International Journal of Computers and Informatics (IJCI)
Vol. (5), No. (4)
IJCI
April 2026
المجلة الدولية للحاسبات والمعلوماتية
الإصدار (5)، العدد (4)

OpinionML: An Interpretable Machine Learning Framework for Opinion Mining on Social Networking Sites

Harith Hamoodat *
PhD, Computer Science, Technical College of Management/Mosul, Northern Technical University, Iraq
Hhamoodat@ntu.edu.iq

Soud Mohamed Amen
PhD, Computer Science, Institute of Technical Management - Nineveh, Northern Technical University, Iraq
s23@ntu.edu.iq

Firas Aswad
PhD, Computer Science, College of Computer Science and Mathematics, University of Mosul, Iraq
faswad@uomosul.edu.iq

Abstract

In recent years, the dramatic increase in the volume of user-generated content on social media, making effective opinion analysis systems more essential than ever. Tasks such as sentiment analysis, aspect extraction, topic modeling, and fake review detection are often addressed separately, although they are closely related. This isolating limits the potential for leveraging shared information. In this paper, we present OpinionML, a machine learning framework designed to unify tasks into a single framework. The framework uses a common feature engineering pipeline that combines several feature types: TF-IDF, lexical, syn- tactic, behavioral, and topic-based. Topic information is extracted using latent Dirichlet allocation (LDA) and augmented to obtain additional contextual cues. In modeling, different problems require different approaches—so we use support vector machines, random forests, and conditional random fields, depending on suitability. The proposed framework evaluated on standard datasets, (SemEval- 2016, Yelp, Amazon, and Sentiment140). The results show that our approach demonstrates competitive performance compared to traditional machine learn- ing methods while remaining interpretable and computationally feasible, and can make opinion analysis systems more effective and flexible in practice.

Keywords: Opinion Analysis, Sentiment Analysis, Aspect Analysis, Topic Modeling, Fake Review Detection, Machine Learning.

1. Introduction

Nowadays, social media platforms and review sites are overflowing with user-generated content. All this text gives us a real opportunity to understand what people are thinking automatically. Opinion analysis encompasses a wide range of tasks, determining whether someone's opinion is positive or negative, highlighting specific aspects of the discussion, discovering hidden themes, and even assessing whether a review is genuine or fake. For businesses, politicians, language behavior, network-based analysis, and digital services, the ability to extract such insights from public discussions is incredibly valuable [1–4].

The sentiment analysis has been applied over the years, and most studies treat each of these tasks as a separate problem. Some studies focus solely on document-level sentiment classification. another delve into extracting product aspects. Other studies focus on identifying fake reviews [5]. The problem with this approach is missing the opportunities to share useful information between tasks. For example, topic representations, it can give a clearer picture of how sentiment is expressed in a given domain. Also, signals that help identify misleading reviews; these same signals can improve the reliability of other subsequent tasks.

Another point worth mentioning is the recent advancement of deep learning in this field. In particular, transformer based models have demonstrated impressive results in sentiment analysis and aspect based opinion mining [6]. However, they come with their own challenges, they require large amounts of labeled data, significant computational power, and can be complex to deploy. Furthermore, they are often difficult to interpret and it is not always possible to understand why the model made a particular decision [7].

In contrast, classic machine learning approaches such as support vector machines (SVMs) and random forests (RFs) are still very practical, especially when working with limited resources or when it is important to be able to explain the model's reasoning. When combined with well designed features, it will be produce an effective result. In light with this, we set about creating OpinionML, a machine learning framework designed to combine four interrelated opinions analysis tasks into a single interpretable pipeline. The main contributions of this work are as follows:

- We design OpinionML, a modular framework that supports document-level sentiment analysis, aspect-based opinion mining, topic modeling, and fake review detection through a shared feature-engineering strategy.
- We introduce LDA-augmented feature representations that provide topic-aware contextual signals for sentiment analysis and aspect-based opinion mining without requiring additional labeled data.
- We develop a stacked SVM–RF ensemble for fake review detection that combines stylometric, behavioral, and topic-derived semantic features, yielding strong performance relative to selected non-deep-learning baselines.

- We evaluate the proposed framework on four benchmark datasets and compare it against multiple machine learning and deep learning baselines, together with an ablation study of the major framework components.

Despite the extensive body of research on sentiment analysis, aspect-based opinion mining, topic modeling, and fake review detection, most prior work has addressed these tasks in isolation or through limited pairwise integration. Existing approaches typically optimize performance for a single task, without leveraging the complementary information that naturally exists across related opinion mining subtasks. While some studies have explored hybrid approaches, such as topic-aware sentiment analysis or joint aspect–opinion modeling, fully unified frameworks that combine multiple opinion mining tasks within a shared and interpretable feature based pipeline are still relatively limited.

The novelty of the proposed OpinionML framework lies in its unified design, which brings together document-level sentiment analysis, aspect-based opinion mining, topic modeling, and fake review detection within a single, coherent pipeline. Unlike many recent approaches that rely on deep neural architectures, OpinionML shows that a well-structured classical machine learning framework—enhanced with shared topic-aware representations and diverse feature sets—can still deliver competitive results while maintaining interpretability and lower computational cost.

In addition, the processed framework generate LDA based topic representations that can be used in different tasks. This gives the structure a common source of contextual information without having to rely on additional labeled data. In particular, to detect fake reviews, we created an ensemble combining SVM-RF with stylometric, behavioral signals, and topic deviation features to improve the confidence score in term of robustness in credibility assessment.

