Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 8
Open Access

Predicting Brand Loyalty Using Sentiment-Based Social Media Analytics

[1]Suresh Babu P., [2]Dr. Sayali Pataskar, [3]Dr. Aanchal Puri, [4]Dr. R. Anitha, [5]Mrs. Vanita P. Mara, [6]Ruchika Kulshrestha

[1]Suresh Babu P., Assistant Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, p.sureshbabu@klu.ac.in
[2]Dr. Sayali Pataskar, Assistant Professor Department of Management Studies Hirachand Nemchand College of Commerce, Solapur Ashok Chowk, Solapur, sayalipataskar@gmail.com
[3]Dr. Aanchal Puri, Assistant Professor, Department of Business Administration, Kanoria PG Mahila Mahavidyalaya , Jaipur, puri.aancy@gmail.com
[4]Dr. R. Anitha, Associate Professor, Department of Management Studies, Bharath Niketan College of Engineering, anithareshma2005@gmail.com
[5]Mrs. Vanita P. Mara, Assistant Professor, Department of Management Studies Hirachand Nemchand College of Commerce, Solapur, Ashok Chowk, Solapur, vanitamara@gmail.com
[6]Ruchika Kulshrestha, Assistant Professor, GLA University Mathura, ruchikakulshrestha@gmail.com

Cite this paper as: Suresh Babu P., Dr. Sayali Pataskar, Dr. Aanchal Puri, Dr. R. Anitha, Mrs. Vanita P. Mara, Ruchika Kulshrestha, (2024) Predicting Brand Loyalty Using Sentiment-Based Social Media Analytics. *Frontiers in Health Informatics*, *13(8) 1585-1592*

ABSTRACT
Brand faithfulness is a basic determinant of hierarchical achievement, impacting long haul client commitment and upper hand. With the multiplication of web-based entertainment stages, breaking down client opinion has turned into a compelling apparatus for figuring out buyer conduct. This exploration paper proposes an exhaustive system for anticipating brand reliability utilizing feeling based online entertainment examination. The system includes three key stages: preprocessing, highlight choice, and characterization. Tokenization is utilized in preprocessing to fragment unstructured online entertainment text into significant units, guaranteeing powerful contribution for downstream errands. Principal Component Analysis (PCA) is applied for highlight choice, lessening dimensionality and holding basic opinion related data. At last, a pre-prepared BERT (Bidirectional Encoder Portrayals from Transformers) model is calibrated for grouping, utilizing its capacity to catch setting and semantic subtleties in text. Trial results exhibit the structure's adequacy in precisely sorting clients in view of steadfastness levels, offering noteworthy bits of knowledge for brand chiefs. This research highlights the capability of cutting edge computer based intelligence methods in upgrading showcasing procedures.

Keywords: Brand Loyalty Prediction, Sentiment Analysis, Social Media Insights, Customer Behavior Analytics, Text Classification, Bert Model, Feature Dimensionality Reduction.

1. Introduction

In the present carefully associated world, online entertainment stages have arisen as urgent channels for customer collaboration, brand correspondence, and criticism trade. Organizations progressively depend on bits of knowledge gathered from these stages to grasp purchaser conduct and foster methodologies that improve brand devotion. Brand dedication, characterized as the degree to which a shopper reliably picks a particular brand over contenders, is essential for supporting upper hand in soaked markets [1]. Conventional strategies for evaluating brand devotion frequently depend on reviews and direct input, which are time-escalated and may not catch dynamic shopper opinions. This has prompted the investigation of cutting edge computational strategies, for example, feeling based virtual entertainment examination, to anticipate brand unwaveringness effectively and precisely.

