**Sentiment Analysis on Social Media Comments: Evaluating Public Perception for Reforms in Uzbekistan**

Table of Contents

[Acknowledgement 2](#_Toc153488609)

[Abstract 3](#_Toc153488610)

[Introduction 4](#_Toc153488611)

[Recent Reforms in Uzbekistan 4](#_Toc153488612)

[Literature review 6](#_Toc153488613)

[Data Collection 8](#_Toc153488614)

[Preprocessing Data 8](#_Toc153488615)

[Methodology 8](#_Toc153488616)

[Discussion 13](#_Toc153488617)

[Conclusion 14](#_Toc153488618)

[Bibliography 15](#_Toc153488619)

[Appendixes 16](#_Toc153488620)

# Acknowledgement

I would like to acknowledge and give my warmest thanks to my supervisor Xuan Lu who made this work possible. Her guidance and advice carried me through all the stages of writing my project.

I would also like to give special thanks to my wife and my family as a whole for their continuous support and understanding when undertaking my research and writing my project. Your prayer for me was what sustained me this far.

Finally, I would like to thank God, for letting me through all the difficulties. I have experienced your guidance day by day. You are the one who let me finish my degree. I will keep on trusting you for my future.

# Abstract

This report outlines the comprehensive analysis undertaken for the Capstone Project by Shakhrukh Bobojanov, a graduate student in the Master's in Data Science program at the University of Arizona, under the supervision of Xuan Lu. This project represents the academic competencies acquired during MS in Data Science program, applying advanced data science techniques and in-depth research to analyze public sentiment in Uzbekistan regarding recent socio-political&economical reforms. Project data and scripts are provided in GitHub[[1]](#footnote-1) repository.

This study employs extensive dataset[[2]](#footnote-2) of over 700,000 comments extracted from more than 23,000 posts from the kun.uz Instagram profile. The main goal of the project is to analyze the public views on the ongoing reforms in Uzbekistan, using both Neural Network-based and traditional Machine Learning-based text classification approaches to classify peoples’ comments into Positive, Negative and Neutral classes. This dual methodology provides a robust framework for sentiment analysis in Uzbek language, providing initial sentiment of public for reforms that resonate positively with the public and for those that might need further engagement and refinement.

The findings of this analysis showed that among both Neural Network and Machine Learning models, CNN model showed best performance with an accuracy of 0.786. The findings obtained from the data represented that negligible gap between positive and negative comments in Society, Politics and Economy suggests both pessimistic and optimistic views. When it comes to categories like Health, Sports, Science and Technology and World, majority of comments and posts are positively discussed, highlighting that these areas are well-accepted by the public. Conversely, Crime content posts have been negatively perceived by the public. Another interesting finding is that during the pandemic peoples’ activity in the media has been reached at its peak as well as with its pessimistic sentiment. Since then, overall public pessimistic sentiment has been gradually decreasing over time after the end of pandemic and positive sentiment has been dominating.

**KEYWORDS:** **UZBEKISTAN, SOCIAL MEDIA, NEURAL NETWORKS,** **NLP, GRU, LTSM, MACHINE LEARNING, LOGISTIC REGRESSION, RANDOM FOREST, KNN**

# Introduction

Since the election of President Shavkat Mirziyoyev in late 2016, Uzbekistan has been experiencing a significant journey of transformation, with wide-ranging reforms in governance, the economy, and social sectors. These significant changes, guiding the country towards a more progressive future, have triggered different reactions from the public. This project seeks to analyze public sentiment regarding these reforms, utilizing a large dataset from the kun.uz Instagram profile. By exploring the online community's views and thoughts by analyzing comments on news posts, the study aims to capture the perspectives of the Uzbek people, providing their collective thinking during this reformative period. Understanding public sentiment is crucial for the success of these policy reforms. It is a fundamental element in assessing the effectiveness of these changes and determining their acceptance among the population. This research focuses on identifying the aspects of the reforms that resonate positively with the public and those that may need further engagement or reassessment. The central research question explores how these reforms are perceived by the online community as the representative of whole population. The findings derived from this analysis could be useful for policymakers, academics, and the general public in understanding and shaping future policy directions.

The rest of the paper is structured as follows: Firstly, an overview of recent reforms in Uzbekistan will be provided. This will be followed by a review of related works in the field of Natural Language Processing (NLP). The data and methods of the study will then be presented. Subsequent sections will provide the results and discuss these results before concluding the paper. The bibliography will finalize the report.

# Recent Reforms in Uzbekistan

In recent years, Uzbekistan has witnessed significant policy changes in several key sectors, marking a massive transformative period in the nation's history. These reforms have been important step ahead in driving economic growth, enhancing governance, and improving the social sectors of the country.

**Social Reforms in Education**

Uzbekistan has made substantial progress in social reforms, particularly in education and healthcare. The government has increased its spending on education, allocating approximately 8% of its GDP to this sector, one of the highest rates in Central Asia. The development strategy named “Action Strategy for Five Priority Areas of Development of the Republic of Uzbekistan in 2017-2021,” adopted at the beginning of a new stage of reforms in Uzbekistan in February 2017, one of the priorities is “to continue the course of further improving the system of lifelong education, increasing the availability of quality educational services, training highly qualified personnel in accordance with the modern needs of the labor market”. Prior to the start of a new stage of reforms, preschool education in Uzbekistan did not receive much attention. If earlier the coverage of children with this form of education was only 27%, then by the end of 2019 it had already increased to 44.5%. During this period, the number of public preschool institutions (preschool institutions) increased 1.5 times (from 4940 to 7500), and private preschool institutions - 3 times (from 269 to 783). If in 2017, 51 thousand teachers worked in the preschool education system, then by the end of 2019 - more than 80 thousand (Burkhanova, 2021).

Regarding high education, the number of higher education institutions has increased by 210 since 2018, resulting in a notable rise in the proportion of the population pursuing higher education, from 9% to 38%. This growth is not limited to public universities, as private universities are also expanding in number. Presently, more than 30 higher education institutions across the country offer foreign language education, whereas in 2016, this number stood at only 7 (Tukhtasinov et al., 2023). Accordingly, number of students in the universities of the republic increased from 140.2 thousand in 2017/2018 to 461 thousand in 2021/2022 academic years.

**Social Reforms in Healthcare and Clean Water**

In 2018 Uzbekistan initiated comprehensive and far-reaching reforms of its health system, including fundamental changes in health financing and service delivery systems, with a primary health care (PHC) approach at its core. Over the past year, more than 20 resolutions and decrees aimed at improving the quality of medical care have been adopted. In 2023, 21 trillion Uzbek soums were allocated from the budget, which is 3 times more than in 2017. Today, modern technologies are being introduced into the activities of medical institutions, which are more than three thousand in the country. Besides, 423 primary care institutions are connected to the Electronic Polyclinic information system which dramatically reduce paperwork, bureaucracy and corruption in the sphere through digitalization (President.uz, 2021). Apart from that, since most of the health problems are caused by consuming not clean water. As of 2023, 70 percent of the Uzbekistan population is supplied with centralized drinking water. According to the statistics, in 2017-2022, more than 25 300 kilometers of networks, more than 1 800 water structures were built and restored. During the past 6 years, additional 6,500,000 people have enjoyed centralized drinking water for the first time, and the water supply of more than 4 million people has been improved (Darakchi.uz, 2023).

**Privatization reforms**

The presence of SOEs in the economy remains pervasive, as little progress has been achieved

in divesting the 2,800 SOEs directly owned by the central government. The privatization efforts in Uzbekistan have been remarkable with the government initiating the privatization of over 1,100 state-owned enterprises since 2017 according to reports of World Bank. Apart from that in the Strategy of Privatization of SOEs 2021-2025, 75% of all SOEs planned to be privatized. During the reforms, a variety of decrees were adopted, culminating in the adoption of the Strategy of Privatization of SOEs 2021-2025 (Corissa, 2021). This move is aimed at boosting the private sector's role in the economy, enhancing efficiency, and attracting foreign investment. The privatization drive has been particularly prominent in sectors like banking, energy, and telecommunications, where private participation was previously limited. For example, a May 2020 decree outlines the plan to reduce the state-share of system assets from 85 percent to 40 percent by 2025 through the full sale of six large state-owned banks. If we look at FDI flows, FDI grew from $1.6 billion in 2018 to $7.81 billion in 2021, according to Uzbekistan Ministry of Finance.

