Personal Pronouns as Features in Polarity Classification of Amazon Reviews

Joseph Blankenship
Dept. of Computer Science
Bowling Green State University
Bowling Green, USA
jkblank@bgsu.edu

Pramiti Barua
Dept. of Computer Science
Bowling Green State University
Bowling Green, USA
pbarua@bgsu.edu

Jaswitha Ravala
Dept. of Computer Science
Bowling Green State University
Bowling Green, USA
jravala@bgsu.edu

***Abstract*—An investigation into how personal pronouns can be used to improve the polarity classification of product review is presented. Polarity classification is the classification of text into categories ranging from positive to negative sentiment. Personal pronouns are often dismissed for use in feature selection in polarity classification models. Our approach explores the use of personal pronouns in several hybrid Naïve Bayes models that include positive and negative sentiment scores as a features. The inclusion of personal pronouns as a feature gives mixed results. Further research on larger datasets is necessary. In addition to this, suggestions are made for including personal pronouns in future polarity classification models.**

Keywords—Naïve Bayes, Polarity Classification, Sentiment, Product Reviews, Natural Language Processing

I. INTRODUCTION

In the fast growing era of different E-commerce websites and social media, sentiment analysis has become an important research area. Sentiment analysis in the domain of E-commerce that strives to use large datasets to increase market share through extracting opinions through reviews, comments, or other types of documents, often text. Natural language processing (NLP) is used in conjunction with machine learning (ML), deep learning, and statistics to analyze human text to understand it. Sentiment analysis is advantageous to organizations that want to produce better products and understand human sentiment toward the organization. By finding the root cause of issues facing customers, organizations may better serve customer interests and discover current trends. Many organizations have hundreds of thousands of products. In addition to this, companies without visibility in the marketplace may have fewer reviews.

Methods to analyze large amounts of texts vary widely, but the Naïve Bayes model can be trained quickly on large amounts of data. This means that the categorization of reviews into one of five-categories, can be done in a computationally inexpensive way. The categorization of reviews by positive or negative sentiment is also called polarity classification. In a hybrid approach, traditional ML methods are used with lexical analysis tools. Sentiwordnet 3.0 is a tool which provides polarity scores or sentiscores. These scores define the positivity or negativity of a given text. To explore effective combinations of features, we will use the Naïve Bayes model along with a polarity score. Through this sentiment analysis framework, we explore how personal pronouns affect polarity classification.

Standard sentiment analysis is mostly used to do research on a particular expression or topic and investigates the tone and/or opinion of products/services. This investigates whether the product/service is good, bad, or neutral. Traditionally, personal pronouns are filtered out when pre-processing review data. They are thought to not add much information to the polarity classification problem. Sentiwordnet 3.0 does not provide a score for personal pronouns. However, we explore if personal pronouns in product reviews have sentiment useful for classification. We perform polarity classification using polarity scores, normalized term frequency (TF), and custom polarity scores for personal pronouns. We assign personal pronouns positive and negative scores in separate models to explore the effect on polarity classification. We compare these four models in their ability to classify Amazon reviews into a five-star rating system. We try to determine if personal pronouns should be considered as features for classifying reviews into star ratings based on sentiment.

This paper is organized as follows: In section II we discuss related work, in section III the features selected for study, in section IV the methodology. Following this, section V presents the results and discussion, and after that section VI includes threats to validity, conclusion, and future research.

II. RELATED WORK

We first discuss related work on polarity classification. Then we discuss pre-processing, feature selection, and model selection. Finally, we identify a research gap.

Polarity classification has been performed at many levels. Shah et al. used three levels when classifying twitter texts for sentiment [1]. While levels can range widely, attempts to classify texts into five-star rating systems usually use five rating levels. Aljuhani et al. classified the reviews into three categories, positive, negative, and neutral [2]. One and two-stars were classified as negative. Three-stars were classified as neutral. Four and five-star were classified as positive. They found the best results with the bag of words trigram. Hogenboom et al. classified reviews into five categories and proposed a "bag-of-sentiword" hybrid method, which reduced documents to vectors of words which carried sentiment and excluded all others [3]. Ground truth data or labeled data used in polarity classification includes many different datasets. However, there have been relatively few studies done which use the users self-reported star rating to train and test statistical models. A review must have a star rating when it is posted so

users must think about it before posting their review. Fang and Zhan suggest that self-reported star ratings gives Amazon review data validity [4].

