Tec Empresarial
P-ISSN: 1659-2395; E-ISSN: 1659-3359

COMPARATIVE ANALYSIS OF TEXT MINING BASED SENTIMENT ANALYSIS MODELS WITH EMOTION DETECTION

Prof. Himashri Purohit

Department of Computer Science, Shree Om College of Computer Science

Email: purohithm@gmail.com

Dr. Ishaan Tamhankar

Department of Information Technology, Surendranagar University

Email : prof.ishaantamhankar@gmail.com

ABSTRACT

The Internet has a significant usage of the applications which generate text based data on a regular basis. This data is unstructured which is not directly available in the form of a tabular representation. From the content of emails to the facebook posts, whatsapp messages to feedback messages on e-commerce website, all these text contents form unstructured data on a regular basis. As the data is unstructured, it is indeed a challenging task to analyze it systematically to get meaningful information. Text mining is an emerging field of computer science, which performs analysis of text data in a systematic manner to get insights and to derive some meaningful information. One use of text mining is to detect sentiment or emotion of the writer from his / her text content. This would help in analyzing different aspect of the content writers – from their states of minds to their satisfaction levels. This research work is based on analyzing performances of some most widely used sentiment analysis models with emotion detection.

Keywords – Corpus, Sentiment Analysis, Text Mining

I. INTRODUCTION

Data mining is the process of extracting valuable patterns and information from large datasets. It involves using various techniques such as machine learning, statistical analysis, and pattern recognition to uncover hidden insights. By analyzing structured data like databases or spreadsheets, data mining helps businesses and organizations make informed decisions, predict future trends, and identify opportunities for improvement. It plays a crucial role in industries like marketing, finance, healthcare, and manufacturing, where data-driven decisions can lead to significant gains in efficiency and profitability [1,2,3,4].

Text mining, on the other hand, focuses specifically on extracting meaningful information from unstructured text data. With the proliferation of digital content such as emails, social media posts, articles, and customer reviews, text mining has become increasingly important. It employs natural language processing (NLP) techniques to analyze and interpret textual data, identifying patterns, sentiments, and trends that might otherwise go unnoticed. Text mining enables businesses to gain valuable insights from vast amounts of unstructured text, helping them understand customer feedback, monitor brand reputation, and extract valuable knowledge from textual sources [2,3,4,5].

In today's era, text mining holds particular significance due to the exponential growth of unstructured textual data on the internet and other digital platforms. Traditional methods of analysis struggle to handle the sheer volume and complexity of this data. Text mining offers a solution by automating the process of extracting insights from text, making it easier and faster for businesses to derive value from their unstructured data. With the ability to analyze sentiments, extract key information, and detect emerging trends, text mining empowers organizations to stay competitive in a rapidly evolving digital landscape, where understanding and leveraging textual data is essential for success [1,2,3,4,5].

II. TEXT MINING ALGORITHMS

There are different types of algorithms are developed for various types of activities [2,3,4,5].

Sentiment Analysis Algorithms: These algorithms figure out if a piece of writing feels happy, sad, or neutral. They use rules or learn from examples to understand how people feel from what they write.

Text Classification Algorithms: These algorithms organize pieces of writing into different groups, like sorting emails into "important" and "not important." They learn from examples to recognize patterns and put similar writings in the same category.

Topic Modeling Algorithms: These algorithms help find what a bunch of writings are mostly about, like figuring out if they talk about sports, food, or movies. They look for words that often appear together to understand the main ideas in the writings.

Named Entity Recognition (NER) Algorithms: These algorithms pick out special things in writings, like names of people, places, or dates. They use rules or learn from examples to find and classify these special things accurately.

Word Embedding Algorithms: These algorithms turn words into special codes that show their meanings based on how they are used in writings. They help computers understand words better by representing them as numbers in a way that captures their meanings.

Text Clustering Algorithms: These algorithms group similar pieces of writing together based on what they talk about. They use different methods to find patterns and put writings with similar content into the same clusters.

III. LITERATURE REVIEW

Over the years, many researchers have started using Text Mining to analyze their unstructured data to determine valuable insights. We have surveyed the recent developments happened over the years and summarized our review in this section.

