International Journal of Advanced Trends in Computer Applications
www.ijatca.com

An Analysis of Multilingual features based Machine Learning Approaches for Sentence Level Sentiment Analysis

[1]Mr. Harjender Singh
[1]Asstt. Professor, Maharaja Surajmal Institute, New Delhi
[1]*harjendersingh@msijanakpuri.com*

Abstract: *Sentiment analysis, a subfield of natural language processing (NLP), involves the automated identification and classification of sentiment or opinion expressed in text. Traditionally, sentiment analysis has focused on English language texts, but with the increasing availability of multilingual data on social media, online reviews, and news articles, there is a growing demand for sentiment analysis in multiple languages. Analyzing sentiment in multiple languages presents unique challenges due to linguistic differences, cultural nuances, and the availability of labeled data.*
This paper provides an analysis of features based machine learning approaches used for sentiment analysis in multiple languages. It discusses the challenges and considerations specific to multilingual sentiment analysis and provides insights into the performance and effectiveness of different machine learning models.
The goal is to explore the performance, effectiveness, and generalization capabilities of different machine learning models across diverse linguistic contexts.

Keywords: multilingual data, social media, online reviews.

I. INTRODUCTION

Sentiment analysis, also known as opinion mining, is a computational technique that involves the use of natural language processing (NLP), machine learning, and text analytics to identify, extract, and classify subjective information expressed in text data. It aims to determine the sentiment or emotion conveyed by individuals towards a particular topic, product, service, or event. The analysis encompasses a range of machine learning techniques, including supervised learning algorithms such as Support Vector Machines (SVM), Naive Bayes, Decision Trees, and neural network architectures like Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). These models are trained and evaluated on labeled datasets in multiple languages, covering a diverse range of domains and sources. There are some key reasons why sentiment analysis is crucial in various domains:

Social media analysis: It refers to the process of extracting valuable insights and information from social media data. It involves analyzing and understanding the vast amount of user-generated content on social media platforms such as Twitter, Facebook, Instagram, LinkedIn, and YouTube. Social media analysis utilizes various techniques, including sentiment analysis, topic modeling, network analysis, and trend detection, to gain actionable insights and understand user behavior, opinions, and trends.

Financial Analysis: Is the process of assessing the financial health, performance, and prospects of a company or organization. It involves the examination and interpretation of financial statements, key financial ratios, market trends, and other relevant financial data to evaluate the company's financial position, profitability, and investment potential. Financial analysis utilizes various quantitative and qualitative techniques, including financial modeling, data analysis, market research, and industry knowledge. It requires expertise in accounting principles, financial management, and investment analysis.

Public Opinion and Policy Making: It refers to the collective views, attitudes, and preferences of the general public on various social, political, and economic issues. Public opinion requires robust research methods, data analysis techniques, and engagement with diverse segments of the population. Public opinion plays a crucial role in policy communication and education. Public opinion serves as

a platform for citizens to voice their opinions, concerns, and demands. Public opinion research involves the systematic collection and analysis of data to understand the attitudes, beliefs, and preferences of the public on specific policy issues.

Business and marketing strategies: Play a critical role in the success and growth of businesses. Business and marketing strategies are developed to attract customers, increase market share, and generate revenue. Effective marketing strategies involve product development and innovation. Marketing strategies incorporate analytics and performance measurement to assess the effectiveness of marketing efforts. Market segmentation helps tailor marketing strategies to specific customer groups, enabling businesses to deliver targeted messages and offerings.

Need for sentiment analysis in multiple languages: Sentiment analysis in multiple languages can be challenging because different languages have different linguistic characteristics, such as grammar, syntax, vocabulary, and idioms. Sentiment analysis models need to be able to handle the complexities and nuances of different languages. Some of the current approaches and challenges for sentiment analysis in multiple languages are:

Using language-specific preprocessing and feature extraction techniques: This involves applying techniques such as part-of-speech tagging, lemmatization, stemming, negation handling, and n-gram extraction to each language separately. This can help capture the relevant linguistic features for sentiment analysis.

Using machine learning and deep learning methods: This involves training supervised or unsupervised models on labeled or unlabeled data to learn the sentiment patterns and rules for each language. Some examples of machine learning methods are support vector machines (SVM), naive Bayes (NB), and decision trees (DT). Some examples of deep learning methods are recurrent neural networks (RNN), convolutional neural networks (CNN), and transformers (BERT).

