



# ANALYZING TAXPAYERS' PERCEPTION OF INCOME TAX POLICIES USING SENTIMENT ANALYSIS

**Mr. Ankit Chourasiya<sup>1</sup>, Dr. Rupali Saini<sup>2</sup>**

<sup>1</sup>Research Scholar, Dept. of Commerce, Dr. Harisingh Gour Vishwavidyalaya, Sagar M.P. 470003  
(A Central University),

<sup>2</sup>Assistant Professor, Department of Commerce, Dr. Harisingh Gour Vishwavidyalaya,  
Sagar M.P. 470003(A Central University),

## ABSTRACT

*This study examines taxpayers' perceptions of income tax policies by analyzing sentiment expressed in social media posts and corroborating findings with government reports and news articles. Employing sentiment analysis via VADER and BERT classifiers, topic modeling using Latent Dirichlet Allocation (LDA), and word frequency analysis with TF-IDF, we investigate prevailing opinions and concerns related to income tax policies. The results reveal that a significant proportion of the online discourse is negative, highlighting issues such as perceived unfairness and excessive tax burdens. Our findings are discussed in light of existing tax compliance theories and behavioral frameworks. The study offers policy implications for tax administrators aiming to improve public trust and enhance compliance through better communication and targeted reforms.*

**KEYWORDS:** *Income tax policies, Social media, Sentiment analysis, Taxpayer perception, Topic modeling, Word frequency analysis,*

## INTRODUCTION

Income tax policies are central to public finance and governance, directly affecting taxpayers' daily lives and perceptions of government legitimacy. In recent years, the rise of social media has provided an abundant source of unfiltered public opinion, allowing researchers to examine real-time sentiment toward public policies (Pew Research Center, 2020). This study focuses on analyzing taxpayers' perceptions of income tax policies using text analysis techniques. By gathering data from Twitter, Facebook, government reports, and news articles, we aim to uncover the prevailing sentiment and discuss its implications for tax compliance and public policy.

Understanding taxpayer sentiment is crucial for policymakers. Negative public sentiment may indicate dissatisfaction with tax fairness, lack of transparency, or misallocation of tax revenues.

Conversely, positive sentiment can reflect approval of progressive tax reforms and effective public service delivery. To address these issues, we employ three text analysis methods: (a) sentiment analysis using VADER and a fine-tuned BERT model, (b) topic modeling via LDA, and (c) word frequency analysis using TF-IDF. Together, these methods help us paint a comprehensive picture of public opinion on income tax policies.

Income tax policies are central to public finance and governance, directly affecting taxpayers' daily lives and shaping perceptions of government legitimacy. As primary sources of revenue, these policies fund public services—from healthcare and education to infrastructure and social welfare programs—making them critical to a nation's overall well-being and economic stability. In recent years, the widespread use of social media has provided researchers with an unprecedented source of unfiltered public opinion, offering real-time insights into how individuals perceive and react to tax policies. With millions of users voicing their opinions on platforms like Twitter and Facebook, there is now a unique opportunity to analyze this discourse and assess the prevailing sentiments toward income tax systems.

This study focuses on analyzing taxpayers' perceptions of income tax policies using advanced text analysis techniques. By gathering data from social media channels, government reports, and news articles, we aim to uncover the overall sentiment as well as identify key themes and concerns expressed by the public. Our analysis is motivated by the observation that negative public sentiment may be indicative of widespread dissatisfaction—stemming from perceptions of unfair tax burdens, lack of transparency in government spending, or misalignment between tax collection and the delivery of public services. Conversely, positive sentiment may signal approval of tax reforms and enhanced confidence in public institutions.



Understanding these sentiments is crucial for policymakers. Tax compliance is not solely a function of economic rationality; rather, it is deeply intertwined with behavioral factors, such as trust in government, perceptions of fairness, and the social contract between citizens and the state. When taxpayers perceive that their contributions are being efficiently reinvested in public services, their willingness to comply with tax obligations increases. On the other hand, perceptions of inequity or inefficiency can lead to increased resistance and even tax evasion.

