Utilizing Topic Modelling in Customer Product Review for Classifying Baby Product

Lay Acheadeth
Computer Science Department,
Faculty of Computing and Media,
Bina Nusantara University,
Jakarta, Indonesia 11480
lay.acheadeth@binus.ac.id

Misa M. Xirinda
Computer Science Department,
Faculty of Computing and Media,
Bina Nusantara University,
Jakarta, Indonesia 11480
misa.xirinda@binus.ac.id

Nunung Nurul Qomariyah
Computer Science Department,
Faculty of Computing and Media,
Bina Nusantara University,
Jakarta, Indonesia 11480
nunung.qomariyah@binus.ac.id

***Abstract*—E-commerce is growing at a breakneck pace. As a result, online shopping has increased, which has resulted in an increase in online product reviews. Often, we come across Amazon products with thousands of reviews, and if we look closely we discover that some of them are completely unrelated to the product. In this study, we conducted a research on how product review classification can assist in resolving the issue of comments on incorrect items. The method used in this research consists of 4 steps which are, data acquisition, data pre-processing, topic modelling, and text classification. Where LDA was used as our topic modelling technique, and for text classification we used Support Vector Machine (SVM), Logistic Regression, and Multi-Layer Perceptron (MLP) classifiers. We found out that by combining both topic modelling and text classification, a powerful tool for handling this kind of problem was developed. Adding the topic modelling can improve the model's accuracy performance from 0.61 to 0.78. So, we can conclude that the topic modelling was useful in classifying the product reviews.**

I. Introduction

The growth and popularity of social media and e-commerce has brought a large amount of data, which is very valuable for companies that want to better understand the behavior of their consumers. However, there are cases where user provides a review on the wrong product, resulting in fault analysis, which is a serious issue for large e-commerce companies like Amazon.

How can we solve the problem? The answer is natural language processing (NLP). Natural language processing is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. Such technologies as NLP can divide the text into components so it could understand the context and intent. The machine can then decide which command to execute - based on the results of the NLP. Ecommerce and retail sectors adopted NLP among the first. It started with chatbots and conversational interfaces and came to automating business processes and enhancing consumer's experience. In this paper, we are using topic modelling, one of the NLP solution, to tackle our challenge.

Topic modelling has increased rapidly over the past years as it becomes more important to businesses, especially e-commerce. When a company wants to understand what is being said about it and the reputation of its products online, one of the ways to do this is using Machine Learning, as well as its sub-area called Deep Learning. Topic modelling is a kind of probabilistic generative model that has been used widely in the field of computer science with a specific focus on text mining and information retrieval in recent years. In technical terms, it is an unsupervised machine learning text mining method, as well as a method of identifying patterns in a cluster. Creating a corpus and running a tool that generates groups of words related to the corpus and distributes them into topics. Topic modelling is also defined as "a method for detecting and tracking clusters of words in large text files".

Since its inception, this model has gotten a lot of attention and piqued the curiosity of academics across a wide range of subjects. Aside from text mining, successful applications have been found in the fields of computer vision, population genetics, and social networks. With topic modelling businesses are able to transfer easy tasks onto robots instead of bombarding their employees with data by employing topic modelling. Consider how much time your staff might save and devote to other essential activities each morning if a computer could filter through endless lists of customer surveys, reviews or support problems.

There are many topic modelling techniques, namely, Latent Dirichlet Allocation (LDA), Non Negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA), Parallel Latent Dirichlet Allocation (PLDA), and, Pachinko Allocation Model (PAM). However, in this paper we only focus on one of them which is LDA [1].

