Investigating bias in different language editions of Wikipedia

by

Archit Aggarwal

Student ID: 1351097

A thesis submitted in total fulfilment for the
degree of Master of Computer Science

in the
Department of Computing and Information System
Melbourne School of Engineering
THE UNIVERSITY OF MELBOURNE

Supervisors:
Dr. Christine de Kock
Dr. Jey Han Lau

June 2024

Declaration of Authorship

I, **Archit Aggarwal**, declare that this thesis titled, 'Investigating bias in different language editions of Wikipedia' and the work presented in it are my own.
I confirm that:

- This thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and to the best of my knowledge and belief, it doesx not contain any material previously published or written by another person where due reference is not made in the text.;
- the thesis is 20,071 in length, excluding text in images table, bibliographies and appendices.

Signed: ____________________________

Date: 2 June 2024

Abstract

Digital information can be inherently biased due to the varying perspectives, cultural contexts, and editorial practices of its contributors. Biases can emerge even on popular platforms like Wikipedia, which is supposed to be neutral. This thesis investigates biases in different language editions of Wikipedia, focusing on English, Hindi, Afrikaans, and Chinese.

Sentences from these language editions were aligned through content alignment techniques, followed by the application of a pre-trained sentiment analysis model to detect divergences in sentiments. Various metrics such as Jensen-Shannon divergence were used to quantify these divergences. Further to this, Named Entity Recognition (NER) was used to identify common entities within the texts to aid with the stance detection process. Stance detection was then conducted through the use of Large Language Models (LLMs). These LLMs were used to assess the stance of authors towards the common entities. These divergences of stances were quantified by the Stance Divergent Score (SDS) across texts. A round-trip translation baseline was prepared to ensure that divergences were due to the text rather than translation artifacts. Additionally, human validation was applied in content alignment and stance detection to assess the efficacy of our approach.

The findings reveal the presence of biases in Wikipedia articles on controversial issues and war-related topics across different language editions. It is worth noting that while sentiment analysis and stance detection are closely related, the observed sentiment divergence does not necessarily correlate with stance divergence. For instance, the English-Hindi language pair exhibited lower sentiment divergence but higher stance divergence. Furthermore, language pairs involving Chinese frequently demonstrated higher divergences in both sentiment and stance. This research demonstrates the nuanced nature of biases in multilingual Wikipedia articles.

Acknowledgements

I would first like to express my sincere gratitude to my supervisors, Dr. Christine de Kock and Dr. Jey Han Lau, for their invaluable patience, thoughtful guidance, motivation, continuous encouragement, and insightful feedback throughout the whole process of research and thesis writing. I am extremely fortunate to have them as my supervisors.

I would like to express my gratitude to my friends, who have become like an extended family to me in Melbourne. Their motivation and encouragement were instrumental in helping me persist with my research, especially during challenging times. I am particularly thankful to Ramah Mandar Gokhale and Shub Mayur Raichand Mardia for tirelessly reviewing my writing and this thesis multiple times, and offering their invaluable suggestions.

Finally, I would like to express my gratitude to the University for providing me with this opportunity and for the incredible supervisors who guided me throughout this research journey, which initially seemed impossible to complete.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Our written records of the world have always been influenced by the powerful, leading to the commonly held belief that *"history is written by the victors."* However, in our modern era of decentralisation of information control (Benkler, 2006), this adage is being challenged. Wikipedia, a free encyclopedia, leads this paradigm shift with over 6.8 million articles in English alone as of 2024 (Wikipedia, 2024a). Wikipedia has become a popular first resort for information seekers across the world (Giles, 2005) supported by principles such as collective editing and unrestricted access. However, concerns have been raised by scholars over its credibility and bias. This decentralisation of information and history raises fundamental questions about who has the authority to define the past, the credibility of open-source platforms, and the biases that may exist in seemingly neutral spaces (Rosenzweig, 2006). This decentralisation signifies a diversion from conventional historical scholarship, allowing different perspectives to emerge.

Wikipedia's Neutral Point of View (NPOV) policy stands as a cornerstone in its mission to provide unbiased information. This policy highlights the importance of conveying even-handed information, free from individual opinions, beliefs, or prejudices (Wikipedia, 2024b). However, the application of this policy differs in practicality. Scholars such as Reagle (2010) and Shaw and Hargittai (2018) have scrutinised the platform, uncovering that even with such rigorous frameworks, subtle biases are present in the content. A stark manifestation of this challenge is observed in the domain of gender bias. Research has pointed out a notable under-representation of women, both in the content itself and among those contributing to it (Cabrera et al., 2018; Sun and Peng, 2021; Graells-Garrido et al., 2015). This gender disparity can lead to considerable consequences concerning the representation of knowledge or information regarding women. This may reinforce stereotypes and exclude significant female figures and perspectives.

Different language editions of Wikipedia are developed by different communities. This contributes to the development of both cultural and geographical biases, with the content often leaning towards Western-centric views. In a study by Callahan and Herring (2011), a comparison was made between articles regarding famous individuals from Poland and the USA in the Polish and English versions of Wikipedia. Through both quantitative and qualitative content analysis, consistent disparities reflecting the distinct cultures, histories, and values of Poland and the United States were found. The study revealed considerable systematic bias favouring the English articles of famous people from the USA.

Hecht and Gergle (2010) examined Wikipedia articles across 25 languages to analyse the information provided through different language translations of the same articles. This study indicated considerable differences in content. To achieve this, the study implemented a tailored algorithm, CONCEPTUALIGN. This algorithm involved translating the content of the article and pinpointing analogous sections across multiple language editions. A quantitative investigation was carried out to assess the similarities and discrepancies in content across the languages. Statistical techniques were utilised to find differences in aspects such as article length, subject matter coverage, and citations, among other aspects. Another work carried out at TU Wein (Rajcic, 2017) undertook a comparative investigation into articles about famous individuals on Wikipedia, contrasting the content across multiple languages. This study also used other quantitative measures such as the languages in which the article is accessible and the number of views the article has received.

Although these studies shed light on the differences between various language editions, they concentrate predominantly on non-linguistic signals and do not integrate the application of natural language processing (NLP) models or techniques for identifying bias. This limitation may overlook aspects related to content bias or subtle variations in meaning across languages. It does not involve limitations associated with the quantitative analysis of the articles.

1.1 Objectives

The project will investigate divergences in the representations of current events across the different language editions of Wikipedia. It is important to note that different language editions of Wikipedia are not translations; rather, they are often generated independently by different communities of contributors. While Wikipedia has a Neutral Point of View (NPOV) policy intended to guide contributors towards objectivity, lapses are inevitable. The objective of this project is to identify explicit preferences in content as well as subtler differences that may influence how information is presented and perceived across different language editions.

A multitude of biases are prevalent in text corpora. Therefore, in the context of this project, whenever the term 'biases' is, it exclusively refers to these sentiment and stance divergences and does not encompass other forms of bias like gender bias or recency bias.

1.1.1 Methods used to identify biases across different language editions of Wikipedia

Research Question 1: What methods can be used to identify biases across different language editions of Wikipedia?

The majority of previous research concerning biased writing has explored different domains such as news articles (Baly et al., 2018) and social media (Mohammad et al., 2017). Baly et al. (2018) delved into the examination of news media articles to identify nuanced textual biases. Specifically, their research used techniques like stance detection (Küçük and Can, 2020) and sentiment analysis (Shen et al., 2018) to dissect these articles. A previous study focused on identifying biased writing across various editing versions on Wikipedia (Recasens et al., 2013). However, this particular study did not address the biases that might be inherently partisan in different language editions of Wikipedia, leaving a gap in the understanding of how language-specific references might influence the presence and expression of bias. In this study, we will specifically assess the sentiment and stances in these articles. Sentiment analysis refers to the process of determining the emotional tone behind a body of text, which promotes understanding the attitudes, opinions, and emotions expressed within it. Stance detection, on the other hand, involves identifying the author's position or attitude towards a particular topic, which can be neutral, in favour, or against it. By focusing on these two aspects, our research aims to provide an understanding of how biases are manifested and perceived in different language editions of Wikipedia.

1.1.2 Measuring biases in a systematic way

Research Question 2: Is it possible to quantify or measure these biases systematically?

To systematically quantify biases in Wikipedia's multilingual content, this study leverages the Jensen-Shannon Divergence (JSD) (Lin, 1991) and other metrics, complemented by human evaluation. Human evaluation acts as a crucial validation method, ensuring that the computational findings align with human judgments. This integrated approach not only quantifies biases effectively but also aids in developing strategies to mitigate them. This allows contributors to work towards a more neutral and inclusive representation of knowledge on Wikipedia.

1.2 Thesis Overview

1.2.1 Chapter 2

In Chapter 2, we provide a review of the literature on key areas relevant to our study: prior works on Wikipedia, language models, sentiment analysis, content alignment, and stance detection. We discuss the contributions and limitations of monolingual and multilingual studies on Wikipedia, focusing on the varying nature of biases across different language editions. In sentiment analysis, we discuss the evolution of methodologies from traditional techniques to advanced neural networks, highlighting their applicability in bias detection. The examination of content alignment techniques accentuates the importance of semantic matching in cross-lingual contexts. Finally, our review of stance detection methodologies discusses the challenges and advancements in identifying authorial positions on contentious issues. This review not only maps out the current landscape but also identifies the critical gaps and challenges that our research aims to address, setting a solid foundation for the subsequent chapters.

1.2.2 Chapter 3

In Chapter 3, we discuss the methodology for obtaining articles from Wikipedia. Along with this, the rationale for selecting topics related to controversial issues and war is also explained. We also provide an account of the number of articles obtained for each language and describe the tools utilized in the data collection process.

1.2.3 Chapter 4

In Chapter 4, we discuss the methodological framework and findings of content alignment within our study, which involves aligning sentences from different Wikipedia language editions. This is an important step for subsequent sentiment analysis.

The chapter also outlines the transition from mBERT to LASER, detailing the rationale behind our methodological pivot and presenting the results obtained through LASER embeddings. Additionally, we describe the use of human validation to determine the cosine similarity threshold for aligning sentences, ensuring accuracy through human evaluation.

1.2.4 Chapter 5

In Chapter 5, we discuss the methodology of sentiment analysis through a pretrained sentiment analysis model. We also define various metrics, such as Mean Jensen-Shannon Divergence (JSD), to quantify biases in Wikipedia language editions. The chapter concludes by presenting the results of our sentiment analysis and discussing the limitations found in our approach.

1.2.5 Chapter 6

In Chapter 6, we first define stance detection and distinguish it from sentiment analysis. We then explain the necessity of identifying targets for stance detection using Named Entity Recognition (NER). Following this, we use a large language model to determine stances towards identified entities. The chapter concludes with a presentation and discussion of the stance detection results.

1.2.6 Chapter 7

In Chapter 7, we conclude our work and outline possible directions for future work.

Chapter 2

Literature Review

This literature review explores the issue of bias within Wikipedia. This situates our research within the broader discourse on information accuracy and neutrality in digital knowledge repositories. We begin by examining prior works that have addressed Wikipedia bias, focusing on gender and societal inequalities and the influence of cultural contexts across different language editions. Following this, we explore sentiment analysis as a tool for detecting biases in textual content, highlighting its applications and challenges, particularly for low-resource languages. The review then discusses content alignment techniques, essential for comparing multilingual Wikipedia articles, and evaluates methods such as cross-lingual embeddings for their effectiveness. Lastly, we examine stance detection, its role in identifying viewpoints in text, and the contributions of advanced NLP techniques, including Named Entity Recognition (NER) and Large Language Models (LLMs), in improving stance detection accuracy. Through this analysis, we aim to provide a robust foundation for our investigation into biases in different language editions of Wikipedia.

2.1 Prior Works on Wikipedia Bias

This section examines work that has previously addressed bias in Wikipedia by reviewing these works.

An article by Koerner (2019) discusses the inherent biases within Wikipedia. It further examines how these biases manifest across the platform's policies, practices, content, and user participation. The article highlights that bias not only exists but also contributes to significant barriers to inclusion, leading to imbalanced participation and distorted portrayal of prominent events and personalities. These barriers are evident in issues such as contributor retention, challenges faced by emerging communities, and the systematic exclusion of certain content.

Previous studies on Wikipedia can be categorized into two main areas: monolingual studies, which concentrate on research within a single language edition, and multilingual studies, which focus on research across multiple language editions.

2.1.1 Monolingual Works

Martin (2018) examines how bias might be imposed while appearing to comply with Wikipedia regulations. For example, biased entries may include citations from credible sources, but these citations may be an unrepresentative sample, or they may deliberately exclude contrary ideas as original research while discreetly providing disputed judgments. Such posts create the illusion of impartiality while incorporating a non-neutral viewpoint. Their research, which includes examples from changing the author's own Wikipedia page, proposes that similar biasing tactics may be used across many domains inside Wikipedia. It follows that practically all entries on difficult topics are likely to contain some amount of personal prejudice, with neutrality reached only when the various biases accidentally nullify one other.

Greenstein and Zhu (2012) study the evolution of bias in Wikipedia articles over a decade, particularly focusing on U.S. political content. Through this study, they discovered a transformation in the political orientation of articles: initially exhibiting a Democratic slant. They also found that Wikipedia articles have progressively moved towards a more balanced representation, approaching the ideal of NPOV. This shift towards neutrality does not predominantly result from revisions of existing articles but rather from the addition of new articles that often present viewpoints contrasting those of earlier pieces. Despite the NPOV guidelines, the challenge of maintaining quality is exacerbated by Wikipedia's reliance on voluntary contributions, resulting in over 40,000 articles currently flagged for NPOV or related quality concerns. Hube and Fetahu (2018) focused on language bias within Wikipedia articles at the sentence level. The researchers proposed a supervised classification approach to identify language bias, utilising an automatically generated lexicon of bias-indicative words leading to syntactical and semantic analysis of statements.

Collier and Bear (2012) surveyed and found a significant gender contribution gap on Wikipedia, where less than 15% of contributors are women. A panel comprising both researchers and practitioners has explored various reasons behind the persistent gender gap in contributions. The gender research literature points to three potential factors contributing to this phenomenon: high levels of conflict during discussions, aversion to critical environments, and a lack of confidence in editing the work of others. Cabrera et al. (2018) further found that there is pervasive gender bias within Wikipedia talk page interactions, which are less frequently examined than article contributions or topic coverage. By analysing a dataset enriched with gender information, it was found that female participation on talk pages is significantly lower than male participation, and the nature of their involvement often aligns with traditional gender stereotypes. Additionally, the likelihood of receiving replies on talk pages was observed to be lower for posts authored by females, with variations depending on the topic. These findings confirm the existence of gender bias in talk page activities. The gender imbalances in contribution circles not only shape the discussions and decision-making processes but also influence the range and depth of content available on Wikipedia.

Graells-Garrido et al. (2015) focused on how men and women are portrayed in biographies. The findings reveal differences in how genders are characterised, with some disparities reflecting societal biases documented by Wikipedia. Their study utilises *Pointwise Mutual Information (PMI)* to identify words strongly associated with each gender based on their occurrence in biographies are displayed as word clouds in Figure 1. The analysis shows a clear gendered division in language usage: biographies of women frequently include terms related to the arts and familial roles such as *"actress"*, *"her husband"*, and *"feminist"*. In contrast, biographies of men are more likely to contain words associated with sports, such as *"footballer"* and *"league"*. These linguistic patterns not only mirror but possibly perpetuate traditional gender roles and stereotypes.

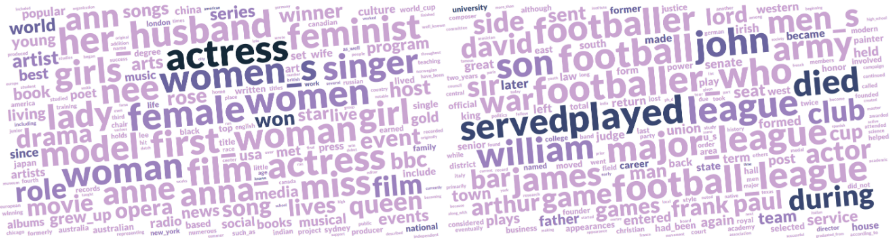


Figure 1: Words most associated with women (left) and men (right), estimated with Pointwise Mutual Information. Font size is inversely proportional to PMI rank. Colour encodes frequency (the darker, the more frequent) (Figure reproduced from Graells-Garrido et al. (2015))

Sun and Peng (2021) utilised a corpus of 7,854 text fragments from the biographies of 10,412 celebrities, enriched with demographic information about career and personal life events. The study used an event detection model and calibrated results using strategically generated templates to identify events with asymmetric associations between genders. The findings reveal a significant bias in the representation of events on Wikipedia: personal life events are more frequently integrated with professional events in the biographies of females than males as shown in Table 1.

Table 1: The marriage events are under different sections on Wikipedia. Yellow background highlights events in the passage,(Table reproduced from (Sun and Peng, 2021))

Name	Wikipedia Description
Loretta Young (F)	**Career**: In 1930, when she was 17, she eloped with 26-year-old actor Grant Withers; they were married in Yuma, Arizona. The marriage was annulled the next year, just as their second movie together (ironically entitled Too Young to Marry) was hlreleased.
Grant Withers (M)	**Personal Life**: In 1930, at 26, he eloped to Yuma, Arizona with 17-year-old actress Loretta Young. The marriage ended in annulment in 1931 just as their second movie together, titled Too Young to Marry, was released.

Wagner et al. (2016) also found linguistic bias where abstract language is used differently across genders, often highlighting positive attributes in men's biographies and negative ones in women's. They also found structural differences in the use of metadata and hyperlinks that affect how information is interconnected and accessed, which could influence the visibility and engagement with content related to different genders.

Young et al. (2016) investigated gender bias in the representation of organizational leaders on Wikipedia. Utilising the bias framework developed by Miranda et al. (2016), the study examines biases stemming from structural limitations and content restrictions inherent in Wikipedia. Structural constraints refer to the inherent design and governance of Wikipedia, which can influence how content is created and edited, including policies, editorial oversight, and community guidelines. Content restrictions include limitations on the types of sources considered reliable and the notability criteria for including information. Through a comparative analysis of Wikipedia profiles of Fortune 1000 CEOs, the study identifies selection, source, and influence biases. Interestingly, these biases appear to favour women and disadvantage men, suggesting that Wikipedia's structural constraints can produce unexpected forms of bias where the typically underrepresented group (women) might receive more positive portrayals than the majority group (men). This could be seen as a counterbalance to the glass ceiling effect, with women facing social barriers that prevent them from being promoted to top jobs in management. This indicates why women in leadership positions might receive more favourable attention due to their minority status. This finding contrasts with other research reviewed, which generally indicates that women are more likely to be perceived with a negative point of view.

Field et al. (2022) study the nature of social biases on Wikipedia, focusing on how different demographic attributes beyond gender, such as race and non-binary gender identities, influence content. They constructed a target corpus of biographies about demographic groups (e.g., women, and racial minorities) and a comparison corpus that matches the target corpus in as many attributes as possible, excluding the attribute of interest. They analyse various dimensions such as article length and recency to identify disparities within the target and comparison corpora. The findings from this analysis revealed significant disparities: for example, articles about Asian Americans and cisgender women are shorter than those of their counterparts. This suggests potential gaps in content that could stem from a lack of thorough editorial efforts. The study also indicated that articles about gender and racial minorities are generally newer, which may be a consequence of recent initiatives to correct recognised biases.

Coverage bias occurs when certain topics or perspectives are underrepresented or over-represented within a content source. In the context of Wikipedia, this type of bias affects the diversity of the information presented (Hinnosaar, 2019). Brown (2011) investigates the accuracy and thoroughness of Wikipedia's political coverage by reviewing thousands of articles about candidates, elections, and officeholders. It finds that while Wikipedia is generally accurate when relevant articles exist, it suffers from significant errors of omission, particularly on older or less prominent topics.

Halavais and Lackaff (2008) examined the breadth of Wikipedia's coverage by comparing it to the distribution of topics in published books and field-specific academic encyclopedias. It highlights that Wikipedia's coverage is influenced by the interests of its contributors, resulting in some subjects being better covered than others. Their study finds that while Wikipedia generally performs well due to its size, it has notable gaps in areas like law and medicine, which are traditionally covered by licensed experts.

This section reviewed studies focused on the English edition of Wikipedia, highlighting issues of gender, coverage bias and societal inequality. The following section will explore previous research that has examined bias across various language editions of Wikipedia.

2.1.2 Multilingual Works

Aleksandrova et al. (2019) examined the revision history of Wikipedia articles in Bulgarian, English and French. This included revisions of sentences that were flagged for bias and subsequently corrected. The researchers identified biased sentences by comparing the last tagged and first untagged revisions. Their approach provided ample data even from smaller Wikipedias, such as Bulgarian, where 62,000 articles yielded 5,000 biased sentences. The method's effectiveness was evaluated through manual annotation of 1,784 sentences across the three languages, revealing that only about half were truly biased. Miquel-Ribé and Laniado (2018) explore the content imbalances in Wikipedia across 40 different language editions, focusing on how cultural contexts influence the scope and nature of the content. They utilised strategies that consider geo-located articles, specific keywords, categories, and inter-article links to identify articles that relate to the cultural context of each Wikipedia language edition and named them Cultural Context Content (CCC). The findings reveal that a significant portion of content within each language edition predominantly reflects its cultural context, with much of this content created in the early years of the project. English only covers about 33.71% of the CCC articles from other languages while the representation of CCC from dominant language editions like English and German in other language editions is minimal, with less than 5% of their CCC articles being featured in other editions. This disparity shows gaps in the intercultural exchange and representation within Wikipedia's global framework.

Kerkhof and Münster (2019) focused on Wikipedia's portrayal of German and French Members of Parliament (MPs) in English, German and French Wikipedia. Their findings indicate a slight to moderate coverage bias against centre-left MPs in both Germany and France. This bias is supported by an analysis of authorship patterns and discussions on Wikipedia's talk pages, particularly for the German MPs. In a comparative study of Wikipedia entries for notable individuals from Poland and the USA, a marked difference was observed between the Polish and English versions. English language entries tend to have more references and external links, exhibit a more positive tone, and provide a broader diversity of information, including mentions of controversy. They are also generally longer than their Polish counterparts, which more frequently focus on professional accomplishments and personal life details (Callahan and Herring, 2011). Prior studies indicate that multilingual editors' cultural and linguistic backgrounds significantly shape their perspectives, influencing their interpretation and documentation of global events (Fichman and Hara, 2014; Kolbitsch and Maurer, 2006).