The most popular tool for NLP tasks in sentiment analysis was the Natural Language Toolkit (NLTK) [5]. Some approaches at pre-processing data include part of speech (POS) tagging, stop word filters, and eliminating duplicate words [5]. Shah et al. compressed words for their large dataset [1]. Song removed numbers from tweets during pre-processing. Aljuhani et al. used stop words on their dataset of Amazon reviews[2].

Another widely varying aspect of polarity classification is feature selection. Combinations of verbs, adverbs, adjectives, and other statistical approaches such as TF-IDF make many combinations possible. No studies could be found on the use of personal pronouns as features within the Naïve Bayes classifier. However, other POS have been investigated. S. Kausar found that comparative adverbs and general superlative adverbs performed well with the Naïve Bayes classifier [8]. Other challenges include a decrease in accuracy due to negations [6]. Sentiwordnet is a simple way to mitigate some of these feature selection challenges.

The most common machine learning strategy for sentiment analysis within the past decade has been probabilistic [2]. This is partially because probabilistic models are simple and effective [3]. Statistical approaches have a benefit in terms of processing speed [4], [5]. The Naïve Bayes model has performed as well as the support vector machine (SVM) model [10]. Hogenboom et al. found that the Naïve Bayes classifier performed better than the nearest neighbor classifiers with their reduced bag of words approach [5]. However, an issue with machine learning is that models must be retrained in order to fit newer trends. Fang et al have had success with continuous learning framework using a new, fine-tuned Naïve Bayes method [4][10]. However, there is still room for improvement in terms of feature selection. With the careful selection of features for use in reviews, probabilistic methods can become even less computationally expensive.

Wankhade et al. compared the Naïve Bayes classifier with Sentiwordnet 3.0 on Amazon reviews[5]. They found that Naïve Bayes had a slightly better F1 score than Sentiwordnet when classifying between positive and negative reviews. They classified the reviews into either positive or negative categories with 79% accuracy on a sample size of 1000. Their Naïve Bayes model performed at 85.34%. Ohana et al. classified movie reviews into either positive or negative categories with the Sentiwordnet score and determined an accuracy of 65.85%[6] . They found that accuracy could be increased slightly when refining with other feature selection methods. They also found that Sentiwordnet scores used as features performed slightly better at 69.35% accuracy.

There is a lack of research on the use of personal pronouns. In addition, there is room for investigation into the combination of Sentiwordnet score and personal pronoun term frequency (TF). We try to discover if personal pronouns perform well as features with Sentiwordnet and the bag of words model with Naïve Bayes.

III. METHODOLOGY

In this section we define our approach in more detail, including scoring and models. This is followed by the metrics used to comparing results.

A. Pre-processing

The dataset contains 5,000 Amazon reviews and their respective star ratings from data.world[1]. The reviews are mostly on electronics, e-readers, and electronic accessories. The entire dataset has 63 one-star reviews, 54 two-star reviews, 197 three-star reviews, 1208 four-star reviews, and 3489 five-star reviews. To deal with our data, we first apply some common NLP techniques. The non-relevant columns are dropped and only the review item, review text, and star ratings are used. The review text do not include the review title. There are no duplicates and the star ratings range from one to five. This dataset is used because the size and format are manageable for the processing power available.

The Natural Language Toolkit NLTK toolkit is used in performing pre-processing tasks. Scikit-learn is also used for lexical resources, ML methods, and analysis [2], [4]. Using Python's NLTK toolkit, each review is tokenized and tagged with the part of speech (POS). We create two stop word lists. One is pulled from the standard NLTK package and includes personal pronouns. Another is created to exclude the personal pronouns. The following personal pronouns are included: "I," "you," "he," "she," "it." "we," "they," "them," "us," "him," "me," and "her."