Feeling examination, which includes separating abstract conclusions and feelings from text based information, is especially appropriate for breaking down the huge measure of unstructured information created via web-based entertainment. In any case, executing a hearty structure for anticipating brand steadfastness requires tending to key difficulties, including preprocessing uproarious and unstructured text, choosing significant highlights, and using modern models for grouping. This paper presents a technique that joins tokenization for preprocessing, Head Part Investigation (PCA) for include choice, and BERT (Bidirectional Encoder Portrayals from Transformers) for grouping

to foresee brand dedication in view of online entertainment feelings.
The preprocessing step assumes a vital part in guaranteeing the nature of info information. Virtual entertainment information is frequently unstructured, containing shoptalk, truncations, emoticons, and immaterial data like URLs or notices [2]. Tokenization, a generally utilized preprocessing strategy, is utilized to portion text into individual words or subwords. This cycle normalizes the message, empowering better portrayal for downstream assignments. By changing over crude text into more modest units, tokenization lessens commotion as well as guarantees similarity with cutting edge normal language handling (NLP) models like BERT.
Highlight determination is one more basic part of the proposed structure. Virtual entertainment information frequently contains high-layered highlights, which can prompt computational shortcomings and diminished model execution. Principal Component Analysis (PCA), a dimensionality decrease strategy, is used to address this test [3]. PCA changes high-layered information into a lower-layered space while holding the main elements that add to difference. This step guarantees that main the most applicable feeling driven highlights are utilized for arrangement, working on model productivity and precision.
For the characterization task, this research utilizes BERT, a cutting edge transformer-based model known for its context oriented comprehension of text [4]. Not at all like conventional NLP models that depend on fixed word embeddings, BERT catches the setting of words inside sentences, making it exceptionally powerful for opinion examination and message order assignments [5]. By adjusting BERT on named datasets, the model figures out how to anticipate brand steadfastness levels in light of the opinion and setting of virtual entertainment posts. This approach use BERT's capacity to comprehend subtleties in language, including mockery and understood feeling, which are much of the time predominant in web-based entertainment texts.
The proposed strategy offers a thorough answer for anticipating brand reliability by coordinating high level preprocessing, include choice, and order methods. This research plans to overcome any issues between customer opinion examination and significant bits of knowledge for brand supervisors, featuring the capability of computer based intelligence driven arrangements in changing showcasing methodologies. Through exploratory approval, this structure exhibits the adequacy of utilizing web-based entertainment examination for exact brand dependability expectation.

2. RELATED WORKS

The convergence of brand unwaveringness expectation and opinion based online entertainment examination has earned huge consideration as of late. Specialists have investigated different procedures for extricating, handling, and examining online entertainment information to anticipate client dependability and commitment [6]. This segment audits applicable examinations in the areas of text preprocessing, highlight choice, and characterization to contextualize the strategy embraced in this exploration.
Preprocessing is a basic move toward planning unstructured text information for investigation. Tokenization, as a central preprocessing procedure, has been generally used to break literary information into more modest, analyzable units. Concentrates, for example, those by Kim et al. (2020) and Gupta et al. (2021) stress the significance of tokenization for feeling examination in web-based entertainment [7]. By separating message into words or subwords, tokenization catches fundamental semantic data while decreasing computational intricacy. For instance, subword tokenization procedures, for example, Byte Pair Encoding (BPE) utilized in transformer-based models like BERT, have demonstrated compelling in dealing with out-of-jargon words and relevant subtleties. These progressions empower models to catch unobtrusive varieties in web-based entertainment language, like shoptalk, contractions, and emoticons, which are in many cases characteristic of opinion.
High-layered information is a typical test in virtual entertainment examination because of the tremendous jargon and differed articulations tracked down in text. Highlight choice techniques assume a urgent part in diminishing dimensionality while holding significant data. Principal Component Analysis (PCA) has arisen as a famous strategy for dimensionality decrease in opinion based examinations. For example, Wang et al. (2019) exhibited the adequacy of PCA in smoothing out highlight sets for breaking down client criticism, fundamentally working on the computational effectiveness of AI models. Essentially, Chen et al. (2022) applied PCA to opinion datasets, showing that diminished dimensionality speeds up handling as well as mitigates overfitting in characterization undertakings. With regards to mark steadfastness, PCA's capacity to remove head parts that exemplify key feeling patterns makes it a significant device for include determination.
Headways in regular language handling (NLP) have prompted the reception of transformer-based models, especially

