Improved Opinion Mining for Unstructured Data Using Machine Learning Enabling Business Intelligence

Ruchi Sharma [1,*] and Pravin Shrinath [2]

[1] Department of Information Technology, Mukesh Patel School of Technology Management and Engineering (MPSTME), Narsee Monjee Institute of Management Studies (NMIMS) University, Mumbai, Maharashtra, India
[2] Department of Computer Engineering, Mukesh Patel School of Technology Management and Engineering (MPSTME), Narsee Monjee Institute of Management Studies (NMIMS) University, Mumbai, Maharashtra, India; Email: pravin.srinath@nmims.edu (P.S)
*Correspondence: ruchi.k6508@gmail.com (R.S.)

***Abstract*—There has been an exponential increase in usage of social informatics in recent years. This makes opinion mining more complex, especially for unstructured data available online. Although a substantial amount of research has been conducted on the COVID pandemic, post-pandemic research is lacking. Our research focuses on design and implementation of opinion mining framework for unstructured data input for business intelligence dealing with post pandemic work environment in industries. In this paper, we implement opinion mining algorithm in combination with machine learning approaches providing a hybrid approach. Transformer architecture Bidirectional Encoder Representations from Transformers language model is implemented to obtain sentence level feature vector of the document corpus and t-distributed stochastic neighbor embedding is implemented for clustering experimental evaluation. In this work, performance evaluation is undertaken using the Intertopic Distance map. By applying a hybrid strategy of natural language processing and machine learning, the results of this study indicate efficient framework development and anticipated to contribute to the improvement of efficacy of opinion mining models compared to existing approaches. This research is significant and will benefit businesses in gaining valuable insights that will lead to improved decision-making and business insights.**

***Keywords*—machine learning, deep learning, natural language processing, artificial intelligence, unstructured data, business intelligence**

I. Introduction

With the increasing influence of internet, users can browse web pages and post their own content. This content can take the shape of messages, images, videos, etc., and is referred to as unstructured data since it does not adhere to a particular standard. This makes the implementation of opinion mining of unstructured data rather challenging [1].

Opinion mining is central to the study of public opinion, feelings, and assessments of any social topic, person, or other institution [2, 3]. The definition of opinion or sentiment from [4] where it is represented as a quintuple (e_j, a_{jk},so_{ijkl}, h_i, t_i) where e_j: Target Entity, a_{jk}: Aspect of the entity, so_{ijkl}: Sentiment value in a granular form, h_i: Opinion holder, and t_i: Opinion Expression Time.

The main research gap identified is sparse data representation and dealing with non-colloquial language for trend analysis.

The following is an outline of the significant contributions made by our research:

(1) We have implemented a hybrid approach using machine learning and natural language processing that allows for more accurate opinion modelling.

(2) An algorithm for feature extraction is presented and implemented.

(3) We have developed curated dataset for the problem statement identified from unstructured data.

(4) Further, transformer architecture BERT (Bidirectional Encoder Representations from Transformers) is adapted and put into effect in order to obtain a sentence-level feature vector of the document corpus.

(5) A demonstration of the design and implementation of our opinion modelling framework and analysis technology is given. This research has significantly added to the body of knowledge in both theoretical and practical ways.

The subsequent is a summary of the research paper:

- A selection of high-quality articles was assessed with the goal to investigate the well-known opinion mining approaches used by researchers to enhance the performance of unstructured data.
- We developed an opinion mining framework that employs machine learning the natural language processing techniques: BOW (Bag of Word), TF-IDF (Term Frequency-Inverse Document Frequency), and BERT.

Manuscript received January 30, 2023; revised March 1, 2023; accepted April 13, 2023; published August 17, 2023.

- The proposed detection model was assessed on the developed corpus natural language processing based feature extraction techniques and machine learning based topic modelling techniques. Furthermore, this research investigates the output of developed opinion mining framework for novel research objective identified.

The remaining sections of the article are divided into the following sections: The second section is a description of previous related work and it evaluates a variety of Natural-Language Processing (NLP) strategies for opinion mining; In the third section, the approach that is proposed by this research is outlined; in the fourth section, the method of research experiments is outlined; in the fifth section, the results of the experiments are presented and discussed; and in the sixth section, the research is concluded.

II. Literature Review

Since opinions are becoming the important source of information for information retrieval, it is finding several areas of potential application sectors. Despite the fact that the Internet is a diverse source of opinions with blogs, forums, and social websites, unstructured text that's unable to be utilized directly for knowledge representation [5]. Quantitative analysis of large corpora is even more complex as discussed in research [6].

Fig. 1 illustrates the various applications of opinion mining in the fields of Banking, Industrial commerce [7], educational analytics, medical domain [8] and Business Intelligence [9], project management [10].

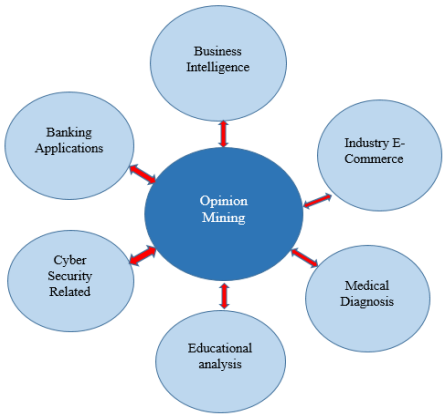


Figure 1. Opinion mining applications.

