Medicine Recommendation System based on Sentiment Analysis of Medicine Reviews using Machine Learning

Lavi Kaushik, Parth Chugh, Mashrufa Mondal, Kashish Verma
Department of Computer Science and Engineering, MIET, Meerut, Uttar Pradesh

Abstract- **The healthcare industry is increasingly adopting intelligent systems to improve drug and disease recommendation processes, ensuring timely and accurate support for both patients and healthcare providers. This paper introduces a dual-function Drug and Disease Recommendation System that integrates sentiment analysis of drug reviews and symptom-based disease prediction using machine learning techniques. The system utilizes patient feedback to assess drug efficacy through sentiment polarity classification, while also mapping user-reported symptoms to probable diseases. Implemented as a real-time multilingual Web application, the system employs Support Vector Machine with TF-IDF vectorization for sentiment analysis (96.35% accuracy) and Multi-Layer Perceptron with TF-IDF for disease prediction (Jaccard Score: 0.7631), offering PDF output and REST API support for seamless integration and usability.**

I. INTRODUCTION

In recent years, the healthcare sector has faced significant challenges due to growing populations, rising demand for medical services, and critical shortages of trained professionals, particularly in rural and underserved regions. The COVID-19 pandemic further intensified these issues, highlighting the urgent need for scalable, technology-driven solutions that can assist both medical professionals and patients in decision-making.

One of the most common problems in medical practice is selecting the most appropriate drug or treatment based on patient symptoms and expected outcomes. While physicians rely on their expertise and clinical history, human error and information overload can lead to prescription mistakes. Simultaneously, patients increasingly rely on online platforms, where reviews and forums provide anecdotal evidence on drug efficacy—though not always in a structured or reliable form.

To address these challenges, our research proposes a dual-function, machine-learning-powered system that performs: [1] Sentiment analysis of drug reviews using Support Vector Machine (SVM) with TF-IDF vectorization (achieving 96.35% accuracy), and [2] Symptom-based disease prediction using a Multi-Layer Perceptron (MLP) with TF-IDF (Jaccard score: 0.7631). The models are deployed via a real-time, multilingual web application featuring REST APIs, SQL database integration, and downloadable PDF reports. This approach not only improves prediction performance but also enhances accessibility, usability, and trust in automated healthcare systems.

II. LITERATURE SURVEY

Recommender systems are widely utilized across diverse domains such as e-commerce, hospitality, and entertainment. These systems provide personalized suggestions based on user behavior, preferences, and previous interactions. However, their application in the medical field, especially for drug and disease recommendation, is still relatively underexplored due to the complexity and sensitivity of healthcare data. Medical data contain terminologies that are highly domain-specific, clinical abbreviations, and expressions that are very intricate, unlike product reviews in commercial online platforms; hence conventional sentiment analysis and recommendation algorithms are inefficient without heavy modifications.

Of the ontology-based systems for clinical drug recommendation, GalenOWL, whose relevance is mooted in [1], is one of the earliest. Patient information, such as allergies, diagnoses, and drug interactions, are converted into formal ontological data using ICD-10 and UNII medical taxonomies. The drugs recommended for a patient are thus mapped accurately to the symptoms and history of that patient which improves decision making in the clinical scenario.

Building on this, semantic analysis, Leilei Sun et al did large-scale mining on treatment records with clustering techniques for optimal therapeutic regimen recommendations. The work relied on multiple hospitals' Electronic Medical Records (EMRs) to personalize treatments with regards to patient demographics and comorbidities. The framework proved the efficacy of semantic similarity and context-aware grouping in medical intervention recommendations. .[2]

Another significant herald of AI into the medical realm is multilingual sentiment analysis. The authors of collected multilingual tweets oriented towards healthcare, translated them into English using the Google Translate API, and used Naïve Bayes and RNN for sentiment classification. The RNN showed greater efficacy than Naïve Bayes, nailing sentiment classification at 95.34% accuracy, while Naïve Bayes lagged far behind, at 77.21%. This then shows the edge of deep learning in modeling syntactic and semantic sentiment patterns even in translated documents. [3]

