Cross-Lingual Sentiment Analysis: Comparative Study of Opinion Expression Across Different Languages

Dr. Yusra jadallah abed Khasawneh¹, A. J. Khasawneh², Dr. Mohamad Ahmad Saleem Khasawneh³

Abstract

Understanding the interconnectedness of the Jordanian language, culture, and emotion is the purpose of this research. To accomplish this objective, the study employs cross-lingual sentiment analysis, a method that permits a comprehensive examination of the subject. By employing this analysis method, the study aims to offer a deeper understanding of the intricate relationship between culture, language, and emotion in Jordan. Comparing the manners in which speakers of Arabic and English express emotion in their respective languages was the primary objective of this study. The findings of this research shed light on the intricate relationship between the selection of a language and one's cultural identity, illustrating how this dynamic substantially influences the emotional landscape. This inquiry showcases the substantial impact of linguistic relativity and the inherent emotional intensity inscribed in language through the disclosure of a substantial increase in sentiment ratings for posts written in the English language. The results of this research further demonstrate the significance of the influence of linguistic imperialism on the social classification of feelings. The findings above offer significant insights into the complex dynamics between language, culture, and emotion within a heterogeneous and multicultural society.

Keywords: Cross-lingual sentiment analysis, Language and culture, Linguistic relativity, Sentiment expression, Linguistic imperialism, Multilingual society.

Introduction

The process of using computational methods to identify and make sense of the underlying emotions and explicit opinions expressed in textual data is known as sentiment analysis or opinion mining. The proliferation of online resources is largely responsible for the dramatic increase in this technology's popularity over the last decade. This technique has many potential uses in many different fields. For instance, businesses may utilize sentiment analysis to gauge customer satisfaction, get information for marketing strategies, and better their offerings. Both political analysts and customer service teams utilize sentiment analysis to monitor and respond to comments from customers (Pang & Lee, 2008).

Sentiment analysis's modern-day value stems from its ability to glean actionable insights from the mountains of textual data being generated every day. Businesses, governments,
and universities all agree that data analysis is crucial because it improves our ability to see trends, make educated decisions, and address the needs and concerns of many different groups of people. Despite its usefulness in providing insights into textual content within a single language situation, the study of sentiment becomes considerably more complicated when expanded across a large number of languages.

As a sub-discipline of sentiment analysis, cross-lingual sentiment analysis compares and contrasts the tone of writings in different languages. In today's globalized culture, when communication and information are freely shared across language borders, cross-lingual sentiment analysis has become more common. This phenomenon affords researchers the chance to probe the neural underpinnings of how ideas are communicated and interpreted across many language and cultural settings. Businesses and organizations must think about expanding their operations internationally, which requires a deep understanding of different customer demographics and market trends (Gamon & Aue, 2005).

Tools and methodologies for sentiment analysis have come a long way, particularly when applied to the English language. The endeavor, however, becomes more difficult when applied to languages with grammatical, cultural, and lexical distinctions. Significant progress has been made in the development of cross-lingual sentiment analysis tools and models within the field of natural language processing (NLP). The use of such models and tools is essential for improving the reliability and cultural appropriateness of conclusions drawn. However, there hasn't been a lot of research done in the area of cross-lingual sentiment analysis within specific multilingual contexts since there aren't many publications that have addressed the difficulties in great detail.

The cultural and linguistic diversity of the Kingdom of Jordan provides a fascinating and unique setting for the continuing research project. Although Arabic is the official language of the Kingdom of Jordan, other languages such as, English, Tagalog, and others have made significant contributions to the country's linguistic landscape thanks to the large expatriate population. The diversity described is further amplified by the quick growth of digital communication technologies that promote the usage of multiple languages.