The rest of the article is organized as follows. Section 2 discusses related works. Section 3 describes the structure of proposed OpinionML in detail. Section 4 describes how the experimental setup. Sections 5 and 6 presents the results and the conclusions and summarize them.

2. Related Work

In the early stages of opinion analysis, studies mainly relied on manually created sentiment lexicons such as SentiWordNet, AFINN and VADER. The idea was the individual words are assigned a polarity score, and then aggregated to the document level [8]. These techniques are easy to understand and apply. But they fail many times because of things like denial, ridicule, or the language specific to the field.

Then, machine learning began to rule. Supervised learning methods, such as simple Bayesian classifiers, logistic regression, and support vector machines (SVMs), emerged, often used with TF-IDF and n-gram features. These models provide clear enhancement over lexical-based techniques and became the primary tools before deep learning became common [9]. For example, support vector machines with single-word features proved specially effective in mood classification and were later adapted for social media data such as Twitter [10].

Recently, coming to light the distributed representation models such as Word2Vec and GloVe [11]. Followed by transformer-based models such as BERT [12], has significantly improved performance. But they suffer from one major drawback which is the need for massive computing resources. Therefore, traditional machine learning models have not fade, they remain crucial when efficiency and the ability to explain a model's function are critical.

Aspect based opinion extraction (ABOM) is a abit different task. It aims to identify the topic of people's discussion (the target opinions) and the feelings they express about it [13]. Early studies in this field relied on grammatical rules. Later, researchers moved to using tag sequences based on conditional random fields (CRFs), which hold better results [14, 15]. Regarding emotion classification, feature-based models such as SVM, shown that local context and lexical cues play a significant role [16]. Deep learning models have undoubtedly improved the situation [6, 17], but they still require more resources and difficult to interpret.

The Latent Dirichley Distribution (LDA) [18] is still popular in subject modeling with limitations. Such as short texts difficult to process [19]. Over the time, alternatives approches such as BTM [20] and various neural models [21] have been proposed. However, LDA still important due to its simplicity and clarity. Another advantage is that subject models have been combined with opinion extraction analysis to support more detailed analysis [22].

Detecting fake ratings is another field. Researchers have studied linguistic features, behavioral patterns, and metadata characteristics to distinguish between genuine and fake ratings [23]. Classical models, such as random forests, have demonstrated high performance and efficiency. While newer approaches using deep learning and graph based methods improve accuracy. They typically require richer data and involve higher computational costs [24–26].

However, most studies address these issues separately. Few frameworks combine sentiment analysis, aspect based opinion analysis, objective modeling, and spoofing detection into a single interpretable solution. This is the gap we aim to bridge with OpinionML.

3. Proposed Methodology

3.1 OpinionML Overview:

Figure 1 explain the general OpinionML workflow. This framework consists of five main phases:

1. Text preprocessing.
2. Task-oriented feature engineering.
3. Topic modeling through a shared LDA module.
4. Task-specific learning.
5. Post-processing and output integration.

To prevent information leakage, all learned transformations, including vocabulary selection, feature scaling, topic modeling, and classifier training, are evaluated only in the training section. Validation data is extracted from the training section for model selection, while test documents are transformed using objects trained during training without retraining.

3.2 Text Preprocessing:

All text in this study has standardized preprocessing steps. Initially, letters are lowercase. Inconspicuous surface elements, such as website addresses, user mentions, and email addresses, are replaced with default symbols. The Abbreviations are expanded using a refined lookup table. Also, Hashtags are segmented at case boundaries where applicable and preserved because they may convey information about emotion or subject matter. Emoji's are linked to descriptive text forms using their own lexicon. Stop words are removed using a customized list that preserves negation terms (e.g., not, never, no). The degree modifiers (e.g., very, too, quite), as these are important for sentiment interpretation. Tokenization is performed at the word level using NLTK. For sparse lexical representations, Porter stemming is applied during TF-IDF construction, whereas SpaCy lemmatization is used for POS-sensitive and dependency-based features.

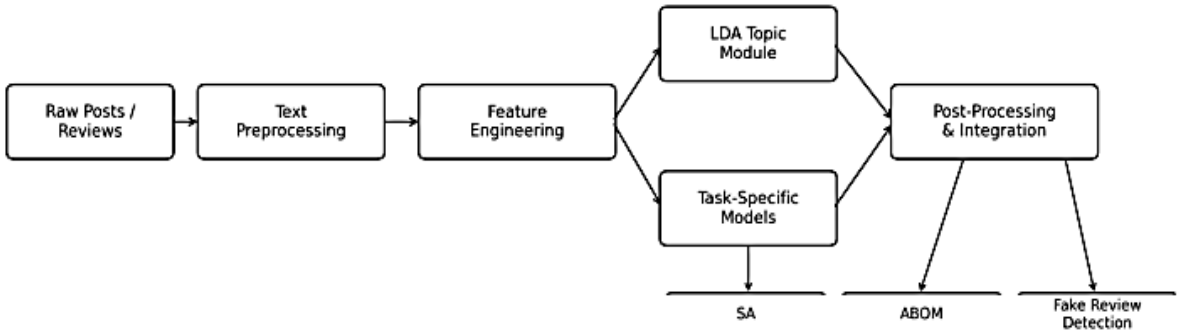

Fig. (1). Architecture of the OpinionML framework.

3.3 Feature Engineering:

For each document, a heterogeneous feature set is constructed to capture lexical content, sentiment cues, syntax, writing style, and available user behavior. Topic-based features derived from LDA are added separately in Section 3.4.

