Score Prediction in Trendyol Product Reviews

CSSM502 - Final Report

Osman Batur İnce · 84457

1 Introduction

Sentiment analysis is one of the most popular subfields of natural language processing. It can be defined as processing a portion of text to accurately predict its emotion, mainly categorized as positive, neutral, and negative. Companies can use sentiment analysis tools to estimate the public reception of a product. Companies can use APIs of social media platforms such as Twitter, Facebook, or other forums to acquire user comments about the product. Then, by using sentiment analysis tools, they can determine how the public perceives the product and what the public likes or dislikes about it. These analyses can result in better customer service and more informed marketing and production decisions.

One downside of sentiment analysis with three classes (positive, neutral, and negative) is that it is too coarse-grained. For example, one consumer might absolutely like a product, and another consumer might perceive the product as just above mediocre. However, in sentiment analysis, these two types of comments are just clumped together into the big umbrella of ”positive” comments. Therefore, the three-class sentiment analysis does not utilize the text information to its full extent.

Therefore, we apply a more fine-grained prediction task to product reviews, namely the score prediction task. In this task, we predict five classes instead of three. Hence, score prediction is equivalent to a fine-grained version of sentiment analysis. The scores of one and two constitute the negative class, the three score is equivalent to the neutral class, and the four and five scores form the positive class. Compared to the three-class sentiment analysis task, we have a more fine-grained output.

For our score prediction task, we needed to collect unstructured text data where we also had access to scores. Keeping the objectives of the sentiment analysis task in mind, we decided to use product reviews. While there are large and popular review datasets in English such as Yelp

[1], Amazon (He and McAuley 2016), and IMDB (Maas et al. 2011) datasets, there is a space for more Turkish review datasets. Moreover, many Turkish review datasets lack information about their data curation strategy and products [2]. Therefore, we opted to scrape product review data from Trendyol.

Trendyol is one of the most popular e-commerce websites in Türkiye. We collected electronic reviews for our main dataset and some out-of-distribution reviews for our additional analysis. We provide more information about these strategies in the following sections.

In this work, the research questions that we seek to answer are as following:

- Does state-of-the-art deep learning models significantly outperform traditional machine learning models? Do the results hold for both in-distribution and out-of-distribution performance?
- Which categories do models have trouble making predictions?

Optional Perform score prediction on tech forums to measure consumer satisfaction

2 Related Work

As the score prediction task can be seen as a more fine-grained version of the sentiment analysis task, we will mainly focus on sentiment analysis literature.

While the automated text analysis can be backtracked to The General Inquirer (Stone et al. 1962), we will mainly focus on the techniques that use machine learning. Thus, we exclude several other works that perform sentiment classification but from other disciplines' points of view such as (Sack 1994; Hearst 1992; Huettner and Subasic 2000).

In the intersection of natural language processing and machine learning, the two seminal works that research sentiment analysis are Pang, Lee, and Vaithyanathan 2002; Turney 2002. Turney 2002 introduce an unsupervised algorithm to classify reviews as recommended or not recommended. The main idea behind this paper is to enable search engines to summarize

[1] https://www.yelp.com/dataset

[2] https://www.kaggle.com/datasets/mustfkeskin/turkish-product-review-dataset, https://huggingface.co/datasets/turkish_product_reviews

reviews. The work consists of three stages: first determining phrases that contain adverbs or adjectives, predicting the sentiment of each phrase with pointwise mutual information, and and averaging the sentiment of the phrases to make a final prediction for the sentence. This work is really simple as it uses basic information theory methods rather than elaborate machine learning techniques. Pang, Lee, and Vaithyanathan 2002 examine sentiment classification in movie reviews. The authors experiment with an IMDB movie review dataset and employ machine learning algorithms, including Maximum Entropy, Naive Bayes, and Support Vector Machines. In addition to experimenting with different algorithms they experiment with different features, such as unigrams, bigrams, and parts of speech to improve performance.