Using feature selection and optimization methods: This involves selecting the most informative and discriminative features for sentiment analysis and reducing the dimensionality of the feature space. This can help improve the performance and efficiency of the models. Some examples of feature selection methods are information gain (IG), chi-square (CHI), and genetic algorithms (GA).

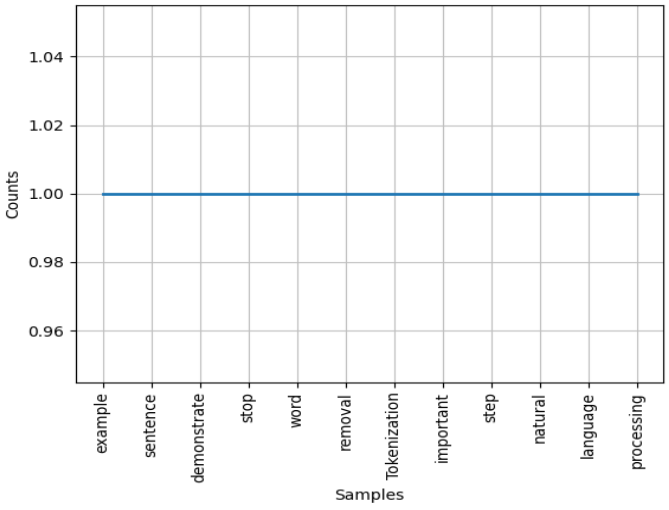
Using cross-lingual and multilingual methods: This involves leveraging the available sentiment resources of a rich-resource language (such as English) to transfer or adapt them to a low-resource language (such as Arabic).

Using multimodal methods: This involves combining textual data with other modalities, such as audio, video, or images, to enrich the sentiment information and capture the multimodal cues for sentiment analysis. This can help enhance the accuracy and robustness of the models.

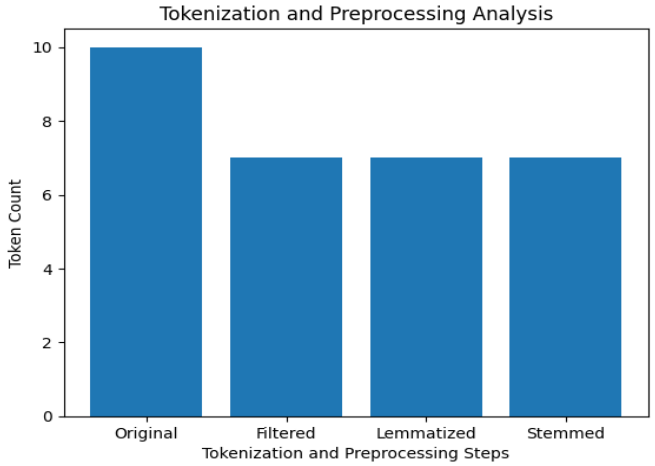
Challenges in Multilingual Sentiment Analysis: Multilingual sentiment analysis is the task of identifying and extracting the emotions, opinions, attitudes, and feelings expressed in text written in different languages. It can be useful for various applications, such as customer service, social media analysis, product reviews, and market research. Some of the challenges in multilingual sentiment analysis are:

1. Stop words and text classifiers
2. Structural differences across languages.
3. Limited Resources
4. Cultural Variation
5. Sentiments Polarity
6. Translation Challenges
7. Sentiment Expression Variations
8. Handling Linguistic Features

Example: using stop words, tokenization, Lemmatization and Stemming

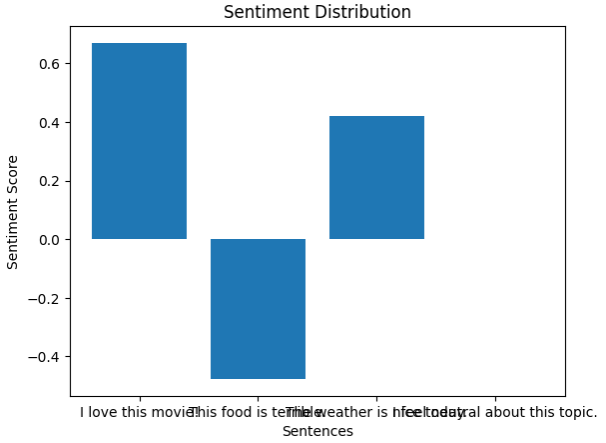


Example: using stop words, tokenization, Lemmatization and Stemming : "The quick brown fox jumps over the lazy dog."