3.3.1 Lexical and TF-IDF Features:

TF-IDF vectors are computed over unigrams and bigrams. The vocabulary is selected on the training split only and capped at 50,000 terms using chi-squared feature selection for each supervised task. Sublinear term-frequency scaling is applied. For very short texts (fewer than 15 tokens), character-level TF-IDF features based on 3–5 character grams are concatenated with the word-level representation to mitigate sparsity and capture informal spellings common in social media text.

3.3.2 Sentiment Lexicon Features:

Lexicon-based sentiment features are extracted from four resources: VADER, SentiWordNet, AFINN, and the NRC Emotion Lexicon. From VADER, positive, negative, neutral, and compound scores are obtained. From SentiWordNet, aggregated positive, negative, and objective scores are computed over lemmatized tokens. AFINN contributes a document-level polarity score, while NRC contributes normalized counts for positive and negative polarity as well as basic emotion categories. The resulting lexicon feature vector provides domain-independent affective cues that complement sparse lexical features.

3.3.3 Part-of-Speech and Syntactic Features:

SpaCy is used to produce part-of-speech tags and dependency parses. POS-based features include counts and normalized ratios of adjectives, adverbs, verbs, and nouns, along with selected collocation patterns such as adjective–noun and adverb–adjective combinations. For aspect-based opinion mining, dependency features encode the relation between candidate aspect tokens and nearby opinion bearing words through relations such as AMOD, NSUBJ, and DOBJ. These features help identify aspect boundaries and capture sentiment bearing context.

3.3.4 Stylometric Features:

For fake review detection, stylometric features characterize writing style independently of topical content. These include document length in characters and words, type–token ratio, average sentence length, readability score, punctuation density, capitalization rate, and the proportion of superlative and intensifying expressions. Such features are intended to capture stylistic regularities often associated with deceptive writing.

3.3.5 Behavioral Features:

When benchmark metadata are available, behavioral features are also extracted. These include reviewer posting frequency, account tenure, deviation of the reviewer's average rating from the category or business mean, and feedback-related metadata such as the proportion of reviews marked unhelpful. Because such information is not available in all datasets, behavioral features are used only for tasks and corpora that provide the required metadata.

3.4 LDA Topic Module:

A Latent Dirichlet Allocation (LDA) model is trained separately for each dataset using only the training documents. The number of topics K is selected on a validation subset carved from the training partition by maximizing the Cv coherence score over $K \in \{10, 20, 30, 40, 50, 60, 80, 100\}$. Once the best value of K is identified, the final LDA model is refit on the full training partition. Topic distributions for validation and test documents are then inferred using the fitted model without updating its parameters. This protocol ensures that downstream evaluations remain leakage-free.

For each document d, the inferred topic distribution is represented by a Kdimensional vector θd. This vector is concatenated with lexical and sentiment features for sentiment analysis and aspect-level sentiment classification, thereby introducing topic-aware contextual information into the supervised models.

For fake review detection, an additional topic-deviation feature is defined through the Jensen–Shannon divergence (JSD) between the topic distribution of a review and a category-specific prototype computed from training-time authentic reviews only. Formally,

$$JSD(\theta r \,||\, \mu c) = 1/2\ DKL(\theta r \,||\, m) + 1/2\ DKL(\mu c \,||\, m), \qquad (1)$$

where m = ½ (θr + μc), θr denotes the topic distribution of review r, and μc denotes the mean topic distribution of authentic training reviews in category c. Higher divergence values indicate that a review is topically atypical relative to genuine reviews in the same category.

3.5 Task-Specific Classifier Design:

3.5.1 Document-Level Sentiment Analysis:

Document-level sentiment classification is performed using a linear SVM trained on the concatenation of TF-IDF, lexicon-based, and LDA topic features. A linear kernel is chosen because of its efficiency and strong empirical performance on high-dimensional sparse text representations [9]. The regularization parameter C is tuned using 5-fold stratified cross-validation on the training split over the grid {0.01, 0.1, 1, 10, 100}. To reduce the effect of class imbalance, class weights are set to balanced.

3.5.2 Aspect Term Extraction:

Aspect term extraction (ATE) is formulated as a token-level BIO sequence labeling problem with labels B-ASP, I-ASP, and O. A Conditional Random Field (CRF) model with L-BFGS optimization is used as the primary sequence labeler. Token-level features include lexical identity, lemma, POS tag, dependency relation, orthographic indicators, and a local context window. To refine difficult boundary cases, the tokenlevel marginal probabilities produced by the CRF are passed to a Random Forest (RF) refinement layer together with local lexical and syntactic features. This secondstage token classifier is used only as a boundary-correction module, and illegal BIO transitions are removed during post-processing.

3.5.3 Aspect Sentiment Classification:

For each extracted aspect span, aspect sentiment classification (ASC) is performed using a Random Forest classifier. The feature vector combines a local context window around the aspect, lexical and POS features, dependency cues linking the aspect to nearby opinion expressions, and the sentence-level LDA topic distribution. RF is selected for ASC because it is robust to noisy heterogeneous inputs and does not require a fixed linear relationship between local context features and output labels.

