

Article

Not peer-reviewed version

Evaluating the Effectiveness of Commonly Used Sentiment Analysis Models for the Second Indochina War

Arijit Chakraborty *

Posted Date: 19 December 2024

doi: 10.20944/preprints202412.1670.v1

Keywords: Sentiment Analysis; VADER; DistilBERT; Vietnam War; Historical Texts; NLP



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Evaluating the Effectiveness of Commonly Used Sentiment Analysis Models for the Second Indochina War

Arijit Chakraborty

Independent Researcher; arijic284@outlook.com

Abstract: Sentiment analysis, a subfield of Natural Language Processing (NLP), is widely used to analyze public opinion, yet its application to historical and political nuanced events are unexplored. This study evaluates the effectiveness of two popular open-source sentiment analysis models—VADER, a lexicon-based approach, and DistilBERT, a transformer-based deep learning model—in analyzing texts related to a historical and politically nuanced event: The Second Indochina War, commonly known as the Vietnam War. Using a dataset categorized into Pro-Vietnamese, Pro-American, and Civilian perspectives, the models were assessed for their ability to capture emotional and ideological nuances. While VADER proved efficient for informal texts, it oversimplified complex narratives. DistilBERT captured subtle context but struggled with ideological undertones. Results highlight the challenges of applying existing sentiment models to historically significant data. The study emphasizes the need for hybrid approaches combining automated tools with human analysis to improve accuracy. Future work could expand datasets and incorporate advanced models for a deeper understanding of politically charged texts.

Keywords: Sentiment Analysis; VADER; DistilBERT; Vietnam War; Historical Texts; NLP

Introduction

Sentiment analysis, a subfield of Natural Language Processing (NLP), has gained significant prominence in recent years for its remarkable ability to extract subjective information from large datasets. It focuses on determining whether an opinion, emotion, or expression is positive, negative, or neutral. This technique is actively used in various domains, including social media analysis, consumer feedback evaluation, and gauging public sentiment.

While sentiment analysis has traditionally been applied to these domains, it is slowly gaining prominence in analyzing historically significant and politically nuanced events. The Vietnam War, also known as the Second Indochina War (1954–1975), stands out as one of the most ideologically and emotionally intricate conflicts of the 20th century. It encompasses diverse viewpoints, ranging from the struggle for national liberation to the broader geopolitical dynamics of the Cold War. Analyzing the sentiments embedded in texts related to the Vietnam War provides valuable insights into the emotional, ideological, and rhetorical dimensions that shaped public opinion and historical narratives. However, the complexity of such scenarios often presents challenges, particularly for tools that are not trained on extensive datasets or lack substantial computing power.

Many existing sentiment analysis models are trained on limited datasets, which do not effectively capture the historical, ideological, and rhetorical undertones of a politically charged narrative such as the Vietnam War. This leads to an important question: Are current language models, such as VADER and DistilBERT, capable of effectively analyzing the sentiments found in such intricate historical situations while also accounting for their emotional and rhetorical undertones? VADER, for example, is a lexicon-based model primarily designed for social media content and has not been specifically trained to analyze politically nuanced historical texts. Despite this, it offers valuable insights into how modern, short-form expressions—often found in social media—differ from the complex, ideologically charged language of historical narratives.

DistilBERT, a smaller variant of the powerful transformer-based BERT model, consumes fewer computing resources due to its reduced number of parameters. While not specifically tailored to historical or political contexts, DistilBERT's efficiency, combined with its ability to capture subtle sentiment nuances, makes it an interesting tool for this analysis. Both VADER and DistilBERT are among the most popular open-source sentiment analysis models available, offering a balance between accessibility and performance for various NLP tasks.

The main goal of this study is to assess how well commonly used sentiment analysis tools—VADER, a lexicon-based model, and DistilBERT, a transformer-based deep learning model—perform when applied to texts that reflect the varied perspectives of the Vietnam War. By examining a dataset divided into pro-Vietnamese, pro-American, and civilian viewpoints, the study seeks to understand how well these models can capture the complex emotional and ideological layers present in such discussions. Additionally, this research explores whether these models, typically designed for modern contexts, can effectively interpret the intricate rhetorical strategies and historical nuances involved.