Mikolov *et al*. proposed the doc2vec paragraph to vector model, which is used by the auto encoder [11]. Semantic weighting can also be implemented for opinion mining using deep learning [4].

Latent Dirichlet Allocation (LDA) [12] is unsupervised method where each document represents a statistical distribution of topics [13] and each topic represents a statistical distribution of words, according to the central assumption. Machine Learning techniques [14] and deep learning techniques were applied in [15].

Sabuj *et al.* [16] have used Support Vector Machine (SVM) to mine views based on information gathered from the internet produced positive outcomes. Few studies have investigated effect of pandemic induced behavior on electronic businesses [17].

This research work focuses on trend analysis of opinion mining using machine learning and the highlights future directions.

III. Proposed Methodology and Research Design

The main steps of the current research approach are visualized in Fig. 2.

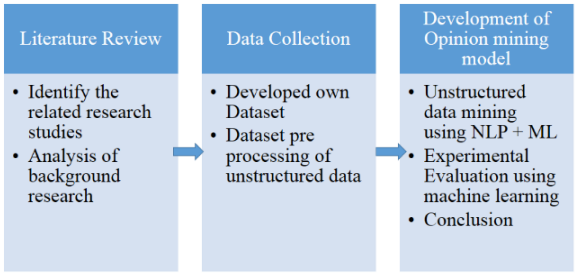


Figure 2. Proposed project methodology.

Although COVID 19 research has been done previously polarity detection [18] social platform analytics [19, 20], post COVID opinion mining has still not been researched extensively especially in context of its effect on business working model.

Proposed project methodology is described in detail as follows:

Step 1: We have conducted an analysis of the background research as well as the terminology.

Step 2: We have created our own dataset using Twitter API and Python 3.9. After the COVID 19 pandemic, key terms were selected to illustrate the working culture of businesses in the post-pandemic era. Hybrid work, hybrid working, remote work, work from home, and similar terms were used as key terms to comprehend the future of the workforce.

Step 3: For unstructured data, a feature extraction approach was used, which will be described in further depth in the following section.

Step 4: Following this, in Step 4, we carried out unstructured data mining using hybridization of NLP (natural language processing) and Machine learning. BERT, Bag of words (BOW) and Linear Discriminant Analysis (LDA) in hybridization with Term frequency inverse document frequency (TFIDF) method is implemented for completion of this phase.

Step 5: We have performed experimental evaluation using t-distributed stochastic neighbor embedding (TSNE) clustering

Step 6: The results and implications of the proposed opinion mining framework are interpreted in relation to the project objective.

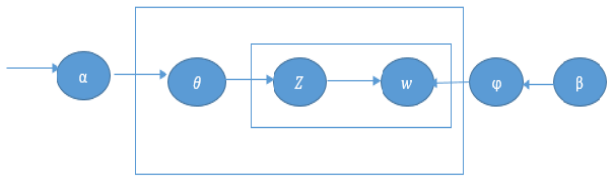


Figure 3. LDA methodology.

As depicted in Fig. 3 following steps are adopted for LDA.

Algorithm 1: LDA methodology

Notations:

α: Parameter Dirichlect distribution;

β: Topic Word Density

k: No of Latent topics in a distribution

D: Number of Documents in a collection

Z: Assignment of the topic

N: No of words in the individual Document provided

I: Index of the word w and the topic z

Input to LDA model: Corpus, Words w ϵ document D

Output: Topic Assignment implemented

Begin

1. Choose a topic for a text's probability distribution using $\theta \sim P(\theta|\alpha)$ where Θ is the topic distribution
2. Repeat the following procedure for each word in a document:

a) Choose a topic using $Z \sim P(z|\theta)$) which computes distribution of topic probability

b) Following equation is employed to pick A new term underneath the topic:

$$W \sim P((w|z), \phi)$$

3. In this manner, we generate a text with N words by executing step (2) zero up to N times. We generate a set of M texts by repeating the operation M times.

End

These opinions have undergone preprocessing procedures for conversion and handling of unstructured data to uncover hidden patterns from the data. The phases in preprocessing are:

Stemming: Words are generated from roots of the provided words.

Lemmatizing: Words are used in their dictionary forms.

Tokenizing: To aid in feature extraction, the original phrases will be divided into a number of tokens. Bigrams and trigrams are also regarded as one word.

Stop-word elimination: This step involves elimination of terms that are frequently used but are not necessary for opinion extraction. Numbers, propositions, and some other words devoid of any meaningful information are the most often used words.