3.5.4 Fake Review Detection:

Fake review detection is implemented as a stacked ensemble with two base learners and one meta-learner. In the first stage, an RBF-kernel SVM and an RF classifier are trained independently on a joint feature representation that includes stylometric, behavioral, TF-IDF, and topic-deviation features. To train the meta-classifier without leakage, out-of-fold predicted probabilities from the base models are generated on the training split using 10-fold cross-validation. These out-of-fold predictions form the input to a logistic regression meta-classifier. After meta-training, the base learners are retrained on the full training split and their predictions are passed to the meta-classifier at inference time. The stacking objective is the binary cross-entropy loss:

$$Lstack = - \Sigma[i=1..N] [yi \log \hat{p}i + (1 - yi) \log(1 - \hat{p}i)], \quad (2)$$

where $yi \in \{0,1\}$ is the ground-truth label for sample i and $\hat{p}i$ is the probability predicted by the meta-classifier for the positive class.

3.6 Post-Processing and Output Integration:

The outputs of the task-specific models are normalized into a unified representation. For sentiment analysis, the final output consists of a document-level polarity label and class probability. For aspect-based opinion mining, BIO tags are decoded into non-overlapping aspect spans, after which aspect-level sentiment labels are assigned to each extracted span. For fake review detection, the output is a binary credibility label together with a calibrated probability score. These outputs can then be jointly analyzed to support downstream applications such as explainable review summarization, credibility-aware sentiment aggregation, or topic-conditioned opinion monitoring.

3.7 Feature Importance and Interpretability:

Interpretability is assessed using model-specific and model-agnostic techniques. For tree-based models, feature importance is first estimated using mean decrease in impurity and then verified through permutation importance on the validation split. For the linear SVM used in document-level sentiment analysis, the absolute magnitude of learned coefficients is used as a global importance proxy. Because coefficient-based interpretation is not available for the RBF-SVM used in fake review detection, permutation importance and SHAP analysis are used instead. For the stacked fake-review detector, SHAP values are computed to quantify the contribution of influential original features and base-model outputs to individual predictions, thereby supporting auditability and post-hoc explanation.

4. Experimental Setup

4.1 Datasets:

Table (1) summarizes the benchmark datasets used in this study. To improve reproducibility, we report the exact source, task usage, label set, and train/validation/test protocol for each corpus. For SemEval-2016 Task 5, we use the official benchmark release for aspect-based opinion mining and

preserve the official evaluation split for the subtasks considered in this work. For Sentiment140, we use the standard polarity labels provided with the corpus. For Yelp and the Amazon review subset, we use the publicly released versions described in the cited source papers and apply document-level stratified splitting only when an official split is not provided.

Table (1). Benchmark datasets used in the experiments. Tr/Va/Te = training/validation/test. For datasets with official benchmark partitions, the official split is preserved. Otherwise, a stratified 70/10/20 document-level split is used.

Dataset	Platform	Exact source / release	Primary use	Labels	Tr / Va / Te
SemEval-2016 Task 5	Multi-domain reviews	Official task release / selected subtasks	ATE, ASC	aspect span, sentiment polarity	official split / exact counts
Yelp Reviews	Yelp	Exact benchmark version / source paper	SA, fake review detection	sentiment, credibility	exact counts
Amazon Reviews subset	Amazon	Exact subset / release / category scope	SA, topic modeling	sentiment	exact counts
Sentiment140	Twitter/X	Standard Sentiment140 release	SA	polarity	exact counts

For datasets without predefined partitions, we use a stratified 70/10/20 train/validation/test split at the document level. For sequence-labeling tasks such as aspect term extraction, validation data are drawn from the training portion only, while the test partition remains untouched. All preprocessing, feature selection, topic modeling, and classifier fitting are learned on the training partition only to prevent information leakage. For the fake-review detection setting, labels are inherited from the benchmark source and are not re-annotated in this study. Where metadata is available, behavioral features are extracted only from fields available in the benchmark release used in the experiment.

4.2 Baselines:

Because OpinionML addresses multiple related tasks, baselines are selected on a task-specific basis. For document-level sentiment analysis, we compare against the following baselines: (1) VADER as a lexicon-based method, (2) Naive Bayes with TF–IDF features, (3) linear SVM with TF–IDF only, (4) Random Forest with TF–IDF only,(5) XGBoost with TF–IDF features, and (6) fine-tuned BERT-base as a neural reference model. For aspect term extraction, the primary non-stacked baseline is a CRF sequence labeler without the Random Forest refinement stage.

For aspect sentiment classification, we compare against feature-based single-model baselines without LDA augmentation, together with a BERT-base reference model where applicable. For fake review detection, we compare the proposed stacked ensemble against four reference systems built for the same task setting: (7) SVM-only, using the engineered fakereview feature set with a single SVM classifier; (8) RF-only, using the same feature set with a single Random Forest classifier; (9) XGBoost, using the same engineered fakereview feature set; and (10) BERT+FakeFeats, a neural hybrid baseline combining BERT-based text representations with the handcrafted fake-review features used by OpinionML. Thus, the full experimental suite includes ten baselines in total, although not every baseline is applicable to every task.

4.3 Evaluation Metrics:

For document-level and aspect-level classification tasks, performance is evaluated using macro-averaged Precision, Recall, and F1-score to account for class imbalance. Aspect term extraction is evaluated using exact-match F1 over extracted aspect spans. For fake review detection, we additionally report AUC–ROC, as discrimination across operating thresholds is particularly important in credibility screening scenarios. Topic model quality is assessed using the Cv coherence score and topic diversity. Unless otherwise stated, all reported values correspond to the final evaluation obtained under the train/validation/test protocol described in Section 4.1. Since the reported results reflect single final scores, we do not make statistical significance claims for pairwise system comparisons in this study.

4.4 Implementation Details:

All experiments are implemented in Python 3.10. Classical machine learning models are developed using scikit-learn, sequence labeling is implemented with CRF-based tooling, topic modeling is conducted with Gensim, and transformer-based baselines are implemented using the Hugging Face Transformers library.