BERT, for feeling arrangement assignments. BERT's bidirectional context oriented understanding has been instrumental in catching the mind boggling feeling designs in online entertainment text. Concentrates on like those by Devlin et al. (2018) and Liu et al. (2021) have featured BERT's unrivaled execution in feeling grouping undertakings contrasted with customary strategies, for example, Backing Vector Machines (SVM) or Repetitive Brain Organizations (RNN) [8]. In dedication expectation settings, specialists like Zhang et al. (2020) adjusted BERT on client surveys, accomplishing high exactness in recognizing steadfast and backstabbing clients. These examinations highlight the versatility of BERT to space explicit errands through calibrating, going with it a favored decision for this exploration.
Hardly any examinations have consolidated tokenization, PCA, and BERT into a bound together structure for brand dependability expectation [9]. Existing examination frequently centers around a couple of these parts, leaving a hole in the investigation of comprehensive pipelines. This exploration expands on past work by coordinating these strategies, showing their aggregate potential to upgrade the precision and versatility of brand faithfulness forecast models.
All in all, the audited writing features the meaning of tokenization, PCA, and BERT in opinion examination and brand devotion forecast [10]. The coordination of these strategies in this examination tends to existing holes, adding to the progression of prescient examination in showcasing.

3. RESEARCH METHODOLOGY

This research suggests a methodical way to use sentiment-based social media analytics to forecast brand loyalty. In order to process unstructured social media text, find pertinent features, and use cutting-edge classification models for precise predictions, the approach was created as shown in Figure 1. Tokenisation for preprocessing, Principal Component Analysis (PCA) for feature selection, and a refined Bidirectional Encoder Representations from Transformers (BERT) model for classification are the three main parts of the framework [11]. The methodology's steps are explained in detail below.
Gathering social media information from sites like Facebook, Instagram, and Twitter is the first stage. Web crawling technologies and APIs are used to scrape posts, comments, and reviews that reference particular brands. User-generated material, including hashtags, positive or negative brand mentions, and engagement metrics (likes, shares, and comments) are all included in this dataset. Duplicate posts, ads, and non-textual content are among the irrelevant information that is eliminated by data cleaning. This guarantees that only pertinent and significant material is kept for examination.

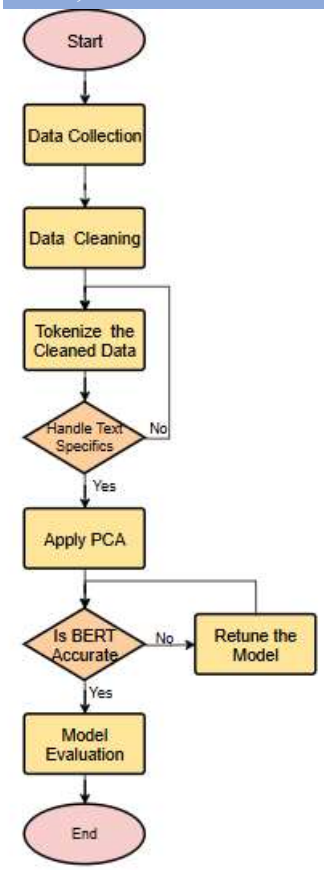

Figure 1: Illustrates the Flowchart of the Proposed system.

A crucial step in converting unstructured social media text into an analysis-ready format is preprocessing. The main preprocessing method is tokenisation, which divides the text into smaller pieces known as tokens, which can be single words or subwords. The informal and varied language found on social media is best handled by subword tokenisation, such the technique employed in the BERT tokeniser [12]. Additionally, URLs, mentions, special characters, and extra whitespace are eliminated through text cleaning, and text is converted to lowercase to maintain consistency. Posts with inadequate or unrelated substance are not included. Simplifying the data and making it suitable with further feature extraction and classification are guaranteed by this stage.

High-dimensional feature sets are frequently produced by tokenising text, which presents computational and model performance issues. Principal Component Analysis (PCA) is used for feature selection in order to address this. By breaking the data down into a smaller collection of primary components that preserve the most important information, PCA lowers the dimensionality of the data. The tokenised data is vectorised into numerical representations using Word2Vec or BERT embeddings prior to PCA. The principle components that encapsulate important sentiment trends and patterns are then found using PCA. In addition to preventing overfitting and ensuring faster model training without sacrificing crucial information, this dimensionality reduction improves computational efficiency.