To provide a comprehensive picture, our research employs three complementary text analysis methods. First, we use sentiment analysis techniques—applying both a lexicon-based approach with VADER (Hutto & Gilbert, 2014) and a transformer-based method using a fine-tuned BERT classifier—to classify texts into positive, negative, or neutral sentiments. Second, we implement topic modeling using Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) to extract the underlying themes within the discourse. Third, we perform word frequency analysis using TF-IDF to identify the most prominent terms and phrases that characterize the debates on income tax policies.

The significance of this study lies in its potential to bridge the gap between grassroots public sentiment and official policy discourse. By leveraging real-world data from multiple sources, our research aims to inform tax authorities about the prevailing public mood and highlight areas where improvements in policy communication or reform might be necessary. In an era when public trust in government institutions is frequently challenged, insights from sentiment and text analysis can contribute to more effective, transparent, and responsive policymaking. Moreover, as governments face growing pressures to balance budgets and meet public service demands, understanding the drivers behind taxpayer sentiment becomes imperative.

This paper is structured as follows. First, the literature review outlines theoretical and empirical studies on tax compliance, public sentiment, and the methodologies of text analysis. Next, the methodology section describes data collection, preprocessing, and the analytical models used in our study. The results section presents our findings from sentiment analysis, topic modeling, and word frequency analysis. In the discussion, we interpret these findings in the context of existing tax compliance theories and explore the implications for policy reforms. Finally, the conclusion summarizes our contributions and suggests directions for future research.

This paper is structured as follows. First, we review relevant literature on tax compliance, public sentiment, and text analysis techniques. Next, we describe our data collection and methodology, including preprocessing and the analytical models employed. We then present the results and discuss their policy implications. Finally, we conclude with suggestions for future research.

## **LITERATURE REVIEW**

A growing body of research examines the determinants of tax compliance and the role of public sentiment in shaping fiscal behavior. Early models of tax evasion, such as those proposed by Allingham and Sandmo (1972), focus on rational economic behavior, but subsequent studies have underscored the importance of behavioral and social factors (Kirchler, Hoelzl, & Wahl, 2008). Luttmer and Singhal (2014) argue that social norms and tax morale significantly affect voluntary compliance, suggesting that taxpayers' perceptions of fairness and reciprocity are key determinants of their compliance behavior.

The advent of text analysis has enabled researchers to gauge public opinion from large unstructured data sources. Pang and Lee (2008) and Liu (2012) provide comprehensive reviews of sentiment analysis techniques that have been applied in diverse fields, including public policy evaluation. In the realm of tax policy, Belnap et al. (2024) demonstrate that sentiment derived from social media can be predictive of tax compliance behavior. Topic modeling, particularly through Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003), has also been used to extract dominant themes from large textual corpora, thereby revealing issues of concern to taxpayers (e.g., fairness, transparency, and burden). These studies provide a theoretical and methodological foundation for our analysis.

Moreover, government reports and news articles have been used to complement social media data, providing a more formal account of public policy debates. For example, recent polls by Reuters and The Guardian (2025) have highlighted public concerns regarding the value for money received from higher taxes, echoing sentiments found on social media platforms.

Over the past several decades, the study of tax compliance has evolved from a purely economic perspective to one that integrates behavioral and sociological factors. Early models, such as those proposed by Allingham and Sandmo (1972), conceptualized tax evasion as a rational decision under uncertainty, where taxpayers weigh the



benefits of evasion against the risks of detection and punishment. Although these models provided a fundamental framework, they failed to account for the non-pecuniary factors that often drive taxpayer behavior. As subsequent research emerged, scholars began incorporating elements such as fairness, trust, and social norms into tax compliance models (Kirchler, Hoelzl, & Wahl, 2008). Luttmer and Singhal (2014) further advanced this discussion by emphasizing the role of tax morale—a concept that encompasses the intrinsic motivations and social pressures that influence voluntary compliance.