In this study, we will go through the methodology used to collect and preprocess datasets, present the results of the sentiment analysis, and conduct an in-depth review of the observations. The paper will conclude by addressing the limitations of the models and the study itself while proposing avenues for future research to enhance the analysis of complex historical and political texts.

Data Collection

The texts utilized in this study were sourced from a diverse range of materials, including academic journals, public records, and military reports, as well as selected interviews and speeches.

The collected texts were categorized into three groups: Pro-Vietnamese, Pro-American, and Civilian. The Pro-Vietnamese texts included reports and other writings that were mostly in favor of the North Vietnamese Pro-Communist Regime, glorifying their struggle against imperialist forces. The Pro-American texts predominantly comprised military reports and several personal accounts that portrayed the war as a brutal conflict fought against communism. The Civilian texts included records of those unfortunate citizens who became ensnared in the war.

The Pro-Vietnamese texts included books such as *People's War, People's Army* by Võ Nguyên Giáp (Giap #), *Our Great Spring Victory* by General Van Tien Dung (Dung #), *The Sorrow of War* by Bao Ninh (Ninh and Bảo #), and Ho Chi Minh's speech against the Americans on July 17, 1966 (*Ho Chi Minh's Appeal to the Vietnamese Nation to Fight Against the Americans* #). These texts were deliberately chosen to provide a comprehensive range of perspectives on the Vietnamese resistance to imperialism, reflecting various facets of the struggle. While works like *The Sorrow of War* are fictional, they offer a profound and insightful portrayal of the Vietnamese struggle during the Second Indochina War, providing a rich and poignant analogy of the emotional and psychological impacts of the conflict.

The Pro-American texts primarily included the National Security Decision Memoranda (National Security Decision Memoranda (NSDM)), the U.S. National Security Council's NSC 5405 from January 16, 1954 (United States National Security Council. NSC 5405: United States Objectives and Courses of Action with Respect to Southeast Asia), History of the Joint Chiefs of Staff (Parts 1 and 2), a speech by the 37th President of the United States, Richard Nixon, on November 3, 1969 (Nixon), and *Surviving Hell: A POW's Journey* by Leo Thorsness (Thorsness #). The National Security Decision Memoranda (NSDM), National Security Council (NSC) 5405, and History of the Joint Chiefs of Staff (Parts 1 and 2) were declassified by the U.S. government and subsequently released to the public. These texts, including official documents such as the NSC 5405 and various military reports, offer valuable insight into the U.S. perspective on the Vietnam War. They shed light on the political mindsets, morale, and rationale of the U.S. government during the conflict.

Finally, the civilian texts primarily comprised memoirs, personal records, and an interview, including *When Heaven and Earth Changed Places* by Le Ly Hayslip (Hayslip and Wurts #), *Last Night I Dreamt of Peace* by Doan Viet Lan (Đặng #), an interview with American veteran nurse Ms. Borg ("Borg, Ms. Vietnam War Veteran Nurse Interview"), and *Vietcong Memoir* by Truong Nhu Tang

(Giap #). These texts provide profound insights into the harsh realities of war and its impact on various groups of people. The interview with Ms. Borg offered a unique perspective on how ordinary American soldiers were affected by the war. In contrast, *Vietcong Memoir* presented the viewpoint of a high-ranking member of the Viet Cong. Texts such as *Last Night I Dreamt of Peace* focused on the experiences of civilians, illustrating the human toll of the conflict from both a personal and societal standpoint.

Data Preprocessing

The texts, categorized into Pro-Vietnamese, Pro-American, and Civilian groups, were in various formats, which included PDFs (a few of which were scanned using OCR), EPUB, and MP3 files. Each format was processed using Python by utilizing appropriate tools.

For text-based PDF files, PyPDF2 was used to extract the text; Tesseract OCR was used to convert scanned documents such as the Ho Chi Minh's Speech into text-based PDF. Tesseract OCR was particularly chosen due to its ease of use and its robust capabilities in recognizing text from scanned images, ensuring that image-based PDFs can be converted into text-based PDFs.