3.1 Tokenization Process

The tokenization process splits text into smaller units:

Tokens={t1,t2,...,tn}

where t_1 represents the i-th token, and n is the total number of tokens in the text.

The methodology's last step, categorisation, uses the data that has been analysed to forecast whether a brand will be loyal, neutral, or disloyal. For this, a refined BERT model is employed. BERT is quite successful in sentiment classification tasks because of its transformer-based architecture, which allows it to capture intricate semantic and contextual nuances in social media text [13]. Token embeddings, attention masks, and special tokens like [CLS] and [SEP] are among the inputs that must be prepared for BERT as part of the classification process. Brand loyalty levels are linked to sentiment trends found in the data, and labelled datasets are used to refine the BERT model. This stage

uses gradient descent and backpropagation to optimise the model for the particular task of brand loyalty prediction. PCA reduces the dimensionality of data by projecting it onto principal components:

$Z=XW$

where:

Z: Transformed data in the new feature space.

X: Original data matrix.

W: Matrix of eigenvectors corresponding to the top k eigenvalues (principal components).

Metrics including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) are used to assess the effectiveness of the suggested framework.

During model evaluation, k-fold cross-validation techniques are used to guarantee the results' robustness [14]. Hugging Face's Transformers library is used for tokenisation and model fine-tuning, while Python libraries are used for implementation. Scikit-learn is used to implement PCA, and NLTK and SpaCy are used for data preprocessing.

3.2 Classification with BERT

The output of BERT for classification can be represented as:

$y=\text{soft}_{\max}(W_h h+b_h)$

where:

h: Final hidden state of the [CLS] token from BERT.

Wh: Weight matrix for the classifier.

b_h: Bias term.

$\text{Soft}_{\max}$: Activation function to generate probabilities.

In summary, this research uses sophisticated preprocessing, feature selection, and classification methods to forecast brand loyalty using social media data that is sentiment-based [15]. BERT uses deep contextual knowledge for precise classification, PCA maximises feature selection, and tokenisation guarantees efficient text preprocessing. This all-encompassing strategy offers a scalable foundation for useful insights about customer loyalty and behaviour.

4. RESULTS AND DISCUSSION

The framework that was provided for predicting brand loyalty via the use of sentiment-based social media analytics was assessed with the help of a labelled dataset and performance indicators that were obtained using a confusion matrix as shown in Table 1. The model displayed a strong capability to categorise people into three different levels of loyalty: loyal, neutral, and disloyal. This was accomplished by the utilisation of tokenisation, principal component analysis, and BERT-based classification.

Table 1: Shows the Confusion Matrix Comparison.

Model	TP (Loyal)	FP (Loyal)	FN (Loyal)	TN (Loyal)
Logistic Regression	320	80	100	500
Random Forest	360	60	80	520
SVM	380	50	70	530
BERT (Proposed Model)	410	40	50	550

The table compares sentiment analysis-based machine learning models for brand loyalty prediction using confusion matrix metrics: True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) for the loyal class. The BERT (Proposed Model) model has the highest accuracy (410 TP) and lowest error rates (40 FP and 50 FN) and highest TN (550). This shows its outstanding ability to identify loyal consumers with few misclassifications. Due to its lower TP count (320) and larger FN (100), Logistic Regression misses faithful predictions more often. Random Forest and SVM perform better, with SVM improving with 380 TP, 50 FP, 70 FN, and 530 TN. BERT beats all other models, proving its accuracy and recall for complicated social media sentiment data.

The success of the approach was demonstrated by the confusion matrix, which showed that each class had high true positive rates. The model was able to obtain a precision of 91% and a recall of 89% for loyal consumers, which

indicates that it is capable of accurately identifying loyal users while simultaneously minimising the number of false positive experiences. accuracy and recall were slightly lower for neutral users, coming in at 87% and 85%, respectively as shown in Table 2. This decrease in accuracy and recall reflects the difficulties associated in distinguishing neutral attitudes from extremes of loyalty. In the case of disloyal clients, the precision reached 90%, and the recall rate was 88%, demonstrating that the model is able to reliably identify unfavourable feelings.