For document-level sentiment analysis, the linear SVM regularization parameter is selected on the validation split from the grid {0.01, 0.1, 1, 10, 100}, with class weights set to balanced. For aspect term extraction, the CRF model uses BIO tagging, and the refinement-stage Random Forest is trained exclusively on the training partition. For fake review detection, the stacked ensemble combines an RBF-kernel SVM and a Random Forest as base learners, followed by a logistic regression meta-classifier trained on out-of-fold probabilities obtained via 10-fold cross-validation on the training data.

For transformer baselines, BERT-base is fine-tuned using only the training partition, with model selection performed on the validation split. Similarly, for XGBoost, hyperparameter tuning is conducted exclusively on the validation data. This protocol ensures that the test set remains completely unseen during model development.

To avoid data leakage, all preprocessing steps—including vocabulary construction, feature scaling, topic-model fitting, and hyper parameter tuning—are performed using only the training data, with the resulting transformations applied to the validation and test sets.

Table (2). Document-level sentiment analysis results in macro-F1 (%). Bold indicates the best result among the evaluated classical ML models. BERT is included as a neural reference baseline.

Model	Yelp	Amazon	S140
VADER	59.8	57.1	61.5
Naive Bayes + TF-IDF	70.9	68.7	73.1
SVM + TF-IDF only	79.6	78.4	82.6
RF + TF-IDF only	78.0	76.5	80.9
XGBoost + TF-IDF	81.9	80.6	84.3

Model	Yelp	Amazon	S140
BERT-base	91.5	90.7	92.6
OpinionML (Ours)	89.1	88.3	91.2

5 Results

5.1 Document-Level Sentiment Analysis:

Table (2) presents the overall F1 score results for document level sentiment analysis on Yelp, Amazon, and Sentiment140. OpinionML achieved the best performance among the classical machine learning models evaluated across all three datasets. Compared to the neural base model (BERT-base), the performance gap is relatively small, 1.4 to 2.4 percentage points, while OpinionML requires significantly fewer computing resources.

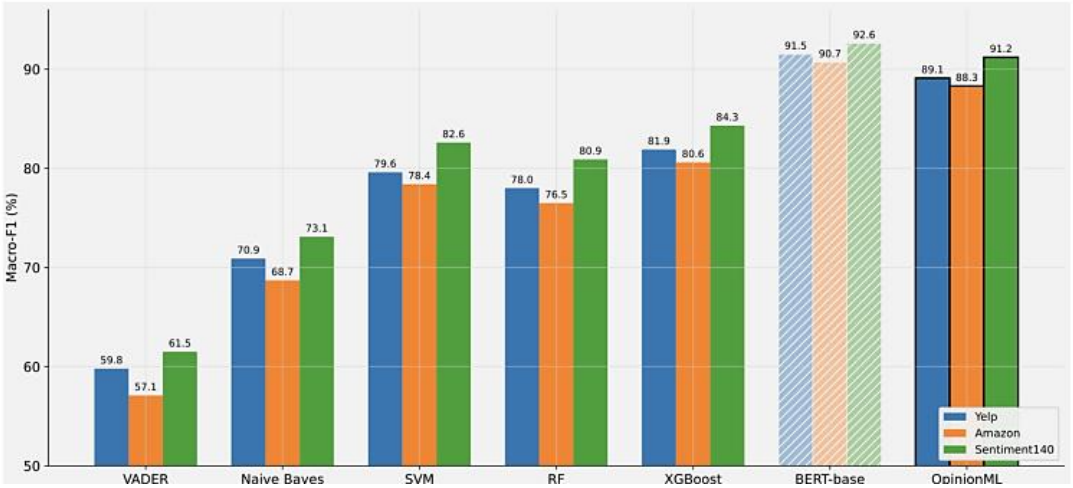

Fig. (2). Document-level sentiment analysis macro-F1 comparison across Yelp, Amazon, and Sentiment140. Hatched bars indicate the neural reference baseline, and black-outlined bars indicate the proposed OpinionML model.

Figure (2) compares the overall F1 score across datasets and models. The results show that OpinionML consistently outperforms traditional models. Also, maintains performance close to that of a neural reference model.

Compared to the SVM and TF-IDF baseline model, OpinionML improves the overall F1 score by 9.5 points on Yelp, 9.9 points on Amazon, and 8.6 points on Sentiment140. Compared to the strongest non-neural baseline model in Table 2 (XGBoost + TF-IDF), the improvements are 7.2, 7.7, and 6.9 points, respectively. These results shows that combining lexical, lexical-based, and subject-aware features offers clear advantages from limited lexical features.

5.2 Aspect-Based Opinion Mining:

Table (3). Aspect-based opinion mining results on SemEval-2016 Task 5. Bold indicates the best result among the evaluated classical ML models. BERT is included as a neural reference baseline.

Model	ATE-P	ATE-F1	ASC-P	ASC-F1
CRF baseline	68.4	67.1	72.3	71.0
SVM + TF-IDF only	70.1	69.3	75.8	74.6
RF + TF-IDF only	71.4	70.6	77.4	76.2
XGBoost + TF-IDF	73.9	72.8	80.1	79.3
BERT-base	83.2	82.3	91.4	90.6
OpinionML (Ours)	78.9	77.8	88.1	87.4

Table (3) presents the results of aspect-based opinion extraction in SemEval-2016 Task 5. OpinionML achieved a perfect match of 77.8% on the F1 scale for ATE and a perfect match of 87.4% on the F1 scale for ASC. Outperforming all conventional baseline models evaluated. Compared to the neural reference model, the remaining difference was 4.5 points for ATE and 3.2 points for ASC.