For EPUB files, ebooklib was used, allowing us to easily extract texts from eBooks in this format. This allowed us to easily clean and parse the HTML-based contents present within the EPUB files using BeautifulSoup.

The audio content was in MP3 format, which was converted into readable text format using Google Speech API, which is a well-regarded speech-text model. These text files were then analyzed using AI models to ensure that these words were transcribed accurately.

All non-text-based file formats were first converted into text files. The converted text files underwent basic cleaning, which included removing empty sentences, eliminating unnecessary spaces, and enhancing the overall readability of the extracted content.

All texts analyzed in this study were either originally in English or reviewed as officially translated English versions. No further translation tools were applied to maintain consistency in linguistic analysis.

Each sentence of the extracted text was then processed and analyzed by two models separately: DistilBERT and VADER. DistilBERT was trained on SST-2 (Stanford Sentiment Treebank 2), as SST-2 provides a reliable, well-curated set of examples to fine-tune models for sentiment recognition in text. This dataset is particularly useful for training sentiment models because it contains a balanced mix of positive and negative sentiment samples.

DistilBERT is a smaller and faster variant of BERT, which, when pre-trained on the SST-2 dataset, enables us to capture nuanced sentiment in texts. The model's ability to capture subtle sentiment variations, especially in text with historical and political context, makes it ideal for analyzing public opinion and discourse about such a significant event.

While VADER may not be traditionally ideal for nuanced sentiment analysis, it proves valuable in this study by offering meaningful insights through its simple, efficient approach. By iterating over individual lines of text, VADER provides deep insights into the topic. Furthermore, VADER's non-binary classification system (positive, negative, and neutral) adds depth to sentiment analysis, making it especially useful for capturing a range of sentiments. This multi-class approach allows for a more balanced understanding of the diverse perspectives surrounding the Vietnam War.

In this study, DistilBERT is constrained to display sentiment classifications based on a score threshold of 0.8, distinguishing between "positive," "negative," and "neutral" sentiment. If the model's confidence score for a given sentence is lower than this threshold, it is classified as "neutral," indicating that the sentiment is neither strongly positive nor negative. This helps capture more nuanced, ambiguous sentiments that may arise in complex historical and political contexts, such as the Vietnam War. Sentences with higher confidence are classified as either "positive" or "negative," depending on the label provided by DistilBERT. This approach allows for a more refined sentiment analysis, particularly useful for texts that involve subtle or polarized viewpoints. This allows us to

enhance the usability of this model for this context, as DistilBERT is bipolar (only classifies as positive or negative) by design.

VADER is used in this study specifically due to its ease of processing large amounts of data, especially the processing of informal language, which is very useful in this study as it includes analysis of personal records and other informal records to capture the essence of war. VADER classifies each of the sentences into “positive,” “negative,” and “neutral.”. It provides insightful results when utilized alongside other models such as DistilBERT.

Sentiment Analysis

To determine the overall sentiment of the collected sources—segregated into Pro-Vietnamese, Pro-American, and Civilian categories—we calculated the aggregate sentiment for each source using VADER and DistilBERT. The processed data were saved into comma-separated value (CSV) files, which facilitated sentiment plotting and comparison across categories.