Product classification is being implemented usually either via images or product name. To the best of our knowledge, there is limited research on classifying review to the right product category. The goal of product classification is being able to put products into the right categories. Usually, it is an automated process. For example, scanning the image and save the product information in a particular category. This product classification is common process in large malls or big logistic companies. It may sound simple, but it is not. There are products that could belong to multiple categories. For example, the item "Mens Basketball Sneakers" could be falling under

many categories such as clothing, shoes and accessories, men shoes, and more. Product categorization is important because it strengthens user experience, improves search relevance, and helps customer find the correct products. Many algorithms can be used to execute this task, ranging from classic machine learning algorithm such as Naive Bayes, decision tree, random forest, to the complex deep learning algorithm such as multi class neural network and Long short-term memory (LSTM). The product review on its own can be processed by using Natural Language Process (NLP) pipeline. But, this does not guarantee the good result, because we still capture the general text. Therefore, we decided to combine both topic modelling and product classification to make a more accurate tool, with topic modelling as the inputs, for solving such problems.

II. Literature Review

There have been multiple different studies conducted by different researchers either on how to extract topic from review, sentiment analysis on the review, or classifying product based on product name. There is also one research related to automatic text classification as well. First example is a research from Dalal and Zaveri [2], on automatic text classification. It is a semi supervised machine learning task that automatically assigns a given document to a set of pre-defined categories based on its textual content and extracted features.

The authors in [3] identified various machine learning algorithms that can be used for sentiment analysis on product reviews.These algorithms include Naive Bayes, SVM, Decision Tree KNN, Neural Network, and, Random Forest. They concluded that its necessary to perform the sentiment analysis before launching the new product, as it will help the company know how the customers feel about it before it is released.

Claus and Charles [4] in their research paper about a practical topic modelling technique to exploratory literature review presented an approach not often found in academia. This approach works by using machine learning to analyze literature in order to identify research paths. The frameworks created had some limitations but the results suggest that topic-modelling-exploratory literature reviews have a promising future.

Bergamaschi, Guerra, and Vincini [5] in their research paper about Data Integration Framework for e-Commerce Product Classification presented an approach suggesting the use of a semi-automatic methodology to define the mapping among different e-commerce product classification standards.

A study on the examination of fake news from a viral perspective was done by Krishnadas, Supriya, Sonia, and, Shahid [6]. And the goal of their research was to investigate the factors that have a major impact on the prediction of fake news. To forecast the likelihood of false news, the researchers looked at a combination of emotion-driven content, emotional resonance, topic modelling, and linguistic aspects of news stories. They found that positive emotions in a text reduce the likelihood of false news. It was also shown that sensational information, such as illicit actions and criminal activity, was linked to false news.

In a paper entitled ”Hierarchical multi-label classification using local neural networks”, Cerri et al. [7] extended their previous works, where they investigated a new local-based classification method that incrementally trains a multi-layer perceptron for each level of the classification hierarchy. They conducted a thorough experimental analysis, demonstrating that their method achieves comparable results to a robust global method in terms of precision and recall.

With this studies in mind, we decided to conduct our research as we noticed that a lot of them did not involve classifying the review into the right category, usually only the product name classification. Furthermore, to the best of our knowledge a study that combines topic modelling and classification model together is very limited. Last but not least, there is an automatic text classification review which is similar to our work, but we focus on the e-commerce side rather than the automation side which involves more complex task to handle. For our research to succeed, we combine topic modelling and multi label classification to achieve product review classification into categories which is our standpoint. Our goal is to create a model that is capable of detecting customer review and alert them if they are giving review on the wrong product. Being able to classify review into category means if the review made by customer is in the wrong category from the one model classify, we alert our user. This is one of Amazon's major issue and what they are trying to tackle.

III. Methodology

This section gives an overall work flow of the product review classification via topic modelling for Amazon baby products reviews. Figure 1 shows the workflow of our research.

A. Data Acquisition

In this paper we used a dataset of Amazon baby products reviews which we acquired from Kaggle[1]. The dataset consist of 3 columns and 183,531 consumer reviews on different Amazon baby products. Those 3 columns are name, review, and rating. With this dataset, we explore the possibility of creating a topic modelling model and classify those dominant topic being extracted into the right category after a bit of research. This way, we can make sure all review are being classified into right categories, and if customers has their review did not match with the product, we will alert them. The snippet of the dataset is shown in Figure 2.

