Term Project: Sentiment Classifier Using Deep Learning Approach

Youngjun Yu

1 Introduction

In these days, people share their thoughts online every day such as in reviews and SNS. A sentiment classifier read these texts and tell us if they are positive or negative. This lets people quickly understand how users feel without reading messages by hand. For example, a restaurant can track reviews to see if customers like their service and food. By using sentiment analysis, organizations can make faster, smarter decisions and keep their customers happy.

2 Problem Definition

Sentiment analysis is a subfield of Computational Linguistics that focuses on automatically identifying the sentiment or opinion expressed in a piece of text. In this task, the system takes as input any text such as a product review, a social media post, or a comment where the writer shares their feelings about a specific item. The model processes this text and outputs a label such as "positive" or "negative" that best matches the writer's attitude. In other words, sentiment analysis turns natural language opinions into clear, structured data that can be used to understand how people feel about products or items.

3 Previous Work

Sentiment analysis has a long history and many different ways to solve the problem. As mentioned above, it is the task of identifying people's opinions or emotions from text, and it has been studied since the early 2000s. One of the most well-known early works is by Pang et al. (2002), who used machine learning algorithms like Naive Bayes and Support Vector Machines to classify movie reviews as positive or negative. Before that, simple rule-based or lexicon-based methods were used, which relied on counting positive or negative words using a predefined dictionary. These methods were easy to apply but could not understand context very well. Later, deep learning models such as CNNs and LSTMs became popular because they could learn patterns in sentences and understand the flow of language better. In recent years, transformer-based models like BERT (Devlin et al., 2018) have achieved the best performance. These models are pre-trained on large text datasets and can be fine-tuned for sentiment analysis with high accuracy.

4 Approach

In this project, I used supervised neural networks built in PyTorch. I designed a baseline system using simple feed-forward DNN with word embeddings. Then, I tested this baseline system with other three different architectures: a bidirectional LSTM, a CNN, and a Transformer. I ran each model on two datasets—the Cornell movie review polarity set (2,000 labeled reviews) and a Yelp restaurant review set (10,391 reviews, binarized to positive if ≥ 4 stars, negative otherwise). Before training, I cleaned and tokenized the text by lowercasing and splitting on whitespace, built a vocabulary of the top 10,000 tokens, and padded or truncated each review to 200 tokens using MAX_SEQ_LEN variable.

For features, I compared three embedding strategies: Random embeddings, GloVe embeddings frozen, GloVe embeddings fine-tuned and a TF-IDF weighted average of pretrained embeddings. Each feature set was fed into our DNN baseline for fair comparison, and I measured both accuracy and training time to identify which

combination of model and feature representation is the best.

In addition to those experiments, I built an improved system by combining the Transformer encoder with fine-tuned GloVe embeddings, which has the highest accuracy in the experiments above. I trained this model on both the movie review and Yelp datasets, using the same preprocessing and tokenization steps, and then also recorded its accuracy and training time on each dataset. Comparing these results to our earlier runs allowed to see how the Transformer + GloVe-ft approach improved overall performance.

5 Results and Discussion

For algorithm comparison, I trained four models (DNN, LSTM, CNN, and Transformer) on the movie review and Yelp datasets. I evaluated each with cross-entropy loss and overall accuracy. I trained the models with epoch 5, learning rate 0.001 and Adam optimizer. On the smaller Pang/Lee movie set, the Transformer led with 59.0% accuracy (loss 0.704) in about 73 s, outperforming the DNN (56.8% in 2 s), CNN (57.0% in 15 s), and LSTM (54.2% in 93 s). On the larger Yelp set, CNN was best at 83.0% accuracy (loss 0.381) in 67 s, slightly ahead of the DNN (81.0% in 5 s), Transformer (81.4% in 365 s), and LSTM (76.7% in 104 s).

