Ashmita Singh

Project title: Sentiment Analysis on social media data

LITERATURE REVIEW :

Methodology, algorithms, advantages, disadvantages, applications etc.

SR.NO	**TITLE OF THE PAPER**	**AUTHORS**	**YEAR**	**REMARKS**
1.	Fine-Grained Sentiment Analysis of Social Media with Emotion Sensing	Zhaoxia WANG, Chee Seng CHONG, Landy LAN, Yinping YANG, Seng Beng HO and Joo Chuan TONG	2016	**METHODLOGY:** The methodology presented in this paper is a social media analytics method that can perform fine-grained sentiment and emotion analysis of text data. The main features of the method are: • It uses a social adaptive fuzzy inference algorithm that mimics human interpretations of the expression of attitudes and emotions in online social network contexts. • It has a built-in advanced linguistic processing unit that handles negation, amplification, diminution, and other linguistic phenomena. • It leverages multiple sources of lexicons, including a dictionary of emotion words and phrases from Standard English, Internet/social media slang and local languages, as well as emoticons. • It can classify text messages into sentiment categories (positive, negative, neutral and mixed), and identify their prevailing emotion categories (e.g., satisfaction, happiness,

				excitement, anger, sadness, and anxiety). • It is implemented within an end-to-end social media analysis system that has the capabilities to collect, filter, classify, and analyze social media text data and display a descriptive and predictive analytics dashboard for a given concept. **ALGORTHM:** The algorithm for fine-grained sentiment analysis of social media with emotion sensing is a social adaptive fuzzy similarity-based classification method. It uses the following steps: • Preprocess the text messages by removing noise, handling negation, amplification, and diminution, and decomposing sentences into clauses. • Compare each clause with a lexicon of emotion words and phrases, and assign a fuzzy similarity score for each emotion category. • Aggregate the fuzzy similarity scores across the clauses of a message, and determine the dominant sentiment category (positive, negative, neutral, or mixed) and the prevailing emotion category (e.g., satisfaction, happiness, excitement, anger, sadness, or anxiety). **ADVANTAGES:** • It does not require any labelled training data, which reduces the cost and effort of annotation

				and makes it adaptable to different domains and languages. • It can perform fine-grained sentiment analysis, which provides more specific and actionable insights than coarse-grained sentiment analysis. • It can handle the linguistic variations and ambiguities of social media text, such as slang, emoticons, and sarcasm, by using advanced linguistic processing and fuzzy logic. **DISADVANTAGES:** • It relies on a lexicon of emotion words and phrases, which may not cover all the possible expressions of sentiment and emotion in social media text, and may need to be updated frequently to capture new terms and trends. • It may not capture the context and intention of the text messages, which may affect the accuracy and reliability of the sentiment and emotion analysis. • It may not account for the influence of external factors, such as the source, time, and location of the text messages, on the sentiment and emotion of the users. **CONCLUSION:** This algorithm is a novel and promising method for fine-grained sentiment analysis of social media with emotion sensing, which can provide valuable information for various applications, such as marketing,

				customer service, and public opinion mining. However, it also has some limitations and challenges that need to be addressed and improved in future research.
2.	Mental Health Sentiment Analysis on social media TikTok with the Naïve Bayes Algorithm	Yuyun Yusnida Lase, Arif Ridho Lubis, Fauziah Elyza, Sekar Arini Syafli	2023	**METHODOLOGY:** • The paper uses sentiment analysis to evaluate the opinions and emotions expressed in TikTok comments related to mental health. • The paper collects data from TikTok using the TikTok API, filters the data using keywords related to mental health, preprocesses the data using techniques such as punctuation removal, stopword removal, and stemming, labels the data manually into three categories (positive, neutral, and negative), and applies the Naïve Bayes algorithm to classify the data into the three categories. • The paper evaluates the performance of the algorithm using metrics such as accuracy, precision, recall, and F1-score. **ALGORITHMS:** • The paper uses the Naïve Bayes algorithm, which is a classification method based on Bayes' theorem. The algorithm calculates the probability of each class (sentiment category) given the features (words) in the data, and assigns the class with the highest probability to the data.

				• The algorithm assumes that the features are independent of each other, which simplifies the computation and reduces the data requirements. **ADVANTAGES:** • The paper claims that the advantages of using the Naïve Bayes algorithm are its fast and easy-to-implement performance, its ability to handle data that has many features, and its suitability for text data analysis. • The paper also claims that the advantages of conducting sentiment analysis on TikTok content related to mental health are to provide insight into the normal use of the platform, to identify important issues that need attention in society, and to increase awareness and understanding of the importance of maintaining mental health. **DISADVANTAGES:** • The Naïve Bayes algorithm may not perform well if the features are not independent of each other, or if the data is imbalanced or noisy. The sentiment analysis may not capture the nuances and contexts of the comments, or may be influenced by the subjectivity and bias of the manual labelling. • The TikTok content related to mental health may not represent the