B. Data Pre-Processing

1) Data Cleaning: The first step to pre-process the dataset is cleaning the data. We did it by dropping the ”name” column since we implemented topic modelling, hence we stayed with ”rating” and ”review” columns. Afterward, we decided to drop the rating also, and instead adding one additional column which is category. It is blank for now. Then, we decide to minimize the data usage to 300 row only. Finally, we annotated those 300 rows of category's data manually. Then, we found

[1]https://www.kaggle.com/sameersmahajan/reviews-of-amazon-baby-products

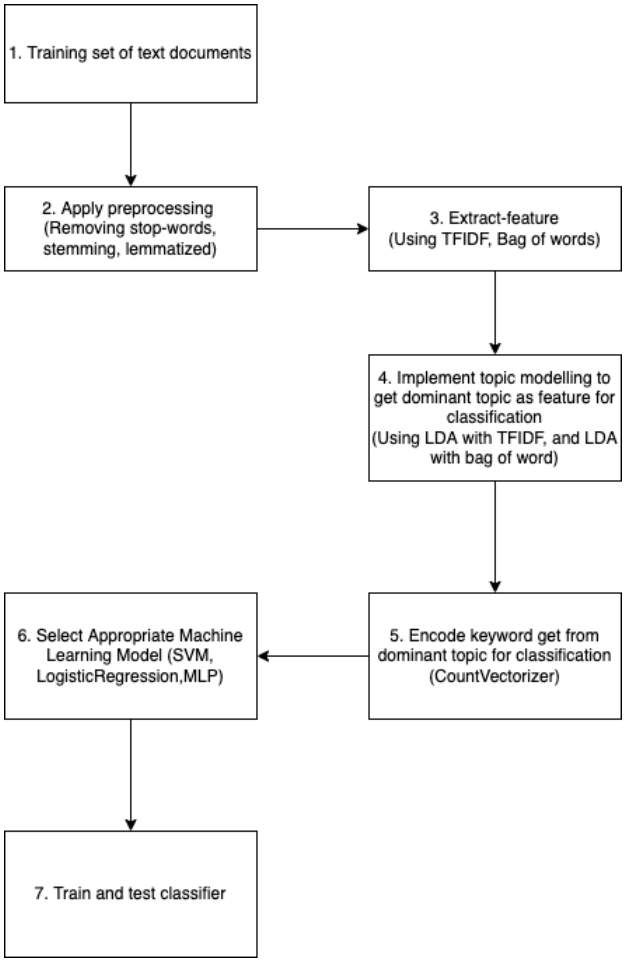

Fig. 1: Research Work Flow

	review	Category	index
0	These flannel wipes are OK, but in my opinion ...	Baby_Care	0
1	it came early and was not disappointed. i love...	Diapering	1
2	Very soft and comfortable and warmer than it l...	Nursery	2
3	This is a product well worth the purchase. I ...	Baby_Care	3
4	All of my kids have cried non-stop when I trie...	Baby_Care	4
...	...	...	...
408	The sweaters are very nice and fit quite well....	Apparel_accessories	408
409	My 13 month old is short and round lol. So I'm...	Apparel_accessories	409
410	This is sp cute and hilarious 100 recomend goo...	Apparel_accessories	410
411	Mr. B is only 4 weeks old and looks so handsom...	Apparel_accessories	411
412	I wanted to dress up my guinea pig but the swe...	Apparel_accessories	412

413 rows × 3 columns

Fig. 2: Dataset Snippet

out that the first 300 rows we annotated manually, the data itself is imbalance. Hence, we try adding more data to balance the data and at last we have 413 rows, with a better balanced data.

2) Tokenization: We split the text into sentences and the sentences into words. And also changed all the letters to lower case and removed punctuations.