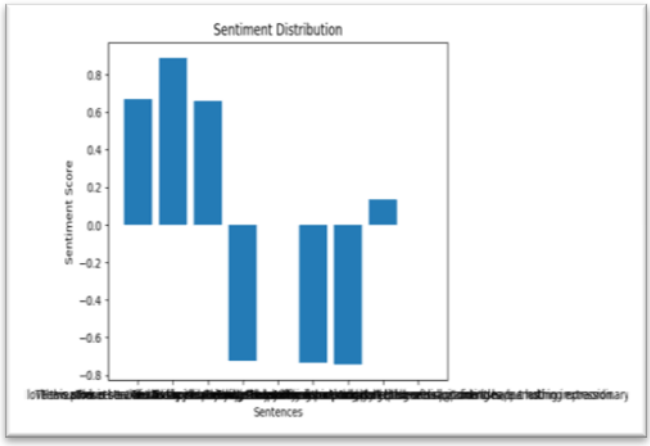
Original Tokens: ['The', 'quick', 'brown', 'fox', 'jumps', 'over', 'the', 'lazy', 'dog', '.']
Filtered Tokens: ['quick', 'brown', 'fox', 'jumps', 'lazy', 'dog', '.']
Lemmatized Tokens: ['quick', 'brown', 'fox', 'jump', 'lazy', 'dog', '.']
Stemmed Tokens: ['quick', 'brown', 'fox', 'jump', 'lazi', 'dog', '.']

Example: Visualize sentiment distribution and calculate the sentiment scores. The sentiment scores range from -1 (negative sentiment) to +1 (positive sentiment), with 0 indicating neutral sentiment.



Example: Visualize sentiment distribution and calculate the sentiment scores. the sentiment scores range from -1 (negative sentiment) to +1 (positive sentiment), with 0 indicating neutral sentiment.

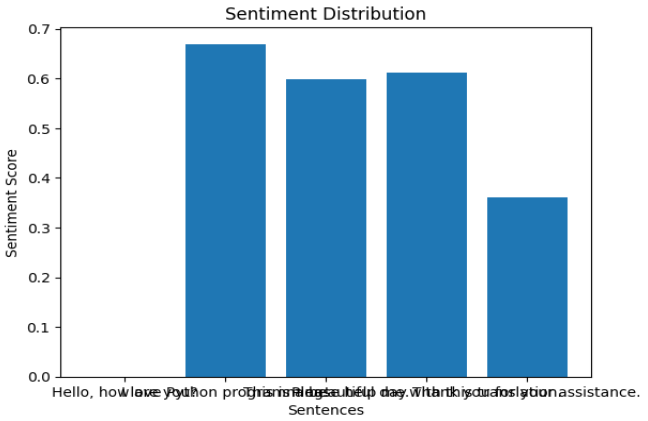
```
import nltk
from nltk.sentiment import SentimentIntensityAnalyzer
import matplotlib.pyplot as plt
# Download required NLTK resources
nltk.download('vader_lexicon')
# Example sentences
sentences = [
    "I love this product! It exceeded my expectations.",
"The weather is beautiful today. I'm enjoying the sunshine.",
"The customer service was excellent. They resolved my issue promptly.",
    "This restaurant has terrible service. The staff was rude and unhelpful.",
"The traffic is unbearable today. It took me hours to get home.",
"I'm so disappointed with the quality of this product. It broke after just one use.",
    "Today is an average day. Nothing particularly exciting or disappointing happened.",
"The book was okay. It had some interesting parts, but overall, it didn't leave a lasting impression.",
"The hotel room was decent. It had all the necessary amenities, but nothing extraordinary."
]

# Create an instance of SentimentIntensityAnalyzer
sia = SentimentIntensityAnalyzer()

# Calculate sentiment scores for each sentence
sentiment_scores = []
for sentence in sentences:
sentiment_score = sia.polarity_scores(sentence)['compound']
sentiment_scores.append(sentiment_score)

# Plot sentiment distribution
plt.bar(sentences, sentiment_scores)
plt.xlabel("Sentences")
plt.ylabel("Sentiment Score")
plt.title("Sentiment Distribution")
plt.show()
```