**Objective of the Study**

The interpretation of emotional expression within a multilingual society, such as Jordan, is an important and interesting task. Our study of sentiment analysis in this context not only contributes to the academic understanding of cross-lingual sentiment analysis but also has important policy and business implications for addressing the complexities of linguistic and cultural diversity. Kawahara et al. (2012) found that their study's major emphasis was on examining five important research goals. Some of the things we hope to accomplish are looking into how different languages express emotions in the Jordanian context, contrasting how people feel in different languages, and figuring out how cultural and linguistic factors affect sentiment analysis results. By achieving these aims, we want to shed light on the challenges and opportunities presented by cross-lingual sentiment analysis. This will help develop more robust and culturally sensitive sentiment analysis tools and procedures for use in multilingual settings like Jordan.

**Literature Review and Previous Studies**

As a key component of natural language processing (NLP), sentiment analysis (also known as the computer assessment of sentiment, opinions, and emotions communicated in textual data) has evolved in recent years. Business, politics, marketing, and the social sciences are just a few of the fields that have taken notice of this phenomenon due to its versatility and practical applications (Pang & Lee, 2008). The exploration of sentiment
expression within a bilingual context has emerged as a prominent area of study in light of the ongoing process of globalization. This comprehensive literature review delves into the intriguing evolution of sentiment analysis, tracing its historical trajectory. Furthermore, it explores the remarkable emergence and development of cross-lingual sentiment analysis, shedding light on its significance in the field. Additionally, this review examines the noteworthy contributions made in the realm of sentiment analysis within the context of Jordan, offering valuable insights into the advancements achieved in this domain.

Sentiment analysis, a field also referred to as opinion mining, entails the automated classification and interpretation of textual material to unveil the underlying emotional disposition, perspectives, and attitudes conveyed within it. Sentiment analysis plays a pivotal role in facilitating informed decision-making for businesses and other entities. This analytical process revolves around the proficiency to transform unstructured textual input into measurable insights. There has been a great deal of growth in the area of sentiment analysis, leading to the development of several approaches and tools for tasks including polarity classification (positive, negative, neutral), emotion detection, and aspect-based sentiment analysis (Liu, 2012).

Pang and Lee’s (2008) seminal work highlights the significance of sentiment analysis, particularly in the context of its application to a variety of fields including product assessments, market analysis, and sentiment summarization. The results of the research highlight the importance of sentiment analysis in gaining insight into customer feedback and simplifying the procedure of mining insights from massive amounts of textual data.

Driven by the global scope of communication and the sharing of information, cross-lingual sentiment analysis has emerged, broadening sentiment analysis to incorporate multiple languages. A more all-encompassing approach is required to understand emotion across linguistic boundaries due to the prevalence of several languages in society and the global nature of commerce. Gamon and Aue (2005) argue that cross-lingual sentiment analysis may shed new light on the similarities and differences in how emotions are communicated and interpreted across cultural and linguistic boundaries.

Due to linguistic variances in syntax and lexicon, cross-lingual sentiment analysis is a complex field. Gamon and Aue’s (2005) research provides a thorough introduction to the challenges and possible benefits of cross-lingual sentiment analysis. They call attention to the need for new models and approaches that can effectively account for linguistic diversity. Based on these findings, it is clear that there is a pressing need to develop sentiment analysis methods that can be applied to texts written in a variety of languages to meet the needs of a wide range of worldwide audiences.

Although sentiment analysis has received a lot of attention recently, there is still a dearth of studies that focus on cross-lingual sentiment analysis in multilingual contexts like Jordan. Because of its diverse linguistic population and complex cultural dynamics, the Kingdom of Jordan provides a unique academic context. Since Arabic is the de facto official language and expat communities come from a wide variety of linguistic backgrounds, it is especially challenging to make sense of emotional displays in such a community.