				general population or the diversity of mental health experiences, or may be affected by the algorithmic personalization and recommendation of the platform. **CONCLUSION:** The paper suggests that the applications of the research are to contribute to the development of knowledge about mental health topics in society, and to help make better decisions and strategies in the business, political, and social fields.
3.	BERT: a sentiment analysis odyssey	Springer Nature Limited	2021	**METHODOLOGY:** • The paper uses four sentiment analysis techniques to classify movie reviews from IMDB as positive or negative. The techniques are: SentiWordNet, logistic regression, LSTM, and BERT. • The paper preprocesses the text data by removing HTML tags, accented characters, contracted words, special characters, punctuations, rare words, and stop words. It also lemmatizes the words and converts them to lower case. • The paper partitions the data into training, validation, and test sets. The ratio of partitioning varies depending on the technique. For SentiWordNet, logistic regression, and LSTM, the ratio is 60:0:40. For BERT, the ratio is 35:15:50. • The paper applies each technique on the training

				data and evaluates its performance on the test data using four metrics: accuracy, precision, recall, and F1 score. The paper also compares the results of the four techniques and discusses their strengths and weaknesses. **ALGORITHMS:** • SentiWordNet: an unsupervised lexicon-based model that assigns positive, negative, and objective scores to WordNet synsets based on a random-walk algorithm. • Logistic regression: a supervised machine learning model that uses a sigmoid function to estimate the probability of a review being positive or negative based on a bag of words representation. • LSTM: a supervised deep learning model that uses a recurrent neural network architecture with long short-term memory units to capture the sequential information in the text. • BERT: an advanced supervised deep learning model that uses a bidirectional transformer to encode the context of each word in the text and pre-trains on two unsupervised tasks: masked language modelling and next sentence prediction. **ADVANTAGES:** • The paper offers a comprehensive comparison of four

				prominent sentiment analysis techniques on a large and balanced dataset of movie reviews. • The paper introduces BERT, a state-of-the-art technique that outperforms the other techniques on all metrics and demonstrates its potential for various natural language understanding tasks. • The paper provides insights and implications for analytics professionals and academicians working on text analysis and sentiment classification. **DISADVANTAGES:** • The paper does not consider the effect of different hyperparameters, such as learning rate, batch size, and number of epochs, on the performance of the techniques. • The paper does not account for the nuances and subtleties of natural language, such as sarcasm, irony, humor, and emotion, that may affect the sentiment of the reviews. • The paper does not explore the generalizability of the techniques to other domains, languages, and types of text data. **APPLICATIONS:** • The paper can help marketers and managers to understand the opinions and preferences of their customers and stakeholders based on online reviews and

				feedback. • The paper can help researchers and scholars to develop and improve sentiment analysis techniques and models for various text analysis problems and challenges. • The paper can help educators and students to learn and teach the fundamentals and applications of sentiment analysis and natural language processing. **CONCLUSION:** This paper has presented a comparative analysis of four sentiment analysis techniques: SentiWordNet, logistic regression, LSTM, and BERT. The analysis was based on a publicly available dataset of 50,000 movie reviews from IMDB. The results showed that BERT outperformed the other three techniques on all four metrics of accuracy, precision, recall, and F1 score. BERT's superiority can be attributed to its ability to capture the bidirectional context of words in a sentence. This paper has demonstrated the usefulness of BERT for sentiment classification and provided a valuable resource for analytics professionals and researchers interested in text mining. Future studies can explore the applications of BERT in other domains and tasks, as well as its limitations and challenges.
4.	CyberHelp: Sentiment Analysis on Social Media Data Using Deep Belief Network to Predict Suicidal Ideation of	U.Sakthi, Thomas M. Chen, Mithileysh Sathiyanarayanan	2023	**METHODLOGY:** • The paper aims to develop a solution called CyberHelp, which can automatically detect and predict suicidal ideation in students based on their social media posts using

	Students			deep learning and optimization techniques. • The paper uses a social media dataset collected from the UCI machine learning repository, which contains various posts by students expressing their emotions, opinions, and feelings. • The paper proposes a novel model based on a Deep Belief Network (DBN) and a Dragonfly Algorithm (DFA) for the task of suicidal ideation prediction. • The DBN is a multi-layer neural network that consists of several Restricted Boltzmann Machines (RBMs), which are unsupervised learning models that can learn the probability distribution of the input data. The DFA is a swarm intelligence meta-heuristic algorithm that mimics the behavior of dragonflies for solving optimization problems. • The DFA is used to tune the hyperparameters of the DBN, such as the number of hidden layers, the number of neurons, the learning rate, and the weight decay. **ALGORITHMS:** • Data preprocessing: The social media posts are preprocessed to remove unwanted data, such as URLs, hashtags, numbers, punctuation, and emojis, and to normalize the data format. • Tokenization: The preprocessed posts are split into words, which are then stored in a