Algorithm 2: Feature Engineering for Opinion mining

Input: Corpus (C), Query Terms

Output: Query, Expanded Query Terms

1: For ∀c; ∈C such that i ∈ (1,2... N)

1. l: Eliminate the tabs and the URLs

1 .2: Delete the digits present

1.3: Filter the initial review opinion in non-English language

2: Define punctuation constant p from string module

For each row r in Dataframe df:

a. For each column c with text data in row r:

i, Replace all punctuation marks in c with empty string and p constant:

c← replace (c, 'p',")

b. End for loop

End for loop

Save pre-processed df as new dataframe

4: Eliminate opinion word repetitions present

5: If word found is in Stop Word List

5.1 Eliminate corresponding stop words,

6: Add specific stop words and repeat the process in Step 5

7: Eliminate any opinion holders

8: for each row r in df:

for each col. consisting opinion O in corresponding r:

tokens = word_tokenize(O)

filtered tokens = {for token is ∉ stop words}

lemmas = [nlp(token) [0]. lemma

for token in filtered tokens]

add lemmas to new column in df

Write row to preprocessed_df

Return preprocessed Data Frame

9: Terminate upon reaching end of file

Return: Document Terms, Preprocessed Data

End

In order to determine the relationship among words and themes, there are constraints on the frequency of words.

Dataset Statistics:
Number of instances: 8922
Average length of text: 26.19379062990361
Vocabulary size: 16508

Sample from our corpus:
Original document:
['Women', 'who', 'work', 'for', 'se', 'companies', 'report', 'far', 'higher', 'levels', 'of', 'engagement', 'trust', 'career', 'satisfaction', 'as', 'well', 'as', 'more', 'positive', 'experiences', 'with', 'hybrid', 'working', 'lower', 'levels', 'of', 'burnout']

Tokenized and lemmatized document:

['women', 'work', 'compani', 'report', 'higher', 'level', 'engag', 'trust', 'career', 'satisfact', 'posit', 'experi', 'hybrid', 'work', 'lower', 'level', 'burnout']

TABLE I. DATASET STATISTICS

Sample Tweet ID	3126
Tokenized_column	[Remote, Work, Saturday, Reads? Lonely, Off…
Flesch reading	48
Smog index	81.8
Lexicon count	39
Syllable count	55
No of char	182
Polysyllable count	3
Mono syllable count	26
Difficult word ratio	0.15385

Table I discusses dataset statistics, provides statistical metrics for Tweet sample 3126. It is challenging to determine with accuracy the relationship underlying word and sentiment polarities when the topic is directly converted to attitude. Since there is less of a connection between words and emotions than there is between them and subjects when the words occur alongside directly.

A multinomial distribution of the corpus is used to select a topic. To select a term, the topic's multinomial distribution is used. The remaining variables are created from the global parameters, or "hyper parameters", and, which are defined for the whole corpus C. A document's probability density is specified; it is stated using Dirichlet's law and other features.

Among the top words listed, C is based on a window function and a pairwise comparison. Top words in comparison with an indirect comparison using normalized pointwise regression as a verification measure for the cosine similarity and Non-Parallel Mutual Information (NPMI).

IV. EXPERIMENTAL AND RESULTS

Discussion of experimental data, their interpretation, and experimental inferences are presented using machine learning and deep learning techniques.

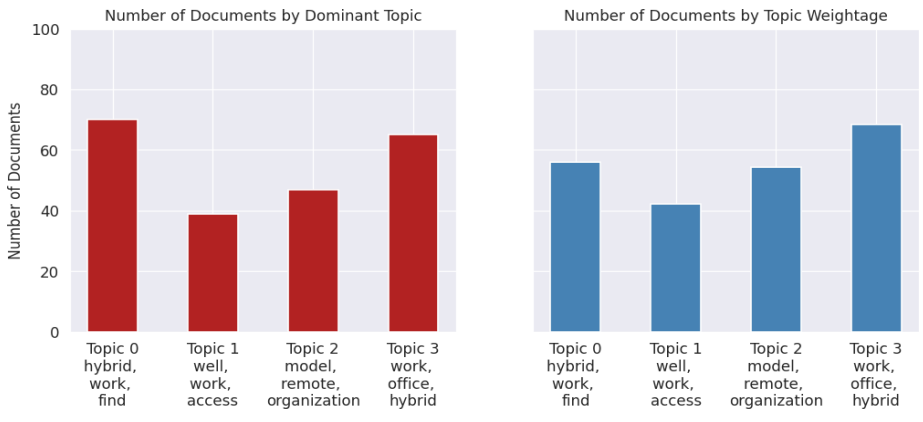
A. LDA + BOW Results

We incorporate Bag of Words (BOW) concept with LDA algorithm model concept to analyze opinion. We aim to analyze opinion related to work environment in 2022. This time frame was chosen since it is the post pandemic era wherein there was conflicting opinion related to unprecedented return to pre pandemic nature of work. We have developed an opinion mining framework from opinions on social media since they are of unstructured format and not having a formal grammatical structure.