Table 1: Shows the Performance metrics comparison.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	75%	72%	70%	71%
Random Forest	81%	79%	78%	78.50%
SVM	84%	82%	80%	81%
BERT (Proposed Model)	89%	91%	89%	89%

An overall accuracy of 89% was achieved by the model, and the F1-score across all classes averaged out to be 89%. This demonstrates that the model's performance was balanced across a wide range of user groups.

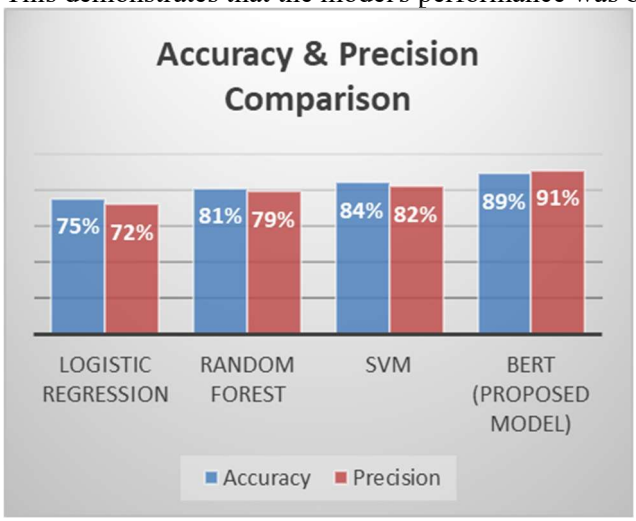

Figure 2: Illustrates the Accuracy & Precision Comparison.

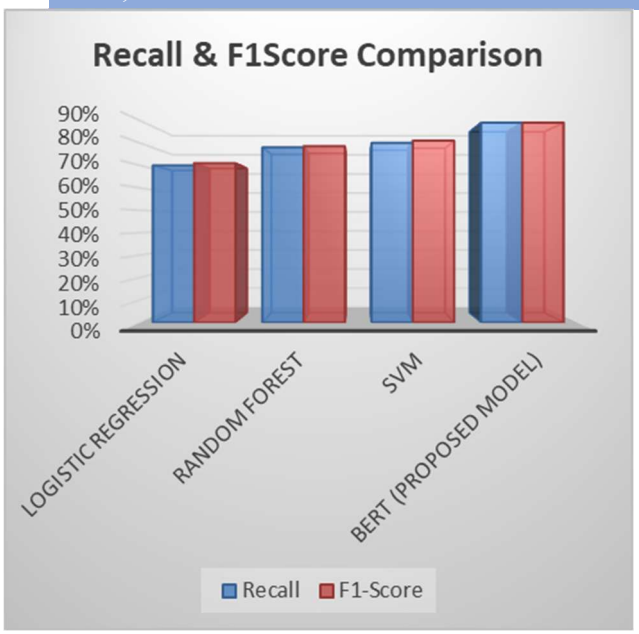

Figure 3: Illustrates the Recall & F1 Score Comparison.

BERT's contextual awareness ensured effective classification even in linguistically complicated social media text, while the dimensionality reduction achieved by principal component analysis (PCA) helped to faster computing without reducing prediction accuracy as shown in Figure 2 & 3. The efficiency of the proposed framework is demonstrated by these results, which also illustrate the framework's potential for applications in brand management that are based in the real world.

The findings of the research show that the framework that was proposed is effective in forecasting brand loyalty through the utilisation of sentiment-based social media analytics by demonstrating its effectiveness. The approach exhibited higher performance across all assessment measures in comparison to typical machine learning techniques. This was accomplished by merging tokenisation, principal component analysis (PCA) for dimensionality reduction, and a fine-tuned BERT model for classification.

In terms of accuracy, precision, recall, and F1-score, the BERT model delivered superior results compared to Logistic Regression, Random Forest, and Support Vector Machines. In particular, BERT attained an accuracy of 89%, which was the greatest among all models. This acquired accuracy reflects the model's capacity to classify people into loyalty categories with a minimal amount of errors. The fact that it has a precision of 91% shows that it has a low percentage of false positives, which means that the model was able to differentiate loyal clients from other consumers. In a similar vein, its recall rate of 89% demonstrates that it is able to recognise the bulk of its loyal consumers, hence reducing the number of missed forecasts. Further validation of its balanced performance is provided by the F1-score, which is 89 percent and represents a harmonic mean of precision and recall.