Model	Train Time (s)	Loss	Accuracy
Pang/Lee Movie Reviews			
DNN	2.10	0.666	0.568
LSTM	92.90	0.807	0.542
CNN	14.88	0.676	0.570
Transformer	73.59	0.704	0.590
Yelp Restaurant Reviews			
DNN	5.43	0.431	0.810
LSTM	104.29	0.566	0.767
CNN	66.98	0.381	0.830
Transformer	364.74	0.461	0.814

Table 1: Training time, loss, and accuracy for various models on the Pang/Lee and Yelp datasets.

Transformer Model outperforms the other architectures because of its self-attention mechanism. It can capture long dependencies and global context more effectively than a simple DNN, a sequential LSTM, or a local-window CNN. Unlike the DNN, which ignores word order, or the CNN, which only sees nearby n-grams, the Transformer attends to every token in the sequence at once. This allows it to understand complex patterns.

For feature engineering on the DNN baseline, I compared random embeddings, frozen GloVe, fine-tuned GloVe, and TF-IDF weighted average embeddings. Fine-tuned GloVe ("glove-ft") yielded 69.3% accuracy on Pang/Lee in 2.2 s and 83.7% on Yelp in 5.6 s, outperforming random and frozen embeddings. TF-IDF averaging was competitive on Pang/Lee (70.8% in 0.16 s) but lagged on Yelp (72.9% in 0.85 s).

Model	Train Time (s)	Accuracy
Pang/Lee Feature-Engineering		
rand-emb	2.43	0.600
glove-fz	1.86	0.500
glove-ft	2.22	0.693
tfidf-avg	0.16	0.708
Yelp Feature-Engineering		
rand-emb	6.08	0.792
glove-fz	3.19	0.605
glove-ft	5.56	0.837
tfidf-avg	0.85	0.729

Table 2: Comparison of training time and accuracy for different feature-engineering strategies on the Pang/Lee and Yelp datasets.

Among feature-engineering strategies, fine-tuned GloVe embeddings ("glove-ft") gave the best results because they start with rich, pretrained word vectors. In contrast, frozen GloVe ("glove-fz") keeps the embeddings fixed, so it cannot learn domain-specific shifts in word usage or sentiment nuance, which leads to weaker performance. Random embeddings must learn all word meanings from scratch, which is slow and inefficient, and TF-IDF weighted averages, while quick to compute, cannot fully exploit the pretrained semantic space or sequential structure.

Based on these results, I selected the Transformer architecture and fine-tuned GloVe embeddings for improved system. This combined model achieved 67.5% accuracy on the movie reviews (loss 2.092, train time 306 s) and 81.8% on Yelp (loss 0.644, train time 725 s). While training time increased substantially compared to our fastest runs, the Transformer with GloVe-ft approach delivered the best balance of accuracy improvements on the smaller dataset and strong performance on the larger Yelp set, demonstrating that fine-tuning pretrained embeddings within a self-attention framework can produce meaningful gains when time and resources permit.

Dataset	Train Time (s)	Loss	Accuracy
Pang/Lee Movie Reviews	306.12	2.092	0.675
Yelp Restaurant Reviews	724.93	0.644	0.818

Table 3: Performance of the improved Transformer+GloVe-ft system on both datasets.

Yelp reviews tend to yield higher accuracy than movie reviews because they are generally shorter, more focused, and use more direct sentiment language. When people review a restaurant, they

often comment on a few clear aspects and they summarize their overall feeling in a single sentence or two. In contrast, movie reviews usually include long plot descriptions, comparisons to other films, and nuanced praise or criticism that can flip several times. Those mixed opinions make it harder for the model to decide on a final label. Another reason to yield higher accuracy in Yelp reviews can be it just has a larger dataset.

Following is 5 misclassified examples on Pang/Lee movie review dataset.