TABLE II. RESULTANT TOPIC DISTRIBUTION AND CORRESPONDING TOPIC WEIGHTAGE

T_1	Topic 1: **Employee** Words: 0.060*"hybrid" + 0.030*"oper" + 0.018×"busi" + 0.015*"help" + 0.014*"technolog" + 0.013×"model" + 0.013*"learn" + 0.010*"latest" + 0.010*"analyst" + 0.010* "check"
T_2	Topic: 2: **Time**: Words: "job", "like", "nypost", "thing", "go", "want", "think", "peopl", "make", "time"
T_3	Topic: 3: **Socialization**: Words: "peopl", "live", "offic", "need", "think", "compani", "like", "great", "space", "social"
T_4	Topic: 4: **Flexibility**: Words: 0.025*"life" + 0.023*"support" + 0.018*"differ" + 0.016*"balanc" + 0.016*"year" + 0.015*"fice" + 0.015*"flexibl" + 0.015*"home" + 0.014*"way" + 0.013*"environ"
T_5	Topic: 5: **Hiring**: Words: 0.122*"hire" + 0.055*"engin" + 0.040*"senior" + 0.040*"remotework" + 0.039*"develop" + 0.027*"softwar" + 0.024*"manag" + 0.023*"remotejob" + 0.016*"appli" + 0.013*"lead"
T_6	Topic: 6: **Hybrid environment**: Words: 0.041*"offic" + 0.027*"meet" + 0.025*"hybrid" + 0.023*"year" + 0.019*"team" + 0.018*"time" + 0.018*"peopl" + 0.013*"love" + 0.013*"person" + 0.012*"hour"
T_7	Topic: 7: **Work from anywhere**: Words: 0.074*"remotework" + 0.033*"servic" + 0.032* "custom" + 0.021*"remotejob" + 0.018*"workanywher" + 0.016*"compani" + 0.014*"busi" + 0.014*"home" + 0.014*"today" + 0.013*"need"
T_8	Topic: 8: **Product Design**: Words: 0.042*"manag" + 0.026*"design" + 0.022*"market" + 0.021*"product" + 0.020*"hybrid" + 0.019*"remotework" + 0.018*"learn" + 0.018*"compani" + 0.016*"hire" + 0.015*"control"
T_9	Topic: 9: **Pay**: Words: 0.038*"video" + 0.035*"remotework" + 0.026*"home" + 0.021*"sale" + 0.019*"countri" + 0.018*"legal" + 0.016*"money" + 0.015*"peopl" + 0.013*"pay" + 0.012*"onlin"

Table II represents the resultant analysis of different topics distribution along with weightage assigned to the topic.



(a) (b)

Figure 4. Topic dominance.

Dominant topic by weightage assigned to topic and number of topics is considered as explained in Fig. 4.

We have also done Topic annotation for each cluster from T_1 to T_9.

B. LDA + TFIDF Performance Evaluation

The BOW approach disregards the relationship and inner semantics between words in sentences. Automatized Topic modeling using LDA and utilizing statistical methods-called term frequency-inverse document frequency (TF-IDF) in next phase.

Intertopic Distance Map, wherein the size of the spheres denotes the predominance and each circle represents a specific topic as illustrated in Fig. 5. The right-hand bar chart shows terms in order of decreasing significance and frequency.

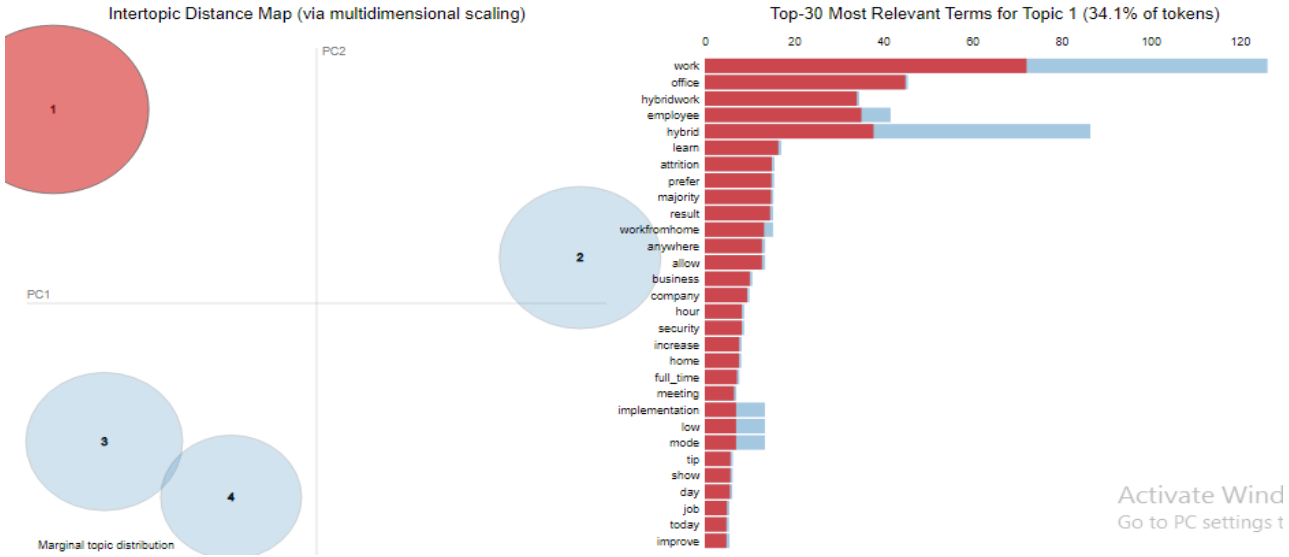


Figure 5. DA visualization.