**Trade Openness**

Uzbekistan’s progress in opening up its economy has been remarkable. According to World Bank report (2021), the government announced unilateral trade liberalization measures through tariff reductions in 2017-2018, reducing average effective tariff rates from 15.3 percent in September 2017 to about 7.5 percent in 2020. These measures lowered most peak tariffs, changed most mixed and compound tariffs to more transparent ad-valorem tariffs, and eliminated most instances of discriminatory excise taxes on certain imports. These measures significantly strengthened trade competitiveness. Simultaneously, the government has enacted other trade liberalization and facilitation measures, including removing export permit requirements for most goods, improving the efficiency of customs procedures, and reducing customs clearance times. The government has also prepared a Comprehensive Legislative Action Plan (LAP) to assess the WTO conformity of domestic trade-related legislation and practices. The LAP serves as a full inventory of WTO-related legislation enacted in Uzbekistan and serves as a roadmap for WTO-related domestic legislative reform. Looking at the numbers, according WITS data, trade turnover with neighboring countries has increased, with a notable 60% increase (from $1.6 billion in 2016 to $3.8 billion in 2021) in trade with Kazakhstan and about 80% rise (from $5.4 billion in 2016 to $9.3 billion in 2022) in trade with Russia in recent years. Moreover, substantial progress has been achieved on non-tariff barriers as well. The main reforms include the unification of the exchange rate the elimination of foreign exchange surrender and advance prepayment requirements, and the removal of most export inspection processes.

**Tax reforms**

The government has significantly reforming tax policy since 2017 and has committed to a medium-term improvement framework to strengthen tax administration. The thrust of the reform in 2018-2019 was on reducing and unifying the tax burden on small and large enterprises,

unifying the rates of corporate profit tax, personal income tax, and the social tax; rationalizing the VAT payments; reducing the number of direct taxes and mandatory payments, and improving tax administration procedures. Despite the lower tax rates, the tax reform significantly improved compliance – as expected – and increased revenues. In late 2019, additional measures were enacted to further improve the tax system. The authorities ended many tax and customs preferences for companies, which cost about 4-6 percent of GDP a year. Measures were also enacted to expand the VAT to include the agriculture sector, a significant part of the economy that had been previously left out of the VAT changes introduced in 2018. A new tax code came into force from January 2020. The new code is a significant improvement over the previously burdensome, complex, and regulatorily invasive code that was inherited from the Soviet era. The most important objective of the new code is to modernize and simplify tax policy and processes. The new code also introduces new measures to expand the VAT to include digital services, modernize the transfer pricing, thin capitalization, and controlled foreign company (CFC) regimes (World Bank Report, 2021).

These developments in Uzbekistan reflect a concerted effort to transition towards a more open, market-oriented, and inclusive society. The numbers and the scale of reforms underscore the country's commitment to fostering sustainable growth and improving the quality of life for its people.

# Literature review

In the text classification within Natural Language Processing (NLP), different methodologies have been found to be effective, ranging from advanced neural networks to traditional machine learning algorithms. This literature review focuses on the contributions of some studies with different models used in each study.

A paper by Yoon Kim (2014) demonstrates the application of Convolutional Neural Networks (CNNs) for sentence text classification. The study reveals that CNNs, traditionally associated with image processing, can be effectively adapted for NLP tasks. The author’s experiments showed that simple CNN models vectorizing with different vectorization with word2vec, with one layer of convolution, could achieve excellent results on multiple benchmark datasets. The paper highlights that baseline model with all randomly initialized words (CNN-rand) does not perform well on its own. However, the versatility of CNNs in capturing semantic patterns in text, making them a powerful tool for NLP applications when utilizing finetuning the pre-trained vectors for each task.

Another paper by Zaremba et al. (2014) offers regularization techniques for Recurrent Neural Networks (RNNs), particularly LSTM networks. The authors presented various strategies to prevent overfitting in these networks, which is crucial for maintaining their performance on NLP tasks. The study's findings showed that dropout, the most successful technique for regularizing neural networks, does not work well with RNNs and LSTMs. Furthermore, they provided a simple way of applying dropout to LSTMs that results in large performance increases on several problems in different domains. They suggested how to properly use dropout for RNNs, and the results suggest that their implementation of dropout could improve performance on a wide variety of applications.

The paper by Jalal et al. (2022) examines the application of Random Forest (RF) algorithms in text classification using feature ranking and optimal number of trees. The authors demonstrates an improved random forest for text classification, called improved random forest for text classification (IRFTC), that incorporates bootstrapping and random subspace methods simultaneously. The IRFTC removes unimportant (less important) features, adds a number of trees in the forest on each iteration, and monitors the classification performance of RF. Their results suggest that IRFTC improves the performance of traditional RF by 6.28% for binary classification while 4.98% with respect to the best performer Naïve Bayes. On the other hand, IRFTC shows 1.50% better performance than traditional RF and 3.37% than the decision tree which is the second best performer after RF for the multiclass classification task. Results show that the proposed IRFTC shows equally superior performance with both binary and multiclass text classification.

Focusing on Logistic Regression, the paper by Prabhat (2017) analyzed twitter reviews categorizing them as positive or negative by employing Naïve Bayes and Logistic Regression. They used Hadoop, Mahout and Eclipse file system interfaces using java programming Language to train and classify the Naïve Bayes and Logistics Regression. The classification task is carried out by both classifiers using unigram technique. The results suggested that analysis with logistics regression gives 10.1% more accurate and 4.34% more precise results with a bit more implementation time for same size of dataset than Naïve Bayes.

The research by Shah (2020) explores BBC news text classification model based on machine learning algorithms. This paper proposes the logistic regression, random forest and K-nearest neighbor algorithms which describes every aspect of model in detail by providing the evaluation

metrics. When machine learning algorithms are implemented on a particular data set, the most important parameter that matters is the accuracy. Hence, the result shows that logistic regression classifier with the TF-IDF Vectorizer feature gets the highest accuracy of 97% for the data set. This algorithm has emerged as the most stable classifier in a small data set. The second best was the

random forest classifier with the accuracy of 93%. The algorithm with the least accuracy among the three was K-nearest neighbors with the overall accuracy of 92%. The logistic regression classifier gave the best performance as in terms of all parameters.

# Data Collection

The primary data source for this study is the kun.uz Instagram profile, a leading public media company in Uzbekistan with over 96,000 posts and 5.1 million followers. Due to funding limitations, I am limited to scrape only dataset comprises over 23,000 posts and their associated comments, amounting to more than 700,000 comments, dating from 2020 to the present. Due to Instagram's API restrictions on data scraping, the posts were collected using the paid version of exportcomments.com and comments were scraped using paid version of apify.com website. Unfortunately, this tool enabled the extraction of a maximum of 50 comments per post, even though some posts has over 1,000 comments. This dataset provides a comprehensive overview of public opinions and reactions to various topics covered by kun.uz.

# Preprocessing Data

The preprocessing[[3]](#footnote-3) of the dataset is critical step to ensure data quality and consistency. Therefore, I implemented some data preprocessing steps:

1. Emoji Removal: All emojis were removed from the dataset to simplify the text data and focus on textual content.
2. Handling NA Values: Any missing or NA values in the dataset were dropped to maintain data integrity.
3. Transliteration: A significant step in the preprocessing was converting Latin characters to Cyrillic. This was essential for consistency, as the data contained a mix of Latin and Cyrillic scripts. The transliteration process ensured uniformity in the dataset, making it suitable for further NLP analysis.
4. Lowercase Conversion: All text data was converted to lowercase to maintain consistency and avoid duplication due to case differences.