Example: To calculate sentiment scores using the VADER sentiment analysis tool

Example: polarity of some sentences

```
"Hello, how are you?",
    "I love Python programming!",
    "This is a beautiful day.",
    "Please help me with this translation.",
    "Thank you for your assistance."
```



Machine Learning Approaches for Multilingual Sentiment Analysis:

Feature-based Approaches: - Feature-based approaches in sentiment analysis typically involve extracting relevant features from text and using them as input to machine learning algorithms.

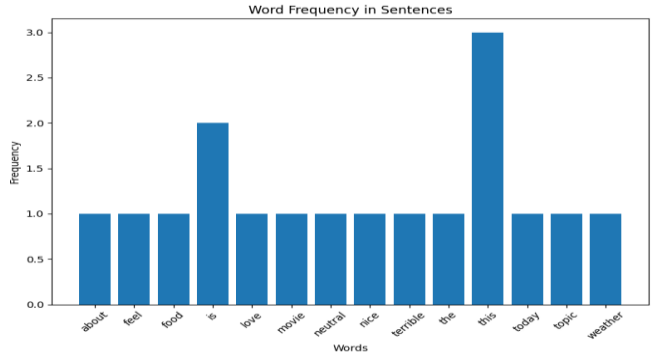
Bag-of-Words models: It represents text as a collection of individual words, disregarding grammar and word order, and focusing only on the presence and frequency of words.

sentences = [
"I love this movie!",
"This food is terrible.",
"The weather is nice today.",
"I feel neutral about this topic."
]
Sentence: I love this movie!
Features: ['about', 'feel', 'food', 'is', 'love', 'movie', 'neutral', 'nice', 'terrible', 'the', 'this', 'today', 'weather']
Vector: [0 0 0 0 1 1 0 0 0 0 1 0 0]

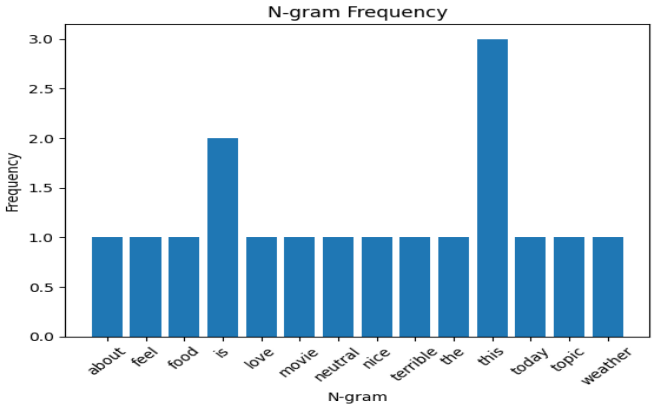
Sentence: This food is terrible.
Features: ['about', 'feel', 'food', 'is', 'love', 'movie', 'neutral', 'nice', 'terrible', 'the', 'this', 'today', 'weather']
Vector: [0 0 1 1 0 0 0 0 1 0 1 0 0]

Sentence: The weather is nice today.
Features: ['about', 'feel', 'food', 'is', 'love', 'movie', 'neutral', 'nice', 'terrible', 'the', 'this', 'today', 'weather']
Vector: [0 0 0 1 0 0 0 1 0 1 0 1 1]

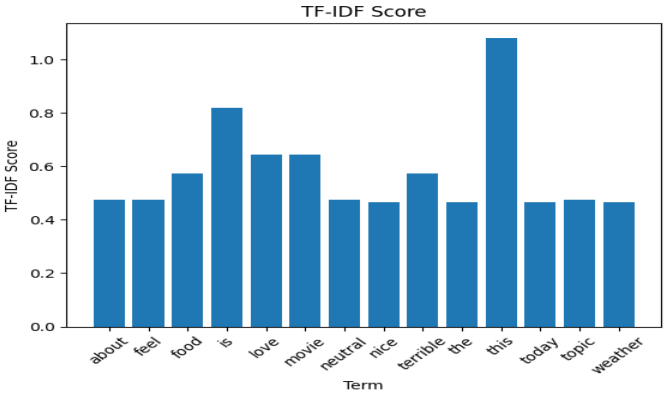
Sentence: I feel neutral about this topic.
Features: ['about', 'feel', 'food', 'is', 'love', 'movie', 'neutral', 'nice', 'terrible', 'the', 'this', 'today', 'weather']
Vector: [1 1 0 0 0 0 1 0 0 0 1 0 0]