The red section shows topic-specific term frequency, whereas the red and blue bars show corpus-wide term frequency considering top 30 relevant terms for Topic 1 from 34.1% of the tokens generated.

Evaluation of opinion mining model is demonstrated in Fig. 6. In order to determine the ideal number of topics, several models are constructed by varying the number of topic parameter.

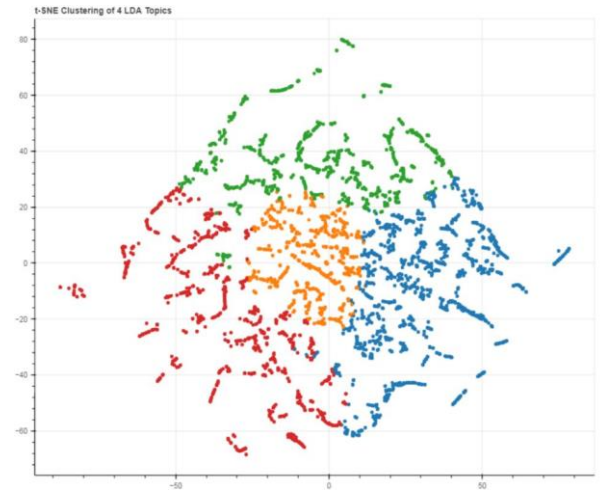


Figure 6. T-SNE clustering results.

Coherence score is computed using the variable topic count. The opinion model is developed and displayed after taking topic count into account.

As illustrated in Fig. 6, Clusters are distinct relatively in T-distributed Stochastic Neighbor Embedding (T-SNE) to each other which means model works well for the given scenario.

Additionally, here we have implemented perplexity metric for efficient model evaluation that allows model generalizability feature on unseen data instances as depicted in Table III.

TABLE III. PROPOSED OPINION MINING MODEL FRAMEWORK TOPIC EVALUATION

Model	Perplexity	Coherence
LDA + BOW	−6.39	0.65
Feature Extraction + LDA + TF-IDF (Proposed)	−7.736	0.72

Perplexity evaluates the ability of a topic model trained on a training set to identify a validation set.

$$\text{Perplexity} = -\frac{\log P(w'|\Phi,\alpha)}{N'} \quad (1)$$

where: w' : Unknown word instances,
Φ: Subsequent estimation of words,
N': Total (Words in w')

Lower the perplexity better the results indicating proposed model performs well on unseen instances.

Evidently, since the proposed model yields a lower perplexity result, it provides a better outcome than the current method.

Here, Coherence Score depicted in Fig. 7 to evaluate our opinion model is computed as:

$$\text{Coherence} = \sum_{i<j} score\ (w_i, w_j) \qquad (2)$$

where w_i, w_j are the topic's key terms.

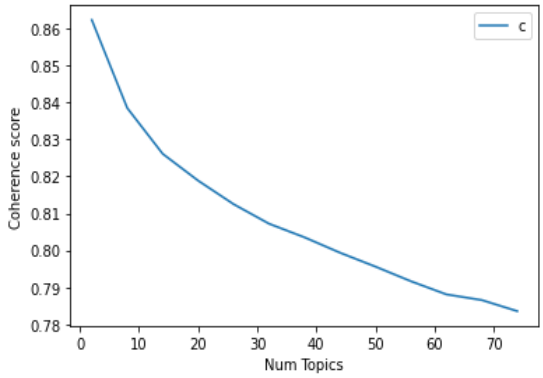
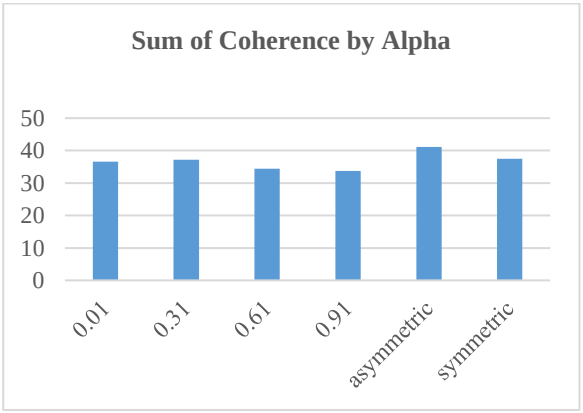
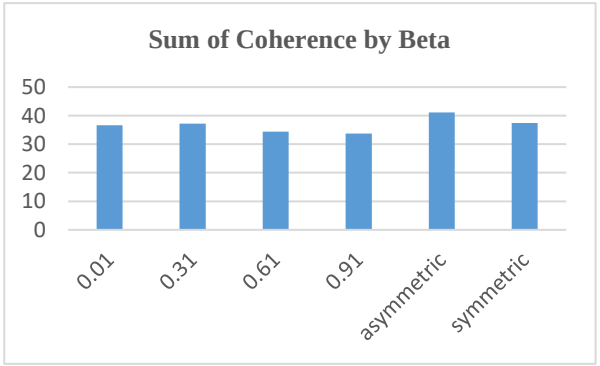


Figure 7. Coherence values of proposed model.