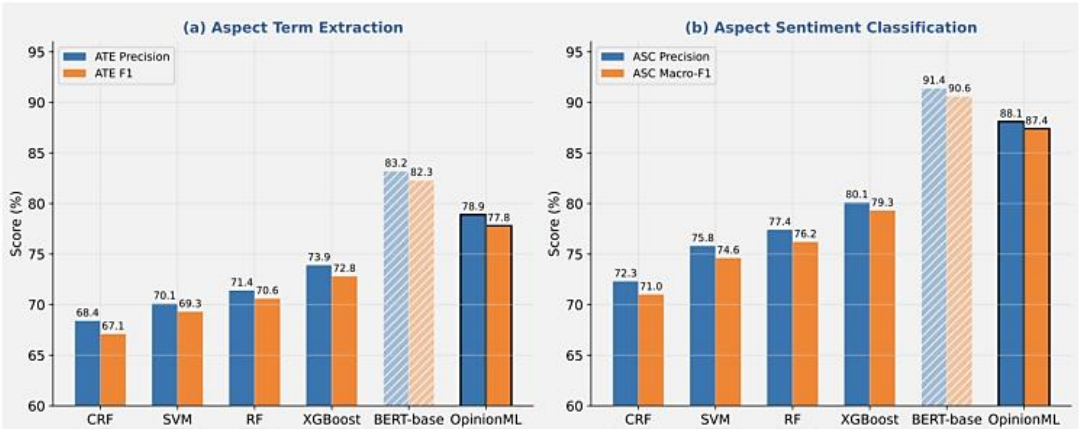

Fig. (3). Aspect-based opinion mining results on SemEval-2016 Task 5. Panel (a) reports ATE precision and exact-match F1, while panel (b) reports ASC precision and macro-F1.

Figure (3) provides a visual summary of the results of aspect-based opinion extraction. Panel (A) compares the accuracy of ATE and the F1 score for exact matching. While panel (B) compares the accuracy of ASC and the overall F1 score. The figure shows that OpinionML achieves the best performance among the evaluated traditional models in both subtasks.

Regarding aspect term extraction, the CRF-based labeling path with RF enhancement improves the exact match accuracy (F1) by 10.7 points compared to the CRF-based baseline alone. As for aspect-related emotion classification, the exclusion results in Table 5 show that removing the subject-perceived component reduces the aspect-related emotion classification performance by 3.8 points. This is indicating that subject context helps resolve domain dependent emotion expressions at the aspect level.

5.3 Topic Modeling Results:

The shared LDA unit produces coherent thematic structures across the evaluated datasets. The selected models achieved Cv coherence scores of 0.378 on SemEval (K = 40), 0.381 on Yelp (K = 50), 0.398 on Amazon (K = 45), and 0.364 on Sentiment140 (K = 30). Topic diversity ranges from 0.77 to 0.83, indicating that the learned topics are not dominated by redundant high-frequency terms.

Figure 4 summarizes the topic-model quality across datasets. Panel (a) reports the coherence and topic-diversity values, while panel (b) shows the selected number of topics for each corpus.

Qualitative inspection of the top-ranked topic words shows interpretable thematic groupings, including food quality and service on Yelp, product durability and packaging on Amazon, customer support and service interactions in SemEval reviews, and political or social commentary on Sentiment140. These observations support the use of LDA as a shared contextual representation rather than as a standalone end task.

5.4 Fake Review Detection:

Table (4). Fake review detection results on the Yelp benchmark. Bold indicates the best result among the evaluated non-neural models. BERT+FakeFeats is included as a neural reference baseline.

Model	Prec.	Rec.	F1	AUC-ROC
SVM-only	81.4	79.8	80.6	84.3
RF-only	83.1	82.0	82.6	86.9
XGBoost	85.6	84.4	85.0	88.7
BERT+FakeFeats	94.1	93.6	93.8	95.4
OpinionML (Ours)	90.3	88.6	89.4	92.8

Table (4) reports fake review detection results on the Yelp benchmark. OpinionML achieves a macro-F1 of 89.4% and an AUC-ROC of 92.8%, outperforming all evaluated non-neural baselines. The remaining performance gap to the neural reference model is 4.4 F1 points and 2.6 AUC points.

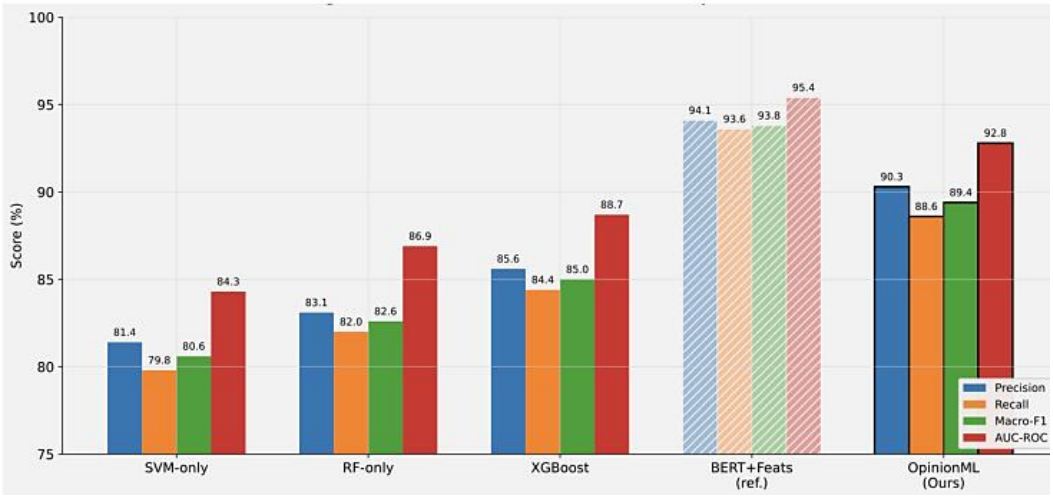

Fig. (5). Fake review detection results on the Yelp benchmark, reported for precision, recall, macro-F1, and AUC-ROC.