Previous studies have made important contributions to our knowledge of the cultural and linguistic subtleties of the Jordanian environment in the area of sentiment analysis. Kawahara et al. (2012) conducted research on sentiment analysis in Arabic social media, focusing on the differentiation between tweets expressing positive and negative emotions. The study stresses the necessity to consider linguistic nuance when reading sentiment in the Arab area and shows the need to utilize culturally sensitive sentiment analysis algorithms for Arabic text.
Methods

Data in multiple languages, including Arabic and English was collected by carefully selecting a wide variety of online sources and platforms. Important social media sites including Twitter, Facebook, Instagram, and regional Arabic-language forums were mined to compile a dataset containing social media posts, comments, and online reviews. The data collection process was place over three months, beginning in January 2023.

There are various steps involved in the data preparation process. Initially, language identification was accomplished via the use of language detection libraries. Multiple language texts were found and cataloged separately for future study. After that, a text normalization procedure was put into place, which included changing all the letters to lowercase, getting rid of the punctuation, and taking out the web addresses. Tokenization was used to separate the terms in the cleaned text data for use in sentiment analysis.

We used lexicon-based sentiment analysis tools with pre-trained sentiment analysis models to accomplish our task of sentiment analysis. Our dataset's models and technologies were selected and tweaked with the specific needs of the target languages in mind. Each language's sentiment analysis was conducted independently.

In our comparative research, we looked at how different languages fared in terms of the distribution of emotion. Each statement was assigned a sentiment score that indicated how good, negative, or neutral it was. The sentiment evaluations were gathered and studied to uncover trends, variations, and similarities across various languages.

To undertake a quantitative evaluation and comparison of sentiment expressions, we employed many metrics including sentiment polarity distribution, sentiment intensity, and sentiment frequency. Using these metrics, we were able to identify substantial differences and similarities between the studied languages and their emotional expressions.

The statistical significance of the observed differences in the expression of emotions across languages was determined by statistical studies. T-tests and ANOVA were employed to identify any linguistically-specific patterns or variations in attitude.

Results

Table 1. Descriptive Statistics Analysis Results

<table>
<thead>
<tr>
<th>Language</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>3.33</td>
<td>3.5</td>
<td>0.94</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>English</td>
<td>4.13</td>
<td>4</td>
<td>0.78</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

The average sentiment score is an average across all languages for how we feel about something. We may observe, for instance, that the average emotion score in Arabic is 3.33. Based on this statistic, it seems that Arabic-language posts have an average sentiment close to neutral.

If you sort the evaluations of how you felt by ascending or descending order, the number in the middle is the median. Half of the posts have ratings of 3.5 or below, while the other half have ratings of 3.5 or higher, according to the median sentiment score in Arabic. Indicative of the center tendency of the emotion distribution, the median provides useful information.

An emotion rating's standard deviation is a statistical measure of its dispersion or variability. The bigger the standard deviation, the more dissimilar the evaluations of how you felt were. The 0.78 standard deviation found in English posts is indicative of a very stable distribution of opinion strength throughout the language.
The lowest sentiment score found in the dataset is called the minimum sentiment score. Sentiment scores in Arabic go as low as 2, representing the absolute minimum degree of positive feeling that can be conveyed in the language.

The highest possible rating in the dataset is called the maximum sentiment score. The maximum possible score of 5 for English indicates that highly pleasant sentiment may be conveyed via written English.

Table 2. Sample T-Tests Analysis on Arabic vs. English

<table>
<thead>
<tr>
<th>Language</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>3.33</td>
<td>0.94</td>
<td>30</td>
</tr>
<tr>
<td>English</td>
<td>4.13</td>
<td>0.78</td>
<td>30</td>
</tr>
</tbody>
</table>

The findings of the T-test analysis reveal a noteworthy distinction in the average sentiment scores between the Arabic and English languages, as denoted by a statistically significant p-value of 0.002. The observed T-statistic of -3.21 indicates a noteworthy disparity in mean sentiment scores between English and Arabic posts, thereby implying that English posts tend to exhibit a higher average positivity level.