While the literature has focused on similar methods such as TF-IDF (Papineni 2001), starting with the impressive performance of (Collobert and Weston 2008) paper, the field moved more towards learned distributed word embeddings approach to represent words (Tomas Mikolov et al. 2013; Tomás Mikolov et al. 2013; Bojanowski et al. 2017). Until the introduction of Transformer architecture (Vaswani et al. 2017), the use of pre-trained word embeddings with sophisticated recurrent neural network architectures provided state-of-the-art results for a lot of tasks in the field, including sentiment analysis (Baziotis, Pelekis, and Doulkeridis 2017; Deshmane and Friedrichs 2017; Hao et al. 2017). After the introduction of Transformers, Devlin et al. 2019 created the BERT model, which provided more accurate and effective word embeddings compared to previous approaches which led to a widespread adoption of the BERT model (Souza and Filho 2022; Mutlu and Özgür 2022). While certain highly-engineered models could outperform these models, their applicability and generalization abities were seriously limited. Then, up until the introduction of GPT-3 (Brown et al. 2020). GPT-3 showed that large language models can be used for zero-shot and few-shot classification tasks. Building on top of GPT-3 with instruction tuning and reinforcement learning with human feedback, ChatGPT is currently the state-of-the-art natural language processing model on numerous tasks. Thus, there are works on sentiment analysis using GPT models (Kheiri and Karimi 2023).

3 Dataset

Trendyol's website is dynamic, not static. Therefore, we needed to wait for the dynamic scroll times to collect reviews. After five days of review collection, we end up with approximately 41000 reviews in total. As review collection took a lot of time, we opted for collecting data for a subset of products rather than the full product range.

We used Python and the Selenium library for the dataset collection. Firstly, I curated a small dataset of about 7000 reviews. However, thinking that 7000 reviews might be small for a task that has too many dimensions, I developed another dataset collection setup after the project proposal. In the first stage of scraping, we scraped 2800 product links from the Trendyol. Then, we fed these links to the second stage to scrape reviews from these products. After that, with some further preprocessing our dataset was ready.

We collected two types of datasets: **1) main dataset** and **2) out-of-distribution (OOD) dataset**.

The main dataset only contains electronics products. It contains 77 types of products encompassing a wide range of types, e.g. television, satellite receiver, and dishwasher. While we collected approximately 70000 reviews, some reviews did not have any text. Moreover, there were multiple identical reviews from different reviewers as well. After cleaning up the dataset, we end up with 40150 reviews from 2800 products and 415 brands. We can group the 77 product types into more high-level nine categories, namely Smart Home Devices, Wearable Tech, Printers & Scanners, TVs & Sound Systems, Personal Care Devices, White Appliances, PC & Tablet, Electronic Home Devices, Games and Game Console, Headphones, Phones & Accessories.

The OOD dataset contains supermarket, clothing, and cosmetics products such as shampoo, dog food, sweater, sunscreen, t-shirt etc. The whole dataset contains 1182 reviews from 43 different brands.

One of the disadvantages of working with reviews is that people usually prefer products with higher reviews. This feedback loop results in significantly larger amount of positive reviews compared to negative reviews. We can see the outcome of this feedback loop better in Figures 1 and 2.

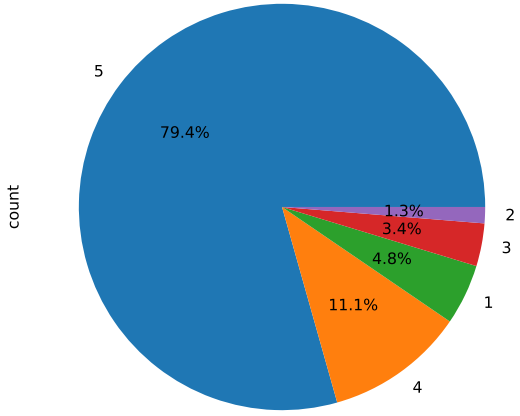

Figure 1: The pie chart to visualize the score distribution of the main dataset

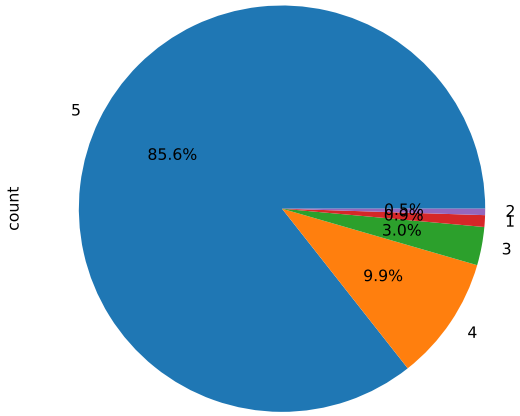

Figure 2: The pie chart to visualize the score distribution of the OOD dataset

Additionally same textual review for the same product might have different scores. For example, for the ”fiyat performans ürünü” review (price performance product in English), three users decided to use three different scores. This ambiguous phenomenon may result in models not achieving their full potential due to inter-personal variances.