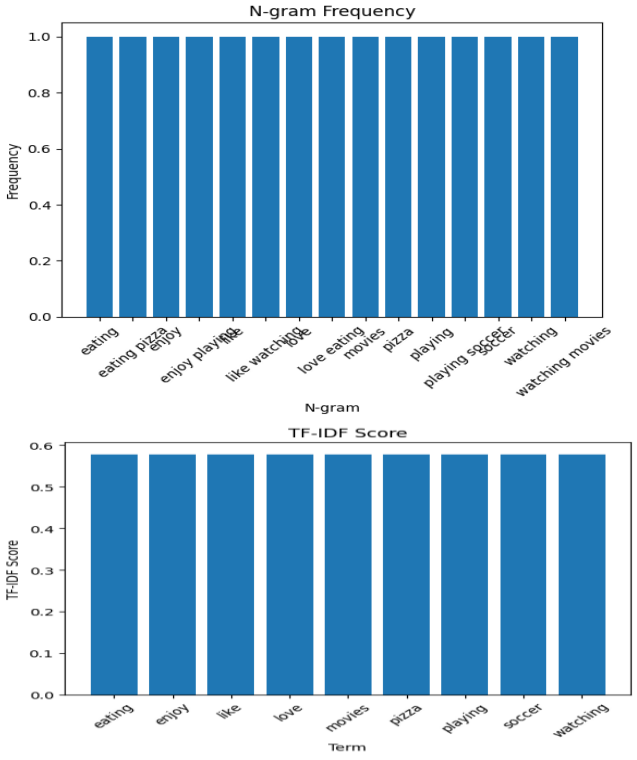
N-grams and TF-IDF representation: N-grams are contiguous sequences of n items from a given text, where the items can be characters, words, or even sentences. N-grams capture the local information and context within the text.
example, in the sentence "I love pizza," the bigrams (2-grams) would be "I love" and "love pizza."



TF-IDF is a numerical statistic that reflects the importance of a word in a document or a corpus. It is calculated by multiplying the term frequency (TF) and inverse document frequency (IDF) of a term. Term Frequency (TF) measures how frequently a term appears in a document. Inverse Document Frequency (IDF) measures the rarity of a term across all documents in the corpus. TF-IDF assigns higher weights to terms that are frequent in a document but rare across the corpus.



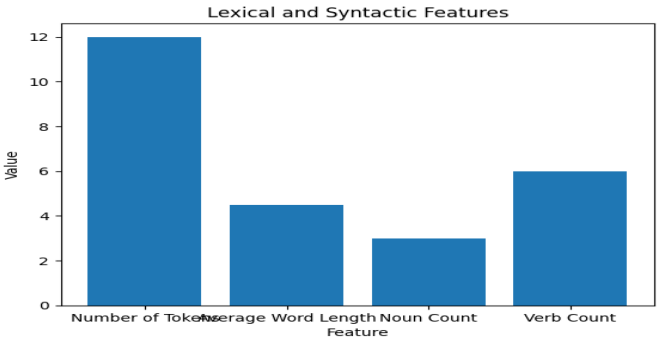
Example : sentences = [
"I enjoy playing soccer.",
"I like watching movies.",
"I love eating pizza."]