General Sentiment Trends

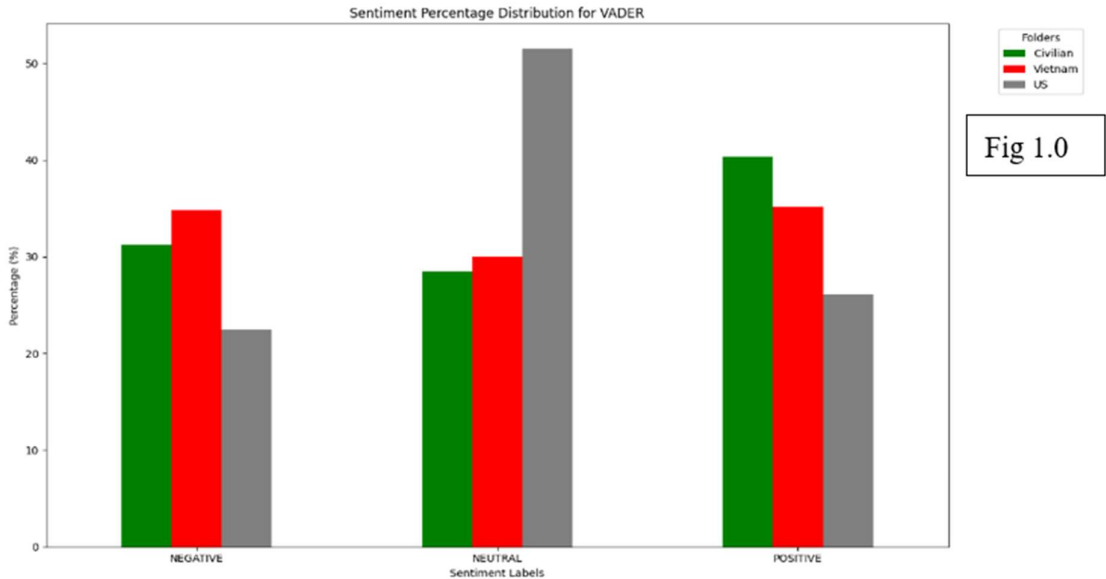


Fig 1.0

Fig. 1.0 illustrates the overall sentiment analysis using VADER. The U.S. sources exhibit a predominantly neutral sentiment, followed by Vietnam and Civilian sources. While the Civilian group shows a predominantly positive sentiment, Vietnam has a more balanced mix, and the U.S. data features the least emotional intensity. Negative sentiment is led by Vietnam, followed by Civilian sources and the U.S., reflecting the emotional tone of conflict narratives.