4 Experimental Setup

For our experimental setup, we considered four models. For traditional machine learning models, we consider Naive Bayes and Random Forest models. For deep learning models, we con-

sider bidirectional LSTM (Bi-LSTM) (Huang, Xu, and Yu 2015) and BERT models. The BERT model performed absolutely worse than other models, therefore it is not included into the analysis. I will only describe the Bi-LSTM model as the remaining models are fairly well-known.

To understand what a Bi-LSTM is, we need to start from understanding recurrent neural networks (RNNs). RNNs are designed to work with sequences of data, where each element in the sequence is processed in a sequential manner. Traditional RNNs have challenges in capturing long-term dependencies due to the vanishing gradient problem. To alleviate this vanishing gradient problem, long short-term memory (Hochreiter and Schmidhuber 1997) is introduced. LSTMs are a specific type of RNN that addresses the vanishing gradient problem by introducing a memory cell mechanism. This allows LSTMs to capture and remember information over long sequences, making them more effective for tasks requiring understanding of context over time. The Bi-LSTM enhances the capabilities of LSTMs by processing the input sequence in both forward and backward directions. This bidirectional approach improves the model's ability to understand context and dependencies in both directions.

For hyper-parameter tuning, we performed 5-fold cross validation for ML models and 90/10 validation for Bi-LSTM. Bi-LSTM models are much larger than traditional ML models, thus they take a lot of time to train. To conduct the experiments in a reasonable amount of time, we performed hyperparameter tuning with 15 different trials. To evaluate our models, we used accuracy, precision, recall, and F1 score metrics.

Lastly, we were curious about the statistical significance of the performance discrepancies. Hence, we performed McNemar's test, which is a nonparametric statistical significance test that can be used for classification tasks. We performed binary statistical significance tests with a 5% confidence level between each pair of models to see whether there was a statistically significant difference between the compared models.

5 Results

5.1 IID Results

When we look at Table 1, we can see that the Bi-LSTM model performs better than other models in three out of four metrics. While Random Forest has a slightly higher precision than Bi-LSTM, Bi-LSTM is better on all other metrics. When we performed statistical significance tests, we observed that Bi-LSTM statistically outperforms the other two models while there is not a statistically significant difference between Naive Bayes and Random Forest.

	Precision	**Recall**	**F1 Score**	**Accuracy**
Naive Bayes	75.99	82.09	76.68	82.09
Bi-LSTM	76.46	**84.43**	**79.31**	**84.43**
Random Forest	**76.92**	81.90	75.40	81.90

Table 1: The results for the main dataset. The best result for each metric is bolded.

5.2 OOD Results

OOD results are shown in Table 2. We can see that Naive Bayes and Random Forest models seem tied as each model performs best on two metrics. Most importantly, while the Naive Bayes model has the highest F1 score, the Random Forest model has the highest accuracy. Our statistical significance tests show that Random Forest performs statistically significantly better than Bi-LSTM while Naive Bayes cannot outperform Bi-LSTM statistically significantly. However, Random Forest cannot outperform Naive Bayes statistically significantly.

	Precision	**Recall**	**F1 Score**	**Accuracy**
Naive Bayes	**80.19**	85.36	**81.02**	85.36
Bi-LSTM	76.46	84.43	79.31	84.43
Random Forest	76.89	**85.62**	79.75	**85.62**

Table 2: The results for the OOD dataset. The best result for each metric is bolded.

	Smart Home Devices	Wearable Tech	Printers & Scanners	TVs & Sound Systems	Personal Care Devices	White Appliances	PC & Tablet	Electronic Home Devices	Games & Game Console	Headphones	Phones & Accessories
Acc	92.39	86.49	85.14	83.73	83.37	82.87	82.32	81.94	81.63	79.79	78.68
Len	80.24	72.78	72.54	71.92	78.76	68.79	67.22	58.79	75.17	75.49	54.10
Score	4.74	4.73	4.53	4.57	4.72	4.58	4.65	4.67	4.53	4.55	4.37

Table 3: The average accuracies for high-level categories. The **Acc** row shows the average accuracies for these categories, the **Len** row shows the average character length of the review, while the **Score** row shows the average score for the reviews that belong to that class.

You will be provided with a product review, and your task is to predict its score as 1, 2, 3, 4, or 5. Give a single score. Do not make any excuses. Predict the most probable score.
Review: <REVIEW> Score:

Figure 3: The prompt template for ChatGPT evaluation.