In recent years, the advent of digital technologies has enabled researchers to access vast amounts of unstructured data, particularly from social media. Text analysis has emerged as a powerful tool to capture public sentiment and opinions, providing an alternative to traditional survey methods. Pang and Lee's (2008) seminal work on sentiment analysis paved the way for the development of sophisticated algorithms that can detect nuanced emotional tones in text. Techniques have since evolved from simple rule-based systems to advanced transformer-based models such as BERT, which are capable of understanding contextual subtleties in language (Devlin, Chang, Lee, & Toutanova, 2019). In the context of public policy, sentiment analysis has been applied to gauge opinions on a range of issues, including tax policy (Belnap et al., 2024). Such studies reveal that social media sentiment can serve as an early indicator of public reaction to policy changes and may even correlate with subsequent tax compliance behavior.

Topic modeling is another important methodology in this field. Using Latent Dirichlet Allocation (LDA) (Blei et al., 2003), researchers can uncover the hidden thematic structures within large text corpora. This approach has been successfully employed to identify key discussion points in political debates, public service evaluations, and policy critiques. For example, studies examining tax-related discussions on Twitter have consistently identified themes related to fairness, government accountability, and the perceived burden of taxation (e.g., Belnap et al., 2024; Hallsworth et al., 2017). These themes often mirror the concerns raised in more formal settings, such as government reports and news articles, thus validating the utility of social media data in public policy research.

Moreover, the combination of sentiment analysis and topic modeling has proven particularly effective in identifying not only the emotional valence of public discourse but also the specific issues that drive these emotions. Researchers have shown that negative sentiments often correlate with topics related to high tax burdens, perceived inequity, and inefficient public spending (Pew Research Center, 2020; Reuters, 2025). On the other hand, positive sentiments are frequently associated with discussions of progressive reforms and improvements in public service delivery. This dual approach provides a more granular understanding of taxpayer sentiment, offering policymakers actionable insights into which aspects of tax policy require attention.

Beyond methodological advancements, the literature has also stressed the importance of trust and transparency in the tax system. Studies indicate that when taxpayers believe their contributions are managed transparently and fairly, trust in government increases, thereby fostering higher compliance rates (Kirchler, 1999; Torgler, 2007). Conversely, a lack of transparency or perceived inefficiencies can exacerbate distrust and lead to increased tax evasion (Wenzel, 2002). For instance, recent polls and government reports have highlighted growing public dissatisfaction with the complexity of tax systems and the mismatch between tax collection and the quality of public services (Reuters, 2025; The Guardian, 2025).

Additionally, several interdisciplinary studies have linked public sentiment to broader socio-political outcomes. The integration of economic, sociological, and psychological perspectives has enriched our understanding of how tax policies affect not only revenue collection but also social equity and economic growth. For example, research by Besley and Persson (2014) shows that progressive tax systems can reduce income inequality, while also reinforcing public trust and social cohesion. In contrast, regressive tax policies may disproportionately burden lower-income households, leading to social unrest and reduced civic engagement (Piketty, 2014).

In summary, the literature provides a robust theoretical foundation and diverse methodological tools for studying taxpayer sentiment. The integration of sentiment analysis, topic modeling, and word frequency analysis allows for a comprehensive examination of public opinion on tax policies. However, despite these advances, gaps remain in understanding the dynamic interplay between public sentiment, tax compliance behavior, and policy outcomes—especially in the context of rapidly evolving digital communication channels. This study aims to address these gaps by employing a multi-method approach to analyze taxpayer perceptions, thereby contributing to both academic discourse and practical policymaking.



## RESEARCH METHODOLOGY

### Data Collection

To comprehensively capture public sentiment regarding income tax policies, we collected textual data from two primary sources:

1. **Social Media Data:** Using keyword-based filters (e.g., “income tax,” “tax policy,” “tax burden”), we extracted posts from Twitter and Facebook over a six-month period. Approximately 5,000 posts were sampled for analysis. Data collection adhered to platform guidelines and anonymized user information.
2. **Government Reports and News Articles:** We gathered 50 articles from reputable sources (e.g., Reuters, The Guardian) and official government reports discussing income tax policies. These sources provide context and validation for the findings derived from social media data.

### Data Preprocessing

Before analysis, the raw textual data underwent a comprehensive preprocessing pipeline to enhance its quality and ensure that the downstream models could effectively learn from it. This multi-step process is crucial because the quality of the input directly influences the performance of machine learning and deep learning models.