Risk-aware recommendation systems are yet another important new development in this area. The authors in determined over 60 physiological and behavioral risk factors (hypertension, alcohol use, and immunity status) and built classifiers to assign risk levels to patients before any drug recommendation. This thus guaranteed that recommendations were accurate and also safe for at-risk populations. [4]

Drug recommendation considers collaborative and content-based filtering. CADRE, or Cloud-Assisted Drug Recommendation Engine, originally designed by Zhang et al. was based on collaborative filtering to group similar medications. The system was then transformed into one that was cloud-assisted and which in turn was able to make use of tensor decomposition to tackle the cold start and sparsity problems usually met in recommendation systems for healthcare. The transformation asserts the necessity of having scalable infrastructure and better representation of real-world application data. [5]

The various types of machine learning algorithms Support Vector Machine, Decision Trees, and Backpropagation Neural Networks were tested for recommending treatments based on clinical datasets. It was found that SVM outperformed all the others in terms of accuracy and efficiency and was also scaled to handle large datasets with ease such that it became the algorithm of choice for many medical classification applications.[6]

Other modern architectures, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have been used for enhanced semantic feature extraction. A good example is Jiugang Li et al. for recommending hashtags, combining skip-gram word embeddings and CNNs with further refinements to LSTM for classification. While these works are not targeted specifically at health care, they nonetheless underscore the incapacity of conventional model SVMs to capture semantic dependencies that can be learned far more effectively by neural networks.[7]

Despite the promising outcomes of these studies, a clear gap remains. Most of the existing research either focuses on sentiment analysis of reviews or clinical treatment recommendations, but not both in a unified system. Moreover, many works remain academic in nature, without public-facing deployment, usability features, or integration into real-time platforms. They often lack multilingual support, RESTful APIs, or integration with web-based interfaces—features necessary for practical use.

Our research addresses this gap by proposing an integrated, dual-function system that performs both sentiment-based drug review classification and symptom-based disease prediction using machine learning. We combine SVM with TF-IDF for sentiment analysis (achieving 96.35% accuracy) and MLP with TF-IDF for disease prediction (Jaccard Score: 0.7631). Additionally, we deploy our models in a full-stack multilingual web platform with REST API integration and PDF download support, thus bridging the gap between research and real-world application

III. METHODOLOGY

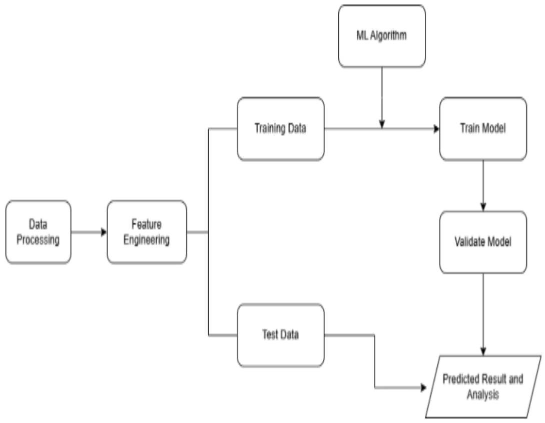


Fig.1 Methodoloy

3.1 Data Processing:

Data processing is a basic and innovative component of this research method. Information is sourced from online material including the Health Forum, drug assessment sites, and drug feedback sites. These sources yield raw data such as user ratings, ratings, names of drugs, side effects and treatment outcomes. This data is collected and processed to make it clean and fit for analysis. Special characters, word text, and words themselves are eliminated during the preliminary processing stage. Trunks or lemmaizations are then employed to transform words into root shapes. Through this process, noise is removed, unstructured data is transformed into structured form, and various analyses and unique extractions are provided.

3.2Feature Engineering:

Following preprocessing, feature engineering is the next stage, during which vectorization techniques are used to transform text data into numerical form. Methods such as:

3.2.1 Bag of Words (BoW):

One of the most basic text vectorization methods in Natural Language Processing (NLP) is called Bag of Words (BoW). It depicts a text (sentence or document) as a collection of words without taking into account the words' context or order. Only the frequency with which each word occurs in the text is counted. BoW is utilized in this project to transform the drug review text data during the feature extraction stage. into in numerical form. For the purpose of training and classifying machine learning models, each review is converted into a vector of word frequencies.