- **Misclassified Example 1 (True Label is *NEG* but predicted *POS*):**
  " spawn " features good guys , bad guys , lots of fighting , bloody violence , a leather-clad machine gun chick , gooey , self-healing bullet holes , scatological humor and a man-eating monster . it not only appears to have been tailor made for a swarm of 12- and 13-year-old boys , it appears to have been made by them . in a classic example of telling and not showing , " spawn " opens with a truckload of mumbo jumbo about forces of darkness , forces of light and how " men are the ones who create evil on earth . " so much for a message . (rest of the content omitted)
- **Misclassified Example 2 (True Label is *NEG* but predicted *POS*):**
  synopsis : al simmons , top-notch assasin with a guilty conscience , dies in a fiery explosion and goes to hell . making a pact with malebolgia , a chief demon there , simmons returns to earth 5 years later reborn as spawn , a general in hell's army donning a necroplasmic costume replete with knives , chains , and a morphing cape . sullen , wise cogliostro and flatulating , wisecracking violator vy for spawn's attention . comments : when todd mcfarlane left marvel comics (where he had made a name for himself as a first-rate comic book penciller on the " spider-man " titles) to join the newly-formed , creator-owned image comics , a new comic book legend was born : spawn . (rest of the content omitted)
- **Misclassified Example 3 (True Label is *NEG* but predicted *POS*):**
  way of the gun is brimming with surprises , some good , most bad . one of the good ones is ryan phillippe's surprisingly halfway decent performance . after the actor gained much attention by posing and preening through teen swill like i know what you did last summer , he hinted at a bit growth in last year's cruel intentions with his amusingly contemptuous john malkovich meets james spader performance , though his acting in that film faltered around the third act mark , precisely when the screenplay made his character grow a heart (rest of the content omitted)
- **Misclassified Example 4 (True Label is *POS* but predicted *NEG*):**
  seen february 15 , 1998 on home video (borrowed from chris wessell) . when it comes to modern gangster movies , it's really difficult to describe and review them without making comparisons to other films of the genre and/or just using the word " routine . " i've always subscribed to the philosophy that any idea (no matter how many times it's been used before) can provide for a good story and " donnie brasco " clinches this idea . it's not unlike most of the great films of the genre , yet it never apes another's style as it has a good layer of authenticity , even if its core is a tad stale . the film starts off in typical fashion by defining its atmosphere of new york city in the late 1970s and the mobsters who inhibit it . we meet lefty (pacino) , an aging wiseguy who can still walk the walk and talk the talk . (rest of the content omitted)
- **Misclassified Example 5 (True Label is *POS* but predicted *NEG*):**
  " take a number , fill out a form , and wait your turn . " starring kati outinen , kari v ? ? n ? nen , sakari kuosmanen , elina salo ; written & directed by aki kaurism ? ki ; cinematography by timo salminen it might be possible to call drifting clouds a satire or a black comedy , but that would imply a sense of anger , of vitriol , of energy ; drifting clouds is what you get when the rage and vitality are gone . it is the sad , slow story of lauri and ilona , a married couple caught between the wheels of capitalism as it grinds inexorably onward . he loses his job as a tram driver , because everyone drives cars nowadays . within a couple of months , she loses her position as a head waiter , when her restaurant is

bought out by a chain and the entire staff replaced . a conversation early in the film reveals a lot about their situation . (rest of the content omitted)

In Example 1, the review ends with a strong negative statement but opens with neutral plot description, so the model labeled it positive. Example 2 begins by praising the comic's art before criticizing the film's execution, and again our model focused on the early praise. In Example 3, phrases like "surprisingly halfway decent performance" masked the later negative critique. Example 4 is a mostly positive review of "Donnie Brasco" that uses words like "routine" to compare it to other films. These slightly negative terms led the model to predict negative. Finally, Example 5 contains nuanced praise of tone and empathy but also words such as "sad" and "futility," which made the prediction negative.

In all cases, the model struggles when positive and negative sentiments are mixed or when the true judgment is buried under plot details. To avoid these errors, I could split each review into sentences and give more weight to the final sentences and train the model on a small set of labeled contrast examples so it could learn to detect when sentiment actually flips.