				dictionary. The words are used for feature extraction and sentiment analysis. • Prediction processing: The proposed DBN model is trained using a greedy layer-by-layer approach, where each RBM is trained separately until all the RBMs are trained in a specific order. • The DBN model uses joint probability to train the stochastic neural network model. The DBN model takes the tokenized posts as input and outputs the predicted labels of suicidal or non-suicidal for each post. • Hyperparameter tuning: The proposed DFA algorithm is used to optimize the feature selection process and to improve the performance of the DBN model. The DFA algorithm uses five coefficients to control the behavior of the dragonflies, such as separation, alignment, cohesion, attraction, and distraction. The DFA algorithm selects the minimal number of features that give the maximum accuracy for the prediction task. **ADVANTAGES:** • Early Intervention: Detecting suicidal ideation early can save lives. • Unstructured Data: Deep learning handles unstructured text data effectively. • Powerful Tool: Deep learning extracts complex

				patterns from social media content. **DISADVANTAGES:** • Lexicon Dependency: Relying on emotion lexicons may miss nuances. • Context Limitations: May not fully capture context and intention. • Data Bias: Social media data may not represent the entire population. **CONCLUSION:** The paper presents an efficient and accurate solution for the automatic detection and prediction of suicidal ideation in students using their social media posts. The paper proposes a novel model based on a Deep Belief Network and a Dragonfly Algorithm, which can learn the probability distribution of the input data and optimize the feature selection and hyperparameter tuning processes. The paper evaluates the performance of the proposed model on a social media dataset and compares it with other existing models, such as Random Forest, Linear Regression, and Recurrent Neural Network. The paper shows that the proposed model achieves the highest accuracy of 96% for the suicidal ideation prediction task, which is significantly better than the other models. The paper demonstrates the potential of the proposed model for saving the lives of students who are at risk of suicide by providing early intervention and treatment.
5.	Studying social media sentiment using human	James Lappeman , Robyn Clark , Jordan Evans,	2023	**METHODOLOGY:** • The paper uses a combination of natural

	validated analysis	Lara Sierra-Rubia, Patrick Gordon		language processing (NLP) and human validation techniques to analyse online sentiment of South Africa's major banks over a 12-month period. • The paper collects social media posts from various platforms, filters them for relevance, assigns them sentiment and topic labels, and performs statistical and thematic analyses to identify trends and drivers of consumer opinions. **ALGORITHMS:** • Data collection and sampling: The paper collects social media posts from various platforms that mention any of the top five retail banks in South Africa over a year period. A filter is used to select relevant posts based on keywords and topics. A non-probability quota sampling technique is used to select a sample of 1 720 810 posts for analysis. • Sentiment analysis: The paper uses a machine learning algorithm to categorize the posts as either positive or negative sentiment, excluding neutral posts. The algorithm is based on a maximum entropy classifier trained on labelled data. The paper also uses human validation to increase the accuracy of the sentiment analysis, by having qualified human contributors rate a sub-sample of 521 326 posts for relevancy, sentiment

				and topics. Each post is rated by multiple contributors and a consensus model is used to resolve conflicts. • Topic analysis: The paper uses another NLP technique to identify and group the posts into 70 predefined banking topics, which are further categorized into seven broad themes. The paper uses Mallet, a tool for topic modelling, to classify the posts according to their relevance to the topics. The paper also uses human validation to assign topics to the posts that were verified for sentiment. The paper analyses the sentiment and volume of the posts within each theme and topic, as well as the main negative themes for each bank. • Analysis procedure: The paper analyses the individual banks and the industry as a whole, using the sentiment and topic data. The paper calculates the net sentiment and weighted net sentiment for each bank over time, and identifies spikes in negative sentiment that indicate online firestorms. The paper also analyses the factors that cause the firestorms and their duration. The paper also analyses the brand content performance of each bank, in terms of volume, engagement and average engagement per post. The paper also compares the net sentiment, share of voice and conversation trends across the banks.

				The paper also analyses the response rates, response times and final interaction sentiment of each bank on Twitter. • The paper uses a machine learning algorithm based on maximum entropy classifier to filter and classify the social media posts. The algorithm is trained on labelled data from the topic analysis and uses features such as words, n-grams, and hashtags to determine the relevance and sentiment of each post. • The algorithm is supplemented by human validation, where a sub-sample of posts is manually verified and assigned to topics by qualified contributors. **ADVANTAGES:** • It can capture unsolicited and non-coercive feedback from consumers, which reflects their lived experiences and emotions. • It can analyse a large and diverse sample of social media data, which increases the generalisability and validity of the results. • It can combine the strengths of both computer and human analysis, which improves the accuracy and reliability of the sentiment and topic labels. • It can provide longitudinal and comparative insights into the online sentiment and reputation of the banks, as well as the factors that