3) Stop Words Removal: Then we removed all stop words and words with less than 3 characters.

4) Lemmatization: Words in third person were changed to first person and verbs in past and future tenses were changed into present.

5) Stemming: Words were reduced to their root form.

We then filtered out tokens that appear in less than 15 documents (absolute number) or more than 0.5 documents (fraction of total corpus size, not absolute number).

6) Feature Extraction: We are using two different feature extraction method, make comparison, and use the one with better performance.

First, it is Bag Of Word (BOW) method. The bag-of-words model is a way of representing text data when modelling text with machine learning algorithms. The bag-of-words model is simple to understand and implement and has seen great success in problems such as language modelling and document classification. In this paper, we implement bag of word by creating a variable called bow-corpus. Bow-corpus store the data of all the word that occur in certain review. An example has been shown in Figure 3.

```
Word 13 ("love") appears 1 time.
Word 24 ("book") appears 2 time.
Word 25 ("easi") appears 1 time.
Word 26 ("expect") appears 1 time.
Word 27 ("gift") appears 1 time.
Word 28 ("go") appears 1 time.
Word 29 ("great") appears 1 time.
Word 30 ("help") appears 1 time.
Word 31 ("kid") appears 2 time.
Word 32 ("parent") appears 1 time.
Word 33 ("tri") appears 1 time.
Word 34 ("work") appears 1 time.
```

Fig. 3: Bag of Word (BOW) Result

Secondly, it is term frequency-inverse document frequency (TF-IDF). It is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents. It has many uses, most importantly in automated text analysis, and is very useful for scoring words in machine learning algorithms for Natural Language Processing (NLP). TF-IDF was invented for document search and information retrieval. It works by increasing proportionally to the number of times a word appears in a document, but is offset by the number of documents that contain the word. So, words that are common in every document, such as this, what, and if, rank low even though they may appear many times, since they do not mean much to that document in particular. In this paper, we implement TF-IDF by first converting the bag of word document into TF-IDF in a variable called corpus-tfidf. An example has been shown in Figure 4.

```
Word 13 ("love") is weighted 0.08715637917887972 in the dataset.
Word 24 ("book") is weighted 0.3507049773213626 in the dataset.
Word 25 ("easi") is weighted 0.17980086488264918 in the dataset.
Word 26 ("expect") is weighted 0.3170277293126538 in the dataset.
Word 27 ("gift") is weighted 0.23331291836146587 in the dataset.
Word 28 ("go") is weighted 0.22814285686849048 in the dataset.
Word 29 ("great") is weighted 0.12142737689587782 in the dataset.
Word 30 ("help") is weighted 0.24761156810927779 in the dataset.
Word 31 ("kid") is weighted 0.6101185471094931 in the dataset.
Word 32 ("parent") is weighted 0.2644307339604643 in the dataset.
Word 33 ("tri") is weighted 0.24457566726673635 in the dataset.
Word 34 ("work") is weighted 0.22814285686849048 in the dataset.
```

Fig. 4: TF-IDF Result

C. Models and Technique

We will be using Latent Dirichlet allocation (LDA) for topic modelling. It is a generative statistical model that allow sets of observations to be explained by unobserved group that explain why some parts of the data are similar. It is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions [8].

We divided the task into 2 parts. Before being able to classify the review, we first find the dominant topic within the review via topic modelling with LDA implementation. We used this algorithm with two different inputs, one from the bag of words, and another one from tfidf. After getting the *dominant topic*, we try merging the keyword and the dominant keyword together as the feature data, with category as the y data. Then, we try dividing the data with 0.1 as test size and 0.9 as our train size. Finally, we are classifying those review based on X to the right categories via three algorithms which are Support Vector Machine, Logistic Regression, and Multi-Layer Perceptron.