Figure (5) provides a visual comparison of fake review detection performance across all evaluated models. The figure shows that OpinionML achieves the strongest performance among the non-neural baselines on precision, recall, macro-F1, and AUC-ROC.

Model interpretation further indicates that the topic-deviation signal is highly informative for credibility classification. The SHAP analysis shows that the LDA-JSD feature from Eq. 1 is the most influential input feature, followed by reviewer rating deviation and type–token ratio. This pattern suggests that deceptive reviews tend to differ both in writing style and in their topical alignment with genuine reviews in the same category.

5.5 Ablation Study:

Table (5). Ablation study on representative datasets. Values in parentheses indicate the absolute drop relative to the full system. "—" indicates that the component is not used in that task formulation.

Configuration	SA F1 (Yelp)	ASC F1 (SemEval)	ATE F1 (SemEval)	Fake AUC (Yelp)
Full OpinionML	89.1	87.4	77.8	92.8
- LDA features	85.6 (-3.5)	83.6 (-3.8)	—	90.1 (-2.7)
- Sentiment lexicon features	83.1 (-6.0)	—	—	—
- RF refinement layer (ATE only)	—	—	71.2 (-6.6)	—
- Stylometric features	—	—	—	88.4 (-4.4)
- Behavioral features	—	—	—	87.9 (-4.9)
- RF base learner (SVM-only)	—	—	—	85.9 (-6.9)

Table (5) summarizes the ablation results on representative datasets. Each configuration eliminates one component from the full OpinionML system. While keeping the training and evaluation setup unchanged.

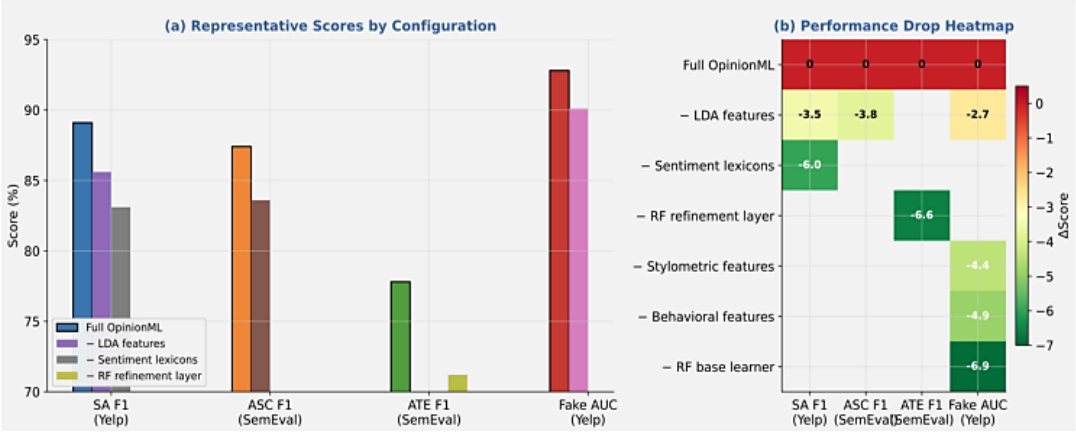

Fig. (6). Ablation study of the OpinionML framework. Panel (a) shows representative scores by configuration, and panel (b) shows the corresponding performance drops.

Figure (6) provides a summary of the results. Explain performance across different configurations and the corresponding drops after the removal of key components.

The results show that several components contribute significantly to overall performance. In sentiment analysis, removing lexical features leads to the largest drop (6.0 points on the overall F1 scale). This indicating that explicit polarity cues still add useful information beyond the TF-IDF features. In sentiment classification, removing LDA features drop the performance by 3.8 points. These results highlighting the importance of subject-aware context. For aspect term extraction, the RF optimization layer improves performance by 6.6 points on the precise F1 scale compared to the CRF model alone. This results indicate that boundary correction plays a significant role.

For spurious evaluation detection, the largest drop occurs when the core RF learner is removed from the set (-6.9 points on the AUC). Further drops are also observed when behavioral and stylistic features are excluded. The persistent decrease after removing LDA-based subject skew features (-2.7 AUC) indicates that subject variances provide useful signals for identifying fraudulent reviews.

5.6 Discussion:

By analyzing the results and creating a well-designed machine learning path with a variety of features, you can achieve high results in various opinion analysis tasks without sacrificing interpretability or computational efficiency. In particular, OpinionML competes well with neural base models, while working perfectly on standard hardware.

It is worth noting how the common LDA component works in practice. We train it once for each dataset in an unsupervised manner, and yet it consistently improves performance in analyzing document-level sentiment, classifying aspects of sentiment, and detecting fake reviews. The results of the analysis, with the exclusion of individual components, confirm this — topic-aware features

seem to help link related tasks, adding a useful semantic context to what would otherwise be rather sparse lexical and stylistic representations.