				influence them. **DISADVANTAGES:** • relies on public and accessible social media data, which may not represent the entire population of consumers or their opinions. • It may encounter ethical issues regarding the use and disclosure of personal information from social media users, especially if they are not aware of the research purposes. • It may face challenges in dealing with the complex and emotive nature of human conversation, such as sarcasm, irony, slang, and cultural references, which may affect the sentiment and topic analysis. **APPLICATIONS:** • Marketing and communication, where it can help to monitor and manage brand image, customer satisfaction, and loyalty, as well as to identify opportunities and threats in the market. • Risk and crisis management, where it can help to detect and respond to online firestorms, negative word-of-mouth, and complaint behaviour, as well as to mitigate the potential damage to the brand reputation and trust. • Social and behavioural science, where it can help to understand and explain the attitudes, preferences, and

				behaviours of consumers and stakeholders, as well as to test and validate theories and hypotheses. **CONCLUSION:** The paper concludes that online sentiment analysis with human validation is a useful and innovative method to study social media data and to gain insights into the consumer opinions and experiences in the banking industry. The paper also suggests some directions for future research, such as exploring other sources and types of data, developing more advanced and robust algorithms, and conducting more in-depth and comparative analyses.
6.	Toward Stopping Incel Rebellion: Detecting Incels in Social Media Using Sentiment Analysis	Mohammad Hajarian, Zahra Khanbabaloo	2021	**METHODOLOGY:** • The paper aims to detect incels (involuntary celibates) in social media using sentiment analysis and offensive language detection on user comments. • The paper collects user comments data from two social media platforms : Facebook and Twitter. It also collects incel comments from Incels.co for evaluation purposes. The paper proposes a four-step algorithm to identify incels based on their comments: • Preprocessing: removing punctuation, symbols, links, numbers, and spaces from the comments. • Sentiment analysis: using TextBlob to assign a polarity score to each comment, ranging from -1 (negative) to 1

				(positive). • Offensive language detection: using profanity-check to check if the comment contains any offensive words. • Incel identification: applying a threshold of -0.5 on the polarity score and a binary value of 1 on the offensive language detection to label a comment as incel or not. The paper evaluates the performance of the algorithm by comparing the results with the actual incel comments from Incels.co. It reports an accuracy of **78.8%** for incel detection, which is higher than the previous work in this field. **ALGORITHMS:** • Read user comments. The method takes the user comments from Facebook and Twitter as input data. • Pre-processing. The method removes any unnecessary data from the comments, such as punctuation marks, symbols, links, numbers, and spaces. • Sentiment analysis. The method uses TextBlob to perform sentiment analysis on the pre-processed comments and assigns a polarity score to each comment, ranging from -1 (negative) to 1 (positive). • Profanity check. The method uses profanity-check to detect any offensive language in the comments and assigns a binary value to each comment, 0 for non-offensive and 1 for offensive.

				Incel detection. The method applies a threshold to the polarity score and the profanity value to identify incels. If the polarity score is less than or equal to -0.5 and the profanity value is 1, the comment is classified as an incel. Otherwise, the comment is classified as a non-incel. **ADVANTAGES:** The paper claims that the proposed method has several advantages, such as: it is the first method to use user comments to identify incels in social media; it outperforms the previous method that used like and fuzzy like data by 10.05% in accuracy; it can help social media platforms and researchers to detect and prevent incel-related problems and violence; and it can be applied to different languages and social media platforms with minor modifications. **DISADVANTAGES:** The paper also acknowledges some limitations and challenges of the proposed method, such as: it relies on the quality and quantity of the collected data, which may be noisy, incomplete, or biased; it may generate false positives or false negatives due to the ambiguity and subjectivity of sentiment analysis and offensive language detection; it may not capture the complex and dynamic nature of incel communities and their interactions; and it may raise ethical and privacy issues regarding the use of user-generated data for incel detection.

				APPLICATIONS: The paper suggests some possible applications of the proposed method, such as: providing early warning and intervention for potential incel-related threats and attacks; developing educational and psychological programs to help incels cope with their problems and improve their social skills; analyzing the incel discourse and behavior to understand their motivations, beliefs, and emotions; and designing gamification and social network analysis techniques to encourage incels to participate in more constructive and positive activities and communities. **CONCLUSION:** The paper concludes that the proposed method is a novel and effective way to identify incels in social media using sentiment analysis and offensive language detection on user comments. The paper also states that the proposed method can be improved and extended by using other machine learning and social network analysis methods, and by considering other media types such as images, videos, and audio files. The paper hopes that the proposed method can contribute to the prevention and reduction of incel-related problems and violence in society.
7.	Emotional sentiment analysis of social media content for mental health safety	Ferdaous Benrouba, Rachid Boudour	2022	**METHODOLOGY AND ALGORITHM:** • The paper proposes an approach to filter the social media content that could be emotionally harmful to the user, based on the analysis and