Figs. 8 and 9 depicts summation of coherence and Topic respectively considering alpha and beta hyper parameters of the model. This shows the term distributions in an interactive bar graph with horizontal bars and a 2D representation of our *n*-dimensional data.

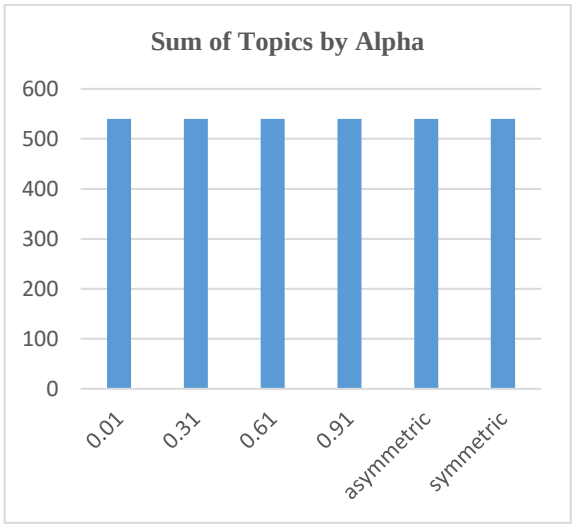


(a)

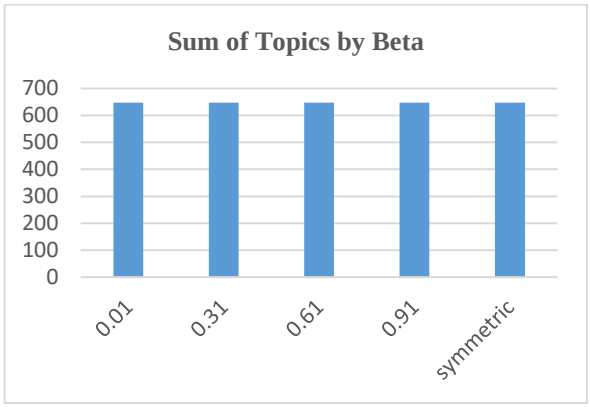


(b)

Figure 8. Summation of coherence by (a) Alpha and (b) Beta.



(a)



(b)

Figure 9. Summation of topic by (a) Alpha and (b) Beta.

V. Discussion

Research discussion and implications are described below:

A. Societal Implication

There are diverse set of opportunities created for prospective employees.

From the standpoint of the business, a larger workforce pool is accessible to choose from.

Moreover, many opinion posts in the dataset featured "Work from home" or "hybrid" positions in order to attract a larger candidate applicant base for the job position. Also Given the increasingly distant nature of business, a greater emphasis was placed on cybersecurity.

These are some of the conclusions drawn after using the proposed opinion mining approach.

B. Practical Implication and Research Significance

Typically, there is a communication gap between employers and potential hires.

- If business leaders are completely cognizant of the present trend in addition to the opinion of their employees, they could opt to focus on the same concerns and proceed to work for solutions in the same direction.

- Employing technology to gain a deeper understanding of each aspect of an industry, making data-driven recommendations based on that information thereby increasing overall productivity.
- Understanding employee requirements and conflicts, and gaining business insights all contribute to a higher level of productivity.
- Complete transparency in the decision-making process that can be adopted using AI technology in proposed opinion mining framework

Fig. 10 shows our implementation of BERT word scores for individual topics from previous phase.

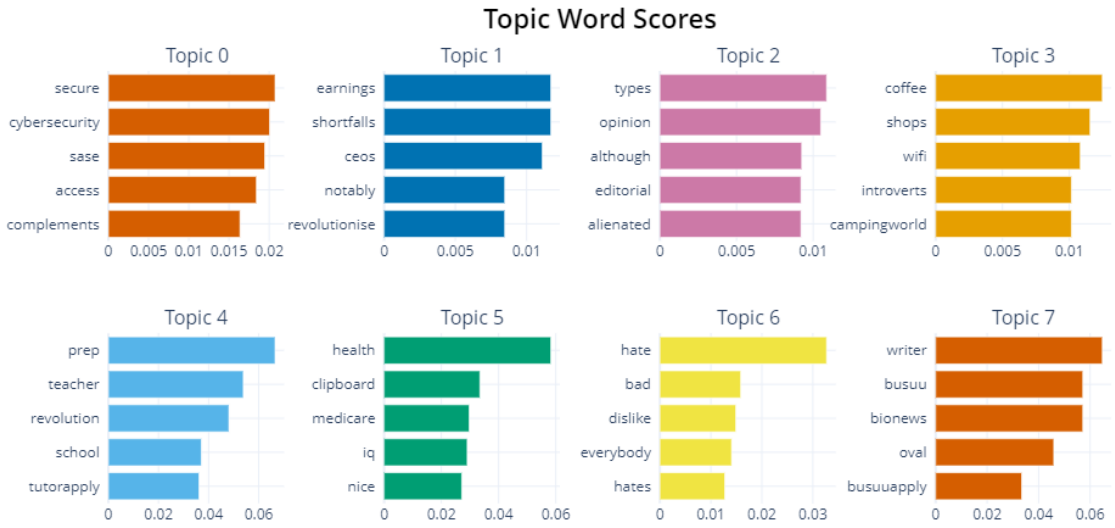


Figure 10. BERT topic word scores.