5.3 ChatGPT Results

The prompt template that I provide is given in Figure 3. This is the prompt that I engineered after careful trial and error. Without the last sentences about prediction, ChatGPT just rambles without making a prediction. As seen in Figure 3, we evaluate our models in a zero-shot manner.

In Table 4, we have the preliminary evaluation results for ChatGPT. As evaluating thousands of examples is not feasible for ChatGPT without paying money, we only evaluate randomly selected 50 examples. However, ChatGPT's performance is limited compared to trained models.

	Precision	**Recall**	**F1 Score**	**Accuracy**
ChatGPT	79.24	56.00	64.00	56.00

Table 4: The results for a subset of the main dataset for only the ChatGPT dataset.

6 Analysis

In Table 3, we can see that there is a performance discrepancy between different categories. The table is created in a way that the categories are sorted from left to right by decreasing their average accuracies.

When we look at the table, we can see that categories that have higher average scores overall have more accuracy. The "Smart home devices" category has the highest accuracy and

it has the highest average score. Similarly, the ”Phones & Accessories” category has the lowest accuracy and it has the lowest average score. When we look at the categories in the middle, we see a few categories that do not fit the pattern such as ”Personal Care Devices” and ”Printers & Scanners”. However, the overall trend suggests that a higher average score results in better average accuracy category-wise.

A similar analysis can be made for length as well. When we look at the table, we can see that categories that have longer reviews overall have more accuracy. The ”Smart home devices” category has the highest accuracy and it has the highest average review length. Similarly, the ”Phones & Accessories” category has the lowest accuracy and it has the lowest average review length. When we look at the categories in the middle, we see a few categories that do not fit the pattern such as ”Personal Care Devices”, ”Headphones” and ”Wearable Tech”. However, the overall trend suggests that a higher average review length results in better average accuracy category-wise.

7 Discussion

Based on the Results section, selecting Bi-LSTM model for in-distribution predictions and Random Forest model for out-of-distribution makes more sense.

On another point, while the scale of our dataset is enough, the class imbalance may inflate unsuccessful models' performance as the class imbalance is more critical in the OOD dataset (see Figures 1 and 2). As the OOD dataset has more reviews with a score of 5, a class that incorrectly predicts the score of 5 in the main dataset might have higher OOD performance just because of the more significant class imbalance.

Similarly, ChatGPT and large language models might probably work better than traditional ML models in deployment while having low accuracies in the evaluation. There might be several reasons for this:

- **Inter-personal disagreement:** As shown in the Dataset section, the same textual review for the same product might have different scores. Even if two people think similarly, they might output different scores for the same reasoning. This reason is in the nature of the

task and not due to our dataset collection.

- **The scale of the data:** Large language models or deep learning models excel in scenarios where an abundant amount of data is available. While 40000 is a nice amount of review for training, deep learning models might need more data to gain upon traditional ML models.

8 Conclusion

In conclusion, this report explored the task of score prediction in Trendyol product reviews, aiming to provide a more fine-grained analysis than traditional sentiment analysis. The motivation behind this work was to overcome the limitations of three-class sentiment analysis by predicting five classes, corresponding to different score ranges.

The dataset was collected from Trendyol, a popular e-commerce website in Türkiye, focusing on electronic product reviews. Two datasets were created: the main dataset, consisting of electronics reviews, and an out-of-distribution (OOD) dataset, including reviews from supermarket, clothing, and cosmetics products.

The experimental setup involved four models: Naive Bayes, Random Forest, bidirectional LSTM (Bi-LSTM), and ChatGPT. Results indicated that the Bi-LSTM model outperformed Naive Bayes and Random Forest in in-distribution (IID) predictions, while Random Forest performed better in OOD predictions. ChatGPT, a large language model, showed promising results in a zero-shot evaluation.

Further analysis revealed performance discrepancies among different product categories, with higher average scores and longer reviews correlating with higher accuracy. Additionally, the class imbalance in the dataset may have influenced model performance, particularly in OOD predictions.

Ultimately, the choice of the best model depends on the specific deployment scenario, with Bi-LSTM excelling in in-distribution predictions and Random Forest showing promise for out-of-distribution scenarios. The findings also suggest that large language models like ChatGPT could be valuable in deployment despite their lower evaluation accuracies.

References

Baziotis, Christos, Nikos Pelekis, and Christos Doulkeridis (Aug. 2017). "DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Ed. by Steven Bethard et al. Vancouver, Canada: Association for Computational Linguistics, pp. 747–754. DOI: 10.18653/v1/S17-2126. URL: https://aclanthology.org/S17-2126.