				classification of the emotions expressed in the text. • The paper uses Twitter API to collect user posts, and then uses IBM natural language understanding API tool to extract and classify the emotions of the Twitter content into five basic emotional categories: joy, sadness, anger, fear, and disgust. • The paper defines a perfect emotion array from 450 words from the English language that are commonly used to express positive emotions, and calculates the Euclidean distance between the emotions array of each tweet content and the perfect emotions array. • The paper sets a threshold value of 1.08, and if the distance is equal to or less than the threshold, the content is considered emotionally safe and can be displayed. Otherwise, a warning message is displayed that the content could be emotionally harmful. **ADVANTAGES:** The paper provides a novel and practical solution to protect the mental health of social media users, by filtering out the content that could trigger negative emotions or affect their well-being. The paper also uses open-source and accessible tools to perform the emotion analysis and classification, which makes the approach easy to implement and replicate. **DISADVANTAGES:**

<table>
<tr><td></td><td></td><td></td><td></td><td>

The paper relies on a fixed and predefined set of words to represent the perfect emotion array, which may not capture the diversity and complexity of human emotions and expressions.
The paper also does not consider other types of content, such as images and videos, that could also be emotionally harmful.
The paper also does not evaluate the accuracy and effectiveness of the proposed approach on a large and diverse dataset of tweets.

APPLICATIONS:

The paper can be applied to create an emotion filter for social media platforms, such as Twitter, Facebook, Instagram, etc., that can help users avoid or reduce exposure to emotionally harmful content and enhance their online experience and mental health.
The paper can also be applied to other domains and scenarios where emotion analysis and classification are needed, such as online customer reviews, product recommendations, sentiment analysis, emotion detection, etc.

CONCLUSION:
The paper presents an approach to filter the social media content that could be emotionally harmful to the user, by analyzing and classifying the emotions of
</td></tr>
</table>

				the text using IBM natural language understanding API tool, and comparing them with a perfect emotion array using Euclidean distance. The paper demonstrates the feasibility and usefulness of the proposed approach, and suggests some future directions for improvement and extension.
8.	A Survey of Sentiment Analysis from Social Media Data	Koyel Chakraborty , Siddhartha Bhattacharyya , Rajib Bag	2020	**METHODOLOGY:** Sentiment analysis: Sentiment analysis, also known as opinion mining, aims to determine the sentiment or emotional tone expressed in a piece of text. Here's a detailed exploration: • Data Collection: Gather relevant text data from various sources such as social media posts, reviews, blogs, and news articles. Preprocess the data by removing noise (e.g., special characters, URLs, and irrelevant content). • Feature Extraction: Convert the text into numerical features that machine learning algorithms can process. • Common techniques include: Bag-of-Words (BoW): Represent each document as a vector of word frequencies. TF-IDF (Term Frequency-Inverse Document Frequency): Weigh words based on their importance in the document and across the entire corpus. • Supervised Learning: Train a machine learning

				model using labeled data (where sentiments are known). Common algorithms include: Naive Bayes: Assumes independence between features. Support Vector Machines (SVM): Finds a hyperplane to separate positive and negative examples. Logistic Regression: Predicts probabilities of sentiment classes. Random Forests: Ensemble of decision trees. • Unsupervised Learning: Discover patterns and clusters in the data without labeled examples. Techniques include: Lexicon-Based Analysis: Use sentiment lexicons (lists of positive and negative words) to score documents. Topic Modeling (e.g., Latent Dirichlet Allocation): Identify underlying topics in the text. • Hybrid Approaches: Combine supervised and unsupervised methods for better accuracy. For instance, use lexicon-based sentiment scores as additional features in a supervised model. **ALGORITHMS:** • Maximum Entropy (MaxEnt) Classifier: - A probabilistic model that assigns probabilities to different classes based on features. - Widely used in sentiment analysis due to its flexibility and ability to handle complex

				relationships. • VADER (Valence Aware Dictionary and sentiment Reasoner): - A lexicon-based approach specifically designed for social media text. - Assigns sentiment scores to words and aggregates them to compute overall sentiment. • Deep Learning Models: - Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks: - Process sequences of words and capture context. - Effective for sentiment analysis but require large amounts of labeled data. - Transformer-based Models (e.g., BERT, GPT): - Pretrained on massive text corpora. - Fine-tuned for specific tasks, including sentiment analysis. **ADVANTAGES:** • Scalability: Can analyze large volumes of text quickly. • Real-Time Insights: Useful for monitoring social media and customer feedback. • Multilingual Support: Works across languages. • Automated Decision-Making: Helps businesses make informed choices based on sentiment trends. **DISADVANTAGES:** • Contextual Challenges: Sarcasm, irony, and context-dependent sentiments can be tricky. • Data Bias: Models may inherit biases present in training data.