Sr. No.	Title of Paper	Findings
1	Using text mining and sentiment analysis for online forums hotspot detection and forecast [6]	Text mining and unsupervised learning are used together to group the forums into various clusters, with the center of each representing a hotspot forum.
2	Text Mining: Sentiment analysis on news classification [7]	A model is built to evaluate the polarity of headlines of news from online reviews by the customers.
3	Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis [8]	A survey to differentiate and compare turkish tweets and English tweets related with Syrian refugee crisis.
4	How to predict explicit recommendations in online reviews using text mining and sentiment analysis [9]	A text mining method was applied to online reviews to identify drivers of explicit recommendations to be useful for the travellers.
5	Sentiment analysis and opinion mining [10]	A detailed discussion on various steps while using sentiment analysis and opinion mining along with the challenges and their solutions.
6	Text mining with sentiment analysis on seafarers' medical documents [11]	Analysis of medical documents in terms of patient records, doctor notes, and prescriptions to get knowledge of medical issues that often happened onboard seafarers was done.
7	Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach. [12]	A systematic way to review effectiveness and popularity of a MOOC course was performed along with deep learning based classification of participants based on their sentiments.
8	Text-based emotion detection: Advances, challenges, and opportunities[13]	A discussion on different types of datasets useful for emotion detection along with detection approaches is given.
9	A review on sentiment analysis and emotion detection from text [14]	A discussion on levels of sentiment analysis, various emotion models is given.

10	A review of different approaches for detecting emotion from text [15]	A discussion on existing approaches, models, datasets, lexicons, metrics and their limitations for emotion detection is given.
11	Emotion detection of textual data: An interdisciplinary survey [16]	A discussion on current literature of text based emotion detection and the psychological models associated with it is given.
12	AI based emotion detection for textual big data: techniques and contribution [17]	Text based emotion detection is explained using Artificial intelligence and big data analytics.
13	Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection [18]	This work discusses the shift of era from the text sentiment analysis to emotion detection and the challenges in these types of work.
14	Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions [19]	A set of techniques for machine learning based emotion detection are explained along with challenges and future directions.

IV. PROPOSED WORK

DataSet: The GoEmotions dataset [20] is formed with 58000 comments from Reddit. These comments are listed from 27 emotion categories or Neutral. For our comparative study, we have worked on a random subset of the dataset consisting of 1000 records.

We have used Orange Data Mining tool to perform comparative study and analysis of various available tools. Sentiment Analysis needs a Corpus – a set of input texts for which we want to predict the corresponding sentiment / emotion. Following lexicon-based sentiment analysis models are used for analysis.

1. Liu & Hu sentiment module from NLTK.

2. Vader sentiment module from NLTK.

3. Multilingual sentiment from the Data Science Lab.

4. SentiArt from Arthur Jacobs

A lexicon-based sentiment analysis model relies on predefined lists of words, often categorized as positive, negative, or neutral, to determine the sentiment of a piece of text. The model calculates the overall sentiment score by counting the occurrences of words from the lexicon in the text and assigning corresponding sentiment scores to them.

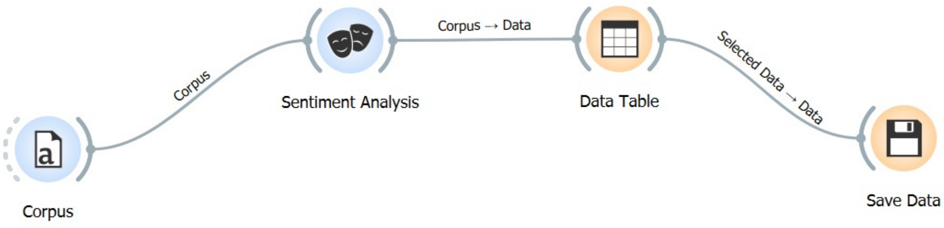


Figure – 1 – Flow of Sentiment Analysis in Orange

Figure (1) shows the flow of sentiment analysis with Orange Data Mining tool. The process starts with arranging a corpus – a set of texts for which we want to sentiment analysis. The corpus is linked with Sentiment Analysis widget from where we need to select model to be used. The results of the semantic analysis can be viewed through Data Table widget where as Save Data widget can be used to save the analysis results.

Liu & Hu sentiment module calculates a sentiment score. This module assigns a score to a given text based on the presence of positive and negative words. The score is calculated by counting the occurrences of positive and negative words in the text and then applying a formula to determine the overall sentiment score. Positive words increase the score, while negative words decrease it. The resulting score can indicate the overall sentiment of the text, with higher scores suggesting a more positive sentiment and lower scores suggesting a more negative sentiment.

The VADER sentiment module from NLTK provides not just a single sentiment score but a breakdown into four scores:

Positive: The proportion of the text that falls into the positive sentiment category.

Negative: The proportion of the text that falls into the negative sentiment category.

Neutral: The proportion of the text that falls into the neutral sentiment category.

Compound: A compound score that summarizes the overall sentiment of the text. This score ranges from -1 (extremely negative) to 1 (extremely positive). It combines the scores of positive, negative, and neutral sentiments, taking into account both the intensity and the polarity of the sentiment words.