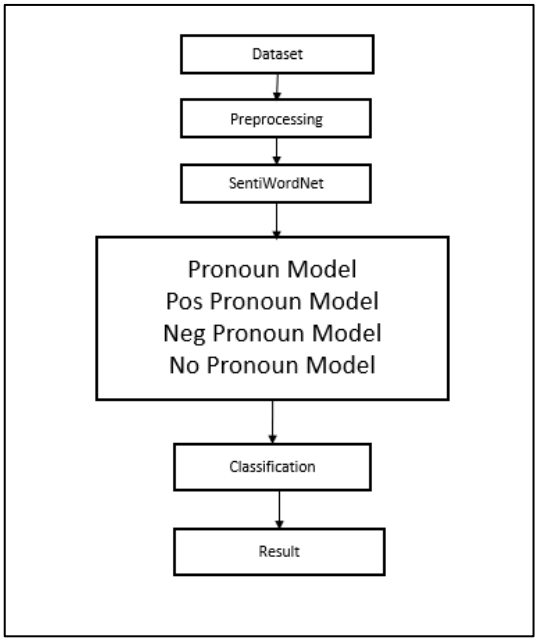

Figure 1: Methodology Process

[1] https://data.world/datafiniti/consumer-reviews-of-amazon-products/workspace/file?filename=Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products.csv

B. Sentiment Detection and Scoring

Sentiwordnet 3.0 is used for sentiment analysis. It uses probabilistic classifiers to create sentiscores. This is a common technique in sentiment analysis [9]. Sentscores measures sentiment in text. It breaks sentences down into synsets and scores them with a positive and negative score (PN scores). The scores range between 0 and 1. PN scores measure the negativity of a text and the positivity in terms of sentiment. The PN scores represent the decisions of several classifiers working in the background of the Sentiwordnet package which considers word relationships and meanings [6]. Sentiwordnet uses a lexical knowledge base for deriving these meanings and relationships [7].

To compute the PN scores, a scoring function is written to identify verbs, adverbs, nouns, adjectives, and pronouns. For each review, the PN scores are aggregated. The function returns PN scores which are then normalized over the length of the review. While Sentiwordnet gives scores to most parts of speech, it does not give a score for personal pronouns. So, we create a custom score for these ranging between 0.3 to 0.9. Depending on the model, these custom scores adjust the aggregate sentiword score for each review into either the positive or negative direction based on frequency of personal pronouns.

C. Model and Evaluation

Using the Scikit-learn toolkit, a term-document matrix is created. The term-document matrix has the frequency of the words within each review. Then the sentiscores are calculated for each review in accordance with the appropriate model and added to the matrix, which can be seen as step three in figure 1. The PN scores of each review from Sentiwordnet are considered features within the bag of words model with Naïve Bayes. In the non-pronoun model, the PN scores are used and pronouns are excluded. In the pronoun model, personal pronouns are not filtered out and they are not given a PN scores through Sentiwordnet. They are included as a TF feature in the bag of words. In the pos-pronoun model, personal pronouns are included just as in the pronoun model, however they are given a small positive score to be included in the PN score feature. In the neg-pronoun model, personal pronouns are included just as in the pronoun model, however they are given a small negative score to be included in the PN score feature.

80% of the dataset is used for training and 20% of the data is used for testing. Once the models have been generated, they are tested on the test dataset. Then the metrics are recorded for the pronoun and non-pronoun model. We sort and extract the most informative features for each class. We find the top ten features for each class and compare the results between the two models.

Precision, recall, macro-averaged accuracy, and F1 score are standard metrics used in measuring text classification. By comparing the statistics, we evaluate the usefulness of personal pronouns as features for classifying reviews into star rating categories. As can be seen in figure 1, classification methods and evaluation are the same for each model, so the pipeline converges. There are various traditional metrics used to evaluate classification, as shown in table 1.

Table 1: Evaluation Metrics

Metric	*Formula*
Precision (P)	$\frac{TP}{(TP+FP)}$
Recall (R)	$\frac{TP}{(TP+FN)}$
F-Measure (F1)	$\frac{2PR}{(P+R)}$
Accuracy	$\frac{(TP+TN)}{(TP+FP+FN+TN)}$

IV. Results and Discussion

A. Pronoun Model vs Non-Pronoun Model

The pronoun model and non-pronoun model have a precision score of 1 for a one-star rating which are shown in table 2. However, the non-pronoun model has a slightly higher recall for a one-star rating. The pronoun model has a score of 0 for both recall and precision for a two star rating. The non-pronoun model has a precision of 1 and similarly low recall. As the rating increases, the pronoun model and non-pronoun model begin to have more similar results.