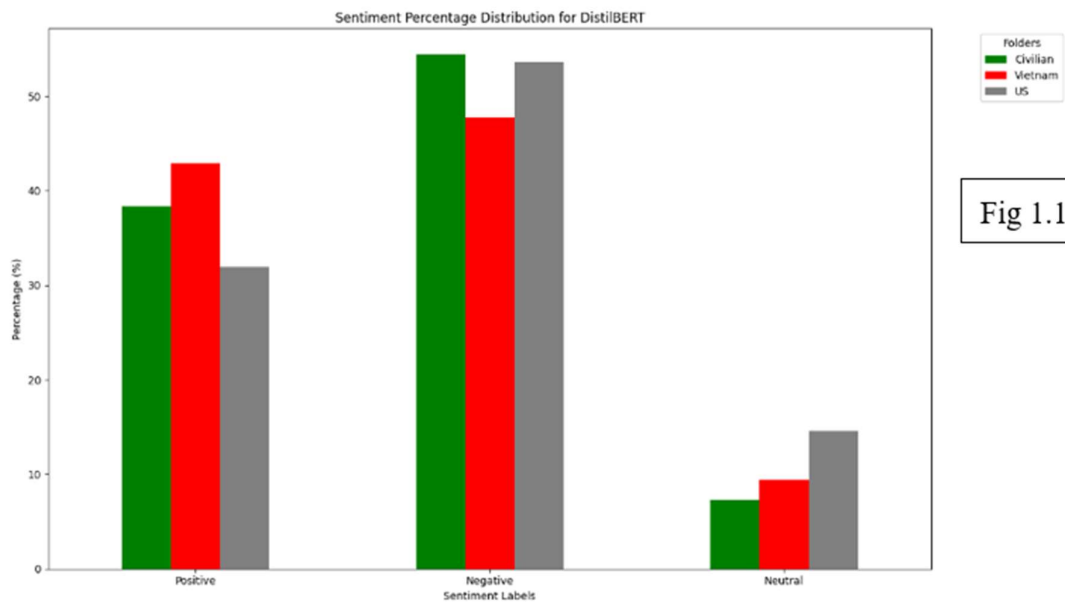


Fig. 1.1 uses DistilBERT fine-tuned on SST-2 and highlights a contrasting sentiment distribution. Here, Vietnam, Civilian, and U.S. categories are predominantly negative, with Civilians leading this group. Positive sentiment remains led by Vietnam, followed by Civilians and the U.S. However, the neutral sentiment in DistilBERT is markedly lower than in VADER due to the stricter classification threshold (confidence score < 0.8). These differences underline the models' distinct methodologies—VADER relies on lexicons and may flag subtler sentiments as neutral, whereas DistilBERT incorporates deeper contextual understanding.

Speech Analysis

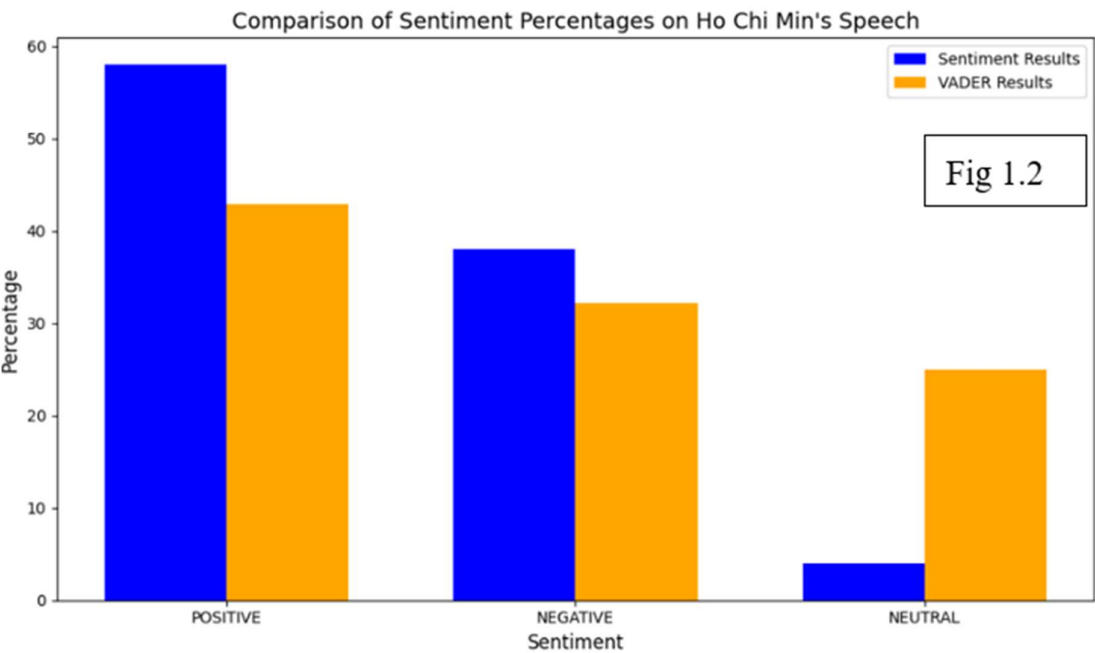


Fig. 1.2 presents the sentiment analysis of Ho Chi Minh's speech delivered on July 17, 1966, aimed at rallying the Vietnamese people. VADER assigns over 40% positive sentiment, while

DistilBERT flags over 55%. Both models converge on the speech’s uplifting and motivational tone, reflecting its intended call for unity and resistance. However, nuanced elements, such as rhetorical strategies evoking nationalistic fervor, may not be fully captured by either model.