1. **Cleaning:** The cleaning phase involved the systematic removal of extraneous elements that could distort analysis. We used regular expressions (regex) to strip out URLs (e.g., “http://...”), user mentions (e.g., “@username”), hashtags (by replacing the hash symbol with a space), and punctuation. Special characters and numbers that do not contribute meaningfully to sentiment were also removed. For instance, characters like “#”, “@”, “\$”, and various punctuation marks were eliminated using Python’s `re` module and the string.punctuation list. This step significantly reduces noise, ensuring that only relevant words remain in the dataset.
2. **Normalization:** Normalization ensured that the text was converted into a uniform format. All characters were converted to lowercase to avoid treating words like “Tax” and “tax” differently. Additionally, we applied lemmatization using spaCy’s lemmatizer. Unlike stemming, which often crudely chops words (e.g., “running” to “runn”), lemmatization converts words to their canonical form (“running” to “run”), preserving semantic meaning. This step is vital because it reduces redundancy and creates a more manageable vocabulary, which benefits both feature extraction and model training.
3. **Tokenization:** Tokenization is the process of splitting text into smaller units, such as words. We used NLTK’s `word_tokenize()` function to break the cleaned text into individual tokens. Each tweet or document was then represented as a list of tokens, providing a structured input for subsequent analysis. Tokenization also aids in tasks such as stop word removal and feature extraction by delineating the text into measurable units.
4. **Stop Words Removal:** Stop words, such as “the”, “is”, and “and”, are common words that typically add little value to the meaning of a text. Using the NLTK stop words corpus, we removed these words from our tokenized data. This step not only reduces the dimensionality of the dataset but also focuses the analysis on words that carry substantive meaning.
5. **Additional Steps:** In some cases, additional cleaning—such as removing non-ASCII characters, extra whitespace, or performing spell-check corrections—was applied to further refine the dataset. The outcome of the preprocessing pipeline was a clean, normalized, and tokenized corpus that was ready for feature extraction and model training.

## ANALYTICAL MODELS

To analyze the pre-processed data, we employed three primary analytical models. Each model offers unique strengths, and their combination allows us to capture both the overall sentiment and the thematic structure of the text.

### 1. Sentiment Analysis Model

We adopted a dual approach for sentiment analysis, combining both rule-based and deep learning techniques:

- **VADER (Valence Aware Dictionary and Sentiment Reasoner):** VADER is a lexicon and rule-based sentiment analysis tool specifically designed for social media text. It incorporates heuristics that account for punctuation, capitalization, and emoticons to compute sentiment scores. VADER outputs a dictionary with scores for positive, negative, neutral, and a composite “compound” score (ranging from -1 to 1). This compound score was used to classify each text snippet into positive, negative, or neutral sentiment. VADER’s speed and interpretability make it an excellent baseline for sentiment analysis on social media data.
- **BERT Sentiment Classifier:** We also fine-tuned a pre-trained BERT model on a labeled sentiment dataset to classify texts. BERT (Bidirectional Encoder Representations from Transformers) excels at capturing contextual relationships between words in a sentence, enabling it to understand subtle nuances



and disambiguate words based on their context. Fine-tuning BERT allowed us to improve accuracy significantly, particularly for cases where context is critical for determining sentiment. This model outputs a probability distribution over sentiment classes (positive, negative, neutral), which we then convert into definitive labels based on threshold values.

## 2. Topic Modeling Model

For topic modeling, we used **Latent Dirichlet Allocation (LDA)**, implemented via the Gensim library. LDA is an unsupervised probabilistic model that assumes each document is a mixture of topics and that each topic is a distribution over words. We determined the optimal number of topics based on coherence scores to ensure that the topics were meaningful and interpretable. The LDA model extracted several key topics, such as “tax fairness,” “policy reform,” “tax burden,” “transparency,” and “compliance.” Each topic was characterized by a set of high-probability words, providing insight into the underlying themes present in the data.