Table 2: Pronoun Model vs Non-Pronoun Model

Rating	*Pronoun Model*		*Non-Pronoun Model*	
	Precision	*Recall*	*Precision*	*Recall*
1	1.00	0.25	1.00	0.35
2	0.00	0.00	1.00	0.09
3	0.60	0.06	0.57	0.10
4	0.51	0.33	0.46	0.28
5	0.77	0.92	0.75	0.92

However, the pronoun model scores just slightly ahead of the non-pronoun model for four and five-star ratings. On the whole, the pronoun model has a slightly higher accuracy than the non-pronoun model which is shown in table 3.

Table 3: Pronoun Model vs Non-Pronoun Model

Model	*Accuracy*
Pronoun Model	0.73
Non-Pronoun Model	0.71

B. Pos-Pronoun Model vs Neg-Pronoun Model

Sentiwordnet does not assign scores for pronouns. In the following models, we assign a series of scores from 0.3 to 0.9, as shown in table 4. Since the best performance between these weights is 0.3, we will use these models to draw contrasts with the pronoun and non-pronoun models.

Excluding a precision of 1 for the one-star ratings, the neg-pronoun model has poor results. The pos-pronoun model has better precision than the neg-pronoun model in almost every

category. Compared to all the models, the pos-pronoun model has the best precision and recall for the five-star ratings. Similar to the pronoun model in the previous section, the neg-pronoun model has 0 for precision and recall for the two-star rating category. Both models performed relatively well in the five-star rating category.

Table 4: Pos-Pronoun Model vs Neg-Pronoun Model

	Pos-Pronoun Model		*Neg-Pronoun Model*	
Rating	***Precision***	***Recall***	***Precision***	***Recall***
1	1.00	0.10	1.00	0.13
2	1.00	0.10	0.00	0.00
3	0.71	0.08	0.20	0.02
4	0.51	0.32	0.54	0.29
5	0.79	0.95	0.76	0.94

The pos-pronoun model has progressively worse accuracy as the positive score of personal pronouns increase. The same trend can be seen in the neg-pronoun model as shown in table 5. However, the pos-pronoun model has a slightly higher accuracy when the score is set to 0.3 and 0.5. The two models perform the same when assigned a score of 0.7. The worst performance is from the neg-pronoun model with an assigned negative score of 0.9 for each pronoun. Since the best performance between these weights are 0.3, we will use these models to draw contrasts with the pronoun and non-pronoun models.

Table 5: Pos-Pronoun Model vs Neg-Pronoun Model

	Accuracy	
Sentiscore for Pronoun	***Pos-Pronoun Model***	***Neg-Pronoun Model***
0.3	0.75	0.73
0.5	0.73	0.71
0.7	0.71	0.71
0.9	0.71	0.70

Some interesting results are shown in figure 2 where we compare the F1 score at each rating classification for each model. The chart includes the 0.3 pos-pronoun, 0.3 neg-pronoun models, non-pronoun, and pronoun models. All of the models converge around the same F1 score from the four to five-star classes. The pos-pronoun model remains steady around an F1 of 0.2 through the two-star rating, while every other model dips around the two-star rating. The non-pronoun model outperforms the pos-pronoun, neg-pronoun, and pronoun model for star ratings one to three.

C. Discussion

The pos-pronoun model has the highest accuracy than all other models. After that, the neg-pronoun model and pronoun model perform the same. The worst performer is the non-pronoun model with an accuracy of 0.71.

Some of the behavior regarding one, two and three-star reviews can possibly be explained through unbalanced classification. Our data is heavily balanced toward more positive reviews. On average, the test set has 850 positive, five-star reviews where there are only 8-10 two-star reviews in other cases. The pronoun model and the 0.3 neg-pronoun model both score 0 for recall and precision on two-star ratings. These models have 8 and 10 two-star reviews in their test set respectively. The entire dataset skews much toward positive reviews. Although this may be a common distribution for Amazon reviews in general. The organization ultimately has control which reviews are displayed for each product or which reviews are included in an average score. In this case, our models are tested under a more realistic scenario and should then be expected to perform more realistically.