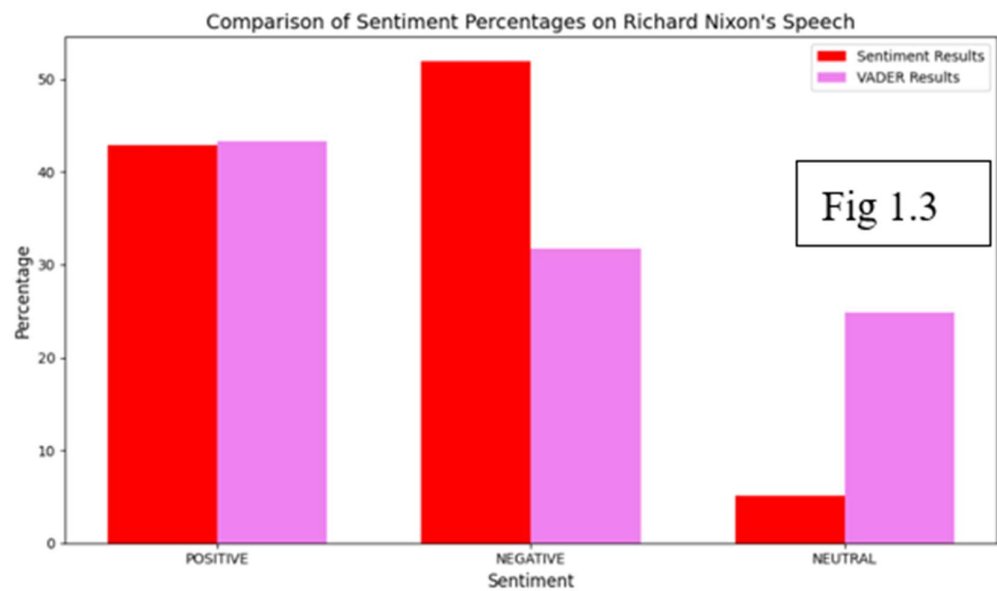


Figure 1. 3 shows the sentiment analysis of Richard Nixon's speech on November 3, 1969, addressing the Vietnam War. Sentiment diverges significantly between the models. VADER identifies a mix of positive and neutral sentiment, aligning with Nixon's reassuring and optimistic appeals. In contrast, DistilBERT assigns a predominantly negative sentiment, likely reflecting the speech's acknowledgment of war challenges and public discontent. Despite DistilBERT's negative classification, the speech contains notable positive undertones, suggesting the need for hybrid approaches to interpret mixed narratives effectively.

Overall Outlook (Analysis of Figs. 1.0 and 1.1):

In this section, we evaluate how traditional sentiment analysis tools, such as VADER and DistilBERT, perform when applied to a complex dataset, such as the Second Indochina War, which presents a variety of nuanced emotional and political outlooks. By doing so, we aim to assess whether these tools can effectively capture the multi-layered perspectives of the different groups involved in this historically significant conflict.

Fig. 1.0 emphasizes resilience and hope amidst adversity; Fig. 1.1, in contrast, shifts the focus to highlight the hardships and negative undertones of the brutal and destructive war.

In Fig. 1.0, the overall outlook of the entire dataset comprising Civilian, Vietnam, and US is stated to be positive. The U.S outlook is predominantly neutral, which can be attributed to the inclusion of military reports. This report is likely to be more formal and neutral, contributing to the overall neutral tone of the dataset. The American government considered the fight to be a fight for democracy and a fight for the eradication of communism and USSR influence on not only Vietnam, but the entirety of southeast asia.

Moving to the Vietnam dataset in Fig. 1.0, we observe a mixed outlook, with sentiments being nearly equally negative and positive. This balance likely stems from an inherent optimism in their narrative, even as it reflects the hardships and challenges they faced. Their persistent struggle against imperial powers and their ambition to create a unified nation contribute to this outlook. They considered the war a fight for their freedom and unity.

Finally, examining the civilian dataset in Fig. 1.0, it is evident that civilians, even amidst the brutal war, remain predominantly positive. Despite a significant negative sentiment, which likely reflects the hardships and struggles brought on by the war, their optimism shines through as they

look forward to a brighter future—just like the overall outlook portrayed in the book *When Heaven and Earth Changed Places*.

Fig. 1.0 emphasizes the presence of resilience and hope even under adversity, while Fig. 1.1 highlights the hardships and negative undertones of the brutal and destructive war. However, when we take a look at Fig. 1.1, where the sentiment analysis is based on DistilBERT, the overall outlook changes. The Civilian outlook shifts from predominantly positive to predominantly negative; the same goes for the U.S. and Vietnam. All of the datasets exhibit an overall negative outlook.