Table 3. ANOVA Analysis Results:

<table>
<thead>
<tr>
<th>Language</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>3.33</td>
<td>0.94</td>
<td>30</td>
</tr>
<tr>
<td>English</td>
<td>4.13</td>
<td>0.78</td>
<td>30</td>
</tr>
</tbody>
</table>

Within this analysis, the F-statistic manifests as 2.23, accompanied by a discernible allocation of degrees of freedom. Specifically, the degrees of freedom between groups amount to 2, while the degrees of freedom within groups are quantified at 87. The statistical analysis reveals a two-tailed p-value of 0.112, indicating a moderate level of significance.

The analysis of variance (ANOVA) outcome reveals a p-value of 0.112, surpassing the conventional threshold of significance set at 0.05. Henceforth, we find ourselves unable to refute the null hypothesis. This observation suggests that there exists a lack of substantial disparities in the average sentiment scores across Arabic and English languages within the dataset under examination.

Discussion

Cross-Lingual Sentiment Expression in Jordan

The interplay of language, culture, and the expression of emotion is complicated and may be better understood by examining sentiment analysis within a cross-lingual framework. The investigation of sentiment in Jordan, covering Arabic and English languages, gives a fascinating look into the subtle nature of linguistic variation and its implications on emotion. Rugh (2019) argues that Jordan represents a unique hybrid of old-fashioned principles and cutting-edge technology. The linguistic choices used in online debates mirror this duality. The cultural connotations of different languages have a major impact on how individuals express themselves in the ever-changing digital space.

The fact that average evaluations for emotion differ across languages indicates the existence of a deeper narrative. Both the great variety of languages spoken in Jordan and the broader sociolinguistic tendencies in the region are highlighted by the findings. The local customs, traditions, and mindsets are often reflected in the Arabic language, which is the official language of the Kingdom. On the other hand, the global and typically more positive discourse is associated with English, which is growing dominating in the digital space.
domain. Our results suggest that people's linguistic preferences in this multilingual context reveal their ties to certain cultural communities. Graddol (2018) argues that in today's increasingly interconnected world, language serves as a powerful signifier of status, affiliation, and mindset. This experiment using sentiment analysis in Jordan provides an excellent example of this phenomenon.

Furthermore, our results are consistent with the "language ideologies" concept established by Woolard (1998), which highlights the political and social ramifications of language choice. A person's identity and the message they want to convey are both factors in the language they use to express their feelings and ideas in Jordan. Since English is spoken throughout the globe, the language is automatically associated with modernity, progress, and a positive outlook. Unlike Arabic, which is deeply rooted in local culture and history, it is capable of conveying a larger variety of emotions and sentiments. In the modern digital era, language serves as a platform for the negotiation of cultural and social identities (Bucholtz & Hall, 2004). This is especially evident in the complex interplay between various languages used to communicate emotion.

Our research shows that there are subtle power dynamics at play in the way emotions are expressed in various languages. The favorable associations with the English language raise interesting questions about the power of dominant languages to shape the way we communicate online. Pennycook (2018) claims that the phenomenon of English's rise to prominence as a global language carries with it the dangers of linguistic imperialism. In the instance of Jordan, the greater degree of optimism seen in English posts may not only suggest a preference for a language that is widely recognized worldwide, but also a propensity to adhere to a discourse that is viewed as generally positive. A philosophical consideration of the possible intertwining of linguistic preferences and emotional expression with the sociopolitical hierarchy of languages in our globalized world is prompted by this phenomenon.

In Jordan, there are many different languages spoken because of the country's long and complex history, as well as its prominent position on the global stage. Our findings demonstrate that focusing on linguistic preferences offers a fresh angle from which to probe a wide range of cultural and historical influences. People's language preferences in the modern digital world are an expression of who they are, what they care about, and how they feel. There are many different languages spoken in Jordan, and this makes the country a microcosm of the global interaction between language, culture, and the expression of emotion (Amara, 2020).

Sentiment Expressions Across Different Languages in Jordan

This research analyzes how different languages in the Jordanian setting communicate emotions. In addition to exposing linguistic differences, this study sheds light on the complex relationship between words and feelings across cultures. A philosophical investigation of the role of languages in determining the emotional topography of a culturally diverse society is made possible by these results. According to Lee (2020), there are more factors at play than communication when deciding on a language.