# Methodology

Before moving onto comment classification, in order to determine the main idea of each post I classified[[4]](#footnote-4) posts into relevant topics using coppercitylabs/uzbek-news-category-classifier hugging face pre-trained model. This model, specifically designed for the Uzbek language, is a pre-trained text classification tool and fine-tuned on approximately 60K news articles over 3 epochs. It categorizes social media posts into various thematic categories such as Society, Health, Sports, World, Science and Technology, Politics, Crime, Economy, Culture, Show Business and Miscellaneous (See the description in Table 1). This classifier uses a combination of advanced NLP techniques, including tokenization, feature extraction, and classification algorithms, to accurately assign categories to each post. The process involves a comprehensive analysis of the text, focusing on key phrases, sentiment, and contextual meanings, before mapping these elements to pre-established categories. Therefore, this approach helps identify the main areas of public interest or concern, shedding light on the subjects most discussed in relation to the reforms. And importantly, categorization allows for the tracking of how public perception evolves over time across different topics, providing overall picture into changing public sentiments. The model returns relevant Post category along with its probability. So, I accepted only predicted post categories with confidence of more than 50% and classified posts with less than 75% confidence as Miscellaneous posts.

Table 1. Description of Categories

| **Category** | **Description** |
| --- | --- |
| Society | Covers a wide range of topics related to the daily lives and concerns of the general public, including social issues and community events.. |
| Health | Includes news related to medical discoveries, health advisories, wellness tips, healthcare policies, and information about diseases and treatments. |
| Sports | Features updates on various sports events, game statistics and analyses of competitions. |
| World | Includes international news, global events, foreign affairs, and stories that have an impact across different countries and cultures. |
| Science and Technology | Reports on developments in science and technology, breakthroughs in research, innovation, and the impact of technology on society. |
| Politics | Includes news about political events, elections, legislative developments, government policies, and political figures. |
| Crime | Reports on criminal activities, law enforcement, judicial proceedings, and any news related to unlawful actions and their consequences. |
| Economy | Deals with financial markets, economic policies, business trends, trade, employment, and economic indicators that affect the economy of country. |
| Culture | Focuses on the arts, literature, music, entertainment, traditions, and cultural events. |
| Show Business | Covers the entertainment industry, including news about celebrities, film and television, music releases, and behind-the-scenes. |
| Miscellaneous | Classified as Miscellaneous with less than 50% confidence predicted by the model |

For classification of comments, I manually labelled 10,000 comments into Positive, Negative and Neutral classes. Because some reactions of users are topic un-related and some reactions were advertisement purposed and some were just emojis and not full words. Therefore, I classified these comments as neutral.

**Machine Learning Models**

I initially implemented traditional machine learning models[[5]](#footnote-5) including Random Forest which is a robust method, useful for handling the high-dimensional data of text classification; Logistic Regression which is a fundamental model for binary and multiclass classification tasks; and KNN (K-Nearest Neighbors) method which is a simple yet effective algorithm for classification problems.

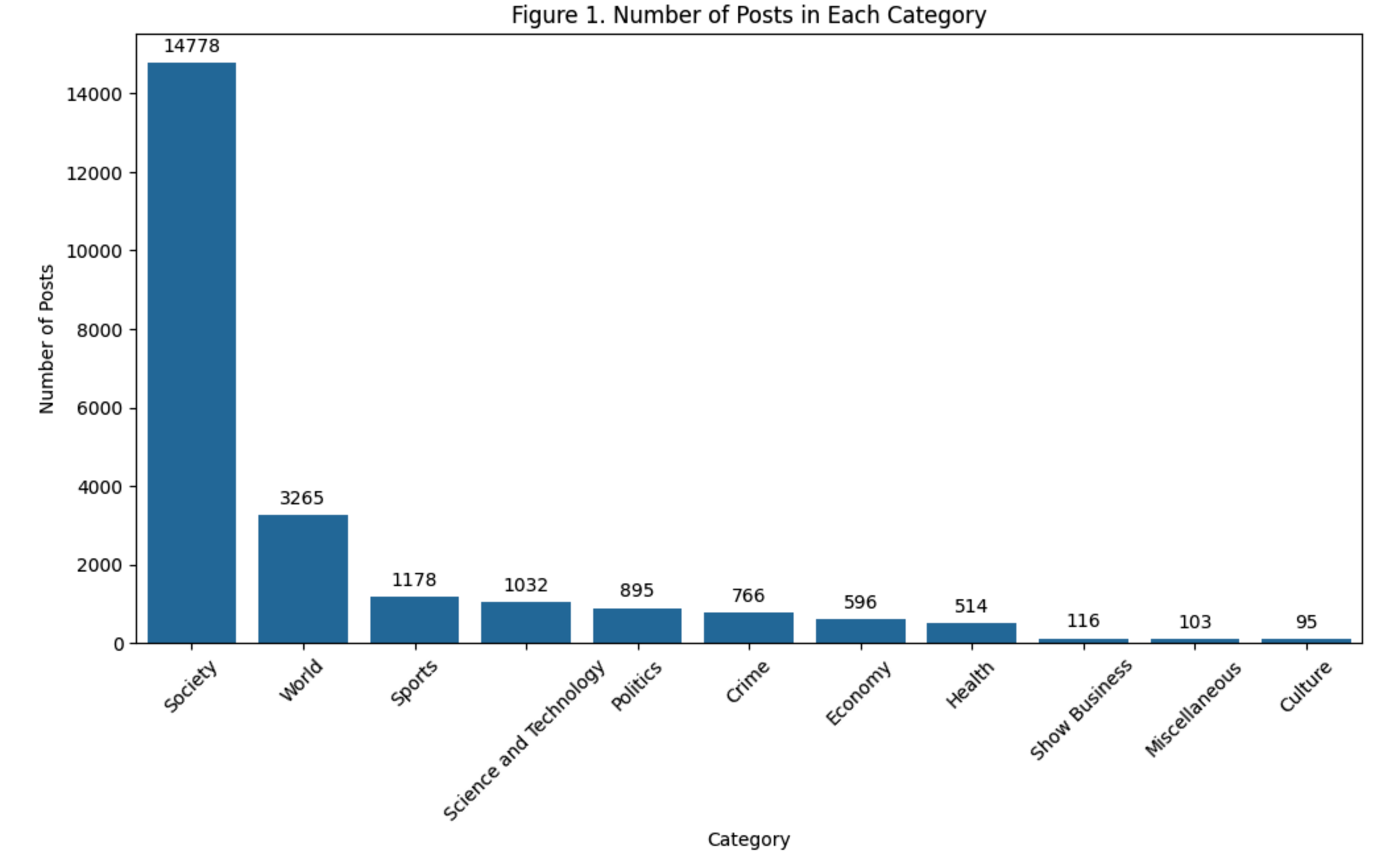
For this, I used TF-IDF (Term Frequency-Inverse Document Frequency) vectorizing method, transforming the text into a format suitable for machine learning models. Furthermore, I evaluated the models' performance using 5-fold cross-validation technique, considering the model’s predictive capabilities. To get higher accuracy of the models, I fine-tuned these models trying different parameters of the models to optimize the performance. This approach provides a comprehensive comparison between traditional machine learning models and advanced neural network architectures in text classification task.

**Neural Network Models**

In addition to traditional machine learning models, I employed advanced Neural Network[[6]](#footnote-6) models based on the coppercitylabs/uzbert-base-uncased pretrained model, specifically designed for the Uzbek language using the Cyrillic script. This model was developed following the masked language modeling and next sentence prediction objectives, providing a sophisticated understanding of language structures and context. It was pretrained on a substantial dataset of approximately 625K news articles, containing around 142 million words, ensuring its effectiveness in interpreting and processing Uzbek text data. So, I tried different Neural Network architectures including CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory) which is a RNN (Recurrent Neural Network) version designed to handle long-term dependencies in sequential data; GRU (Gated Recurrent Unit) which is similar in function to LSTM but with a more streamlined architecture; and Feed-Forward Neural Network which serves as a baseline model, providing a comparative perspective against other architectures.

For training these models, I used the preprocessed dataset, fine-tuning various hyperparameters to optimize each model's performance. This approach, integrating the advanced capabilities of the coppercitylabs/uzbert-base-uncased model with different neural network architectures, gave me a detailed exploration of text classification nuances. So that I had a chance to compare the performance of traditional machine learning and neural network models, highlighting the strengths and applications of each in the context of analyzing public sentiment in Uzbekistan.  
  