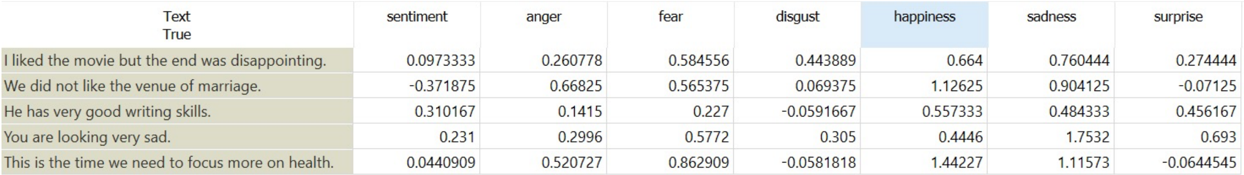
Result of SentiArt Module

In the Multilingual Sentiment Analysis tool from the Data Science Lab, the sentiment score likely represents the intensity or polarity of sentiment expressed in the text. This score can range from negative to positive, with values closer to negative indicating a more negative sentiment and values closer to positive indicating a more positive sentiment.

In SentiArt, along with the sentiment score, emotion wise score will also be provided. Following emotions are identified – anger, fear, disgust, happiness, sadness, and surprise. Each emotion-wise score represents the intensity or likelihood of the corresponding emotion being expressed in the input text. These scores can provide more detailed insights into the emotional content of the text beyond just sentiment analysis.

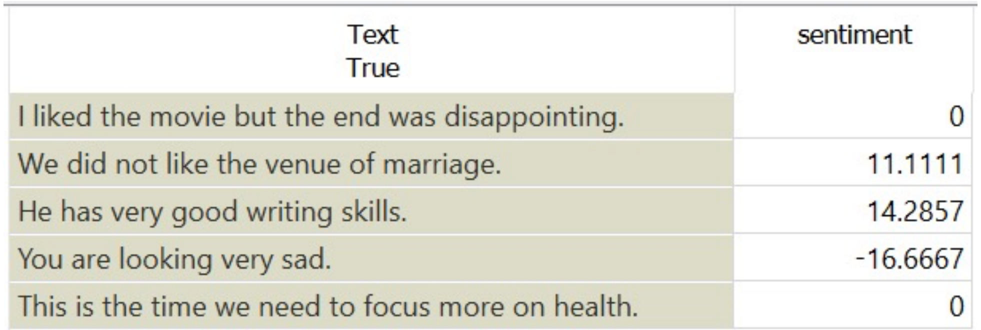
To discuss the input-output mapping, we are presenting examples of 5 texts and how each of the above mentioned module performs sentiment analysis with emotion detection.

For simplicity of discussion the five sentences shown here are very simple without any ambiguity or grammatical mistake. The purpose is to understand how text mining models react to the given sentences. Based on referring to the result we can easily identify that the predictions seem to be confusing at times. These approaches need to be formally evaluated for larger amount of data to find accuracy.



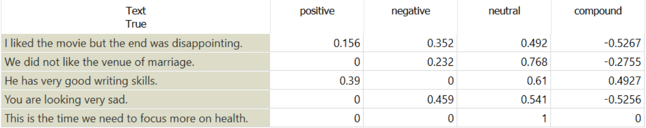
Text True	sentiment	anger	fear	disgust	happiness	sadness	surprise
I liked the movie but the end was disappointing.	0.0973333	0.260778	0.584556	0.443889	0.664	0.760444	0.274444
We did not like the venue of marriage.	-0.371875	0.66825	0.565375	0.069375	1.12625	0.904125	-0.07125
He has very good writing skills.	0.310167	0.1415	0.227	-0.0591667	0.557333	0.484333	0.456167
You are looking very sad.	0.231	0.2996	0.5772	0.305	0.4446	1.7532	0.693
This is the time we need to focus more on health.	0.0440909	0.520727	0.862909	-0.0581818	1.44227	1.11573	-0.0644545

Result of Liu & Hu Sentiment Module



Text True	sentiment
I liked the movie but the end was disappointing.	0
We did not like the venue of marriage.	11.1111
He has very good writing skills.	14.2857
You are looking very sad.	-16.6667
This is the time we need to focus more on health.	0

Result of Vader Sentiment Module



Text True	positive	negative	neutral	compound
I liked the movie but the end was disappointing.	0.156	0.352	0.492	-0.5267
We did not like the venue of marriage.	0	0.232	0.768	-0.2755
He has very good writing skills.	0.39	0	0.61	0.4927
You are looking very sad.	0	0.459	0.541	-0.5256
This is the time we need to focus more on health.	0	0	1	0

Result of Multilingual Sentiment

Performance Evaluation: We have performed several tests to check performances of various methods. Here we are sharing the results of one analysis where we selected 50 sentences of Happy mood. We applied there four modules to check what they predict.