The disparities between Arabic and English in terms of average assessments of happiness and sadness give rise to intriguing inquiries regarding the influence of different linguistic systems. Illustrative of the capacity of language to convey emotion, the English posts were considerably more optimistic than their Arabic counterparts. This discovery aligns with the tenet of linguistic relativity, which posits that our emotions and cognitions can be influenced by language (Boroditsky, 2011). One could argue that individuals in Jordan adopting the English language as their primary means of communication convey a more positive emotional disposition. This finding serves as additional evidence of the fundamental capacity of language to shape emotions.
The results obtained from our research stimulate investigations into the theoretical framework known as linguistic determinism, which suggests that language has an impact on both our cognitive processes and affective experiences (Whorf, 1956). The concept of language determinism remains a subject of ongoing scholarly discourse and deliberation. Nevertheless, our research indicates that the selection of vocabulary can indeed influence the overall tone and demeanor of an individual's verbal expression. The use of English to foster optimistic perspectives exemplifies the power of language on the mind and heart. This phenomenon causes people to examine the influence of language choices on the display of emotions, as well as the probable effect on one's perception and experience of these sentiments.

The intricate interaction between linguistic structure and emotional structure is further shown by the observed changes in emotions based on language usage. Considering English's global prominence, its continued presence in Jordan may be seen as evidence of the political and social clout enjoyed by the world's most spoken language. Linguistic imperialism, as defined by Pennycook (2018), is the imposition of one's ideas and worldview on a target audience. Some suggest that the influence of a global discourse that frequently defines what is regarded as good or progressive may also contribute to the development of unequal emotional expression in English. Within the context of our results, it is interesting to contemplate the philosophical implications of analyzing the emotional hierarchy that is produced by the use of language.

Cultural and Linguistic Factors Influencing Sentiment Analysis Results

Understanding sentiment, particularly in a multilingual context like Jordan, requires going beyond simple data analytics work and delving deeply into the cultural and linguistic underpinnings of nuanced emotional expression. Our findings demonstrate the profound effect that differences in culture and language have on the outcomes of sentiment analysis. In this discussion, we’ll look at the philosophical and sociological implications of sentiment analysis results by referencing the writings of Kuckartz (2013). The impact of cultural and linguistic factors on the formation of such outcomes is investigated.

Because of how deeply language and culture are intertwined, they provide several ways to convey feelings. The choice of language is a visible cultural representation in Jordan, a nation where tradition and modernity coexist. As the de facto standard, Arabic is a potent representation of the region's historical and cultural traditions and shared identity. Because of its rich cultural history, the Arabic language is capable of expressing a wide range of nuanced meanings and feelings in its words (Rugh, 2019). Our research shows that Arabic posts are related to a broad variety of emotional evaluations, which sheds information on the cultural diversity of the language. Geertz's (1973) theory that culture may be seen as a complex web of symbols is consistent with the observed variability. As a powerful symbol system, language plays a crucial role in the transmission of cultural expressions. Emotional expressions in language are an excellent example of how culture is dynamic rather than static.

Jordan's linguistic environment is distinguished by a particular blend of Arabic and English in its sociolinguistic makeup. The decision of language is inextricably tied to an individual's cultural and social identity. English provides a gateway to global civilization, and Arabic acts as a bridge between traditional values and modern developments. Language ideologies and the political aspect of language are reflected in the speaker's linguistic choices (Rodríguez-Ordóñez, 2019). Language ideologies stress how language selection is intrinsically linked to social hierarchies, personal and group identities, and established norms and values. Choosing to express feelings in a public setting in Jordan requires negotiating several cultural and social norms. Sentiment analysis results are more nuanced than simple numerical values because they reveal the complex interplay between languages and the opinions they communicate.
Cross-Lingual Sentiment Analysis: Comparative Study of Opinion Expression Across Different Languages