Bojanowski, Piotr et al. (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5. Ed. by Lillian Lee, Mark Johnson, and Kristina Toutanova, pp. 135–146. DOI: 10.1162/tacl_a_00051. URL: https://aclanthology.org/Q17-1010.

Brown, Tom et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Collobert, Ronan and Jason Weston (2008). "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning". In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland: Association for Computing Machinery, pp. 160–167. ISBN: 9781605582054. DOI: 10.1145/1390156.1390177. URL: https://doi.org/10.1145/1390156.1390177.

Deshmane, Amit Ajit and Jasper Friedrichs (Aug. 2017). "TSA-INF at SemEval-2017 Task 4: An Ensemble of Deep Learning Architectures Including Lexicon Features for Twitter Sentiment Analysis". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Ed. by Steven Bethard et al. Vancouver, Canada: Association for Computational Linguistics, pp. 802–806. DOI: 10.18653/v1/S17-2135. URL: https://aclanthology.org/S17-2135.

Devlin, Jacob et al. (June 2019). "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). Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Hao, Yazhou et al. (Aug. 2017). "XJSA at SemEval-2017 Task 4: A Deep System for Sentiment Classification in Twitter". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Ed. by Steven Bethard et al. Vancouver, Canada: Association for Computational Linguistics, pp. 728–731. DOI: 10.18653/v1/S17-2122. URL: https://aclanthology.org/S17-2122.

He, Ruining and Julian J. McAuley (2016). "Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering". In: *CoRR* abs/1602.01585. arXiv: 1602.01585. URL: http://arxiv.org/abs/1602.01585.

Hearst, Marti A. (1992). "Direction-Based Text Interpretation as an Information Access Refinement". In: *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*. USA: L. Erlbaum Associates Inc., pp. 257–274. ISBN: 0805811893.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Comput.* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Huang, Zhiheng, Wei Xu, and Kai Yu (2015). "Bidirectional LSTM-CRF Models for Sequence Tagging". In: *CoRR* abs/1508.01991. arXiv: 1508.01991. URL: http://arxiv.org/abs/1508.01991.

Huettner, Alison and Pero Subasic (Jan. 2000). "Fuzzy Typing for Document Management". In: Association for Computational Linguistics.

Kheiri, Kiana and Hamid Karimi (2023). *SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning*. arXiv: 2307.10234 [cs.CL].

Maas, Andrew L. et al. (June 2011). "Learning Word Vectors for Sentiment Analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:*

Human Language Technologies. Ed. by Dekang Lin, Yuji Matsumoto, and Rada Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, pp. 142–150. URL: https://aclanthology.org/P11-1015.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Mikolov, Tomás et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1301.3781.

Mutlu, Mustafa Melih and Arzucan Özgür (May 2022). "A Dataset and BERT-based Models for Targeted Sentiment Analysis on Turkish Texts". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Ed. by Samuel Louvan, Andrea Madotto, and Brielen Madureira. Dublin, Ireland: Association for Computational Linguistics, pp. 467–472. DOI: 10.18653/v1/2022.acl-srw.39. URL: https://aclanthology.org/2022.acl-srw.39.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan (July 2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Association for Computational Linguistics, pp. 79–86. DOI: 10.3115/1118693.1118704. URL: https://aclanthology.org/W02-1011.

Papineni, Kishore (2001). "Why Inverse Document Frequency?" In: *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. URL: https://aclanthology.org/N01-1004.

Sack, Warren (1994). "On the Computation of Point of View". In: *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 2)*. AAAI'94. Seattle, Washington, USA: American Association for Artificial Intelligence, p. 1488. ISBN: 0262611023.

Souza, Frederico Dias and João B. O. Souza Filho (2022). “BERT for Sentiment Analysis: Pre-trained and Fine-Tuned Alternatives”. In: *CoRR* abs/2201.03382. arXiv: 2201.03382. URL: https://arxiv.org/abs/2201.03382.

Stone, Philip J. et al. (1962). “The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information”. In: *Behavioral Science* 7.4, pp. 484–498. DOI: https://doi.org/10.1002/bs.3830070412. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bs.3830070412. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bs.3830070412.

Turney, Peter (July 2002). “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 417–424. DOI: 10.3115/1073083.1073153. URL: https://aclanthology.org/P02-1053.

Vaswani, Ashish et al. (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.