Results  
**Post Categorization Analysis**

The analysis of post categorization represents diversity of the post content of kun.uz Instagram profile (Figure 1). The categorization process classified a total of 23,204 posts into different post categories based on their content. The results show a predominant focus on Society-related news, with 14,778 posts falling under this category, which is approximately 63.7% of the total posts analyzed. This high number indicates a strong emphasis on societal issues and events, reflecting the media's focus on topics that resonate closely with the daily lives and concerns of the general public.



Other notable categories include World (3,265 posts), Sports (1,178 posts), Science and Technology (1032 posts), Politics (895 posts) and Crime (766 posts). These categories, while less predominant than Society, indicate a diverse range of interests provided to by the media channel. Interestingly, categories like Health (514 posts), Economy (596 posts), Show Business (116 posts), and Culture (95 posts) received relatively lower representation, suggesting these topics might not be the primary focus of the media channel or might not resonate as strongly with the audience.

The category 'Miscellaneous', which includes posts with less than 50% confidence in classification, comprised 103 posts. This could indicate either a diverse range of topics not fitting into predefined categories or a limitation in the classifier's ability to categorize certain posts accurately.

**Machine Learning and Neural Network Models**

The performance of the machine learning models was evaluated using a 5-fold cross-validation technique (See Table 2). The models showed varying levels of accuracy in classifying the sentiment of the comments into ‘Positive’, ‘Negative’ and ‘Neutral’. Logistic Regression (LR) emerged as the most effective model, with an average accuracy score of 0.610 and F1 score of 0.595. This suggests that LR, despite its simplicity, is quite effective for the binary and multiclass classification tasks in this dataset.

Random Forest (RF) followed closely with an accuracy score of 0.598, indicating its capability to handle the high-dimensional data characteristic of text classification. However, K-Nearest Neighbors (KNN) lagged behind with an accuracy of 0.469, possibly due to its simplicity and the nature of the text data, which might not be best suited for distance-based classification methods like KNN.

Table 2. Accuracy metrics obtained from models

| Model | Accuracy | Precision | Recall | F1 Score |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.610 | 0.679 | 0.570 | 0.595 |
| Random Forest | 0.598 | 0.643 | 0.569 | 0.587 |
| KNN | 0.469 | 0.636 | 0.397 | 0.335 |
| Feed Forward | 0.765 | 0.671 | 0.579 | 0.642 |
| LSTM | 0.745 | 0.619 | 0.615 | 0.619 |
| GRU | 0.757 | 0.644 | 0.609 | 0.635 |
| CNN | 0.784 | 0.696 | 0.623 | 0.668 |

In terms of Neural Network models, the performance was measured using accuracy and F1 scores. The Convolutional Neural Network (CNN) model showed the highest performance with accuracy of 0.784 and an F1 score of 0.668. This indicates the significant accuracy of CNN in capturing the contextual relationships in text data for classification purposes.

The Gated Recurrent Unit (GRU) model also showed high performance with an accuracy of 0.757 and F1 score of 0.635. GRU's streamlined architecture, which is similar to LSTM but more efficient, proved to be effective in handling sequential data, which is a common characteristic of text.

The Long Short-Term Memory (LSTM) model, another type of RNN, scored an accuracy of 0.745 and F1 of 0.619 which is very close to GRU. While LSTM is designed to handle long-term dependencies in text, its performance was slightly lower than GRU in this project, possibly due to the nature of the data or the specific architecture of the model used.

The baseline model, a simple Feed-Forward Neural Network, surprisingly showed a strong performance with an accuracy of 0.765 and F1 score of 0.642. This indicates that even basic neural network structures can be quite effective for text classification tasks, especially when dealing with well-preprocessed and structured data.

The results from both machine learning and neural network models suggested useful findings. The CNN and the baseline feed-forward model in the neural network category and the Logistic Regression model in the traditional machine learning category surpasses other models in terms of accuracy. Since CNN model exhibited the highest accuracy with an accuracy of 0.784, I will further use it for classifying comments to assess public sentiment. Moving forward, the CNN model will be applied to classify the comments into positive, neutral, and negative sentiments since this classification will provide a deeper understanding of public sentiment in each category of posts and further discuss some surprising findings come from public views in terms of each category of the post.

# Discussion

After classifying the posts into categories and applying Neural Networks model to comments classification into Positive, Negative and Neutral classes, I found some interesting trends in the data. In the following paragraphs, I will provide these insights.

An examination of the classified comments reveals a diverse public view (See Figure 2 in Appendixes). Almost in all categories, the number of positive predicted comments exceeds the number of predicted negatives except crime. Overall sentiment exhibits an active and positive engagement, with receiving 339,628 positives versus 277,017 negative comments. A noteworthy finding is that except Health, Sports and Science and Technology, the difference between negatives and positives is slightly low indicating that the public sentiment is diverse supporting the idea that people have not either predominantly optimistic or pessimistic perspectives on all categories of news. The Society category stands out with a significantly high number of positive comments (211,755), overshadowing the negative (184,053) and neutral (59,130) ones. This is because the number of posts is highly correlated with the number of comments and more 67% of posts are classified in Society category. The Economy and Politics categories, while the difference between positives and negatives is smaller in volume, also display a positive sentiment, indicative of a supportive for economics and politics related topics. Conversely, Crime posts showed a higher volume of negative comments (11,946) compared to positive ones (10,920), probably reflecting public hate speech with criminal issues or cases.

After reviewing the comments in each category, the predominant[[7]](#footnote-7) sentiment for each post has been assigned according to the difference between number of positives and negatives in the post (See Figure 3 in Appendixes). The post sentiment also draws a detailed picture of public opinion. Society, once again, leads with 8,075 posts marked positive, compared to 5,387 negatives, suggesting a tendency towards a favorable view of social developments. Health, Sports, Science and Technology and World categories show a majority of positively perceived posts, reinforcing the notion that these areas are well-regarded by the public. Reflecting to finding above, Politics and Economy categories have a surprisingly very close number of positives (485 and 326) and negatives (319 and 229), suggesting that topics in these categories can resonate almost equally positively and negatively with the audience. On another hand, Crime category has predominantly high number of negative sentiment posts with 404 negatives and 287 positives showing that people tend to react negatively to Crime related posts.

Media activity from a trend perspective over time (See Figure 4 in appendixes), measured as the frequency of posts published in the channel started to surge around the beginning of the COVID-19 pandemic, due to the lockdown measures by the government and ultimately increased public activity on media as well. This heightened level of posting was sustained until the end of 2020, after which there was a gradual decrease, stabilizing at an average of around 400 posts per month in early 2022.

The volume of comments highly correlated with post published frequency, spiked during the pandemic's peak and then receded to more regular levels. Notably, the gap between positive and negative comments&posts fluctuated over time, especially during 2020, suggesting that stages of the pandemic may have polarized public negative opinion more strongly (See Figure 5 and Figure 6 in Appendixes). Both figure 5 and figure 6 suggest that since the end of 2020, public pessimistic sentiment has been gradually decreasing over time making the gap very negligible till present. It highlights the concerns of people during the pandemic and prevalence of optimistic views emergence after the pandemic.

# Conclusion

The analysis aimed to examine the public sentiment in Uzbekistan using kun.uz Instagram profile posts and people’s comments. This study provides valuable contribution into the public's reaction to media content, highlighting areas of strong engagement and sentiment. For the analysis, I used both Machine Learning and Neural Network models to classify peoples’ comments into Positive, Negative and Neutral classes. Among both Neural Network and Machine Learning models, CNN model showed best performance with an accuracy of 0.786. The findings obtained from the data represented that negligible gap between positive and negative comments in Society, Politics and Economy suggests both pessimistic and optimistic views. When it comes to categories like Health, Sports, Science and Technology and World, majority of comments and posts are positively discussed, highlighting that these areas are well-accepted by the public. Conversely, Crime content posts have been negatively perceived by the public. Another interesting finding is that during the pandemic peoples’ activity in the media has been reached at its peak as well as with its pessimistic sentiment. Since then, overall public pessimistic sentiment has been gradually decreasing over time after the end of pandemic and positive sentiment has been dominating.