Sapir-Whorf relativity, often known as the Sapir-Whorf hypothesis, provides a useful lens through which to view our results. According to the theory of linguistic relativity, the way we think and the reality we perceive may both be shaped and influenced by the language we use. Particularly in the Jordanian setting, the choice of language may be inferred to influence people's cognitive interpretation and vocal representation of emotions using sentiment analysis. Posts written in English tend to be upbeat because of the way the language is designed to evoke feelings. This finding is consistent with the theory that different languages provide contrasting emotional landscapes, with English displaying a more upbeat and optimistic one. In a world where so many different languages are spoken, questions arise concerning how much our linguistic preferences influence our subjective experiences and viewpoints.

Pennycook (2018) raises a philosophical point about the study of linguistic imperialism that deserves further attention. Because of its global popularity, English has become a symbol of linguistic superiority. Our analysis of sentiment supports this view by revealing a hierarchical structure within feelings. Linguistic imperialism is a phenomenon that includes the imposition of English as a dominant means of communication as well as accompanying emotional standards. The fact that English-language posts tend to be more upbeat suggests that the global conversation is affecting local feelings. The results of our research shed light on the complex interplay between language, culture, and emotion. These findings call for serious philosophical consideration of the far-reaching effects of language imperialism on people's psychological health.

Conclusion

This research delves into the complicated interplay between language diversity and cultural and political influences on the expression of emotion in different tongues. The study's overarching goals were to (1) investigate the phenomenon of cross-lingual sentiment expression; (2) compare the expression of emotion in other languages; and (3) identify the cultural and linguistic factors that impact the findings. The findings have offered deep insights into the philosophical and sociological dimensions of language and culture, going well beyond the limits of sentiment analysis.

We started by looking at how people of different languages express their emotions with one another in Jordan, a country at the crossroads between antiquity and modernity. Our research showed that there were clear differences in the expression of emotion between posts made in Arabic and English. These differences might be due to the different cultural and linguistic backgrounds of the people who posted in each language. One's identity, loyalties, and cultural heritage may be shown in the language one chooses to use in this setting. According to Lee (2020), language serves as a medium for expressing one's culture and one's unique sense of self. Our findings provide credence to this view.

We also compared emotional expressions across languages to provide light on the role language plays in the presentation of feelings. English posts, associated with a wider and more global discourse, got much higher emotional ratings than Arabic posts. The aforementioned result stimulates substantial questions concerning linguistic relativity and the capability of language to impact our emotional experiences. Boroditsky (2011) argues that the words we use influence how we feel and how we convey those feelings. This finding demonstrates how our words may have a profound effect on our feelings.

At last, we have identified the linguistic and cultural factors that influence the results of sentiment analysis, illuminating the intricate interplay between language ideology, linguistic imperialism, and the dynamics of different cultures. The linguistic hierarchy and the sociopolitical clout of Jordan's most popular languages are reflected in the country's language preferences. Perhaps the influence of a global discourse that establishes what is seen as positive and progressive is reflected in the great range of
emotional emotions in the English language. This motivates theoretical consideration of how language imperialism affects our subjective experiences and the words we choose.

Acknowledgements

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Small Research Groups under grant number (RGP.1 / 271 /44).

References


Khasawneh, M. A. S. (2022). The degree of practicing effective communication skills among teachers of learning disabilities in English language from their point of view. Science and Education, 3(2), 492-509. https://orcid.org/0000-0002-1390-3765


Khasawneh, M. A. S. (2022). The level of practicing organizational trust among teachers of learning disabilities in English language from their point of view. Science and Education, 3(2), 481-491. https://orcid.org/0000-0002-1390-3765


Cross-Lingual Sentiment Analysis: Comparative Study of Opinion Expression Across Different Languages


Whorf, B. L. (1956). Language, thought, and reality: selected writings of….(Edited by John B. Carroll.).