Performance Evaluation for testing is done using random unseen document, namely "what future of workforce will be?" is the query is inputted to our proposed opinion mining framework.

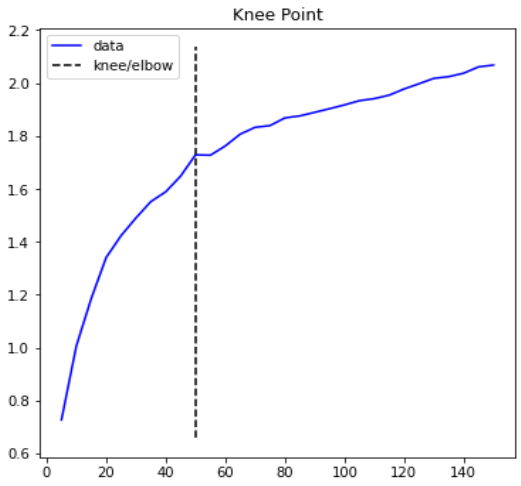


Figure 11. Knee point results.

The normalised curve's "knee" locations is calculated at the local maxima of the difference curve in Fig. 11
Result vector is described as:

Score: 0.6999387145042419

Topic: 0.060*"hybrid" + 0.030*"oper" + 0.018*"busi" + 0.015*"help" + 0.014*"technolog

Which implies that hybrid operations for businesses and technologies is going to be immediate future in the post pandemic era.

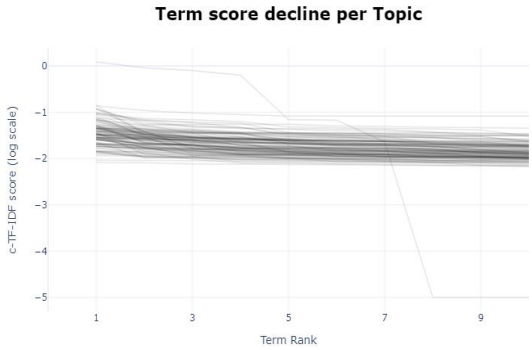


Figure 12. Term score decline topic wise.

In Fig. 12, Class-based TF-IDF (C–TFIDF) is implemented which is class based TFIDF demonstration C-TFIDF score with respect to the term rank.

Each class is converted into set of documents instead of single documents thus providing an efficient representation.

$$W_{x,c} = \|t_{x,c}\| \times \log\left(1 + \frac{Av}{f_w}\right) \quad (3)$$

tf_{xc} = frequency of word w in class c
f_w = frequency of word w across all classes
Av = average number of words per class

The aspect that makes c-TF-IDF stand out in comparison to TF-IDF is the fact that we can adapt it in such a way that we can search for words that make up particular classes, which offers both consistency and efficient speedup.

Opinion mining is used for software engineering research [21] and for media and communication [22–24] and for topic modeling [25] it is implemented in one of the phases; however, to the best of the author's knowledge there does not exist a research for analyzing work trends in post pandemic era using opinion mining of data in the proposed approach. Second, we have created a novel curated dataset to address the stated issue. Furthermore, the data employed is in an unstructured data format, thus makes the process of analysis more challenging.

Thirdly, a unique issue description and research goals are outlined. The existing approach is to rely on internal surveys and human judgement to make business decisions. However, we have presented an automated approach using hybridization of natural language processing and machine learning for proposed opinion mining framework. We have utilized a variety of performance evaluation metrics to demonstrate that the proposed approach is superior to existing approaches namely—perplexity, coherence and TSNE. This would aid in assessing the influence of specific cases on broader rhetoric that would effectively lead to effective business intelligence.

VI. Conclusion

This research focusses on development of novel opinion mining framework using hybridization of deep learning word embedding and topic modeling processes. It focusses on current and near future working trend in enterprises which is clearly near remote or hybrid working environment as demonstrated in opinion mining experimentation which offers greater flexibility and work life balance. The experimental results and discussion depict that the proposed approach provide better feature vector using text feature weighted approach based classifiers. Several industry trends have been identified. Another trend evaluation was growth in cybersecurity domain in industries and greater awareness creation.

The research can be developed further through the development of a customized data analysis engine. Further research directions can include automated query analysis for evaluation using large-scale big data analytics. Our future work focuses on semantic opinion conceptualization and query expansion, classification and inference.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Ruchi K Sharma conducted the research, analyzed the data, and wrote the paper. Dr. Pravin Shrinath supervised the research process and overall review process. All authors had approved the final version.