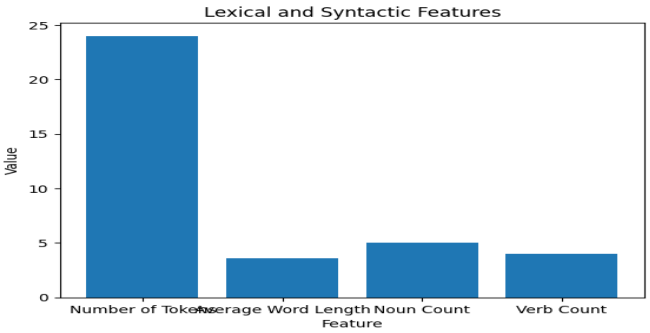
Lexical and syntactic features: Lexical features focus on the individual words or tokens in the text. They capture information related to word choice, frequency, and characteristics.
Examples: lexical features include word frequency, word length, word type, and word context.

Syntactic features involve analyzing the grammatical structure and relationships between words in a sentence. They capture information related to parts of speech, syntactic patterns, and grammatical dependencies.
Examples : syntactic features include part-of-speech tags, syntactic parse trees, noun phrases, verb phrases, and syntactic relations.

```
from nltk import word_tokenize, pos_tag
import matplotlib.pyplot as plt
# Example sentence
sentence = "I like watching movies I enjoy playing soccer I love eating pizza "
# Tokenize the sentence
tokens = word_tokenize(sentence)
# Calculate the lexical features
num_tokens = len(tokens)
avg_word_length = sum(len(word) for word in tokens) / num_tokens
# Calculate the syntactic features
pos_tags = pos_tag(tokens)
noun_count = sum(1 for tag in pos_tags if tag[1].startswith('N'))
verb_count = sum(1 for tag in pos_tags if tag[1].startswith('V'))
# Plot the graph
labels = ['Number of Tokens', 'Average Word Length', 'Noun Count', 'Verb Count']
values = [num_tokens, avg_word_length, noun_count, verb_count]
plt.bar(labels, values)
plt.xlabel("Feature")
plt.ylabel("Value")
plt.title("Lexical and Syntactic Features")
plt.show()
```



Example: to calculate syntactic and lexical features for the sentence = "I love this movie!, This food is terrible. The weather is nice today. I feel neutral about this topic."



II. CONCLUSION

The objective of this analysis is to investigate the effectiveness of using multilingual features in machine learning models to accurately classify the sentiment expressed in sentences. This paper provides an analysis of various machine learning approaches used for sentiment analysis in multiple languages. The analysis reveals that machine learning approaches for sentiment analysis can achieve high accuracy and performance across multiple languages, indicating their general applicability. It also highlights the significance of feature engineering, including lexical, syntactic, and semantic features, for capturing the nuances of sentiment in diverse languages. In future we can further extend our research using pre-trained models like BERT, to improve sentiment analysis performance, especially with limited labeled data. This paper aims for researchers and practitioners to provide valuable insights and guidelines for effectively performing sentiment analysis in diverse linguistic contexts.

REFERENCES

[1]. Wankhade, M., Rao, A. C. S., &Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 1-50.

[2]. Rosa, R. L., Schwartz, G. M., Ruggiero, W. V., & Rodríguez, D. Z. (2018). A knowledge-based recommendation system that includes sentiment analysis and deep learning. IEEE Transactions on Industrial Informatics, 15(4), 2124-2135.
[3]. https://www.gmrwebteam.com/blog/understanding-the-importance-of-sentiment-analysis-in-healthcare
[4]. Lai, S. T., &Mafas, R. (2022, April). Sentiment Analysis in Healthcare: Motives, Challenges & Opportunities pertaining to Machine Learning. In 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-4).IEEE.
[5]. https://monkeylearn.com/blog/intent-classification
[6]. Nandwani, P., &Verma, R. (2021). A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining, 11(1), 1-19.
[7]. Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013, June). What yelp fake review filter might be doing?.In Proceedings of the international AAAI conference on web and social media (Vol. 7, No. 1).
[8]. Wankhade, M., Annavarapu, C. S. R., &Verma, M. K. (2022). CBVoSD: context based vectors over sentiment domain ensemble model for review classification. The Journal of Supercomputing, 78(5), 6411-6447.
[9]. Eke, C. I., Norman, A. A., Shuib, L., &Nweke, H. F. (2020). Sarcasm identification in textual data: systematic review, research challenges and open directions. Artificial Intelligence Review, 53(6), 4215-4258.
[10]. Poria, S., Hazarika, D., Majumder, N., &Mihalcea, R. (2020). Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. IEEE Transactions on Affective Computing.