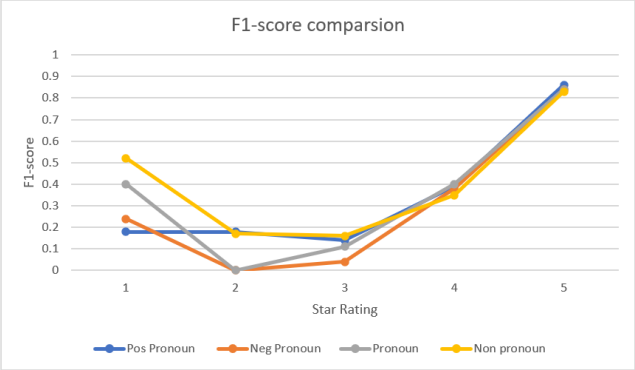

Figure 2: F1-Score Comparison

When considering the F1 score between the pronoun and non-pronoun model, it is important to note that there may be only a one or two review difference between them. So, while the F1 score of the non-pronoun model may look better overall in figure 2, the slight difference in F1 score for the four-star ratings between the pronoun and non-pronoun models most likely contributes more to the overall accuracy score.

It is also important to notice that the difference in F1 scores at the one-star rating level is much closer. There are more one-star reviews in the dataset than two-star reviews. So, this explains the higher accuracy score for the pronoun model versus the non-pronoun model and suggests that the inclusion of the personal pronoun term frequency in the bag of words feature matrix may provide slightly more accurate results when classifying product reviews by polarity. The higher accuracy in the pronoun model also suggests that the repeated use of personal pronouns may give statically significant information to the model. However, the difference is not large enough to draw clear conclusions.

To replicate the sentiment scoring of Sentiwordnet, we add a sentiscore for personal pronouns. For the pos-pronoun model, a positive score was added to the respective review for each personal pronoun that the review contained. For the neg-pronoun model, a negative score is added in the same way. The values for these scores are intended to have a small effect on the overall score for each review. However, the custom sentiscores do not consider the relationships between the words beside them.

The Sentiwordnet score uses the concept of synsets to capture sentiment and has gone through several versions to improve its sentiment capturing ability. Our rudimentary system is designed to test the validity of pronouns as features and not for implementation into other models. As the custom sentiscore values increased, the accuracy decreased. This means the model favors smaller scores for pronouns. This was expected because other parts of speech are well known to be more relevant features for polarity categorization.

For our purposes, the pronoun sentiscores draw contrasts between pronouns as negative or positive words within a review. The pos-pronoun model performs much better than the neg-pronoun model in terms of accuracy, precision, recall, and F1 score. For both models, the smaller score of 0.3 has the greatest effect. This suggests that the multiplicative results of a smaller score led to better predictions under the Naïve Bayes model. This also suggests that reviews with the repetitive use of pronouns are more likely to be positive than negative. However, our model do favor positive reviews. So, this could also only mean that a classifier that skews positive performs better when classifying sentiment in a mostly positive environment.

The neg-pronoun model and pronoun model performed similarly. This may mean that the negative scores on the neg-pronoun model are neutral overall but does not explain why the neg-pronoun model performed worse than the pos-pronoun on a majority of the negative reviews for the one to three star reviews.

When the features are examined for importance within the model, the sentiscores are highest in each instance. The negative sentiscores are the top features for the one and two-star reviews and the positive sentiscores are the second most important feature. The positive sentiscores are the top feature for the three, four, and five-star reviews and the negative sentiscores are the second most important features in these categories.

Other important features for the model included some pronouns such as “you” and “it.” “It” is the top feature for each class in the pronoun model, so it is debatable how much information it actually contributes towards classification between categories. “You” was in the top features for each class, but the position is different. For example, the five-star class has “you” as the ninth top feature but for the two and four-star classes it is the fourth top feature. For one and three-star classes “you” is in the sixth position. So, the classifiers within the Sentiwordnet framework do much of the work in the model followed by some pronouns.

V. Conclusions and Future Work

A. Threats to Validity

Our models are built on a relatively small amount of data within just one subject domain of consumer reviews. The data we use only includes reviews on electronic devices and accessories. There may be bias towards the type of consumers who often buy electronics online. Also, star ratings are not universal across individuals or cultures. One consumer may feel three-stars is an average score where another may feel that two-stars may also represent an average score. Our dataset was not balanced for each star rating. We have many more positive reviews than negative reviews. The custom sentiscores for the pos-pronoun model and neg-pronoun model are assigned at random based on experience. That range might misrepresent the effect pronouns have on polarity classification. This may affect our metrics and skew some our models in an unnatural direction. We use the Naïve Bayes model but there are many other models that can be used to test the use of pronouns and may provide different results.

B. Conclusions

The careful selection of features is an important and growing area of sentiment analysis. It is important to consider subject domain when selecting features. Certain features may be more important in one context over another. We examine features for use in polarity classification of product reviews. When balancing speed and effectiveness, some approaches may be better than others.