Following is another 5 misclassified examples on Yelp reviews.

- **Misclassified Example 1 (True Label is *NEG* but predicted *POS*):**
  Severely overrated place. Their sugar-coated corn pancake topped with green onions was the most confusing appetizer I'd ever had. Their side dishes lack variety and freshness, chicken cutlet from their bentos was awfully hard, and their dak-bokkum-tang was a disaster. I tried here because my favorite teacher was a friend of the manager, but sorry I'm not going there again. I get why people like it though; a good place to start if you have zero tolerance for the authentic savory Korean flavor and all you need is an instagram photo of 'exotic' food.
- **Misclassified Example 2 (True Label is *NEG* but predicted *POS*):**
  I thought this was going to be the Za's from a couple years back but it is not. They took a proven concept and have butchered it with subpar ingredients and the use of a microwave. They used to cook the pasta with the sauce along with the ingredients which led to a great product that I was willing to pay for. This new Za's is putting ingredients on pasta, microwaving it, then pouring sauce and spices on after. I don't know who thought of this cooking process but it wasn't someone who knows how to cook. It was a bad experience, and now I feel a bit sick from it. If you value your stomach and tastebuds do not go to this establishment.
- **Misclassified Example 3 (True Label is *NEG* but predicted *POS*):**
  After going to Sushi Avenue, man does this place drop in my book. I already have serious qualms with that guy at the counter. And now I realize that Sushi Ichiban has the aesthetic appeal of a hospital waiting room. Unless you want a foam box full of average teriyaki, save your prime dollars for a sushi place that is worth going to... I'll give you a hint, it's on Green Street.
- **Misclassified Example 4 (True Label is *NEG* but predicted *POS*):**
  Restaurant review I recently visited this restaurant with a friend that came from out of town. The server was friendly and gave us plenty of time to order. While you wait you get complimentary popcorn with seasoning. Although it is good, I can't say I would fall head over heels for this. I ended up getting the sea scallops. Although I think they were fresh, the mixture of ingredients including blood orange and fennel puree was okay, but on the salty side. The highlight of the meal was the bread pudding that came out piping hot and was the perfect blend of chocolate and vanilla cream. (rest of the content omitted)
- **Missclassified Example 5 (True Label is *NEG* but predicted *POS*):**
  This place has so much potential, so it's kind of sad that they don't have better service.... My fiance and I waited at the bar for a drink for about 15 minutes before anyone asked us what we wanted. That may not sound too long, but the place wasn't even busy. And what's funny is the bartender was like purposely ignoring us. There's no way she didn't see us politely flagging her down. (rest of the content omitted)

On the Yelp dataset, similar issues appear. In Example 1, the reviewer lists strong negatives later ("disaster," "hard") but has some positive sentences like "I get why people like it", causing a false positive. Example 2 recounts past good experiences before condemning the cooking process, and Example 3 delays any negative comment until well into the review. In Example 4, a mix of pros like "friendly server" and "perfect bread pudding" precedes cons like "salty"), so positives dominated. Example 5 praises the ambiance and beer list before describing very poor service.

Here again, simple averaging treats early positive words equally with later negatives. To fix this, I could prioritize sentiment in the final sentences.

6 Conclusion and Future Work

Through this project, I came to appreciate the importance of sentiment analysis. I observed how performance and training time vary depending on the choice of algorithm or model, the dataset used, and the feature engineering techniques applied. By conducting error analysis, I identified the weaknesses of my models and explored ways to address them.

In future work, I would use these error analysis insights to design sentiment-specific models that pay special attention to contrastive phrases like "but" and "although," which often flip the overall sentiment. I also expect that fine-tuning large, pretrained language models such as BERT on our datasets would further improve performance.

References

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment Classification using Machine Learning Techniques.*

Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification.*

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*