Wagner et al. (2015) assess bias across on Wikipedia, evaluating coverage and visibility within six language editions. Their findings reveal an unexpected pattern: contrary to common perceptions of male bias, there is a slight over-representation of women, with no significant differences in the proportional coverage between men and women. This suggests that both genders are equally represented in terms of article quantity and visibility, including the selection of featured articles on the English Wikipedia's start page. However, the study also uncovers gender differences in content presentation. Structurally, women are more frequently linked to men than men are to women, indicating a potential bias in how relational networks are constructed within Wikipedia. Lexically, there is a notable disparity in how genders are described: articles about women disproportionately focus on romantic relationships and family matters compared to those about men. These findings highlight the presence of both structural and lexical gender biases on Wikipedia, suggesting areas where the Wikipedia community could improve to achieve a more balanced portrayal of men and women.

Alvarez et al. (2020) used a qualitative exploratory approach to examine in-group bias in English and Spanish Wikipedia articles about international conflicts, focusing on two engagement resources from *Appraisal Analysis*: attribution and counter. In-group bias refers to the tendency to favour one's own group over others. Attribution refers to how information is sourced or credited, while counter involves presenting one idea and contrasting it with another. Their analysis involved identifying these resources in 14 articles and examining their use to reveal bias. The study found differences in how Spanish and English versions referred to sources and structured content. Spanish texts often provided detailed references to individual leaders and their arguments, highlighting disputes, while English versions grouped countries into broad categories, minimising disagreements. For instance, in the Free Trade Area of the Americas (FTAA) articles, Spanish texts emphasised resistance from Spanish-speaking countries, whereas English texts downplayed the controversy. Similarly, counters were used differently to alter narratives subtly. Spanish versions highlighted comprehensive embargoes and justified actions by Spanish-speaking countries, while English versions minimised these aspects. They concluded that attribution and counter-resources serve as linguistic indicators of in-group bias. This subtly shapes narratives in favour of the preferred in-group. This exploratory qualitative research is based on naturalistic observations of 14 selected articles in Wikipedia and does not use any advanced NLP methods.

2.1.3 Identifying Research Gap

In Sections 2.1.1 and 2.1.2, we reviewed existing studies that explore biases in Wikipedia within both monolingual and multilingual contexts. While previous research has highlighted the presence of biases, it primarily focuses on language-agnostic metrics such as article lengths and hyperlink structures, without focusing on the content itself or assessing the presence of ideological bias in the text. Furthermore, these studies often neglect to utilise NLP techniques for a more comprehensive bias analysis. This oversight restricts the potential for both quantitative and qualitative linguistic analyses, thereby potentially missing subtle content biases or variations in meaning across different languages.

Moreover, prior research has predominantly concentrated on gender bias within biographical articles. In contrast, this study expands the scope by examining controversial and war-related topics, aiming to uncover ideological biases and differing perspectives within these more contentious areas. This approach addresses a significant gap in the literature by providing insights into how ideological and stance biases manifest in discussions about sensitive topics on Wikipedia.

2.2 Language Models

This section will briefly explain the language models that are widely used in the field of NLP and will be mentioned in this literature review. This section will encompass the evolution of language models, starting with traditional methods, such as statistical approaches and early machine learning techniques, which laid the groundwork for language processing. This will be followed by pre-trained models, which introduced the concept of leveraging large corpora for generating word embeddings. The section will then explore transformer-based models, which revolutionised NLP with their ability to capture context through advanced mechanisms. Finally, the section will end with a discussion about the emergence of large language models, highlighting their scale and capabilities in understanding and generating human-like text.

2.2.1 Earlier works in language models

One of the earliest methods for language modelling involved rule-based systems. These systems relied on manually crafted rules to process and generate text. While they were effective for specific tasks, their rigidity and lack of scalability posed limitations Jelinek (1998). As computational resources grew, researchers shifted towards more automated approaches.

The introduction of statistical methods to probabilistic models of natural languages marked a significant advancement in language modelling. These methods utilised probabilistic models to predict the likelihood of a sequence of words. One of the most well-known statistical models is the n-gram model. An n-gram model predicts the occurrence of a word based on the preceding $n-1$ words. For instance, a bigram model (n=2) considers the previous word to predict the next word, while a trigram model (n=3) considers the previous two words (Cavnar et al., 1994). Despite their simplicity, n-gram models proved effective in various applications such as speech recognition and machine translation (Jelinek, 1998). However, these models suffered from data sparsity, where many possible word combinations were not present in the training data, leading to poor performance in predicting rare or unseen sequences.

Hidden Markov Models (HMMs) extended the capabilities of n-gram models by incorporating hidden states to capture the underlying structure of language. HMMs became widely used in tasks such as part-of-speech tagging and named entity recognition (Rabiner, 1989). The hidden states in HMMs allowed for better handling of sequential data, improving the performance of language models on more complex tasks. However, HMMs also had limitations. They assumed that the probability of a word depended only on a limited number of preceding words, similar to n-gram models. This assumption restricted their ability to capture long-range dependencies in text (Manning and Schutze, 1999).

The application of machine learning to language modelling began with the use of feed-forward neural networks. These early neural language models aimed to learn word representations and predict the next word in a sequence based on a fixed context window. While they showed promise, they were computationally expensive and struggled with longer contexts (Bengio et al., 2000). Recurrent Neural Networks (RNNs) addressed some of the limitations of earlier neural language models by introducing recurrent connections. This allowed the model to maintain a hidden state that could capture information from previous time steps. RNNs proved effective in handling sequential data and long-range dependencies (Elman, 1990). However, RNNs also faced challenges, such as the vanishing gradient problem, which made training deep RNNs difficult. This issue was partially mitigated by the development of Long Short-Term Memory (LSTM) networks. The architecture of LSTMs maintains and updates a cell state over long sequences, making them particularly effective for tasks that require memory of the previous context. As a result, LSTMs are widely used in NLP applications such as machine translation (Ramadhan et al., 2022) and sentiment analysis (Murthy et al., 2020).

2.2.2 Transformer-based models

Transformer-based models represent another advancement in NLP, building on the limitations of the previously described statistical models HMMs and RNNs. These previous models struggled with capturing long-range dependencies and required sequential processing, which limited their efficiency and effectiveness. The transformer architecture, introduced by Vaswani et al. (2017) addresses these issues through its self-attention mechanism, allowing for parallel processing of input sequences.

Self-attention enables the model to weigh the importance of each word in a sequence relative to all other words, regardless of their position. This mechanism involves projecting each word into query, key, and value vectors, computing attention scores through dot products, and generating a weighted sum of value vectors. This allows transformers to capture context more effectively and handle long-range dependencies with greater accuracy. The ability to process sequences in parallel enhances computational efficiency, making transformers more scalable than RNNs and LSTMs. By overcoming the constraints of traditional methods and early neural networks, transformer-based models have revolutionized NLP, leading to substantial improvements in tasks such as machine translation (Zhu et al., 2020), text generation (Koncel-Kedziorski et al., 2019), and more.

2.2.3 Pre-trained language models

Models that are initially trained on vast amounts of text data to learn general language representations, can then be fine-tuned for specific applications. This approach contrasts with traditional methods that rely heavily on task-specific training data and often require feature engineering.

One of the early breakthroughs in pre-trained language models was the development of word embeddings, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). These models represented words as dense vectors in a continuous vector space, capturing semantic relationships between words based on their co-occurrence patterns in large corpora. Word embeddings enabled more effective and efficient use of linguistic data, leading to improved performance in various NLP tasks such as sentiment analysis and machine translation.

The advent of transformer-based architectures as discussed in Section 2.2.2, further advanced pre-trained models. One of the most notable examples is BERT (Bidirectional Encoder Representations from Transformers), which introduced a bidirectional training approach, allowing the model to consider the context from both the left and right of a target word. This bidirectional nature enabled BERT to achieve state-of-the-art results on a wide range of NLP benchmarks (Devlin et al., 2019).

RoBERTa (Robustly optimized BERT approach) architecture, which is an optimised version of BERT (Vaswani et al., 2017). RoBERTa enhances BERT by training with larger mini-batches and learning rates, as well as removing the next sentence prediction objective. This was found to be unnecessary for robust performance of this model. The RoBERTa architecture relies on transformer networks that utilise self-attention mechanisms to understand contextual relations between words in a sentence. This architecture helps the model to capture intricate language patterns and context, making it highly effective for various NLP tasks, including sentiment analysis (Liu et al., 2019).

2.2.4 Multilingual language models

Multilingual language models are designed to learn language representations from diverse linguistic datasets, allowing them to perform NLP tasks across different languages without the need for separate models for each language.

One of the pioneering multilingual models was Multilingual BERT (mBERT), a variant of BERT pre-trained on a large corpus comprising of more than 100 languages[1], has capabilities in understanding and processing multiple languages within a singular model architecture. mBERT inherits its architectural foundation from BERT, which is built on the Transformer model introduced by Vaswani et al. (2017). Following mBERT, several other multilingual models have been developed, each contributing to the field in different ways. XLM (Cross-lingual Language Model) and its successor XLM-R (XLM-Roberta) are notable examples. XLM uses a combination of masked language modelling and translation language modelling objectives to improve cross-lingual understanding (Conneau and Lample, 2019). XLM-R, an enhanced version, is trained on a larger and more diverse dataset, achieving state-of-the-art results on various multilingual benchmarks (Conneau et al., 2019).

[1] https://github.com/google-research/bert/blob/master/multilingual.md

LASER, on the other hand, is explicitly designed for creating language-agnostic sentence embeddings. It utilizes a single BiLSTM (Bidirectional Long Short-Term Memory) network that encodes 93 languages, trained on parallel texts to optimize for cross-lingual semantic similarity as shown in Figure 2. LASER's training regime includes a task-specific objective that encourages the model to align semantically equivalent sentences across languages in the embedding space, leading to its high performance in cross-lingual sentence alignment tasks (Artetxe and Schwenk, 2019).

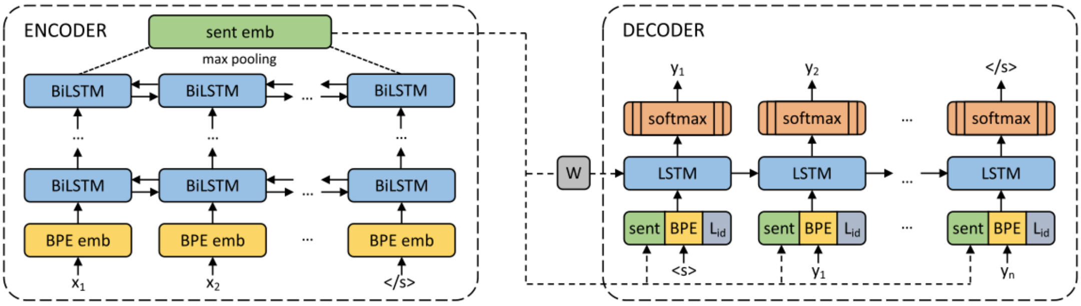


Figure 2: LASER architecture (Figure reproduced from Artetxe and Schwenk (2019))

2.2.5 Large Language Models (LLMs)

Large Language Models (LLMs) are advanced deep-learning models designed to understand and generate human-like text based on the input they receive. These models are trained on vast amounts of text data, enabling them to learn patterns, structures, and subtleties of human language. LLMs utilise architectures such as transformers, which allow them to capture long-range dependencies in text more effectively than traditional models (Chang et al., 2024).

LLMs have found applications across various fields. In healthcare, LLMs are used for processing clinical notes, predicting patient outcomes, and assisting in medical research (Lee et al., 2020). In the legal field, they assist in document review, contract analysis, and legal research (Chalkidis et al., 2020). LLMs can also analyse financial reports, and conduct market sentiment analysis (Yang et al., 2020). Additionally, LLMs are also used in the educational field by helping in grading essays, generating educational content, and providing personalized tutoring (Moore et al., 2023). Recently, LLMs have been known to power virtual assistants and chatbots (Freire et al., 2024).

Recent advancements in LLMs have pushed the boundaries of what these models can achieve. GPT (Generative Pre-trained Transformer) models, for example, focus on unidirectional training but are particularly effective at generating coherent and contextually relevant text. These models have been used in applications ranging from text completion to creative writing (Radford et al., 2019). GPT-3 by OpenAI, with 175 billion parameters, has demonstrated remarkable capabilities in text generation, translation, and summarisation (Brown et al., 2020). T5 (Text-To-Text Transfer Transformer) by Google has unified NLP tasks under a single framework, showing strong performance across various benchmarks (Raffel et al., 2020). Newer models like Gemini by Google (Team et al., 2023) and LLaMA by Meta (Touvron et al., 2023) have further expanded the capabilities and applications of LLMs in various domains.

2.3 Sentiment Analysis

Sentiment analysis aims to find the emotional tone behind textual data to understand opinions and attitudes. This section will introduce sentiment analysis and the models used to perform sentiment analysis. It will also explore the application of sentiment analysis in detecting biases within textual content, highlighting how it has been used to detect prejudices in various domains.

Additionally, we will discuss the challenges and advancements in applying sentiment analysis to low-resource languages, as they often lack extensive linguistic data and tools. This exploration will provide insights into the implications of sentiment analysis.

2.3.1 Introduction to Sentiment Analysis

Sentiment Analysis is a field within NLP that focuses on identifying and categorising opinions expressed in text, particularly to determine whether the writer's attitude towards a particular topic, product, or service is positive, negative, or neutral.

Currently, sentiment analysis is widely applied across diverse sectors. In the business domain (Rambocas and Gama, 2013), companies employ sentiment analysis to evaluate customer reviews (Gräbner et al., 2012), and social media comments (Yue et al., 2019) to gauge public opinion about their brand and products. This analysis assists in enhancing customer service, tailoring marketing efforts, and improving product offerings.

In media and politics, sentiment analysis tools analyse news articles (Taj et al., 2019), speeches (Rudkowsky et al., 2017), and social media posts (Sharma et al., 2020) to capture public sentiment regarding political events, policy decisions, or social issues. This application is particularly useful in predicting election results, understanding public policy impact, and crafting effective public communications.

Technologically, the landscape of sentiment analysis has evolved with the advancement of machine learning algorithms and deep learning techniques. Modern approaches often involve sophisticated models like LSTM (Long Short-Term Memory networks)(Murthy et al., 2020), BERT (Bidirectional Encoder Representations from Transformers) (Hoang et al., 2019), or GPT (Generative Pre-trained Transformer) (Leippold, 2023), which have improved the accuracy and granularity of sentiment detection.

Moreover, the integration of sentiment analysis with other technologies such as artificial intelligence (AI) (Ahmed et al., 2022) and big data analytics (Rokade and Aruna, 2019) has led to more robust and scalable solutions capable of handling vast amounts of unstructured text data across various digital platforms. This integration enables real-time sentiment analysis, providing instant insights into consumer behaviour and public opinion (Sharma and Goyal, 2023).

As sentiment analysis continues to grow, its applications are becoming more innovative and impactful. This prompts ongoing research and development to further increase the capabilities and accuracy of this analysis. This ongoing development highlights the importance of sentiment analysis in today's digital and data-driven landscape, where understanding human emotions and opinions plays a critical role in shaping people's opinions (Tan et al., 2023).

2.3.2 Uncovering ideological and political biases

Finding bias in published articles is difficult due to the huge size of text corpus from the articles and sparse hyperlink information (Gupta, 2009). Enevoldsen and Hansen (2017) focused on applying sentiment analysis to identify political biases in Danish newspapers, specifically analysing the representation of two political parties. Their research disclosed that the Danish newspaper *Berlingske* demonstrated a more positive slant for the Liberal Alliance party, thus shedding light on the newspaper's political leanings and biases. This indicates that sentiment analysis can serve as a tool for revealing emotional tone as well as political leanings within media publications. In another similar study, Al-Sarraj and Lubbad (2018) systematically crawled news articles from key Western media outlets to ascertain biases concerning the Israeli-Palestinian conflict. The study uncovered a significant pro-Israeli bias and subsequently constructed sentiment classifiers that predicted article bias using various machine-learning techniques. Rawat and Vadivu (2022) used sentiment analysis, bias scoring, and clustering to investigate biases in political news articles to categorise English news articles from India and the USA based on their affiliated media outlets and assess the extent of their biases towards specific political parties in India. This analysis helped to better understand relationships between media houses and political entities, potentially predicting the degree of bias exhibited by different publishers.

Smirnova et al. (2017) focused on how Russia and Islam are portrayed in The New York Times before and after significant events—the annexation of Crimea and the 9/11 attacks, respectively. The methodology involves extracting ideological cues from a corpus of political writings and representing the data as sequences of cues. This approach is enhanced by using a domain-informed Bayesian Hidden Markov Model (HMM) as done by Sim et al. (2013) to infer ideological proportions. This enables the quantification of how these representations shift in response to the events. Their findings reveal a distinct increase in negative sentiment in media portrayals post-event, evidenced by a rise in negative emotional language, rather than a reduction in positive expressions.

2.3.3 Multilingual Sentiment Analysis

The studies mentioned before provide critical insights into biases within English-language texts. The primary objective of our current study is to extend this to explore biases in different language editions of the same Wikipedia article. Languages are classified as *"high-resource"* or *"low-resource"* based on the availability of annotated datasets for developing language processing models. High-resource languages like English have extensive datasets that enable the use of diverse models with high accuracy. Conversely, low-resource languages, such as Afrikaans and Hindi lack sufficient data and thus hinder NLP applications (Nankani et al., 2020).

Sentiment analysis techniques require an accurate interpretation of word sentiments and contextual information to predict sentence polarity effectively. However, as Altowayan and Tao (2016) noted, this requirement poses a challenge for low-resource languages, where the creation of sentiment dictionaries is not feasible. The substantial investment in man-hours needed to develop these resources, along with the time required to tailor dictionaries to individual languages, makes the process impractical for low-resource languages.

Multilingual embeddings provide an alternative for sentiment analysis across diverse languages, these embeddings are trained to have representations in the same space over different languages. Such embeddings facilitate the transfer of knowledge from high-resource languages to low-resource ones. One of the leading models frequently used for multilingual studies is Multilingual BERT (mBERT) (Devlin et al., 2019). Many studies have conducted extensive empirical analyses for multiple NLP tasks and linguistic settings to determine the efficacy of this model. Their findings revealed that mBERT performs well in cross-lingual generalisation, especially when applied to languages with structural similarities (Wu and Dredze, 2019; K et al., 2020). In a related study, Li et al. (2022) applied sentiment analysis to political posts in Cantonese, specifically focusing on local forums in Hong Kong. Their study evaluates a variety of methods, including dictionary-based sentiment analysis, traditional machine learning models, fine-tuned BERT, and fine-tuned multilingual BERT (mBERT). According to the results obtained from their study, the fine-tuned mBERT model exhibited the best performance compared to the other methodologies.

Despite these advancements, the approach of using multilingual embeddings is not without its challenges. One limitation is the performance disparity across languages. Languages with linguistic structures that were vastly different from the languages most represented in the training corpus often do not benefit as much as those closely related (Pires et al., 2019). This results in sub-optimal performance for languages like Chinese, which has a unique script and syntactic structure, or Hindi, where the divergence in syntax and morphology from English or Roman languages can lead to poor transfer of sentiment analysis capabilities. Conneau et al. (2019) have shown that low-resource languages frequently exhibit more noise and variance in performance metrics compared to high-resource languages. This highlights the need for tailored approaches to mitigate these disparities for low-resource languages.

Another widely accepted method is translating datasets into a high-resource language, typically English, and then conducting sentiment analysis on the translated data (Chew et al., 2023). Denecke (2008) used the lexical resource SentiWordNet to find word sentiments, particularly focusing on adjectives to ascertain document polarity. The study was initially focused on German reviews, which were then translated into English for the SentiWordNet analysis. Using translatable sentiment dictionaries in multilingual sentiment analysis offers another framework that was used in the study of legislative conflict across different languages and countries. This approach allows for the extraction and analysis of sentiment from legislative bill debates, providing insights into the nature and intensity of conflicts within parliaments. The applications of this method in the study confirm its effectiveness in recovering the dynamics between government and opposition parties (Proksch et al., 2019). Baliyan et al. (2021) tackled the challenge of classifying vast amounts of multilingual text data for sentiment analysis by using machine translation to generate labels based on an abundantly available labelled English dataset. It involves employing Global Vectors for Word Representation (GloVe) for word embeddings, which are then processed through a Recurrent Neural Network-Long Short-Term Memory (RNN-LSTM) framework. Balahur and Turchi (2012) advocate for the utilisation of machine translation systems, such as Google Translate[1], as a preliminary step for conducting sentiment analysis across multiple languages. They translated texts from the annotated English dataset to German, Spanish, and French in their case study and then trained the sentiment classifier on the translated dataset. The rationale behind this strategy is the application of a stand-alone sentiment analysis model to a uniform dataset, thereby increasing the process's efficiency and scalability. According to their evaluations, these machine translation systems have reached a level of maturity that is sufficient for reliable deployment to generate training data for languages other than English. Moreover, they demonstrate that the performance metrics for sentiment analysis in translated text are comparable to those achieved when analysing English-language texts directly. Prettenhofer and Stein (2010) of-

[1] `https://translate.google.com`

fered an alternative approach from (Balahur and Turchi, 2012), they introduced Structural Correspondence Learning (SCL) and avoided the need for translation by training a classifier in the source language and subsequently transferring its learned features for text classification in the target language.

Upon reviewing existing research, it is evident that multilingual sentiment analysis presents considerable challenges. While various strategies have been explored in Section 2.3.3 to address these issues, machine translation emerged as the most feasible approach for our study. This method aligns well with the specific linguistic contexts of Afrikaans, Chinese, and Hindi, which are the languages used in our research.

2.3.4 Sentiment Analysis for longer texts

Our research focuses on the analysis of Wikipedia articles across different languages to examine potential biases. This requires a review of prior studies that have conducted sentiment analysis on longer texts, rather than sentences. Hong and Fang (2015) suggest that while paragraph vectors can achieve impressive results, they are notably challenging to tune, particularly for datasets like the Stanford Sentiment Treebank (Socher et al., 2013), where fine-grained and binary classifications are required. This difficulty sheds light on a broader issue in sentiment analysis: longer textual formats, such as paragraphs or extended documents, present greater obstacles due to their complex semantic structures that are harder to capture with current modelling techniques.

In longer texts, sentiment can vary between sections. For instance, an article might begin with a positive tone, shift to a critical or negative tone in the middle, and conclude on a hopeful note. Current sentiment analysis models may struggle to accurately aggregate these different sentiment expressions to provide an overall sentiment score. This variability demands advanced algorithms capable of segmenting the text and applying differential analysis to understand how sentiments evolve and interact throughout the text (Reagan et al., 2017). Traditional sentiment analysis tools, designed primarily for brief snippets like tweets or reviews, might not adapt well to the context shifts and the layered narrative structures found in longer documents. This can result in a loss of nuance, where subtle shifts in sentiment are overlooked or misinterpreted (Tsirmpas et al., 2023). Furthermore, in longer texts, the context or the way sentiments are expressed may change, making it difficult for sentiment analysis methods to maintain consistent accuracy. Sentiments expressed through irony, metaphor, or complex literary devices further challenge these tools, requiring more sophisticated natural language understanding capabilities that go beyond simple keyword recognition (Sharma and Goyal, 2023).