We explore sentiscore features, personal pronouns in the bag of words approach using Naïve Bayes with term frequencies (TF), and a custom sentiscore for personal pronouns. The pronoun model includes pronoun TF for each review. The non-pronoun model do not include pronoun TF for pronouns. The pos-pronoun model includes a small custom positive sentiscore for each pronoun added to the review sentiscore and also includes the pronoun TF. The neg-pronoun model includes a small negative custom score and TF. For each model, the greatest predictors are the sentiscores. For the pronoun model, several pronouns are in the top features. The pronoun “it” is the top of each class. However, the pronoun “you” is in unique positions for each class.

The model that gives pronouns a positive sentiscore has the greatest accuracy and the model without pronouns has the worst accuracy. However. unforeseen difficulties with an unbalanced dataset may have introduce error into our exploration. Further study is necessary with these models on larger datasets.

C. Future Work

The pos-pronoun model has the highest accuracy than any other model, but it does not compare well with previous polarity classification models. While the Naïve Bayes model performs well enough with just a few features, other models have been known to have better performance. In the future, personal pronouns should be included in other polarity classification models as a Sentiwordnet score to test if they remain useful. A better sentiword score for personal pronouns may be helpful. Otherwise, at least the inclusion of personal pronoun term frequency into future polarity classification models would be useful. In addition to this, identification of the computation time increase versus benefit of personal pronouns might also be an area of interest.

Source code can be found on github.[2]

[2] https://github.com/JoeKBlank

REFERENCES

[1] S. Shah, K. K, and S. Ra. K, "A novel classification approach based on Naïve Bayes for Twitter sentiment analysis," *KSII Trans. Internet Inf. Syst.*, vol. 11, no. 6, Jun. 2017, doi: 10.3837/tiis.2017.06.011.

[2] S. A. Aljuhani and N. Saleh, "A Comparison of Sentiment Analysis Methods on Amazon Reviews of Mobile Phones," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, 2019, doi: 10.14569/IJACSA.2019.0100678.

[3] A. Hokenboom, F. Boon, and F. Frasincar, "A Statistical Approach to Star Rating Classification of Sentiment," in *Management Intelligent Systems*, vol. 171, Springer Berlin Heidelberg, pp. 242–251.

[4] X. Fang and J. Zhan, "Sentiment analysis using product review data," *J. Big Data*, vol. 2, no. 1, p. 5, Dec. 2015, doi: 10.1186/s40537-015-0015-2.

[5] B. Wankhade and Gupta, Sunil R., "Analysis and Prediction of Customer Reviews Rating using Machine Learning Classifiers and Sentiwordnet," *Infokara Res.*, vol. 8, no. 11, pp. 218–228, 2019.

[6] B. Ohana, "Sentiment Classification of Reviews Using Sentiwordnet," 2009, doi: 10.21427/D77S56.

[7] E. I. Elmurngi and A. Gherbi, "Unfair Reviews Detection on Amazon Reviews using Sentiment Analysis with Supervised Learning Techniques," *J. Comput. Sci.*, vol. 14, no. 5, pp. 714–726, May 2018, doi: 10.3844/jcssp.2018.714.726.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, Not Known, 2002, vol. 10, pp. 79–86. doi: 10.3115/1118693.1118704.

[9] K. Cortis and B. Davis, "Over a decade of social opinion mining: a systematic review," *Artif. Intell. Rev.*, vol. 54, no. 7, pp. 4873–4965, Oct. 2021, doi: 10.1007/s10462-021-10030-2.

[10] J. Casillas, F. J. Martínez-López, and J. M. Corchado Rodríguez, Eds., *Management Intelligent Systems: First International Symposium*, vol. 171. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-30864-2.

[11] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, pp. 117–126, Sep. 2016, doi: 10.1016/j.eswa.2016.03.028.

[12] S. Kausar, X. Huahu, W. Ahmad, M. Y. Shabir, and W. Ahmad, "A Sentiment Polarity Categorization Technique for Online Product Reviews," *IEEE Access*, vol. 8, pp. 3594–3605, 2020, doi: 10.1109/ACCESS.2019.2963020.

[13] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. New York: Cambridge University Press, 2008.

[14] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," *Eur. Lang. Resour. Assoc. ELRA*, vol. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), May 2010, [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

.