				- Privacy Concerns: Handling user-generated content requires ethical considerations. - Domain-Specific Adaptation: Models may not generalize well to specialized domains. **CONCLUSION:** Sentiment analysis is a dynamic field with ongoing research. Combining various methodologies and algorithms can lead to more accurate results. Future directions include addressing bias, improving context understanding, and exploring multimodal approaches.
9.	Socio-Analyzer: A Sentiment Analysis Using Social Media Data	Ajay Bandi and Aziz Fellah	2019	**METHODOLOGY:** - The paper aims to develop a Socio-Analyzer that can analyze the sentiment of social media data related to social movements, using the MeToo movement as a case study. - The paper follows a four-phase approach to implement the Socio-Analyzer: 1) Data Collection, 2) Preprocessing and Storage, 3) Sentiment Analysis, and 4) Data Validation. - In the Data Collection phase, the paper collects two types of data from Twitter: static data and stream data. Static data is retrieved from the data world website, while stream data is scraped periodically using the Python-Tweepy library.

				The input for the data collection is the hashtag #MeToo and the date range. The output is a tweet object that contains the tweet ID, date, and text. The paper collects 393,869 tweets from October 2017 to February 2018. • In the Preprocessing and Storage phase, the paper uses the Python-re library to clean the tweet object by removing emoticons, URLs, and hashtags. The cleaned tweets are stored in an Excel file that contains the tweet ID, date, and text. • In the Sentiment Analysis phase, the paper develops the Socio-Analyzer using Natural Language Processing (NLP) techniques with the Python-NLTK library. The Socio-Analyzer takes the Excel file as input and outputs the sentiment count of tweets in three categories: positive, neutral, and negative. • The Socio-Analyzer performs the following steps to analyze the sentiment of each tweet: - Tokenize the tweet text by splitting it into words based on space as a delimiter. - Remove the stop words, which are common words that do not convey much meaning, such as "the", "is", "a", etc. - Stem the words, which means reducing them to their root form, such as "fishing" to "fish", "beautiful" to "beauti", etc. - Compare each word with a

				sentiment dictionary, which is a list of words with positive or negative polarity scores, and increment the positive or negative count accordingly. If the word is not in the dictionary, it is considered neutral. - In the Data Validation phase, the paper compares the results of the Socio-Analyzer with the results of TextBlob, which is another sentiment analysis tool. The paper uses a weather dataset of 765 tweets as a benchmark, and manually labels them as positive, neutral, or negative. The paper calculates the precision of both tools, which is the ratio of true positives to the total number of predicted positives. The paper reports that the precision values of Socio-Analyzer and TextBlob are 70.74% and 72.92%, respectively, when considering neutral tweets as positive, and 68.94% and 65.82%, respectively, when considering neutral tweets as negative. **ADVANTAGES:** • Real-Time Insights: The Socio-Analyzer provides real-time sentiment analysis of social media content, allowing organizations and individuals to monitor public opinion promptly. • Cost-Effective: Leveraging existing NLP libraries and APIs (such as NLTK and IBM NLU) makes the approach cost-effective and accessible. • Customizable: The methodology allows customization by adjusting the emotion dictionary or incorporating domain-specific lexicons. • Multilingual Support:

				The approach can be extended to analyze sentiments in multiple languages. **DISADVANTAGES:** • Fixed Lexicon Limitations: Relying on a predefined set of words for sentiment analysis may not capture nuanced emotions or context-specific expressions. • Data Bias: The approach inherits biases present in the sentiment dictionary, affecting accuracy.Complex Emotions: Difficulty in handling complex emotions like sarcasm, irony, and cultural nuances. • Limited to Text: The method does not consider other media types (images, videos) that convey emotions. **APPLICATIONS:** • Brand Reputation Management: Organizations can use sentiment analysis to monitor their brand perception and respond to negative sentiment promptly. • Crisis Detection: Early detection of negative sentiment can help prevent or mitigate crises. • Market Research: Understand consumer opinions, preferences, and trends. • Healthcare: Monitor patient sentiments in online health forums. **CONCLUSION:** The paper demonstrates a

				practical approach to sentiment analysis using social media data. While it has limitations, such as fixed lexicons and data bias, it provides valuable insights for various applications. Future research should focus on improving accuracy, handling complex emotions, and expanding to multimodal content.
10.	Aspect-level Sentiment Analysis for Social Media Data in the Political Domain using Hierarchical Attention and Position Embeddings	Renny Pradina Kusumawardani, Muhammad Wildan Maulidani		**METHODOLOGY AND ALGORITHMS:** • The paper proposes a deep learning architecture called hierarchical attention position-aware network (HAPN) for aspect-level sentiment analysis on social media data in the political domain. • HAPN consists of four major parts: input embeddings, sentence encoder, fusion layer, and output layer. • The input embeddings layer uses both word embeddings and position embeddings to represent the words and their distances from the target aspect in a sentence. • The sentence encoder layer uses a bidirectional recurrent neural network (RNN) to produce an abstract representation of the sentence based on the input embeddings. The paper experiments with different RNN variants, such as simple RNN, GRU, and LSTM. • The fusion layer uses two types of **attention mechanisms**: **source2aspect attention** and **source2context attention**. The former