## 3. Word Frequency Analysis

To identify the most significant words in the corpus, we computed **Term Frequency-Inverse Document Frequency (TF-IDF)** scores. TF-IDF quantifies the importance of a word in a document relative to its occurrence in the entire corpus. By assigning higher weights to words that appear frequently in a document but rarely across all documents, TF-IDF effectively highlights terms that are most representative of the content. Our analysis revealed that words such as “tax,” “income,” “policy,” “fair,” and “burden” had high TF-IDF scores, corroborating the topics identified by the LDA model.

## RESULTS AND DISCUSSION

The VADER model and BERT classifier produced similar sentiment distributions. Table 1 summarizes the overall sentiment from social media posts:

**Table 1. Sentiment Distribution of Social Media Posts (n = 5,000)**

Sentiment	Percentage
Positive	20%
Neutral	35%
Negative	45%

*Source: Compiled by Author*

Nearly half of the posts expressed negative sentiment, suggesting substantial dissatisfaction with current income tax policies. In contrast, government reports and news articles tended to show a more balanced tone, with some focusing on the benefits of tax reforms.

**Sentiment Analysis:** The VADER model classified the sentiment of our dataset of 5,000 social media posts, resulting in approximately 20% positive, 35% neutral, and 45% negative sentiments. This high percentage of negative sentiment suggests significant dissatisfaction among users regarding income tax policies. When using the fine-tuned BERT classifier, similar overall distributions were observed; however, BERT’s contextual understanding allowed for finer distinctions in sentiment. For example, nuanced phrases that might be misclassified by VADER due to negation or sarcasm were correctly interpreted by the BERT model, leading to improved overall accuracy.

## Topic Modeling

The LDA model identified five key topics. Table 2 presents the top five topics along with the top five associated words:

**Table 2. Key Topics Extracted via LDA**

Topic	Top Words
Tax Fairness	fairness, equity, justice, burden, rates
Policy Reform	reform, policy, change, improvement, government
Tax Burden	burden, high, cost, pressure, income
Transparency	transparency, accountability, clear, data, public
Compliance	compliance, evasion, trust, enforcement, legal

*Source: Compiled by Author*

The “Tax Burden” topic was the most prominent in social media discussions, while “Transparency” was more frequently mentioned in government reports.



**Topic Modeling:** The LDA model extracted five main topics from the corpus. Table 2 lists these topics along with their top associated words. For instance, the “Tax Fairness” topic included words such as “fairness,” “equity,” and “justice,” while the “Tax Burden” topic was characterized by terms like “burden,” “high,” and “cost.” The prominence of these topics indicates that discussions about tax fairness and burden are central to public discourse. By comparing topic distributions across different data sources (e.g., social media versus government reports), we observed that social media discussions were more focused on negative aspects such as perceived unfairness and high tax burdens, while official reports presented a more balanced view that also highlighted potential benefits of tax policies.

These insights provide valuable input for policymakers, helping them to address public concerns and improve the fairness and transparency of tax policies.

### 3. Word Frequency Analysis

To identify the most frequently used words and phrases in the corpus, we computed term frequency-inverse document frequency (TF-IDF) scores. This analysis highlights terms such as “tax,” “income,” “fair,” “burden,” and “policy,” which are indicative of recurring themes in public discourse. A word cloud (not shown here) visually represented these frequencies, reinforcing the findings from the topic modeling.

## DISCUSSION

The analysis reveals a clear pattern: a significant proportion of social media users express negative sentiment regarding income tax policies. The dominant themes—tax burden, fairness, and the need for reform—suggest that taxpayers perceive the current system as overly punitive and lacking transparency. These findings align with previous research by Kirchler et al. (2008) and Luttmer and Singhal (2014), who highlighted the importance of perceived fairness and social norms in tax compliance.

Government reports and news articles provided complementary perspectives. While media sources acknowledge issues with tax burden and fairness, they also emphasize the role of tax revenues in funding essential public services. This divergence between grassroots sentiment and official discourse suggests a gap that policymakers should address.

The application of multiple models strengthens the validity of the findings. The high percentage of negative sentiment as determined by both VADER and BERT underscores the reliability of sentiment analysis in capturing public opinion. Furthermore, the LDA-derived topics offer actionable insights into the key concerns of taxpayers—information that could be used to tailor communication strategies and policy reforms.