The feature vector is as follows if a document is D and the total vocabulary size is V:

$$\mathrm{BoW}(D)=[f1,f2,f3,...,fn]$$

Where:

fn = Frequency of the nth word from the vocabulary in document *D*

3.2.2 Word2Vec:

Word2Vec is a sophisticated word embedding method that uses semantic meaning and context to transform words into dense vector representations. Word2Vec records the relationships between words, in contrast to BoW, which merely counts words.

Word2Vec is utilized in this project to record the context of words in medication reviews. It improves sentiment analysis and recommendation generation by assisting the model in comprehending word similarities (for example, "tablet" and "medicine" will have closer vectors).Word2Vec has two main architectures:

CBOW (Continuous Bag of Words) — Predicts the target word from surrounding words.

Skip-Gram — Predicts surrounding words from the target word.

Formula for Word2Vec:

The probability of a target word w□ given the context words is:

P(w□ | context) = exp(v□w□ □ . v□context□) / Σ□w ∈ V□ exp(v□w□ . v□context□)

Where,

v□w□ □ = Vector representation of the target word

v□context□ = Vector representation of the context words

V = Total vocabulary size

3.2.3 TF-IDF (Term Frequency - Inverse Document Frequency):

A numerical metric called TF-IDF is used to assess a word's significance in a document in relation to a corpus, or group of documents.

Words are given weight according to their importance rather than just their frequency count in this widely used weighting technique. Search engines, information retrieval, and text mining all frequently use TF-IDF.

The idea behind TF-IDF is to give terms that appear frequently in the dataset a lower priority because they offer less useful information. As a result, TF-IDF calculates a word's relevance in a document rather than just how often it appears.

$$TF(t, d) = \log(1 + freq(t, d))$$

Where,

t=Term(word)

d=Document

freq(t, d) = Number of times term t appears in document d.

The formula for Inverse Document Frequency (IDF) is:

IDF(t) = log (N / n□)

Where,

N=Total number of documents

n□ = Number of documents containing term t

The final TF-IDF formula is:

$$TF\text{-}IDF(t, d) = TF(t, d) \times IDF(t)$$

3.3 Training Data and Test Data:

After the data underwent processing and feature engineering, it was separated into training and testing sets. The training data provide the machine learning models with the means to learn how to categorize the sentiments based on patterns in the reviews. The testing data are separated for the evaluation of the models' accuracy and performance. This separation ensures that the models can generalize well to the new and unseen data.

3.4 ML Algorithm:

After obtaining the vectorized data, various machine learning (ML) algorithms are being applied for the correct classification of sentiment. A number of algorithms are being tested to identify the best method for sentiment analysis.

Algorithms Used:

3.4.1 Support Vector Machine (SVM):

Among the other machine learning kernels, the widely preferred one for classification and regression applications includes the Support Vector Machine (SVM))

SVM is specially designed to establish the best separation, which is also known as the decision facet or hyperplane, between the two sets of data.

SVM aims to maximize the margin between data points of different classes in order to improve generalization with respect to unseen data.

3.4.2 MLP (Multi-Layer Perceptron):

Multi-layer perceptron (MLP) is an advanced artificial neural network architecture with numerous layers of nodes (neurons). It is commonly used for regression and many classification problems. An MLP consists of an input layer, one or more hidden layers, and an output layer. Each neuron in a layer is connected to every neuron in the next layer, forming a fully connected network. MLPs use activation functions like ReLU or sigmoid and are trained using backpropagation and gradient descent.

3.5 Train Model:

Once the algorithms are selected, the models undergo training using the training dataset. During this process, feature vectors are fed into the model allowing it to learn and adjust its parameters to minimize errors. By fine-tuning the models, we make sure they can classify the sentiments in drug reviews...

3.6 Validate Model:

The testing dataset checks how well the models work after training. This step looks at how the models handle data they haven't seen before. We use different ways to measure how good the models are, like Jaccard Score, F1-Score, Hamming Loss, Precision, and Recall. Checking the models this way helps make sure they're not too fitted to the training data or not fitted enough..