				captures the relation between the target aspect and the words in the sentence, while the latter captures the most indicative words in the context. The fusion layer produces a final representation of the sentence based on a weighted combination of the hidden states from the RNN layer. • The output layer uses a **softmax function** to predict the sentiment polarity of the target aspect based on the final representation from the fusion layer. The paper uses three sentiment classes: **positive, negative, and neutral**. **ADVANTAGES:** The model can capture the sentiment polarity of specific aspects in a text, which may vary depending on the position and context of the words. The model also uses a trainable word embedding that is adapted to the social media and political domain, which improves the performance and coverage of the vocabulary. **DISADVANTAGES:** • The model requires a large amount of annotated data for training, which may not be available for other domains or languages. • The model also does not handle the cases where the sentiment polarity is implicit, neutral, or mixed, which are common in political opinions.

				APPLICATIONS: • The model can be used for analyzing the public opinion on political topics, such as candidates, parties, issues, or policies, based on social media data. • The model can also be applied to other domains that involve aspect-based sentiment analysis, such as product reviews, customer feedback, or news articles. **EXAMPLES:** • The paper provides some examples of the model's predictions on tweets that contain different aspects and sentiments, such as "Infrastructure development is now very advanced and good, but unfortunately the country's finances are bad because of continuing debt". • The paper also shows the attention weights assigned to each word by the model, which indicate the importance of the word for the sentiment prediction. **CONCLUSION:** The paper concludes that the hierarchical attention with position embeddings is an effective architecture for aspect-level sentiment analysis on social media data in the political domain. The paper also suggests some future directions for improving the model, such as incorporating external knowledge, handling negation and sarcasm, and using multi-task learning.

11.	Deep Learning for Automated Sentiment Analysis of Social Media	Li-Chen Cheng, Song-Lin Tsai	2019	**METHODOLOGY AND ALGORITHMS:** • Web Crawler: The paper employs a custom Python web crawler to collect review data from various online sources, with a focus on movie reviews. The crawler navigates through websites, extracts relevant content, and compiles a dataset for further analysis. • Preprocessing: - Before feeding the data to the sentiment analysis models, the paper performs preprocessing steps: -Tokenization: The text is split into individual words or tokens. -Stop Word Removal: Common words (such as "the," "and," "is") are removed to reduce noise. - Lemmatization/Stemming: Words are transformed to their base forms (lemmas) or stems. - Handling Special Characters: Emoticons, slang, abbreviations, and repeated letters are addressed. - Sentiment Labeling: Each sentence is labeled as positive or negative based on sentiment scores from tools like NLTK, Textblob, and Google Cloud Natural Language API. • Word Embedding: - Word embedding techniques are used to represent words as dense vectors in a continuous space. Two popular methods are: - Word2Vec: It learns word representations by predicting context words given a target word. - GloVe (Global Vectors for Word Representation): It constructs word vectors based on global word co-occurrence statistics. - These embeddings capture

				semantic similarity and context, enabling the models to understand relationships between words. • Sentiment Analysis Models: - The paper proposes three deep learning models: - Long Short-Term Memory (LSTM) - An RNN variant that captures long-range dependencies in sequential data. - Learns patterns in the text by maintaining hidden states over time. - Bidirectional LSTM (BiLSTM): - Enhances LSTM by processing input sequences in both forward and backward directions. - Captures context from both past and future words. - Gated Recurrent Unit (GRU): - Similar to LSTM but with fewer parameters. - Efficiently captures sequential information. - These models learn to classify reviews as positive or negative based on the learned representations. • Evaluation: - The paper evaluates model performance using a dataset of 6000 sentences (3000 positive and 3000 negative) crawled from YouTube. - Metrics include accuracy, precision, recall, specificity, and F1 score. - BiLSTM achieves the best results among the three models, demonstrating its effectiveness in sentiment classification. **EXAMPLES:** The paper gives some examples of social media language, such as