As explored in this section, using longer texts for sentiment analysis presents considerable challenges, particularly for low-resource languages as used in this project. These challenges require an approach to effectively capture sentiment. To address this, the text must be segmented into smaller units like sentences. This segmentation will help in the comparison of sentence-level sentiment across different language versions. The methodology for comparing these sentences involves content alignment, which is reviewed in Section 2.4. To obtain sentiments from sentences, we opted to utilise a sentiment analysis model that has been described in the next section.

2.4 Content Alignment

Content alignment is a key component in our research project, acting as the preliminary step that allows for comparison and subsequent sentiment analysis of Wikipedia articles across different language editions. This section offers an overview of existing literature concerning content alignment, particularly focusing on multilingual alignments.

2.4.1 Monolingual Alignment

Monolingual alignment involves the process of aligning sentences or words within a single language. Words with similar meanings in similar contexts are potential pairs for alignment. By using contextual evidence, we can identify and align these pairs (Sultan et al., 2014). Elhadad and Barzilay (2003) incorporate context into their alignment method through two complementary approaches: learning rules for matching paragraphs based on topic structure and refining these matches through local alignment to identify optimal sentence pairs. Their evaluation shows that this context-aware alignment method surpasses the performance of systems specifically designed for content alignment tasks. Their research demonstrates that even a weak sentence similarity measure, when combined with contextual information, outperforms methods that rely solely on advanced sentence similarity functions.

Some studies have approached the task of sentence alignment by framing it as a parse forest mapping problem. This methodology treats sentences as parse trees, where each node represents a grammatical component, and the structure reflects the syntactic organisation of the sentence. The alignment process then involves mapping these parse trees from one sentence to their counterparts in another, identifying correspondences between nodes across the forests. Gildea (2003) explored syntactic dependencies within aligned sentences to improve translation models. This was extended by Kadotani and Arase (2023) by aligning parse forests rather than just the best trees. This approach of conforming to syntactic structures significantly improves phrase alignment quality by efficiently mapping forests on a structured setup. This also addresses the challenge of syntactic ambiguities.

2.4.2 Multilingual Alignment

Earlier methods for content alignment often used statistical models and rule-based algorithms, such as the work by Gale et al. (1994). Despite their contribution, these methods faced challenges in scalability and robustness. For example, these methods were not able to account for multiple interpretations of the same phrases depending on the context of the article and hence had issues with content alignment after translation.

The advent of deep learning brought forth neural network-based methods, headed by the transformer model (Vaswani et al., 2017). These were applied to sentence alignment in bilingual corpora by Yang et al. (2018). These models were found to be computationally demanding with long sequences despite being effective (Wang et al., 2019). Transformer-based models are used to learn the mapping between parse forests of aligned sentences (Arase and Tsujii, 2020; Kadotani and Arase, 2023).

Cross-lingual word embeddings

Another approach involves the exploration of cross-lingual word embeddings (Conneau et al., 2017). They provide a shared semantic space across different languages. This method has opened new pathways, although they depend on hyperparameters and the quality of the embeddings, as noted by Ruder et al. (2019). These techniques have since been extensively applied in bilingual corpora, demonstrating their effectiveness in bridging linguistic variations and identifying relationships between languages.

In recent years, methods using mBERT for cross-language span prediction have emerged. Nagata et al. (2020) used a supervised word alignment method by using mBERT (mBERT) for cross-language span prediction, modelled as a question-answering task. Utilising token context for the predictions and using a dataset with gold standard alignments increased word alignment accuracy across multiple language pairs. This included Chinese, Japanese, German, Romanian, French, and English. This utilisation of token context achieved better performance over existing methods with minimal training data and offered competitive zero-shot alignment accuracy, surpassing traditional statistical techniques. Another method used for aligning sentences in noisy parallel texts is integer linear programming (ILP) (Chousa et al., 2020). mBERT brings new perspectives but might still present challenges in specific contexts and languages.

Zha et al. (2024) used LASER embeddings for tasks like paraphrase detection, semantic textual similarity, and co-reference resolution. This study highlights LASER's ability to generalise well across different NLP tasks and its effectiveness in aligning text pairs for a variety of applications.

2.5 Stance Detection

Stance detection aims to determine the position or viewpoint expressed in textual data concerning a specific target or entity. This section will introduce stance detection and explore its application in analysing textual content to identify biases and ideological leanings.

We will discuss the role of Named Entity Recognition (NER) in identifying targets for stance detection and the techniques used to handle entity ambiguity. Additionally, the section will cover the use of Large Language Models (LLMs) and specific models such as Generative Pre-trained Transformers (GPT) in stance detection tasks.

2.5.1 Introduction to Stance Detection

Stance detection seeks to identify the position or viewpoint conveyed by the text on a particular topic or target, regardless of the sentiment expressed which requires a detailed understanding of the text by capturing the underlying attitude, which can be independent of the emotional tone (Mohammad et al., 2017; Küçük and Can, 2020).

In news articles, stance detection can help identify biases or perspectives of different media outlets towards specific events or figures, providing insights into media bias and helping in the critical evaluation of information (Pomerleau and Rao, 2017). In social media, stance detection aids in understanding public opinion on various issues by analysing user-generated content, which is crucial for tasks such as sentiment analysis, opinion mining, and political analysis (Küçük and Can, 2020). Additionally, it plays a critical role in fact-checking systems by identifying articles that support or contradict a given claim, thereby enhancing the accuracy and reliability of automated fact-checking processes (Hanselowski et al., 2019).

Initial works in the field of stance detection have utilised traditional feature engineering techniques. HaCohen-Kerner et al. (2017) introduced skip character n-grams that permit intervals between characters or words within a defined scope. This methodology provides a more adaptable and potentially richer textual representation. In another study, Sen et al. (2018) formulated a set of features that they combined with a Support Vector Machine (SVM) model as well as a feed-forward neural network model.

Neural network models have also become widespread in stance detection, often outperforming earlier methods (Röchert et al., 2020). Du et al. (2017) put forth a neural network model to integrate information specific to the target for stance classification, leveraging an attention mechanism. This mechanism aimed to identify the key text segments that affect the target. Their approach achieved peak performance on both English and Chinese stance detection datasets. Umer et al. (2020) extended the approach of stance detection in fake news classification by integrating CNN and LSTM networks. The study used dimensionality reduction techniques such as Principal Component Analysis (PCA) and Chi-Square Test for feature selection by reducing the feature space through these dimensionality reduction techniques. The model thus gains computational efficiency thereby potentially reducing overfitting. This allows the CNN and LSTM components to focus on the most salient features for stance classification.

The majority of existing work in stance detection primarily focuses on modelling the sequence of words to learn a document's representation. However, this approach often overlooks other linguistic information, such as the polarity and the argumentative structure of the text, which can be beneficial in determining the document's stance. In response to this gap, Sun et al. (2018) introduced a neural model designed to consider these forms of linguistic information recognising that different linguistic elements exert varying degrees of influence on the document's stance. They then proposed a hierarchical attention network that is responsible for assigning weight to different linguistic features.

Due to the supervised nature of stance detection, it presents challenges in predicting stance toward articles that have not been previously analysed. Under earlier approaches, each new target requires the development of a new classifier based on an annotated dataset. Sun et al. (2022) introduced an adversarial attention network that aims to mitigate this limitation by incorporating multi-target data. This network serves to identify and link both target and sentiment information within the text. The adversarial mechanism is used to ascertain the topic and sentiment conveyed in each post, thus capturing some target-invariant information crucial for stance detection. Furthermore, the attention mechanism is deployed to associate articles sharing similar topics or sentiments, thus harvesting essential information for stance detection.

Identifying targets or entities within the text is crucial for stance detection, as it directly influences the accuracy and relevance of the analysis. This foundational step ensures that the stance is correctly attributed to the appropriate entity, which is particularly important in texts mentioning multiple entities. The following section on Named Entity Recognition (NER) delves into the techniques used for accurately identifying and categorising these targets, thereby laying the groundwork for stance detection.

2.5.2 Named Entity Recognition (NER)

Named Entity Recognition (NER) is a task in NLP that involves identifying and classifying entities in text into predefined categories such as person names, organizations, locations, dates, and other relevant entities. NER is crucial for various downstream tasks in NLP, including information extraction (Perera et al., 2020), question answering (Wongso et al., 2016), and stance detection (De Magistris et al., 2022), as it helps in structuring unstructured text data by tagging the entities mentioned within the text.

The early works in NER date back to the 1990s, when rule-based systems and handcrafted features were predominantly used. Rule-based approaches, which rely on manually crafted linguistic rules and patterns, were among the first methods used for NER. For instance, Appelt et al. (1993) developed a rule-based system that used linguistic patterns to identify entities in text. However, these approaches were limited by their reliance on domain-specific rules and lack of scalability. With the advent of machine learning, the focus shifted from rule-based approaches to statistical methods. Conditional Random Fields (CRFs) (McCallum and Li, 2003) and Hidden Markov Models (HMMs) (Morwal et al., 2012) became popular for sequence labelling tasks, including NER. McCallum and Li (2003) introduced a CRF-based model for NER that outperformed earlier methods by modelling the dependencies between labels in the sequence. Subsequently, deep learning techniques have improved NER by using neural networks to learn complex patterns from large datasets. Lample et al. (2016) proposed a neural architecture that combines bidirectional Long Short-Term Memory (BiLSTM) networks with CRFs, achieving state-of-the-art results on several benchmark datasets. This approach models both character-level and word-level features, enabling the system to capture intricate patterns in the data.

NER systems have predominantly focused on high-resource languages such as English, resulting in limited performance for low-resource languages. Multilingual

NER aims to extend the capabilities of NER systems to multiple languages. However, building NER systems for low-resource languages poses significant challenges due to the scarcity of annotated data and linguistic resources. One approach to address this issue is to use cross-lingual transfer learning, where models trained on high-resource languages are adapted to low-resource languages. Rahimi et al. (2019) demonstrated the effectiveness of cross-lingual transfer learning for NER in low-resource languages by leveraging bilingual word embeddings and pre-trained multilingual models. However, the complexity and variability of languages pose significant hurdles, making it challenging to achieve consistent performance across different languages. Another approach is to use translation-based methods, where text in low-resource languages is translated into a high-resource language like English, and then NER is performed using established models for the high-resource language.

SpaCy is an open-source NLP library that has gained popularity for its efficiency and ease of use in performing various NLP tasks, including NER. SpaCy provides pre-trained NER models for several languages, which have been trained on large annotated corpora and can be easily integrated into NLP pipelines. SpaCy's NER models leverage deep learning architectures, specifically, Convolutional Neural Networks (CNNs), to achieve high accuracy in entity recognition. Honnibal and Montani (2017) developed SpaCy with a focus on practical use cases, providing a robust and scalable solution for NER. Several studies have demonstrated the effectiveness of SpaCy for NER in diverse applications. For instance, Śniegula et al. (2019) evaluated the SpaCy library against other NER libraries and determined that SpaCy not only excelled in accuracy but was also the fastest in processing biomedical texts.

2.5.3 Handling Entity Ambiguity

Entity ambiguity is a common challenge in NER, referring to the difficulty in accurately identifying and categorizing entities that may have multiple meanings or representations. This can occur due to various reasons, such as abbreviations or different forms of names. For instance, “NLP” could refer to Natural Language Processing, or National Literacy Program, depending on the context. Similarly, an individual’s name might appear in different forms, such as “Bill Clinton” and “William Jefferson Clinton,” which can create ambiguity in identifying the same entity across different texts.

Wikipedia has emerged as a resource for handling entity ambiguity due to its extensive and well-organized content. The structure of Wikipedia articles, which includes hyperlinks, categories, and infoboxes, provides a rich source of contextual information that can be leveraged to disambiguate entities. Milne and Witten (2008) developed an approach that uses Wikipedia links to disambiguate entities by considering the context provided by linked articles. This method improves the accuracy of NER systems by utilising the data available in Wikipedia. Cucerzan (2007) introduced a method that utilises Wikipedia as a knowledge base to enhance NER. By linking entities in the text to their corresponding entries in Wikipedia, the system can disambiguate entities based on the contextual information provided by the encyclopedia.

2.5.4 Zero-Shot and Few-Shot Learning

Zero-shot learning is a paradigm in machine learning where a model is required to make predictions for classes or targets that it has never encountered during its training. In the context of stance detection, zero-shot learning means predicting the stance towards new targets without having seen any labelled examples for those targets during training. On the other hand, many large language models (LLMs), such as BERT are trained using data from Wikipedia, although the specific subsets or versions of Wikipedia used are not detailed. These models typically use few-shot learning, where they are provided with a small number of examples to adapt to new tasks or domains.

In this study, we are performing zero-shot and few-shot learning as we are trying to predict the stance towards common entities of different language editions of Wikipedia articles without altering the training data of the model for our study. Several studies have investigated the limitations of BERT in zero-shot and few-shot learning for stance detection. For instance, Sun et al. (2019) utilised BERT for aspect-based sentiment analysis, a task closely related to stance detection. They found that while BERT performed well with ample training data, its accuracy diminished in scenarios with limited data. Another research proposed WS-BERT, a model that infuses background knowledge from Wikipedia into the stance detection process. By incorporating this additional knowledge, WS-BERT enhanced the model's ability to accurately infer stances. Despite incorporating background knowledge, WS-BERT produced a Macro F1 score of 75.3 for zero-shot learning and 73.6 for few-shot learning He et al. (2022). Moreover, Liu et al. (2019) in their study on RoBERTa highlighted similar limitations. While RoBERTa outperformed BERT in many NLP tasks, its performance in zero-shot and few-shot learning scenarios remained sub-optimal.

The study highlighted that both BERT and RoBERTa require large annotated datasets to achieve high accuracy, which is not feasible in zero-shot learning contexts.

2.5.5 LLMs for Stance Detection

LLMs have shown great promise in stance detection, where the goal is to determine the position or attitude expressed in the text towards a particular target (Cruickshank and Ng, 2024). Another study showed that training a BERT model to focus on tokens identified by the log-odds-ratio on Twitter data increased stance detection performance (Kawintiranon and Singh, 2021). Schiller et al. (2021) utilised BERT for stance detection in social media, showing that LLMs can effectively capture the subtle differences in opinion expressed in short, informal texts. Tran et al. (2022) introduced an architecture that also uses transformers for stance detection in Vietnamese claims. The architecture also takes advantage of BERT for the extraction of context-aware word embeddings, as opposed to resorting to conventional Word2Vec models. These context-rich embeddings are then processed through Convolutional Neural Networks (CNNs) to capture local features, which serve as the training input for the stance detection model.

Generative Pre-trained Transformer (GPT), developed by OpenAI, has advanced the field of NLP by making use of extensive pre-training on diverse textual data followed by task-specific fine-tuning. This approach allows GPT models, such as GPT-3 (Brown et al., 2020) and GPT-4 (Achiam et al., 2023), to generate coherent and contextually relevant text, making them highly effective for various NLP tasks, including stance detection. Zhang et al. (2023) used a chain-of-thought (CoT) prompting strategy, which enhances the model's reasoning by breaking down complex queries into manageable steps. This method has been effective in improving the accuracy of stance predictions, especially in zero-shot scenarios where the model is not fine-tuned on the specific task but still needs to make accurate predictions based on its pre-existing knowledge. Burnham (2024) utilized GPT-3.5 for zero-shot stance detection, emphasising its capability to perform stance detection without relying on extensive labelled datasets.

While supervised classification requires training a model on annotated data, natural language inference (NLI) leverages pre-trained models to infer stance without the need for extensive labelled data. In-context learning with generative language models like GPT involves using prompts within the model to classify stances based on the provided context. The study finds that NLI classifiers can perform comparably to supervised classifiers when the document itself contains sufficient information for accurate classification, thereby negating the need for model training. However, when the external context is necessary, supervised and in-context classifiers prove to be more effective. Mets et al. (2024) annotated a large set of pro- and anti-immigration examples to train and compare the performance of multiple language models, including GPT-3.5 as a zero-shot classifier.

The results indicate that supervised models achieve acceptable performance, with Est-RoBERTa achieving an F1 macro score of 0.66, while GPT-3.5 yields a similar accuracy of 0.65 without requiring annotated data. This suggests that GPT-3.5 could be a simpler and more cost-effective alternative for text classification tasks in lower-resource languages.

2.6 Summary of Literature Review

This literature review has identified key areas of prior research on Wikipedia bias, sentiment analysis, content alignment, and stance detection, highlighting their relevance to our study. It reveals significant issues of gender bias, societal inequalities, and cultural disparities in Wikipedia content. Moreover, it discusses the challenges of applying sentiment analysis to low-resource languages and longer texts and the role of advanced NLP techniques, such as LLMs, in the field of stance detection.

Despite the progress made in prior research on Wikipedia, gaps remain in the nuanced analysis of ideological biases in Wikipedia, which our research aims to fill. By utilising NLP methodologies, we seek to provide deeper insights into how biases manifest across different language editions of Wikipedia, thereby contributing to a more global understanding of digital information accuracy and neutrality.

Chapter 3

Data Collection

In this study, we aim to obtain a selection of Wikipedia articles that are intrinsically polarising or open to interpretation. Such topics are usually fertile grounds for the emergence of divergent viewpoints, particularly when authored by contributors from diverse backgrounds.

Each contributor brings their own cultural, political, and ideological leanings into the editorial process. This often creates inconsistent or differing perspectives, which could manifest as subtle or overt biases within the articles. Such articles generally revolve around socially debatable topics, such as politically controversial issues[1], ongoing conflicts[2], or past wars[3]. The rationale for focusing on these specific categories of topics is twofold :

1. Controversial issues and conflicts are often subject to multiple interpretations, offering a fertile ground for studying systematic biases.

2. The contentious nature of these topics ensures that they will be covered across multiple language editions of Wikipedia, providing a rich data set for comparative analysis.

To facilitate the extraction of Wikipedia articles from the previously mentioned lists of controversial topics such as ongoing conflicts and wars, we have used the web scraping tool *BeautifulSoup*. This tool has been chosen due to its effectiveness in parsing HTML and XML documents.

[1] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

[2] https://en.wikipedia.org/wiki/List_of_ongoing_armed_conflicts

[3] https://en.wikipedia.org/wiki/Lists_of_wars

The subsequent steps were implemented to extract text from Wikipedia articles and determine the availability of articles in various language editions specifically Afrikaans, Chinese, and Hindi were utilised in this study.

3.1 Handling article title ambiguity

In collecting articles, the initial approach based solely on title searches proved inadequate. This was mainly because not all articles possess direct title translations across different languages. This is exemplified by the English article about the 1989 protests in China titled ”1989 Tiananmen Square protests and massacre”, while its Chinese counterpart is titled ” 六四事件” (June Fourth Incident). This linguistic variation could potentially result in the omission of pertinent articles with varying translated titles. To mitigate this challenge, the Wikidata IDs[1] were utilised as a more reliable method of retrieval. Wikidata IDs act as a common identifier across all language editions, ensuring that articles corresponding to the same concept or event, despite linguistic variations in titles, are accurately located.

3.2 Utilizing chunking

To manage data collection from Wikipedia effectively and to reduce the likelihood of request timeouts from the Wikipedia API[2] , data was gathered in segmented ’chunks.’ This method also known as chunking (Dozza et al., 2013) facilitated the handling of articles in smaller subsets, which improved the processing of data across multiple languages and increased the efficiency of the data retrieval process. This approach ensures that the system remains responsive and can handle large volumes of data without performance degradation.

[1] https://www.wikidata.org/wiki/Wikidata:Identifiers

[2] https://api.wikimedia.org/wiki/Main_Page

3.3 Extracting articles from Wikipedia lists

We utilized Wikipedia lists of controversial issues and past wars, as described earlier, to identify relevant articles. By checking their availability in different languages using Wikidata IDs, we ensured accurate identification of corresponding articles across different language editions. The text of these articles was then extracted using *BeautifulSoup*. The quantity of articles obtained for different language editions is presented in Table 2.

Language Edition	**Number of Articles**
Available in Afrikaans	615
Available in Chinese	1265
Available in English	1985
Available in Hindi	612
Available in all four languages	**452**

Table 2: Number of Wikipedia articles available in English, Hindi, Afrikaans, and Chinese language editions

3.4 Data Cleaning

This step involved cleaning the article text and selectively extracting text from the summary section of Wikipedia articles. An examination of the Wikipedia markup was conducted to use *BeautifulSoup* with precision, ensuring that only the desired text was retrieved without irrelevant content. Text cleaning was carried out in two stages:

1. Excise citations and citation numerals from the text since the research focuses solely on the article's textual information, not the bibliographic or numerical details. It is important to clarify that while all citations and most citation numerical data were removed, certain context-specific numerical information, such as the number of reported deaths in an event, may be retained if it

contributes to the article’ s content. This is because numerical information can contribute to bias; the presence or absence of statistics in articles can reflect inherent biases. Therefore, retaining numerical data adds substantive context to the topic.

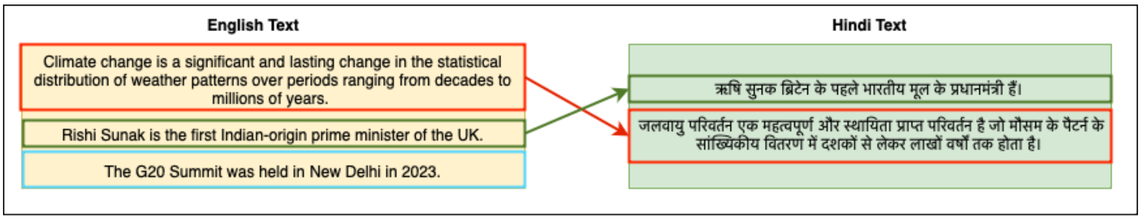
2. The second stage included stripping away any text within parentheses, as these often include translations or phonetic instructions that could cause issues during the subsequent content alignment phase. Such text in parentheses tends to correspond more closely across languages due to the use of identical characters, potentially complicating the alignment process described in Chapter 4.

Chapter 4

Content Alignment

This chapter discusses the methodological framework and findings of content alignment within our study, This is a key component that can help uncover biases in different Wikipedia language editions. Content alignment assists in aligning sentences of articles across languages. Initially, our methodology used Multilingual BERT (mBERT) (Devlin et al., 2019) embedding, a model designed to understand multiple languages. However, despite mBERT's capabilities in past research (Wu and Dredze, 2019; K et al., 2020; Li et al., 2022), it fell short of providing the desired level of accuracy in aligning content across different languages. This prompted a shift towards LASER (Language-Agnostic SEntence Representations) (Artetxe and Schwenk, 2019; Schwenk and Douze, 2017), a more adept tool for our specific requirements. This chapter outlines the journey from mBERT to LASER, detailing the rationale behind our methodological pivot and presenting the results obtained through LASER embeddings.