In conclusion, the findings from this sentiment analysis offer overall picture of public sentiment in Uzbekistan, reflecting both the immediate reactions to current events and the broader trends in public discussion. These findings are essential for understanding the trend of public opinion and useful for policymakers as this analysis can help the development of policies that resonates with public needs.

# Bibliography

1. A.Prabhat and V. Khullar, "Sentiment classification on big data using Naïve bayes and logistic regression," 2017 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2017, pp. 1-5, doi: 10.1109/ICCCI.2017.8117734.
2. Burkhanova, S.T. (2021) Education reforms in Uzbekistan during the years of Independence. Pindus Journal of Culture, Literature, and ELT ISSN: 2792 – 1883 Volume 11. https://literature.academicjournal.io/index.php/literature/article/download/150/158
3. Corissa Steiner (2021). A Closer Look at Uzbekistan’s Privatization Push. Tashkent has promoted privatization schemes before. Is this time different? https://thediplomat.com/2021/02/a-closer-look-at-uzbekistans-privatization-push/
4. Darakchi.uz (2023). The coverage of providing the population with clean drinking water is increasing. https://darakchi.uz/oz/168088
5. Jalal, N., Mehmood, A., Choi, G. S., & Ashraf, I. (2022). A novel improved random forest for text classification using feature ranking and optimal number of trees. In Journal of King Saud University - Computer and Information Sciences (Vol. 34, Issue 6, pp. 2733–2742). Elsevier BV. https://doi.org/10.1016/j.jksuci.2022.03.012
6. President.uz (2021). Important Transformations in the Healthcare Sector Identified. https://president.uz/en/lists/view/4372
7. Shah, K., Patel, H., Sanghvi, D. et al. A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. Augment Hum Res 5, 12 (2020). <https://doi.org/10.1007/s41133-020-00032-0>
8. Tukhtasinov, I. and Asif Noor, M. (2023) Significance of higher education in advancing growth and progress in Uzbekistan. https://thediplomaticinsight.com/significance-of-higher-education-in-advancing-growth-and-progress-in-uzbekistan.
9. World Bank Report (2021). Assessing Uzbekistan’s Transition. https://documents1.worldbank.org/curated/en/862261637233938240/pdf/Full-Report.pdf
10. Yoon Kim. 2014. [Convolutional Neural Networks for Sentence Classification](https://aclanthology.org/D14-1181). In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.
11. Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent Neural Network Regularization (Version 5). arXiv. https://doi.org/10.48550/ARXIV.1409.2329

# Appendixes

Figure 2. The distribution of comments predicted as ‘Positive’, ‘Negative’ and ‘Neutral’ across categories