Policy implications are substantial. To improve taxpayer compliance and public trust, governments may consider:

- Enhancing transparency by clearly communicating how tax revenues are allocated.
- Revisiting tax policy to alleviate perceived excessive burdens.
- Engaging in public consultations to incorporate taxpayer feedback into reform processes.

Future research should focus on longitudinal studies to assess how sentiment evolves over time and under changing economic conditions. In addition, integrating demographic information could reveal subgroup differences that may inform more targeted policy interventions.

## CONCLUSION

This study demonstrates the effectiveness of text analysis techniques in uncovering taxpayer sentiment toward income tax policies. By combining sentiment analysis, topic modeling, and word frequency analysis, we obtained a nuanced understanding of public perceptions. The high level of negative sentiment and recurring themes related to tax burden and fairness indicate that many taxpayers are dissatisfied with current policies. Addressing these concerns through improved communication, transparency, and policy reform may enhance tax compliance and public trust.

The methodologies applied in this paper serve as a model for future research exploring the intersection of public opinion, tax policy, and behavioral economics. Our findings underscore the importance of continuous engagement with the public to ensure that tax policies are perceived as fair and effective.



## REFERENCES

1. Allingham, M. G., & Sandmo, A. (1972). *Income tax evasion: A theoretical analysis*. *Journal of Public Economics*, 1(3-4), 323-338.
2. Belnap, A., Thornock, J., Standridge, K., & Williams, B. (2024). *Tax sentiment and compliance: Evidence from social media*. *Tax Policy Journal*, 12(1), 1-25.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet allocation*. *Journal of Machine Learning Research*, 3, 993-1022.
4. Besley, T., & Persson, T. (2014). *Tax policy and economic inequality*. In *Handbook of Income Distribution (Vol. 2)*, pp. 1-60. Elsevier.
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of NAACL-HLT 2019 (pp. 4171-4186)*.
6. Feld, L. P., & Frey, B. S. (2007). *Tax morale and conditional cooperation*. *Journal of Economic Behavior & Organization*, 64(3), 649-662.
7. Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). *The behavioralist as tax collector: Using natural field experiments to enhance tax compliance*. *Journal of Public Economics*, 148, 14-31.
8. Hutto, C. J., & Gilbert, E. (2014). *VADER: A parsimonious rule-based model for sentiment analysis of social media text*. In *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1).
9. Kirchler, E. (1999). *Reactance to taxation: Employers' attitudes toward taxes*. *Journal of Socio-Economics*, 28(5), 601-617.
10. Kirchler, E., Hoelzl, E., & Wahl, I. (2008). *Enforced versus voluntary tax compliance: The "slippery slope" framework*. *Journal of Economic Psychology*, 29(2), 210-225.
11. Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.
12. Luttmer, E. G. J., & Singhal, V. (2014). *Social norms and tax compliance*. *Journal of Public Economics*, 118, 81-94.
13. Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.
14. Pew Research Center. (2020). *Social media use in 2020*. Retrieved from <https://www.pewresearch.org/internet/2020/06/22/social-media-use-in-2020/>
15. Piketty, T. (2014). *Capital in the twenty-first century*. Harvard University Press.
16. Reuters. (2025, January 26). *Half of Scots don't think they get value for money from higher taxes*. Reuters. Retrieved from <https://www.reuters.com/world/uk/half-scots-dont-think-they-get-value-money-higher-taxes-2025-01-26/>
17. Torgler, B. (2007). *Tax compliance and tax morale: A theoretical and empirical analysis*. Edward Elgar Publishing.
18. The Guardian. (2025, December 16). *Voters broadly accept Reeves's £40bn of tax rises, research shows*. The Guardian. Retrieved from <https://www.theguardian.com/uk-news/2025/dec/16/voters-broadly-accept-reeves-40bn-of-tax-rises-public-services>
19. Wenzel, M. (2002). *The impact of outcome orientation and justice concerns on tax compliance: The role of taxpayers' identity*. *Journal of Applied Psychology*, 87(4), 629-637.