3.7 Predicted Result:

The model predicts the mood of new drug reviews after validation. Its output indicates whether the review is neutral, negative, or positive. To provide a more complete assessment of a specific drug, the system also calculates a sentiment score by averaging the expected results.

3.8 Analysis:

Evaluating what the expected results will be is the final step. Because the sentiment analysis was performed earlier, the system suggests tailored drugs to the users. As for the drug recommendation system, it considers the patient's preferences, side effects, and prevailing views regarding the drug in order to recommend the most suitable one. This assists patients and healthcare practitioners in making appropriate selections, hence resulting in better treatment outcomes and greater trust in recommendation systems.

IV. RESULT AND DISCUSSION

4.1 Hamming Loss:

In multilabel classification, the percentage of label s that are predicted wrongly is measured using a st atistic known as Hamming Loss.It calculates the p roportion of labels that differ between the true and expected labels.A smaller Hamming Loss indicat es better model performance, or fewer incorrect pr edictions.

4.2 Jaccard Score:

The Jaccard Score, also known as the Jaccard Index or Intersection over Union (IoU), is a similarity metric for comparing sets. It evaluates the number of elements shared by the true and anticipated sets.

4.3 F1 Score:

The F1 Score, a performance indicator that balances recall and precision, is especially useful for unbalanced datasets. It is the harmonic mean of precision and recall to account for both false positives and false negatives. A higher F1 Score indicates better classification.

4.4 Precision:

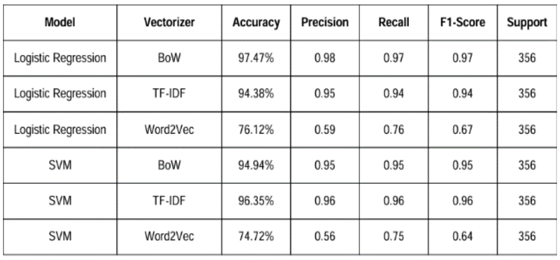
The accuracy of positive forecasts is measured by a parameter known as precision. It determines the proportion of successfully identified positive events out of all anticipated positives. In applications where false positives are costly, a model with high precision generates fewer false positive mistakes.

4.5 Recall:

In machine learning, recall is a crucial evaluation metric that assesses a model's capacity to accurately identify every pertinent positive instance in the dataset. It is computed as the ratio of true positives to the total of false negatives and true positives. The model is successfully capturing the majority of real positive cases when the recall value is higher performance.

4.6 Sentiment Analysis Results

To evaluate the performance of sentiment classification, multiple machine learning models were tested using three vectorization techniques: Bag of Words (BoW), TF-IDF, and Word2Vec. Support Vector Machine (SVM) with TF-IDF yielded the best results with an accuracy of 96.35% and F1 Score of 0.96



Model	Vectorizer	Accuracy	Precision	Recall	F1-Score	Support
Logistic Regression	BoW	97.47%	0.98	0.97	0.97	356
Logistic Regression	TF-IDF	94.38%	0.95	0.94	0.94	356
Logistic Regression	Word2Vec	76.12%	0.59	0.76	0.67	356
SVM	BoW	94.94%	0.95	0.95	0.95	356
SVM	TF-IDF	96.35%	0.96	0.96	0.96	356
SVM	Word2Vec	74.72%	0.56	0.75	0.64	356

Fig.2 Sentiment Analysis Result

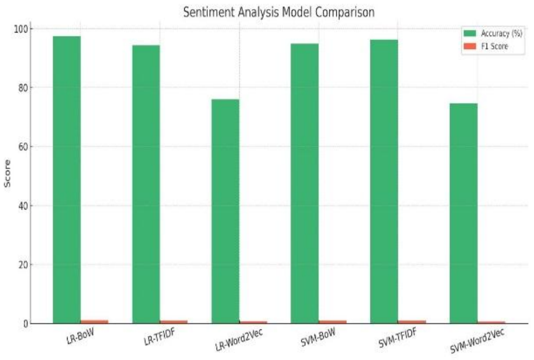
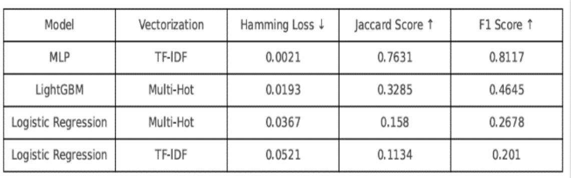


Fig.3 Sentiment Analysis Model compariosn

4.7 Disease Prediction Results

For disease prediction based on symptom inputs, four combinations of model and vectorization techniques were tested. The MLP with TF-IDF model demonstrated superior performance across all metrics.