4.1 What is Content Alignment?



Climate change is a significant and lasting change in the statistical distribution of weather patterns over periods ranging from decades to millions of years.

Rishi Sunak is the first Indian-origin prime minister of the UK.

The G20 Summit was held in New Delhi in 2023.

ऋषि सुनक ब्रिटेन के पहले भारतीय मूल के प्रधानमंत्री हैं।

जलवायु परिवर्तन एक महत्वपूर्ण और स्थायिता प्राप्त परिवर्तन है जो मौसम के पैटर्न के सांख्यिकीय वितरण में दशकों से लेकर लाखों वर्षों तक होता है।

Figure 3: Content alignment between sample English and Hindi texts.

The role of content alignment in this research project is to conduct a comparison of articles from various language editions of Wikipedia. As depicted in Figure

3, a straightforward comparison between the sample texts in English and Hindi presents complexities. These complexities arise because the sentences that convey equivalent meanings are frequently not positioned adjacently within the texts. Furthermore, some sentences conveying critical information in one language edition may be completely absent in another, underscoring the challenges in achieving an accurate alignment. This irregularity is mainly because articles in different language editions are not translations of each other, but independently authored articles by different contributors. This provides a foundation for further investigations into biases to ensure that articles from different languages are matched to ensure analytical accuracy.

4.2 Experimental Setup

In the development of a methodology for cross-lingual sentence alignment, our initial approach used the Multilingual BERT (mBERT) model, pre-trained across a concatenated corpus from 104 languages. mBERT's design leverages a shared sub-word vocabulary and helps it to understand and process text from multiple languages, capturing linguistic characteristics without being constrained by any specific language (Devlin et al., 2019). However, our initial investigations revealed limitations in mBERT's performance concerning semantic alignment across languages (Artetxe and Schwenk, 2019). This can be attributed to mBERT's training methodology, which does not explicitly focus on parallel data nor is it directly optimised for cross-lingual alignment objectives. Past research indicates that the inherent structure of mBERT's multilingual embeddings does not uniformly benefit all languages. The performance disparities are particularly evident in tasks like Named Entity Recognition (NER) and part-of-speech tagging, where the bottom 30% of languages in mBERT's training set perform worse than models specifically trained for those languages (Wang et al., 2020). Given these challenges, we pivoted to utilising the

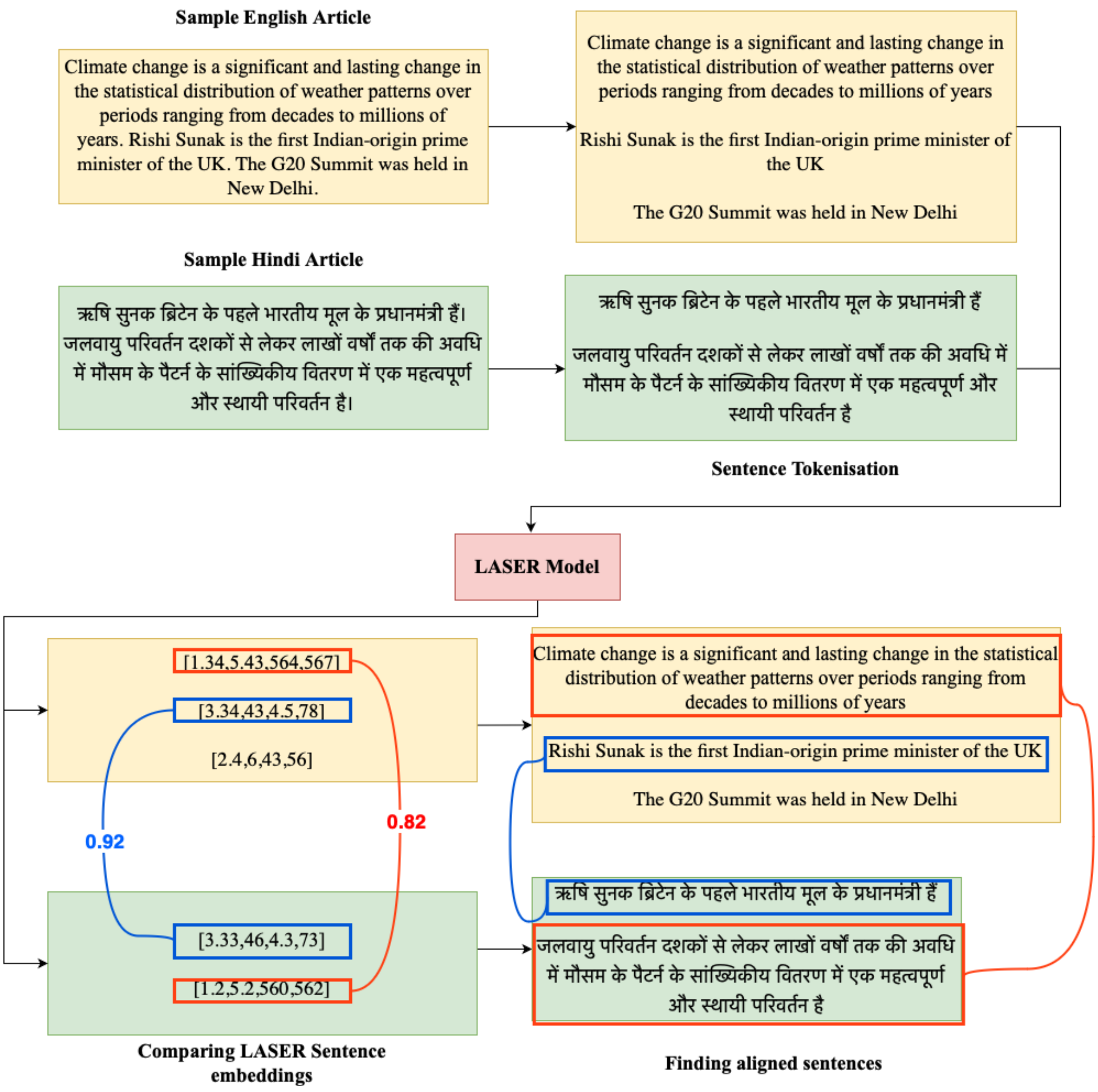

Sample English Article
Climate change is a significant and lasting change in the statistical distribution of weather patterns over periods ranging from decades to millions of years. Rishi Sunak is the first Indian-origin prime minister of the UK. The G20 Summit was held in New Delhi.
Climate change is a significant and lasting change in the statistical distribution of weather patterns over periods ranging from decades to millions of years
Rishi Sunak is the first Indian-origin prime minister of the UK
The G20 Summit was held in New Delhi
Sample Hindi Article
ऋषि सुनक ब्रिटेन के पहले भारतीय मूल के प्रधानमंत्री हैं। जलवायु परिवर्तन दशकों से लेकर लाखों वर्षों तक की अवधि में मौसम के पैटर्न के सांख्यिकीय वितरण में एक महत्वपूर्ण और स्थायी परिवर्तन है।
ऋषि सुनक ब्रिटेन के पहले भारतीय मूल के प्रधानमंत्री हैं
जलवायु परिवर्तन दशकों से लेकर लाखों वर्षों तक की अवधि में मौसम के पैटर्न के सांख्यिकीय वितरण में एक महत्वपूर्ण और स्थायी परिवर्तन है
Sentence Tokenisation
LASER Model
[1.34,5.43,564,567]
[3.34,43,4.5,78]
[2.4,6,43,56]
0.92
0.82
[3.33,46,4.3,73]
[1.2,5.2,560,562]
Comparing LASER Sentence embeddings
Climate change is a significant and lasting change in the statistical distribution of weather patterns over periods ranging from decades to millions of years
Rishi Sunak is the first Indian-origin prime minister of the UK
The G20 Summit was held in New Delhi
ऋषि सुनक ब्रिटेन के पहले भारतीय मूल के प्रधानमंत्री हैं
जलवायु परिवर्तन दशकों से लेकर लाखों वर्षों तक की अवधि में मौसम के पैटर्न के सांख्यिकीय वितरण में एक महत्वपूर्ण और स्थायी परिवर्तन है
Finding aligned sentences

Figure 4: Content alignment methodology

Language-Agnostic SEntence Representations (LASER) model as the foundational technology for our sentence alignment methodology (Schwenk and Douze, 2017). The architectural underpinnings of LASER offer an advantage for our objectives: it is trained on parallel corpora and explicitly optimised for cross-lingual semantic similarity. This training approach equips LASER with the capability to produce language-agnostic sentence embeddings, thereby enabling a more accurate comparison of semantic content across languages (Artetxe and Schwenk, 2019). This section explains the process used to align sentences from multilingual editions of Wikipedia articles and the steps involved in achieving the same as shown in Figure 4.

For each language pair under consideration, we collected Wikipedia articles about identical topics as explained in Chapter 3. The process of segmenting the articles into individual sentences, known as tokenisation, was conducted with careful consideration of the grammatical conventions specific to each language. Tokenisation for English and Afrikaans was relatively direct, owing to the transparent sentence demarcation. For the tokenisation of English and Afrikaans sentences specifically, we used the sentence tokeniser from the Natural Language Toolkit (NLTK) (Bird et al., 2009). On the other hand, the segmentation of Chinese text involved identifying sentence separators such as full stops (。), exclamation marks (！), question marks (？), and semicolons (；), followed by the deployment of regular expressions to separate sentences. In the case of Hindi, the sentence tokenisation was done by using the Indic NLP library (Kunchukuttan, 2020), which is tailored for processing Indian languages.

Once tokenised, each sentence is encoded using the LASER model to generate fixed-length embeddings that capture the essential meaning of the sentences (Artetxe and Schwenk, 2019). Once embeddings have been obtained for sentences across the various languages under study, we proceed to measure similarity using metrics such as cosine similarity (Hadj Taieb et al., 2013) between the embeddings of sentences from the respective languages.

For the pairing of sentences, a similarity matrix is constructed where each entry represents the cosine similarity between two articles. For instance, consider Article A with sentences $\boldsymbol{A}$ and $\boldsymbol{B}$, and Article B with sentences $\boldsymbol{a}$ and $\boldsymbol{b}$. The similarity matrix is represented as follows:

$$\begin{array}{c|cc} & a & b \\ \hline A & Aa & Ab \\ B & Ba & Bb \end{array}$$

Each element of the matrix, such as $\boldsymbol{Aa}$ or $\boldsymbol{Ab}$, quantifies the cosine similarity between sentences from the two articles. Higher cosine similarity means that the sentences are similar. The alignment process involves selecting the pair with the highest cosine similarity score from the matrix for matching, thus ensuring that each sentence is aligned with its most semantically similar counterpart in the other article. Once a pairing is made, those sentences are removed from consideration in subsequent rounds of alignment, maintaining a strict one-to-one mapping. This procedure is repeated iteratively until all sentences that exceed a specified similarity threshold are aligned, or all the sentences of the shorter article are aligned.

We then conducted content alignment on three sample articles of different topics to assess the efficiency and accuracy of content alignment across all languages. We then proceeded with human evaluation to verify the outcomes.

4.3 Human Evaluation Results

The process of human evaluation consisted of involving native speakers of the three examined languages—Afrikaans, Chinese, and Hindi—to scrutinise the articles that had been aligned with English texts. The starting point for the cosine similarity threshold was chosen to be on the lower side. To test the adequacy of this threshold, the reviewers were presented with three content-aligned articles that had different numbers of sentence pairs in each: 66 pairs for English-Afrikaans, 76 for English-Chinese, and 47 for English-Hindi article sets. The sample outcomes of this final annotation process are presented in Table3.

Upon obtaining the annotated data, a precision-recall analysis was conducted to empirically establish an optimal cosine similarity threshold. The construction of the precision-recall curve is a widely recognized method for evaluation in information retrieval and natural language processing tasks (Davis and Goadrich, 2006). The precision-recall curve illustrates the trade-off between the accuracy (precision) of the alignment against its completeness (recall) across various threshold settings.

The annotated results were graphically represented in a precision-recall graph, as depicted in Figure 5. The graph is used in discerning the threshold at which the precision is maximized while still retaining a reasonable level of recall. A threshold value that is too low would yield high recall but low precision, thereby introducing noise in the form of misaligned sentences into the dataset. Conversely, a threshold that is too high may increase precision but at the cost of excluding relevant alignments, hence reducing recall.

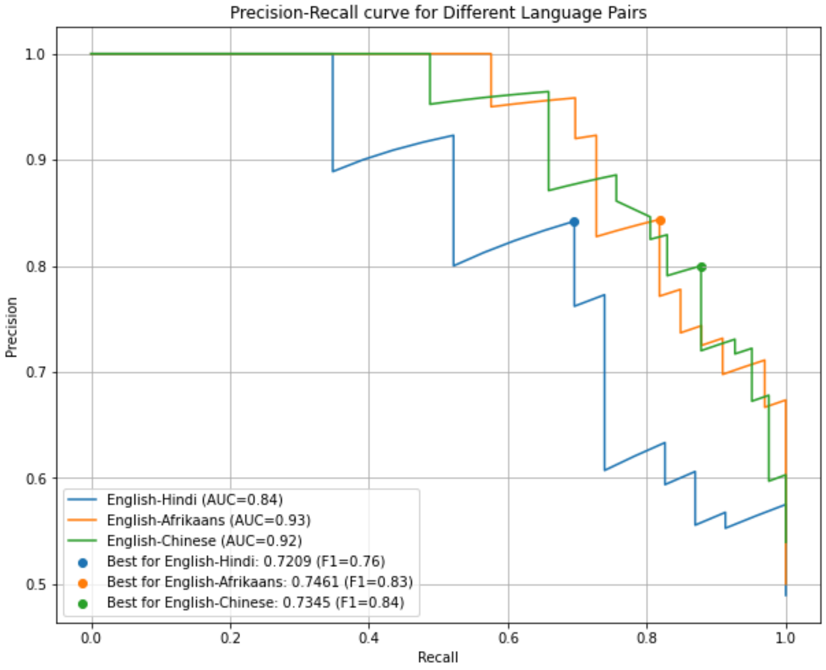


Figure 5: Precision-Recall Curve for annotated data

From the analysis, it emerged that a cosine similarity threshold of at least ***0.75*** is judicious across all language pairs. At this juncture, the precision is sufficiently high to ensure that the sentiment analysis is not adversely affected by the inclusion of misaligned sentence pairs. This threshold was selected on the premise that high precision is paramount, as it guarantees the reliability of aligned sentences for subsequent processing tasks. The recall is accepted at a lower value to maintain this reliability, which is a standard approach in tasks where the quality of the output is more critical than the quantity (Schütze et al., 2008). This balance is crucial for sentiment analysis, where the cost of including a misaligned pair is significantly higher than excluding a correctly aligned one (Buckland and Gey, 1994).

To further substantiate this finding, an analysis of the F1 score, which is the harmonic mean of precision and recall, was undertaken. This measure is an indicator of the test's accuracy as it considers both precision and recall (Van Rijsbergen, 1974). The best F1 scores for each language pair, which align closely with the 0.75 threshold, lend additional credence to its selection. The F1 score, a harmonic mean of precision and recall, is a critical metric in the evaluation of classification models, especially in the context of NLP. Its significance is underscored in empirical studies, such as the one by Schütze et al. (2008), which demonstrated that an F1 score effectively balances the trade-off between precision and recall.

After implementing a cosine similarity threshold of 0.75 for all articles across different languages, articles were excluded because none of their sentences met this criterion. The count of remaining articles is presented in Table4 below.

4.4 Limitations

A notable limitation in our study is the omission of unaligned sentences, as depicted in Figure 3, which could significantly impact the analysis. The disparity in content volume between Wikipedia articles—where English versions are often more detailed than their counterparts in other languages (Rajcic, 2017) —necessitates the identification of sentences that are mismatched or absent in translations for a balanced comparative analysis.

Moreover, another limitation is in handling compound sentences. In some instances, a sentence in the source text may be split into multiple sentences in another language. For example, a sentence like *"Jane works in the NLP Research wing and has been elected as the team lead for 2 consecutive years"* may be present in another language edition in two distinct sentences such as *"Jane presently works in the NLP Research Wing. She has been elected to lead the team for 2 consecutive*

Table 3
Cosine similarity scores between English and non-English sentences

English sentence	Non-English sentence	Cosine Similarity Score	Human Annotator (Y/N)
The Mexican Revolution was an extended sequence of armed regional conflicts in Mexico from approximately 1910 to 1920.	Die Meksikaanse Rewolusie was ’n burgeroorlog in Meksiko tussen konserwatiewe teen-rewolusionêre magte en liberale rewolusionêre magte vanaf 20 Februarie 1910 tot 1920.	0.8210	Y
In November 2022, he announced his candidacy for the Republican nomination in the 2024 presidential election.	Hy het in Junie 2015 sy kandidatuur vir die Amerikaanse presidentskap van November 2016 aangekondig en is op 19 Julie 2016 deur die Republikeinse Party amptelik as die party se presidentskandidaat benoem.	0.7871	N
A hundred hours after the beginning of the ground campaign, the coalition ceased its advance into Iraq and declared a ceasefire.	Een honderd uur na die grondinval verklaar die koalisie ’n skietstilstand.	0.7636	Y
These health effects can reduce life expectancy by 10 years.	酗酒會讓個人的預期壽命縮短大約十年。	0.7345	Y
The meaning of life can be derived from philosophical and religious contemplation of, and scientific inquiries about, existence, social ties, consciousness, and happiness.	从狭义上讲，它探究的是生物和社会文化的演变中，特别是探究智人可能意义的问题。	0.6951	N
The coalition’s efforts against Iraq were carried out in two key phases: Operation Desert Shield, which marked the military buildup from August 1990 to January 1991; and Operation Desert Storm, which began with the aerial bombing campaign against Iraq on 17 January 1991 and came to a close with the American-led Liberation of Kuwait on 28 February 1991.	以美國🄵首的多國部隊在取得聯合國授權後，於 1991 年 1 月 17 日開始對科威特和伊拉克境🄵的伊拉克軍隊發動軍事進攻，主要戰鬥包括歷時 42 天的空襲、在伊拉克、科威特和沙特阿拉伯邊境地帶展開的歷時 100 小時的陸戰。	0.7902	N
The bulk of the coalition’s military power was from the United States, with Saudi Arabia, the United Kingdom, and Egypt as the largest lead-up contributors, in that order; Saudi Arabia and the Kuwaiti government-in-exile paid around US$32 billion of the US$60 billion cost to mobilize the coalition against Iraq.	गठबंधन में सैन्य बलों का बहुमत संयुक्त राज्य अमेरिका, सऊदी अरब, संयुक्त राष्ट्र और इजिप्ट से प्राप्त हुआ, ये इसी क्रम में अग्रणी योगदानकर्ता देश थे।	0.7636	Y
From 1920 to 1940, revolutionary generals held office, a period when state power became more centralized and revolutionary reforms were implemented, bringing the military under the control of the civilian government.	मध्य वर्ग पहले की तुलना में सशक्त हुआ और सेना का भी देश की राजनीति पर प्रभाव बढ़ गया।	0.6796	N
Many other related questions include: “Why are we here?”	कई अन्य संबंधित प्रश्नों में सम्मलित हैं: □□हम यहाँ क्यों हैं?"	0.9126	Y

years.” This kind of structural discrepancy can result in the fragmentation of aligned content, where one part of the original sentence is matched while the other segment

Table 4: Number of Aligned Sentences Across Language Pairs

Language Pair	Number of Articles	Number of Aligned Sentences
English and Hindi	449	1855
English and Chinese	433	2328
English and Afrikaans	450	1704
Chinese and Hindi	360	993
Afrikaans and Hindi	295	605
Afrikaans and Chinese	400	1132
Common in all six language pairs		**243 common articles**

is disregarded. Such occurrences can distort the sentiment analysis, as they may inadvertently exclude significant portions of the text that contribute to the overall sentiment. Therefore, this aspect of sentence structure and its potential to skew analysis should be carefully considered in future research to enhance the accuracy and reliability of cross-lingual sentiment analysis.

Additionally, our methodology involved the participation of three native speakers to evaluate the alignment accuracy from English to the tested languages. However, we were unable to assess the content alignment quality for translations between languages that do not include English, resorting to a cosine similarity threshold of 0.75 for these language pairs.

Chapter 5

Sentiment Analysis

This chapter presents an examination of the sentiment analysis conducted on Wikipedia articles across different language editions, a step towards understanding the nuances of sentiment variation and bias. To achieve this, the study harnesses the RoBERTa-based sentiment analysis model. Given the multilingual nature of the data, a pre-processing step involved translating the aligned content pairs into English using Google Translate[1]. This ensured compatibility with the sentiment analysis model, which was trained predominantly on English data.

Subsequently, the study used the Jensen-Shannon Divergence (JSD) (Lin, 1991), a method for measuring the similarity between two probability distributions, to quantify the sentiment variations across the language pairs.

5.1 Preparing data for the sentiment analysis model

As shown in Table 4, we gathered **243** content-aligned articles common to the six language pairs under investigation. These articles were divided into two predominant categories: controversial topics and war topics, comprising of **212** and **31** content-aligned articles respectively.

The non-English were then translated via Google Translate[1]. This step was essential, as the sentiment model articles have been primarily trained on English language data. Given that there are limited resources available for multilingual and language-specific sentiment analysis in the languages we are studying - Afrikaans, Chinese and Hindi, we opted to utilise translation services.

[1] `https://translate.google.com`

Research has demonstrated that commercial translation tools like Google Translate do not significantly distort the semantic integrity of the text, preserving essential contextual elements (Artetxe et al., 2017). Therefore, using such services can be an effective solution for conducting sentiment analysis across languages with lower resources. Furthermore, studies have suggested that, despite some inherent translation inaccuracies, the overall sentiment of phrases tends to be preserved, making this approach viable for initial analyses intended to identify broad sentiment trends rather than nuanced interpretations (Poncelas et al., 2020; Mohammad et al., 2016).

5.2 Performing sentiment analysis

We have used a sentiment analysis model trained on 58 million tweets and fine-tuned it for sentiment analysis with the TweetEval benchmark (Barbieri et al., 2020). For each sentence, we used the RoBERTa-based tokenizer (Liu et al., 2019) to convert the text into tokens compatible with the model. The model then performs sequence classification, outputting logits for each sentiment category. These logits are converted into probabilities, which signify the model's confidence across different sentiment classes, namely positive, neutral, and negative. The tokenizer and model are applied with their default configurations as provided by Hugging Face[2] , ensuring that the analysis benefits from the pre-tuned parameters optimised for sentiment analysis tasks for English. We apply this sentiment model to classify our data without any further tuning.