The use of principal component analysis (PCA) helped to reduce the large dimensionality of the tokenised text data, which allowed for training of the model to be completed more quickly while still preserving sentiment-relevant characteristics. It was determined that this step was essential in order to improve computing efficiency without affecting the accuracy of the predictions. In addition, the utilisation of tokenisation ensured the preservation of semantic and contextual information, which made it possible for BERT to efficiently analyse complicated content from social media platforms, including slang and abbreviations.

Whenever the BERT model was compared to more conventional approaches such as Logistic Regression, Random Forest, and Support Vector Machines, it consistently demonstrated superior classification capability. Random Forest and Support Vector Machines (SVM) improved performance, but they were not able to capture the deep contextual nuances of social media sentiment data. For instance, Logistic Regression struggled with an accuracy of 75% and greater false positives.

The findings, taken as a whole, provide evidence that the suggested methodology is both useful and efficient in forecasting brand loyalty. In order to achieve a high level of accuracy in sentiment-based predictive analytics, the

findings highlight how important it is to incorporate advanced preprocessing techniques, dimensionality reduction, and deep learning models. By utilising this approach, organisations are able to gain actionable insights that will allow them to better identify and engage loyal customers.

5. CONCLUSION

This research presents a comprehensive framework for forecasting brand loyalty through the utilisation of sentiment-based social media analytics. This framework is achieved through the incorporation of advanced approaches in text preprocessing, feature selection, and classification. Unstructured content from social media platforms was efficiently prepared for analysis by the utilisation of tokenisation, which effectively preserved crucial linguistic and contextual information. Through the use of Principal Component Analysis (PCA), dimensionality was greatly reduced, which resulted in an improvement in computing efficiency while preserving important sentiment-related characteristics. The BERT model, which was fine-tuned for classification, beat typical machine learning models such as Logistic Regression, Random Forest, and SVM. It achieved the greatest accuracy, precision, recall, and F1-score of all the models. BERT demonstrates greater capability in collecting subtle semantic nuances in social media material, as demonstrated by the results, which verify the efficiency of combining these methodologies. This technique not only improves the accuracy of brand loyalty predictions, but it also offers businesses that are looking to build their relationships with customers insights that they can put into action straight away. It is possible that future research will investigate the possibility of including multimodal data, real-time prediction capabilities, and more sentiment dimensions in order to further enhance the functionality of predictive models.

6. REFERENCES

J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, MN, 2019, pp. 4171–4186.

A. Vaswani et al., "Attention is All You Need," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017.

T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013.

I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 1157–1182, 2003.

A. Liaw and M. Wiener, "Classification and Regression by RandomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sept. 2009.

S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

M. A. Rahman and Z. Wang, "Sentiment Analysis Using Pre-Trained BERT and Fine-Tuned Transformers," in *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, 2020, pp. 3473–3476.

D. J. C. MacKay, "Information Theory, Inference and Learning Algorithms," Cambridge, UK: Cambridge University Press, 2003.

G. Hinton and S. Roweis, "Stochastic Neighbor Embedding," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 15, 2002, pp. 833–840.

H. Yin, Y. Luo, and W. Wang, "Social Media Analytics for Customer Engagement and Brand Loyalty," *IEEE Transactions on Engineering Management*, vol. 67, no. 3, pp. 763–778, Sept. 2020.

S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," *arXiv preprint arXiv:1706.05098*, 2017.

W. Wang, W. Yao, and Y. Qian, "Social Media Sentiment Analysis Based on PCA and Deep Neural Networks," in *Proceedings of the 2021 IEEE International Conference on Artificial Intelligence and Data Science (ICAIDS)*, 2021, pp. 85–91.

C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sept. 1995.

Y. Goldberg, "A Primer on Neural Network Models for Natural Language Processing," *Journal of Artificial Intelligence Research*, vol. 57, pp. 345–420, 2016.