These results differ significantly from Fig. 1.0, and they can be attributed to the way VADER and DistilBERT classify text as positive, neutral, and negative. VADER, a lexicon-based model that relies on predefined word lists and their associated sentiment scores, makes it more sensitive to context and linguistic changes. In contrast, DistilBERT, fine-tuned on SST-2, uses a deep learning-based approach that considers individual words as well as sentence-level semantics. It also has a stricter threshold for classifying texts as neutral (scores less than 0.8), which contributes to these differences.

Outlook of Political Speeches (Fig. 1.2 and Fig. 1.3):

The speeches by Ho Chi Minh and Richard Nixon provide an opportunity to test whether traditional sentiment analysis tools can navigate through complex political texts and offer valuable insights. By analyzing the sentiment of these speeches, we can assess the effectiveness of VADER and DistilBERT in capturing the ideological and emotional nuances of politically charged narratives.

In Fig. 1.2, the sentiment analysis of Ho Chi Minh's speech 'Call of President Ho' on July 17, 1966, reveals a predominantly positive sentiment. Both VADER and DistilBERT identified this positive tone, likely reflecting the speech's powerful call for national unity and resistance. This resonates with the morale and determination of the North Vietnamese, emphasizing their fight for freedom and sovereignty. The positive sentiment in this case aligns with the emotional intensity of the rhetoric, which resonates with the Vietnamese people's aspirations for independence.

In contrast, Nixon's speech on November 3, 1969, as seen in Fig. 1.3, shows predominantly mixed sentiments, with DistilBERT pointing to a negative outlook, while VADER suggests a more positive tone. DistilBERT's negative classification seems to reflect the somber and harsh realities of the Vietnam War, including the public's growing resentment and war fatigue. VADER, on the other hand, captures the reassurance Nixon attempts to provide, resulting in a more optimistic interpretation of the speech. These contrasting results emphasize the differing approaches of the models, with VADER focusing on individual word sentiment, while DistilBERT incorporates sentence-level semantics and context.

These findings, however, also highlight a limitation of traditional sentiment analysis models. While they provide useful insights into the general sentiment of these speeches, classifying them as positive, negative, or neutral oversimplifies the complexity of the rhetoric. Sentiment models, especially when used in isolation, fail to capture the full depth of emotional and ideological nuances inherent in such politically charged discourse. Nonetheless, these tools can still offer a high-level overview of the emotional landscape and serve as a starting point for understanding the general tone of the speeches, especially in contexts where detailed, nuanced analysis is not required. Given the nature of the Vietnam War and its widely recognized historical significance, sentiment analysis models offer a useful tool for gaining initial insights, but they fall short when it comes to capturing the deeper layers of meaning.

Limitations

While this study offers valuable insights into how well sentiment analysis tools work in understanding complex historical and political situations, there are a few limitations to consider.

First, the datasets used—comprising speeches, interviews, and various texts—were categorized to help clarify and represent the different perspectives on the conflict. However, during the data selection process, it's possible that some key viewpoints were left out, which could introduce bias into the results. The study could benefit from a broader dataset that includes a wider range of voices, which would make the analysis more comprehensive.

Additionally, the sentiment analysis models used—VADER and DistilBERT—are effective for identifying general sentiment but have limitations when it comes to grasping the complexity of historical and ideological contexts. These models rely on predefined word lists and training data, which can lead to biases that overlook the subtle nuances in wartime language. The choice of these models was due to the lack of powerful computing resources. Had more advanced models like GPT or BERT been available, more accurate results could have been achieved. However, the verdict remains the same: these tools cannot be used on their own. Instead, a more nuanced approach, perhaps by combining them with human analysis or utilizing specialized models, would offer a richer and more accurate understanding of the underlying sentiment.

Conclusion

To summarize, this study examines the effectiveness of sentiment analysis tools, particularly VADER and DistilBERT trained on the SST-2 dataset, in analyzing complex historical and political texts related to the Vietnam War. Our findings reflect that while both models are capable of providing a high-level overview of the subject, they fail to capture the full depth of emotional and ideological nuance present in the text.