The classification process uses the model's understanding of language nuances and sentiment expressions learned during its training on a large corpus of Twitter data. This training dataset is particularly appropriate for our analysis because, like Twitter data, the sentences we are examining are short in length.

[2] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

5.3 Outcomes of Sentiment Analysis Experiments

In this section, we present the experimental setup, results, and subsequent analyses conducted to explore sentiment differences across various language editions of Wikipedia using a pre-trained sentiment analysis model. The primary objective of these experiments was to investigate how sentiments expressed in Wikipedia articles vary linguistically and culturally. Our experimental framework involved six different language pairs, where sentences of an article for a language pair are aligned as detailed in Chapter 4. These pairs include alignments between English, Afrikaans, Chinese and Hindi. The sentiment analysis model as described earlier in Section 5.2 was utilised to determine the divergences in sentiments in the content-aligned articles.

5.3.1 Mean Jensen-Shannon Divergence (JSD)

Jensen-Shannon Divergence (JSD) is a method used to measure the similarity between two probability distributions. It is a symmetrized and smoothed version of the Kullback-Leibler (KL) divergence (Ji et al., 2020). This makes it more suitable for practical applications in comparing probability distributions derived from different sources. In this study, JSD is used to calculate the divergence in sentiment probabilities across different language editions of Wikipedia articles.

Jensen-Shannon Divergence is defined as follows:

$$JSD(P||Q) = \frac{1}{2}\left(KL(P||M) + KL(Q||M)\right) \tag{5.1}$$

where P and Q are the probability distributions to be compared, and M is the average distribution:

$$M = \frac{1}{2}(P + Q) \tag{5.2}$$

KL divergence, which measures the difference between two probability distributions, is defined as:

$$KL(P||M) = \sum_i P(i) \log \left(\frac{P(i)}{M(i)} \right) \tag{5.3}$$

$$KL(Q||M) = \sum_i Q(i) \log \left(\frac{Q(i)}{M(i)} \right) \tag{5.4}$$

Thus, the JSD can be expressed in terms of these KL divergences. This method is particularly useful in our context because it handles the divergence between the sentiment distributions of aligned sentences from different language editions in a symmetric and interpretable manner (Lin, 1991).

In the context of this research, the sentiment distributions P and Q represent the probabilities of sentences expressing positive, negative, and neutral sentiments in different language editions. By calculating the JSD for these distributions, we can quantify how differently the same content is perceived sentiment-wise across languages.

It is particularly useful in assessing the variability in text data across different contexts (Pechenick et al., 2015). Figure 6 illustrates the methodological framework used to compute the mean JSD for a sample article within our study. For each pair of sentences in each article, a JSD value is calculated based on the sentiment probabilities provided by the model. These values are then aggregated across all articles for a language pair, resulting in an overall mean JSD. Additionally, the dataset was categorised into two primary groups: controversial topics and war-related topics. Detailed discussions of the results and their inferences are discussed below.

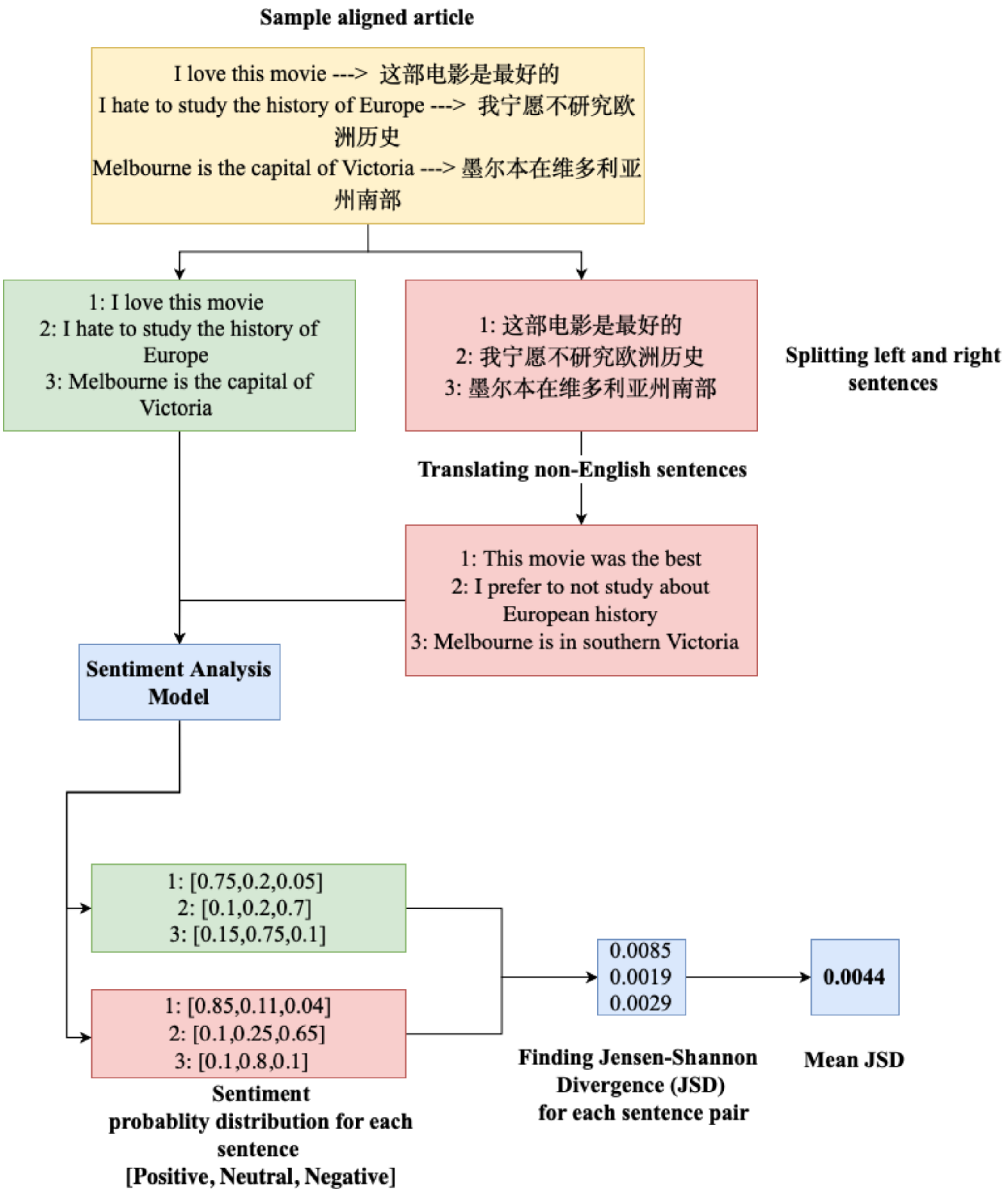

Sample aligned article
I love this movie ---> 这部电影是最好的
I hate to study the history of Europe ---> 我宁愿不研究欧洲历史
Melbourne is the capital of Victoria ---> 墨尔本在维多利亚州南部
1: I love this movie
2: I hate to study the history of Europe
3: Melbourne is the capital of Victoria
1: 这部电影是最好的
2: 我宁愿不研究欧洲历史
3: 墨尔本在维多利亚州南部
Splitting left and right sentences
Translating non-English sentences
1: This movie was the best
2: I prefer to not study about European history
3: Melbourne is in southern Victoria
Sentiment Analysis Model
1: [0.75,0.2,0.05]
2: [0.1,0.2,0.7]
3: [0.15,0.75,0.1]
1: [0.85,0.11,0.04]
2: [0.1,0.25,0.65]
3: [0.1,0.8,0.1]
Sentiment probablity distribution for each sentence [Positive, Neutral, Negative]
0.0085
0.0019
0.0029
Finding Jensen-Shannon Divergence (JSD) for each sentence pair
0.0044
Mean JSD

Figure 6: Mean Jensen-Shannon Divergence (JSD) Methodology

Round-trip translation

To ensure that the observed differences in sentiment across language pairs are due to variations in sentiment expression rather than artifacts introduced by translation processes, we used a round-trip translation technique. This involved translating English articles into Hindi and then back into English. This allowed us to compare the original English text with its round-trip translated version to identify any alterations in meaning or sentiment that could occur due to the translation process alone.

Results and Inference

Table 5: Mean Jensen-Shannon Divergence (JSD) for Different Topics (RTT = Round-trip translated)

Language Pair	Overall Mean JSD	Controversial JSD	War JSD
Afrikaans-Chinese	0.1616	0.1603	0.1705
Afrikaans-Hindi	0.1411	0.1391	0.1659
Chinese-Hindi	0.1578	0.1566	0.1694
English-Afrikaans	0.1354	0.1377	0.1049
English-Chinese	0.1490	0.1490	0.1574
English-Hindi	0.1306	0.1315	0.1255
English-RTTEnglish	0.0383	0.0377	0.0443

Table 5 reports the result of the mean JSD values across language pairs. The results from the round-trip translation are considerably lower compared to other language pairs, indicating that the observed sentiment differences are due to variations in the sentiment conveyed by the text itself, rather than being artifacts of the translation process.

As one can see for some language pairs, there is consistency in their relative positions across different topics. For example, English-Hindi consistently exhibits the lowest JSD across all three categories, indicating agreement in these languages on these topics. We further note that language pairs involving Chinese tend to exhibit greater JSD compared to other combinations. This indicates a marked divergence in the use of Chinese compared to the other languages assessed. On the other hand, English when paired with other languages, tends to be on the lower end of JSD values, indicating that English may have a more universal discourse or is more often used as a common reference point in global discussions.

The English-Afrikaans pair displays a decrease in JSD when moving from controversial to war topics, which could imply a similar understanding or usage of language when discussing war-related content. On the other hand, both pairs involving Hindi with Afrikaans and Chinese see a notable increase in JSD in war topics compared to controversial topics, suggesting that discussions about wars may differ more significantly in these language pairs, possibly due to cultural or historical perspectives.

Furthermore, a review of Figure 7 and Figure 8 reveals a uniform dispersion of JSD values across various language pairs, despite the different subject matters under consideration The frequency of outliers also remains stable between these figures. However, it is to be noted that articles related to controversial topics exhibit a higher occurrence of outliers in comparison to those associated with war topics. This discrepancy in outlier prevalence may be due to the substantial difference in the volume of articles analyzed for each category, with 212 articles about controversial topics and 31 articles about war topics.

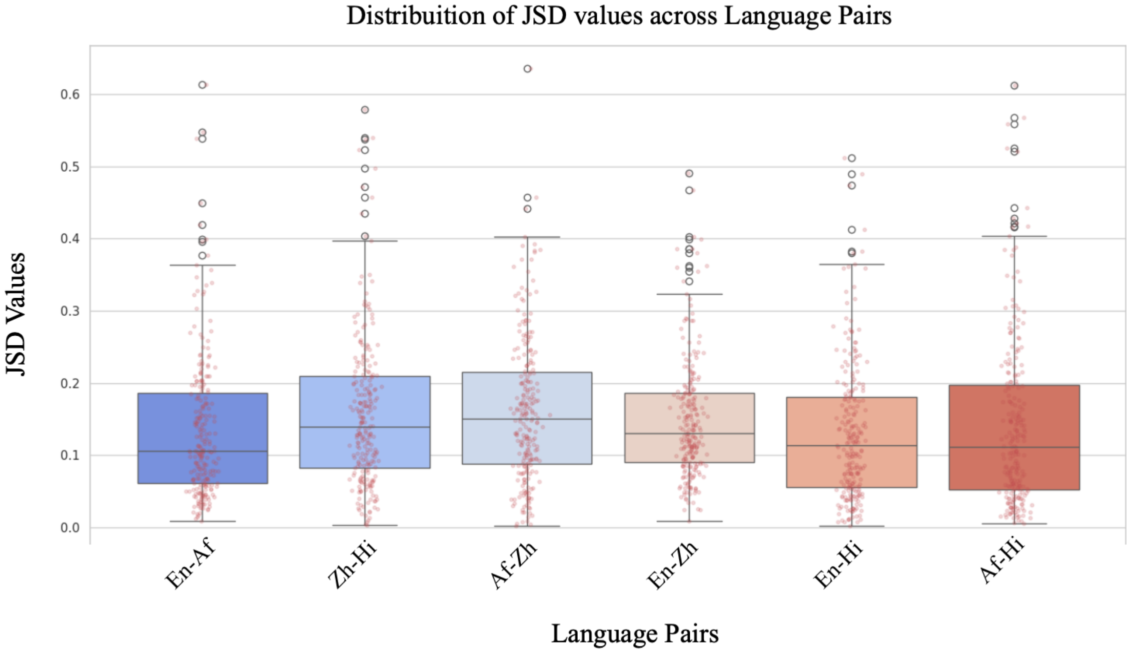
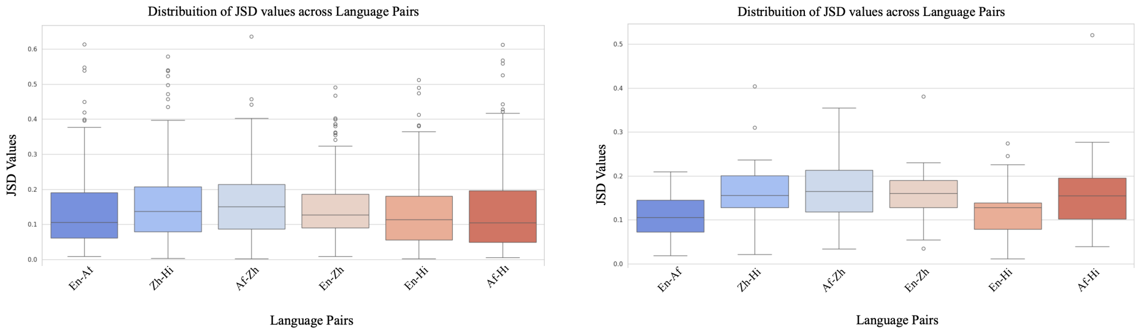


Figure 7: Overall Mean JSD across language pairs



(a) Controversial Topics

(b) War Topics

Figure 8: Mean JSD across language pairs for Controversial and War Topics

5.3.2 Different Sentiment Proportion (DSP)

As discussed in Section 5.3.1, the mean Jensen-Shannon Divergence (JSD) indicates that while there are differences in sentiment between language editions, quantifying the extent of these differences is challenging due to the probabilistic characteristics of divergence metrics. To deepen our understanding of the nuances in sentiment divergence, we have introduced an additional experimental approach termed Different Sentiment Probability (DSP).

Contrary to the mean JSD, which calculates the divergence between the probability distributions of sentiments, DSP simplifies the analysis by assigning a binary value to each sentence pair. In this method, each sentence in a pair is classified into a single sentiment category; if both sentences share the same sentiment, the pair is labelled as "0" (indicating agreement), otherwise, it is labelled as "1" (indicating disagreement). This binary classification is then aggregated over all aligned sentence pairs to determine the proportion of sentences exhibiting sentiment differences. A sample framework for this analysis is illustrated in Figure 9.

The binary and discrete nature of DSP offers a clearer perspective on the prevalence of sentiment agreement or disagreement. This provides insights into the extent of sentiment alignment across different languages.

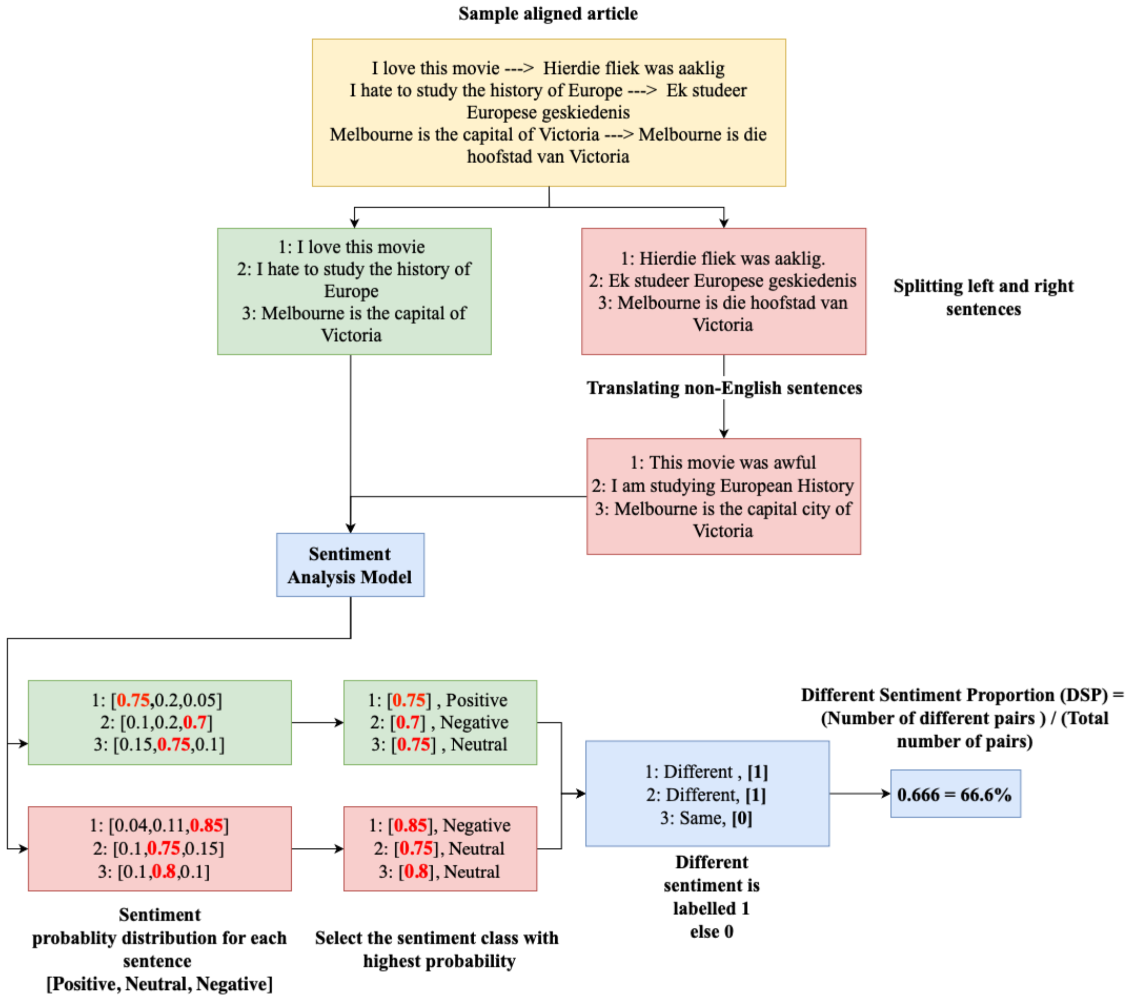

Sample aligned article
I love this movie ---> Hierdie fliek was aaklig
I hate to study the history of Europe ---> Ek studeer Europese geskiedenis
Melbourne is the capital of Victoria ---> Melbourne is die hoofstad van Victoria
1: I love this movie
2: I hate to study the history of Europe
3: Melbourne is the capital of Victoria
1: Hierdie fliek was aaklig.
2: Ek studeer Europese geskiedenis
3: Melbourne is die hoofstad van Victoria
Splitting left and right sentences
Translating non-English sentences
1: This movie was awful
2: I am studying European History
3: Melbourne is the capital city of Victoria
Sentiment Analysis Model
1: [0.75,0.2,0.05]
2: [0.1,0.2,0.7]
3: [0.15,0.75,0.1]
1: [0.75] , Positive
2: [0.7] , Negative
3: [0.75] , Neutral
1: [0.04,0.11,0.85]
2: [0.1,0.75,0.15]
3: [0.1,0.8,0.1]
1: [0.85], Negative
2: [0.75], Neutral
3: [0.8], Neutral
1: Different , [1]
2: Different, [1]
3: Same, [0]
Different Sentiment Proportion (DSP) = (Number of different pairs) / (Total number of pairs)
0.666 = 66.6%
Different sentiment is labelled 1 else 0
Sentiment probablity distribution for each sentence [Positive, Neutral, Negative]
Select the sentiment class with highest probability

Figure 9: Different Sentiment Proportion Methodology

Results and Inference

The results from the DSP experiment can be analysed by two different methods, with or without article boundaries — these use the binary labels assigned to each sentence pair as shown in sample Figure 9. This binary classification spans multiple articles grouped by language pairs, further categorised into topics such as war and controversial topics. Each method provides unique insights into the sentiment alignment across these categories.

1. *Without Article Boundary:* This broader approach aggregates the data across all articles within the same language pair, regardless of the article boundaries. It measures the overall proportion of sentence pairs with differing sentiments

across a larger corpus. This method offers a macro-level perspective on sentiment divergence and is suitable for assessing general trends in sentiment alignment across larger datasets.

2. *With Article Boundary:* This method computes the average number of sentence pairs within a single article that exhibits differing sentiments. It provides a localised view of sentiment disparity, allowing for an article-by-article analysis of sentiment alignment.

Table 6: Different Sentiment Proportion across language pairs

Language Pair	**Without Article Boundary DSP**			**With Article Boundary DSP**		
	Overall	**Controversial**	**War**	**Overall**	**Controversial**	**War**
Afrikaans-Chinese	0.2480	0.2403	0.3222	0.2217	0.2158	0.2897
Afrikaans-Hindi	0.1857	0.1795	0.250	0.1581	0.1554	0.1694
Chinese-Hindi	0.2229	0.2191	0.2639	0.2199	0.2198	0.2157
English-Afrikaans	0.1901	0.1854	0.2558	0.1967	0.1969	0.2310
English-Chinese	0.1814	0.1749	0.2719	0.1972	0.1946	0.2724
English-Hindi	0.1655	0.1634	0.250	0.1816	0.1819	0.2280

As observed in Table 6 in the analysis without considering article boundaries, a macroscopic observation suggests that sentiment alignment varies across language pairs. The English-Hindi pair exhibits the lowest proportion of sentiment disagreement, indicating a higher degree of agreement in sentiment across this particular language pair. This finding is congruent with the Mean JSD results, which also showed a lower divergence for English-Hindi, reinforcing the notion of a closer sentiment alignment between these languages. Conversely, the Afrikaans-Chinese pair displays the highest DSP value, reflecting a pronounced sentiment misalignment. This trend again echoes the observations from the Mean JSD results, where the same language pair exhibited the highest mean divergence, suggesting a consistent disparity in sentiment across different analytical methods.

Within the war-related topics, an increased DSP value is observed across all language pairs in the absence of article boundaries. This elevation may be due to the difference in sample size compared to controversial topics.

