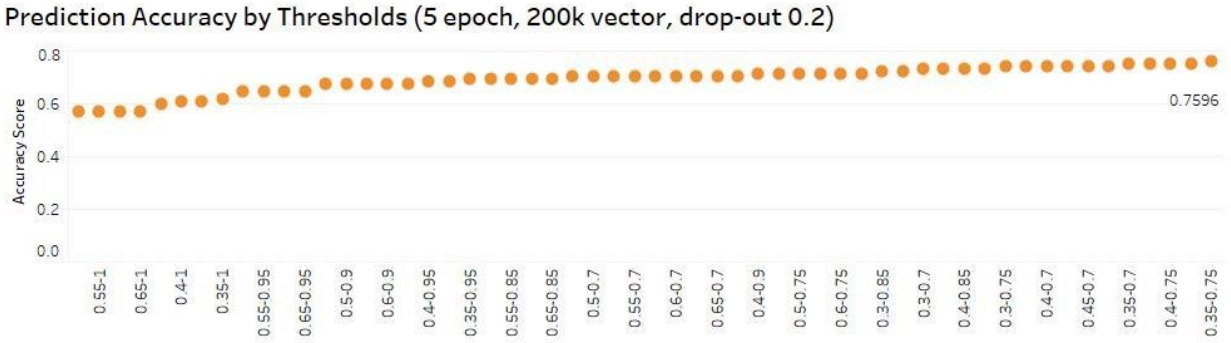


Figure SA2: Bottom chart presents prediction accuracy by different thresholds.
(0.35-0.75 means comments with probability <0.35 is negative, >0.75 is positive and neutral in between)