Acknowledgment

The authors value the reviewer's insightful comments, which helped them to better their work.

References

[1] K. Khan, B. Baharum, A. Khan, and A. Ullah., "Mining opinion components from unstructured reviews: A review," *J. King Saud Univ.—Computer and Information Sciences*, vol. 26, issue 3, pp 258–275, 2014.

[2] V. Singh and S. K. Dubey, "Opinion mining and analysis: A literature review," in *Proc. 2014 5th International Conference of the Next Generation Information Technology Summit (Confluence)*, 2014, pp. 232–239.

[3] H. Chen and D. Zimbra, "AI and opinion mining," *IEEE Intell. Syst*., vol. 25, no. 3, pp. 74–80, 2010.

[4] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Inf. Fusion*, vol. 36, pp. 10–25, 2017.

[5] M. S. Evans, "A computational approach to qualitative analysis in large textual datasets," *PLoS ONE*, vol. 9, no. 2, Feb. 2014, https://doi.org/10.1371/journal.pone.0087908

[6] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *J. Big Data*, vol. 6, no. 91, 2019, https://doi.org/10.1186/s40537-019-0254-8

[7] S. U. Maheswari and S. S. Dhenakaran, "Opinion mining on integrated social networks and e-commerce blog," *IETE Journal of Research*, vol. 69, no. 4, pp. 2080–2088, 2021, https://doi.org/10.1080/03772063.2021.1886603

[8] A. M. Shah, X. Yan, S. A. A. Shah, *et al*., "Mining patient opinion to evaluate the service quality in healthcare: A deep-learning approach," *J. Ambient Intell. Human Comput.*, vol. 11, pp. 2925–2942, 2020.

[9] R. Sharma and P. Srinath, "Business intelligence using machine learning and data mining techniques-an analysis," in *Proc. 2018 Second International Conference on Electronics Communication and Aerospace Technology (ICECA)*, 2018, pp. 1473–1478.

[10] S. Ruchi and P. Srinath, "Big data platform for enterprise project management digitization using machine learning," in *Proc. 2018 Second International Conference on Electronics Communication and Aerospace Technology (ICECA)*, 2018, pp. 1479–1484.

[11] T. L. Mikolov and V. Quoc, "Distributed representations of sentences and documents," arXiv preprint, arXiv: 1405.4053, 2014.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[13] R. Annisa, I. Surjandari, and Zulkarnain, "Opinion mining on Mandalika hotel reviews using latent Dirichlet allocation," in *Proc. Comput. Sci.*, vol. 161, pp. 739–746, 2019.

[14] N. Banik and M. H. H. Rahman, "Evaluation of naïve bayes and support vector machines on Bangla textual movie reviews," in *Proc. 2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–6, Sep. 2018.

[15] M. Rodrigo, V. João, and P. G. Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Systems with Applications*, vol. 40, pp. 621–633, 2013.

[16] M. S. Sabuj, Z. Afrin, and K. M. A. Hasan, "Opinion mining using support vector machine with web based diverse data," in *Lecture Notes in Computer Science*, Springer International Publishing, pp. 673–678, 2017.

[17] C. Luo, "Analyzing the impact of social networks and social behavior on electronic business during COVID-19 pandemic," *Inf. Process. Manage.*, vol. 58, no. 5, 102667, Sep. 2021.

[18] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets," *IEEE Access*, vol. 8, pp. 181074–181090, 2020.

[19] J. G. Harb, R. Ebeling, and K. Becker, "A framework to analyze the emotional reactions to very violent events on Twitter and influential factors," *Information Processing & Management*, vol. 57, no. 6, 2020.

[20] S. N. Saleh, N. Sameh, C. U. Lehmann, S. McDonald, A. Basit, A. Mujeeb and J. R. Medford, "Understanding public perception of coronavirus disease 2019 (COVID-19) social distancing on Twitter,"

Infect Control Hosp Epidemiol, vol. 42, no. 2, pp. 131–138, Feb. 2021.
[21] B. F. Demissie, M. Ceccato, and L. K. Shar, "Security analysis of permission re-delegation vulnerabilities in Android apps," *Empir. Software Eng.*, vol. 25, pp 5084–5136, 2020.
[22] C. Puschmann and T. Scheffler, "Topic modeling for media and communication research: A short primer," *SSRN*, Aug. 2016, https://dx.doi.org/10.2139/ssrn.2836478
[23] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, *et al*., "Applying LDA topic modeling in communication research: Toward a valid and reliable methodology," *Commun. Methods Measures*, vol. 12, no. 2, pp. 93–118, 2018.
[24] C. Jacobi, W. Van Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," *Digit. Journalism*, vol. 4, no. 1, pp. 89–106, 2016.
[25] S. Koltcov, S. I. Nikolenko, O. Koltsova, V. Filippov, and S. Bodrunova, "Stable topic modeling with local density regularization," in *Proc. INSCI 2016: Lecture Notes in Computer Science*, Springer, Cham, 2016, vol. 9934, https://doi.org/10.1007/978-3-319-45982-0_16