When juxtaposing overall and controversial topics, we found higher sentiment agreement for controversial topics, except for the English-Afrikaans pair. Such an increase may imply that controversial topics express a more uniform sentiment expression within these language communities, although the English-Afrikaans language maintains a more consistent sentiment alignment across both topic categories.

Incorporating article boundaries into the analysis yields a DSP value that does not substantially deviate from the aforementioned broader analysis. This suggests that sentiment alignment maintains a level of consistency, irrespective of whether the sentiment is assessed within individual articles or across a more expansive dataset.

The results of the DSP analysis, particularly when observed in conjunction with the Mean JSD data, emphasise the importance of applying diverse methodological approaches to gain an understanding of sentiment alignment. While the Mean JSD offers insights into the probabilistic distribution of sentiments, the DSP provides a more dichotomous perspective. Together, these methodologies offer a multi-faceted view of sentiment divergence. Importantly, the inclusion of DSP simplifies the interpretation of Mean JSD results by providing clearer insights into sentiment divergence. The key point is that the numbers provided here are more interpretable: a value of 0.25 indicates that, on average, one-quarter of the sentences exhibit differing sentiments. JSD does not convey this level of detail. Overall, approximately 20% of the sentences display differing sentiments across various language editions. This clarification facilitates a more straightforward understanding of the analyses presented in Section 5.3.3, increasing the understanding of our findings and illustrating the practical significance of using diverse analytical techniques.

5.3.3 Mean JSD and Sentence Length Difference Correlation

This section extends the analysis presented in Figure 6, where we investigate the potential correlation between sentiment divergence and sentence length differences within aligned sentence pairs. The decision to explore this relationship was prompted by observations from a set of content-aligned sentence pairs, where JSD values exceeded 0.4, indicating sentiment differences. Some examples are displayed in Table 7, illustrating the basis for our analytical focus on sentence length difference.

Table 7: Sentence Pairs and Their Length Differences

Sentence Pair	Length Difference
”Elvis Aaron Presley was an American musician and actor.” ”Elvis Aaron Presley, an American rock singer, musician and film actor, is regarded as one of the most important cultural icons of the 20th century.”	14 words
”Henry Ford was the founder of the Ford Motor Company.” ”Henry Ford introduced the Ford Model T, which revolutionized transportation and American industry.”	3 words
”It originated in 1982 through the amalgamation of various Shiite groups to fight against the Israeli Invasion of Lebanon.” ”Hezbollah is a Shiite Islamic political and military organization funded by Iran in 1982 with a purpose of eliminating Israel and expel Western forces from Lebanon.”	7 words
”Nonbelievers contend that atheism is a more parsimonious position than theism and that everyone is born without beliefs in deities; therefore, they argue that the burden of proof lies not on the atheist to disprove the existence of gods but on the theist to provide a rationale for theism.” ”Atheism or atheism or atheism is the doctrine that does not accept the existence of any God who created the world, operates it and controls it, on the basis of lack of universally accepted evidence.”	18 words
”Bernard Lawrence Madoff was an American former market player, investment advisor, financier, fraudster and convicted felon, who had to serve a federal prison sentence for offenses related to a massive Ponzi scheme.” ”Bernard Lawrence Madoff was an American financial broker and former chairman of Nasdaq.”	19 words

Li et al. (2020) conducted a study on online reviews and discovered that classification performance is particularly susceptible to changes in the dataset at the word count level of the reviews. Another study by Chen (2023) indicates that for Twitter tweets, RNNs with LSTM and pre-trained embeddings yield the highest accuracy; for IMDB reviews, logistic regression paired with TF-IDF is most accurate; and for Yelp reviews, RNNs with LSTM and trainable embeddings perform best. While comparing these three datasets, it becomes evident that the differences among the models are small however, the size of the training set impacts prediction accuracy. This suggests that text length plays a crucial role in sentiment classification outcomes.

This method involves calculating the absolute difference in sentence length between each aligned pair and then determining the Pearson correlation between these length differences and the JSD for each sentence pair within a language pair. The Pearson correlation, a common statistical method in NLP to assess the strength and direction of linear relationships between variables (Cohen et al., 2009), is used for our analysis.

Results and Inference

The result of the analysis of the Pearson correlation coefficients, as shown in Table 8, presents the relationship between sentence length differences and Mean JSD across language pairs and topics.

Table 8: Different Sentiment Proportion across language pairs

Language Pair	Pearson Correlation			P-Value		
	Overall	Controversial	War	Overall	Controversial	War
Afrikaans-Chinese	0.1290	0.1428	0.0071	0.0003	0.0000	0.9467
Afrikaans-Hindi	0.1744	0.1824	-0.0936	0.0001	0.0001	0.5270
Chinese-Hindi	0.1700	0.1820	-0.0068	0.0000	0.0000	0.9545
English-Afrikaans	0.1764	0.1818	0.0305	0.0000	0.0002	0.7318
English-Chinese	0.1916	0.2005	0.1023	0.0000	0.0000	0.2789
English-Hindi	0.1947	0.2048	-0.1069	0.0000	0.0000	0.3854

The correlation coefficients across the majority of language pairs and topics are positive, albeit small in magnitude. This suggests a tendency for greater differences in sentence length to correspond with an increase in sentiment divergence.

A closer inspection of the English-Chinese and English-Hindi pairs reveals a more pronounced correlation in the context of controversial topics. This may imply that in these linguistic pairings, sentence length variations are potentially more impactful on sentiment divergence.

A noteworthy deviation is observed in war topics. The English-Hindi and Afrikaans-Hindi language pairs show a negative correlation. This anomaly indicates that for the corpus, sentence length discrepancies are not predominantly responsible for sentiment divergence in war-related discourse. The subdued correlation coefficients within war topics could be a function of the relatively limited dataset size, consisting of only 31 articles, compared to the 212 articles examined for controversial topics. This low volume of data may restrict the establishment of a robust correlation. It may also hint at the presence of alternative determinants of sentiment divergence within war-related content.

The modesty of the correlation coefficients across all language pairs and topics indicates that sentence length variances, while contributory, are not a strong factor in influencing sentiment divergence. Additional language features that may be useful for correlating sentiment divergence are outlined in Section 5.5.

5.4 Conclusion

In conclusion, the exploration of sentiment analysis across different language pairs and topical categories in this chapter has provided insights into the nature of cross-linguistic sentiment alignment. Utilising the Mean Jensen-Shannon Divergence (JSD) and the Different Sentiment Probability (DSP), along with correlational analysis between sentence length disparities and sentiment divergence, has increased our understanding of the complex dynamics in multilingual sentiment expression.

The Mean JSD analysis indicated varying degrees of sentiment divergence across language pairs, with English-Hindi consistently demonstrating the closest alignment. This was further corroborated by DSP results, which highlighted patterns of sentiment agreement, particularly within controversial and war-related topics. The use of DSP analysis, by simplifying sentiment to binary classification, has offered a clearer perspective on sentiment alignment that complements the probabilistic nature of the Mean JSD.

Moreover, the correlation analysis shed light on the relationship between linguistic features such as sentence length and sentiment divergence, revealing that while there is an association, it is not an overwhelmingly determinant factor. This suggests that sentiment divergence is influenced by several language factors.

The findings from this chapter highlight the importance of adopting a different methodological approach when conducting sentiment analysis across languages. The complexities inherent in natural language and the subtleties of human emotion require thorough analysis and methodological diversity to capture the essence of sentiment alignment and divergence accurately.

Given that our project focuses on the examination of Wikipedia articles, the role of entity recognition is crucial for discerning sentiment divergence across different linguistic editions. It is essential to identify named entities and their associative sentiment within the text. Chapter 6 extends this examination by exploring stance detection, which evaluates the alignment or disagreement with specific entities or topics.

5.5 Limitations and Future Work

While this chapter has provided insights into sentiment divergence across various language pairs using sentence length differences, mean Jensen-Shannon Divergence (JSD), and different sentiment proportions (DSP) several limitations must be addressed.

Firstly, the reliance on translation for low-resource languages like Afrikaans and Hindi can introduce inaccuracies due to the loss of linguistic and cultural nuances. Future studies might benefit from using advanced multilingual models that are capable of processing text in native languages, thereby preserving original sentiment contexts. Additionally, the use of binary classification in sentiment analysis, though effective for simplifying the computational process, may overlook subtleties in sentiment expression.

Another limitation of this work is that it focuses solely on aligned sentence pairs, disregarding unaligned sentence pairs. However, this limitation will be addressed in Chapter 6, where the analysis will expand to include common entities, ensuring that unaligned sentences are also considered.

Future research into syntactic complexity (Lu, 2010), lexical diversity (McCarthy and Jarvis, 2010), and part-of-speech (POS) tags (Nerabie et al., 2021) correlating to sentiment divergence could also provide deeper insights into the intricacies of language and sentiment.

Chapter 6

Stance Detection

This chapter addresses the task of stance detection, which distinguishes itself from sentiment analysis by determining the text's attitude towards specific topics or entities. Stance detection is particularly useful for analysing texts where multiple viewpoints may be present, such as multilingual Wikipedia articles.

The methodology adopted for this analysis involves several key steps. Initially, non-English Wikipedia articles are translated into English to standardise the input data, helping in consistent entity identification across languages. Then, generative pre-trained transformer (GPT-3) (Brown et al., 2020) is utilised to identify stances towards these recognised entities. To quantitatively evaluate the variations in these stances, a specific metric has been developed to quantify stance divergence.

The validity of this metric and the overall approach is confirmed through human validation. This chapter outlines these procedures and prepares the ground for discussion in subsequent sections, highlighting the importance of stance detection in the analysis of multilingual divergences in Wikipedia.

6.1 What is Stance Detection and how is it different from Sentiment Analysis?

Sentiment analysis is primarily concerned with identifying the polarity of a text—determining if the text conveys a positive, negative, or neutral tone (Nandwani and Verma, 2021). Its main objective is to find the emotional or subjective elements embedded within the text. However, sentiment analysis might not adequately address or detect biases that arise when the same content is presented differently, potentially affecting the perception of the target based on its portrayal.

In contrast, stance detection specifically aims to ascertain the explicit stance or viewpoint that the text adopts towards a particular topic or entity, which is not necessarily tied to the emotional tone conveyed. This distinction allows stance detection to offer a more nuanced understanding of the text by capturing the underlying attitudes that may remain distinct from the emotive content as illustrated in Table 9 (Küçük and Can, 2020; Mohammad et al., 2017). Considerable research within stance detection has focused on various applications, such as analysing the stance in news articles (Ghanem et al., 2018) and evaluating the orientation towards rumours (Derczynski et al., 2017). These studies highlight the critical role of stance detection in discerning not just the sentiment, but the intended implication and viewpoint within complex texts of multiple domains.

Stance analysis requires a specific "target" to be meaningful (i.e., identifying a stance as "against" is nonsensical without a specified target). In this study, the targets are significant "entities" within the articles. Consequently, the initial step involves extracting these entities. Additionally, the implication of content alignment performed during sentiment analysis can be problematic. The approach often discards unaligned sentences, which could provide valuable insights. For example, if one language variant specifically discusses the atrocities of war while another variant does not, our previous analyses would miss this crucial difference.

6.2 Entity Identification

Entity identification comprises two stages. The initial stage is translating articles from non-English languages into English. Then, a named entity recognition (NER) model is applied to identify entities within the text as explained in Section 6.2.1. The next stage is the normalisation of entities. This addresses the consolidation of

Table 9: **Example of stance versus sentiment**

Example Sentence	Sentiment	Stance	Target
”The novel was brilliantly written, but I disagree with its glorification of war.”	Positive	Against	Glorification of War
”I hate to say it, but the movie was good.”	Positive	Support	The movie
”I hate how the new tax law forces us to change, but I understand it’s essential for the economy.”	Negative	Support	New Tax Law

entities that appear in multiple forms or are mentioned multiple times, as detailed in Section 6.2.2. Figure 10 illustrates the entity identification process for a sample article, demonstrating the sequential steps involved in translating, recognising, and normalising entities.

6.2.1 Need for translation and applying NER Model

The initial phase of processing non-English articles involves translation using Google Translate[1]. This step has two purposes:

1. *Consistency in Entity Identification:* Translation standardises the text across various languages, ensuring that the entity identification process remains uniform regardless of the original language. This consistency is crucial for comparative analysis across different linguistic datasets.

2. *Addressing the Lack of NER Models for Low-Resource Languages:* The languages involved in this project are considered low-resource in the context of computational linguistic tools such as Named Entity Recognition (NER) models. Specifically, no NER models for these languages have sufficient annotated data, comparable to resources like the CoNLL datasets (Tjong Kim Sang and De Meulder, 2003), which are used to train and deploy NER systems. This

[1] https://translate.google.com/

scarcity necessitates using translation to leverage more developed NER capabilities available for English.

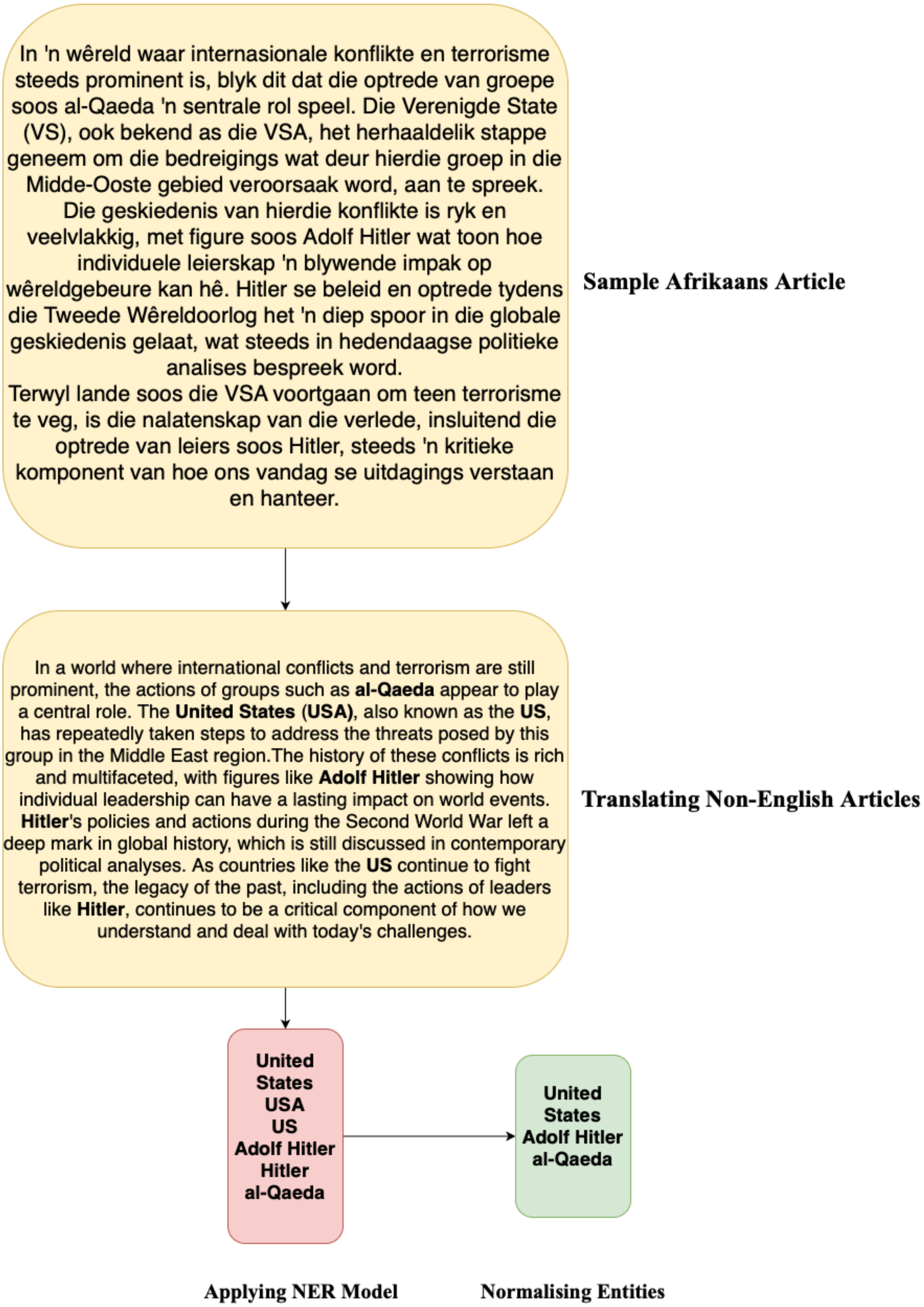
Following the translation, the articles—except those originally in English, which do not require translation—are processed using the SpaCy model for NER. SpaCy is selected for its efficiency and accuracy in detecting named entities within large text corpora, making it a suitable choice for extracting entities from the translated articles (Honnibal and Montani, 2017).

6.2.2 Normalisation of entities

For the normalisation and disambiguation of entities identified in the texts, we used the Wikipedia library (Cucerzan, 2007; Bunescu and Paşca, 2006). This tool facilitated the elimination of duplicates and abbreviations, ensuring a clearer and more precise entity recognition process, as illustrated in Figure 10. This approach has been effective in resolving ambiguities associated with entity names.

The normalisation process involves converting various representations of the same entity into a single, standardised form. For example, different references to the same person (e.g., *"Barack Obama"*, *"Obama"*, *"President Obama"*) are normalised to a single canonical name. This is achieved by querying Wikipedia's API[1] to retrieve the canonical name for each entity, thereby ensuring consistency and accuracy in the entity recognition process. The tool also helps disambiguate entities by distinguishing between entities with similar or identical names based on context, such as distinguishing between *"Apple"* the technology company and *"apple"* the fruit.

[1] `https://en.wikipedia.org/w/api.php`



In 'n wêreld waar internasionale konflikte en terrorisme steeds prominent is, blyk dit dat die optrede van groepe soos al-Qaeda 'n sentrale rol speel. Die Verenigde State (VS), ook bekend as die VSA, het herhaaldelik stappe geneem om die bedreigings wat deur hierdie groep in die Midde-Ooste gebied veroorsaak word, aan te spreek.

Die geskiedenis van hierdie konflikte is ryk en veelvlakkig, met figure soos Adolf Hitler wat toon hoe individuele leierskap 'n blywende impak op wêreldgebeure kan hê. Hitler se beleid en optrede tydens die Tweede Wêreldoorlog het 'n diep spoor in die globale geskiedenis gelaat, wat steeds in hedendaagse politieke analises bespreek word.

Terwyl lande soos die VSA voortgaan om teen terrorisme te veg, is die nalatenskap van die verlede, insluitend die optrede van leiers soos Hitler, steeds 'n kritieke komponent van hoe ons vandag se uitdagings verstaan en hanteer.

In a world where international conflicts and terrorism are still prominent, the actions of groups such as **al-Qaeda** appear to play a central role. The **United States** (**USA)**, also known as the **US**, has repeatedly taken steps to address the threats posed by this group in the Middle East region.The history of these conflicts is rich and multifaceted, with figures like **Adolf Hitler** showing how individual leadership can have a lasting impact on world events. **Hitler**'s policies and actions during the Second World War left a deep mark in global history, which is still discussed in contemporary political analyses. As countries like the **US** continue to fight terrorism, the legacy of the past, including the actions of leaders like **Hitler**, continues to be a critical component of how we understand and deal with today's challenges.

Figure 10: Entity Identification for Sample Afrikaans Article

6.3 Finding stance towards entities

6.3.1 Identifying common entities

Following the identification of entities as explained in Section 6.2, the next step was recognising common entities across articles from different languages. After entities are extracted and normalised for each article, we then compute the shared entities for a language pair of an article by collecting the set of entities that have occurred at least once in both languages. This selection criterion was implemented to recognise and focus exclusively on significant entities for upcoming analytical processes.

6.3.2 Using Large Language Models (LLMs)

GPT-3, developed by OpenAI, is suitable for this study due to its comprehensive training on a diverse range of textual data. This extensive training allows GPT-3 to handle a variety of NLP tasks. The ability of GPT-3 to produce context-sensitive responses is useful for analysing complex textual data such as Wikipedia articles, where stances are intricately interwoven with cultural and linguistic nuances (Brown et al., 2020). Figure 11 illustrates the setup used for GPT-3, detailing the prompts and inputs utilised in determining stances within the model.

In this study, we are using GPT-3 in a zero-shot manner, where our prompt includes an explanation of the task in natural language. This means we do not use in-context learning to include example input-output pairs, nor do we perform any fine-tuning. Instead, we rely on the model's pre-trained capabilities to understand and perform the task based on the natural language instructions provided in the prompt.

The specific model used for this study is GPT-3.5 Turbo, which we accessed via the OpenAI API on 27 January 2024. We send a prompt to the GPT-3.5 Turbo model that explains the task in natural language, asking the model to identify the stance expressed towards a specified target within a given text. The prompt specifies three possible stances: 'Favour,' 'Against,' or 'Neutral,' and the model's response is interpreted accordingly.

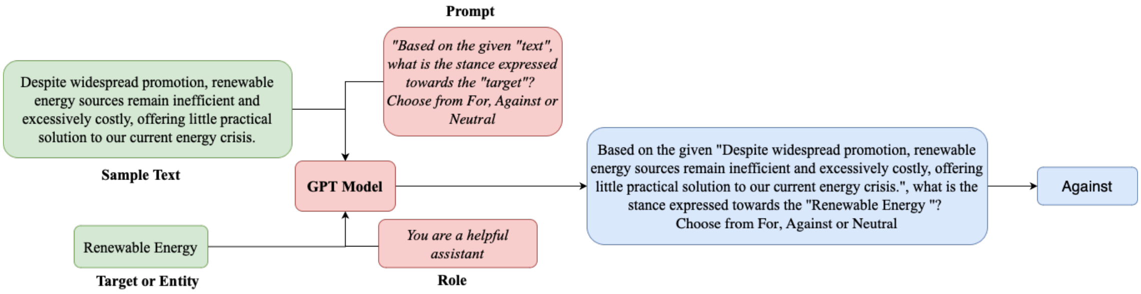

Prompt
"Based on the given "text", what is the stance expressed towards the "target"? Choose from For, Against or Neutral
Despite widespread promotion, renewable energy sources remain inefficient and excessively costly, offering little practical solution to our current energy crisis.
Sample Text
GPT Model
Based on the given "Despite widespread promotion, renewable energy sources remain inefficient and excessively costly, offering little practical solution to our current energy crisis.", what is the stance expressed towards the "Renewable Energy "? Choose from For, Against or Neutral
Against
Renewable Energy
Target or Entity
You are a helpful assistant
Role

Figure 11: Stance detection methodology

6.3.3 Stance Divergent Score (SDS)

We have developed a metric, Stance Divergent Score (SDS), as we want to quantitatively assess the divergence in stance orientation towards specific entities across different language versions of Wikipedia articles. The metric is predicated on the initial stance detection performed by the GPT model as discussed in Section 6.3.2.