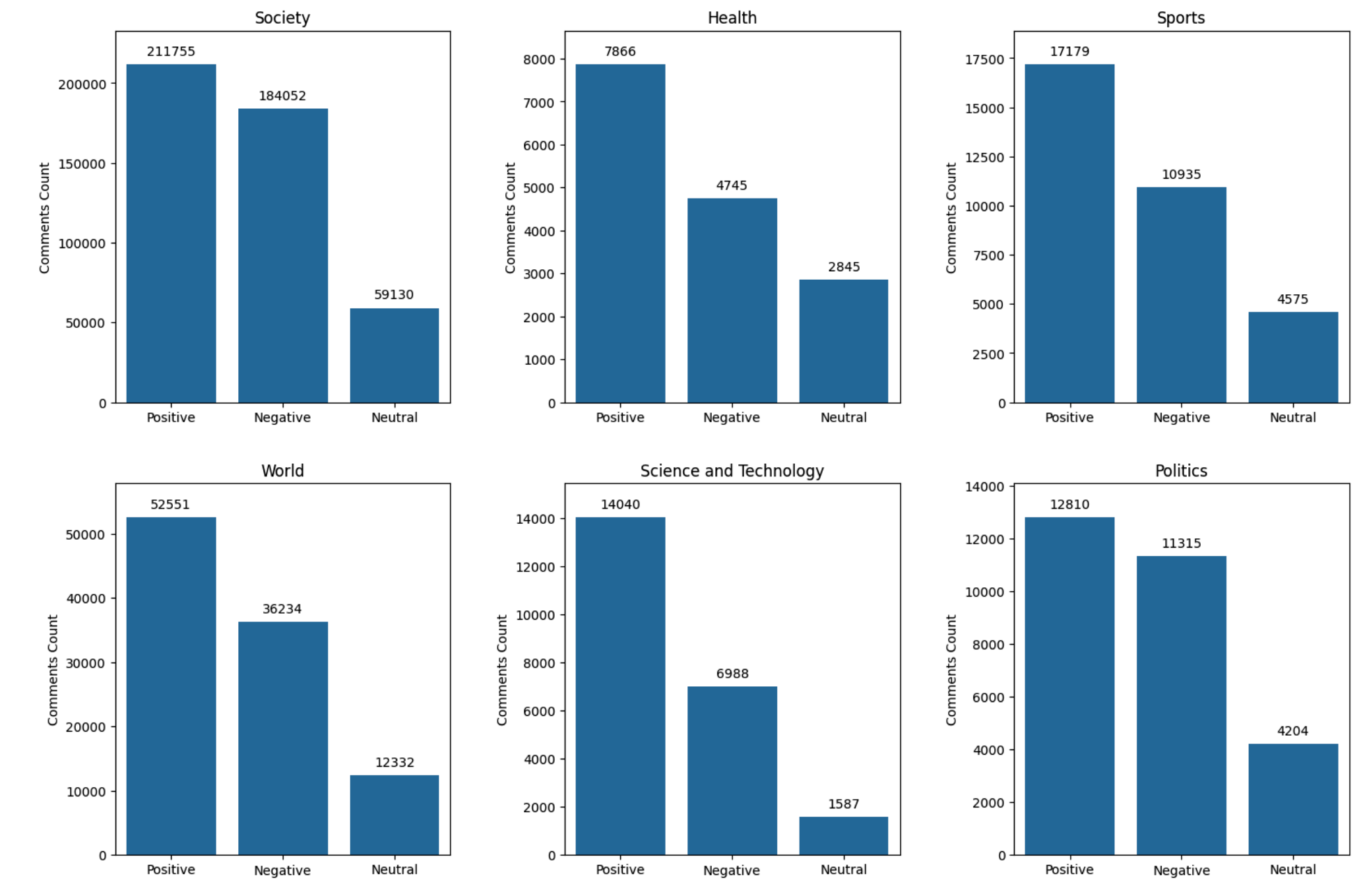
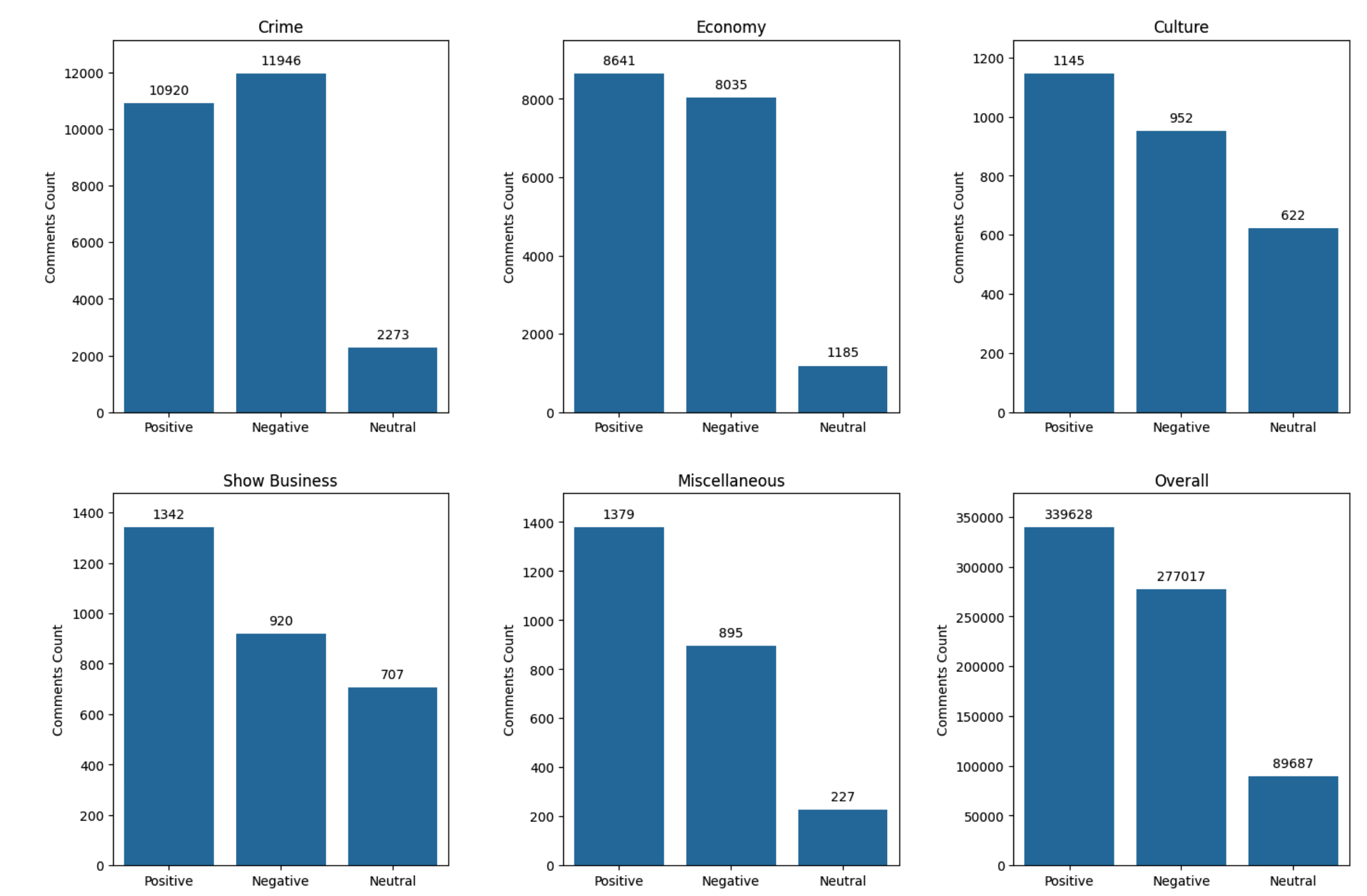
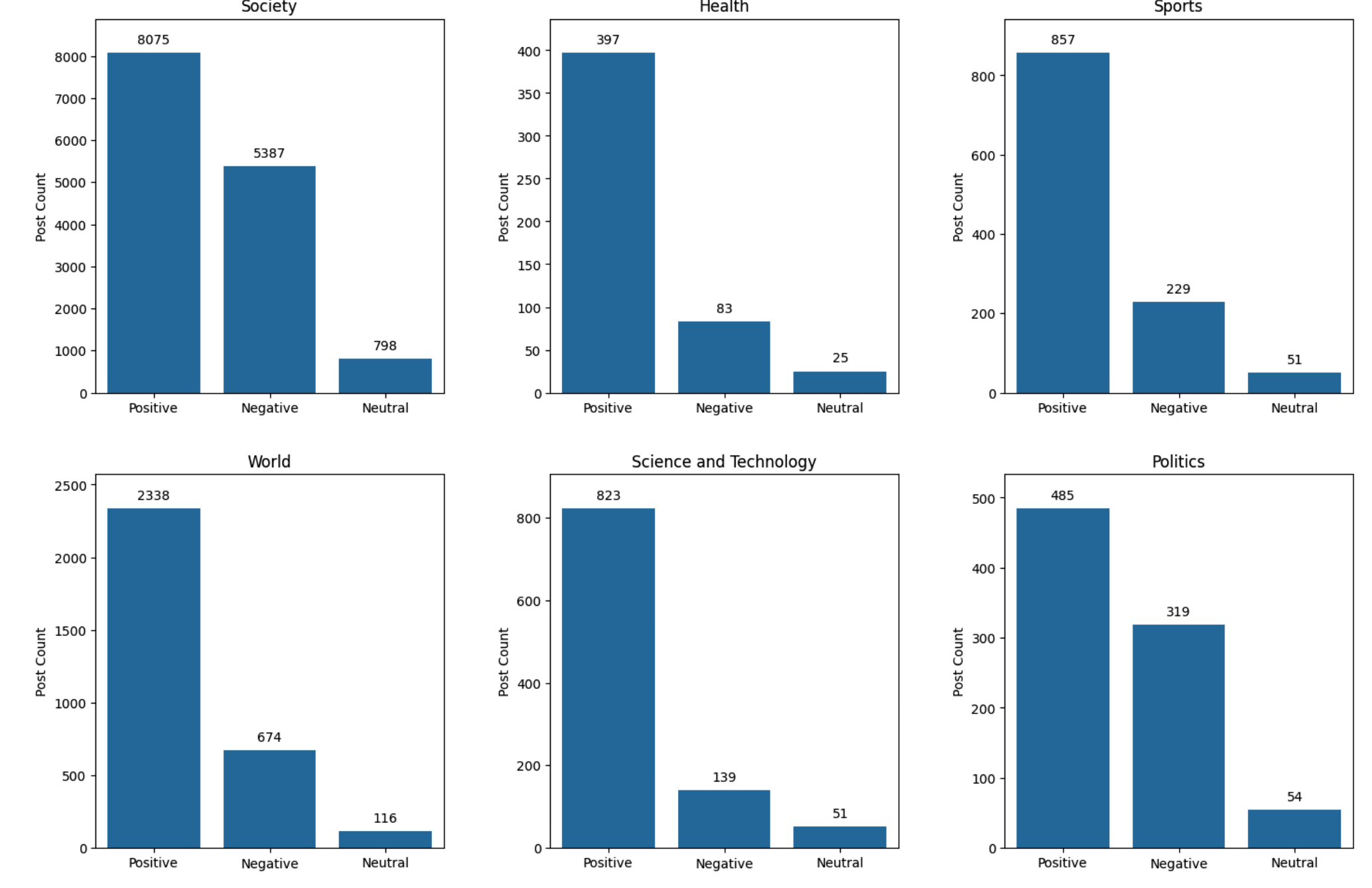
 

Figure 3. The distribution of posts by Predominant Labels assigned as ‘Positive’, ‘Negative’ and ‘Neutral’ across categories