IV. Results and Discussion

After training our models, we used the sklearn's built in accuracy-score method to determine how accurate our classification models are. Aside from accuracy-score from sklearn, we also used a few different metrics, including but not limited to: precision, recall, and F1 scores. Before determining the accuracy score, we need to have the dominant topic first. We get the dominant topic from LDA algorithm that has both words' vectorized values, by using Bag of Words (BOW) and TF-IDF calculation, as its inputs.

Overall, the result from LDA with TF-IDF perform better with the three algorithms mentioned, with the best accuracy of 0.79 for Multi-Layer Perceptron. The accuracy of the classifiers is shown in Table I.

TABLE I: Accuracy Scores of the 3 classifiers

	Accuracy Score
Multi-Layer Perceptron	0.79
Logistic Regression	0.76
Support Vector Machine	0.71

The result show two different keywords and dominant topic due to different feature extraction. However, LDA model using Tfidf seems to generate a higher accuracy when generating the dominant topic. In addition, when using both dominant topic from LDA with bag of word and LDA with tfidf for classification, we discover that LDA with tfidf performs way better than LDA with bag of words when taking its result to implement multi label classification.

Table II shows the results of the dominant topic created by LDA model using tfidf.

TABLE II: F1 Scores of the 3 classifiers

	MLP	**LR**	**SVM**
Activity_entertainment	0.67	0.75	0.75
Apparel accessories	1.00	1.00	1.00
Baby_Care	0.88	0.86	0.77
Baby_stationary	0.71	0.77	0.62
Baby_toddler_toys	0.71	0.71	0.67
Diapering	1.00	0.86	0.86
Gift	0.75	0.57	0.50
Nursery	0.77	0.71	0.77
macro avg	0.81	0.78	0.74
weighted avg	0.78	0.76	0.71

V. Conclusion and Future Work

We began our research with the goal of creating an accurate classification model for classifying the review. After seeing the result, we realized that we are not yet close to our goal. As seen from our best performance, the accuracy is only 0.78 which is from Multiple-Layer Perceptrons algorithm with TF-IDF as the vectorize technique. That will not be acceptable for working in a real environment. The performance is poor due to the imbalance of our data which we already stated. For our future work we could use Neural Network technique on this since Neural Network has more potential in parameter tuning and improving performance. In addition, we can try adjusting NLP stemming process. Instead of getting all type of word, we should only get the noun and verb. That may yield better result. Finally, we can try to balance the data even more for better model performance.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[2] M. K. Dalal and M. A. Zaveri, "Automatic text classification: a technical review," *International Journal of Computer Applications*, vol. 28, no. 2, pp. 37–40, 2011.

[3] H. Paruchuri, S. Vadlamudi, A. Ahmed, W. Eid, and P. K. Donepudi, "Product reviews sentiment analysis using machine learning: A systematic literature review," *Turkish Journal of Physiotherapy and Rehabilitation*, vol. 23, no. 2, pp. 2362–2368, 2021.

[4] C. B. Asmussen and C. Møller, "Smart literature review: a practical topic modelling approach to exploratory literature review," *Journal of Big Data*, vol. 6, no. 1, pp. 1–18, 2019.

[5] S. Bergamaschi, F. Guerra, and M. Vincini, "A data integration framework for e-commerce product classification," in *International Semantic Web Conference*. Springer, 2002, pp. 379–393.

[6] K. Nanath, S. Kaitheri, S. Malik, and S. Mustafa, "Examination of fake news from a viral perspective: an interplay of emotions, resonance, and sentiments," *Journal of Systems and Information Technology*, 2022.

[7] R. Cerri, R. C. Barros, and A. C. De Carvalho, "Hierarchical multi-label classification using local neural networks," *Journal of Computer and System Sciences*, vol. 80, no. 1, pp. 39–56, 2014.

[8] E. Cambria and B. White, "Jumping nlp curves: A review of natural language processing research," *IEEE Computational intelligence magazine*, vol. 9, no. 2, pp. 48–57, 2014.