Each stance label for each entity is compared between the languages under study. This comparison transforms the stance labels into a binary format where "1" represents stance mismatch and "0" represents stance match. Then we compute the average stance for each entity within individual articles. This average is aggregated across all articles to ascertain a mean stance divergent score for each language pair.

The Stance Divergent Score (SDS) for each language pair is computed by averaging these mean scores across all common entities for each article in the dataset. This metric provides a standardised measure of the stance alignment or discrepancy between two languages regarding the same entities, as depicted in Figure 12. The SDS quantifies the degree of stance agreement offering an understanding of cross-lingual stance dynamics.

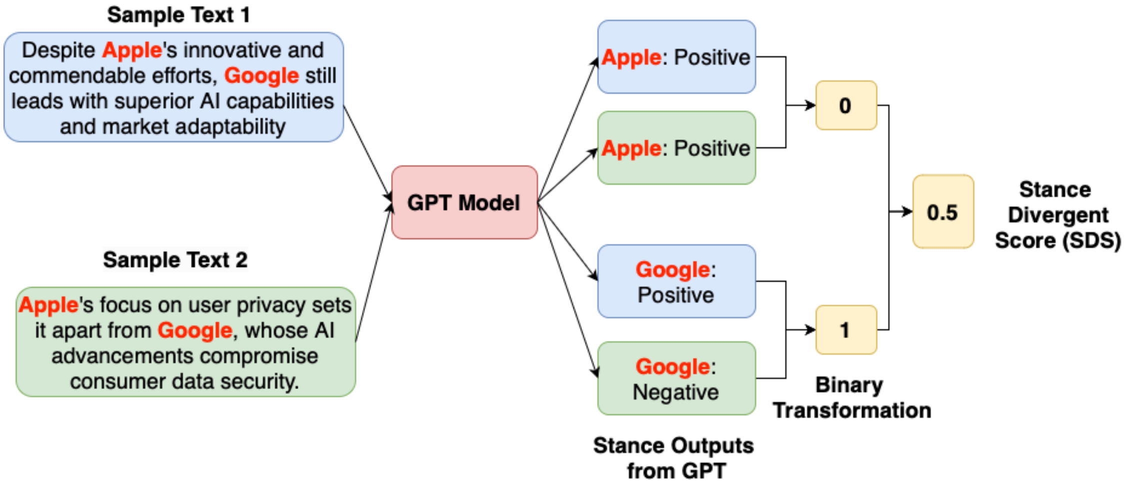

Sample Text 1
Despite Apple's innovative and commendable efforts, Google still leads with superior AI capabilities and market adaptability
Sample Text 2
Apple's focus on user privacy sets it apart from Google, whose AI advancements compromise consumer data security.
GPT Model
Apple: Positive
Apple: Positive
Google: Positive
Google: Negative
Stance Outputs from GPT
0
1
Binary Transformation
0.5
Stance Divergent Score (SDS)

Figure 12: Stance Divergent Score

6.4 Results and Inference

This section presents the outcomes derived from using stance detection methodologies, detailing the findings we have made through the approach that we have described in the previous sections.

6.4.1 Stance and Common Entities Variance across language pairs

Upon identifying the stances of common entities across various language pairs, we sought to analyse the overall distribution of stances and entities.

As one can see in Figure 13, the stance variation across language pairs highlights a pattern in the distribution of stances, with a prevalence of neutral stances compared to those for or against. Despite the predominance of neutral stances, some variations in stance distribution are still evident, indicating subtle differences in how topics are approached across language pairs. These variations, while not overwhelming, are crucial for understanding the depth and nature of neutrality in Wikipedia articles.

Language pairs involving English tend to have higher neutral stance counts, possibly indicating a more detached or objective treatment of topics typical of English Wikipedia's editorial style. In contrast, pairs involving non-Western languages like Hindi and Chinese might reflect different narrative styles or cultural emphases, affecting both the quantity and nature of entity discussion. We also find that an overall higher entity count in English-involving language pairs emphasises the role of English in multilingual NLP research and application.

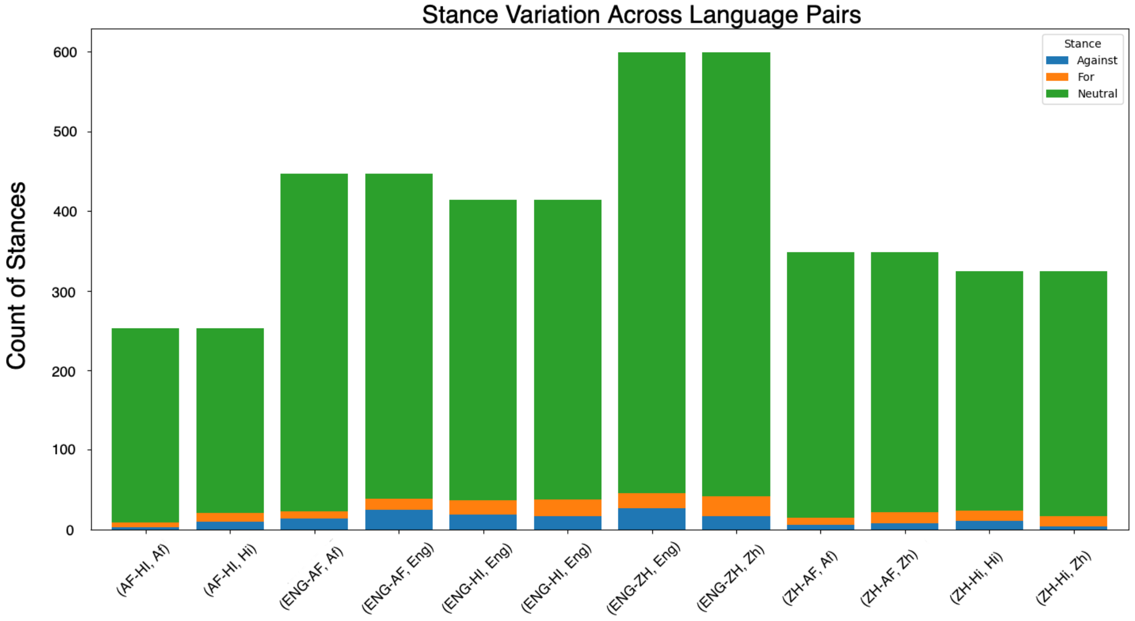


Figure 13: Stance Variation across Language Pairs

6.4.2 Stance Divergent Score (SDS)

The mean SDS for language pairs are shown in Table 10.

Language pairs involving Chinese exhibit a higher degree of stance divergence, reinforcing the findings presented in Chapter 5.

While language pairs involving English generally exhibit lower SDS values, a notable exception is the English-Hindi pair, which presents the highest Mean SDS among all the language pairs analysed, including those involving non-Western languages. This outcome is particularly surprising given that, as shown in Table 5, the English-Hindi pair displayed the lowest mean JSD. This indicates that although

Table 10: Mean Stance Divergent Score (SDS) for Different Topics (RTT = Round-trip translated)

Language Pair	Overall SDS	Controversial SDS	War SDS
Afrikaans-Chinese	0.0888	0.0802	0.1770
Afrikaans-Hindi	0.0843	0.0756	0.1510
Chinese-Hindi	0.1018	0.1010	0.0735
English-Afrikaans	0.0742	0.0708	0.0992
English-Chinese	0.0677	0.0658	0.0750
English-Hindi	0.1278	0.1339	0.0357
English-RTTEnglish	0.00002	0.0017	0.00004

there is a sentiment alignment between the articles in these two languages, the agreement on stance towards specific entities is significantly divergent. This divergence highlights the complexity of stance detection.

We also make the general observation that overall the divergence is small, that is less than 10% of entities have divergent stances, which is a positive finding for Wikipedia. This indicates that despite the presence of some stance divergence, the majority of entities across different language editions exhibit a consistent stance, suggesting a high level of coherence in Wikipedia's coverage.

The results from the round-trip translation are considerably lower compared to other language pairs, indicating that the observed sentiment differences are due to variations in the sentiment conveyed by the text itself, rather than being artifacts of the translation process.

Figure 14 displays the mean SDS of language pairs across controversial and war topics. It is important to note that the number of war articles analyzed is considerably lower than that of controversial topics. This discrepancy in sample size must be considered when interpreting the results, as it could influence the apparent stance divergence observed in the data.

In the analysis, language pairs involving war topics often show higher SDS values compared to those involving controversial topics. This suggests a greater divergence in stance for war-related articles, which may stem from the inherently sensitive and

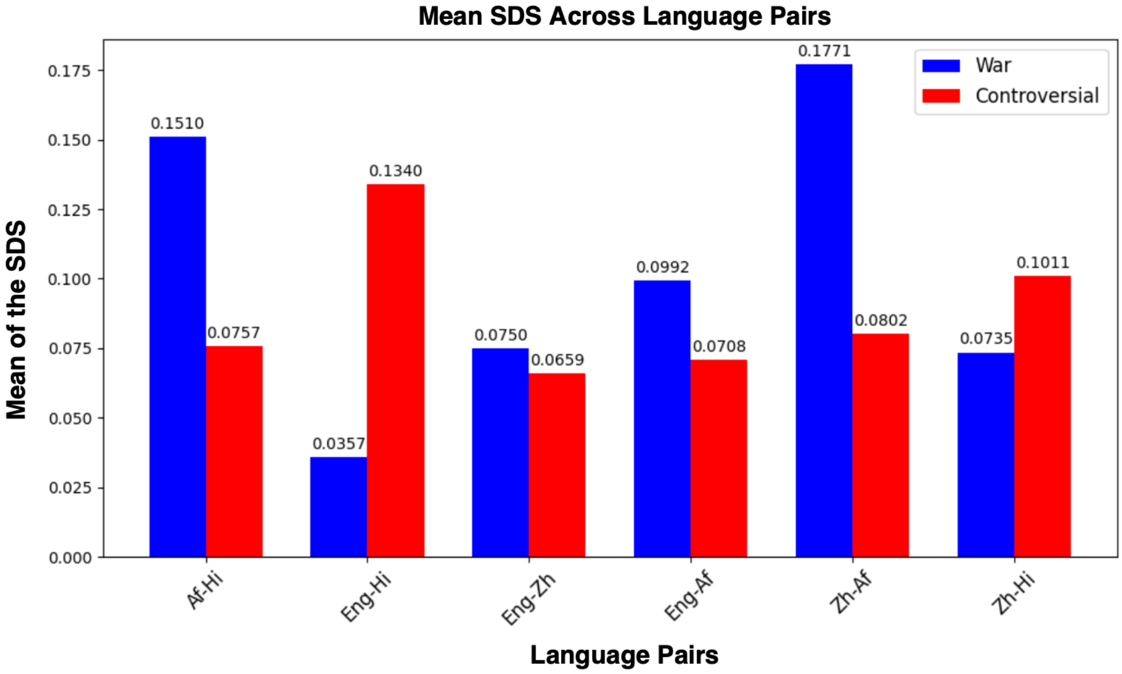


Figure 14: Mean SDS across language pairs and categories

polarising nature of war, potentially evoking stronger nationalistic or culturally specific sentiments that influence the portrayal of entities and events. For example, the Afrikaans-Chinese pair exhibits a marked increase in SDS for war topics compared to controversial topics, indicating significant differences in perspective or editorial stance that are likely influenced by distinct historical narratives or political contexts specific to each language community.

Conversely, controversial topics, while also subject to cultural and contextual influences, tend to show lower stance divergence as indicated by lower SDS values. This could suggest that the topics classified as controversial might still be subject to global norms or more universal views that lead to a more aligned stance across different languages.

6.5 Human Evaluation

We also decided to conduct a human evaluation of our stance detection methodology, focusing on assessing the efficacy of the GPT model and examining whether the SDS is a viable metric for quantifying stance divergences.

This evaluative step compares the automated stance detection results generated by the GPT model with human judgments to determine the model's accuracy and reliability in reflecting true sentiment and stance.

Additionally, by calculating SDS for human and model results, we provide a quantifiable measure that can potentially capture the complexities of stance variations as perceived by human interpreters. This combined approach seeks not only to validate the computational methods used but also to refine our understanding of how automated tools compare with human cognition in the task of stance detection.

6.5.1 Human Evaluation Setup

The human evaluation involved assigning each participant a language pair undisclosed to them to maintain objectivity. Participants were provided with an online form and tasked with reading two versions of articles, both translated into English to standardise the assessment process. We pre-identified and highlighted common entities in these articles based on our preliminary analysis as described in Section 6.3.1.

The study involved six participants, all of whom were students from diverse academic fields, none directly related to NLP. Each participant was required to assess a set of five pairs of articles. These pairs were selected to include eight or nine common entities, identified across different language pairs to obtain a comprehensive review. The evaluation task replicated the prompt given to the GPT model,

> *"Based on the given "[TEXT]", what is the stance expressed towards the "[TARGET]"? Choose from For, Against or Neutral."*

This ensures that both human and machine assessments are based on the same criteria.

Participants were instructed to complete the form for all identified entities across all provided articles. The selection of articles intentionally included a mix with varying stances as well as articles where stances were aligned. This strategy was designed to test the accuracy of our model's predictions under different conditions, allowing us to assess its efficacy in detecting stance variations and its reliability across a range of expressions.

6.5.2 Human Evaluation Results

Upon receiving the responses from the participants, we proceeded to calculate the SDS for each participant's responses concerning the articles. This calculation was mirrored by applying the approach as shown in Figure 12.

Entity-level Results

Upon analyzing the results at the entity level, it was observed that the model predicted the Sentiment Divergence Score (SDS) with an accuracy of **81.58%**, aligning closely with human judgment. This high level of accuracy demonstrates that our model is capable of effectively predicting stances in a manner that corresponds well with human assessments. Furthermore, a statistically significant Pearson correlation (Cohen et al., 2009) of **0.655** with a p-value of **0.000** was established between the model predictions and human results. This correlation further demonstrates the reliability of the model in mirroring human cognitive evaluations of stance within the text.

Upon closer examination between model predictions and human judgments at the entity level, a pattern emerged regarding the types of entities assessed. The model and human judgments aligned closely for entities representing nations and geographic locations. However, discrepancies were observed for entities representing human personalities as shown in Table 11.

These differences could stem from the participants' preconceived notions about certain personalities, influencing their interpretations of the texts. Additionally, the length of texts concerning personalities might lead survey participants to misinterpret the overall sentiment as positive or negative while the model predicted them as neutral. This observation suggests that while the model is adept at assessing neutral or fact-based content, such as locations and nations, it may not fully capture the sentiments associated with more subjective and character-driven content.

Table 11: Comparison of Stance Detection Outputs

Text A	Text B	Entity	Category	GPT Model Results		Human Results	
				Stance A	Stance B	Stance A	Stance B
”Freddie Mercury, formerly known as Farooq Bulesara, was an Indian-British singer and songwriter, best known as the lead singer and pianist of the rock band Queen [...]”	”Freddie Mercury was a British musician of Indian-Zoroastrian descent. He was known as the lead singer of the rock group Queen and developed into one of the most popular pop artists of all [...]”	Freddie Mercury	Human	Positive	Neutral	Neutral	Neutral
”Member countries Candidate countries Promised invitation No membership planned Attitude on accession unknown The North Atlantic Treaty Organization is a military [...]”	”The North Atlantic Treaty Organization is a military alliance established on 4 April 1949. Its headquarters is in Brussels. The organization has created a system of collective security, under [...]”	NATO	Organization	Neutral	Neutral	Neutral	Neutral
”Nero Claudius Caesar Augustus Germanicus was Roman emperor and the final emperor of the Julio-Claudian dynasty, reigning from AD 54 gain [...]”	”Nero Claudius Caesar Augustus Germanicus, born Lucius Domitius Ahenobarbus, took the name Nero Claudius [...]”	Nero	Human	Negative	Neutral	Neutral	Neutral

Language Pair Results

The mean SDS was then calculated for language pairs which are presented in Table 12. For consistency, the SDS was computed for the identical set of articles across each language pair that had been used for human evaluation.

Table 12: Human Evaluation Results

Language Pair	Model Results	Human Results
Afrikaans-Chinese	0.7	0.5
Afrikaans-Hindi	0.6	0.6
Chinese-Hindi	0.5	0.35
English-Afrikaans	0.6	0.6
English-Chinese	0.7	0.5
English-Hindi	0.63	0.43

From Table 12 we found that the Pearson correlation (Cohen et al., 2009), at approximately **0.366**, which demonstrates a moderate positive correlation, indicating a partial alignment between the model's output and human evaluations. However, the strength of this correlation is not robust, which implies that while the model successfully captures certain facets of human judgment, divergences remain. This divergence may also be attributed to the limited scope of our evaluation, involving only five sets of articles per language pair, yet it illustrates the extent of stance divergence.

It was also observed that when calculating SDS at the language pair level, the model mostly recorded higher SDS values compared to those derived from human judgments in most cases. This discrepancy suggests that the model may be more sensitive to detecting subtle differences in stances between the language pairs than human evaluators are. This heightened sensitivity could be attributed to the model's ability to process and analyse linguistic differences more systematically than humans, who might overlook or interpret these differences differently due to subjective biases or variations in individual perception.

6.6 Conclusion

The findings from our stance detection analysis across multiple language pairs using both automated models and human evaluations have provided insights into the complexities involved in accurately determining and comparing stances on Wikipedia. While the methodologies used in this study have shown effectiveness in certain contexts, the results also point to substantial room for improvement, particularly in aligning model outputs more closely with human judgment.

One of the key observations from our study is the distinction between sentiment and stance, which, although related, do not always correlate as one might expect. The analysis revealed instances where language pairs, such as English-Hindi, displayed low sentiment divergence but significant stance divergence. This phenomenon highlights the nuanced differences between sentiment, which may express general feelings or attitudes, and stance, which is specifically related to positions taken on particular entities or topics. Our results underscore that sentiment alignment does not necessarily guarantee stance agreement, suggesting that these aspects may be influenced by different factors, such as cultural nuances, contextual understanding, and the specific nature of the topics discussed.

Furthermore, the varying degrees of stance agreement, as quantified by the SDS, across different topics and language pairs point to the influence of cultural and contextual variables on how stances are formed and expressed. The moderate correlation between human evaluations and model predictions also indicates the challenges in automating stance detection to the level of human cognition. The differences observed serve as a reminder of the complexities inherent in processing and interpreting multilingual and multicultural content. It suggests that future improvements in stance detection models should focus on better understanding and integrating the cultural and contextual layers that significantly impact how stances are articulated and understood across different languages.

In conclusion, this study lays foundational work for advancing the field of multilingual stance detection by highlighting the critical differences between sentiment and stance and the impact of cultural contexts on these phenomena. Moving forward, enhancing the precision of stance detection tools will require a concerted effort to incorporate more sophisticated linguistic models, a broader and more balanced dataset across languages, and a deeper integration of cultural intelligence into computational models. This will not only improve the alignment with human judgment but also enhance our understanding of globally controversial and war-related topics that invite multiple perspectives and opinions from around the world.

Chapter 7

Conclusion

7.1 Overview

In this research, we set out to examine potential biases in Wikipedia articles across various language versions by applying sentiment analysis and stance detection methodologies. The initial phase involved content alignment using LASER embeddings to ensure that the articles from different language versions were contextually comparable for accurate sentiment analysis. This step was crucial for establishing a reliable foundation for further analysis.

Following the alignment, our analysis revealed sentiment divergences between the language pairs. These findings were quantified through several specifically designed metrics that accurately captured the extent and nature of sentiment differences across languages, providing a structured approach to understanding these variations.

The research then progressed to a more detailed examination of the stance within the texts. We identified and compared stances towards common entities across the different language versions using large language models like GPT. This process involved stance detection analysis followed by quantification of stance divergences. To ensure the validity of our approach, we supplemented our analysis with human evaluations, which helped confirm the consistency and reliability of the detected patterns.

Our investigation into the biases present in Wikipedia challenged the platform' s claim to uphold a Neutral Point of View (NPOV). Despite Wikipedia' s guidelines to maintain objectivity, our findings found discrepancies in how different language versions handle the portrayal of similar topics. By quantifying these discrepancies, this research contributes to insights into the bias inherent in global information sources. This demonstrates the need for ongoing scrutiny and adjustment to increase the accuracy and fairness of content presented on widely used digital platforms like Wikipedia.

Overall, the divergence between language variants appears relatively low, with approximately 20% difference in terms of sentiment and less than 10% difference in terms of stance. This indicates that, on average, the majority of entities across different language editions exhibit consistent sentiment and stance. Such findings are encouraging for Wikipedia, as they suggest a relatively high level of coherence and neutrality in its content. However, the identified discrepancies, although not predominant, highlight areas where Wikipedia’ s NPOV can be improved. This duality emphasises the importance of continual monitoring and refinement to ensure Wikipedia remains a reliable and unbiased information source.

7.2 Limitations of this study

One significant limitation in our study is the exclusion of unaligned sentences, which could substantially impact the analysis. The disparity in content volume between Wikipedia articles—where English versions often contain more detail than their counterparts in other languages (Rajcic, 2017)—highlights the need to identify sentences that are mismatched or absent in translations to achieve a balanced comparative analysis. Additionally, handling compound sentences presents another challenge. For instance, an English sentence might be split into multiple sentences in another language, leading to fragmented content alignment. This fragmentation can distort sentiment analysis by excluding significant portions of the text that contribute to the overall sentiment. Our methodology also involved native speakers evaluating alignment from English to other languages, but we could not assess content alignment quality for translations between non-English language pairs, relying instead on a cosine similarity threshold of 0.75.

In sentiment analysis, the reliance on translation for low-resource languages like Afrikaans and Hindi can introduce inaccuracies due to the loss of linguistic and cultural nuances. Additionally, using binary metrics like DSP in sentiment analysis may overlook subtleties in sentiment expression although they simplify the computational process.

Another limitation of our current study is the exclusion of non-overlapping content. Specifically, unaligned sentences in sentiment analysis and non-common entities in stance detection. This limitation may result in an incomplete understanding of the sentiment and stance across different language pairs, as it ignores substantial portions of the text that do not have direct equivalents in the translations.

Overall, while the methodologies used in this study have shown effectiveness in certain contexts, there is substantial room for improvement, particularly in aligning model outputs more closely with human judgment. The nuanced differences between sentiment and stance, as well as the impact of cultural and contextual variables, highlight the inherent complexities of multilingual sentiment and stance detection.

7.3 Future Works

In future work, several avenues could be elaborated on while using the findings of this study. Firstly, it would be beneficial to explore additional language pairs, particularly those that represent non-Western perspectives. Investigating languages from diverse cultural backgrounds could provide deeper insights into the global biases present in Wikipedia articles and increase our understanding of cross-cultural information representation.