Model	Vectorization	Hamming Loss ↓	Jaccard Score ↑	F1 Score ↑
MLP	TF-IDF	0.0021	0.7631	0.8117
LightGBM	Multi-Hot	0.0193	0.3285	0.4645
Logistic Regression	Multi-Hot	0.0367	0.158	0.2678
Logistic Regression	TF-IDF	0.0521	0.1134	0.201

Fig.4 Disease Prediction Results

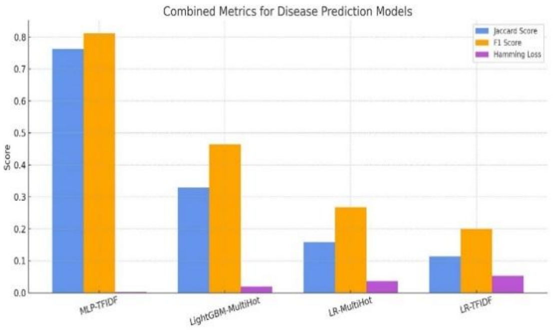


Fig.5 Combined Metrics for Disease Prediction Model

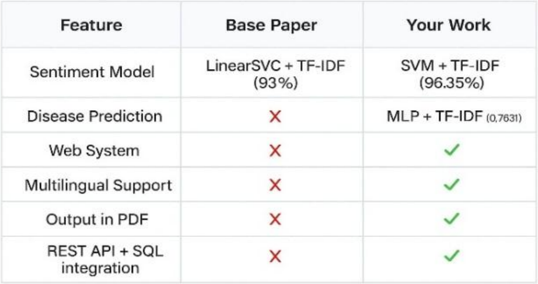
4.8 Evaluation Summary

The analysis shows that:

With high accuracy and F1 scores across both BoW and TF-IDF, SVM with TF-IDF is the best option for sentiment classification.

With the lowest Hamming Loss and a marked improvement over other models in Jaccard and F1 Scores, MLP with TF-IDF is unquestionably the best model for disease prediction.

Both models were incorporated into the deployed system for ultimate use in light of these findings.



Feature	Base Paper	Your Work
Sentiment Model	LinearSVC + TF-IDF (93%)	SVM + TF-IDF (96.35%)
Disease Prediction	✗	MLP + TF-IDF (0.7631)
Web System	✗	✓
Multilingual Support	✗	✓
Output in PDF	✗	✓
REST API + SQL integration	✗	✓

Fig.6 Comparison between Base Paper and our work

V. CONCLUSION AND FUTURE WORK

In this study, we developed and put into practice a real-time web-based drug and disease recommendation system that combines sentiment analysis with disease prediction based on symptoms. Our system outperformed numerous existing methods documented in previous studies with a high accuracy of 96.35% when using Support Vector Machine (SVM) with TF-IDF for sentiment classification. Furthermore, we used Multi-Layer Perceptron (MLP) with TF-IDF to introduce a novel disease prediction component. Its efficacy for multi-class classification problems was validated with an F1 Score of 0.8117 and a Jaccard Score of 0.7631.

Our dual-model approach bridges both approaches, providing comprehensive support for patients who might not have access to professional healthcare advice, in contrast to the majority of existing models that only focus on sentiment or treatment prediction. Using HTML, CSS, JavaScript, Node.js, and SQL, the system is implemented as a multilingual web application with intuitive features like downloadable PDF reports.

A third module that makes recommendations for appropriate therapies or medication combinations based on the anticipated disease can be added to our system to increase its functionality and scalability. For more precise forecasts, transformer-based models such as BERT or Bio BERT can also improve sentiment and symptom comprehension.

Voice input and data from wearable devices can be combined to improve accessibility, particularly for older or disabled users. Furthermore, the system will become more dependable and user-friendly by connecting to reputable medical APIs like Drugs.com or MedlinePlus, which will offer the most recent data on drug safety and side effects.