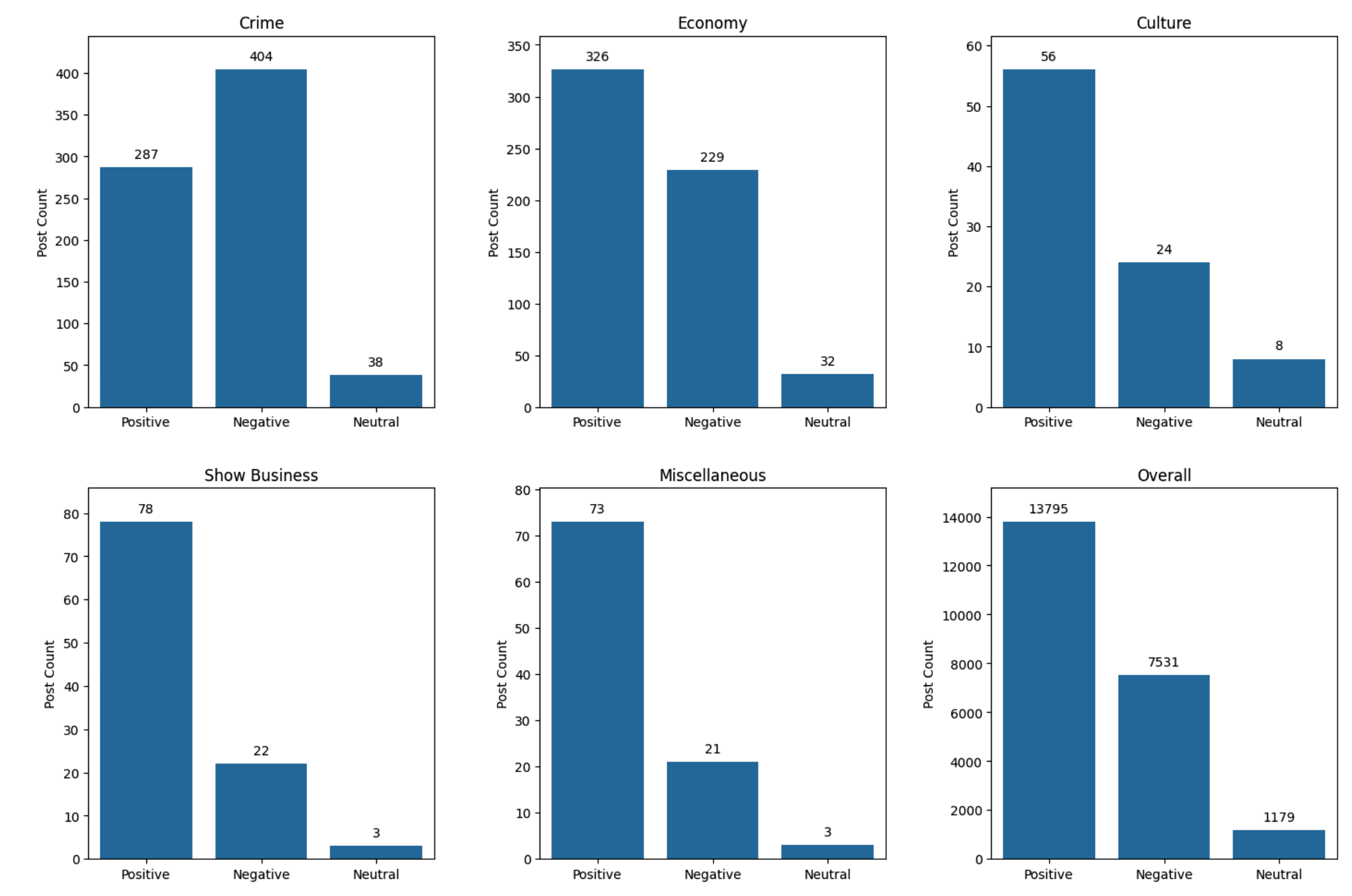
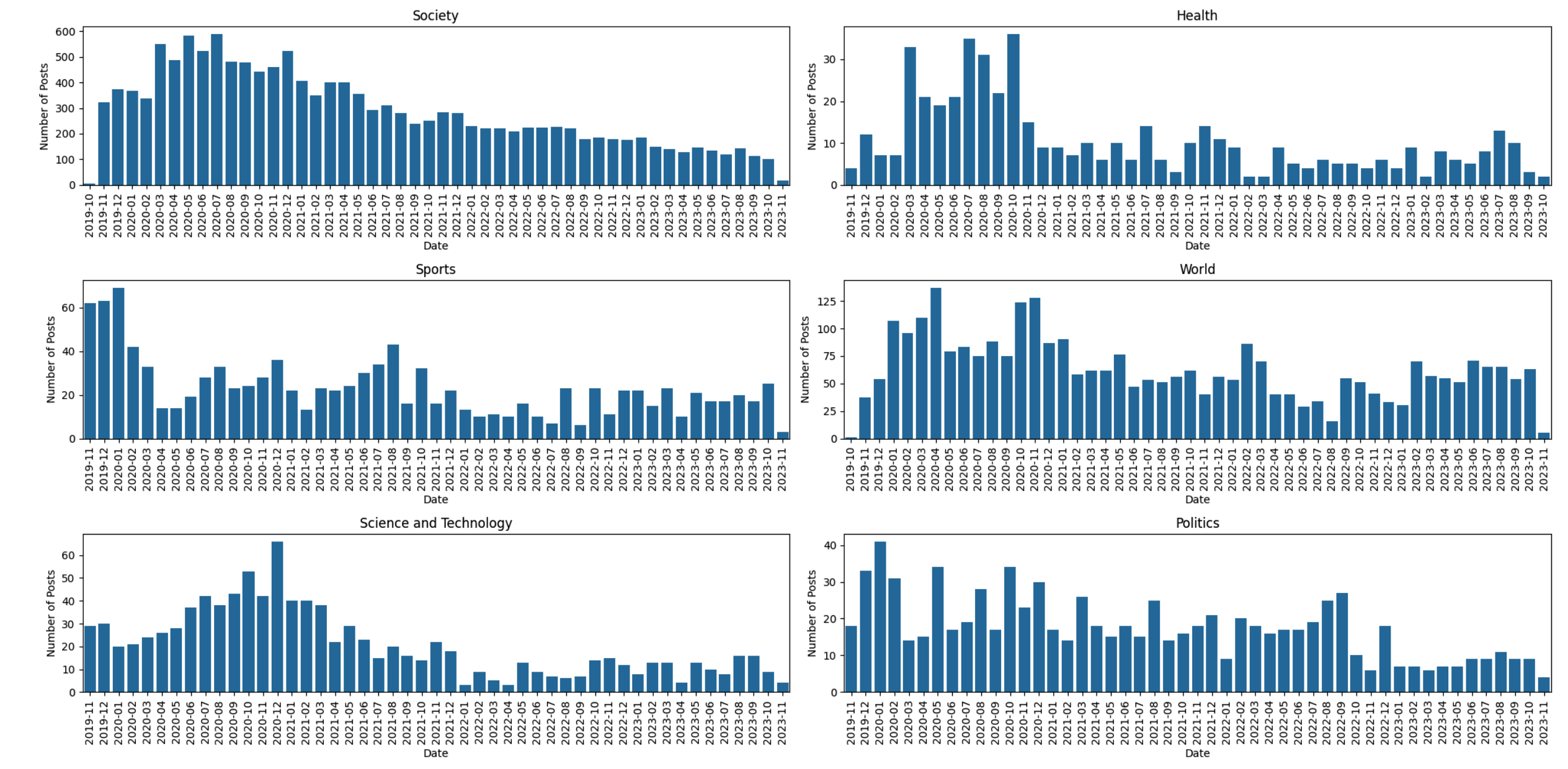


Figure 4. The media activity measured by frequency of published posts in the media channel.