Total Number of Sentences with Happy Mood: 50

Sr. No.	Module	Findings
1	Liu & Hu sentiment module from NLTK.	45 Sentences got positive sentiment score.
2	Vader sentiment module from NLTK.	47 Sentences got positive sentiment score.
3	Multilingual sentiment from the Data Science Lab.	45 Sentences got positive sentiment score.
4	SentiArt from Arthur Jacobs	50 Sentences got positive sentiment score.

CONCLUSION

This research work is based on exploring how text mining methods can be used to perform sentiment analysis with specific to detect mood of given text. We have understood the systematic approach and implemented a workflow using Orange tool. Four most widely accepted modules are selected and evaluated on various available datasets. We have observed that all these four modules perform the best. But they differ in what type of other information is provided.

Liu & Hu sentiment module from NLTK and Multilingual sentiment from the Data Science Lab provide sentiment score only where as Vader sentiment module from NLTK and SentiArt from Arthur Jacobs be more specific in finding mood of the text also. We have observed that Liu & Hu sentiment module and Multilingual sentiment module provide the same results. We have observed that SentiArt from Arthur Jacobs module provide most accurate result with detailed analysis of various moods and their applicability for input texts.

FUTURE WORK

This research work can be further extended with analysis of more complicated real world datasets. We can try to find the accuracy of different modules based on different moods. We can also extend this work to perform evaluations by setting different parameters of modules and by visualizing the results.

REFERENCES

[1] Han, Jiawei, Micheline Kamber, and Data Mining. "Concepts and techniques." Morgan Kaufmann 340 (2006): 94104-3205.

[2] Chakrabarti, Soumen, et al. Data mining: know it all. Morgan Kaufmann, 2008.

[3] Zong, Chengqing, Rui Xia, and Jiajun Zhang. Text data mining. Vol. 711. Singapore: Springer, 2021.

[4] Jo, Taeho. "Text mining." Studies in Big Data 45 (2019).

[5] Gaikwad, Sonali Vijay, Archana Chaugule, and Pramod Patil. "Text mining methods and techniques." International Journal of Computer Applications 85.17 (2014).

[6] Li, Nan, and Desheng Dash Wu. "Using text mining and sentiment analysis for online forums hotspot detection and forecast." Decision support systems 48.2 (2010): 354-368.

[7] Gomes, Helder, Miguel de Castro Neto, and Roberto Henriques. "Text Mining: Sentiment analysis on news classification." 2013 8th Iberian Conference on Information Systems and Technologies (CISTI). IEEE, 2013.

[8] Öztürk, Nazan, and Serkan Ayvaz. "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis." Telematics and Informatics 35.1 (2018): 136-147.

[9] Guerreiro, Joao, and Paulo Rita. "How to predict explicit recommendations in online reviews using text mining and sentiment analysis." Journal of Hospitality and Tourism Management 43 (2020): 269-272.

[10] Liu, Bing. Sentiment analysis and opinion mining. Springer Nature, 2022.

[11] Chintalapudi, Nalini, et al. "Text mining with sentiment analysis on seafarers' medical documents." International Journal of Information Management Data Insights 1.1 (2021): 100005.

[12] Onan, Aytuğ. "Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach." Computer Applications in Engineering Education 29.3 (2021): 572-589.

[13] Acheampong, Francisca Adoma, Chen Wenyu, and Henry Nunoo-Mensah. "Text-based emotion detection: Advances, challenges, and opportunities." Engineering Reports 2.7 (2020): e12189.

[14] Nandwani, Pansy, and Rupali Verma. "A review on sentiment analysis and emotion detection from text." Social network analysis and mining 11.1 (2021): 81.

[15] Murthy, Ashritha R., and KM Anil Kumar. "A review of different approaches for detecting emotion from text." IOP Conference Series: Materials Science and Engineering. Vol. 1110. No. 1. IOP Publishing, 2021.

[16] Zad, Samira, et al. "Emotion detection of textual data: An interdisciplinary survey." 2021 IEEE World AI IoT Congress (AIIoT). IEEE, 2021.

[17] Kusal, Sheetal, et al. "AI based emotion detection for textual big data: techniques and contribution." Big Data and Cognitive Computing 5.3 (2021): 43.

[18] Hung, Lai Po, and Suraya Alias. "Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection." Journal of Advanced Computational Intelligence and Intelligent Informatics 27.1 (2023): 84-95.

[19] Alslaity, Alaa, and Rita Orji. "Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions." Behaviour & Information Technology 43.1 (2024): 139-164.

[20] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4040–4054, Online. Association for Computational Linguistics

Biographies and Photographs

Short biographies (120-150 words) should be provided that detail the authors' education and work histories as well as their research interests. The authors' names are italicized. Small (3.5 X 4.8 cm), black-and-white pictures/digitized images of the authors can be included.