Secondly, expanding the scope of analysis beyond the summary sections of Wikipedia articles to include entire pages could offer a more comprehensive view of content biases. This broader analysis would involve examining various elements such as images, info boxes, and different sections of each article, providing a holistic view of how information is structured and presented across different languages.

Thirdly, a longitudinal study of Wikipedia articles, analysing how content evolves across different editing versions, could reveal trends and shifts in sentiment and stance over time. Such an investigation would shed light on the dynamics of content modification and the factors influencing changes in portrayal. This could in turn provide insights to understand editorial behaviours and community influences on Wikipedia.

To address the limitation of non-overlapping content, future research could incorporate more sophisticated alignment techniques that account for unaligned sentences. One approach is the use of advanced machine translation models, such as neural machine translation (NMT) systems, which can better handle sentence fragmentation and alignment discrepancies (Bahdanau et al., 2014). By improving the alignment accuracy, we can ensure that more content is included in the analysis, providing a fuller picture of the sentiment distribution. For stance detection, incorporating non-common entities can be achieved by broadening the scope of entity recognition and alignment. Techniques such as entity linking and cross-lingual entity extraction (Upadhyay et al., 2016) can help identify and align entities that are not explicitly mentioned in all language versions. This approach would involve using pre-trained language models, like BERT or mBERT, to improve entity recognition across languages. By expanding the range of recognised entities, we can capture more data points for stance analysis, thus improving the robustness of our findings. Furthermore, leveraging unsupervised learning methods to identify and analyse unaligned sentences and non-common entities can provide additional insights. Techniques such as topic modelling (Blei et al., 2003) can be applied to detect underlying themes in unaligned content, offering a deeper understanding of the divergences in information representation. These methods can be complemented by human-in-the-loop approaches, where native speakers assist in refining the alignment and interpretation of complex cases (Snow et al., 2008).

Finally, applying the methodologies developed in this research to other platforms that claim neutrality, such as other encyclopedias or informational websites, would test the applicability of our findings and methodologies in different contexts. This extension could validate the robustness of the analytical techniques used and potentially reveal systemic biases in other widely used informational resources.

Bibliography

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 165–174, 2015.

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288. URL `https://aclanthology.org/Q19-1038`.

Jiao Sun and Nanyun Peng. Men are elected, women are married: Events gender bias on wikipedia. *arXiv preprint arXiv:2106.01601*, 2021.

Yochai Benkler. *Introduction: A Moment of Opportunity and Challenge*, pages 1–28. Yale University Press, 2006. ISBN 9780300110562. URL `http://www.jstor.org/stable/j.ctt1njknw.4`.

Wikipedia. Wikipedia:Size of Wikipedia — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Wikipedia%3ASize%20of%20Wikipedia&oldid=1216603163`, 2024a. [Online; accessed 08-April-2024].

Jim Giles. Special report internet encyclopaedias go head to head. *nature*, 438(15): 900–901, 2005.

Roy Rosenzweig. Can history be open source? wikipedia and the future of the past. *The Journal of American History*, 93(1):117–146, 2006.

Wikipedia. Wikipedia:Neutral point of view — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Wikipedia%3ANeutral%20point%20of%20view&oldid=1210967506`, 2024b. [Online; accessed 08-April-2024].

Joseph Michael Reagle. *Good faith collaboration: The culture of Wikipedia*. MIT press, 2010.

Aaron Shaw and Eszter Hargittai. The pipeline of online participation inequalities: The case of wikipedia editing. *Journal of communication*, 68(1):143–168, 2018.

Benjamin Cabrera, Björn Ross, Marielle Dado, and Maritta Heisel. The gender gap in wikipedia talk pages. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

Ewa S Callahan and Susan C Herring. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915, 2011.

Brent Hecht and Darren Gergle. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 291–300, 2010.

Nemanja Rajcic. Comparison of Wikipedia articles in different languages. page 98 pages, 2017. doi: 10.34726/HSS.2017.35937. URL `https://repositum.tuwien.at/handle/20.500.12708/5121`. Artwork Size: 98 pages Medium: application/pdf Publisher: TU Wien.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765*, 2018.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23, 2017.

Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37, 2020.

Judy Hanwen Shen, Lauren Fratamico, Iyad Rahwan, and Alexander M Rush. Darling or babygirl? investigating stylistic bias in sentiment analysis. *Proc. of FATML*, 2018.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 1650–1659, 2013.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Jackie Koerner. Wikipedia Has a Bias Problem. *Wikipedia @ 20*, June 2019. https://wikipedia20.mitpress.mit.edu/pub/u5vsaip5.

Brian Martin. Persistent bias on wikipedia: Methods and responses. *Social Science Computer Review*, 36(3):379–388, 2018.

Shane Greenstein and Feng Zhu. Is wikipedia biased? *American Economic Review*, 102(3):343–348, 2012.

Christoph Hube and Besnik Fetahu. Detecting biased statements in wikipedia. In *Companion proceedings of the the web conference 2018*, pages 1779–1786, 2018.

Benjamin Collier and Julia Bear. Conflict, criticism, or confidence: an empirical examination of the gender gap in wikipedia contributions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 383–392, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310864. doi: 10.1145/2145204.2145265. URL `https://doi.org/10.1145/2145204.2145265`.

Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer.

Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ data science*, 5:1–24, 2016.

Amber Young, Ari D Wigdor, and Gerald Kane. It’ s not what you think: Gender bias in information about fortune 1000 ceos on wikipedia. 2016.

Shaila M Miranda, Amber Young, and Emre Yetgin. Are social media emancipatory or hegemonic? societal effects of mass media digitization in the case of the sopa discourse. *MIS quarterly*, 40(2):303–330, 2016.

Anjalie Field, Chan Young Park, Kevin Z Lin, and Yulia Tsvetkov. Controlled analyses of social biases in wikipedia bios. In *Proceedings of the ACM Web Conference 2022*, pages 2624–2635, 2022.

Marit Hinnosaar. Gender inequality in new media: Evidence from wikipedia. *Journal of economic behavior & organization*, 163:262–276, 2019.

Adam R Brown. Wikipedia as a data source for political scientists: Accuracy and completeness of coverage. *PS: Political Science & Politics*, 44(2):339–343, 2011.

Alexander Halavais and Derek Lackaff. An analysis of topical coverage of wikipedia. *Journal of computer-mediated communication*, 13(2):429–440, 2008.

Desislava Aleksandrova, François Lareau, and Pierre André Ménard. Multilingual sentence-level bias detection in Wikipedia. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 42–51, Varna, Bulgaria, September 2019. INCOMA Ltd. doi: 10.26615/978-954-452-056-4_006. URL `https://aclanthology.org/R19-1006`.

Marc Miquel-Ribé and David Laniado. Wikipedia culture gap: quantifying content imbalances across 40 language editions. *Frontiers in physics*, 6:54, 2018.

Anna Kerkhof and Johannes Münster. Detecting coverage bias in user-generated content. *Journal of Media Economics*, 32(3-4):99–130, 2019.

Pnina Fichman and Noriko Hara. Introduction in global wikipedia: International and cross-cultural issues in online collaboration. Rowman & Littlefield, 2014.

Josef Kolbitsch and Hermann A Maurer. The transformation of the web: How emerging communities shape the information we consume. *J. Univers. Comput. Sci.*, 12(2):187–213, 2006.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463, 2015.

Guadalupe Alvarez, Aileen Oeberst, Ulrike Cress, and Laura Ferrari. Linguistic evidence of in-group bias in english and spanish wikipedia articles about international conflicts. *Discourse, Context & Media*, 35:100391, 2020.

Frederick Jelinek. *Statistical methods for speech recognition*. MIT press, 1998.

William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175, page 14. Las Vegas, NV, 1994.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Teguh Ikhlas Ramadhan, Nur Ghaniaviyanto Ramadhan, and Agus Supriatman. Implementation of neural machine translation for english-sundanese language using long short term memory (lstm). *Building of Informatics, Technology and Science (BITS)*, 4(3):1438–1446, 2022.

GSN Murthy, Shanmukha Rao Allu, Bhargavi Andhavarapu, Mounika Bagadi, and Mounika Belusonti. Text based sentiment analysis using lstm. *Int. J. Eng. Res. Tech. Res*, 9(05), 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*, 2020.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. Text generation from knowledge graphs with graph transformers. *arXiv preprint arXiv:1904.02342*, 2019.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and

Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*, 2020.

Steven Moore, Richard Tong, Anjali Singh, Zitao Liu, Xiangen Hu, Yu Lu, Joleen Liang, Chen Cao, Hassan Khosravi, Paul Denny, et al. Empowering education with llms-the next-gen interface and content generation. In *International Conference on Artificial Intelligence in Education*, pages 32–37. Springer, 2023.

Samuel Kernan Freire, Chaofan Wang, and Evangelos Niforatos. Chatbots in knowledge-intensive contexts: Comparing intent and llm-based systems. *arXiv preprint arXiv:2402.04955*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Meena Rambocas and João Gama. Marketing research: The role of sentiment analysis. Technical report, Universidade do Porto, Faculdade de Economia do Porto, 2013.

Dietmar Gräbner, Markus Zanker, Günther Fliedl, and Matthias Fuchs. Classification of customer reviews based on sentiment analysis. In *Information and communication technologies in tourism 2012*, pages 460–470. Springer, 2012.

Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60:617–663, 2019.

Soonh Taj, Baby Bakhtawer Shaikh, and Areej Fatemah Meghji. Sentiment analysis of news articles: a lexicon based approach. In *2019 2nd international conference on computing, mathematics and engineering technologies (iCoMET)*, pages 1–5. IEEE, 2019.

Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair. Supervised sentiment analysis of parliamentary speeches and news reports. In *67th Annual Conference of the International Communication Association (ICA), Panel on Automatic Sentiment Analysis*, 2017.

Chanakya Sharma, Samuel Whittle, Pari D Haghighi, Frada Burstein, and Helen Keen. Sentiment analysis of social media posts on pharmacotherapy: A scoping review. *Pharmacology Research & Perspectives*, 8(5):e00640, 2020.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd nordic conference on computational linguistics*, pages 187–196, 2019.

Markus Leippold. Sentiment spin: Attacking financial sentiment with gpt-3. *Finance Research Letters*, 55:103957, 2023.

Alim Al Ayub Ahmed, Sugandha Agarwal, IMade Gede Ariestova Kurniawan, Samuel PD Anantadjaya, and Chitra Krishnan. Business boosting through sentiment analysis using artificial intelligence approach. *International Journal of System Assurance Engineering and Management*, 13(Suppl 1):699–709, 2022.

Prakash P Rokade and Kumari D Aruna. Business intelligence analytics using sentiment analysis-a survey. *International Journal of Electrical and Computer Engineering*, 9(1):613, 2019.

Harish Dutt Sharma and Parul Goyal. An analysis of sentiment: Methods, applications, and challenges. *Engineering Proceedings*, 59(1):68, 2023.

Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, 13(7):4550, 2023.

Sonal Gupta. Finding bias in political news and blog websites, 2009.

Kenneth C Enevoldsen and Lasse Hansen. Analysing political biases in danish newspapers using sentiment analysis. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, 2(2):87–98, 2017.

Wael F Al-Sarraj and Heba M Lubbad. Bias detection of palestinian/israeli conflict in western media: A sentiment analysis experimental study. In *2018 International Conference on Promising Electronic Technologies (ICPET)*, pages 98–103. IEEE, 2018.

Sachin Rawat and G Vadivu. Media bias detection using sentimental analysis and clustering algorithms. In *Proceedings of international conference on deep learning, computing and intelligence: ICDCI 2021*, pages 485–494. Springer, 2022.

Anastasia Smirnova, Helena Laranetto, and Nicholas Kolenda. Ideology through sentiment analysis: A changing perspective on russia and islam in nyt. *Discourse & Communication*, 11(3):296–313, 2017.

Yanchuan Sim, Brice DL Acree, Justin H Gross, and Noah A Smith. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 91–101, 2013.

Hitesh Nankani, Hritwik Dutta, Harsh Shrivastava, PVNS Rama Krishna, Debanjan Mahata, and Rajiv Ratn Shah. Multilingual sentiment analysis. *Deep learning-based approaches for sentiment analysis*, pages 193–236, 2020.

A Aziz Altowayan and Lixin Tao. Word embeddings for arabic sentiment analysis. In *2016 IEEE international conference on big data (big data)*, pages 3820–3825. IEEE, 2016.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1077. URL `https://aclanthology.org/D19-1077`.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study, 2020.

Guanrong Li, Ziwei Wang, Minzhu Zhao, Yunya Song, and Liang Lan. Sentiment analysis of political posts on hong kong local forums using fine-tuned mbert. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 6763–6765, 2022. doi: 10.1109/BigData55660.2022.10020704.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL `https://aclanthology.org/P19-1493`.

Edward W Chew, William D Weisman, Jingying Huang, and Seth Frey. Machine translation for accessible multi-language text analysis. *arXiv preprint arXiv:2301.08416*, 2023.

Kerstin Denecke. Using sentiwordnet for multilingual sentiment analysis. In *2008 IEEE 24th international conference on data engineering workshop*, pages 507–512. IEEE, 2008.

Sven-Oliver Proksch, Will Lowe, Jens Wäckerle, and Stuart Soroka. Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, 44(1):97–131, 2019.

Anupam Baliyan, Akshit Batra, and Sunil Pratap Singh. Multilingual sentiment analysis using rnn-lstm and neural machine translation. In *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 710–713. IEEE, 2021.

Alexandra Balahur and Marco Turchi. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 52–60, 2012.

Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre, editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `https://aclanthology.org/P10-1114`.

James Hong and Michael Fang. Sentiment analysis with deeply learned distributed representations of variable length texts. *Stanford University Report*, pages 1–9, 2015.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://aclanthology.org/D13-1170`.

Andrew J Reagan, Christopher M Danforth, Brian Tivnan, Jake Ryland Williams, and Peter Sheridan Dodds. Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. *EPJ Data Science*, 6:1–21, 2017.

Dimitrios Tsirmpas, Ioannis Gkionis, and Ioannis Mademlis. Neural natural language processing for long texts: A survey of the state-of-the-art. *arXiv preprint arXiv:2305.16259*, 2023.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230, 2014.

Noemie Elhadad and Regina Barzilay. Sentence alignment for monolingual comparable corpora. 2003.

Daniel Gildea. Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 80–87, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1075096.1075107. URL `https://aclanthology.org/P03-1011`.

Sora Kadotani and Yuki Arase. Monolingual phrase alignment as parse forest mapping. In Alexis Palmer and Jose Camacho-collados, editors, *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 449–455, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.starsem-1.39. URL `https://aclanthology.org/2023.starsem-1.39`.

William A. Gale, Kenneth Ward Church, et al. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102, 1994.

Baosong Yang, Zhaopeng Tu, Derek F Wong, Fandong Meng, Lidia S Chao, and Tong Zhang. Modeling localness for self-attention networks. *arXiv preprint arXiv:1810.10182*, 2018.

Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi. Self-attention with structural position representations. *arXiv preprint arXiv:1909.00383*, 2019.

Yuki Arase and Jun’ ichi Tsujii. Compositional phrase alignment and beyond. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1611–1623, 2020.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019.

Masaaki Nagata, Chousa Katsuki, and Masaaki Nishino. A supervised word alignment method based on cross-language span prediction using multilingual bert. *arXiv preprint arXiv:2004.14516*, 2020.

Katsuki Chousa, Masaaki Nagata, and Masaaki Nishino. Spanalign: Sentence alignment method based on cross-language span prediction and ilp. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4750–4761, 2020.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Text alignment is an efficient unified model for massive nlp tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

Dean Pomerleau and Delip Rao. Fake news challenge stage 1 (fnc-i): Stance detection. *URL www. fakenewschallenge. org*, 2017.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. A richly annotated corpus for different tasks in automated fact-checking. In Mohit Bansal and Aline Villavicencio, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1046. URL `https://aclanthology.org/K19-1046`.

Yaakov HaCohen-Kerner, Ziv Ido, and Ronen Ya' akobov. Stance classification of tweets using skip char ngrams. In *Machine Learning and Knowledge Discovery* *in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part III 10*, pages 266–278. Springer, 2017.

Anirban Sen, Manjira Sinha, Sandya Mannarswamy, and Shourya Roy. Stance classification of multi-perspective consumer health information. In *Proceedings of the ACM India joint international conference on data science and management of data*, pages 273–281, 2018.

Daniel Röchert, German Neubaum, and Stefan Stieglitz. Identifying political sentiments on youtube: a systematic comparison regarding the accuracy of recurrent neural network and machine learning models. In *Disinformation in Open Online Media: Second Multidisciplinary International Symposium, MISDOOM 2020, Leiden, The Netherlands, October 26–27, 2020, Proceedings 2*, pages 107–121. Springer, 2020.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. Stance classification with target-specific neural attention networks. In *26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 3988–3994. International Joint Conferences on Artificial Intelligence, 2017.

Muhammad Umer, Zainab Imtiaz, Saleem Ullah, Arif Mehmood, Gyu Sang Choi, and Byung-Won On. Fake news stance detection using deep learning architecture (cnn-lstm). *IEEE Access*, 8:156695–156706, 2020.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. Stance detection with hierarchical attention network. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1203.

Qingying Sun, Xuefeng Xi, Jiajun Sun, Zhongqing Wang, and Huiyan Xu. Stance detection with a multi-target adversarial attention network. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–21, 2022.

Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, 8:673, 2020.

Rini Wongso, Derwin Suhartono, et al. A literature review of question answering system using named entity recognition. In *2016 3rd international conference on information technology, computer, and electrical engineering (ICITACEE)*, pages 274–277. IEEE, 2016.

Giorgio De Magistris, Samuele Russo, Paolo Roma, Janusz T Starczewski, and Christian Napoli. An explainable fake news detector based on named entity recognition and stance classification applied to covid-19. *Information*, 13(3):137, 2022.

Douglas E Appelt, Jerry R Hobbs, John Bear, David Israel, and Mabry Tyson. Fastus: A finite-state processor for information extraction from real-world text. In *IJCAI*, volume 93, pages 1172–1178, 1993.

Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. 2003.

Sudha Morwal, Nusrat Jahan, and Deepti Chopra. Named entity recognition using hidden markov model (hmm). *International Journal on Natural Language Computing (IJNLC) Vol*, 1, 2012.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for NER. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 151–164, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1015. URL https://aclanthology.org/P19-1015.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

Anna Śniegula, Aneta Poniszewska-Marańda, and Łukasz Chomątek. Study of named entity recognition methods in biomedical field. *Procedia Computer Science*, 160:260–265, 2019.

David Milne and Ian H Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518, 2008.

Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In Jason Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/D07-1074.

Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1035. URL https://aclanthology.org/N19-1035.

Zihao He, Negar Mokhberian, and Kristina Lerman. Infusing knowledge from wikipedia to enhance stance detection, 2022.

Iain J. Cruickshank and Lynnette Hui Xian Ng. Prompting and fine-tuning open-sourced large language models for stance classification, 2024.

Kornraphop Kawintiranon and Lisa Singh. Knowledge enhanced masked language model for stance detection. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.376. URL `https://aclanthology.org/2021.naacl-main.376`.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz*, pages 1–13, 2021.

Oanh Tran, Anh Cong Phung, and Bach Xuan Ngo. Using convolution neural network with bert for stance detection in vietnamese. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7220–7225, 2022.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Yangyang Li, and Liwen Jing. Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*, 2023.

Michael Burnham. Stance detection: A practical guide to classifying political beliefs in text, 2024.

Mark Mets, Andres Karjus, Indrek Ibrus, and Maximilian Schich. Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media. *Plos one*, 19(4):e0302380, 2024.

Marco Dozza, Jonas Bärgman, and John D. Lee. Chunking: A procedure to improve naturalistic data analysis. *Accident Analysis & Prevention*, 58:309–317, 2013. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2012.03.020. URL `https://www.sciencedirect.com/science/article/pii/S0001457512001091`.

Holger Schwenk and Matthijs Douze. Learning joint multilingual sentence representations with neural machine translation. In Phil Blunsom, Antoine Bordes, Kyunghyun Cho, Shay Cohen, Chris Dyer, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, and Scott Yih, editors, *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2619. URL `https://aclanthology.org/W17-2619`.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. Extending multilingual BERT to low-resource languages. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.240. URL `https://aclanthology.org/2020.findings-emnlp.240`.

Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

Anoop Kunchukuttan. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`, 2020.

Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, and Abdelmajid Ben Hamadou.

Computing semantic relatedness using wikipedia features. *Knowledge-Based Systems*, 50:260–278, 2013. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2013.06.015. URL `https://www.sciencedirect.com/science/article/pii/S0950705113001913`.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Michael Buckland and Fredric Gey. The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19, 1994.

Cornelis Joost Van Rijsbergen. Foundation of evaluation. *Journal of documentation*, 30(4):365–373, 1974.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

Alberto Poncelas, Pintu Lohar, Andy Way, and James Hadley. The impact of indirect machine translation on sentiment classification. *arXiv preprint arXiv:2008.11257*, 2020.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130, 2016.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics:*

EMNLP 2020, pages 1644–1650, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL `https://aclanthology.org/2020.findings-emnlp.148`.

Shuyi Ji, Zizhao Zhang, Shihui Ying, Liejun Wang, Xibin Zhao, and Yue Gao. Kullback–leibler divergence metric learning. *IEEE transactions on cybernetics*, 52(4):2047–2058, 2020.

Eitan Adam Pechenick, Christopher M Danforth, and Peter Sheridan Dodds. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041, 2015.

Lin Li, Tiong-Thye Goh, and Dawei Jin. How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis. *Neural Computing and Applications*, 32:4387–4415, 2020.

Jiwen Chen. Sentiment analysis behind text with different length and formality. *CS230: Deep Learning.(Fall 2021). Retrieved*, 26, 2023.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

Xiaofei Lu. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496, 2010.

Philip M McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392, 2010.

Abdul Munem Nerabie, Manar AlKhatib, Sujith Samuel Mathew, May El Barachi, and Farhad Oroumchian. The impact of arabic part of speech tagging on sentiment analysis: A new corpus and deep learning approach. *Procedia Computer Science*, 184:148–155, 2021.

Pansy Nandwani and Rupali Verma. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81, 2021.

Bilal Ghanem, Paolo Rosso, and Francisco Rangel. Stance detection in fake news a combined feature representation. In *Proceedings of the first workshop on fact extraction and VERification (FEVER)*, pages 66–71, 2018.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972*, 2017.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL https://aclanthology.org/W03-0419.

Razvan Bunescu and Marius Paşca. Using encyclopedic knowledge for named entity disambiguation. In Diana McCarthy and Shuly Wintner, editors, *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://aclanthology.org/E06-1002.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425*, 2016.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Rion Snow, Brendan O' connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263, 2008.