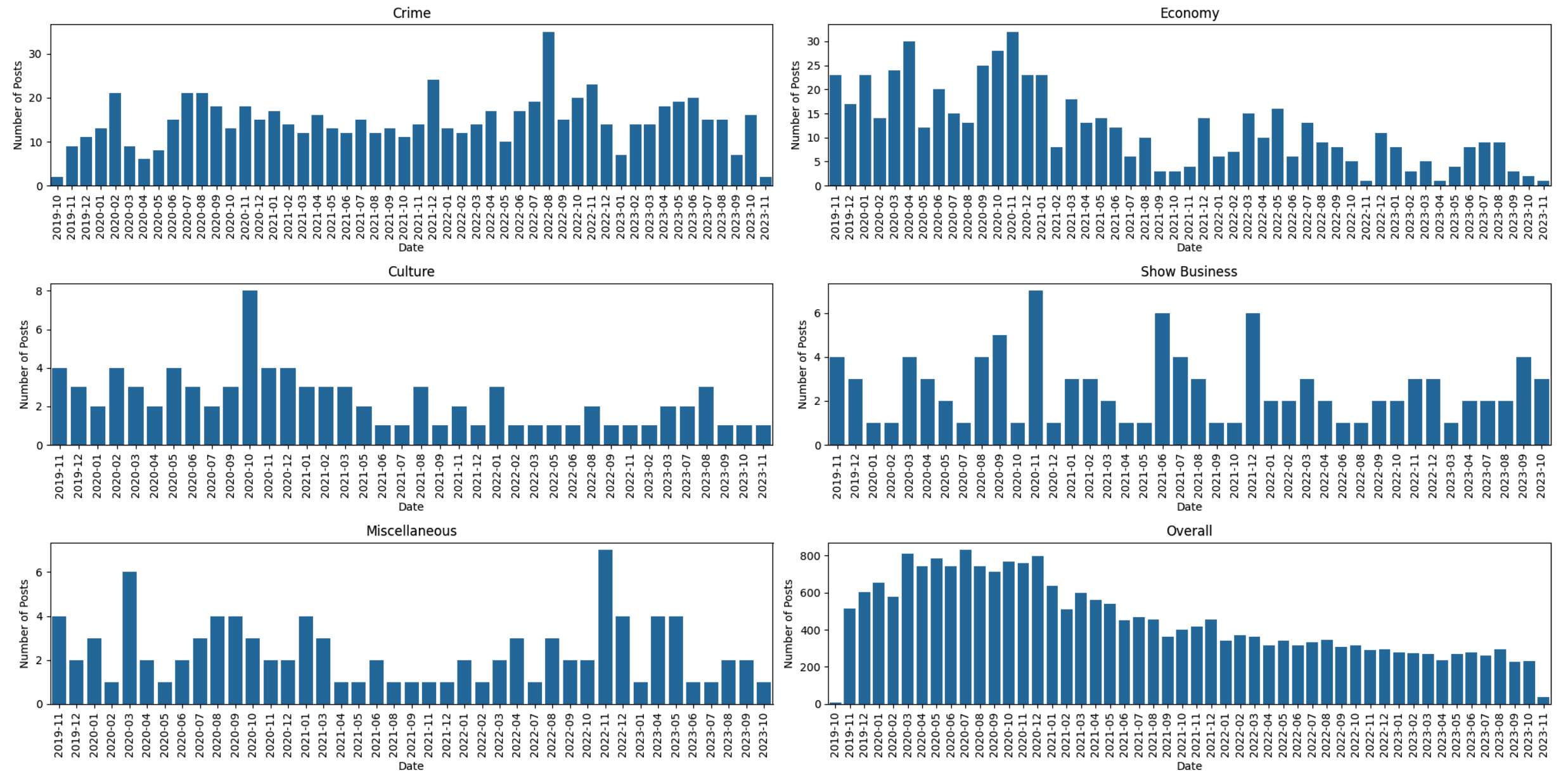


Figure 5. Percentage of Negative, Positive and Neutral Comments over time

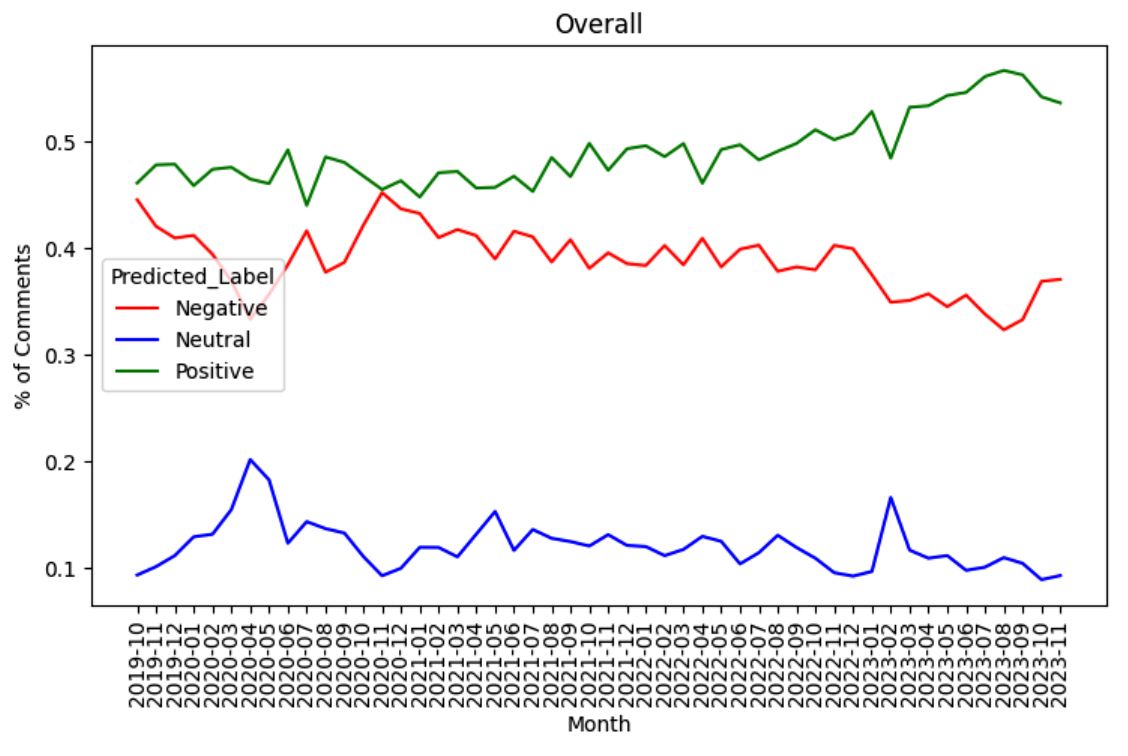
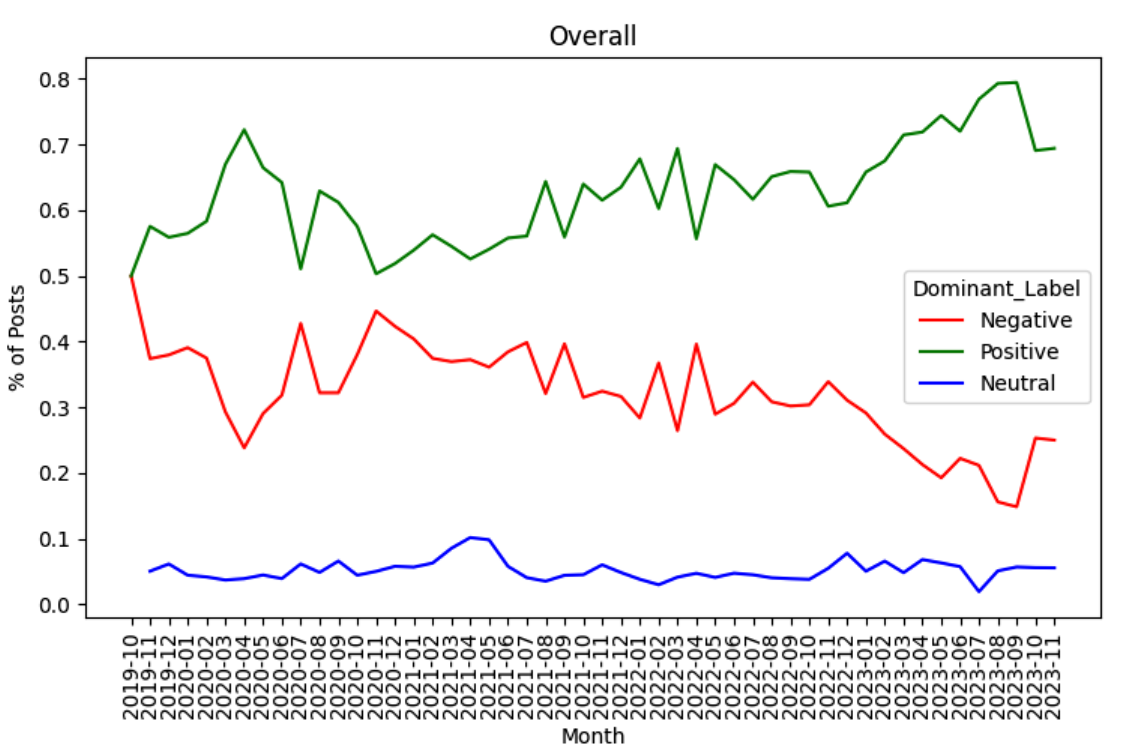


Figure 6. Percentage of Negative, Positive and Neutral Posts over time



1. GitHub repository for project: https://github.com/shakhrukhb/UzbekistanReformSentimentAnalysis [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)
3. Preprocessing steps are provided in preprocess.py script. [↑](#footnote-ref-3)
4. Classification of posts into categories is provided in the classification.py script. [↑](#footnote-ref-4)
5. Machine Learning models are provided in ml.py script. [↑](#footnote-ref-5)
6. The implementation of Neural Network models is provided nn.py script. [↑](#footnote-ref-6)
7. If the number of positive comments exceeds that of negatives, Positive label has been assigned to the post, if not exceeds, Negative label. If they are equal, Neutral label has been assigned to the post. [↑](#footnote-ref-7)