VADER, due to its lexicon-based approach, is effective for high-level sentiment analysis but may oversimplify texts, especially in texts that have an underlying emotional undertone. On the other hand, DistilBERT uses a deep learning method, allowing it to capture a more nuanced view of the text; however, it still fails to understand the intricate historical and ideological contexts.

While classifying these sentiments as positive, negative, or neutral may prove helpful while conducting a high-level analysis. It fails to capture the true essence of the situation

Future research could address these limitations by exploring hybrid approaches that combine sentiment analysis tools with human input. This would allow for a more comprehensive and nuanced analysis, helping to capture the diverse perspectives and deeper meanings in historical events like the Vietnam War. Expanding the dataset and incorporating additional models could also enhance the depth and accuracy of sentiment analysis in politically and emotionally charged contexts.

Ultimately, the complexity of historical events, especially in topics as multifaceted as war, cannot be captured by sentiment categories alone. War holds a deeply personal meaning, shaped by the unique roles, experiences, and perspectives of those involved—civilians enduring profound loss and hardship, soldiers grappling with duty and sacrifice, and leaders navigating strategic decisions influenced by ideological and national interests. This highlights the necessity of combining human interpretation with automated tools to achieve a more comprehensive understanding of such emotionally and ideologically charged narratives.

References

1. Đặng, Thùy Trâm. *Last night I dreamed of peace: the diary of Dang Thuy Tram*. Harmony Books, 2007.
2. Dung, Van Tien. *Our Great Spring Victory*. Monthly Review Press, 1977. Accessed 2 December 2024.
3. Giap, Vo Nguyen. *People's War People's Army: The Viet Cong Insurrection Manual for Underdeveloped Countries*. University Press of the Pacific, 2001. Accessed 2 December 2024.
4. Giap, Vo Nguyen. *People's War People's Army: The Viet Cong Insurrection Manual for Underdeveloped Countries*. University Press of the Pacific, 2001. Accessed 2 December 2024.
5. Hayslip, Le Ly, and Jay Wurts. *When Heaven and Earth Changed Places: A Vietnamese Woman's Journey from War to Peace*. Knopf Doubleday Publishing Group, 2017. Accessed 2 December 2024.
6. *Ho Chi Minh's Appeal to the Vietnamese Nation to Fight Against the Americans*. Digital Archive, Wilson Center, <https://digitalarchive.wilsoncenter.org/document/ho-chi-minhs-appeal-vietnamese-nation-fight-against-americans>. Accessed 11 2024.

7. *National Security Decision Memoranda (NSDM)*. Nixon Presidential Library, <https://www.nixonlibrary.gov/national-security-decision-memoranda-nsdm>.
8. Ninh, Bao, and Bảo Ninh. *The sorrow of war: a novel*. Edited by Frank Palmos, translated by Frank Palmos, Minerva, 1994. Accessed 2 December 2024.
9. Nixon, Richard. *Address to the Nation on the War in Vietnam*. 3 11 1969, <https://www.presidency.ucsb.edu/documents/address-the-nation-the-war-vietnam>. Accessed 11 2024.
10. Thorsness, Leo. *Surviving Hell: A POW's Journey*. Encounter Books, 2011. Accessed 2 December 2024.
11. *United States National Security Council*. NSC 5405: *United States Objectives and Courses of Action with Respect to Southeast Asia*. https://www.vietnamwar50th.com/assets/1/7/US_National_Security_Council,_NSC_5405,_United_States_Objectives_and_Courses_of_Action_with_respect_to_Southeast_Asia_16_Jan_1954.pdf. Accessed 11 2024.
12. "Vietnam War Veteran Nurse Interview." *Library of Congress, Veterans History Project*, <https://www.loc.gov/collections/veterans-history-project-collection/serving-our-voices/vietnam-war/vietnam-war-looking-back/item/afc2001001.46805/>. Accessed 11 2024.