REFERENCES

[1] Aggarwal, Charu C. Recommender Systems. Vol. Cham: Springer International Publishing, 2016.

[2] Brownlee, Jason. *Master Machine Learning Algorithms*. Machine Learning Mastery, 2020.

[3] Goldberg, David, David Nichols, Brian M. Oki, and Douglas Terry. "Using Collaborative Filtering to Weave an Information Tapestry." *Communications of the ACM* 35, no. 12 (1992): 61–70.

[4] Lu, Jie, Di Wu, Min Mao, Wei Wang, and Guanghua Zhang. "Recommender System Application Developments: A Survey." *Decision Support Systems* 74 (2015): 12–32.

[5] Kim, Yoon. "Convolutional Neural Networks for Sentence Classification." *arXiv preprint* arXiv:1408.5882 (2014).

[6] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543, 2014.

[7] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[8] He, Xiangnan, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. "Neural Collaborative Filtering." In *Proceedings of the 26th International Conference on World Wide Web*, 173–182, 2017.

[9] Johnson, Jeff, and Yizhou Zhang. "Learning User Representations in Personalized Recommender Systems." In *Proceedings of the International Conference on Big Data*, 1231–1240, 2015.

[10] Schafer, J. Ben, Joseph A. Konstan, and John Riedl. "E-Commerce Recommendation Applications." *Data Mining and Knowledge Discovery* 5, no. 1–2 (2001): 115–153.

[11] Resnick, Paul, and Hal R. Varian. "Recommender Systems." *Communications of the ACM* 40, no. 3 (1997): 56–58.

[12] Liu, Jing, Hongzhi Zhou, and Lei Chen. "Context-Aware Recommender Systems: A Survey." *Journal of Software Engineering and Applications* 13, no. 4 (2020): 123–134.

[13] Ruder, Sebastian. "Transfer Learning in Natural Language Processing." In *Proceedings of the NAACL-HLT 2019*, 1–4, 2019.

[14] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30 (2017).

[15] Chen, Tianqi, and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794, 2016.

[16] Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. "LIBLINEAR: A Library for Large Linear

Classification." *Journal of Machine Learning Research* 9 (2008): 1871–1874.

[17] Hinton, Geoffrey, Nitish Srivastava, and Kevin Swersky. *Neural Networks for Machine Learning: Lecture 6a – Overview of Mini-Batch Gradient Descent*. University of Toronto, 2012.

[18] Li, Hang, and Jun Xu. "Semantic Matching in Search." *Foundations and Trends® in Information Retrieval* 7, no. 5 (2014): 343–469.

[19] Tang, Duyu, Bing Qin, and Ting Liu. "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1422–1432.

[20] Medsker, Larry, and Lakhmi C. Jain. "Recurrent Neural Networks: Design and Applications." *Design and Applications* 5 (1999): 64–67.

[21] Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. "Learning Word Vectors for Sentiment Analysis." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150, 2011.

[22] Nguyen, Thien Huu, and Kazunari Shirai. "PhraseRNN: Phrase Recursive Neural Networks for Aspect-Based Sentiment Analysis." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2500–2505.

[23] Pang, Bo, and Lillian Lee. "Opinion Mining and Sentiment Analysis." *Foundations and Trends® in Information Retrieval* 2, no. 1–2 (2008): 1–135.

[24] Rajpurkar, Pranav, Robin Jia, and Percy Liang. "Know What You Don't Know: Unanswerable Questions for SQuAD." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 784–789, 2018.

[25] Wang, Chong, David M. Blei, and Fei-Fei Li. "Simultaneous Image Classification and Annotation." In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1903–1910, 2017.

[26] Zhang, Xiang, Junbo Zhao, and Yann LeCun. "Character-Level Convolutional Networks for Text Classification." *Advances in Neural Information Processing Systems (NIPS)* 28 (2015): 649–657.

[27] Ioffe, Sergey, and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 448–456, 2015.

[28] Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. "Learning Long-Term Dependencies with Gradient Descent Is Difficult." *IEEE Transactions on Neural Networks* 5, no. 2 (1994): 157–166.