				LOOOL, gr8, and emoticons, and how they are processed by the proposed framework. It also shows some examples of sentences labeled as positive or negative by different sentiment analysis tools. **ADVANTAGES:** The paper claims that the proposed framework has some advantages over traditional natural language processing methods, such as handling slang and special terms, using word embeddings to represent text, and applying deep learning models to classify sentiment. **DISADVANTAGES:** Requiring a large and balanced dataset for training, being sensitive to noise and ambiguity in social media text, and lacking explainability of the deep learning models. **APPLICATIONS:** The paper suggests that the proposed framework can be used to analyze the sentiment of user-generated content on social media sites, such as movie reviews, product feedback, and customer opinions. The paper also states that the extracted information will be useful for many future applications, such as marketing, recommendation, and decision making. **CONCLUSION:** The paper concludes that sentiment analysis is a useful tool to mine useful insights from social media data, but it faces

				some challenges due to the complexity and dynamic nature of social media language[1][1]. The paper proposes a deep-learning based framework to overcome these challenges and achieve high accuracy in sentiment classification. The paper also collects a dataset of social media reviews for future research.
12.	Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data	George Manias, Argyro Mavrogiorgou, Athanasios Kiourtis, Chrysostomos Symvoulidis, Dimosthenis Kyriazis	2023	**METHODLOGY:** • The paper uses two different approaches for multilingual text classification: BERT-based multilingual classification (BERT-MC) and zero-shot sentiment classification (ZSSC). • BERT-MC uses four different multilingual BERT-based models: mBERT (cased and uncased), XLM-RoBERTa, and DistilmBERT[1][1]. These models are fine-tuned on a multilingual Amazon reviews corpus (MARC) for two tasks: multilingual text categorization and multilingual sentiment analysis. • ZSSC uses a pre-trained XLM-RoBERTa model that is trained on a large cross-lingual natural language inference dataset (XNLI). This model is used to perform zero-shot classification by using natural language descriptions of the target classes as input. • The paper compares the performance of the two approaches on the MARC dataset and on a real-world multilingual Twitter dataset that contains tweets related to

				wine products in six languages. The paper evaluates the accuracy, precision, recall, and F1-score of the models on both tasks and datasets. **ALGORITHMS:** • BERT-MC stands for BERT-based multilingual classification. It is an approach that fine-tunes four different multilingual BERT-based models on a multilingual Amazon reviews corpus (MARC) for two tasks: multilingual text categorization and multilingual sentiment analysis. The four models are: • mBERT (cased and uncased): A multilingual version of BERT that supports 104 languages and was trained on Wikipedia and BooksCorpus. • XLM-RoBERTa: A multilingual version of RoBERTa that supports 100 languages and was trained on Common Crawl. • DistilmBERT: A smaller and faster version of mBERT that retains most of its performance. • ZSSC stands for zero-shot sentiment classification. It is an approach that uses a pre-trained XLM-RoBERTa model that was trained on a large cross-lingual natural language inference dataset (XNLI). This model is used to perform zero-shot classification by using natural language descriptions of the target classes as input. For

				example, to classify a sentence as positive or negative, the model would take the sentence and the descriptions "This text expresses a positive opinion" and "This text expresses a negative opinion" as input, and output the most likely class. The paper compares the performance of the two approaches on the MARC dataset and on a real-world multilingual Twitter dataset that contains tweets related to wine products in six languages. The paper evaluates the accuracy, precision, recall, and F1-score of the models on both tasks and datasets. The paper finds that BERT-MC outperforms ZSSC on both datasets and tasks, but ZSSC has the advantage of being more efficient and scalable, as it does not require any fine-tuning or labeled data. **ADVANTAGES:** • The paper leverages the state-of-the-art multilingual BERT models and evaluates them on two different tasks: multilingual text categorization and multilingual sentiment analysis. • The paper also explores the use of a zero-shot classification approach that does not require any training data and can be applied to any language and task. • The paper performs experiments on both a large multilingual corpus of Amazon reviews and a real-world dataset of multilingual tweets related to wine

				products. **DISADVANTAGES:** • The paper does not compare the multilingual BERT models with other multilingual models or methods, such as LASER or MultiFit, that have been proposed for multilingual text classification. • The paper does not provide a detailed analysis of the performance of the models across different languages and domains, and how they handle linguistic variations and challenges. • The paper does not address the ethical and social implications of using multilingual text classification for public policy making and other sensitive applications. **APPLICATIONS:** • The paper demonstrates the potential of multilingual text classification for various domains and scenarios, such as health care, policy making, digital marketing, and social media analysis. • The paper shows how multilingual text classification can help extract valuable knowledge and insights from user-generated data in multiple languages, such as reviews, tweets, comments, etc. • The paper suggests that multilingual text classification can enable cross-lingual transfer

				learning and inference, and facilitate the development of language-agnostic and scalable solutions. **CONCLUSION:** • The paper concludes that multilingual BERT models achieve high performance and transfer inference when trained and fine-tuned on multilingual data, and outperform the zero-shot classification approach in terms of accuracy. • The paper also concludes that the zero-shot classification approach presents a novel technique for creating multilingual solutions in a faster, more efficient, and scalable way, but it may not be comparable to fine-tuned multilingual BERT models when accuracy is more important than efficiency. • The paper highlights the need for further research and evaluation of multilingual text classification approaches on different domains, tasks, and languages, and the challenges and opportunities that they pose for the NLP field.