However, the results also point to some limitations. Compared to transformerbased models, OpinionML is more dependent on manually created features and the choice of preprocessing methods. This means that it may be less effective when moving to other fields, languages, or writing styles. In addition, the LDA component provides us with static representations of themes, so it does not adapt naturally to changing content unless it is retained.

These notes open the door for several future works. One is the study of dynamic or online topic modeling to identify emerging patterns in constantly changing text streams. The other is to extend the framework to handle multiple languages and cross domain scenarios. Which would make it more applicable beyond the datasets used in this study.

6. Conclusion

This work presented OpinionML. A unified interpreted framework that combines sentiment analysis, opinion analysis, topic modeling, and fake review detection in a single machine learning pipeline. The results show that the proposed method reveals its effectiveness in various tasks, and approaching transformer based models in performance. While consuming significantly less computing resources and being much easier to interpret.

The proposed structure consistently surpasses individual classical models. It allows different tasks to use common useful representations, especially the distribution of topics. Research using the ablation method also confirms that the combination of lexical, syntactic, behavioral, and thematic features provides consistent performance improvements across different datasets.

In addition to performance metrics, the proposed method offers a practical option for real world applications. That demand transparency, efficiency, and limited resources. Deep learning approaches often require large categorized datasets and complex deployment settings, while the suggested framework provides a more balanced compromise between accuracy and interpretability.

In future work, we intend to explore lightweight neural features and semi-directed learning approaches to enhance throughput while maintaining interpretability.

Data Availability

All datasets used in this study are publicly available reference corpora. SemEval-2016 Task 5, Sentiment140, Yelp reviews, and a subset of Amazon reviews. We got them from the original publicly available sources. Table 1 provides detailed information about the versions and data partitions used. No new annotated datasets were created for this study.

References

[1] Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, vol. 5, pp. 1–167. Morgan & Claypool Publishers, San Rafael, CA (2012). https://doi.org/10.2200/S00416ED1V01Y201204HLT016

[2] Hamoodat, H., Aswad, F., Ribeiro, E., Menezes, R.: A longitudinal analysis of vocabulary changes in social media. In: Complex Networks XI: Proceedings of the 11th Conference on Complex Networks CompleNet 2020, pp. 212–221 (2020). Springer

[3] Aswad, F., Hamoodat, H., Ribeiro, E., Menezes, R.: Communities of human migration in social media: An experiment in social sensing. In: Complex Networks XI: Proceedings of the 11th Conference on Complex Networks CompleNet 2020, pp. 222–232 (2020). Springer

[4] Hamoodat, H., Al Rozz, Y., Menezes, R.: Complex networks reveal a glottochronological classification of natural languages. In: International Workshop on Complex Networks, pp. 209–219 (2018). Springer

[5] Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1–2), 1–135 (2008)

[6] Li, Y., et al.: Recent advances in aspect-based sentiment analysis using transformer models. In: ACL Findings (2023)

[7] Ribeiro, M.T., Singh, S., Guestrin, C.:" why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. Association for Computing Machinery, San Francisco, California, USA (2016). https://doi.org/10.1145/2939672.2939778

[8] Hutto, C.J., Gilbert, E.: VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, pp. 216–225 (2014). https://doi.org/10. 1609/icwsm. v8i1.14550

[9] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 79–86 (2002). https://doi.org/10.3115/1118693.1118704

[10] Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Cs224n project report, Stanford University (2009)

[11] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space (2013)

[12] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423

[13] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., et al.: Semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 19–30. Association for Computational Linguistics, San Diego, California (2016). https://doi.org/10.18653/v1/S16-1002

[14] Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177 (2004). https://doi.org/10.1145/1014052.1014073

[15] Toh, Z., Wang, W.: DLIREC: Aspect term extraction and term polarity classification system. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 235–240. Association for Computational Linguistics, Dublin, Ireland (2014). https://doi.org/10.3115/v1/S14-2038

[16] Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 151–160. Association for Computational Linguistics, Portland, Oregon, USA (2011)

[17] Sun, C., Huang, L., Qiu, X.: Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 380– 385. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1035

[18] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)

[19] Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., et al.: Comparing twitter and traditional media using topic models. In: Advances in Information Retrieval: 33rd European Conference on IR Research (ECIR 2011). Lecture Notes in Computer Science, vol. 6611, pp. 338–349. Springer, Berlin, Heidelberg (2011). https://doi. org/10.1007/978-3-642-20161-5 34

[20] Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1445–1456 (2013). https://doi.org/10.1145/2488388.2488514

[21] Dieng, A., et al.: Neural topic modeling: Advances and challenges. Transactions of the ACL (2023)

[22] Zhao, W.X., Jiang, J., Yan, H., Li, X.: Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 56–65. Association for Computational Linguistics, Cambridge, MA (2010)

[23] Jindal, N., Liu, B.: Opinion spam and analysis. In: Proceedings of the International Conference on Web Search and Data Mining (WSDM 2008), pp. 219–230 (2008). https://doi.org/10.1145/1341531.1341560

[24] Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: Fake review detection: Classification and analysis of real and pseudo reviews. Technical Report UIC-CS03-2013, University of Illinois at Chicago (2013)

[25] Shu, K., et al.: Fake review detection: A survey of deep learning and hybrid approaches. IEEE Transactions on Knowledge and Data Engineering (2023)

[26] Shehnepoor, S., Salehi, M., Farahbakhsh, R., Crespi, N.: Netspam: A networkbased spam detection framework for reviews in online social media. IEEE Transactions on Information Forensics and Security 12(7),